# Towards Practical Sketch-Based 3D Shape Generation: The Role of Professional Sketches

Yue Zhong, *Member, IEEE*, Yonggang Qi, *Member, IEEE*, Yulia Gryaditskaya, *Member, IEEE*, Honggang Zhang, *Senior Member, IEEE*, and Yi-Zhe Song, *Senior Member, IEEE*

***Abstract*— In this paper, for the first time, we investigate the problem of generating 3D shapes from professional 2D sketches via deep learning. We target sketches done by professional artists, as these sketches are likely to contain more details than the ones produced by novices, and thus the reconstruction from such sketches poses a higher demand on the level of detail in the reconstructed models. This is importantly different to previous work, where the training and testing was conducted on either synthetic sketches or sketches done by novices. Novices sketches often depict shapes that are physically unrealistic, while models trained with synthetic sketches could not cope with the level of abstraction and style found in real sketches. To address this problem, we collected the first large-scale dataset of professional sketches, where each sketch is paired with a reference 3D shape, with a total of 1,500 professional sketches collected across 500 3D shapes. The dataset is available at http://sketchx.ai/downloads/. We introduce two bespoke designs within a deep adversarial network to tackle the imprecision of human sketches and the unique figure/ground ambiguity problem inherent to sketch-based reconstruction. We show that existing 3D shapes generation methods designed for images fail to be naively applied to our problem, and demonstrate the effectiveness of our method both qualitatively and quantitatively.**

***Index Terms*— Professional sketch dataset, deep sketch modeling.**

## I. Introduction

SKETCHING has long been used as a creative tool that significantly benefits animation, movie and building industries. Professional artists and architects typically initialize their design by sketching, which can later be converted into full 3D shapes (*e.g.*, for 3D animation). This process is however labor-intensive, and ultimately involves creating 3D shapes from scratch without leveraging the ready-made sketches.

While automatic conversion of an input sketch to 3D was studied for a long while, see [1] for a detailed overview, many methods either rely on certain assumption on sketching style, or require dedicated-user interface, or manual annotation of an input sketch. Recent work [2]–[5] leverages deep learning to automate the process of sketch to 3D shape generation, as means to improve efficiency and save cost. Yet, efforts have focused on using novices [2], [3] and synthetic [4], [5] sketches as input, which albeit being a good starting point, do not bring enough insights on how practical demands are met. This is because (i) novices sketches often result in shapes that are physically unrealistic, therefore requiring considerable efforts in view of utilizing them for practical tasks such as animation and 3D printing, where starting from scratch might often be a more effective approach, and (ii) synthetic sketches, *i.e.*, sketches rendered from the actual 3D shapes, exhibit perfect sketch-3D shape alignment, making computational models trained on them unable to handle the level of abstraction and style variations in real human sketches, as was shown in [6].

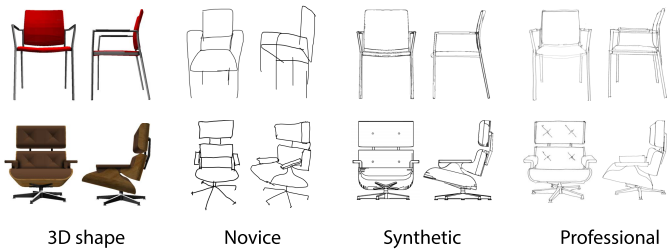



Fig. 1. Comparison of sketches done by professionals and novices with 2D projections of 3D shapes and synthetic sketches generated with non-photo-realistic rendering algorithms from reference 3D shapes.

In this paper, we study the problem of professional sketch to 3D shape generation with deep learning. Professional sketches carefully depict shape details and volume, thus require higher quality of the generated 3D shapes than those that can be created from novices sketches. The use of professional sketches in the pipeline of automatic 3D shape generation is motivated by the fact that in most practical scenarios, where its application is required, content creators are often trained artists. Fig. 1 illustrates some of the differences among novices, synthetic and professional sketches.

The study of professional sketches to 3D shape generation has not been possible till now due to a lack of datasets that exhibit one-to-one sketch-3D correspondences. Gryaditskaya *et al.* [6] collected a detailed dataset of professional concept sketches, yet it contains in total around

Manuscript received January 2, 2020; revised October 8, 2020; accepted November 12, 2020. Date of publication November 26, 2020; date of current version September 3, 2021. This work was funded by National Natural Science Foundation of China under Grant No. 62076034. This article was recommended by Associate Editor G. Zhao. *(Corresponding author: Honggang Zhang.)*

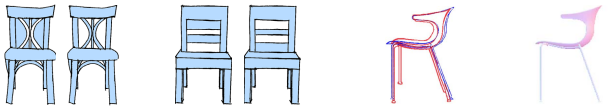
Yue Zhong, Yonggang Qi, and Honggang Zhang are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zhhg@bupt.edu.cn).

Yulia Gryaditskaya is with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K.

Yi-Zhe Song is with the SketchX Laboratory, University of Surrey, Guildford GU2 7XH, U.K.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2020.3040900.

Digital Object Identifier 10.1109/TCSVT.2020.3040900



(a) Foreground Ambiguity (b) Stroke Misalignment

Fig. 2. Main challenges of a 2D professional sketch to 3D shape generation. (a) A toy example of figure/ground ambiguity: exactly the same sketch drawing could result in quite different 3D shapes and (b) an example of typical misalignment between a professional drawing (left strokes in red) and the synthetic sketch (left lines in blue) matching a 2D projection of a 3D shape (right).

400 sketches, which is not sufficient to train deep models. In their work they showed that the model trained on synthetic data does not generalize well to real-world sketches, motivating the need for larger-scale human sketches datasets. The first key contribution of our paper is a professional sketch and 3D shape dataset, consisting in total of 1,500 sketch-3D pairings across 500 3D shapes for the chair object category. Collecting such a dataset is challenging in that different to collecting free-hand sketches as per sketch recognition [7]–[9] and sketch-based image retrieval [10]–[13], where sketches can be relatively cheaply sourced (*e.g.*, under 20 seconds per sketch for QuickDraw [14]). We are faced with the more difficult task of sourcing professional artists to produce high-quality sketches, which is both a lengthy (each sketch takes considerably more time to draw) and costly process (reward per sketch is higher).

Even with the dataset, the problem of professional sketch to 3D shape generation is non-trivial. There are two unique difficulties brought by professional sketch that need to be addressed. First, sketches naturally exhibit figure/ground ambiguity, *i.e.*, the same sketch can lead to different 3D shape interpretations (Fig. 2(a) offers examples). This ambiguity is largely caused by foreground being hard to distinguish relying solely on a few lines in absence of colour and texture. Second, despite the sketches are professionally produced, misalignment still exists between sketch and 3D shape as a result of different drawing skills and styles of the artists.

As the second contribution, we design a deep adversarial network to specifically tackle these challenges, where (i) a discrimination-attention mechanism is developed to help figure/ground estimation; we re-purpose the self-attention mechanism in [15] and introduce a novel attention loss to ensure that an automatically generated attention map aligns with that of the ground-truth 3D shape, and (ii) in order to tackle the inherent sketch-3D shape misalignment, we learn a global non-linear geometric transformation between an input sketch and its 3D shape counterpart via a spatial transform network. Our method builds on a recent work [4], where an adversarial learning strategy is adopted to train a conditional GAN, which is able to generate 2D representations of surface normal and depth maps describing a 3D shape from different viewpoints, which can then be fused into a full 3D representation.

Our contribution is summarized as follows:

- We contribute the first professional sketch and 3D shape dataset, that contains 500 3D shapes and 1,500 professional sketches, to drive future research.
- We propose a deep adversarial network with specific designs to tackle the unique traits of the sketch-based 3D model generation from professional detailed sketches.

## II. Related Work

*a) Image-based 3D shape modeling:* A large body of literature on image-based 3D shape modeling exists, which can be mainly categorized into single-view [16]–[23] and multi-view 3D shape modeling [16], [18], [24]. The existing approaches can be further categorized based on the used 3D representation. Voxel-based models have been the most widely studied, where CNN has been successfully applied to learn probabilistic latent space of 3D object reconstruction given 2D view images [16]. However, these approaches often suffer from limitation on computation efficiency [22], [25] hence often lead to unsatisfied number of voxels produced. To alleviate it, octree-based networks [26], [27] and point cloud models [28] have been proposed to generate higher-resolution models. More recently, polygon meshes based methods [24], [29], which encode both geometrical (point cloud) and topological (surface connectivity) cues, gained popularity. These approaches are still limited by the topology of the template shape and struggle to accurately reproduce shapes with genius higher than 0. Some works [25], [30], [31] have been proposed to train 3D modeling systems using multiple views of surface normal and depth maps as a supervision signal. An additional fusion step is utilized to merge them into point clouds and to prune outliers using the predicted silhouettes.

*b) Sketch-based 3D shape modeling:* Estimating a 3D shape from a human sketch is even more challenging than from a single image due to two main factors: (i) *ambiguity*: multiple 3D shapes can potentially project onto the same drawing (Fig. 2 (a)), and (ii) *distortion*: misalignment exists between a sketch and a 3D shape (Fig. 2 (b)). Early works infer 3D shapes by leveraging geometric properties. Thus, [32], [33] predict 3D shape by inferring local geometric properties. Hand-crafted rules are usually devised to extrude or elevate a smooth 3D surface from contours [34]. Polyhedral scaffolds [35], cross-section lines [36] and curvature flow lines [37] have also been exploited to create free form surfaces. Different from the above methods, our approach does not require explicit principles to describe geometric cues for input sketches or to generate 3D shape views.

More recently, learning-based methods were proposed, where sketch-based 3D shape modeling has been commonly treated as an image-to-image translation problem [38]. Su *et al.* [2] proposed to train a GAN framework to generate normal maps, and an interactive process of incorporating user-specific normal is utilized to improve the quality. It is however expensive to generalize as it requires additional input of point mask and user guidance. Wang *et al.* [3] employ adversarial training and utilize pairs of synthesized images and reference 3D shapes from a 3D shape gallery as sources to pooling together as a union feature. Albeit with satisfactory results obtained, it heavily relies on template shapes in the gallery, which largely limits its general applicability. Reference [5] investigates 3D shape modeling by training on

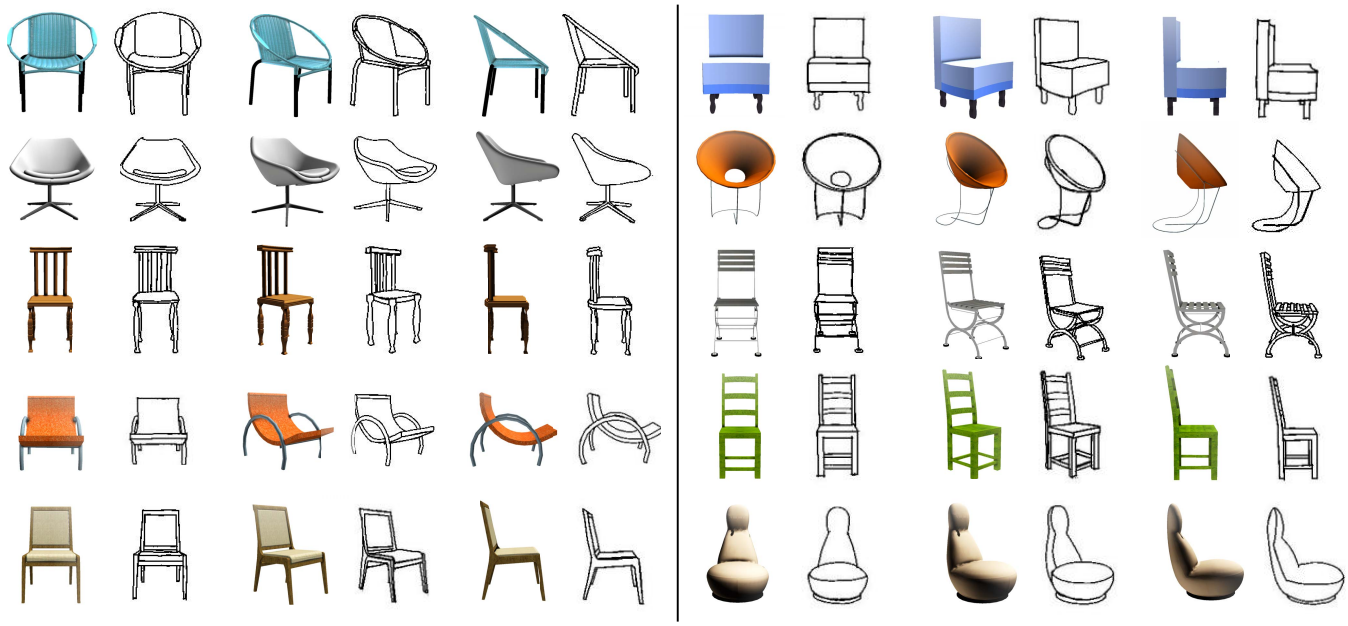


Fig. 3. Example ProSketch-3D pairs. Each row shows the ground-truth 3D shapes and their corresponding drawn professional sketches in three views (front, side and 45°).

synthetic sketches obtained by image-space contour rendering [39]. A single-view CNN and an updater CNN are trained to generate occupancy of a voxel grid from a sketch. The limitation however lies in reconstructing thin structures. Concurrent to our work by Jin *et al.* [40] proposed an alternative voxel-based method, while it proposes an interesting idea of inferring the two-additional viewpoints given the distance to the encoding of the reference view in the embedded space, it inherits the limitations of all voxel-based methods – the limited spatial resolution. The work by Li *et al.* [41] is targeted to designing free-form surfaces with complex curvature patterns. This work requires careful user annotation of used line types, such as silhouettes, ridges, valleys, sharp feature and some others. Moreover, compared to our work they only reconstruct a height field from the reference viewpoint, while we aim at a full 3D model reconstruction. Han *et al.* [42] proposed a dedicated network for interactive face sketch-based modeling that outputs a set of coefficients for bilinear face representation, and thus can not be trivially extended to an arbitrary shape class.

We build on the recent work by Lun *et al.* [4] which takes as input two orthographic views of a shape, and predicts a set of views (including unseen views) of depth and normal maps by training an encoder-decoder network through adversarial learning. Note that this allows us to exploit the success of image-to-image translation networks, which are for instance were applied to photo-to-sketch generation [38], [43], [44]. We show that our method allows to more accurately reproduce shape details than previous work, and results in substantially increased robustness of sketch-based reconstruction.

## III. Professional Sketch and 3D Shape Dataset

Our proposed professional sketch and 3D chair dataset, *ProSketch3D*, is the first dataset that exhibits one-to-one correspondences between professional-drawn sketches and 3D shapes. It contains 500 3D shapes of chairs and 1,500 corresponding professional sketches under three views (front, side and 45°). Some example drawings are shown in Fig. 3.

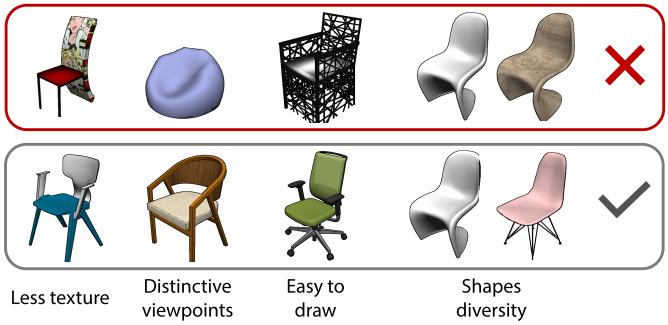



Fig. 4. Illustration of four 3D shape selection criteria: uniform textures, view distinguishability, easy to draw and clear shape diversity.

*c) 3D shapes:* Our dataset is built on ShapeNetCore [45], the largest 3D shape dataset to date. Category chair is carefully picked out due to its large variance in an intra-class appearance. In particular, 500 3D chairs are collected from a set of 6,778 available 3D chair shapes in ShapeNetCore [45] following the four criteria as shown in Fig. 4: (1) We select shapes which contain little texture, since it can distract the artist from the shape itself. (2) We were attempting to avoid 3D shapes with a similar appearance from different views, such as a round lazy sofa. (3) We selected the shapes that are not too difficult to draw to make the task of large-scale dataset collection feasible. (4) We selected shapes with distinctive shape topologies.

*d) Sketching device:* Collecting professional sketches from 3D models is a non-trivial endeavour. First and foremost,

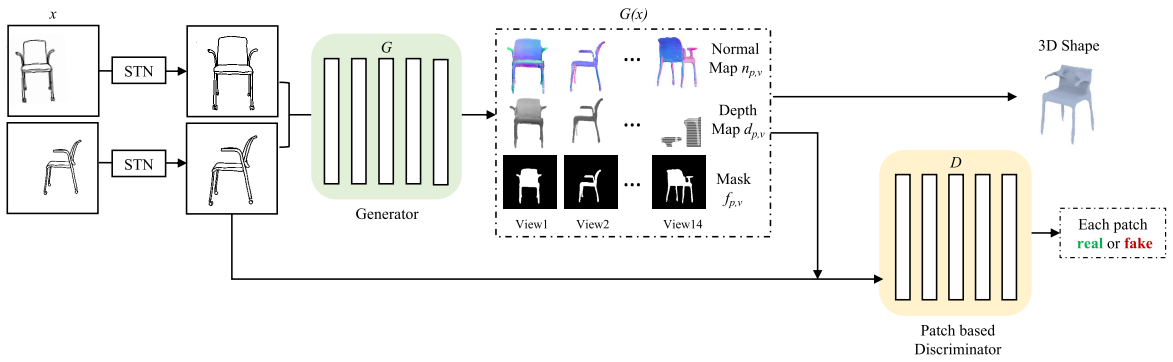



Fig. 5. Network architecture, described in detail in Sec. IV: Each of two input sketches is first fed into one out of two STN modules, followed by the Generator $G$. Al last the generator output and sketches transformed with STN are fed into the Discriminator $D$ to predict if each $N \times N$ patch is real or fake. The final 3D shape is reconstructed from a set of normal and depth maps and foreground masks which are output from the generator.

it is important that one uses the right drawing device so that (i) artists can achieve best sketching performance (e.g., drawing on an iPad is completely different to drawing on a paper), and (ii) the detailed drawing information is faithfully preserved. We opted to use the ISKN Slate 2,[1] a digital drawing tablet that uses a real paper and a real pencil as the drawing interface. This maximally facilitates artists to produce drawings in a more authentic creative environment, hence leads to exactly the same quality drawings with the habitually production for artists. Moreover, except pixel-based drawings, the ISKN Slate 2 also supports a vector-sketch format which captures stroke-level data with ordering information.

*e) Style unification:* Other than choosing the right input device, a key problem faced by all sketch research is that of style [46] – different artists tend to draw in distinct styles, which means that an enormous amount of data is needed to be collected to faithfully reproduce the real-world style diversity. To make the data collection task attainable, we focus here on collecting the sketches completed in a similar style, yet containing realistic perspective and mechanical inaccuracies inherent to human sketches, and unique to each person. We recruited 36 experienced professionals who have similar art training. Every professional has major in arts and more than 5 years of specialized drawing experience. At the first stage, we prepared 30 views of 10 3D shapes, and ask the participants to sketch in their habitual sketching style. We then chose 10 artists from the 36, whose sketches are of a similar drawing style, and employed them to produce sketches for our dataset.

*f) Sketch collection:* Before the final data collection stage, all the professionals were given at least two days to get used to the ISKN slate 2 sketching device. We selected one of the sketches from the first stage as a sketching style exemplar. All 10 artists were gathered together and were given more than an hour to study and imitate the prepared exemplar. We evenly divided the 500 models into 10 groups, resulting in total of 50 models per a professional. For each 3D model, firstly, 5 mins are given for participants to view the model using a 3D viewer under free rotation and scaling, so that they can get a "sufficient understanding" of the shape. Then, three predefined viewpoints of the models are presented to them on a computer screen, front, side and 45°-view, and they are required to produce a sketch for these views. Professionals were only allowed to move on to the next view/model when they were satisfied with the current drawing. A post-processing check is performed for quality control, where we asked participant to redo a sketch if it had some obvious defects, such as blurriness. The average sketching time for each 3D shape in all 3 views constitutes around 20 minutes.

[1]Explore more about the ISKN Slate 2 on: https://www.iskn.co/uk/

## IV. Method

### A. Network Overview

Our network is designed to take as an input two orthographic views of the shape, such views are often created by designers and architectures prior to 3D shape modeling or fabrication. Reconstructing a 3D model from such sketches is challenging, since (i) sketches naturally exhibit figure/ground ambiguity, caused by the foreground being hard to distinguish relying solely on a few lines in absence of color and texture (Fig. 2(a)); (ii) sketches are imprecise and commonly contain a misalignment between a sketch and a 3D shape, and between two views (Fig.2(a)).

To deal with misalignment, we leverage a Spatial Transform Module (STN) [47], placed before the network generator (Section IV-B). To address figure/ground ambiguity, we employ a self-attention mechanism [15] in conjunction with an attention-based loss aiming at better capturing the global structure of a sketch (Section IV-E). Furthermore, to improve the reconstruction of details we utilize a patch-based discriminator (Section IV-D).

The overview of our network is shown in Fig. 5. A GAN-based deep convolutional neural network is developed to translate input 2D sketches $S_{u_i}$, $i = \{front, side\}$ into 2D images of surface normals, depth maps and masks over 14 viewpoints, which jointly describe the target 3D shape. These 14 viewpoints, including the front and side views, and the other 12 viewpoints, are located at the equidistant vertices of a regular icosahedron [4]. All our training shapes are normalised to fit inside the icosahedron and are consistently oriented. Details of the key components of our network are described below.

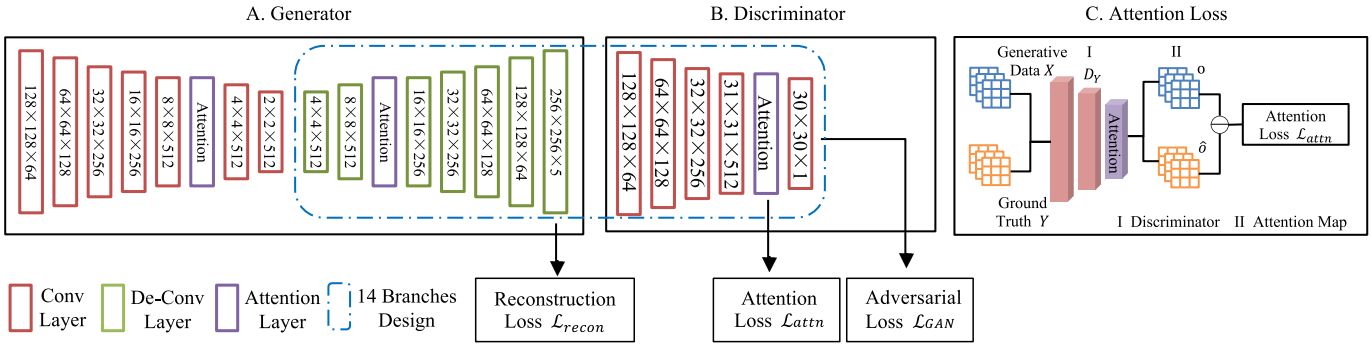



Fig. 6. Network overview: the Generator (A), the Discriminator (B), and detailed illustration of the proposed Attention Loss (C). Best viewed in color.

### *B. Spatial Transform Network (STN): Alignment*

To tackle the inherent sketch-3D shape misalignment, two *STNs* [47] are employed before input sketches are passed to the generator, as shown in Fig. 5. An *STN* contains two fully-connected layers, followed by a spatial transformer layer. The two fully-connected layers take the input sketch $S \in \mathbb{R}^{h\times w\times c}$, of height $h$, width $w$ and $c = 1$ color channels, and return the transform variables $\theta$ of a global affine transformation matrix. The input sketch and the transform variables are then passed to the spatial transformer layer to produce a transformed sketch.

### *C. Generator*

Our generator is trained to produce normal maps, depth maps and foreground masks from observed sketches to fool a discriminator. The generator is an encoder-decoder architecture, shown in Fig. 6. The *encoder* takes concatenated views $S_u \in R^{256\times256\times2}$ as an input, which are passed through the seven convolutional layers and an attention layer, a 2048-d feature vector ($2 \times 2 \times 512$) is obtained, which is then passed to the decoder.

The *decoder* contains 14 branches, each corresponding to a specific 3D viewpoint. The output of each branch $X_v \in R^{256\times256\times5}$ for each viewpoint $v$, consists of 5 channels, three of which encode a normal map $N_v$, one a depth map $D_v$ and one a foreground mask $M_v$, similar to [4]. All the 14 branches of the decoder are learned individually and do not share weights.

### *D. Patch-Based Discriminator: Details*

Inspired by [38], we use a patch-based discriminator architecture to encourage the GAN discriminator to model high-frequency structures. Patch-based discriminator tries to classify whether each $N \times N$ image patch is real or fake rather than directly discriminating on a whole image. In our case, the discriminator is run convolutionally on the maps $X_v \in R^{256\times256\times5}$ from each viewpoints. Effectively, for each patch $P_j \in R^{N\times N\times5}$ of $X_v$, with a $N = 70$ patch size, it predicts if this patch is real or fake. Thus, the patch-based discriminator first maps the $X_v$ to a $30 \times 30$ array, where each value signifies whether the corresponding patch in the $X_v$ is real or fake, as shown in Fig. 6. The output of our discriminator is, then, the average discrimination value of patches. Note that normal and depth maps are filtered by foreground masks before being passed to the discriminator.

### *E. Attention Layer: Global Sketch/Shape Structure*

Since sketches contain only sparse lines compared to rich color and texture available in photographs, we argue for the need of the mechanism that would allow to model long range and multi-level dependencies across a sketch, lifting figure/ground ambiguity. A self-attention mechanism [15] effectively models such dependencies. This mechanism was previously proved in [48], [49], and further exploration of long range dependencies is discussed in [50].

We are the first to exploit a self-attention mechanism [15] for a sketch-based modeling. The attention layers in the generator allow to predict maps in which fine details in each location are consistent with details in distant regions. The discriminator equipped with an attention layer can more reliably predict if each patch is fake or real by taking into account global information.

The idea behind a self-attention layer is to calculate the response at a given position as a weighted sum of the features at all positions. Let $X \in \mathbb{R}^{n\times c}$ be a features map from the hidden layer preceding an attention layer, where $n = h \times w$ is the feature locations and $c$ is a number of channels. These features $X$ are first transformed into three new feature maps, which are referred to in [51] as queries $Q \in \mathbb{R}^{n\times\hat{c}}$, keys $K \in \mathbb{R}^{n\times\hat{c}}$ and values $V \in \mathbb{R}^{n\times\hat{c}}$ via $1\times1$ convolutions, where $\hat{c} = \frac{c}{8}$. Then the self-attention map $A \in \mathbb{R}^{n\times n}$ is computed, that represents a normalized score for each position $j$ against the query position $i$:

$$A_{j,i} = \frac{\exp\left(Q_i \cdot K_j\right)}{\sum_{i=1}^{N} \exp\left(Q_i \cdot K_j\right)}. \tag{1}$$

To obtain the final output, the values $V$ are weighted by the self-attention scores and are summed up. The output of the this convolutional layer is then scaled and added to the input feature map $X$:

$$O_j = \alpha \sum_{i=1}^{n} (A_{j,i} V_i) + X_j. \tag{2}$$

Scaling parameter $\alpha$ is a learnable parameter, initialized with 0 as in [15].

### *F. Objectives*

Our full loss consist of three losses: the reconstruction loss, the attention loss and the adversarial loss, detailed below.

*g) Reconstruction loss:* The reconstruction loss of generator enforces the encoder-decoder to reliably reconstruct 2D images of normal, depth maps and foreground masks across viewpoints:

$$\mathcal{L}_{recon} = \mathcal{L}_{recon}^{normal} + \mathcal{L}_{recon}^{depth} + \mathcal{L}_{recon}^{foreground}. \tag{3}$$

The first two terms consider a per-pixel error between generated data and ground-truth:

$$\mathcal{L}_{recon}^{normal} = \sum_{p,v}(1 - n_{p,v}(S) \cdot \hat{n}_{p,v})\hat{f}_{p,v}, \tag{4}$$

$$\mathcal{L}_{recon}^{depth} = \sum_{p,v}(|d_{p,v}(S) - \hat{d}_{p,v}|)\hat{f}_{p,v}, \tag{5}$$

where $\hat{n}_{p,v}$ and $\hat{d}_{p,v}$ are the ground-truth normal and depth for the pixel $p$ in the viewpoint $v$ for the input sketch $S$, while $n_{p,v}$ and $d_{p,v}$ are the predicted normal and depth values. The loss for a foreground mask prediction is defines as:

$$\mathcal{L}_{recon}^{foreground} = \sum_{p,v}(|f(S) - \hat{f}_{p,v}|)., \tag{6}$$

where $\hat{f}$ is a predicted foreground mask, and $f$ is a ground-truth.

*h) Attention loss:* In the discriminator, a novel attention loss is developed to penalize the disagreement of attention maps produced by the generated data and ground-truth 3D shape as show in Fig. 6 to help resolve figure/ground ambiguity, defined as an $L_2$ norm:

$$\mathcal{L}_{attn} = \sum_{p,v} \|o_{p,v} - \hat{o}_{p,v}\|_2, \tag{7}$$

where $o_{p,v}, \hat{o}_{p,v}$ are attention layers outputs at the pixel $p$ and view $v$ of a discriminator and generator, accordingly.

*i) Adversarial loss:* Given the output of generator $X$, and suppose that the target data is a ground-truth normal map $N$ and depth map $D$, denoted as $Y = \{N, D\}$, then for the generator $G : X \rightarrow Y$ and its discriminator $D_Y$, the adversarial loss is defined as:

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y\sim p(y)}[log D_Y(y)] + \mathbb{E}_{x\sim p(x)}[log(1 - D_Y(G(x))], \tag{8}$$

where $G$ tries to generate images $G(x)$ that can mimic images from domain $Y$, while $D_Y$ aims to distinguish $G(x)$ from real ones $y$, which is a min-max optimization on $G$ and $D_Y$.

*j) Full objective:* Our full objective function is:

$$\mathcal{L}(S, Y) = \mathcal{L}_{attn} + \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{recon}. \tag{9}$$

We jointly train the network to optimize this final objective with an Adam solver:

$$G^* = \arg\min_G \max_{D_Y} \mathcal{L}(S, Y). \tag{10}$$

### *G. Generating a 3D Shape*

Given the 14 viewpoints of depth and normal maps generated by our network at inference stage, we merge them to obtain a 3D shape. This is achieved by following the approach from [4]: Firstly, the depth and normal maps are fused to produce corresponding 3D point clouds by: (i) generating a point set per-view given the known camera extrinsic parameters, (ii) running a conventional iterative closest point (ICP) algorithm to align all the point sets, and (iii) executing an optimization step to refine the alignment. Secondly, the 3D point clouds are converted into a surface mesh by applying the screened Poisson Surface Reconstruction algorithm [52]. We refer readers to [4] for further details.

## V. Evaluation

We first compare our approach against recent work for 3D shape reconstruction from images and sketches, qualitatively and quantitatively. We then evaluate the importance of each component of our method. We additionally demonstrate how our approach can be used for instance-level sketch-based 3D shape retrieval.

### *A. Competitors*

We compare to six alternative methods. *ShapeMVD* [4] is a state-of-the-art on sketch-based 3D shape reconstruction. It technically falls into the category of cross-domain translation models [38], [53] [54], but ShapeMVD is specially designed for multi-view-based 3D shape reconstruction. *ONet* [55] serves as a strong baseline as it achieves state-of-the-art performance on 3D shape reconstruction by predicting point occupancy in a 3D space. *Pixel2Mesh* [24] is a one-way an image to a 3D shape supervised translation model which outputs a 3D mesh. It was previously evaluated only on RGB images. *PSGN* [28] is another state-of-the-art work for RGB image-based 3D shape reconstruction, where a conditional shape sampler is employed to predict plausible 3D point clouds given an input image. *3D-R2N2* [16], a pioneering work on deep image-based 3D modeling, leverages a 3D recurrent neural network to predict $32^3$ voxel-grid occupancy from either a single or multiple views. Delanoy *et al.* [5] proposed a U-Net type architecture for $64^3$ voxel-grid occupancy prediction. We refer to this work as *3DSkVP*. The input to this method can be a single image or multiple views, where the multi-view reconstruction is done in an iterative way: At each new iteration the input are the new view and the previously reconstructed 3D shape. This design was optimized for iterative sketching process. Since in our work we assume that the two views are available simultaneously, instead of performing iterative refining, we concatenate the two views along the third dimension before feeding them to the network.

### *B. Data Split and Implementation Details*

We randomly split our ProSketch3D dataset into 450 3D shapes used for training and 50 3D shapes used for testing. Due to the limited size of the ProSketch3D dataset, we pre-train our network and all the competitors with

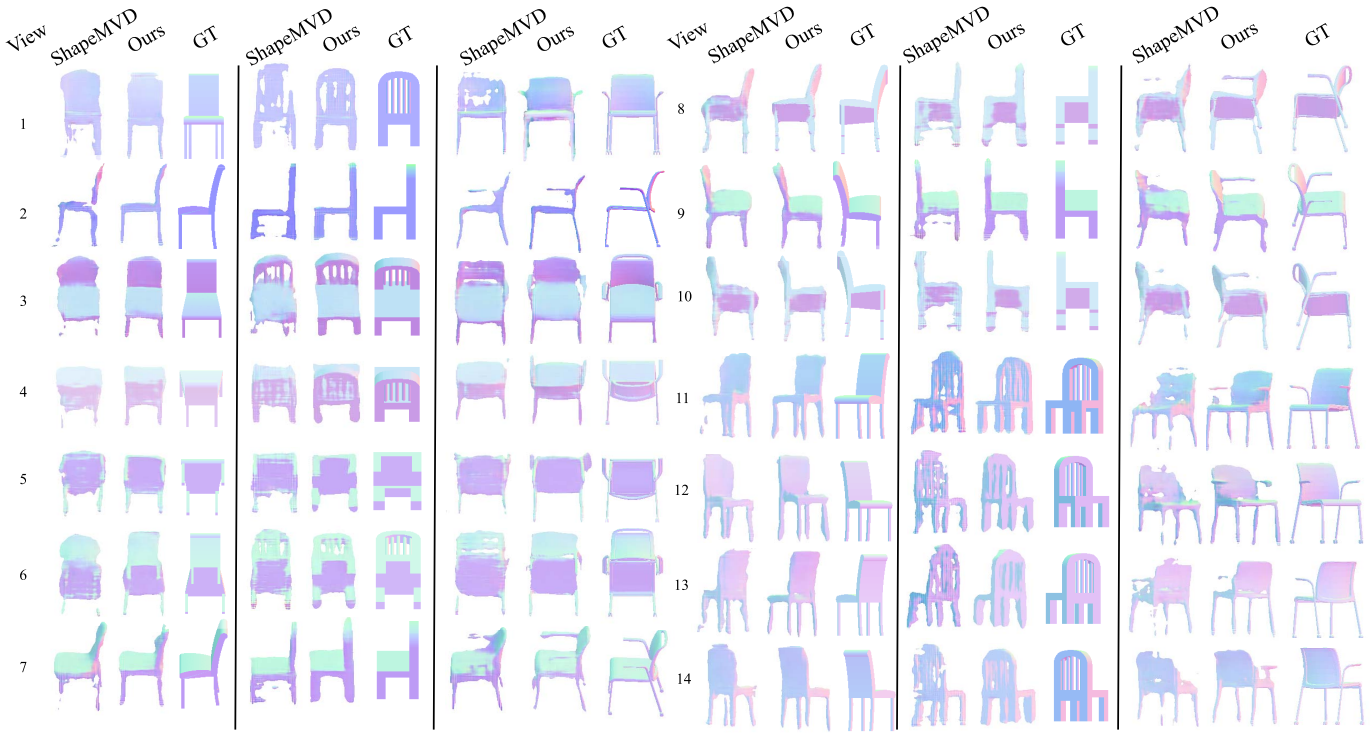



Fig. 7. Comparisons of generated views. Our generated views have finer details and sharper outlines. Zoom-in for details.

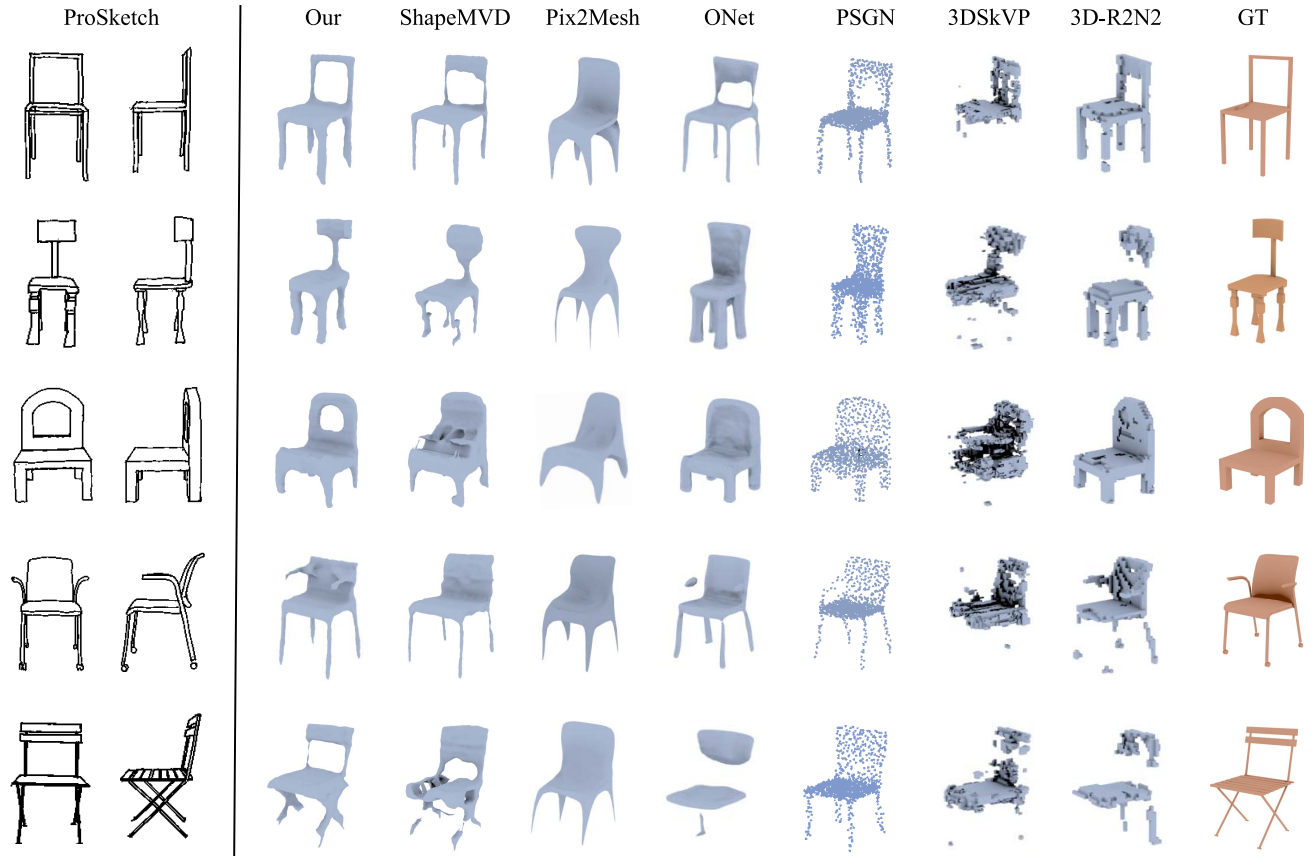



Fig. 8. Comparison of generated whole 3D shapes. Our produced 3D shapes have richer details and sharper shape, which are better than those given by other methods.

synthetically generated sketches from 6,278 3D chairs from the ShapeNet [45] dataset. For each 3D shape we generate its corresponding front and side orthographic projections using non-photo-realistic rendering as in [4]. After pre-training, we train the network on the train split of the ProSketch3D dataset.

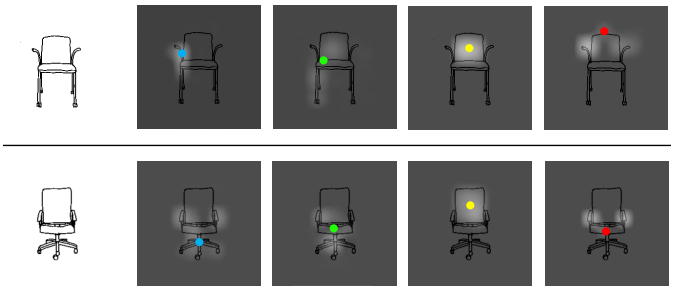


Fig. 9. Example visualizations of attention maps in the generator. White color intensity indicates a relative attention of a location with respect to a query location. Each color dot shows a query location.

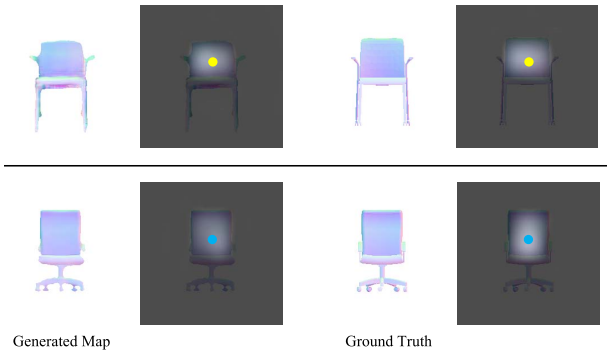



Fig. 10. Example visualizations of attention maps in the discriminator for generated maps and ground-truth maps. Trained network tends to attend to similar regions on the generated and ground-truth 3D shapes. Each color dot shows a query location.

Our network is implemented in TensorFlow on a single 1080Ti GPU. Adam optimizer [56] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$ is used. The initial learning rate is set to 0.0001 and the batch size is set to 2.

For all the competitors we used the authors provided implementations, apart from for the *3DSkVP* [5]. Due to the outdated version of the custom modified Caffe libraries used in the original implementation we re-implemented their work in PyTorch. We used Adam optimizer with the learning rate set to 0.0005, and the batch size set to 64.

### C. Results and Discussions

*1) Qualitative Results:* We first compare the individual viewpoints generated by our method with those generated by ShapeMVD [4]. For both methods, each view is produced by fusing per view normal and depth maps, and a foreground mask, as described in [4]. It can be seen in Fig. 7 that the views generated with our method have finer details and sharper outlines.

We then compare the 3D models generated with our method with those produces by all the baseline models. The 3D shapes reconstructed with our method more accurately represent the input sketch and capture fine details better than any alternative method (Fig. 8).

To reveal how and why attention the layer works, we visualize the attention maps. Firstly, we visualize the attention maps produced by the generator. In fig. 9, we can observe that the network learns meaningful structural dependencies. Secondly, to validate the effectiveness of the proposed attention loss in the discriminator, we visualize in Fig. 10 the attention maps for the generated data and its ground-truth. It conforms to our intention: the trained network tends to attend to similar regions of the generated and ground-truth 3D shapes.

*2) Quantitative Results:* We use five metrics to quantitatively evaluate the quality of the generated 3D shape: Normal angle distance (**NAD**) measures the average normal angle distance for every nearest pair of points between generated and ground truth 3D shape. The Hausdorff distance (**HD**) measures the maximum distance from each surface point on the reconstructed shape to the nearest surface point on the reference shape. Depth error (**DE**) is a Euclidean pixel-wise distance between the generated and reconstructed depth maps. Mask error (**ME**) is computed by measuring the pixel-wise average error between predicted and ground-truth masks. The Chamfer distance (**CD**) evaluates the average distance of each of the points on the generated shape (we randomly sampled 5K points for each shape) to the nearest surface point on the ground-truth shape. We can observe from the Table I that our approach outperforms all the other competitors.

TABLE I
QUANTITATIVE COMPARISONS

| Methods | NAD | HD | CD |
|---|---|---|---|
| ShapeMVD [4] | 33.8402 | 0.4152 | 0.0681 |
| Pixel2Mesh [24] | 35.9844 | 0.5218 | 0.0858 |
| ONet [55] | 34.3463 | 0.2154 | 0.0337 |
| PSGN [28] | - | 0.2471 | 0.0368 |
| 3DSkVP [5] | 35.3724 | 0.3625 | 0.0533 |
| 3D-R2N2 [16] | 34.5192 | 0.2614 | 0.0385 |
| Our-Synthetic | 33.9621 | 0.4536 | 0.0714 |
| **Our** | **33.0102** | **0.1973** | **0.0297** |

TABLE II
QUANTITATIVE RESULTS FOR ABLATION STUDY

| Methods | ME | DE | NAD | HD | CD |
|---|---|---|---|---|---|
| **Base** | 0.0303 | 0.9044 | 33.8402 | 0.4152 | 0.0681 |
| **+ AL** | 0.0275 | 0.8069 | 33.6973 | 0.3724 | 0.0593 |
| **+ ALoss** | 0.0270 | 0.7992 | 33.5821 | 0.3068 | 0.0472 |
| **+ PD** | 0.0258 | 0.7647 | 33.3305 | 0.2415 | 0.0356 |
| **+ STN** | **0.0243** | **0.7259** | **33.0102** | **0.1973** | **0.0297** |

*3) Ablation Studies:* We study the effectiveness of each of the key components of our model: an attention layer (**AL**), a discrimination-attention loss (**Aloss**), a spatial transform network (**STN**) and a patch based discriminator (**PD**). We compare our full model with alternatives: (i) **Base** becomes ShapeMVD model without all key components, (ii) base model with an attention layer **+ AL**, (iii) additionally plus an attention loss **+ Aloss**, (iv) further plus a patch based discriminator **+ PD** and (v) with an addition of a spatial transform network **+ STN**. Quantitative results in Table II and qualitative results in views Fig. 12 and 3D shapes Fig. 13 validate the effectiveness of each component.


Fig. 11. Comparisons of training solely on synthetic data and ours: training solely with synthetic sketch is inferior to that using professional sketches.

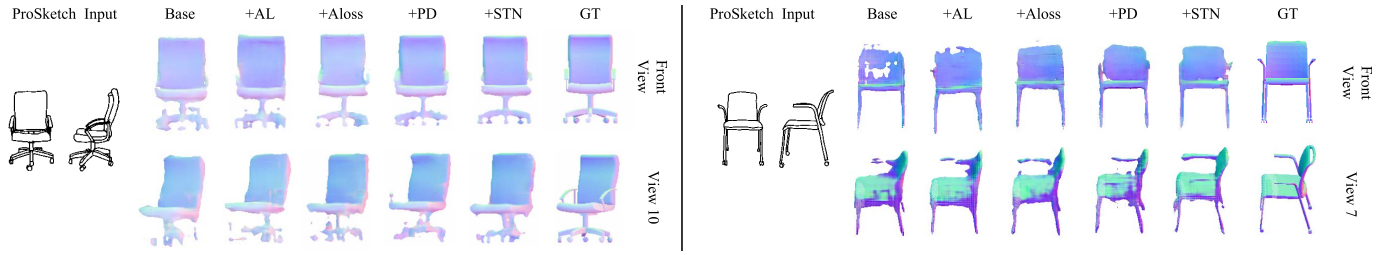



Fig. 12. More comparison on generation results in different 2D views, which further demonstrate the effectiveness of each component when generating views. All the abbreviations used in this figure are defined in Section V-C.3.


Fig. 13. Generated example 3D shapes are presented when adding each key component on the base model one by one. It turns out that finer details and sharper shape can be obtained with our full model. All the abbreviations used in this figure are defined in Section V-C.3.


Fig. 14. Examples of sketch-based 3D shape retrieval. All top 4 3D shapes are quite relevant, with the top ranked result being most similar (except groundtruth) to the automatically generated result.

*4) Training With Synthetic Data:* As previously discussed, sketches by professional artists, albeit being more accurate than novices ones, still exhibit heavy distortions compared to naive synthetic sketches. To evaluate the importance of real human sketches at training stage we compare the results obtained with fine-tuning on the training split of our ProSketch3D dataset with the results obtained by fine-tuning with the synthetic sketches of the 3D shapes from the training split. The results in Fig. 11 and Table I (Our-Synthetic) show that training solely with naive synthetic sketches does not generalise well to human sketches, even as carefully created as sketches in our ProSketch3D dataset.

TABLE III
RECOGNITION ACCURACY OF GENERATED 3D SHAPE

| Methods | Recognition Accuracy |
|---|---|
| ShapeMVD [4] | 0.6600 |
| Pixel2Mesh [24] | 0.4500 |
| ONet [55] | 0.7900 |
| PSGN [28] | 0.8100 |
| 3DSkVP [5] | 0.5300 |
| 3D-R2N2 [16] | 0.7100 |
| **Our** | **0.8700** |

### *D. 3D Shape Retrieval and Recognition*

We show how our shape-generation network can be re-purposed for 3D retrieval and recognition. For retrieval we utilize the encoder as a feature descriptor for both query sketch and 2D projections of a 3D shape, and compute the Euclidean distances in this feature space. In our experiment, all the chairs in the test set are used as query sketches, and the gallery set contains all the 6,778 3D chair shapes from ShapeNetCore.

It can be observed in Fig. 14 that all top 4 retrieved 3D shapes are visually similar to the ground-truth 3D shape.

For recognition, PointNet++ [57] is employed to measure how well our generated 3D chair could be recognised. Results of recognition accuracy show in the Table III. We can observe that our generated data is more easily to be correctly classified comparing with competitors, which in turn shows its superiority.

## VI. Conclusion, Discussion, and Future Work

In this work, we introduce the first large-scale paired professional sketch-3D dataset, termed ProSketch3D. We are the first to source sketches with real pen and paper by using the ISKN slate 2, allowing artists to make drawings in an authentic creative environment leading to high quality human sketches. ProSketch3D consists of a total of 1,500 professional sketches of 500 chairs drawn from 3 different viewpoints. We argue for the importance of such data since (i) professionals are the ones who have a potential to benefit from an automatic sketch-based reconstruction the most, their sketches offer rich object details and many fine structures which need to be carefully reconstructed by the automatic algorithms; (ii) Despite the accuracy of such sketches compared to sketches by novices, professional sketches contain misalignments caused by perceptual and mechanical inaccuracies – we showed that the algorithms trained on synthetic data that does not take this into account do not generalize to the real human sketches. Using the dataset, we investigated the problem of professional sketch-based 3D shape generation, where we proposed specific mechanisms to deal with the misalignment and foreground/background discrimination problems. Experimental results on our proposed ProSketch3D dataset validate the effectiveness of our approach. In this work we investigated the reconstruction from two orthographic views, in the future we would like to investigate a single input reconstruction from a $45^o$ input, as the informative view. Moreover, ProSketch3D dataset will benefit 3D shape retrieval. We will endeavor to release all the code and dataset, and continue to grow the dataset in the future to encourage research in this area.

## References


[1] A. Bonnici *et al.*, "Sketch-based interaction and modeling: Where do we stand?" *Anal. Manuf. Artif. Intell. Eng. Des.*, vol. 33, no. 4, pp. 370–388, Nov. 2019.

[2] W. Su, D. Du, X. Yang, S. Zhou, and H. Fu, "Interactive sketch-based normal map generation with deep neural networks," *Comput. Graph. Interact. Techn.*, vol. 1, no. 1, pp. 1–17, 2018.

[3] L. Wang, C. Qian, J. Wang, and Y. Fang, "Unsupervised learning of 3D model reconstruction from hand-drawn sketches," in *Proc. ACM Multimedia Conf. Multimedia Conf.*, 2018, pp. 1820–1828.

[4] Z. Lun, M. Gadelha, E. Kalogerakis, S. Maji, and R. Wang, "3D shape reconstruction from sketches via multi-view convolutional networks," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 67–77.

[5] J. Delanoy, M. Aubry, P. Isola, A. Efros, and A. Bousseau, "3D sketching using multi-view deep, volumetric prediction," *Comput. Graph. Interact. Techn.*, vol. 1, pp. 1–22, Jul. 2018.

[6] Y. Gryaditskaya, M. Sypesteyn, J. W. Hoftijzer, S. Pont, F. Durand, and A. Bousseau, "Opensketch: A richly-annotated dataset of product design sketches," *ACM Trans. Graph.*, vol. 38, no. 6, p. 232, 2019.

[7] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales, "Sketch-a-net that beats humans," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–5.

[8] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, Jul. 2012.

[9] P. Xu, C. K. Joshi, and X. Bresson, "Multi-graph transformer for free-hand sketch recognition," 2019, *arXiv:1912.11258*. [Online]. Available: http://arxiv.org/abs/1912.11258

[10] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2862–2871.

[11] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5551–5560.

[12] K. Pang *et al.*, "Generalising fine-grained sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 677–689.

[13] P. Xu *et al.*, "SketchMate: Deep hashing for million-scale human sketch retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8090–8098.

[14] D. Ha and D. Eck, "A neural representation of sketch drawings," in *Proc. ICLR*, 2018, pp. 1–4.

[15] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. CoRR*, 2018, pp. 7354–7363.

[16] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," in *Proc. ECCV*, 2016, pp. 628–644.

[17] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1746–1754.

[18] H. Kato and T. Harada, "Learning view priors for single-view 3D reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9778–9787.

[19] B. Yang, H. Wen, S. Wang, R. Clark, A. Markham, and N. Trigoni, "3D object reconstruction from a single depth view with adversarial learning," in *Proc. Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2017.

[20] J. Wu, Y. Wang, T. Xue, X. Sun, B. Freeman, and J. Tenenbaum, "MarrNet: 3D shape reconstruction via 2.5D sketches," in *Proc. NeurIPS*, 2017, pp. 540–550.

[21] X. Sun *et al.*, "Pix3D: Dataset and methods for single-image 3D shape modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2974–2983.

[22] D. Jack *et al.*, "Learning free-form deformations for 3D object reconstruction," in *Proc. ACCV*, 2018, pp. 317–333.

[23] L. Jiang, S. Shi, X. Qi, and J. Jia, "GAL: Geometric adversarial loss for single-view 3D-object reconstruction," in *Proc. ECCV*, 2018, pp. 802–816.

[24] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 52–67.

[25] S. Roth and S. R. Richter, "Matryoshka networks: Predicting 3D geometry via nested shape layers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1936–1944.

[26] G. Riegler, A. O. Ulusoy, and A. Geiger, "OctNet: Learning deep 3D representations at high resolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3577–3586.

[27] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong, "O-CNN: Octree-based convolutional neural networks for 3D shape analysis," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–7, 2017.

[28] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 605–613.

[29] J. Pan, X. Han, W. Chen, J. Tang, and K. Jia, "Deep mesh reconstruction from single RGB images via topology modification networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9964–9973.

[30] A. A. Soltani, H. Huang, J. Wu, T. D. Kulkarni, and J. B. Tenenbaum, "Synthesizing 3D shapes via modeling multi-view depth maps and silhouettes with deep generative networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1511–1519.

[31] C. Lin, C. Kong, and S. Lucey, "Learning efficient point cloud generation for dense 3D object reconstruction," in *Proc. AAAI*, 2018, pp. 1–10.

[32] H. Lipson and M. Shpitalni, "Optimization-based reconstruction of a 3D object from a single freehand line drawing," *Comput.-Aided Des.*, vol. 28, no. 8, pp. 651–663, 2007.

[33] R. C. Zeleznik, K. P. Herndon, and J. F. Hughes, "SKETCH: An interface for sketching 3D scenes," in *Proc. SIGGRAPH*, 2007, p. 9.

[34] T. Igarashi, S. Matsuoka, and H. Tanaka, "Teddy: A sketching interface for 3D freeform design," in *Proc. SIGGRAPH*, 2006, p. 11.

[35] R. Schmidt, A. Khan, K. Singh, and G. Kurtenbach, "Analytic drawing of 3D scaffolds," in *Proc. ACM SIGGRAPH*, 2009, pp. 1–10.
[36] B. Xu, W. Chang, A. Sheffer, A. Bousseau, J. McCrae, and K. Singh, "True2Form: 3D curve networks from 2D sketches via selective regularization," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 131:1–131:13, Dec. 2014.
[37] H. Pan, Y. Liu, A. Sheffer, N. Vining, C.-J. Li, and W. Wang, "Flow aligned surfacing of curve networks," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–10, Jul. 2015.
[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
[39] T. Saito and T. Takahashi, "Comprehensible rendering of 3-D shapes," in *Proc. 17th Annu. Conf. Comput. Graph. Interact. Techn.*, 1990, pp. 197–206.
[40] A. Jin, Q. Fu, and Z. Deng, "Contour-based 3D modeling through joint embedding of shapes and contours," in *Proc. Symp. Interact. 3D Graph. Games*, May 2020, pp. 1–10.
[41] C. Li, H. Pan, Y. Liu, X. Tong, A. Sheffer, and W. Wang, "BendSketch: Modeling freeform surfaces through 2D sketching," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017.
[42] X. Han, C. Gao, and Y. Yu, "DeepSketch2Face: A deep learning based sketching system for 3D face and caricature modeling," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, Jul. 2017.
[43] H. Tang *et al.*, "Attribute-guided sketch generation," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–7.
[44] C. Peng, N. Wang, J. Li, and X. Gao, "Universal face photo-sketch style transfer via multiview domain translation," *IEEE Trans. Image Process.*, vol. 29, pp. 8519–8534, Aug. 2020.
[45] A. X. Chang *et al.*, "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*. [Online]. Available: https://arxiv.org/abs/1512.03012
[46] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy, "Sketch me that shoe," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 799–807.
[47] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. ICONIP*, 2015, pp. 2017–2025.
[48] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[49] Y. Sun, Y. Wang, Z. Liu, J. E. Siegel, and S. E. Sarma, "PointGrow: Autoregressively learned point cloud generation with self-attention," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 61–70.
[50] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," 2019, *arXiv:1904.11492*. [Online]. Available: http://arxiv.org/abs/1904.11492
[51] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[52] M. Kazhdan and H. Hoppe, "Screened Poisson surface reconstruction," *ACM Trans. Graph.*, vol. 32, no. 3, pp. 1–13, Jun. 2013.
[53] Y. Güçlütürk, U. Güçlü, R. van Lier, and M. A. van Gerven, "Convolutional sketch inversion," in *Proc. ECCV*, 2016, pp. 810–824.
[54] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5400–5409.
[55] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4460–4470.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.
[57] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. NeurIPS*, 2017, pp. 5099–5108.

**Yue Zhong** (Member, IEEE) is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. Her research interests include computer vision, with particular focus on human free-hand sketches-based 3D shape reconstruction, and how they can be transferred into novel commercial applications.

**Yonggang Qi** (Member, IEEE) received the Ph.D. degree in signal processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2015. He was a joint Ph.D. Student with the SketchX Lab, Queen Mary University of London (QMUL). He also worked as a Guest Ph.D. Student with Aalborg University, Demark, in 2013, and a Visiting Researcher with Sun Yat-sen University, China, in 2014. He is currently an Assistant Professor (a Lecturer) with BUPT. His research interests include perceptual grouping, sketch-based tasks evolving sketch-based image retrieval (SBIR), sketch recognition, sketch generation, and language-based sketch understanding.

**Yulia Gryaditskaya** (Member, IEEE) received the Diploma degree from Lomonosov Moscow State University, Moscow, Russia, and the Ph.D. degree in computer vision and graphics from the Max-Planck Institute for Informatics, Saarbruecken, Germany, in 2017. She was a Post-Doctoral Researcher with Inria, Sophia Antipolis, France. She is currently a Senior Research Fellow in computer vision and machine learning with the Centre for Vision Speech and Signal Processing (CVSSP), U.K.'s Largest Academic Research Center for Artificial Intelligence with approx. 200 researchers.

**Honggang Zhang** (Senior Member, IEEE) received the B.S. degree from the Department of Electrical Engineering, Shandong University, in 1996, and the master's and Ph.D. degrees from the School of Information Engineering, Beijing University of Posts and Telecommunications (BUPT), in 1999 and 2003, respectively. He was a Visiting Scholar with the School of Computer Science, Carnegie Mellon University, from 2007 to 2008. He is currently an Associate Professor and the Director of the Web Search Center, BUPT. He has published more than 30 articles on TPAMI, SCIENCE, machine vision and applications, AAAI, ICPR, and ICIP. His research interests include image retrieval, computer vision, and pattern recognition.

**Yi-Zhe Song** (Senior Member, IEEE) received the M.Sc. degree from the University of Cambridge in 2004 and the Ph.D. degree in computer vision and machine learning from the University of Bath, in 2008. He was a Senior Lecturer with the Queen Mary University of London, and a Research and Teaching Fellow with the University of Bath. He is currently a Reader of Computer Vision and Machine Learning with the Centre for Vision Speech and Signal Processing (CVSSP), U.K.'s Largest Academic Research Centre for Artificial Intelligence with approx. 200 researchers. He received the Best Dissertation Award from his M.Sc. degree from the University of Cambridge in 2004, after getting a First Class Honours degree from the University of Bath in 2003. He is a fellow of the Higher Education Academy. He is a Full Member of the Review College of the Engineering and Physical Sciences Research Council (EPSRC), the U.K.'s main agency for funding research in engineering and the physical sciences, and serves as an Expert Reviewer for the Czech National Science Foundation.