









# Image-to-Voxel Model Translation for 3D Scene Reconstruction and Segmentation

Vladimir V. Kniaz<sup>1,2</sup> , Vladimir A. Kniaz<sup>1,2</sup>  , Fabio Remondino<sup>3</sup> ,  
Artem Bordodymov<sup>1</sup> , and Petr Moshkantsev<sup>1</sup> 

<sup>1</sup> State Research Institute of Aviation Systems (GosNIIAS), Moscow, Russia  
{knyaz, v1.kniaz, bordodymov, moshkantsev}@gosniias.ru

<sup>2</sup> Moscow Institute of Physics and Technology (MIPT), Dolgoprudny, Russia

<sup>3</sup> Bruno Kessler Foundation (FBK), Trento, Italy

remondino@fbk.eu

**Abstract.** Objects class, depth, and shape are instantly reconstructed by a human looking at a 2D image. While modern deep models solve each of these challenging tasks separately, they struggle to perform simultaneous scene 3D reconstruction and segmentation. We propose a single shot image-to-semantic voxel model translation framework. We train a generator adversarially against a discriminator that verifies the object's poses. Furthermore, trapezium-shaped voxels, volumetric residual blocks, and 2D-to-3D skip connections facilitate our model learning explicit reasoning about 3D scene structure. We collected a SemanticVoxels dataset with 116k images, ground-truth semantic voxel models, depth maps, and 6D object poses. Experiments on ShapeNet and our SemanticVoxels datasets demonstrate that our framework achieves and surpasses state-of-the-art in the reconstruction of scenes with multiple non-rigid objects of different classes. We made our model and dataset publicly available (<http://www.zefirus.org/SSZ>).

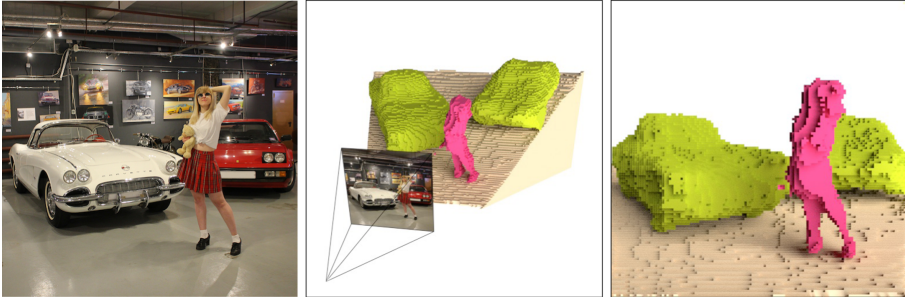
**Keywords:** Single photo 3D reconstruction · 3D semantic segmentation

## 1 Introduction

While humans live and navigate in the 3D world, they reason about it semantically. Given only a class of an object, a human could easily imagine its 3D shape. Object's class, depth, and shape are closely related to each other, and a deep model should reason explicitly about them to truly understand a 3D scene.

There have been exciting recent progress in single image 3D object reconstruction [1–4]. While modern models can reconstruct the human body [5] or arbitrary object [3] from a single view, they are usually focused on the prediction of a single instance of a single object class. Recently proposed multilayer

**Electronic supplementary material** The online version of this chapter ([https://doi.org/10.1007/978-3-030-58571-6\\_7](https://doi.org/10.1007/978-3-030-58571-6_7)) contains supplementary material, which is available to authorized users.



**Fig. 1.** Image-to-semantic voxel model translation using our SSZ model. Input color image (left), 2D-to-3D contour alignment (center), semantic voxel model output (right).

depth maps [6] make a step towards the 3D reconstruction of the whole scene. Still, they do not provide semantic labeling of the 3D scene. On the other hand, 3D scene semantic segmentation models [7] require a 3D model as input.

In this paper, we propose a Single Shot Z-space segmentation and 3D reconstruction model (SSZ) for single image-to-semantic voxel model translation. Different from modern baselines, our SSZ model performs joint 3D voxel model reconstruction and 3D scene semantic segmentation from a single image. Moreover, a modern architecture based on volumetric residual blocks allows our SSZ model to provide near-real-time performance at inference.

We hypothesize that semantic labeling of 3D object classes could aid a deep model learning explicit reasoning about 3D scene structure. To this end, we propose a multiclass semantic voxel model that represents the whole 3D scene visible by the camera. In our semantic voxel model, each voxel holds the ID of its class. Moreover, we leverage trapezium-shaped voxels to keep each voxel aligned with a corresponding pixel (see Fig. 1). Such 3D representation allows us to design direct 2D-to-3D skip connections, that leverage contour correspondences between an image and a 3D model. We use assumptions of Ronneberger et al. [8] and Sandler et al. [9] as a starting point to incorporate a U-net-like generator with inverted residuals blocks and skip connections into our framework.

Generative modeling [10] of 3D shapes has demonstrated promising progress recently [11]. Inspired by adversarial learning of 3D shapes, we incorporate a 3D pose discriminator into our framework. Specifically, we simultaneously train two models: an SSZ generator and an adversarial Pose6DoF discriminator (see Fig. 2). The aim of our Pose6DoF discriminator is twofold. Firstly, it estimates the poses of all object instances in the SSZ generator’s output. Secondly, it qualifies each object instance as either being ‘real’ or ‘fake.’ The aim of our SSZ generator is fooling the discriminator Pose6DoF by producing a realistic and geometrically accurate semantic voxel model.

We collected a large SemanticVoxels dataset to train and evaluate our model and baselines. Our SemanticVoxels dataset includes 116k color images and pixel-level aligned semantic voxel models of nine object classes: person, car, truck, van, bus, building, tree, bicycle, ground.

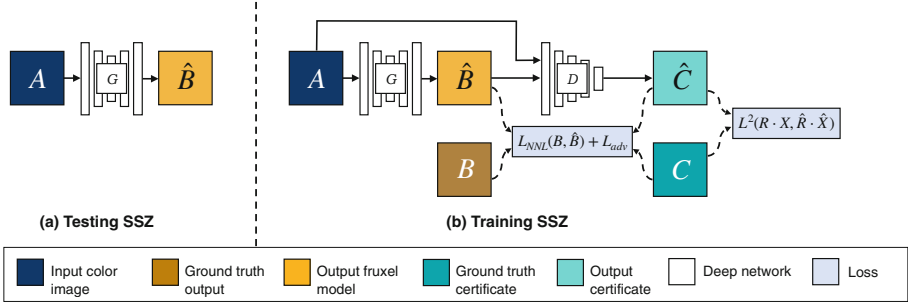


Fig. 2. SSZ framework.

Experiments on our SemanticVoxels dataset and various public benchmarks demonstrate that our SSZ model achieves the state-of-the-art in single-image 3D scene reconstruction. We show quantitative and qualitative results demonstrating our SSZ model ability to reconstruct a detailed voxel model of the whole scene from a single image. Moreover, our SSZ model produces both high-resolution 3D model and multiclass 3D semantic segmentation from a single image.

The developed model will be able to estimate shape, pose, and a class of all objects in the scene in such applications such as autonomous driving, robotics, and single photo 3D scene reconstruction.

We present four key technical contributions: (1) An SSZ generator architecture for single-shot 3D scene reconstruction and segmentation from a single image with 2D-to-3D skip connections and volumetric inverted residual blocks, (2) a generative-adversarial framework for training a volumetric generator against 6DoF pose reasoning discriminator, (3) a large SemanticVoxels dataset with 116k samples. Each sample includes color image, view-centered semantic voxel model, depth map, pose annotations of nine objects classes: person, car, truck, van, bus, building, tree, bicycle, ground, (4) an evaluation of our SSZ model and state-of-the-art baselines on ShapeNet, and our SemanticVoxels dataset.

## 2 Related Work

**Single-Photo 3D Reconstruction.** Deep networks for generation of 3D models from a single photo fall into two groups: object-centered models [12] and view-centered models [2, 3, 6, 13]. Object-centered models [12] reconstruct object 3D model in the same coordinate system for any camera pose with respect to the object. While the object-centered setting is generally easier in terms of data collection and model structure, most of the object-centered models fail to generalize to new object classes. The main reason for this is the absence of explicit reasoning about connections between object shape in the image and the reconstructed 3D shape.

View-centered models [1, 3, 13–15] overcome this problem using paired datasets. Such datasets include a separate 3D model in the camera coordinate

system for each image. The collection of view-centered 3D shape datasets is challenging as the camera pose must be recovered for each image. Still, explicit coding of the camera pose in the dataset allows a model to learn complicated 2D-to-3D reconstruction techniques. Hence, view-centered models are generally more robust to new object classes and backgrounds [13].

Multi-view models [13, 14, 16–18] leverage multiple images of a single object to improve 3D reconstruction accuracy. Related to our semantic frustum voxel models are projective convolutional networks (PCN) [14] that use view-centered frame projection for 3D model reconstruction and segmentation from multiple images. Unlike PCN, our SSZ model uses a view-centered frame during the training time. Closely related to our Pose6DoF discriminator is geometric adversarial loss (GAL) [19] focused on the consistency of reconstructed 3D shapes. Unlike the GAL, our pose adversarial loss function is designed for multiple objects and focused on the scene structure.

**3D Model Representations.** While images are commonly represented as multichannel 2D tensors to train deep models, volumetric 3D shapes are more challenging to incorporate in deep learning pipeline. Therefore 3D reconstruction deep models could be divided into groups by the 3D model representation they use. **Voxel Models** divide object space into equal volume elements that encode probability  $p$  of space being either empty or occupied by an object. While voxel models are the most straightforward data representation for volumetric convolutional neural networks [12, 20–31], they consume large amounts of GPU memory. Hence, the resolution of most modern methods is limited to  $128 \times 128 \times 128$  voxels. **Matryoshka** networks [32] overcome this problem leveraging a memory-efficient shape encoding, which recursively decomposes a 3D shape into nested shape layers. Leveraging the semantic annotations for improving 3D reconstruction accuracy demonstrated promising results recently [33]. **Depth Maps** estimation methods [6, 34–38] are closely connected to 3D model reconstruction. Still, only the visible surface of the object is being reconstructed in such methods. Closely related to our SSZ model is the property of depth maps to preserve contour correspondence between the input image and the reconstructed depth map. This correspondence allows using of skip connections between generator layers [8, 39] to increase model resolution and robustness to new object classes. **Deformable Meshes** allow to use polygonal models for network training [40–48]. While this representation consumes less GPU memory than voxel models, it is best suited for symmetric, smooth objects such as hair [42] or human face [35, 49–53]. The semantic description of the scene at the object level [54] is related to multiclass semantic voxel models in our SSZ model. Similar to our semantic voxel model is 3D-RCNN [55] for instance-level 3D object reconstruction. Unlike 3D-RCNN, our SSZ is a single-shot detector. **Frustum Voxel Models** [56–58] are similar to voxel models but utilize view-oriented projection similar to depth maps. Being designed specifically for single-photo 3D reconstruction, frustum voxel models (fruxel models) can significantly improve model performance for generator with skip connections. In this paper, we extend the fruxel model 3D representations for multiclass 3D scene reconstruction. We train our generator to produce

tensors of  $n \times w \times h \times d$  elements, where  $n$  is the number of classes,  $w, h, d$  number of elements for the width, height, and depth of a fruxel model.

### 3 Method

Our goal is training an SSZ generator  $G : (\mathbf{A}) \rightarrow \mathbf{B}$  translating an input image  $\mathbf{A}$  into a multiclass frustum voxel model of the scene  $\mathbf{F}$ . Specifically, for an input image  $\mathbf{A} \in \mathbb{R}^{w \times h \times 3}$  our model predicts a probability tensor  $\mathbf{B} \in [0, 1]^{n \times w \times h \times d}$ , where  $n$  is the number of classes. Each element in  $\mathbf{B}$  represents a probability  $p(x, y, z)$  of point with coordinates  $(x, y, z)$  belonging to object class  $i$ . We found the resulting fruxel model  $\mathbf{F} \in \{0, 1, \dots, n - 1\}^{w \times h \times d}$  as an arg max of the probability map  $\mathbf{B}$ .

$$\mathbf{F}(x, y, z) = \arg \max_i \mathbf{B}(i, x, y, z). \quad (1)$$

Inspired by generative models for 3D reconstruction, we train two models simultaneously: a generator network  $G$  and an adversarial discriminator  $D$  (see Fig. 2). The aim of our Pose6DoF discriminator  $D : (\mathbf{A}, \mathbf{F}) \rightarrow \mathbf{C}$  is predicting a certificate  $\mathbf{C} \in \{t, q, r\}^{u, v, w}$ , where  $u, v, w$  is dimensions of the discriminator output,  $t \in R^3$  is object translation in the view-centered coordinate frame,  $q \in R^4$  is the object rotation quaternion,  $r \in [0, 1]$  is the probability of object being ‘real’ or ‘fake’. Certificate  $\mathbf{C}$  describes the poses of object instances in the scene and qualifies them as either ‘real’ or ‘fake.’ The aim of our generator  $G$  is generating a realistic and geometrically accurate semantic voxel model  $\mathbf{F}$ . To this end, the objective of our generator  $G$  is maximizing the probability of discriminator  $D$  making a mistake in certificate  $\mathbf{C}$  qualifying a synthesized semantic voxel  $\hat{\mathbf{F}}$  as a real sample  $\mathbf{F}$  from the training dataset. On the other hand, the generator is forced to minimize the error between ground truth object poses  $(t, q)$  and the predicted poses  $(\hat{t}, \hat{q})$ .

Two loss functions govern the training process of our framework: a negative log-likelihood loss  $\mathcal{L}_{NLL}(\mathbf{B}, \hat{\mathbf{B}})$  and a pose adversarial loss  $\mathcal{L}_{adv}(\mathbf{C}, \hat{\mathbf{C}})$ . Inspired by the efficiency of negative log-likelihood loss for the task of 2D semantic segmentation [59], we leverage a similar loss function for our 3D semantic labeling. The aim of our  $\mathcal{L}_{NLL}(\mathbf{B}, \hat{\mathbf{B}})$  loss is maximizing the probability  $p(x, y, z)$  of voxel being labeled with the correct object class

$$\mathcal{L}_{NLL}(\mathbf{B}, \hat{\mathbf{B}}) = \frac{1}{q \cdot w \cdot h \cdot d} \sum_{x=0}^w \sum_{y=0}^h \sum_{z=0}^d \sum_{i=0}^n -k_i \cdot \log \left( \hat{\mathbf{B}}(f, x, y, z) \right), \quad (2)$$

where  $k_i$  is a scalar weight of an object class  $i$ ,  $q = \sum_{i=0}^n k_i$  is the sum of weights for all classes,  $f = \mathbf{F}(x, y, z)$  is the index of the correct object class for point  $(x, y, z)$ ,  $\sum_{f=1}^n \hat{\mathbf{B}}(f, x, y, z) = 1$ . The negative log-likelihood loss introduces a penalty only for voxels, where the predicted class does not equal to the target class. Hence, under such an objective, the voxels representing the empty space of the scene could be filled with any class without any penalty. To avoid such a

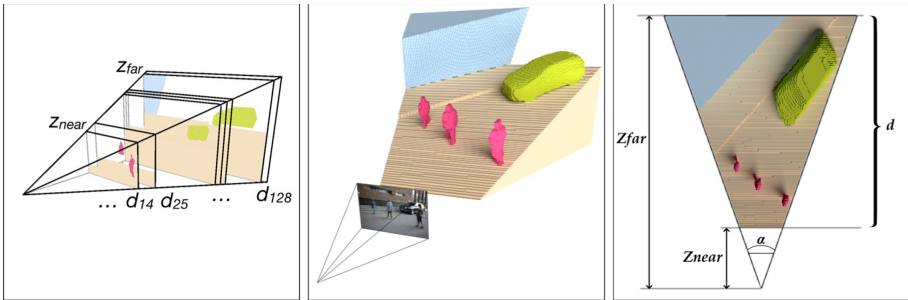
scenario, we use an additional ‘air’ class that forces the loss function to include empty voxels in the training process.

We firstly present our semantic frustum voxel, in Sect. 3.1, and then discuss our SSZ generator in Sect. 3.2. After that, in Sect. 3.3, we introduce our Pose6DoF discriminator that provides the adversarial loss. Finally, we present our SemanticVoxels dataset in Sect. 3.4.

### 3.1 Semantic Frustum Voxel Model

Unlike the rectangular voxel model, the fruxel model leverages trapezium-shaped voxels. The trapezium of each fruxel lies on the ray that connects a pixel on the sensor matrix and a point on an object (see Fig. 3). Let  $I = \{0, 1, \dots, n - 1\}$  be the set of  $n$  classes that the deep model has to predict in the image. Then the semantic voxel model  $F \in \{0, 1, \dots, n - 1\}^{w \times h \times d}$  is a 3D tensor in which each element contains the index  $i \in I$  of the class of an object located in the given fruxel.

To this end, the fruxel model can be regarded as a multilayer 3D semantic segmentation. Each slice is a boolean intersection of an object and a thin box orthogonal to the camera optical axis located at a given distance. A fruxel model can be described by the following set of parameters  $\{z_n, z_f, d, \alpha\}$ , where  $z_n$  is the distance from the camera to the nearest frustum clipping plane,  $z_f$  is the distance to the far clipping plane,  $d$  is the number of slices, and  $\alpha$  is the camera’s horizontal field of view (see Fig. 3).



**Fig. 3. Frustum voxel model:** Slices generation by the boolean intersection of a cutting plane with 3D objects (left). A 3D model composed of trapezium-shaped elements (middle). Top view illustrating fruxel model parameters (right).

### 3.2 SSZ Generator

A defining feature of image-to-voxel translation problems is that they transform high-resolution 2D features to their 3D counterparts. While such translation can be achieved using hidden embedded representations [12], explicit feature translation using skip connections improves model generalization ability. We use

assumptions made by Ronneberger et al. [8] and Sandler et al. [9] as a starting point for our SSZ generator. Namely, we connect the corresponding layers of an encoder and a decoder using skip connections that we term ‘copy-inflate.’

While feature maps in the encoder are 3D tensors  $\mathbf{M}_e \in R^{w \times h \times c}$ , their corresponding feature maps in the decoder are 4D tensors  $\mathbf{M}_d \in R^{w \times h \times d \times c}$ , where  $c$  is the number of channels in a feature map. To match the dimensions, our ‘copy-inflate’ skip connections expand the new dimension by copying  $d$  times 2D slices of each channel in an encoder feature map  $\mathbf{M}_e$ . While the ‘copy-inflate’ connection does not add new information to the expanded feature maps  $\mathbf{M}_d$ , the pixel level contour correspondence between  $\mathbf{M}_e$  and  $\mathbf{M}_d$  allows the model to reason explicitly about relationships between 2D contours and the corresponding 3D shape.

We build the encoder and decoder of our model using inverted residual blocks [60, 61]. This stimulates effective gradient propagation through our model. Moreover, modified inverted residual blocks allow near real-time inference time of the trained model. Each block of the encoder includes inverted residual blocks similar to [61] and an additional pointwise and depthwise convolutions that downscale the feature map.

We use volumetric inverted residual blocks to construct our decoder. Each volumetric inverted residual block includes a volumetric depth separable deconvolution layer followed by a Leaky ReLU activation and a pointwise volumetric convolution. We believe that depth separable convolution in our volumetric inverted residual blocks facilitates learning diverse filters for 2D and 3D features maps. The resulting generator architecture is presented in Fig. 4.

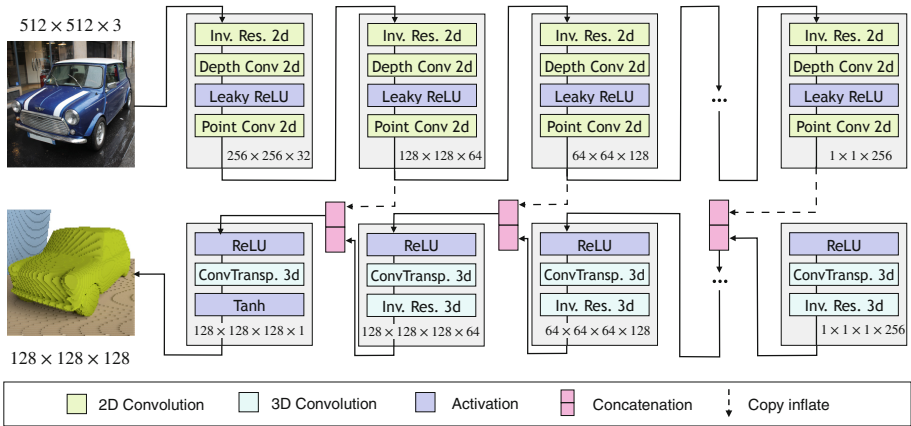


Fig. 4. SSZ generator.

### 3.3 Pose6DoF Discriminator

Our **Pose6DoF** discriminator aims to provide an adversarial loss function focused on the pose accuracy of the objects predicted by our **SSZ** generator. Different from modern volumetric discriminators [11], that qualify the input voxel model as being either ‘real’ or ‘fake,’ our **Pose6DoF** discriminator estimates 6DoF poses of objects in the scene and their perceptual realism. Hence, the architecture of our **Pose6DoF** discriminator fuses a pose estimation model and a discriminator.

We hypothesize that an additional pose term in an adversarial loss will facilitate the accuracy of our **SSZ** generator in terms of depth estimation. During training, our **Pose6DoF** discriminator receives either real fruxel model  $\mathbf{F}$  from the dataset or a generator output  $\hat{\mathbf{F}}$ . The objective of our **Pose6DoF** discriminator is twofold. Firstly, it must detect all instances of objects of all classes and predict their 6DoF poses. Secondly, for each instance it must predict if the instance is ‘real’ or ‘fake.’

We use a PatchGAN discriminator [39] as a starting point for our **Pose6DoF** discriminator. Specifically, our architecture is similar to the encoder part of our **SSZ** generator with 2D convolutions replaced by volumetric convolutions. Our **Pose6DoF** is a conditional discriminator  $D : (\mathbf{A}, \mathbf{F}) \rightarrow \mathcal{C}$  that receives an image  $\mathbf{A}$  and fruxel model  $\mathbf{F}$  concatenated to a single tensor. Given the input  $(\mathbf{A}, \mathbf{F})$  the model predicts a certificate  $\mathcal{C} \in \{t, q, r\}^{u,v,w}$ . The discriminator output’s structure is inspired by single-shot object detection models [62].

The aim of our adversarial loss  $\mathcal{L}_{adv}(G, D)$  is twofold. Firstly, it introduces a penalty for incorrect object poses. Secondly, it penalizes unrealistic 3D object instances predicted by  $G$

$$\begin{aligned} \mathcal{L}_{adv}(G, D) = \mathbb{E}_{\mathbf{F}}[\log D(\mathbf{F})] + \mathbb{E}_{\mathbf{A}}[\log(1 - D(G(\mathbf{A})))] \\ + \sum_{j=0}^m \|R(\hat{q}_j)\hat{t}_j - R(q_j)t_j\|^2, \end{aligned} \quad (3)$$

where  $R(q)$  – is the mapping from quaternion  $q$  to rotation matrix. Please see Supplementary material for details on our **Pose6DoF** discriminator.

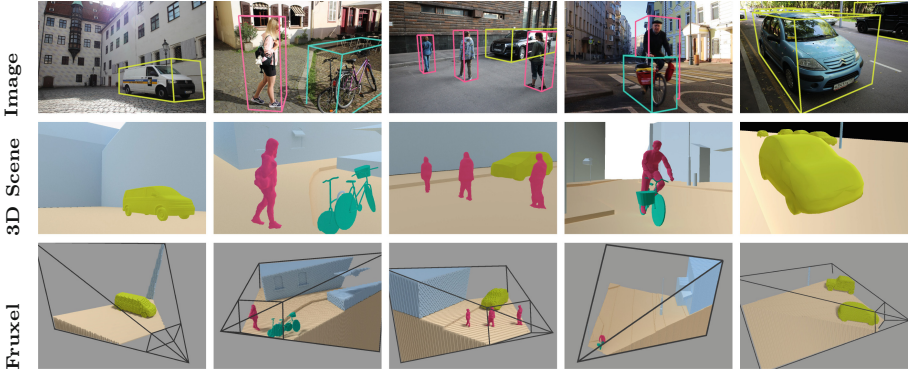
### 3.4 SemanticVoxels Dataset

Our **SemanticVoxels** dataset was inspired by the **VoxelCity** dataset [56]. It includes 116k samples of 3D and 2D data. Each data sample represents a single camera pose. It includes a color image, a semantic frustum voxel model, a depth map, a camera pose, and an object pose annotations for all classes. We used 8k images of 10 street scenes from [56] to increase the diversity of the dataset. **SemanticVoxels** dataset make the following contributions to the **VoxelCity** dataset: (1) 8k new real images of 20 street scenes, (2) 100k synthetic images of 200 scenes, (3) 116k new semantic voxel annotations for 9 object classes

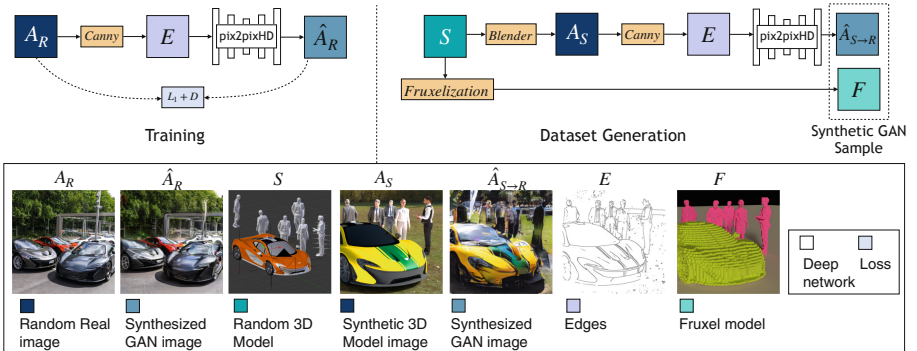
We made our dataset consistent with the **NuScenes** dataset format [63]. Our dataset is divided into two splits: real and synthetic. The real split was generated using a Structure-from-Motion (SfM) technique similar to [64, 65]. It contains



16k images. We present additional details on our SemanticVoxels dataset in the Supplementary material. Example scenes from the dataset are shown in Fig. 5.



**Fig. 5.** Examples of color images with 6D pose annotations and ground truth semantic voxel models from our SemanticVoxels dataset.



**Fig. 6. Synthetic data generation using GAN.** Training pix2pixHD to generate realistic color images from edges (left). Generating paired data samples by rendering a non-realistic 3D model  $A_S$ , calculating its edges  $E$ , and generating a realistic GAN image  $\hat{A}_{S \rightarrow R}$  (right).

**Synthetic Data Generation Using GANs.** Generation of 3D datasets is challenging if it is required to obtain paired images and view-centered 3D models [66]. To overcome this problem, we developed a method based on generative modeling. Inspired by recent advances in generating realistic images from object contours [67–70], we hypothesize that object edges are very similar for real images and non-realistic images generated using the 3D model. Therefore, a ground truth color image for a voxel model could be generated from edges of a 3D model rendered in a non-realistic setup. Our pipeline is presented in Fig. 6.

Firstly, we generate a training dataset from random images of objects of given classes from the COCO dataset [71]. For each real image  $A$ , we generate contours  $E$  using a Canny operator [72]. We train the `pix2pixHD` [67] model on the task of edges-to-image translation.

We generate the dataset samples by creating virtual scenes  $S$  containing 3D models of various classes of objects. For each scene, we render a non-realistic image of the scene  $A_S$  and a corresponding frustum voxel model  $F$ . We extract the edges  $E$  from the image  $A_S$  and generate a realistic color image  $\hat{A}_{S \rightarrow R}$  using the `pix2pixHD` [67] model.

## 4 Experiments

We evaluate our `SSZ` model and baselines on our `SemanticVoxels` dataset, the `ShapeNet` dataset [73], and the `ScanNet` dataset [74]. We train all models on the train split of `ShapeNet` and our `SemanticVoxels` datasets for the tasks of outdoor single photo 3D reconstruction. For the task of 3D `Semantic Scene Completion`, we use train and test splits of `ScanNet` dataset [74]. While our `SSZ` model simultaneously predicts voxel models for  $N$  classes of objects, all baselines predict only single class of object for a single photo. Therefore, we perform per-class accuracy comparison with baselines models. We use 3D Intersection over Union (IoU) metric. Our experiments are threefold. Firstly, we perform a qualitative evaluation to demonstrate rich 3D scene model details and multiclass reconstruction provided by our `SSZ`. Then, we evaluate our model and baselines quantitatively to prove the accuracy of 3D shape and pose of reconstructed 3D models. Finally, we demonstrate the necessity of all components in our `SSZ` model by performing an ablation study.

### 4.1 Baselines

We compare our `SSZ` model to four baselines `DISN` [4], `Pix2Vox` [3], `3D-R2N2` [2] and one 3D semantic scene completion baseline `TS3DSC` [75]. `Deep Implicit Surface Network (DISN)` [4] for high-quality single-view 3D reconstruction predicts a high-quality detail-rich 3D mesh from a single 2D image. The `DISN` model allows capturing the holes in a 3D shape using signed distance fields. `Pix2Vox` [3] exploits an encoder-decoder architecture to generate a coarse 3D volumes and refine them using a fusion block. `3D-R2N2` [2] utilizes a view-based generator that allows tackling single or multiview reconstruction problem. `Two Stream 3D Semantic Scene Completion (TS3DSC)` [75] leverages two stream model that uses the input depth and color modalities to perform semantic segmentation of indoor scenes. We train `DISN`, [3], `3D-R2N2` and our `SSZ` model on train splits of `ShapeNet` and our `SemanticVoxels` datasets. We train `TS3DSC` and our `SSZ` model on train split of `ScanNet`. We test all models on the test split of `ShapeNet`, `ScanNet` and our `SemanticVoxels` datasets.

## 4.2 Training Details

Our SSZ framework was trained on the SemanticVoxels dataset using the PyTorch library [76]. For training on the ShapeNet dataset, we convert ground truth 3D models to fruxel models with parameters  $\{z_n = 3, z_f = 10, d = 128, \alpha = 60^\circ\}$ . For training on the SemanticVoxels dataset, we use fruxel models with parameters  $\{z_n = 2, z_f = 12, d = 128, \alpha = 40^\circ\}$ . The training was performed using the NVIDIA 2080 RTX GPU and took 82h for the ShapeNet dataset and 173h for our SemanticVoxels dataset. For network optimization, we use minibatch SGD with an Adam solver. We set the learning rate to 0.0002 with momentum parameters  $\beta_1 = 0.5, \beta_2 = 0.999$  similar to [39].

## 4.3 Qualitative Evaluation

We evaluate our model and baselines qualitatively by reconstructing 3D scenes with multiple objects from single images. None of the compared baselines can to perform semantic segmentation of the resulting 3D model. Hence, to perform a fair evaluation, we extract a single class from our resulting fruxel model and compare it to the output of baselines. Qualitative results for ShapeNet [73] dataset are presented in Fig. 7. Pix2Vox and 3D-R2N2 models are the best competing baselines demonstrating the correct structure of the 3D shape. While the DISN model attempts to reconstruct the interior structure of the 3D model, its shape differs from the ground-truth model. The voxel model generated by our SSZ framework demonstrates more details and pose correspondence to the input image. The results for our SemanticVoxels dataset are presented in Fig. 8. Unlike the ShapeNet dataset our, SemanticVoxels dataset includes images with multiple objects. During the training stage, we use single-class ground truth 3D models for baselines. We select the 3D model of the object that occupies the largest area in the image. Only the Pix2Vox model can reconstruct the rough shape of the object. We believe that our ‘copy-inflate’ skip connections allow our model to reconstruct 3D scenes with multiple images. For more qualitative results on our SemanticVoxels dataset, see Supplementary material. Qualitative results for ScanNet [74] are given in Fig. 9. While the baseline TS3DSC [75] model receives both depth and color information as an input, our SSZ model still leverages only single color input image. Still our framework outperforms the TS3DSC both in fine details and number of reconstructed object classes.

## 4.4 Quantitative Results

We compare quantitative results in terms of 3D IoU. We present per-class 3D IoU for the ShapeNet dataset in Table 1. Pix2Vox and 3D-R2N2 are the next best performing models after our SSZ model. Pix2Vox model performs the best on plane models and boat models outperforming our model for these classes. Our SSZ model demonstrates the best mean IoU compared to baselines. Quantitative results on our SemanticVoxels dataset demonstrate that our SSZ model successfully reconstructs complex scenes with multiple non-rigid objects of different

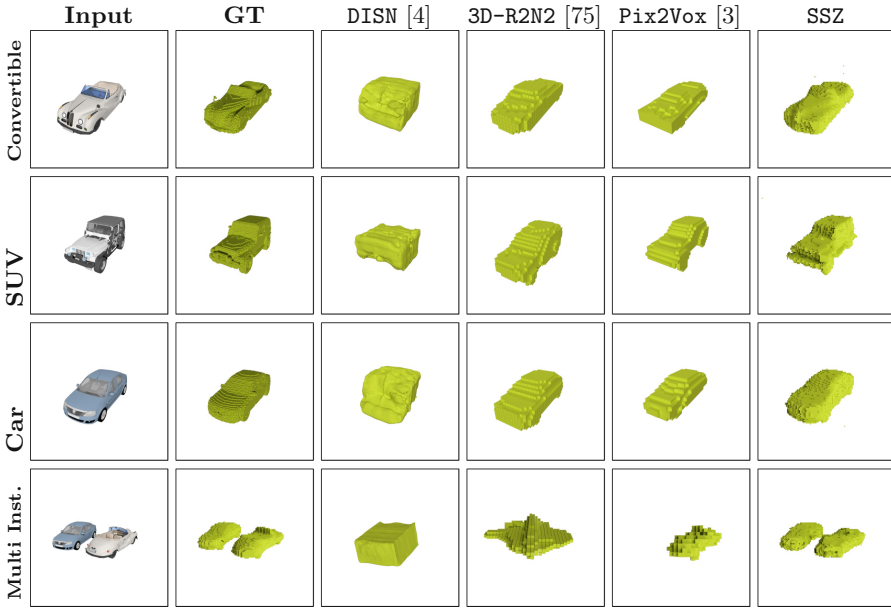


Fig. 7. Examples of 3D reconstruction using DISN [4], Pix2Vox [3], 3D-R2N2 [2], and our SSZ model on ShapeNet [77] dataset. Note that all baselines fail to reconstruct multi instance input images.

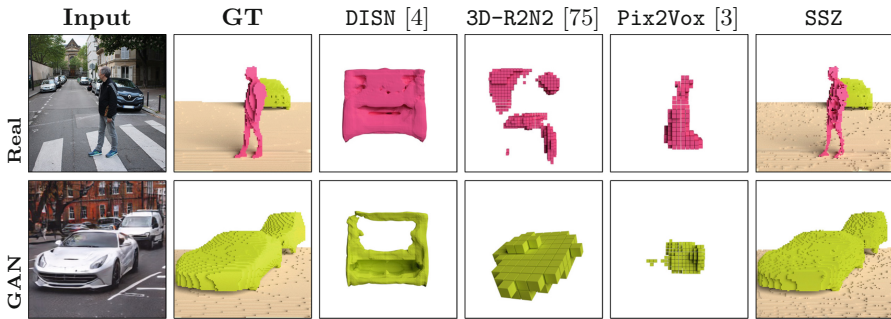
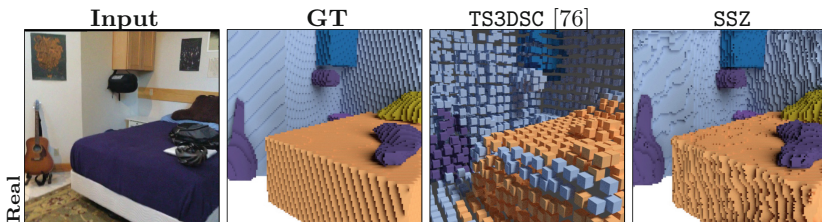


Fig. 8. Example of 3D reconstruction using DISN [4], Pix2Vox [3], 3D-R2N2 [2] and our SSZ on our SemanticVoxels dataset.

classes (see Table 2). 3D-R2N2 is the next best performing model for challenging non-rigid classes such as a human. Pix2Vox model demonstrates the next best results in mean IoU. Our SSZ model demonstrates best results in reconstructing non-rigid objects with complex structures such as humans.



**Fig. 9.** Example of 3D reconstruction using TS3DSC [75] and our SSZ on ScanNet [74] dataset.

**Table 1.** Per-category IoU for different object classes for ShapeNet images.

Object class									
Method	Car	SUV	Conv.	Bike	Bus	Truck	Plane	Boat	Mean
DISN [4]	0.563	0.484	0.427	0.487	0.531	0.522	0.575	0.559	0.519
3D-R2N2 [13]	0.698	0.722	0.515	0.312	0.468	0.455	0.513	0.513	0.525
Pix2Vox [3]	0.732	0.714	0.577	0.356	0.471	0.465	<b>0.598</b>	<b>0.582</b>	0.562
SSZ	<b>0.804</b>	<b>0.745</b>	<b>0.653</b>	<b>0.531</b>	<b>0.562</b>	<b>0.518</b>	0.558	0.539	<b>0.614</b>
SSZ no 6D	0.597	0.586	0.463	0.393	0.412	0.386	0.408	0.395	0.455
SSZ no IR2D	0.682	0.598	0.544	0.433	0.474	0.428	0.457	0.431	0.506
SSZ no IR2D3D	0.594	0.604	0.474	0.429	0.418	0.401	0.441	0.390	0.469

**Table 2.** Per-category IoU for different object classes on our SemanticVoxels dataset.

Object Class										
Method	Person	Car	Van	Build.	Bicycle	Bus	Truck	Tree	Ground	Mean
DISN [4]	0.128	0.270	0.272	0.213	0.171	0.121	0.142	0.178	0.298	0.199
3D-R2N2 [3]	0.225	0.354	0.341	0.214	0.278	0.169	0.138	0.101	0.194	0.224
Pix2Vox [13]	0.140	0.286	0.247	0.306	0.246	0.267	0.256	0.282	0.253	0.254
SSZ	<b>0.618</b>	<b>0.822</b>	<b>0.745</b>	<b>0.585</b>	<b>0.531</b>	<b>0.662</b>	<b>0.518</b>	<b>0.558</b>	<b>0.539</b>	<b>0.620</b>
SSZ no 6D	0.452	0.611	0.538	0.440	0.407	0.495	0.400	0.418	0.380	0.461
SSZ no IR2D	0.502	0.702	0.604	0.458	0.000	0.558	0.432	0.469	0.436	0.462
SSZ no IR2D3D	0.499	0.611	0.607	0.433	0.000	0.505	0.388	0.430	0.438	0.435

#### 4.5 Ablation Studies

We evaluate the necessity of all components of our model by performing 3D scene reconstructions using an ablated version of our model. We firstly remove our Pose6DoF discriminator to check the geometric accuracy of the reconstructed scene (see Fig. 10). The qualitative comparison demonstrates that the ablated version of our model introduces distortions of the scene geometry. Therefore, our pose loss forces the generator to learn to reconstruct invisible parts of an object and their dimensions along the camera’s optical axis.

Secondly, we compare the performance of the SSZ generator without 2D and 3D inverted residual blocks. The ablated version of our model fails to reconstruct textureless objects such as ground and fine shape details. Furthermore, the ablated version could not reconstruct rare object classes such as bicycle (see Table 2). Therefore, all components of our SSZ framework contribute to the accuracy of the trained generator that allows it to achieve the state-of-the-art performance for the task of single-photo 3D reconstruction of multiclass non-rigid objects.

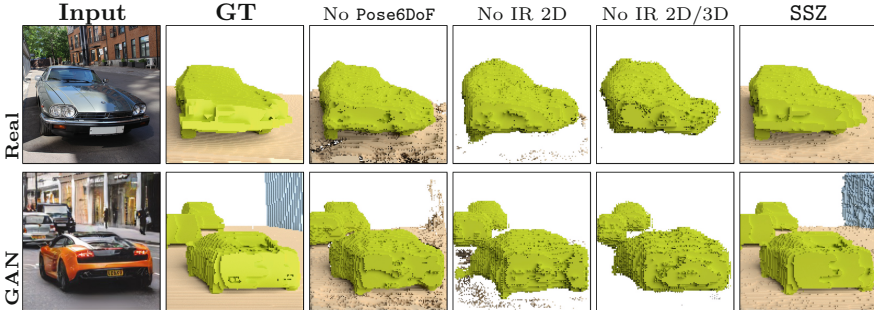


Fig. 10. Evaluation of ablated versions of our SSZ model.

## 5 Conclusions

We demonstrated that volumetric residual blocks could learn reconstruction and segmentation of 3D scenes from a single image. Furthermore, our frustum voxel model 3D scene representation allows using 2D-to-3D skip connections, facilitating the generalization ability of our SSZ model and robust reconstruction of previously unseen objects. Our main observation is that multiclass 3D scene reconstruction and semantic segmentation requires a similar number of model parameters compared to single class image-to-voxel model translation task. Moreover, rich semantic data in the training dataset allows our model to reason explicitly about geometric relationships between object classes.

Compared to state-of-the-art image-to-voxel model translation models, our SSZ framework surpasses leading results in both 3D IoU and pose accuracy for multiclass 3D scene reconstruction. Moreover, our SSZ model is end-to-end trainable. While modern GPUs pose hardware challenges for increasing voxel model resolution, graph convolution networks demonstrate promising results in voxel model super-resolution. The development of a mixed image-to-voxel model with graph convolution super-resolution is an exciting project that requires further work.

**Acknowledgments.** The reported study was funded by Russian Foundation for Basic Research (RFBR) according to the research project N° 17-29-04509.

## References

1. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree generating networks: efficient convolutional architectures for high-resolution 3D outputs. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017, pp. 2107–2115 (2017)
2. Choy, C.B., Xu, D., Gwak, J.Y., Chen, K., Savarese, S.: 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 628–644. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46484-8\\_38](https://doi.org/10.1007/978-3-319-46484-8_38). As references [2] and [75] are same, we have deleted the duplicate reference and renumbered accordingly. Please check and confirm.
3. Xie, H., Yao, H., Sun, X., Zhou, S., Zhang, S.: Pix2Vox: context-aware 3D reconstruction from single and multi-view images. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
4. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: DISN: deep implicit surface network for high-quality single-view 3D reconstruction. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 492–502. Curran Associates, Inc. (2019)
5. Jackson, A.S., Manafas, C., Tzimiropoulos, G.: 3D human body reconstruction from a single image via volumetric regression. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11132, pp. 64–77. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-11018-5\\_6](https://doi.org/10.1007/978-3-030-11018-5_6)
6. Shin, D., Ren, Z., Sudderth, E.B., Fowlkes, C.C.: 3D scene reconstruction with multi-layer depth and epipolar transformers. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
7. Choy, C.B., Gwak, J., Savarese, S.: 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019, pp. 3075–3084 (2019)
8. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28)
9. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks, pp. 4510–4520 (2018)
10. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
11. Wu, J., Zhang, C., Xue, T., Freeman, W.T., Tenenbaum, J.B.: Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In: Advances in Neural Information Processing Systems, pp. 82–90 (2016)
12. Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 484–499. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46466-4\\_29](https://doi.org/10.1007/978-3-319-46466-4_29)
13. Shin, D., Fowlkes, C., Hoiem, D.: Pixels, voxels, and views: a study of shape representations for single view 3D object shape prediction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

14. Kalogerakis, E., Averkiou, M., Maji, S., Chaudhuri, S.: 3D shape segmentation with projective convolutional networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
15. Zhu, R., Kiani Galoogahi, H., Wang, C., Lucey, S.: Rethinking reprojection: closing the loop for pose-aware shape reconstruction from a single image. In: The IEEE International Conference on Computer Vision (ICCV) (October 2017)
16. Leroy, V., Franco, J.-S., Boyer, E.: Shape reconstruction using volume sweeping and learned photoconsistency. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 796–811. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01240-3\\_48](https://doi.org/10.1007/978-3-030-01240-3_48)
17. Sridhar, S., Rempe, D., Valentin, J., Sofien, B., Guibas, L.J.: Multiview aggregation for learning category-specific shape reconstruction. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 2351–2362. Curran Associates, Inc. (2019)
18. Insafutdinov, E., Dosovitskiy, A.: Unsupervised learning of shape and pose with differentiable point clouds. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31, pp. 2802–2812. Curran Associates, Inc. (2018)
19. Jiang, L., Shi, S., Qi, X., Jia, J.: GAL: geometric adversarial loss for single-view 3D-object reconstruction. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11212, pp. 820–834. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01237-3\\_49](https://doi.org/10.1007/978-3-030-01237-3_49)
20. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, W.T., Tenenbaum, J.B.: MarrNet: 3D shape reconstruction via 2.5D sketches. In: Advances In Neural Information Processing Systems (2017)
21. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3D object reconstruction from a single image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
22. Li, K., Pham, T., Zhan, H., Reid, I.: Efficient dense point cloud object reconstruction using deformation vector fields. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11216, pp. 508–524. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01258-8\\_31](https://doi.org/10.1007/978-3-030-01258-8_31)
23. Zhang, X., Zhang, Z., Zhang, C., Tenenbaum, J., Freeman, B., Wu, J.: Learning to reconstruct shapes from unseen classes. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31, pp. 2257–2268. Curran Associates, Inc. (2018)
24. Yang, G., Cui, Y., Belongie, S., Hariharan, B.: Learning single-view 3D reconstruction with limited pose supervision. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11219, pp. 90–105. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01267-0\\_6](https://doi.org/10.1007/978-3-030-01267-0_6)
25. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
26. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
27. Zhou, Y., Tuzel, O.: Voxnet: end-to-end learning for point cloud based 3D object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)



28. Moon, G., Yong Chang, J., Mu Lee, K.: V2V-PoseNet: voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
29. Sitzmann, V., Thies, J., Heide, F., Niessner, M., Wetzstein, G., Zollhofer, M.: DeepVoxels: Learning persistent 3D feature embeddings. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
30. Gadelha, M., Wang, R., Maji, S.: Shape reconstruction using differentiable projections and deep priors. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
31. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: DeepHuman: 3D human reconstruction from a single image. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
32. Richter, S.R., Roth, S.: Matryoshka networks: predicting 3D geometry via nested shape layers. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 1936–1944 (2018)
33. Zhang, D., Han, J., Yang, Y., Huang, D.: Learning category-specific 3D shape models from weakly labeled 2D images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
34. Zheng, C., Cham, T.-J., Cai, J.: T<sup>2</sup>Net: synthetic-to-realistic translation for solving single-image depth estimation tasks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 798–814. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01234-2\\_47](https://doi.org/10.1007/978-3-030-01234-2_47)
35. Feng, M., Gilani, S.Z., Wang, Y., Mian, A.: 3D face reconstruction from light field images: a model-free approach. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 508–526. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01249-6\\_31](https://doi.org/10.1007/978-3-030-01249-6_31)
36. Kumar, S., Dai, Y., Li, H.: Monocular dense 3D reconstruction of a complex dynamic scene from two perspective frames. In: The IEEE International Conference on Computer Vision (ICCV) (October 2017)
37. Zhan, H., et al.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
38. Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., Fan, X.: Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
39. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976. IEEE (2017)
40. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11219, pp. 386–402. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01267-0\\_23](https://doi.org/10.1007/978-3-030-01267-0_23)
41. Shimada, S., Golyanik, V., Theobalt, C., Stricker, D.: IsMo-GAN: adversarial learning for monocular non-rigid 3D reconstruction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
42. Zhou, Y., et al.: HairNet: single-view hair reconstruction using convolutional neural networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 249–265. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01252-6\\_15](https://doi.org/10.1007/978-3-030-01252-6_15)

43. Alp Guler, R., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., Kokkinos, I.: DenseReg: fully convolutional dense shape regression in-the-wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
44. Shi, Y., Xu, K., Nießner, M., Rusinkiewicz, S., Funkhouser, T.: PlaneMatch: patch coplanarity prediction for robust RGB-D reconstruction. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11212, pp. 767–784. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01237-3\\_46](https://doi.org/10.1007/978-3-030-01237-3_46)
45. Wu, J., Zhang, C., Zhang, X., Zhang, Z., Freeman, W.T., Tenenbaum, J.B.: Learning shape priors for single-view 3D completion and reconstruction. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 673–691. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01252-6\\_40](https://doi.org/10.1007/978-3-030-01252-6_40)
46. Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y.: PlaneNet: piece-wise planar reconstruction from a single RGB image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
47. Agudo, A., Pijoan, M., Moreno-Noguer, F.: Image collection pop-up: 3D reconstruction and clustering of rigid and non-rigid categories. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
48. Sinha, A., Unmesh, A., Huang, Q., Ramani, K.: SurfNet: generating 3D shape surfaces using deep residual networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
49. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
50. Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3D face reconstruction with deep neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
51. Tewari, A., et al.: MoFA: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: The IEEE International Conference on Computer Vision (ICCV) (October 2017)
52. Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In: The IEEE International Conference on Computer Vision (ICCV) (October 2017)
53. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: The IEEE International Conference on Computer Vision (ICCV) (October 2017)
54. Huang, S., Qi, S., Zhu, Y., Xiao, Y., Xu, Y., Zhu, S.-C.: Holistic 3D scene parsing and reconstruction from a single RGB image. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 194–211. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01234-2\\_12](https://doi.org/10.1007/978-3-030-01234-2_12)
55. Kundu, A., Li, Y., Rehg, J.M.: 3D-RCNN: instance-level 3D object reconstruction via render-and-compare. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
56. Knyaz, V.A., Kniaz, V.V., Remondino, F.: Image-to-voxel model translation with conditional adversarial networks. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11129, pp. 601–618. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-11009-3\\_37](https://doi.org/10.1007/978-3-030-11009-3_37)
57. Kniaz, V.V., Moshkantsev, P.V., Mizginov, V.A.: Deep learning a single photo voxel model prediction from real and synthetic images. In: Kryzhanovsky, B., Dunin-Barkowski, W., Redko, V., Tiumentsev, Y. (eds.) NEUROINFORMATICS 2019. SCI, vol. 856, pp. 3–16. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-30425-6\\_1](https://doi.org/10.1007/978-3-030-30425-6_1)

58. Kniaz, V.V., Remondino, F., Knyaz, V.A.: Generative adversarial networks for single photo 3D reconstruction. In: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-2/W9, pp. 403–408 (2019)
59. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11209, pp. 432–448. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01228-1\\_26](https://doi.org/10.1007/978-3-030-01228-1_26)
60. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015)
61. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 4510–4520 (2018)
62. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 6517–6525 (2017)
63. Caesar, H., et al.: nuScenes: A multimodal dataset for autonomous driving. arXiv preprint [arXiv:1903.11027](https://arxiv.org/abs/1903.11027) (2019)
64. Locher, A., Havlena, M., Van Gool, L.: Progressive structure from motion. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 22–38. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01225-0\\_2](https://doi.org/10.1007/978-3-030-01225-0_2)
65. Mizginov, V.A., Kniaz, V.V.: Evaluating the accuracy of 3D object reconstruction from thermal images. In: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-2/W18, pp. 129–134 (2019)
66. Sun, X., et al.: Pix3D: dataset and methods for single-image 3D shape modeling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
67. Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 8798–8807 (2018)
68. Kniaz, V.V., Knyaz, V.A., Remondino, F.: The point where reality meets fantasy: mixed adversarial generators for image splice detection. In: Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8–14 December 2019, Vancouver, BC, Canada, vol. 32, pp. 215–226 (2019)
69. Kniaz, V.V., Knyaz, V.A., Hladůvka, J., Kropatsch, W.G., Mizginov, V.: Thermal-GAN: multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11134, pp. 606–624. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-11024-6\\_46](https://doi.org/10.1007/978-3-030-11024-6_46)
70. Kniaz, V.V., Bordodymov, A.N.: Long wave infrared image colorization for person re-identification. In: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-2/W12, pp. 111–116 (2019)
71. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-10602-1\\_48](https://doi.org/10.1007/978-3-319-10602-1_48)

72. Canny, J.F.: A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **8**(6), 679–698 (1986)
73. Chang, A.X., Funkhouser, T.A., et al.: ShapeNet: An information-rich 3D model repository. *CoRR* abs/1512.03012 (2015)
74. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*. IEEE (2017)
75. Garbade, M., Chen, Y., Sawatzky, J., Gall, J.: Two stream 3D semantic scene completion. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, 16–20 June 2019*, pp. 416–425 (2019)
76. Paszke, A., et al.: *Automatic differentiation in PyTorch* (2017)
77. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond PASCAL: a benchmark for 3D object detection in the wild. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)* (2014)