3D-FUTURE: 3D Furniture shape with TextURE

Huan Fu · Rongfei Jia · Lin Gao · Mingming Gong · Binqiang Zhao · Steve Maybank · Dacheng Tao

Received: date / Accepted: date

Abstract The 3D CAD shapes in current 3D benchmarks are mostly collected from online model repositories. Thus, they typically have insufficient geometric details and less informative textures, making them less attractive for comprehensive and subtle research in areas such as high-quality 3D mesh and texture recovery. This paper presents 3D Furniture shape with TextURE (3D-FUTURE): a richly-annotated and large-scale repository of 3D furniture shapes in the household scenario. At the time of this technical report, 3D-FUTURE contains 20,240 clean and realistic synthetic images of 5,000 different rooms. There are 9,992 unique detailed 3D instances of furniture with high-resolution textures. Experienced designers developed the room scenes, and the 3D CAD shapes in the scene are used for industrial production. Given the well-organized 3D-FUTURE, we provide baseline experiments on several widely studied tasks, such as joint 2D instance segmentation and 3D object pose estimation, image-based 3D shape retrieval, 3D object reconstruction from a single image, and texture recovery for 3D shapes, to facilitate related future researches on our database.

Keywords 3D-FUTURE · Furniture Shapes · Textures · Interior Designs · Synthetic Images

1 Introduction

The rapid progress of modern machine learning methods, such as deep neural models, has led to various impressive breakthroughs towards 2D computer vision and natural language processing (NLP). One key to facilitating the advancement of these approaches is the availability of large-scale labeled benchmarks. Mirroring this pattern, the computer graphics and 3D vision communities have put tremendous efforts in establishing 3D datasets over the past years, expecting to enable and innovate the avenues of future research.

Chang et al. [2015], Wu et al. [2015], Xiao et al. [2013], Song et al. [2015], Xiao et al. [2016], Sun et al. [2018a], Xiang et al. [2014, 2016], Silberman et al. [2012], Dai et al. [2017], Hua et al. [2016]. For example, the largest 3D repositories, like ShapeNet (Chang et al. [2015]) and ModelNet (Wu et al. [2015]), collected massive 3D shapes from online repositories and organized them under the WordNet taxonomy. Relying on the repositories, several works, such as Pascal 3D+ (Xiang et al. [2014]), ObjectNet3D (Xiang et al. [2016]), Pix3D
In contrast to the 2D counterparts (Krizhevsky et al., 2012; Lin et al., 2014; Geiger et al., 2012), we realize that there is still a big gap between 3D academic research and industrial productions. For instance, the 3D CAD models in existing datasets mainly come from public online repositories like Trimble 3D Warehouse\(^1\) and Yobi3D\(^2\). These 3D shapes typically have fewer geometry details and uninformative textures or even no textures. Specific to shapes in the household scenario, most of them are outdated and dull furniture deprecated by modern professional designers. Therefore, the current 3D shapes are inadequate for comprehensive and subtle research in areas such as industry closely related fine-grained 3D shape understanding and texture recovery. Besides, existing benchmarks only provide pseudo image or shape alignments, and the estimated camera pose annotations. Namely, the benchmark designers manually choose a roughly matched 3D CAD model from available 3D shape benchmarks according to the object in the image. Thus, annotators may largely ignore some local shape details, which prevents the progress of fundamental data-driven studies such as high-quality 3D reconstruction from real-world images and high-accuracy image-based 3D shape retrieval. Last but not least, there is no well-organized benchmark that offers realistic synthetic indoor images with both instance-level semantic annotations and the involved 3D shapes with textures.

Motivated by the observations, we present 3D Furniture shape with TextURE (3D-FUTURE): a richly-annotated, large-scale repository of 3D furniture shapes specific to the household scenario as shown in Figure 1. At this time, 3D-FUTURE provides 20,240 realistic indoor images and the associated 9,992 unique 3D furniture models with rich geometry details and informative textures. We render these images via one of the most advanced industrial 3D rendering engines based on 5,000 exquisite room scenes developed by experienced designers. The 3D furniture shapes are used for modern industrial productions and have fine-grained geometry and texture related attributes such as category, style, theme, and material. Further, 3D-FUTURE offers instance segmentation annotation and the rendering information, including six degrees of freedom (6DoF) pose and camera field of view (FoV). Apart from these highlight features, another compelling part of 3D-FUTURE is that it enables many fundamental studies and new research opportunities such as furnishing composition, texture recovery, and other interior understanding subjects.

It is, however, nontrivial to collect thousands of aesthetic interior designs. To the best of our knowledge, it takes a designer several days to complete a house’s interior design. Thus, we considered two main research questions when establishing 3D-FUTURE: 1) can we develop a framework that allows creators to design delicate rooms efficiently? 2) can we automatically create some aesthetic designs based on the professional layout information? To investigate the former question, we build a furnishing suit composition (FSC) platform\(^3\). The system recurrently recommends visually matched furniture by considering instance aesthetics and compatibility during the design progress. For the latter one, we reuse the expert layouts, generate multiple furnishing suit candidates with some rules and the FSC approach, render the scene, and manually select visually appealing ones. These AI-created designs will also be reviewed by designers to ensure good quality.

The remainder of this paper is organized as follows. First, we briefly review the public 3D benchmarks and discuss their imperfections. Second, we present the data acquisition process and the FSC pipeline. Third, we introduce the properties and statistics of 3D-FUTURE. Finally, we conduct various experiments leveraging on the properties. These experiments can serve as baselines for subsequent research on 3D-FUTURE.

### 2 Related Work

Lots of 3D benchmarks have been established and made publicly available over the past decades (Chang et al., 2015; Wu et al., 2015; Xiao et al., 2013; Song et al., 2015; Xiao et al., 2016; Sun et al., 2018a; Xiang et al., 2014; 2016; Silberman et al., 2012; Dai et al., 2017; Hua et al., 2016; Choi et al., 2016; Shilane et al., 2004). These datasets can be mainly divided into two groups, including 3D models and RGB-D scenes. We will briefly review some representative 3D benchmarks in the following.

---

1. https://3dwarehouse.sketchup.com
2. https://yobi3d.com
3. https://3d.shejji.com/
2.1 3D Models

One of the large and exhaustively studied 3D shape repositories is ShapeNet \cite{Chang_2015}. It collected millions of raw 3D CAD models from public online repositories such as Warehouse3D and Yobi3D. By re-organizing the datasets, the subsets ShapeNetCore and ShapeNetSem have been made available, including 51,300 and 12,000 models. ShapeNet assigned rich semantic annotations to part of the shapes, such as synsets in the WordNet taxonomy, functional patterns, parts, keypoints, and categories. 3D shape repositories like ModelNet \cite{Wu_2015} and Princeton Shape Benchmark \cite{Shilane_2004} also share similar content as ShapeNet. Several other works like \cite{Cho_2016} and ScanObjectNN \cite{Uy_2019} create the datasets of 3D scans of real objects based on state of the art (SOTA) RGB-D reconstruction approaches. These benchmarks have largely driven the fundamental 3D studies, including 3D representation, 3D shape recognition, 3D object reconstruction, and part segmentation. However, since the 3D shapes are collected online, many may lack geometry details and have dreamlike or no textures.

Relying on these large-scale 3D shape databases, the community also builds benchmarks with image and shape associations to facilitate the research of 3D object understanding from images. For example, PASCAL3D+ \cite{Xiang_2014} and ObjectNet3D \cite{Xiang_2016} aligned objects in the 2D images with the 3D shapes and provided raw 3D pose annotation. Further, Pix3D \cite{Sun_2018a} contributed more accurate 2D-3D alignment for 395 3D shapes of nine object categories. Unluckily, these pseudo alignments may largely ignore some local shape details. Moreover, the expensive labor efforts make it difficult to build a large-scale benchmark with precise pixel-level 2D-3D alignment.

2.2 RGB-D Scenes

In recent years, the community has put significant efforts into building RGB-D datasets to expand researches on 3D scene understanding. For example, NYU Depth V2 \cite{Silberman_2012} captured 464 short Kinect RGB-D sequences from 464 different indoor scenes, where 1,449 images are with dense per-pixel labeling, including depth, surface normal, and semantic labels. SUN RGB-D \cite{Song_2015} followed the pattern by annotating 10,335 RGB-D frames, and offered 3D bounding boxes. To capture the full 3D extent of indoor environments, SUN3D \cite{Xiao_2013} obtained 415 long sequences in 254 unique spaces with comprehensive views. Further, Dai et al. established ScanNet \cite{Dai_2017}, an RGB-D video dataset containing 2.5M views in 1513 scenes annotated with estimated 3D camera poses, surface reconstructions, semantic segmentation, and a broad set of CAD model alignments. Later, a more extensive dataset Matterport3D \cite{Chang_2017} was made publicly available, contributing to panoramic HDR color images with 3D scene annotations. Different from these RGB-D real-world databases, we focus on experienced exquisite interior designs used in industrial productions.
The works most closely related to ours are InteriorNet (Li et al., 2018) and Structured3D (Zheng et al., 2019), which also offer photorealistic images by rendering professional house designs. However, there are two significant differences. First of all, we provide furniture shapes with textures in the scenes. The 6DoF pose and camera FoV are shared in 3D-FUTURE. Second, 3D-FUTURE additionally expects to foster studies of exquisite interior design understanding. Thus, for each room, the camera viewpoints are suggested by designers, so that the captured images contain the whole design idea.

3 Data Acquisition Process

In this section, we introduce the pipeline of our dataset construction procedure. We mainly address the two issues, i.e., designing efficiency and aesthetic design creation, as stated in Sec. 1.

3.1 Large-scale Interior Database

We construct a 3D pool containing a large amount of industrial 3D computer-aided design (CAD) furnishing and interior finish models. We associate each shape
with multiple textures and materials, resulting in enlarged shape repositories. The models are richly annotated with diverse attributes, including theme color, style, material, brand, real-world size, and category in the WordNet taxonomy. There are 500 fine-grained categories in five levels of the taxonomy. High-resolution 2D rendering for each textured model is also available in the database. Based on these objects, hundreds of experienced designers have created \( \sim 60,000 \) decorative houses for different scenarios in several years, where \( \sim 30,000 \) homes have been evaluated as excellent or brilliant designs. A design sample is shown in Figure \[1\]. The large-scale interior data is offered by Alibaba Topping Homestyler\footnote{https://www.tangping.com/} \textsuperscript{4}. We set up a project based on the large-scale interior data to build 3D-FUTURE.

### 3.2 Furnishing Suit Composition (FSC)

One of the main challenges in establishing 3D-FUTURE is how to collect many exquisite interior designs in an acceptable project cycle. To address...
this issue, we develop a high-performing furnishing suit composition framework. We mainly borrow the concepts of attribute-based interpretable compatibility methods (Yang et al., 2019; Wang et al., 2018b; Chen et al., 2019a) in fashion outfit compatibility learning. An overview of this framework is presented in Figure 2.

Our training set for FSC has the form as \( \{X_1, X_2, \ldots, X^N\} \), and \( X^i = \{x^i_1, x^i_2, \ldots, x^i_{m_i}\} \). Here, \( N \) is the total number of experienced house designs. \( m_i \) is the number of items in house \( X^i \), and \( x^i_j \) is a specific furnishing item contained in house \( X^i \). Note that the elements in \( X^i \) are in order, which means \( x^i_j \) is a former furnishing item selected by designers followed by \( x^i_{j+1} \).

### 3.2.1 Deep Visual Embedding

As aforementioned, we have rich attribute annotations for each furnishing item. These interpretable attributes show significance in understanding the item’s content and are thus beneficial to FSC learning. However, we cannot expect a limited number of attributes to represent an item comprehensively. Therefore, we propose a Deep Furnishing Suit Model (DFSM), which consists of a visual embedding network (VEN) and two transformer encoders (TransEnc1 and TransEnc2), to learn representative deep visual embedding leveraging on the large-scale excellent house designs. DFSM is driven by a margin ranking loss with hard sample mining and a variant of classification loss.

In specific, given a furnishing suit \( X^i \), we randomly capture a subset \( X^i_{j-k} = \{x^i_j, x^i_{j+1}, \ldots, x^i_k\} \), where \( 1 \leq j \leq k < m_i \). Our goal is to predict \( x^i_{k+1} \) given \( X^i_{j-k} \). According to the category label of \( X^i_{j-k} \), we randomly choose three negative examples from the furnishing pool to construct a candidate set \( C = \{x^i_{k+1}, z^0_{k+1}, z^1_{k+1}, z^2_{k+1}\} \) in an online manner. We also ensure that the negative examples \( z^0 \), \( z^1 \), \( z^2 \) have the same style / color / material as \( x^i_{k+1} \). We feed both
X_{j-k}^i$ and the candidate images into VEN to extract visual features. In our paper, we take the CNN part of MobileNetV2 followed by a projection layer as VEN, and pre-train it via the unsupervised learning strategy stated in (Wu et al., 2018).

After obtaining the image features, we construct two tasks, i.e., mask prediction and compatibility scoring, based on the impressive transformer architecture in NLP (Devlin et al., 2018; Vaswani et al., 2017). For the former one, we have a sequence of feature vectors $F_i = \{VEN(X_{j-k}^i), [Mask]\}$ with dimension $d$, where $[Mask]$ denotes a particular mask embedding. The task is to predict the masked item given the previous ones. We thus feed $F_i$ into TransEnc1 to capture the enhanced feature $\tilde{F}_i$, and optimize the model via the following loss:

$$L_{mp} = -\frac{1}{N} \sum_{i=1}^{N} \log(\mathcal{P}(x_{k+1}^i | \tilde{F}_i; \Theta, \Phi)),$$

$$\mathcal{P}(x_{k+1}^i | \tilde{F}_i; \Theta, \Phi) = \frac{\exp(\tilde{f}_i^{T} f_{x_{k+1}^i})}{\sum_{c} \exp(\tilde{f}_i^{T} f_{c}^T)},$$

where $\Theta$ and $\Phi$ are the learnable parameters of VEN and TransEnc1, respectively; $c \in C$ is a candidate; $f_{x}^T$ is the transpose of $f_x$; $f_x$ denotes the visual embedding of item $x$, i.e., $f_x = VEN(x)$; and $\tilde{f}_i^{mask}$ represents the feature vector of the [Mask] token from TransEnc1.

For the second task, we take the candidate suits as inputs and directly learn their compatibility scores. Let $F_i(X_{j-k}^i, c)$ be the visual feature vectors of a candidate suit $O_i(X_{j-k}^i, c)$, where $[Start]$ is a particular start token embedding. To estimate the compatibility score of a suit, we need to first capture an embedding that can represent it. We thus employ TransEnc2 to acquire $\tilde{F}_i(X_{j-k}^i, c)$, and use the feature vector of the [Start] token as the representation of suit $O_i(X_{j-k}^i, c)$ (denoted as $r_i(X_{j-k}^i, c)$). Further, we utilize two fully connected layers and a sigmoid function to secure a score ($s_i(X_{j-k}^i, c)$), which is the measure of the quality of the suit. For conventional presentation, $s_i(X_{j-k}^i, c)$ is abbreviated as $s_i(x_{k+1}^i)$ hereafter. Since the ground truth compatibility scores are not available, we minimize a margin ranking loss with a simple hard sample mining policy. The objective is expressed as:

$$L_{cs} = -\frac{1}{N} \sum_{i=1}^{N} \max(0, -s_i(x_{k+1}^i) + s_i(z_{k+1}^i) + \alpha),$$

$$s_i(z_{k+1}^i) = \max(s_i(z_{k+1}^0), s_i(z_{k+1}^1), s_i(z_{k+1}^2)).$$
Fig. 5 Samples of the high-quality 3D shapes and their informative textures in 3D-FUTURE.

Fig. 6 An illustration of decision tree based FSC presented in Sec. 3.2.2. Orange: The visual embedding for each item is obtained from the trained DFSM in Figure 2. Blue: The attribute embedding obtained from the attribute labels for each furniture item.

where $\alpha$ is set to 0.1 in our experiments. The trained visual embedding network (VEN) is used to extract the visual feature for each furniture item.

3.2.2 Decision Tree Based FSC

The main goal here is to infer attribute-based matching patterns, i.e., attribute crosses, for FSC. Considering both interpretability and scalability, we utilize GBDT (Friedman, 2001) to automatically construct attribute crosses as shown in Figure 6. We will not introduce the details of GBDT here, but only present some facts in training the decision trees.

We employ six attributes to represent a specific item, including theme color, style, material, real-world size, the second-level category, and visual information (the learned visual embedding). Here, we denote the learned visual embedding as an attribute. For the discrete attributes (style, material, and the second-level category), we directly convert them to one-hot vectors. For theme color and real-world size, we first adopt k-Means Clustering (Kanungo et al., 2002) to discretize real values and then transform them into one-hot vectors. By further considering the visual embedding, we can represent each item as a feature vector. We assign a label (positive or negative) to each specific furnishing suit to train the decision trees. Both the
positive and negative suits are constructed similarly in Sec. 3.2.1.

3.2.3 Re-training via Hard Sample Mining

The negative furnishing suits construction strategy may return some naive negative samples, due to the large-scale furnishing pool, causing some inaccuracies in both the deep embedding networks and the decision trees. We fine-tune the visual embedding network and re-train the decision tree model via a straightforward hard sample mining strategy to address the issue. Specifically, given $X_{j \sim k}$, we can have the TopK recommendations using the trained DFSM. We then randomly select negative samples from the TopK pool. After the re-training stage, we fix VEN’s parameters and establish an automatically re-training system to update the decision tree model daily using continuously enlarged online designs.

3.3 Topping Homestyler Design Platform

Our DFSM is integrated into the online Topping Homestyler Design Platform[^5] to improve the house design efficiency. There are also other highlight features that can facilitate the design procedure, such as the large-scale shape pool, image-based furniture retrieval, 2D display, 3D Roaming, various professional design templates, texture and item replacement, and online rendering.

3.4 Create Aesthetic Interior Design

We have collected 5,000 exquisite interior room designs in the 3D-FUTURE project. We do not plan to provide several synthetic images in different viewpoints for each

[^5]: http://3d.shejijia.com
Fig. 8 The statistics of the attribute annotations of the 9,992 shapes in 3D-FUTURE. Furniture shapes with the attributes such as “Modern”, “Japanese”, “Smooth Net”, “Texture Mark”, “Rough Cloth”, and “Wood” may be more welcomed by designers when designing the rooms. Besides, except for some special cases, each attribute category has at least 90 shapes.

room. Instead, we expect to deliver more superlative designs to bring more research possibilities. Thus, we take these experienced designs as templates and create several aesthetic interior designs for each template.

For example, given a template room with professional design information, we first replace the interior finishing according to the materials, room style, and other descriptions. Second, we choose a furniture seed (e.g., bed) based on the interior finishing information. Third, we iteratively perform recommendations based on our DFSM and other rules to generate a furnishing list. Finally, we put the items contained in the furnishing list into their corresponding positions. In the third step, we also learn one-to-one visual compatibility models (e.g., bed-nightstand and sofa-coffee table) as additional rules to improve the robustness.

With the pipeline, we can automatically create many interior designs as shown in Figure 3. To ensure the quality, we render an image for each design and manually select 15,240 visually appealing designs. Our experienced designers further review these designs to assure good quality.

The 3D designs are rendered by one of the most advanced computer-generated imagery rendering software applications, V-Ray[6]. To ensure reality, we enable as many functions as possible supported by V-Ray.

4 Properties of 3D-FUTURE

In this section, we summarize the properties of our 3D-FUTURE database. Compared to previous 3D benchmarks, 3D-FUTURE has some prominent properties that can bring more possibilities for future 3D research.

4.1 Photo-realistic Synthetic Images

3D-FUTURE offers 20,240 photo-realistic synthetic images corresponding to 20,240 interior designs. As

[6] https://www.chaosgroup.com/
Fig. 9 The shape number of the 34 categories in 3D-FUTURE. These categories are verified and used by experienced designers in their daily works. The figure also implies the frequency of furniture selected by designers to design the room scenes. There are only 7 dressing chairs because designers commonly choose other chairs as the replacements of dressing chairs when designing a room. For example, Classic Chinese Chair and Chaise Longue Sofa only appear in some special designs.

Fig. 10 The percentile plot of the number of vertices and faces over ShapeNetCore (Chang et al., 2015), ModelNet40 (Wu et al., 2015) and 3D-FUTURE. While other datasets have some extremely low-resolution shapes, 3D shapes in 3D-FUTURE show uniformed distributions on both vertices and faces.

aforementioned, we have 5,000 experienced designs and 15,240 automatically created aesthetic designs. We render one image for each design. Previous datasets, such as Structured3D (Zheng et al., 2019) and InteriorNet (Li et al., 2018), also provide realistic indoor images and scene parsing annotations. However, they put cameras in random positions and capture redundant images for each house. These images were not manually verified, thus suffer from unexpected viewpoints.

In contrast, 3D-FUTURE focuses more on inspiring the understanding of exquisite interior designs. Thus, the camera positions are suggested by professional designers to obtain the best viewpoint for each room. Besides, 3D-FUTURE provides instance semantic labels of 34 categories and ten supper-categories. Moreover, the images contained in 3D-FUTURE are visually more appealing and realistic compared to previous ones.
4.2 2D-3D Alignments

Previous benchmarks only provide pseudo 2D-3D alignment annotations (Xiang et al., 2016; Sun et al., 2018a; Dai et al., 2017; Krause et al., 2013; Xiang et al., 2014). Namely, they manually choose a roughly matched 3D CAD model from public 3D shape benchmarks according to the object contained in an image. Annotators thus may largely ignore some local shape details. As a result, these benchmarks offer a small number of matched 3D shape and 2D image pairs. Besides, previous benchmarks with alignment annotations do not come with scene images. In contrast, 3D-FUTURE provides precious 2D-3D alignments and 3D pose annotations. It contains 9,992 unique 3D shapes and 20,240 scene images. By cropping instances from the scene images, we can further secure 37,441 image and shape pairs with slight occlusions, as reported in Table 3. Some samples are presented in Figure 4.
Fig. 13 An illustration of the network for joint instance segmentation and pose estimation. B: region proposals. C: object recognition. H: network head. M: mask prediction. R&T: pose estimation. Seg: the network in the instance segmentation branch. Pose: the network in the pose estimation branch.

| Category             | MVCNN | PointNet++ |
|----------------------|-------|------------|
| Children Cabinet     | 72.0% | 32.1%      |
| Nightstand           | 75.0% | 71.8%      |
| Bookcase             | 66.7% | 52.3%      |
| Wardrobe             | 56.7% | 82.0%      |
| Coffee Table         | 69.7% | 82.6%      |
| Corner/Side Table    | 64.5% | 74.7%      |
| Side Cabinet         | 49.7% | 65.2%      |
| Wine Cabinet         | 62.9% | 67.1%      |
| TV Stand             | 73.5% | 73.6%      |
| Drawer Chest         | 67.5% | 55.2%      |
| Shelf                | 51.9% | 48.4%      |
| Round End Table      | 52.2% | 75.0%      |
| Double/Queen/King Bed| 78.6% | 91.2%      |
| Bunk Bed             | 57.1% | 77.8%      |
| Bed Frame            | 93.8% | 93.8%      |
| Single Bed           | 69.7% | 68.9%      |
| Kids Bed             | 12.5% | 14.3%      |
| Dining Chair         | 50.5% | 63.9%      |
| Lounge/Office Chair  | 60.3% | 60.5%      |
| Classic Chinese Chair| 57.1% | 62.5%      |
| Barstool             | 32.0% | 66.7%      |
| Dressing Table       | 73.7% | 68.2%      |
| Dining Table         | 84.3% | 61.1%      |
| Desk                 | 54.0% | 20.4%      |
| Three-Seat Sofa      | 71.7% | 82.6%      |
| Armchair             | 72.5% | 68.0%      |
| Loveseat Sofa        | 62.9% | 60.4%      |
| L-shaped Sofa        | 83.3% | 85.9%      |
| Lazy Sofa            | 66.7% | 50.0%      |
| Stool                | 91.9% | 75.8%      |
| Pendant Lamp         | 89.8% | 90.9%      |
| Ceiling Lamp         | 63.0% | 70.7%      |

Table 2 Classification accuracy on 3D-FUTURE. MVCNN: 12 view + ResNet50 backbone. PointNet++: 1024 points + MSG + normal.

4.3 High-quality Shapes with Informative Textures

The 3D shapes contained in previous large-scale shape repositories (Chang et al. 2015; Shilane et al. 2004; Wu et al. 2015) are mainly collected from online repositories. These 3D CAD models usually contain few geometry details and low informative textures. Luckily, 3D-FUTURE provides high-quality 3D furniture shapes with rich details in various styles, including European furniture, which often contains intricate carvings. All the shapes come with informative textures and have been used for modern industrial productions. We show some samples in Figure 5. We believe these features can potentially facilitate innovative research on...
Fig. 15  The retrieval sequences for several query images. 3D-FUTURE contains fine-grained shapes for each furniture category.

high-quality 3D shape understanding and generation. In Figure 10, we compare the proportion of different numbers of vertices and faces over ShapeNetCore (Chang et al. 2015), ModelNet40 (Wu et al. 2015) and our dataset. While other datasets have some extremely low-resolution shapes, 3D shapes in 3D-FUTURE show uniform distributions in both vertices and faces.

4.4 Fine-Grained Attributes

Previous 3D benchmark provides functional attribute annotations in WordNet taxonomy for 3D shapes (Chang et al. 2015). However, these attributes are not well organized and do not have corresponding textures. In contrast, for each textured shape in 3D-FUTURE, we provide four types of attributes verified by professional designers. We have 34 shape categories, 8 super-categories, 19 styles, 15 materials, and 16 themes. These attributes have been demonstrated valuable for interior designs and content understanding by industrial productions. We present the statistics of these attributes in Figure 9 and Figure 8. These figures imply the preferences of experienced modern designers when designing the rooms.

5 Baseline Experiments

In the section, we conduct several baseline experiments by leveraging the properties of 3D-FUTURE, including shape recognition, joint 2D instance segmentation and 3D pose estimation, image-based shape retrieval, 3D object reconstruction, and texture synthesis. We split our 3D shapes into a training set with 6,699 models, and a test set with 3,293 models. The scene images are divided according to the training and test splits of 3D shapes. There are 14,761 images for training and 5,479 images for test. We will briefly present the experimental details for each task and report the scores.

5.1 Fine-grained 3D Object Recognition

Over the past years, most 3D object recognition methods extend deep convolutional neural networks (DCNNs) to modeling 3D data. Because 3D CNNs are too memory intensive (Ji et al. 2012), some researchers prefer to either develop special deep learning operations on point clouds and mesh surfaces (Qi et al. 2017a,b; Hanocka et al. 2019; Feng et al. 2019), or project 3D shapes to several 2D images and then apply 2D convolutional networks (Su et al. 2015).
Fig. 16 Instance segmentation results. The images are captured under suggested viewpoint (by designer) for design exhibition. Zoom in for better view.

| Category            | Top1@R | Top3@R | F-score |
|---------------------|--------|--------|---------|
| Children Cabinet    | 26.9   | 53.4   | 23.7    |
| Nightstand          | 37.4   | 64.1   | 32.6    |
| Bookcase            | 8.7    | 8.7    | 23.9    |
| Wardrobe            | 21.7   | 43.3   | 42.2    |
| Coffee Table        | 27.7   | 57.1   | 31.3    |
| Corner/Side Table   | 32.0   | 49.8   | 32.4    |
| Wine Cabinet        | 4.9    | 7.3    | 22.1    |
| TV Stand            | 16.4   | 27.5   | 37.1    |
| Drawer Chest        | 15.0   | 31.8   | 41.6    |
| Shelf               | 19.4   | 27.3   | 35.6    |
| Round End Table     | 20.0   | 21.1   | 26.1    |
| Double/Queen/King Bed| 23.8  | 48.7   | 78.8    |
| Bunk Bed            | 13.0   | 26.1   | 50.0    |
| Bed Frame           | 26.0   | 47.1   | 52.6    |
| Single Bed          | 16.5   | 34.7   | 63.2    |
| Kids Bed            | 18.2   | 45.5   | 65.5    |
| Dining Chair        | 16.1   | 38.4   | 50.5    |
| Lounge/Office Chair | 33.7   | 64.7   | 56.4    |
| Classic Chinese Chair| 20.0  | 80.0   | 49.8    |
| Barstool            | 52.8   | 58.3   | 22.1    |
| Dressing Table      | 53.9   | 65.4   | 31.8    |
| Dining Table        | 13.9   | 21.5   | 26.0    |
| Desk                | 13.2   | 27.9   | 18.6    |
| Three-Seat Sofa     | 5.0    | 10.5   | 59.8    |
| Armchair            | 45.6   | 68.7   | 56.9    |
| Loveseat Sofa       | 6.7    | 17.5   | 56.1    |
| Lazy Sofa           | 19.4   | 48.4   | 51.7    |
| Chaise Longue Sofa  | 16.7   | 16.7   | 31.4    |
| Stool               | 31.6   | 63.3   | 48.9    |
| Pendant Lamp        | 29.4   | 56.4   | 48.8    |
| Ceiling Lamp        | 31.4   | 59.1   | 37.6    |
| mean                | 23.4   | 40.6   | 47.1    |

Table 4 Numerical retrieval results on 3D-FUTURE for category level. We train a single model and perform retrieval in the full 3D-FUTURE pool. F-score here represents Top5 average F-score.

However, it is nontrivial to extend the projection-based methods to high-resolution 3D scene understanding. Moreover, point and mesh-based approaches suffer from computation bottlenecks and are thus limited to sparse point clouds and a small number of surfaces.

In contrast to ShapeNet (Chang et al., 2015) and ModelNet (Wu et al., 2015), 3D-FUTURE enables the study of fine-grained 3D furniture recognition, which requires the networks to capture more local and global geometric details. Here we consider the well-known MVCNN (Su et al., 2015) and PointNet++ (Qi et al., 2017b) as the baselines. In specific, we train a 12-view MVCNN with ResNet50 as the backbone. For PointNet++, we sample 1024 points for each shape instance and adopt the multi-scale grouping (MSG) strategy (Qi et al., 2017b) and normal vectors to secure the best performance. We train the networks using 6,699 shapes and evaluate the trained models via the remaining 3293 shapes. The classification accuracy for each category is presented in Table 2 and Figure 11. While these methods can reach 90% accuracy on ModelNet40 and ShapenetCore, they do not perform well (69.2% ~ 69.9%) on 3D-FUTURE, due to the presence of fine-grained furniture categories. This observation would motivate researchers to exploit more efficient 3D representation learning approaches for deeper 3D shape analysis.

5.2 Image-based 3D Shape Retrieval

Cross-domain image-based 3D shape retrieval (IBSR) is to identify the CAD models of the objects contained in query images. The primary issue in IBSR is the large appearance gaps between 3D shapes and 2D images. To tackle this challenge, early works made efforts to map cross-domain representations into a
unified constrained embedding space via adaptation techniques such as weight-sharing constraints, metric learning, and distance matching (Li et al., 2015; Aubry et al., 2014; Lee et al., 2018; Massa et al., 2016; Tasse and Dodgson, 2016; Girdhar et al., 2016). Recent works (Sun et al., 2018a; Huang et al., 2018; Wu et al., 2017; Bansal et al., 2016; Bachman, 1978; Grabner et al., 2018 [2019]) predict 2.5D sketches from images, such as surface normal, depth, and location field, to bridge the gaps between 3D and 2D domains. However, the performance of state-of-the-art IBSR methods show a large gap than its 2D counterpart, i.e., content-based image retrieval. This is because there are no large-scale benchmarks that offer large amounts of precious 2D-3D alignment annotations.

In this experiment, we train the baseline using 31,444 image-shape pairs and evaluate the retrieval algorithm via the other 5,994 image-shape pairs. Then we crop the furniture instances with occlusion levels of “NO”, “Slight” and “Standard” from the scene images to produce the image-shape pairs. The statistics of the train and test sets are presented in Table. 5. We develop a DCNN based metric learning network to study the cross-domain shape similarities, as shown in Fig. 17. Specifically, we first project the selected 3D shapes

| Category          | AP  | AR  | AP^100 | AR^100 | AP^75  | AR^75  | AOS  | AVP  | RMSE  |
|-------------------|-----|-----|--------|--------|--------|--------|------|------|-------|
| Children Cabinet  | 0.69| 0.75| 0.83   | 0.86   | 0.79   | 0.83   | 0.53 | 0.54 | 0.38  |
| Nightstand        | 0.68| 0.75| 0.94   | 0.97   | 0.81   | 0.86   | 0.47 | 0.48 | 0.83  |
| Bookcase          | 0.25| 0.41| 0.52   | 0.68   | 0.21   | 0.43   | 0.17 | 0.17 | 0.93  |
| Wardrobe          | 0.80| 0.86| 0.93   | 0.96   | 0.90   | 0.94   | 0.60 | 0.60 | 0.43  |
| Coffee Table      | 0.59| 0.67| 0.94   | 0.96   | 0.64   | 0.75   | 0.41 | 0.40 | 0.26  |
| Corner/Side Table | 0.37| 0.49| 0.74   | 0.82   | 0.32   | 0.49   | 0.30 | 0.29 | 0.29  |
| Side Cabinet      | 0.55| 0.65| 0.81   | 0.88   | 0.60   | 0.71   | 0.43 | 0.43 | 0.28  |
| Wine Cabinet      | 0.58| 0.65| 0.96   | 0.90   | 0.65   | 0.74   | 0.42 | 0.42 | 0.52  |
| TV Stand          | 0.73| 0.79| 0.95   | 0.96   | 0.89   | 0.91   | 0.61 | 0.62 | 0.38  |
| Drawer Chest      | 0.39| 0.77| 0.83   | 0.89   | 0.79   | 0.86   | 0.57 | 0.58 | 0.45  |
| Shelf             | 0.11| 0.35| 0.25   | 0.55   | 0.03   | 0.17   | 0.06 | 0.06 | 0.09  |
| Round End Table   | 0.35| 0.43| 0.80   | 0.84   | 0.20   | 0.37   | 0.32 | 0.32 | 0.22  |
| Double/Queen/King Bed | 0.85| 0.92| 0.95   | 0.98   | 0.91   | 0.96   | 0.68 | 0.69 | 0.24  |
| Bunk Bed          | 0.60| 0.68| 0.85   | 0.93   | 0.78   | 0.88   | 0.45 | 0.46 | 0.21  |
| Bed Frame         | 0.79| 0.87| 0.94   | 0.99   | 0.89   | 0.94   | 0.70 | 0.71 | 0.20  |
| Single Bed        | 0.67| 0.79| 0.79   | 0.88   | 0.73   | 0.84   | 0.60 | 0.61 | 0.28  |
| Kids Bed          | 0.32| 0.59| 0.47   | 0.76   | 0.36   | 0.65   | 0.28 | 0.28 | 0.33  |
| Dining Chair      | 0.41| 0.50| 0.79   | 0.84   | 0.37   | 0.53   | 0.30 | 0.29 | 0.14  |
| Lounge/Office Chair | 0.60| 0.72| 0.85   | 0.93   | 0.71   | 0.83   | 0.49 | 0.49 | 0.43  |
| Dressing Chair    | 0.58| 0.65| 0.92   | 0.94   | 0.68   | 0.78   | 0.11 | 0.11 | 0.07  |
| Classic Chinese Chair | 0.37| 0.40| 0.77   | 0.81   | 0.28   | 0.33   | 0.35 | 0.36 | 0.07  |
| Barstool          | 0.09| 0.19| 0.35   | 0.51   | 0.02   | 0.11   | 0.06 | 0.05 | 0.05  |
| Dressing Table    | 0.43| 0.50| 0.91   | 0.93   | 0.30   | 0.50   | 0.38 | 0.38 | 0.22  |
| Dining Table      | 0.19| 0.32| 0.63   | 0.76   | 0.06   | 0.22   | 0.14 | 0.13 | 0.11  |
| Desk              | 0.28| 0.43| 0.72   | 0.82   | 0.15   | 0.35   | 0.23 | 0.23 | 0.15  |
| Three-Seat Sofa   | 0.73| 0.84| 0.90   | 0.98   | 0.88   | 0.96   | 0.66 | 0.66 | 0.28  |
| Armchair          | 0.79| 0.86| 0.91   | 0.95   | 0.89   | 0.93   | 0.61 | 0.60 | 0.26  |
| Loveseat Sofa     | 0.64| 0.79| 0.77   | 0.91   | 0.73   | 0.87   | 0.49 | 0.48 | 0.35  |
| L-shaped Sofa     | 0.67| 0.79| 0.80   | 0.91   | 0.79   | 0.90   | 0.58 | 0.59 | 0.59  |
| Lazy Sofa         | 0.86| 0.88| 0.90   | 0.91   | 0.89   | 0.91   | 0.81 | 0.82 | 0.18  |
| Chaise Longue Sofa| 0.66| 0.76| 0.84   | 0.91   | 0.84   | 0.91   | 0.56 | 0.52 | 0.42  |
| Stool             | 0.76| 0.82| 0.90   | 0.93   | 0.85   | 0.89   | 0.54 | 0.52 | 0.22  |
| Pendant Lamp      | 0.33| 0.47| 0.70   | 0.79   | 0.28   | 0.47   | 0.22 | 0.21 | 0.21  |
| Ceiling Lamp      | 0.55| 0.65| 0.84   | 0.87   | 0.63   | 0.73   | 0.36 | 0.35 | 0.28  |
| mean              | 0.55| 0.65| 0.79   | 0.84   | 0.58   | 0.69   | 0.43 | 0.43 | 0.30  |

Table 5 Quantitative results of the Cascade-Mask R-CNN baseline for joint instance segmentation and 3D pose estimation. AOS and AVP: Higher is better. RMSE: Lower is better.
Table 6 Numerical comparison of our several baselines for single image 3D reconstruction on our 3D-FUTURE dataset. Metrics are IoU (%), CD ($\times 10^{-3}$), and F-score (threshold is 1%, the reconstruction volume side length defined in (Tatarchenko et al., 2019)). IoU and F-score: Higher is better. CD: Lower is better.

| Category               | IoU (%) | CD ($\times 10^{-3}$) | F-score (%) |
|------------------------|---------|-----------------------|-------------|
|                        | Pixel2Mesh | ONet | DISN | Pixel2Mesh | ONet | DISN | Pixel2Mesh | ONet | DISN |
| Children Cabinet       | 65.54   | 35.96 | 67.50 | 64.00 | 172.36 | 100.56 | 43.19 | 13.76 | 34.96 |
| Nightstand             | 53.42   | 40.75 | 60.10 | 69.95 | 177.66 | 134.67 | 41.01 | 18.60 | 30.86 |
| Bookcase               | 52.74   | 18.15 | 50.89 | 33.04 | 132.13 | 46.14 | 59.03 | 15.76 | 52.06 |
| Wardrobe               | 64.65   | 34.10 | 66.63 | 47.69 | 147.30 | 79.37 | 46.19 | 15.20 | 38.71 |
| Coffee Table           | 42.74   | 57.17 | 41.89 | 58.55 | 165.69 | 92.69 | 49.24 | 17.24 | 39.83 |
| Corner/Side Table      | 38.08   | 17.55 | 42.37 | 51.02 | 213.03 | 123.63 | 59.26 | 14.93 | 35.00 |
| Side Cabinet           | 53.42   | 40.75 | 60.10 | 44.34 | 121.10 | 60.33 | 47.49 | 15.86 | 34.86 |
| Wine Cabinet           | 64.77   | 26.41 | 58.66 | 18.76 | 119.21 | 27.12 | 68.70 | 19.39 | 59.33 |
| TV Stand               | 63.81   | 20.87 | 68.45 | 43.36 | 133.41 | 41.67 | 51.02 | 14.61 | 53.41 |
| Drawer Chest           | 59.99   | 35.19 | 65.12 | 39.93 | 128.75 | 59.92 | 59.55 | 21.19 | 53.30 |
| Shelf                  | 29.71   | 1.62  | 15.59 | 17.47 | 170.99 | 23.86 | 70.46 | 12.03 | 60.87 |
| Round End Table        | 24.88   | 7.49  | 27.15 | 45.24 | 186.17 | 82.00 | 66.47 | 14.93 | 50.10 |
| Double/Queen/King Bed  | 52.63   | 19.02 | 45.85 | 13.08 | 131.75 | 28.42 | 83.82 | 23.74 | 68.75 |
| Bunk Bed               | 39.75   | 23.20 | 35.08 | 19.70 | 132.13 | 28.42 | 69.06 | 12.03 | 50.25 |
| Bed Frame              | 50.92   | 11.20 | 41.58 | 59.03 | 170.99 | 23.86 | 70.46 | 12.03 | 60.87 |
| Single Bed             | 55.32   | 16.70 | 48.72 | 11.97 | 192.22 | 24.52 | 83.82 | 16.47 | 68.04 |
| Kids Bed               | 42.21   | 16.45 | 36.57 | 17.22 | 145.30 | 38.55 | 74.68 | 22.13 | 56.00 |
| Dining Chair           | 40.76   | 15.87 | 40.80 | 11.73 | 100.80 | 30.77 | 86.13 | 27.18 | 69.33 |
| Lounge/Office Chair    | 45.65   | 23.85 | 47.31 | 12.89 | 100.45 | 31.15 | 82.75 | 29.76 | 65.17 |
| Dressing Chair         | 39.47   | 23.20 | 31.63 | 23.22 | 107.77 | 50.03 | 64.97 | 21.79 | 45.94 |
| Classic Chinese Chair  | 23.77   | 13.88 | 39.90 | 21.30 | 108.05 | 50.03 | 71.11 | 24.55 | 50.90 |
| Barstool               | 23.32   | 6.85  | 37.43 | 20.28 | 162.84 | 57.72 | 76.95 | 14.28 | 50.33 |
| Dressing Table         | 42.18   | 18.57 | 44.51 | 29.53 | 152.15 | 49.97 | 64.97 | 21.79 | 45.94 |
| Dining Table           | 43.07   | 10.36 | 40.39 | 56.83 | 171.71 | 85.80 | 45.81 | 16.51 | 42.71 |
| Desk                   | 41.41   | 12.18 | 37.92 | 67.41 | 170.40 | 96.82 | 43.51 | 16.03 | 36.12 |
| Three-Seat Sofa        | 59.60   | 24.39 | 59.06 | 12.16 | 89.77 | 16.24 | 83.81 | 33.43 | 77.42 |
| Armchair               | 51.27   | 33.34 | 50.63 | 16.01 | 87.83 | 33.13 | 76.77 | 32.94 | 59.18 |
| Loveseat Sofa          | 56.53   | 29.01 | 57.14 | 13.31 | 72.93 | 17.77 | 81.55 | 37.11 | 75.03 |
| L-shaped Sofa          | 61.79   | 20.13 | 35.21 | 9.74  | 125.71 | 29.81 | 85.34 | 28.06 | 68.38 |
| Lazy Sofa              | 45.21   | 33.80 | 54.93 | 17.57 | 106.72 | 30.54 | 75.89 | 30.15 | 64.28 |
| Chaise Longue Sofa     | 52.57   | 21.69 | 40.85 | 19.94 | 117.47 | 34.11 | 69.22 | 26.60 | 57.92 |
| Stool                  | 44.74   | 39.51 | 61.92 | 20.82 | 96.18  | 39.55 | 78.61 | 39.51 | 61.92 |
| Pendant Lamp           | 25.37   | 4.73  | 20.64 | 30.52 | 215.87 | 54.45 | 69.97 | 16.73 | 53.06 |
| Ceiling Lamp           | 50.94   | 22.44 | 50.03 | 45.70 | 170.39 | 71.14 | 57.71 | 20.14 | 46.74 |
| Mean                   | 47.32   | 21.32 | 46.50 | 32.35 | 144.76 | 37.33 | 65.69 | 21.42 | 53.46 |

5.3 Jointly 2D Instance Segmentation and 3D Pose Estimation

Image-based 6DoF pose estimation is a fundamental 3D vision task that can benefit many intelligent applications such as autonomous driving, augmented reality, and robotic manipulation. Typical methods 6DoF pose estimation first build point-wise correspondences between 3D models and 2D images, followed by the Perspective-n-Point (PnP) algorithm to compute pose parameters (Collet et al., 2011).
These approaches perform well for objects with rich textures but are not robust to featureless or occluded cases. Recent works thus employ RGB-D sensors and deep learning to improve keypoints detection or directly predict 6DoF pose from images (Kehl et al., 2016; Brachmann et al., 2014; Bo et al., 2014; Hinterstoisser et al., 2012; Xiang et al., 2017; Peng et al., 2019; Song et al., 2020; Tekin et al., 2018; Rad and Lepetit, 2017; Park et al., 2020). Nevertheless, the main issues such as occlusion and clutter, scalability to multiple objects, and symmetries have not been well addressed.

Instance segmentation is the task of detecting and delineating each distinct object of interest appearing in an image. Current instance segmentation methods can be roughly categorized into two paradigms: segmentation-based methods and detection-based methods. The former category of approaches group the predicted category labels via techniques such as clustering (Dhanachandra et al., 2015), metric learning (Fathi et al., 2017), and watershed algorithms (Najman and Schmitt, 1994), to form instance segmentation results. The latter predicts the mask for region instances detected by SOTA object detectors. Methods such as Mask R-CNN series (He et al., 2017; Huang et al., 2019; Cai and Vasconcelos, 2019) have achieved impressive performance for daily objects.

In this experiment, we learn to predict instance segmentation in 2D images and estimate their 6DoF poses in a unified framework. In contrast to the well-studied benchmarks such as ObjectNet3D (Xiang et al., 2016), PASCAL3D+ (Xiang et al., 2014), and Pix3D (Sun et al., 2018a), 3D-FUTURE encourages estimating pose parameters for multiple objects with occlusions in diverse indoor scenes. We provide 3D pose annotations for 100K+ objects in the scene images. The objects are further divided into five occlusion levels, including “NO”, “Slight”, “Standard”, “Heavy”, and “N/A”. Here, an object labeled as “N/A” means that its corresponding 3D shape is not available, or a part of the object is out of the camera view. We train our model on the 14,761 training images and test it on the remaining 5,479 test images.

| Category | Train | Test |
|----------|-------|------|
| Image Shape | Image Shape |
| Sofa     | 49,056 | 8,460 |
| Bed      | 20,032 | 3,912 |
| Chair    | 26,208 | 4,152 |
| Table    | 12,320 | 2,700 |
| Total    | 107,616 | 19,224 |

Table 7: The statistics of the training and test sets for the subject of texture synthesis for 3D shapes.

We modify Cascade Mask-RCNN (Cai and Vasconcelos, 2019; He et al., 2017) as our baseline. The network architecture is shown in Figure. 13. Specifically, we take ResNeXt-101 (Xie et al., 2017) with the setting of 64x4d (group number: 64, width of group: 4) as the backbone, and adopt FPN (Lin et al., 2017)
3D-FUTURE: 3D Furniture shape with TextURE

Fig. 19 An illustration of our BicycleGAN++ baseline. The input are rendered images from 3D shapes. E: Texture Encoder. G: Generator.

To extract the dense features. Then, we utilize a three-stage cascade architecture to perform bounding box regression and object classification. Finally, we add two branches that consist of several fully connected layers to predict the instance masks and their 6DoF poses simultaneously. We cast rotation estimation as a viewpoint classification problem. In detail, we convert the rotation matrices to Euler angles and divide the 360-degree azimuth, 180-degree elevation, and 360-degree in-plane rotation into 18 bins, 9 bins, and 18 bins, respectively. For translation estimation, we use L1 smooth loss to regress the translation parameters directly.

For 2D instance segmentation, we report Average Precision (AP) and Average Recall (AR) over different IoU thresholds following (He et al., 2017). For 3D pose estimation, we take both Average Viewpoint Precision (AVP) in PASCAL3D+ (Xiang et al., 2014) and Average Orientation Similarity (AOS) in KITTI (Geiger et al., 2012) to measure the rotation predictions. In specific, we define the difference between an estimated rotation matrix $R$ and its ground truth $R_{gt}$ as $\nabla(R, R_{gt}) = \frac{1}{\sqrt{2}} \| \log(R^T R_{gt}) \|_F$. In AVP, a correct estimation should satisfy $\nabla(R, R_{gt}) < \frac{\pi}{6}$. The cosine similarity between rotations in AOS is computed as $\cos(\nabla(R, R_{gt}))$.

We present the instance segmentation and pose estimation results in Table 5. Here, the metrics for camera poses are with respect to AP and AR, where the IoU thresholds range from 0.5 to 0.95. For instance segmentation, our baseline captures a mean AP of 0.55 on 3D-FUTURE. The score is at a similar level to those reported on the MSCOCO leaderboard achieved by recent SOTA methods. For 3D pose estimation, our baseline yields a mean AVP of 43%. Besides, as analyzed in (Xiang et al., 2016), AP is an upper bound of AOS. This means the closer AOS is to AP, the more accurate the rotation estimation is. By showing the gaps between AOS and AP in Figure 12, we can see that the estimated rotation (0.43) can be further improved. From the observations, we conclude that most objects’ 3D poses are not well modeled in our challenging setting. This suggests that researchers may need to carefully study 3D pose estimation with different levels of occlusions based on 3D-FUTURE. Some qualitative results are shown in Figure 16 and Figure 17 to further justify our conclusions.

5.4 Single-View 3D Object Reconstruction

Inferring 3D structure from a single image has been an active research area for a long time. In the supervised setting, traditional methods investigated shape from shading (Durou et al., 2008; Zhang et al., 1999) and defocse (Favaro and Soatto, 2005) to reason the visible parts of objects. Leveraging on large-scale shape repositories, various works examined deep architectures to produce shapes in 3D volume (Choy et al., 2016), point cloud (Fan et al., 2017), and mesh surface (Groueix et al., 2018) directly. Recently, several SOTA methods recovered 3D meshes from initializations using shape deformation based on deep networks (Wang et al., 2018a). In the unsupervised setting, 3D recovery has been recast as a 2D image reconstruction progress of unobserved views with differentiable rendering (Liu et al., 2019; Chen et al., 2019b).

In this paper, we examine several SOTA reconstruction algorithms as the baselines, including
| Category | Texture Field | BicyleGAN++ |
|----------|---------------|-------------|
|          | FID | SSIM | L1 | Feat1 | FID | SSIM | L1 | Feat1 |
| Sofa     | 22.01 | 0.959 | 0.013 | 0.168 | 10.01 | 0.951 | 0.019 | 0.146 |
| Bed      | 37.22 | 0.924 | 0.024 | 0.190 | 18.06 | 0.916 | 0.030 | 0.172 |
| Chair    | 15.36 | 0.951 | 0.017 | 0.131 | 10.65 | 0.941 | 0.022 | 0.120 |
| Table    | 29.45 | 0.964 | 0.011 | 0.149 | 21.78 | 0.958 | 0.016 | 0.137 |
| mean     | 26.01 | 0.952 | 0.016 | 0.160 | 15.12 | 0.942 | 0.022 | 0.144 |

Table 8 Quantitative Evaluation using the FID, SSIM, L1, and Feat1 metrics. FID, L1, Feat1: lower is better. SSIM: higher is better.

Fig. 20 The multi-view texture synthesis results. Top: Texture Fields (Oechsle et al., 2019). Bottom: Our BicycleGAN++ based on BicycleGAN (Zhu et al., 2017).

5.5 Texture Synthesis For 3D Shapes

Unlike geometry reconstruction, texture reconstruction of 3D objects has received less attention from the community. Previous works studied the subject by learning colored 3D reconstruction on voxels or point clouds (Sun et al., 2018b; Tulsiani et al., 2017) based on view synthesis and multi-view geometry. While voxel representations are limited to the low resolutions, point representations are sparse and thus ignore geometric details. Recent approaches alternatively learned a 2D texture atlas (UV mapping) for 3D meshes to map a point on the shape manifold to a pixel in the texture atlas. These methods mainly take advantage of differentiable rendering to recast the problem as an unobserved view synthesis problem (Raj et al., 2019; Oechsle et al., 2019).

Existing 3D repositories contain less dreamlike or uninformative textures and cannot support high-quality texture recovery studies. In contrast, 3D-FUTURE provides furniture shapes with informative textures, which are widely used in industrial productions. We examine two baselines for texture synthesis, i.e., Texture Fields (Oechsle et al., 2019) and a novel BicycleGAN++ method. Here,
BicycleGAN++ extends BicycleGAN [Zhu et al., 2017] for texture synthesis. An illustration of the network is shown in Figure 19. In specific, we incorporate a texture encoder such that the learned model can perform controllable texture synthesis. Importantly, by enlarging the weights of the reconstruction losses and introducing a texture consistency loss, we find that the produced multi-view textured images will show preferable consistency in overlap regions.

We conduct experiments on four super-categories, including Sofa, Bed, Chair, and Table. The details of our train and test splits are reported in Table 7. We randomly render 32 views of images for each shape to enlarge the training set. For each baseline, we first train them on the whole train set and then perform category-specific fine-tuning. Following [Oechsle et al., 2019], we use structure similarity image metric (SSIM) [Wang et al., 2004], L1, Frechet Inception Distance (FID) [Heusel et al., 2017], and Feat1 as our metrics to evaluate the quality of the synthetic texture. Here, L1 is the L1 distance between the ground-truth view rendering and the produced textured image under the same viewpoint. Feat1 is a global perceptual measure operated on the Inception-net [Szegedy et al., 2015] feature space using the L1 distance. As shown in Table 8 while BicycleGAN++ earns higher scores on FID and Feat1, Texture Fields performs better in terms of SSIM and L1, indicating that BicycleGAN++ produces more realistic images with higher quality and Texture Fields focuses more on structured texture details. We also give some qualitative results in Figure 20 and Figure 21. We can see that BicycleGAN++ can only learn the main color information while largely ignores the semantic parts of objects. Texture Fields can partially preserve the structured texture details but produces dreamlike textures. These observations demonstrate that achieving visually appealing texture recovery for 3D meshes is still very challenging, especially for the industrial 3D shapes with informative texture details.

6 Conclusion

In this paper, we have built the large-scale 3D-FUTURE benchmark specific to the household scenario with rich 3D and 2D annotations. 3D-FUTURE contains 20,240 realistic synthetic images and 9,992 high-quality 3D CAD furniture shapes. The exciting features include but are not limited to the exhausting interior designs by experienced designers, photo-realistic renderings, 2D-3D alignments, and most significantly the industrial 3D furniture shapes with informative textures. We conduct several experiments to show the remarkable properties of 3D-FUTURE. The experiments can serve as baselines for future research using our database. We hope that 3D-FUTURE can facilitate innovative research on high-quality 3D shape understanding and generation, bring new research
opportunities for 3D vision, and build a bridge between academic study and 3D industrial applications.

References

Aubry M, Maturana D, Efros AA, Russell BC, Sivic J (2014) Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3762–3769

Bachman CW (1978) Data processing system utilizing data field descriptors for processing data files. US Patent 4,068,300

Bansal A, Russell B, Gupta A (2016) Marr revisited: 2d-3d alignment via surface normal prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5965–5974

Bo L, Ren X, Fox D (2014) Learning hierarchical sparse features for rgb-(d) object recognition. The International Journal of Robotics Research 33(4):581–599

Brachmann E, Krull A, Michel F, Gumhold S, Shotton J, Rother C (2014) Learning 6d object pose estimation using 3d object coordinates. The International Journal of Robotics Research 33(4):581–599

Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, Savarese S, Savva M, Song S, Su H, et al. (2015) Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:151203012

Chen W, Huang P, Xu J, Guo X, Guo C, Sun F, Li C, Pfader A, Zhao H, Zhao B (2019a) Pog: Personalized outfit generation for fashion recommendation at alibaba ifashion. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 2662–2670

Chen W, Ling H, Gao J, Smith E, Lehtinen J, Jacobson A, Fidler S (2019b) Learning to predict 3d objects with an interpolation-based differentiable renderer. In: Advances in Neural Information Processing Systems, pp 9609–9619

Choi S, Zhou QY, Miller S, Koltun V (2016) A large dataset of object scans. arXiv preprint arXiv:160202481

Choy CB, Xu D, Gwak J, Chen K, Savarese S (2016) 3D-R2N2: A unified approach for single and multi-view 3d object reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV)

Collet A, Martinez M, Srinivasa SS (2011) The moped framework: Object recognition and pose estimation for manipulation. The international journal of robotics research 30(10):1284–1306

Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Nießner M (2017) Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5828–5839

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

Dhanachandra N, Manglem K, Chau YJ (2015) Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. Procedia Computer Science 54:764–771

Durou JD, Falcone M, Sagona M (2008) Numerical methods for shape-from-shading: A new survey with benchmarks. Computer Vision and Image Understanding 109(1):22–43

Fan H, Su H, Guibas LJ (2017) A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 605–613

Fathi A, Wojna Z, Rathod V, Wang P, Song HO, Guadarrama S, Murphy KP (2017) Semantic instance segmentation via deep metric learning. arXiv preprint arXiv:170310277

Favaro P, Soatto S (2005) A geometric approach to shape from defocus. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3):406–417

Feng Y, Feng Y, You H, Zhao X, Gao Y (2019) Meshnet: Mesh neural network for 3d shape representation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 8279–8286

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of statistics pp 1189–1232

Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 3354–3361

Girdhar R, Fouhey DF, Rodriguez M, Gupta A (2016) Learning a predictable and generative vector representation for objects. In: European Conference on Computer Vision, Springer, pp 484–499

Grabner A, Roth PM, Lepetit V (2018) 3d pose estimation and 3d model retrieval for objects in the wild. In: Proceedings of the IEEE Conference
Grabner A, Roth PM, Lepetit V (2019) Location field descriptors: Single image 3d model retrieval in the wild. In: 2019 International Conference on 3D Vision (3DV), IEEE, pp 583–593

Groueix T, Fisher M, Kim VG, Russell B, Aubry M (2018) AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)

Hanocka R, Hertz A, Fish N, Giryes R, Fleishman S, Cohen-Or D (2019) Meshcnn: a network with an edge. ACM Transactions on Graphics (TOG) 38(4):1–12

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE conference on computer vision, pp 2961–2969

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems, pp 6626–6637

Hinterstoisser S, Lepetit V, Ilic S, Holzer S, Bradski G, Konolige K, Navab N (2012) Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: Asian conference on computer vision, Springer, pp 548–562

Huang S, Qi S, Zhu Y, Xiao Y, Xu Y, Zhu SC (2018) Holistic 3d scene parsing and reconstruction from a single rgb image. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 187–203

Huang Z, Huang L, Gong Y, Huang C, Wang X (2019) Mask scoring r-cnn. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6409–6418

Ji S, Xu W, Yang M, Yu K (2012) 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence 35(1):221–231

Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY (2002) An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence 24(7):881–892

Kehl W, Milletari F, Tombari F, Ilic S, Navab N (2016) Deep learning of local rgb-d patches for 3d object detection and 6d pose estimation. In: European conference on computer vision, Springer, pp 205–220

Krause J, Stark M, Deng J, Fei-Fei L (2013) 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105

Lee T, Lin YL, Chiang H, Chiu MW, Hsu W, Huang P (2018) Cross-domain image-based 3d shape retrieval by view sequence learning. In: 2018 International Conference on 3D Vision (3DV), IEEE, pp 258–266

Li W, Saeedi S, McCormac J, Clark R, Tzoumanikas D, Ye Q, Huang Y, Tang R, Leutenegger S (2018) Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. arXiv preprint arXiv:180900716

Li Y, Su H, Qi CR, Fish N, Cohen-Or D, Guibas LJ (2015) Joint embeddings of shapes and images via cnn image purification. ACM transactions on graphics (TOG) 34(6):234

Lim JJ, Pirsiavash H, Torralba A (2013) Parsing ikea objects: Fine pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2992–2999

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, pp 740–755

Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125

Liu S, Li T, Chen W, Li H (2019) Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: Proceedings of the IEEE International Conference on Computer Vision, pp 7708–7717

Massa F, Russell BC, Aubry M (2016) Deep exemplar 2d-3d detection by adapting from real to rendered views. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6024–6033

Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A (2019) Occupancy Networks: Learning 3D reconstruction in function space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4460–4470

Najman L, Schmitt M (1994) Watershed of a continuous function. Signal Processing 38(1):99–112

Oechsle M, Mescheder L, Niemeyer M, Strauss T, Geiger A (2019) Texture fields: Learning texture
representations in function space. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4531–4540
Park K, Mousavian A, Xiang Y, Fox D (2020) Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10710–10719
Peng S, Liu Y, Huang Q, Zhou X, Bao H (2019) Pvnnet: Pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4561–4570
Qi CR, Su H, Mo K, Guibas LJ (2017a) Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 652–660
Qi CR, Yi L, Su H, Guibas LJ (2017b) Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems, pp 5099–5108
Rad M, Lepetit V (2017) Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3828–3836
Raj A, Ham C, Barnes C, Kim V, Lu J, Hays J (2019) Learning to generate textures on 3d meshes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 32–38
Rothganger F, Lazebnik S, Schmid C, Ponce J (2006) 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. International journal of computer vision 66(3):231–259
Shilane P, Min P, Kazhdan M, Funkhouser T (2004) The princeton shape benchmark. In: Proceedings of Shape Modeling Applications, 2004., IEEE, pp 167–178
Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from rgbd images. In: European conference on computer vision, Springer, pp 746–760
Song C, Song J, Huang Q (2020) Hybridpose: 6d object pose estimation under hybrid representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 431–440
Song S, Lichtenberg SP, Xiao J (2015) Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 567–576
Su H, Maji S, Kalogerakis E, Learned-Miller E (2015) Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision, pp 945–953
Sun X, Wu J, Zhang X, Zhang Z, Zhang C, Xue T, Tenenbaum JB, Freeman WT (2018a) Pix3d: Dataset and methods for single-image 3d shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2974–2983
Sun Y, Liu Z, Wang Y, Sarma SE (2018b) Im2avatar: Colorful 3d reconstruction from a single image. arXiv preprint arXiv:180406375
Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
Tasse FP, Dodgson N (2016) Shape2vec: semantic-based descriptors for 3d shapes, sketches and images. ACM Transactions on Graphics (TOG) 35(6):208
Tatarchenko M, Richter SR, Ranftl R, Li Z, Koltun V, Brox T (2019) What do single-view 3d reconstruction networks learn? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3405–3414
Tekin B, Sinha SN, Fua P (2018) Real-time seamless single shot 6d object pose prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 292–301
Tulsiani S, Zhou T, Efros AA, Malik J (2017) Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2626–2634
Uy MA, Pham QH, Hua BS, Nguyen T, Yeung SK (2019) Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1588–1597
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
Wang N, Zhang Y, Li Z, Fu Y, Liu W, Jiang YG (2018a) Pixel2Mesh: Generating 3D mesh models from single RGB images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 52–67
Wang X, He X, Feng F, Nie L, Chua TS (2018b) Tem: Tree-enhanced embedding model for explainable
Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13(4):600–612

Wu J, Wang Y, Xue T, Sun X, Freeman B, Tenenbaum J (2017) Marrnet: 3d shape reconstruction via 2.5 d sketches. In: Advances in neural information processing systems, pp 540–550

Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J (2015) 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1912–1920

Wu Z, Xiong Y, Yu SX, Lin D (2018) Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3733–3742

Xiang Y, Mottaghi R, Savarese S (2014) Beyond pascal: A benchmark for 3d object detection in the wild. In: IEEE winter conference on applications of computer vision, IEEE, pp 75–82

Xiang Y, Kim W, Chen W, Ji J, Choy C, Su H, Mottaghi R, Guibas L, Savarese S (2016) Objectnet3d: A large scale database for 3d object recognition. In: European Conference on Computer Vision, Springer, pp 160–176

Xiang Y, Schmidt T, Narayanan V, Fox D (2017) Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:171100199

Xiao J, Owens A, Torralba A (2013) Sun3d: A database of big spaces reconstructed using sfm and object labels. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1625–1632

Xiao J, Ehinger KA, Hays J, Torralba A, Oliva A (2016) Sun database: Exploring a large collection of scene categories. International Journal of Computer Vision 119(1):3–22

Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1492–1500

Xu Q, Wang W, Ceylan D, Mech R, Neumann U (2019) DISN: Deep implicit surface network for high-quality single-view 3D reconstruction. In: Advances in Neural Information Processing Systems, pp 492–502

Yang L, Luo P, Change Loy C, Tang X (2015) A large-scale car dataset for fine-grained categorization and verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3973–3981

Yang X, He X, Wang X, Ma Y, Feng F, Wang M, Chua TS (2019) Interpretable fashion matching with rich attributes. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 775–784

Zhang R, Tsai PS, Cryer JE, Shah M (1999) Shape-from-shading: a survey. IEEE transactions on pattern analysis and machine intelligence 21(8):690–706

Zheng J, Zhang J, Li J, Tang R, Gao S, Zhou Z (2019) Structured3d: A large photo-realistic dataset for structured 3d modeling. arXiv preprint arXiv:190800222

Zhou B, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A (2017) Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 633–641

Zhu JY, Zhang R, Pathak D, Darrell T, Efros AA, Wang O, Shechtman E (2017) Toward multimodal image-to-image translation. In: Advances in neural information processing systems, pp 465–476