Towards Implicit Text-Guided 3D Shape Generation: Supplementary Material

Zhengzhe Liu[1] Yi Wang[2] Xiaojuan Qi[3*] Chi-Wing Fu[1*]
[1]The Chinese University of Hong Kong [2]Shanghai AI Laboratory [3]The University of Hong Kong
{zzliu,cwfu}@cse.cuhk.edu.hk wangyi@pjlab.org.cn xjqi@eee.hku.hk

1. Evaluation Metrics

This section introduces the evaluation metrics employed in the experiments. Below, we first introduce the metrics we formulated/extended from the existing ones for the evaluations, and then introduce other metrics from Text2Shape [3].

- **PS** and **FPD**

  Point Score (PS) and Fréchet Point Distance (FPD) measure shape diversity and quality.

  Existing works [10, 12] often utilize the Fréchet Point Distance (FPD) to evaluate the quality of the generated 3D shapes. However, such metric cannot account for color, which is one of the important characteristics in our results that is not present in the previous works. To jointly evaluate the shape and color in the generated results, we formulate PS and extend FPD for shape and color evaluations based on Inception Score (IS) [13] and Fréchet Inception Distance (FID) [7].

  PS measures the KL-divergence between the conditional probabilities of the generated shapes and their marginal probabilities. On the other hand, FPD measures the Wasserstein distance between the distribution of the generated shapes and that of the real samples. The mentioned probabilities are inferred from a pre-trained classification network (*e.g.*, Inception v3 [13] on ImageNet for image generation). In our case, PS and FPD are built upon a newly-trained PointNet [11], since no existing 3D classification network can simultaneously consider both shape and color as far as we know. Specifically, we train a classification-based PointNet on ScanObjectNN [14] for 200 epochs, with a validation classification accuracy of 84.85%.

- **IoU**

  Intersection over Union (IoU) measures the similarity to the ground truth. We evaluate the IoU between the generated shape and ground truth, by measuring the similarity of the occupancy between them.

- **R-precision**

  We adopt **R-precision** [15] to measure the consistency between the generated shape **S** and input text **T**. Specifically, we extract shape and text features using E and B, respectively, then evaluate R-Precision three times with different random seeds to reduce the randomness.

- Metrics in Text2Shape [3]

  Chen *et al*. [3] adopt four metrics, including **IoU**, **EMD**, **IS**, and **Acc** (**Err**=1-Acc), for evaluating their results. IoU and EMD measure the shape and color similarity between the generated shape and ground truth, respectively. IS measures the diversity and quality of the generated shapes, and Err (Acc) measures the quality. Please refer to [3] for the details. For a pair comparison, we train a classification model using the official code released by the author of [3] to evaluate our IS and Acc.

2. Details of the Baselines

2.1. Text-Guided Shape Generation

We create the following baselines to evaluate the key modules of our text-guided shape generation framework, including the auto-encoder (AE), decoupled shape-color decoder (DSCD), WLST module (WLST), and cyclic loss (CL).

(i) "Without AE." In this setting, the network is composed of text encoder B and decoder D but without shape encoder E. It is optimized to regress the the occupancy and color values of the target shape I with an L_2 loss. It serves as our preliminary baseline for text-guided shape generation, where B maps the text **T** to a latent space and D reconstructs the shape and color.

(ii) "+AE." The network is composed of the auto-encoder E, D_{share}, and text encoder B. E encodes the input shape I as a joint shape-color feature f_{share} and D reconstructs the shape and color of I. B extracts the text feature $\bar{f}_{share}$ and minimizes the mean squared difference between $\bar{f}_{share}$ and f_{share}. It serves as a baseline to directly adopt the auto-encoder-based approach [4] to our task, as introduced in Section 1 of the main paper, so this baseline demonstrates the necessity of the auto-encoder in our approach.

(iii) "Further +DSCD." In this setting, we replace the shared decoder D_{share} with a pair of decoupled shape-color decoders $D = \{D_s, D_c\}$ and replace shared features f_{share} and $\hat{f}_{share}$ in (ii) with a pair of decoupled features $f = \{f_s, f_c\}$ and $\hat{f} = \{\hat{f}_s, \hat{f}_c\}$, respectively. This baseline manifests the effectiveness of the decoupled shape-color decoder (DSCD) in our framework.

(iv) "Further +WLST." Based on (iii), the spatial-aware decoder D' with the WLST is adopted for text-guided shape generation in place of D. This baseline manifests the effectiveness of WLST.

(v) "Further +CL" (our full model). Model (iv) is trained with an additional cyclic loss (which is Eq.(5) in the main paper), whereas Model (v) is our full model. Comparing between model (iv) and model (v) verifies the applicability of the cyclic loss.

2.2. Diversified Generation

To evaluate the core modules for diversified shape generation, We compare our Shape IMLE with two other approaches: (i) Latent GAN and (ii) fully-connected IMLE (FC IMLE). Besides, we evaluate the performance gain of our proposed WLST and cyclic loss (CL) for diversified shape generation. For each baseline, we generate three different samples for each text with random noises z_1 to z_3.

(i) "Latent-GAN." We adopt Latent-GAN [1, 2] conditioned on the input text to generate diversified results. We adopt our style-based latent shape-IMLE generator G (Figure 5 in the main paper) as generator and a small network with three fully-connected layers as discriminator D_{latent}. We train the generator and discriminator iteratively using adversarial training with all the other modules frozen.

(ii) "FC IMLE." In this setting, we introduce the IMLE framework for diversified generation. As shown in Figure 1, the IMLE generator G_{simple} is composed of six fully-connected layers that take $\bar{f} \oplus z$ as input. This baseline aims to show the superiority of IMLE.

(iii) "Shape IMLE." In place of FC-IMLE in (ii), we adopt the style-based shape-IMLE generator G shown in Figure 5 in the main paper. This baseline manifests the effectiveness of our shape-IMLE generator G.

(iv) "+WLST" and "+CL." Similar to "+WLST" and "+CL" in Section 2.1, we again evaluate their capability on improving the diversified generation.

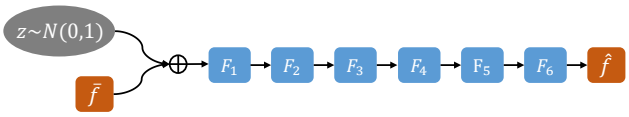

Figure 1. The FC generator architecture in FC IMLE.

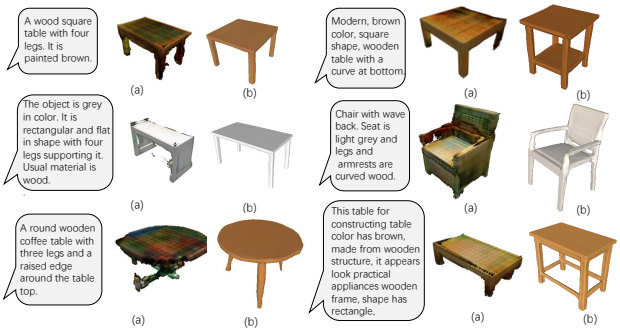

Figure 2. Additional text-guided generation results compared with Text2Shape [3].

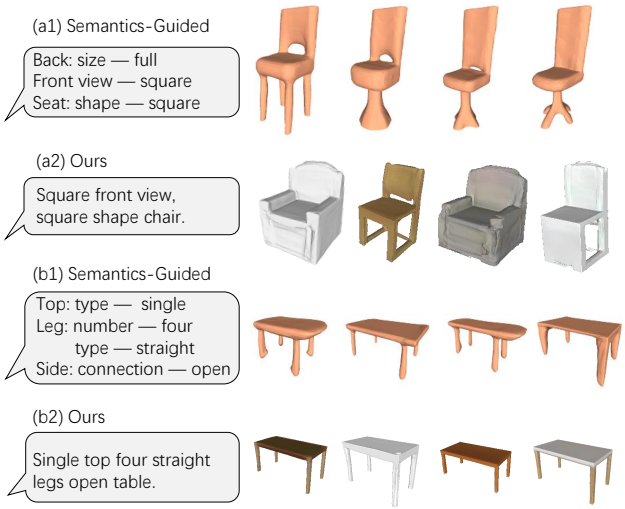

Figure 3. Additional text-guided generation results compared with [8].

3. Results of Text-Guided Shape Generation

3.1. Comparison with Existing Works

In this section, we show more results on comparing our method with [3] and [8]. As shown in Figure 2, our approach is able to generate shapes with higher fidelity compared with [3]. Also, our results are more consistent to the input texts. As shown in Figure 2 (e) on the bottom left of the figure, our approach is able to create a folding chair following the text description, where [3] can only output a regular chair.

The most recent work [8] takes only pre-defined semantic labels as inputs, unlike our approach, which can take natural language as inputs. As shown in Figure 3, our approach can

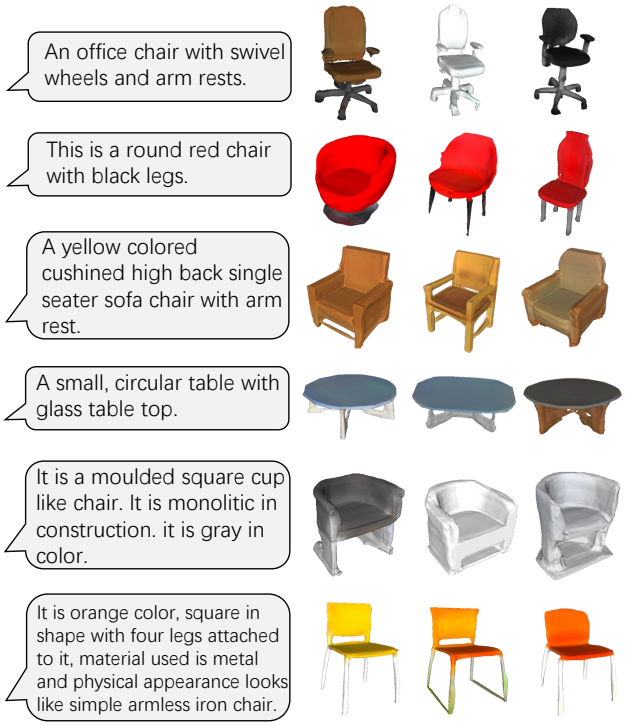

Figure 4. Our text-guided shape generation results.

generate more diversified shapes that better match the input text description ("square shape, square view" in Figure 3 (a1, a2)), compared with [8].

3.2. Additional Generation Results

Further, we show more text-guided shape generation results in Figures 4. These results again manifest the superiority of our approach on diversity, fidelity, and text-shape consistency, demonstrating the capability of our method over the previous ones.

4. Text-Guided Shape Manipulation

4.1. Color Manipulation Framework

In this section, we introduce our framework for text-guided color manipulation with shape unchanged. As shown in Figure 5, we feed shape feature $\bar{f}_{1s}$ (extracted from $\mathbf{T}_1$) and color feature $\bar{f}_{2c}$ (extracted from $\mathbf{T}_2$) to G_3 to predict the manipulated feature $\hat{f}_{1s}, \hat{f}_{2c}$, then feed it to D' to produce the edited shape $\dot{S}$. We then extract manipulated feature $\dot{f} = \{\dot{f}_s, \dot{f}_c\}$ from $\dot{S}$ using E and use the two-way cyclic loss, *i.e.*, L_{cyc_s} to encourage shape consistency ($\dot{f}_s$ and $\hat{f}_{1s}$) and L_{cyc_c} to encourage color consistency ($\dot{f}_c$ and $\hat{f}_{2c}$).

Similar to the shape manipulation framework shown in the main paper, we train our color manipulation framework using the following loss:

$$L^c_{mani} = (||\dot{f}_s - \hat{f}_{1s}||_2^2 + ||\dot{f}_c - \hat{f}_{2c}||_2^2)\mathbb{1}(\text{IoU}(I_1, I_2) > t) + L_{G_1} + L_{G_2}, \tag{1}$$

where the terms have the same definition as Eq.(8) in the main paper.

4.2. Comparison with the Existing Work

In this section, we compare our method with [3] on shape manipulation capability. As shown in Figure 6, inserting or editing words related to the color attribute leads to undesirable changes in the other attributes, as shown in the results produced by [3], *e.g.*, the shape of the chair back and the table leg, whereas our approach is able to better preserve the shapes (geometries and structures).

4.3. Ablation Studies

In this section, we evaluate the different strategies for text-guided manipulation quantitatively. To measure the quality of the manipulated shapes, We adopt PS and FPD as the evaluate metrics. To further evaluate the consistency before and after the manipulation, we calculate R-Precision$_1$ based on $\dot{f}$ (feature from the manipulated shape) and $\hat{f}_1$ (feature from the original text), and assess R-Precision$_2$ based on $\dot{f}$ (feature from the manipulated shape) and $E(D'(\hat{f}_1))$ (feature from the generated shape by the original text). We build a small dataset containing 50 pairs of original and manipulated texts for the evaluation.

(i) Baseline 1. As shown in Figure 7(b) of the main paper, we directly feed the feature from the edited text $\hat{f}_2 = \{\hat{f}_{2s}, \hat{f}_{2c}\}$ to our generation framework. It is the primitive baseline for manipulation because it adopts no mechanism for consistency preserving, but it serves as the upper bound of the shape manipulation quality, since it directly adopts our generation framework (Figure 2 in the main paper) to produce the result without any constraints on the manipulation consistency.

(ii) Baseline 2. As shown in Figure 7(c) of the main paper, the shape is generated by a mixture of $\hat{f}_1$ and $\hat{f}_2$. Specifically, for the shape manipulation, we feed $\hat{f}_{2s} \oplus \hat{f}_{1c}$ to D'; and for the color manipulation, we use $\hat{f}_{1s} \oplus \hat{f}_{2c}$.

(iii) Baseline 3. As shown in Figure 7(d) of the main paper, we feed a mixture of $\bar{f}_1$ and $\bar{f}_2$ to G to boost the shape-color alignment. For the shape manipulation, we feed $\bar{f}_{2s} \oplus \bar{f}_{1c}$ to G to derive $\hat{f}_{2s}, \hat{f}_{1c}$; and for the color manipulation, we predict $\hat{f}_{1s}, \hat{f}_{2c}$ from $\hat{f}_{1s} \oplus \hat{f}_{2c}$ with G.

(iv) Our full model (Ours). Built upon (iii), we further incorporate the two-way cyclic loss shown in Eq.(8) of

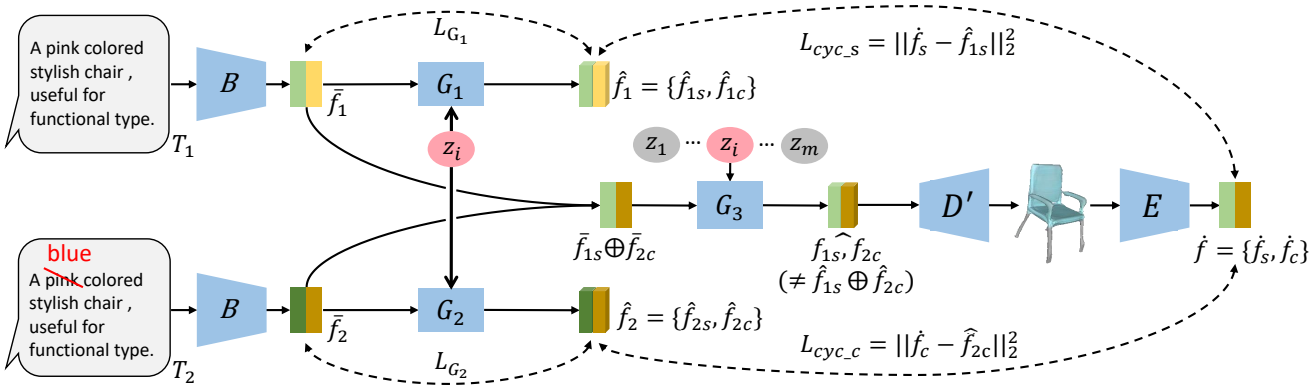

Figure 5. Overview of our text-guided color manipulation framework (with shape unchanged). Given two pieces of text $\mathbf{T}_1, \mathbf{T}_2$, shape IMLE G_1 and G_2 use the same random noise z_i for shape generation. G_3 takes $\{\bar{f}_{1s}, \bar{f}_{2c}\}$ and z_i as input to generate shape $\dot{\mathbf{S}}$ with feature $\{\dot{f}_s, \dot{f}_c\}$ (encoded by E), such that $\dot{f}_s$ and $\dot{f}_c$ should be similar to $\hat{f}_{1s}$ and $\hat{f}_{2c}$, respectively. To this end, we propose a two-way cyclic loss to encourage the shape consistency between $\dot{\mathbf{S}}$ and $\mathbf{T}_1$, and the color consistency between $\dot{\mathbf{S}}$ and $\mathbf{T}_2$. G_1, G_2, G_3 share the same weights.

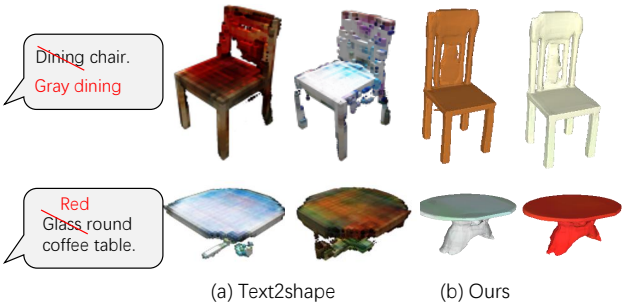

Figure 6. Text-guided manipulation results on comparing our method with Text2Shape [3].

Table 1. Ablation studies on text-guided manipulation.

Method	PS (↑)	FPD (↓)	R-Precision$_1$ (↑)	R-Precision$_2$ (↑)
Baseline 1	**2.80 ± 3.03**	**31.70**	36.00 ± 2.83	48.67 ± 0.94
Baseline 2	2.73 ± 0.39	33.77	43.33 ± 0.94	52.00 ± 3.26
Baseline 3	2.75 ± 0.48	35.74	42.66 ± 3.77	56.67 ± 2.49
Ours	2.76 ± 0.53	32.03	**58.00 ± 2.82**	**67.33 ± 1.89**

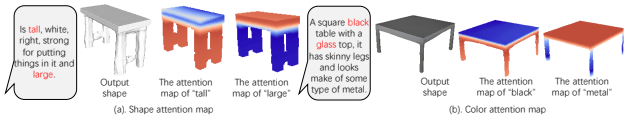

Figure 7. Attention map of the end-to-end training strategy.

the main paper for shape manipulation, and Eq. (1) in this supplementary document for color manipulation.

"Baseline 1" generates a new shape using the edited text without considering what the original shape is. As shown in Table 1, despite of the best diversity and quality it achieves, the lowest R-Precisions indicate the unsatisfying consistency before and after the manipulation (see Figure 7 (b) in the main paper).

On the other hand, "Baseline 2" and "Baseline 3" attain better consistency at the expense of the generation quality, and our manipulation framework with the two-way cyclic loss is able to achieve the best consistency before and after the manipulation, while having better generation quality compared with both "Baseline 2" and "Baseline 3," even being close to "Baseline 1."

5. Alternative Training Strategy

In this section, we discuss an optional training strategy. Specifically, we jointly train the shape auto-encoder and text encoder E, D' and B end-to-end, instead of following the training strategy presented in the main paper that first trains E, D, and then jointly trains E, D', B. This strategy includes fewer training steps, but needs much more training time because B continuously optimized in the whole training process. This training strategy achieves comparable performance as presented in the main paper, and can generate attention maps that are more consistent with the semantic meaningful shape parts as shown in Figure 7.

6. Limitations and Future Work

Our current approach still has some limitations. First, text-guided shape generation is a very challenging task as discussed in Section 2 of the main paper. For example, some attributes, such as "ten legs" in Figure 8 (a), are extremely challenging to generate. Our framework fails to generate a shape that faithfully follows such description. Second, shapes with long, thin, and fine structures may get distorted or become noisy, as shown in Figure 8 (b, c). To address this issue, we plan to explore more recent 3D implicit representations [5, 6, 9]. Also, the manipulation performance is limited by the inherent bias of the dataset. If we add an armrest to the chair in Figure 8 (d), the manipulated chair will have become wider than the original one. Such a result is par-

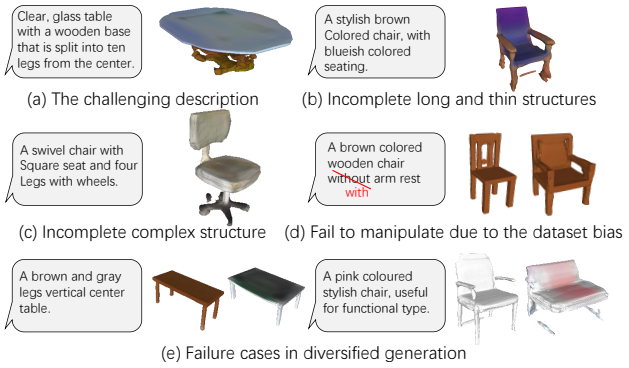

Figure 8. Illustrating the limitations of our current approach.

tially due to the dataset bias, since armed chairs are typically wider (like sofa) than those without armrests in the dataset. To resolve it, the manipulation process requires a topological understanding of the shapes. In addition, our metrics have some limitations. On the one hand, IoU may not be a good metric for text-to-shape generation task, because a slight difference in height/position between the generated shapes and GT shape can cause a low IoU; particularly, this is beyond the representative ability of a small piece of text. On the other hand, PS and FPD cannot fully reflect the generative quality, because these two metrics are based on the ScanObjectNN dataset [14], which has a large domain gap from our training dataset ShapeNet. In other words, better PS and FPD simply indicate that the generated shapes are more similar to the ScanObjectNN shapes, not necessarily meaning better quality. Also, there is a trade-off between the diversity and fidelity in our diversified generation. When stronger noise added to encourage the diversity, the quality of some generations cannot be ensured, and some are inconsistent with the text description as shown in Figure 8 (e). It gets more serious when the text is long and contains descriptions on shape details. Last, our approach needs paired text-shape data for training, so we temporarily only explore shapes of table and chair, since the largest existing dataset [3] only provides samples of these two categories. In the future, we will plan to explore zero-shot text-guided shape generation to extend the applicability of this work.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J. Guibas. Learning representations and generative models for 3D point clouds. In *ICML*, 2018.

[2] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *ICLR*, 2017.

[3] Kevin Chen, Christopher B. Choy, Manolis Savva, Angel X. Chang, Thomas Funkhouser, and Silvio Savarese. Text2Shape: Generating shapes from natural language by learning joint embeddings. In *ACCV*, 2018.

[4] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019.

[5] Zhang Chen, Yinda Zhang, Kyle Genova, Sean Fanello, Sofien Bouaziz, Christian Hane, Ruofei Du, Cem Keskin, Thomas Funkhouser, and Danhang Tang. Multiresolution deep implicit functions for 3d shape representation. In *ICCV*, 2021.

[6] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3D shape. In *CVPR*, 2020.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *NIPS*, 2017.

[8] Tansin Jahan, Yanran Guan, and Oliver van Kaick. Semantics-guided latent space exploration for shape generation. In *COMPUT GRAPH FORUM*, 2021.

[9] Manyi Li and Hao Zhang. D^2IM-Net: Learning detail disentangled implicit fields from single images. *CVPR*, 2021.

[10] Ruihui Li, Xianzhi Li, Ka-Hei Hui, and Chi-Wing Fu. SP-GAN: sphere-guided 3D shape generation and manipulation. *ACM TOG (SIGGRAPH)*, 2021.

[11] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017.

[12] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3D point cloud generative adversarial network based on tree structured graph convolutions. In *ICCV*, 2019.

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

[14] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, 2019.

[15] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018.