Adversarial Open Domain Adaptation for Sketch-to-Photo Synthesis

Xiaoyu Xiang[1]*, Ding Liu[2], Xiao Yang[2], Yiheng Zhu[2], Xiaohui Shen[2], Jan P. Allebach[1]
[1]Purdue University, [2]ByteDance Inc.
{xiang43,allebach}@purdue.edu,
{liuding,yangxiao.0,yiheng.zhu,shenxiaohui}@bytedance.com

Abstract

In this paper, we explore open-domain sketch-to-photo translation, which aims to synthesize a realistic photo from a freehand sketch with its class label, even if the sketches of that class are missing in the training data. It is challenging due to the lack of training supervision and the large geometric distortion between the freehand sketch and photo domains. To synthesize the absent freehand sketches from photos, we propose a framework that jointly learns sketch-to-photo and photo-to-sketch generation. However, the generator trained from fake sketches might lead to unsatisfying results when dealing with sketches of missing classes, due to the domain gap between synthesized sketches and real ones. To alleviate this issue, we further propose a simple yet effective open-domain sampling and optimization strategy to "fool" the generator into treating fake sketches as real ones. Our method takes advantage of the learned sketch-to-photo and photo-to-sketch mapping of in-domain data and generalizes it to the open-domain classes. We validate our method on the Scribble and SketchyCOCO datasets. Compared with the recent competing methods, our approach shows impressive results in synthesizing realistic color, texture, and maintaining the geometric composition for various categories of open-domain sketches.

1. Introduction

Freehand sketch is an intuitive way for users to interact on visual media and express their intentions. The popularization of touch screens provides more and more scenarios for sketch-based application, *e.g.* sketch-based photo-editing [52, 12, 26, 47, 62], sketch-based image retrieval for 2D images [65, 56, 38, 66, 63, 50, 14, 10, 15, 3, 37, 2] and 3D shapes [59, 68, 11, 61, 5], and 3D modeling from sketches [43, 22, 54].

Sketch-to-photo translation aims to automatically translate a sketch in the source domain S to the target photo-realistic domain P. Many existing works [25, 9, 42, 19, 34, 18, 6, 35] adopt generative adversarial networks (GAN) [20] to learn the sketch-to-image process from paired data. However, the sketch-to-photo translation task suffers from the open-domain adaptation problem, where the majority of data is unlabeled and unpaired [16, 36, 32, 21, 71, 4, 33], and the freehand sketch covers only a small portion of the photo categories [53, 65, 56, 41, 19] due to the fact that they require a large number of human annotations. Therefore, some works [25, 34, 6, 35] use edges extracted from the target photos as substitution. Still, edges and freehand sketches are very different: freehand sketches are human abstractions of an object, usually with more deformations. Due to this domain gap, models trained on the edge inputs easily fail to generalize to freehand sketches. A good sketch-based image generator should not only fill the correct textures within the lines, but also correct the object structure conditioned on the input composition.

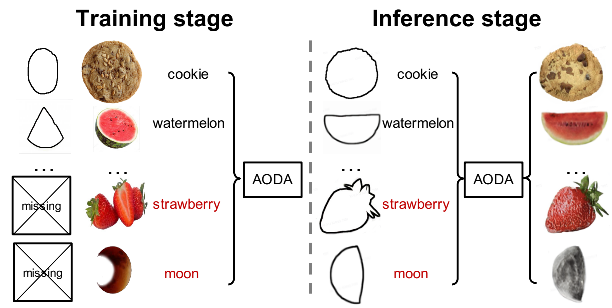

Figure 1: Illustration of open-domain sketch-to-photo synthesis problem. During the training stage of multi-class sketch-to-photo generation, sketches of some categories are missing. In the inference stage, our algorithm synthesizes photos from the input sketches for not only known classes, but also the classes that were missing during the training.

Well-labeled freehand sketches and photos can help the translation model better understand the geometry correspondence. In recent years, [69, 39, 24, 31, 29, 41] aim to learn from unpaired sketches and photos collected separately. Even so, the existing sketch datasets cannot cover all types of photos in the open domain [48]: the largest

*The author was with Purdue University when conducting the work in this paper during an internship at ByteDance. She is now with Facebook.

sketch dataset *Quick, Draw!* [21] has 345 categories, while the full ImageNet [13] has as many as 21,841 class labels. Therefore, most categories even lack corresponding freehand sketches to train a sketch-to-image translation model.

To resolve this challenging task, we propose an Adversarial Open Domain Adaptation (AODA) framework that for the first time learns to synthesize the absent freehand sketches and makes the unsupervised open-domain adaptation possible, as illustrated in Figure 1. We propose to jointly learn a sketch-to-photo translation network and a photo-to-sketch translation network for mapping the open-domain photos into the sketches with the GAN priors. With the bridge of the photo-to-sketch generation, we can generalize the learned correspondence between in-domain freehand sketches and photos to open-domain categories. Still, there is an unignorable domain gap between synthesized sketches and real ones, which prevents the generator from generalizing the learned correspondence to real sketches and synthesizing realistic photos for open-domain classes.

To further mitigate its influence on the generator and leverage the output quality of open-domain translation, we introduce a simple yet effective random-mixed sampling strategy that considers a certain proportion of fake sketches as real ones blindly for all categories. With the proposed framework and training strategy, our model is able to synthesize a photo-realistic output even for sketches of unseen classes. We compare the proposed AODA to existing unpaired sketch-to-image generation approaches. Both qualitative and quantitative results show that our proposed method achieves significantly superior performance on both seen and unseen data.

The main contributions of this paper are three-fold: (1) We propose the adversarial open-domain adaptation (AODA) framework as the first attempt to solve the open-domain multi-class sketch-to-photo synthesis problem by learning to generate the missing freehand sketches. (2) We introduce an open-domain training strategy by considering certain fake sketches as real ones to reduce the generator's bias of synthesized sketches and leverage the generalization of adversarial domain adaptation, thus achieving more faithful generation for open-domain classes. (3) Our network provides, as a byproduct, a high-quality freehand sketch extractor for arbitrary photos. Extensive experiments and user studies on diverse datasets demonstrate that our model can faithfully synthesize realistic photos for different categories of open-domain freehand sketches. The source code and pre-trained models are available at https://github.com/Mukosame/AODA.

2. Related Work

Sketch-Based Image Synthesis The goal of sketch-based image synthesis is to output a target image from a given sketch. Early works [7, 17, 8] regard freehand sketches as queries or constraints to retrieve each composition and stitch them into a picture. In recent years, an increasing number of works adopt GAN-based models [20] to learn pixel-wise translation between sketches and photos directly. [69, 34, 6] train their networks with pairs of photos and corresponding edge maps due to the lack of real sketch data. However, the freehand sketches are usually distorted in shape compared with the target photo. Even when depicting the same object, the sketches from different users vary in appearance due to differences in their drawing skills and the levels of abstractness. To make the model applicable to freehand sketches, SketchyGAN [9] trained with both sketches and augmented edge maps. ContextualGAN [42] turns the image generation problem into an image completion problem: the network learns the joint distribution of sketch and image pairs and acquires the result by iteratively traversing the manifold. iSketchNFill [19] uses simple outlines to represent freehand sketches and generates photos from partial strokes with two-stage generators. Gao *et al*. [18] applies two generators to synthesize the foreground and background respectively and proposes a novel GAN structure to encode the edge maps and corresponding photos into a shared latent space. The above works are supervised based on paired data. Liu *et al*. [41] proposes a two-stage model for the unsupervised sketch-to-photo generation with reference images in a single class. Compared with these works, our problem setting is more challenging: we aim to learn the multi-class generation without supervision using paired data from an incomplete and heavily unbalanced dataset.

Conditional Image Generation Image generation can be controlled by class-condition [19, 18], reference images [42, 40, 41], or specific semantic features [27, 51, 70], *etc*. The pioneering work cGAN [44] combines the input noise with the condition for generator and discriminator. To help the generator synthesize images based on the input label, AC-GAN [46] makes the discriminator also predict the class labels. SGAN [45] unifies the idea of discriminator and classifier by including the fake images as a new class. In this paper, we adopt a photo classifier that is jointly trained with the generator and discriminator to supervise the sketch-to-photo generator's training.

3. Adversarial Open Domain Adaptation

First, we discuss the challenge of the open-domain generation problem and the limitation of previous methods in Section 3.1. Then we introduce our proposed solution, including our AODA framework and the proposed training strategy in Section 3.2.

3.1. Challenge

Unlike previous sketch-to-photo synthesis works [9, 19] that can directly learn the mapping between the input sketch and its corresponding photo, during training, the sketches of

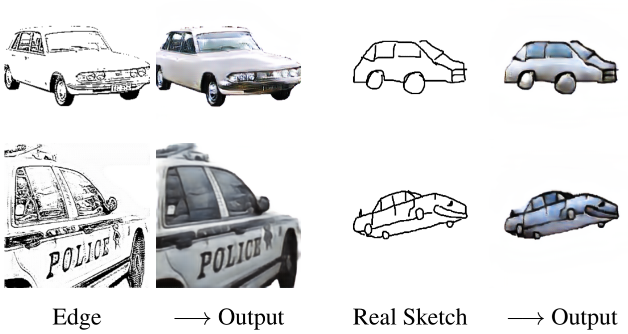

Figure 2: Results of photo synthesis from edge inputs and real sketch inputs generated by a model trained with xDoG edges and photos from the SketchyCOCO dataset [18]. The left two columns show the xDoG inputs and their outputs, and the right two columns are the real freehand sketch inputs and the corresponding unsatisfactory outputs, which shows that the model simply trained with edges cannot rectify the distorted shapes of freehand sketches.

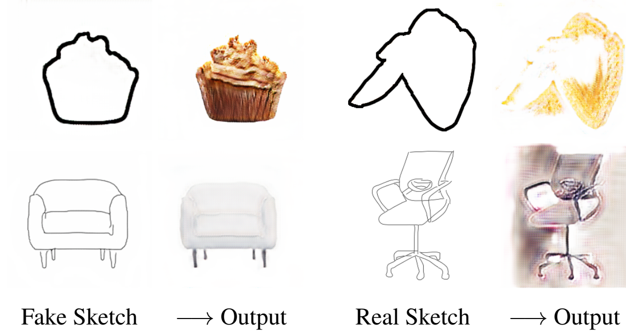

Figure 3: Results of photo synthesis from fake sketch inputs and real sketch inputs on Scribble [19] and QMUL-Sketch datasets [65, 56, 41]. The outputs are generated by a model trained with synthesized sketches, and the setting remains the same as in [41], where the fake sketches are generated using a sketch extractor trained on the in-domain data. The left two columns show the fake sketch inputs and their outputs, and the right two columns are the real freehand sketch inputs and the corresponding unsatisfactory outputs. Comparing the outputs, we can see this training strategy makes the model fail to generalize on real sketches.

open-domain classes are missing. To enable the network to learn to synthesize photos from sketches of both in-domain and open-domain classes, there are two ways to solve this problem: (1) training with extracted edge maps and (2) enriching the open-domain classes with synthesized sketches from a pre-trained photo-to-sketch extractor. We show the results of these two methods and discuss their limitations.

Edge Maps. Figure 2 shows the results of a model trained on edges extracted by XDoG [60]. While the model can generate vivid highlights and shadows and fine details from the edge inputs, the images generated from the actual freehand sketches are not that photo-realistic, but more like a colored drawing. This is because edges and freehand sketches are very different: freehand sketches are human abstractions of an object, usually with more deformations. The connections between the target photos and the input sketch are looser than with edges. Due to this domain gap, sketch-to-photo generators trained on the edge inputs usually cannot learn shape rectification, and thus fail to generalize to freehand sketches.

Synthesized sketches. Another intuitive solution for open-domain generation is to enrich the training set of unseen classes $\mathcal{M}$ with sketches synthesized by a pre-trained photo-to-sketch generator [41]. Figure 3 shows the result from a model trained with pre-extracted sketches on Scribble [19] and QMUL-Sketch dataset [65, 56, 41], where the photo-to-sketch extractor is trained with the in-domain classes of the training set. From the left two columns in Figure 3, we can see that the model is able to generate photo-realistic outputs from synthesized sketches. However, it fails on real freehand sketches, as shown in the right two columns: even though it can generate the correct color and texture conditioned by the input label, it cannot understand the basic structure of real sketches (*e.g.* distinguish the object from the background). The reason is, even though they are visually similar, the real and fake sketches are still distinguishable for the model. This strategy cannot guarantee that the model can generalize from the synthesized data to the real freehand sketches, especially for the multi-class generation. Thus, simply using the synthesized sketch to replace the missing freehand sketches cannot ensure photo-realistic generation.

3.2. Our Method

To solve this problem, we propose to learn the photo-to-sketch and sketch-to-photo translation jointly and to narrow the domain gap between the synthesized and real sketches.

Framework. As shown in Figure 4, our framework mainly consists of the following parts: *two generators*: a photo-to-sketch generator G_s, and a multi-class sketch-to-photo generator G_p that takes sketch s and class label η_s as input; *two discriminators* D_s and D_p that encourage the generators to synthesize outputs indistinguishable from the sketch domain S and photo domain P, respectively; and a *classifier* R that predicts class labels for both real photos p and fake photos $G_p(s, \eta_s)$ to ensure that the output is truly conditioned on the input label η_s. Our AODA framework is trained with the unpaired sketch and photo data.

During the training process, G_p extracts the sketch $G_s(p)$ from the given photo p. Then, the synthesized sketch $G_s(p)$ and the real sketch s are sent to G_p along with their labels η_p and η_s, and turned into the reconstructed photo $G_p(G_s(p), \eta_p)$ and the synthesized photo $G_p(s, \eta_s)$, respec-

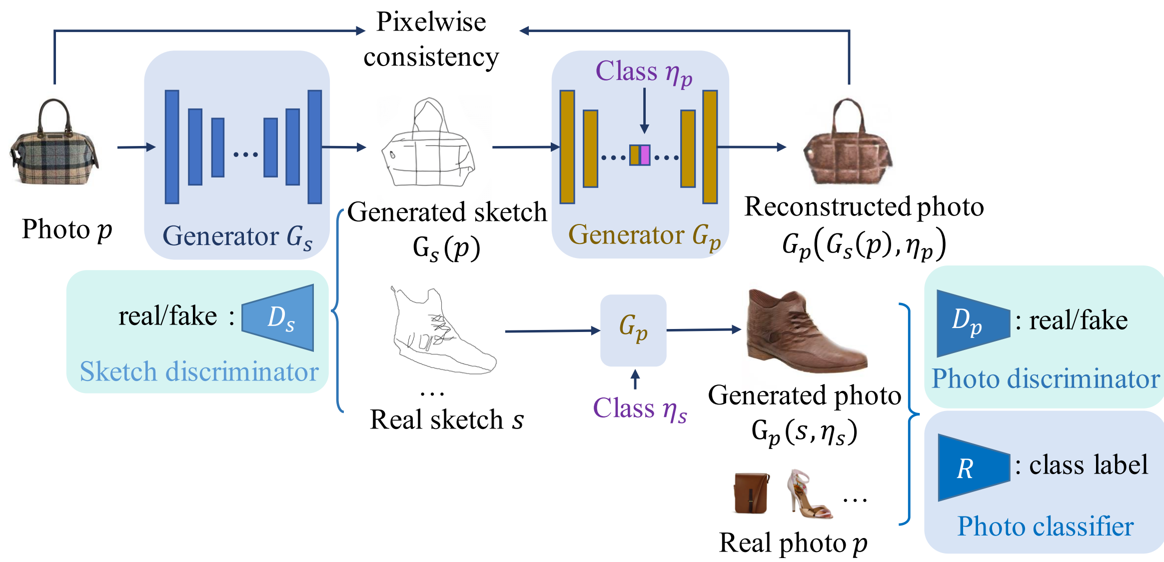

Figure 4: AODA framework overview. It has two generators G_s : photo $\rightarrow$ sketch and G_p : sketch $\rightarrow$ photo conditioned on the input label, and two discriminators D_s and D_p for the sketch and photo domains, respectively. In addition, we use a photo classifier R to encourage G_p to generate indistinguishable photos from the real ones of the same class.

tively. Note that we only send the sketch with its true label to ensure that G_p learns the correct shape rectification from sketch to image domain for each class. The reconstructed photo is supposed to look similar to the original photo, which is imposed by a pixel-wise consistency loss. We do not add such a consistency constraint onto the sketch domain since we wish the synthesized sketches to be diverse. The generated photo is finally sent to the discriminator D_p to ensure that it is photo-realistic, and to the classifier R to ensure that it has the same perceptual features as the target class. In summary, the generator loss includes four parts: the adversarial loss of photo-to-sketch generation $\mathcal{L}_{G_s}$, the adversarial loss of sketch-to-photo translation $\mathcal{L}_{G_p}$, the pixel-wise consistency of photo reconstruction $\mathcal{L}_{pix}$, and the classification loss for the synthesized photo $\mathcal{L}_\eta$:

$$\mathcal{L}_{GAN} = \lambda_s \mathcal{L}_{G_s}(G_s, D_s, p) + \lambda_p \mathcal{L}_{G_p}(G_p, D_p, s, \eta_s) + \lambda_{pix} \mathcal{L}_{pix}(G_s, G_p, p, \eta_p) + \lambda_\eta \mathcal{L}_\eta(R, G_p, s, \eta_s). \quad (1)$$

Please see our supplementary materials for more details. All parts of our framework are trained jointly from scratch.

However, if we directly train the multi-class generator with the loss defined in Equation 1, the training objective for open-domain classes $\mathcal{M}$ becomes the following form due to the missing sketches s:

$$\mathcal{L}_{GAN}^{\mathcal{M}} = \lambda_s \mathcal{L}_{G_s}(G_s, D_s, p) + \lambda_{pix} \mathcal{L}_{pix}(G_s, G_p, p, \eta_p), \quad (2)$$

where $\eta_p \in \mathcal{M}$. As a result, the sketch-to-photo generator G_p is solely supervised by the pixel-wise consistency. Since the commonly used $\mathcal{L}_1$ and $\mathcal{L}_2$ loss lead to the median and mean of pixels, respectively, this bias will make G_p generate blurry photos for the open-domain classes.

To solve this problem, we propose the random-mixed sampling strategy to minimize the domain gap between real and fake sketch inputs for the generator and to improve its output quality with the open-domain classes.

Random-mixed strategy. This strategy aims to "fool" the generator into treating fake sketches as real ones. Algorithm 1 describes the detailed steps for the random-mixed sampling and modified optimization: *Pool* denotes the buffer that stores the minibatch of sketch-label pairs. Querying the pool returns either the current minibatch or a previously stored one (and inserts the current minibatch in the pool) with a certain likelihood. U denotes uniform sampling in the given range, and t denotes the threshold that is set according to the ratio of open-domain classes and in-domain classes to match the photo data distribution.

One key operation of this strategy is to construct pseudo sketches for G_p by randomly mixing the synthesized sketches with real ones in a batch-wise manner. In this step, the pseudo sketches are treated as the real ones by the generator. Thus, the open-domain classes' $\mathcal{L}_{GAN}^{\mathcal{M}}$ becomes:

$$\mathcal{L}_{GAN}^{\mathcal{M}} = \lambda_s \mathcal{L}_{G_s}(G_s, D_s, p) + \lambda_p \mathcal{L}_{G_p}(G_p, D_p, s_{fake}, \eta_p) + \lambda_{pix} \mathcal{L}_{pix}(G_s, G_p, p, \eta_p) + \lambda_\eta \mathcal{L}_\eta(R, G_p, s_{fake}, \eta_p), \quad (3)$$

where $\eta_p \in \mathcal{M}$. Another key of the strategy is on optimization: the sampling strategy is only for G_p. The classifier and discriminators are still updated with real/fake data to guarantee their discriminative powers.

The random mixing operation is blind to in-domain and open-domain classes. As a result, the training sketches include both real and pseudo sketches from all categories. By including pseudo sketches from both the in-domain and

Algorithm 1: Minibatch Random-Mixed Sampling and Updating

Input: In training set $\mathcal{D}$, each minibatch contains photo set p, freehand sketch set s, the class label of photo η_p, and the class label of sketch η_s;

for *number of training iterations* **do**

- $s_{fake} \leftarrow G_s(p)$;
- $s_c \leftarrow s$;
- $\eta_c \leftarrow \eta_s$;
- **if** $t < u \sim U(0,1)$ **then**
  - $s_c, \eta_c \leftarrow pool.\text{query}(s_{fake}, \eta_p)$;
- **end**
- $p_{rec} \leftarrow G_p(s_{fake}, \eta_p)$;
- $p_{fake} \leftarrow G_p(s, \eta_s)$
- Calculate $\mathcal{L}_{GAN}$ with $(p, s_c, p_{rec}, \eta_c)$ and update G_s and G_p;
- Calculate $\mathcal{L}_{D_s}(s, s_{fake})$ and $\mathcal{L}_{D_s}(p, p_{fake})$, update D_s and D_p;
- Calculate $\mathcal{L}_R(p, p_{fake}, \eta_p, \eta_s)$ and update the classifier.

end

open-domain classes, it would further enforce the sketch-to-image generator to ignore the domain gap in the inputs and synthesize realistic photos from both real and fake sketches. Note that since G_s's parameters are consistently updated during training, the pseudo sketches also change for each batch. Moreover, the pseudo sketch-label pairs are acquired from a history of generated sketches and their labels rather than the latest produced ones by G_s. We maintain a buffer that stores the 50 previously added minibatch of sketch-label pairs [55, 69].

Mixing real sketches with fake ones can be regarded as an online data augmentation technique for training G_p. Compared with augmentation using edges, G_s can learn the distortions from real freehand sketches by approaching the real data distribution [20, 30, 67], and enable G_p to learn shape rectification on the fly. Benefiting from the joint training mechanism, as the training progresses, the sketches generated by G_s change epoch by epoch. The loose consistency constraint on sketch generation further increases the diverseness of the sketch data in the open-domain. Compared with using pre-extracted sketches, the open-domain buffer maintains a broad spectrum of sketches: from the very coarse ones generated in early epochs to the more human-like sketches in later epochs as G_s converges.

4. Experiments

4.1. Experiment Setup

Dataset. We train and evaluate the performance of sketch-to-photo synthesis methods on two datasets: Scribble [19] (10 classes), and SketchyCOCO [18] (14 classes of objects).

Scribble contains ten object classes with photos and simple outline sketches. Six out of ten classes have similar round outlines, which imposes more stringent requirements on the network: whether it can generate the correct structure and texture conditioned on the input class label. In the open-domain setting, we only have the sketches of four classes for training: *pineapple*, *cookie*, *orange*, and *watermelon*, which means that 60% of the classes are open-domain.

SketchyCOCO includes 14 object classes, where the sketches are collected from the Sketchy dataset [53], TU-Berlin dataset [16], and *Quick! Draw* dataset [21]. The 14,081 photos for each object class are segmented from the natural images of COCO Stuff [4] under unconstrained conditions, thereby making it more difficult for existing methods to map the freehand sketches to the photo domain. The two open-domain classes are: *sheep* and *giraffe*.

Metrics. We quantitatively evaluate the generation results with three different metrics: 1) Fréchet Inception Distance (FID) [23] that measures the feature similarity between generated and real image sets. A Low FID score means the generated images are less different from the real ones and thus have high fidelity; 2) Classification Accuracy (Acc) [1] of generated images with a pre-trained classifier in the same manner as [19, 18]. Higher accuracy indicates better image realism; 3) User Preference Study (Human): we show the participants a given sketch and the class label, and ask them to pick one photo with the best quality and realism from generated results. We randomly sample 31 groups of images. For each evaluation, we shuffle the options and show them to 25 users. We collect 775 answers in total.

4.2. Sketch-to-Photo Synthesis

4.2.1 Comparison to Other Methods

To better illustrate the effectiveness of our proposed solution, here we adopt CycleGAN [69] as the baseline in building our network and include the original CycleGAN in the following comparison. To make it be able to accept sketch class labels, we modified the sketch-to-photo translator to be a conditional generator. Besides, we also compare to a recent work EdgeGAN [18] on each dataset. We mark the open-domain sketch with a ★ for better visualization.

Scribble. Figure 5 shows the qualitative results of (a) CycleGAN, (b) conditional CycleGAN, (c) conditional CycleGAN with classification loss, (d) EdgeGAN and our method, where the bottom three rows are open-domain. The original CycleGAN exhibits mode collapse and synthesizes identical textures for all categories, probably due to the fact that the sketches in the Scribble dataset barely imply their class labels. This problem is alleviated in (b). Still, it fails to synthesize natural photos for some categories due to the gap between open-domain and in-domain data. Such a domain gap is even worse in (c), where the in-domain result is with

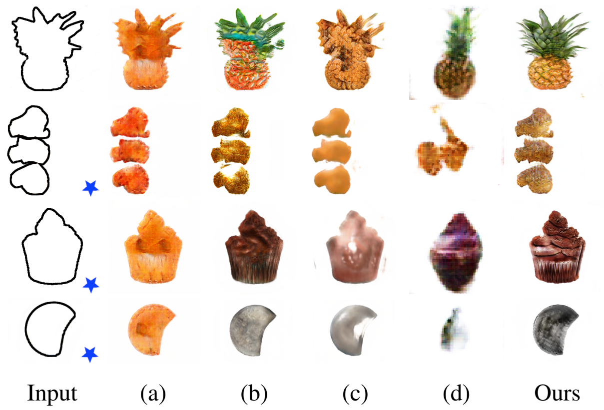

Figure 5: Results on Scribble dataset [19]. We mark the open-domain inputs with ★. The following columns are outputs of (a) CycleGAN [69], (b) conditional CycleGAN, (c) classifier+(b), (d) EdgeGAN [18], and ours.

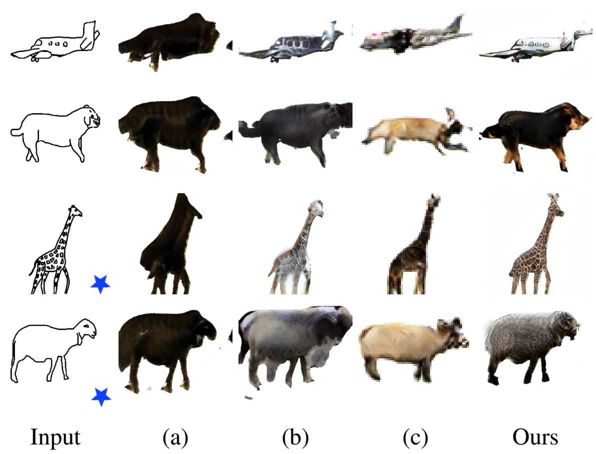

Figure 6: Results on SketchyCOCO dataset [18] for the compared methods: (a) CycleGAN [69], (b) conditional CycleGAN, (c) EdgeGAN [18], and ours. The open-domain inputs are marked with ★.

realistic but wrong texture, and the open-domain results are texture-less. This might be because that classifier implicitly increases the domain gap while maximizing the class discrepancy. Thus, we do not include this model for comparison on the SketchyCOCO dataset. Compared with (d), our results are more consistent with the input sketch shape, demonstrating that our model is better at understanding the composition in sketches and learning more faithful shape rectification in sketch-to-photo domain mapping.

SketchyCOCO The qualitative results are shown in Figure 6, where the top two rows are of in-domain categories, and the bottom two are open-domain. The photos generated by CycleGAN suffer from mode collapse. As shown in column (b), conditional CycleGAN cannot generate vivid textures for open-domain categories. Compared with EdgeGAN in (c), the poses in our generated photos are more faithful to the input sketches.

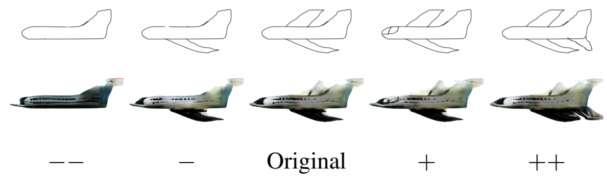

Figure 7: Our model works well for the sketches that are modified by removing strokes (left two columns) and adding strokes (right two columns).

The quantitative results for the three datasets are summarized in Table 1. We can see that our model is preferred by more users than the other compared methods, and achieves the best results in terms of the FID score and classification accuracy on all datasets. These results confirm our observations of the qualitative outputs, as discussed above. Besides, we have an interesting observation: compared with the baseline CycleGAN and conditional CycleGAN, our random-mixed strategy improves not only the open-domain results, but also in-domain results. We find a possible explanation from [57]: the "fake-as-real" operation can effectively alleviate the gradient exploding issue during GAN training and result in a more faithful generated distribution.

4.2.2 Robustness

We test our sketch-to-photo generator's robustness to the inputs and show the visual results in Figure 7: the left two columns show partial sketches that are generated by removing some strokes from the original one, and the right two columns are enriched sketches that are generated by adding extra strokes to the original ones. The original sketch from the SketchyCOCO [18] test set and its output are shown in the middle column. Our model can synthesize realistic airplanes, even when the image composition is changed by adding or removing strokes.

4.3. Photo-to-Sketch Synthesis

As a byproduct, our network can also provide a high-quality freehand sketch generator G_s for a given photo [49, 64, 28]. We run our sketch extractor on COCO objects (top two rows) and web images (bottom two rows) and display the results in Figure 8. Our model can generate different styles of freehand sketches like human drawers beyond the edge map of a photo, even for unseen objects.

Characterized by the joint training, the weights of the photo-to-sketch generator are constantly updated as the training progresses. As a result, the sketches generated by G_s change epoch by epoch. Figure 9 shows the extracted sketches at different epochs. The changing sketches increase the diverseness of the sketch, and thus can further augment the data and help the sketch-to-photo generator to better generalize to various freehand sketch inputs.

Dataset	Method / Metric	CycleGAN [69] full	CycleGAN [69] in-domain	CycleGAN [69] open-domain	conditional CycleGAN full	conditional CycleGAN in-domain	conditional CycleGAN open-domain	EdgeGAN [18] full	EdgeGAN [18] in-domain	EdgeGAN [18] open-domain	Ours full	Ours in-domain	Ours open-domain
Scribble	FID ↓	279.5	252.7	355.9	213.6	210.9	253.6	259.7	256.3	298.5	**209.5**	**204.6**	**252.8**
	Acc (%) ↑	16.0	30.0	6.7	68.0	70.0	66.7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	Human (%) ↑	5.60	1.00	8.67	19.20	17.00	20.67	25.20	17.00	30.67	**48.80**	**65.00**	**38.00**
SketchyCOCO	FID ↓	201.7	218.7	237.2	124.3	138.7	171.6	169.7	177.8	221	**114.8**	**128.4**	**139.2**
	Acc (%) ↑	8.4	10.8	1.9	57.0	58.7	52.4	75.8	68.8	98.3	**78.3**	**70.5**	**100.0**
	Human (%) ↑	0.36	0.00	0.67	5.09	5.60	4.67	22.55	32.00	14.67	**72.00**	**59.20**	**82.67**

Table 1: Quantitative evaluation and user study on Scribble and SketchyCOCO datasets. We show the full testset results, in-domain results, and open-domain results, respectively. Best results are shown in **bold**.

Figure 8: Photo-based sketch synthesis results. Given a photo input, as shown in the first column, our photo-to-sketch generator can translate it into sketches in different styles. Our model is able to generate freehand sketches like human drawers on both seen classes and unseen classes.

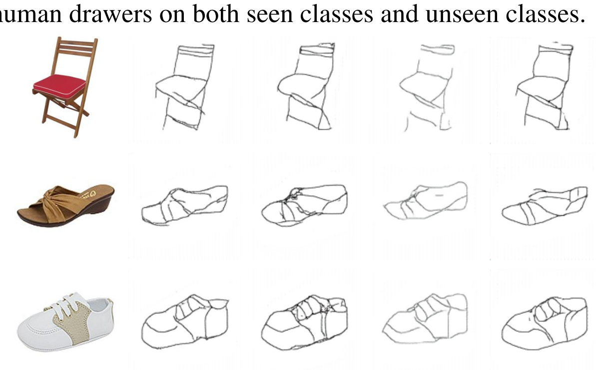

Figure 9: Photo-based sketch synthesis results at different epochs. Given an input shown in the first column, synthesized sketches from our model change at different epochs.

4.4. Ablation Study

Effectiveness of AODA To illustrate the effect of the proposed open-domain training strategy, we simplify the dataset to two classes, including the in-domain class

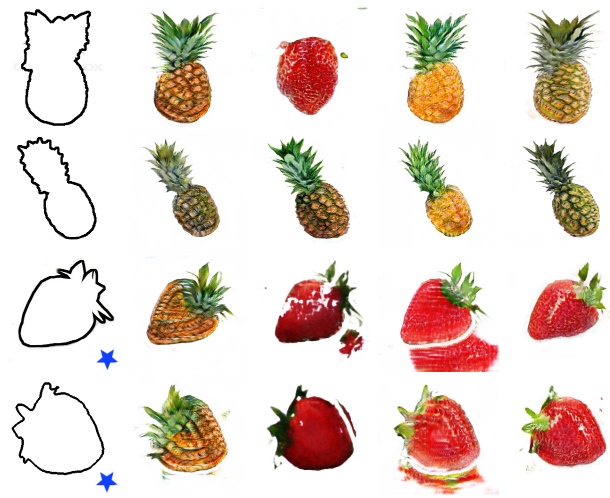

Figure 10: Ablation study of the proposed solution. (a): baseline without classifier or strategy; (b): our framework without strategy; (c) trained with pre-extracted open-domain and real in-domain sketches; (d): random-mixed sampling strategy. Open-domain class is marked with ★.

pineapple and the open-domain class *strawberry*. We compare four models: (a) baseline CycleGAN without classifier or sampling strategy; (b) AODA framework without sampling strategy; (c) AODA trained with synthesized open-domain sketches and real in-domain sketches; (d) AODA trained with the random-mixed sampling strategy as described in Algorithm 1. The results are shown in Figure 10,

From Figure 10, we can see that the base model in column (a) translates all inputs to the in-domain category; (b) generates texture-less images with correct colors for the open-domain class due to the pixel-wise consistency, as discussed in Equation 2. For in-domain sketches, it generates photo-realistic outputs with the shape and texture of any category, which indicates that the model associates the class label with real/fake sketches, and thus fails to generalize to open-domain data. For column (c), the model trained with fake open-domain sketches can barely generate realistic textures for *strawberries*. Besides, it fails to distinguish the object region from the background due to the weak generalization capability, as the extracted sketches actually impair the discriminative power of D_s. Column (d) shows that our open-domain sampling and training strategy can alleviate the above issues, and improve multi-class generation.

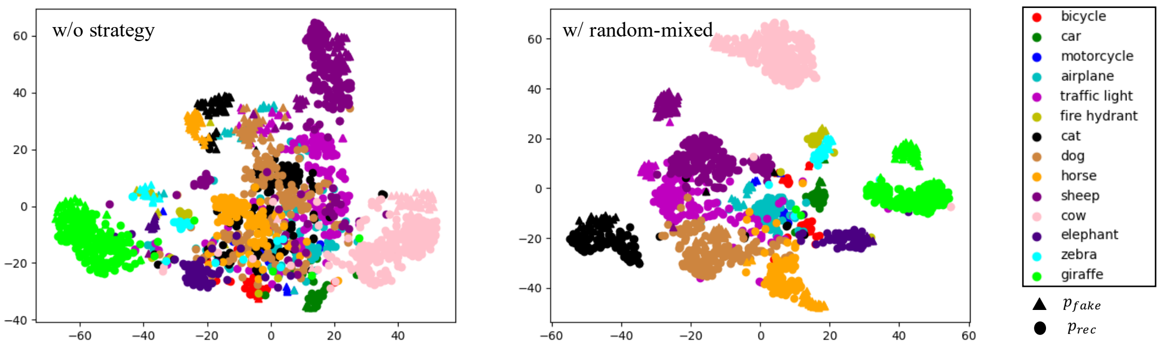

Figure 11: t-SNE visualization of photo embeddings from without any strategy, and with the random-mixed sampling strategy models. Different colors refer to different categories. Our strategies can make the generator learn more separable embeddings for different categories, regardless of in-domain or open-domain data.

To better understand the effect of the random-mixed strategy, we visualize the embedding of generated photos using the t-SNE [58] on SketchyCOCO [18]. We compare the outputs of the AODA framework trained with/without the strategy in Figure 11. We plot both photos p_{fake} synthesized from real sketches (●), and photos p_{rec} reconstructed from fake sketches (▲).As shown in the left plot, for the model trained without any strategy, even with class label conditioning, embeddings of different categories severely overlap. For most in-domain classes, the distance between p_{fake} and p_{rec} is much larger than the inter-class distance. At the same time, the distribution of open-domain classes (● and ●) is well-separated from the in-domain classes, which implies that this model cannot overcome the gap between the in-domain and open-domain data thus fails to synthesize realistic and distinct photos for multiple classes. Instead, it associates the open-domain generation's regressed objective function (Equation 2) with the class label conditioning. As a result, the bias caused by missing sketches in the training set is amplified. While in the right plot, those issues are greatly alleviated with our proposed training strategy. The inter-class distances are greatly maximized with the aid of the random-mixed sampling strategy, which corresponds to more distinctive visual features (textures, colors, shapes, *etc.*) for each category. The intra-class distances are minimized, as shown in the right figure. This is likely due to the blind mixed sampling implicitly encouraging the sketch-to-image generator to ignore the domain gap between real and fake sketch inputs for all classes.

Influence of Missing Classes We study the influence of missing sketches by changing the number of open-domain classes $n_{\mathcal{M}}$ on the Scribble dataset. $n_{\mathcal{M}}$ increases from 0 to 6 by the following order: $strawberry$, $chicken$, $cupcake$, $moon$, $soccer$, and $basketball$. As shown in Table 2, when the number of missing classes becomes larger, the FID score increases, which means that overall output quality degrades due to the fewer real sketches in the training set. But the classification accuracy does not show such a decreasing trend thanks to the classifier's supervision. Figure 12 provides visual results showing that the quality degradation exists in both in- and open-domain classes. Even so, the most representative color composition and textures of each category are maintained, making the synthesized photos recognizable for human viewers and the trained classifier.

$n_{\mathcal{M}} =$	0	1	2	3	4	5	6
FID ↓	167.8	182.6	202.0	207.2	204.2	183.2	209.5
Acc (%) ↑	88.0	80.0	88.0	90.0	76.0	86.0	100.0

Table 2: Influence of missing class number on Scribble [19].

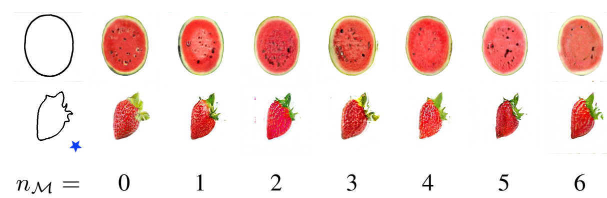

Figure 12: Examples for the influence of missing sketches on Scribble [19]. The output quality of both in-domain and open-domain (★) classes degrades with the increase of $n_{\mathcal{M}}$.

5. Conclusion and Future Work

In this paper, we propose an adversarial open domain adaptation framework to synthesize realistic photos from freehand sketches with class labels even if the training sketches are absent for the class. The two key ideas are that our framework (1) jointly learns sketch-to-photo and photo-to-sketch translation to make unsupervised open-domain adaptation possible, and (2) applies the proposed open-domain training strategy to minimize the domain gap's influence on the generator and better generalize the learned correspondence of in-domain sketch-photo samples to open-domain categories. Extensive experiments and user studies on diverse datasets demonstrate that our model can faithfully synthesize realistic photos for different categories of open-domain freehand sketches. We believe that AODA provides a novel idea to utilize scarce data in real-world scenarios. In future works, we will expand our method to handle more categories of natural images and explore a more efficient design to generate higher-resolution photos.

References

[1] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4561–4569, 2019. 5

[2] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Vectorization and rasterization: Self-supervised learning for sketch and handwriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5672–5681, June 2021. 1

[3] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9779–9788, 2020. 1

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 1, 5

[5] Jiaxin Chen and Yi Fang. Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 605–620, 2018. 1

[6] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. Deepfacedrawing: deep generation of face images from sketches. *ACM Transactions on Graphics (TOG)*, 39(4):72–1, 2020. 1, 2

[7] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. *ACM Transactions on Graphics (TOG)*, 28(5):1–10, 2009. 2

[8] Tao Chen, Ping Tan, Li-Qian Ma, Ming-Ming Cheng, Ariel Shamir, and Shi-Min Hu. Poseshop: Human image database construction and personalized content synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 19(5):824–837, 2012. 2

[9] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9416–9425, 2018. 1, 2

[10] John Collomosse, Tu Bui, and Hailin Jin. Livesketch: Query perturbations for guided sketch-based visual search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2887, 2019. 1

[11] Guoxian Dai, Jin Xie, Fan Zhu, and Yi Fang. Deep correlated metric learning for sketch-based 3d shape retrieval. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 1

[12] Tali Dekel, Chuang Gan, Dilip Krishnan, Ce Liu, and William T Freeman. Sparse, smart contours to represent and edit images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3511–3520, 2018. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 2

[14] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2179–2188, 2019. 1

[15] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5098, 2019. 1

[16] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012. 1, 5

[17] Mathias Eitz, Ronald Richter, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Photosketcher: interactive sketch-based image synthesis. *IEEE Computer Graphics and Applications*, 31(6):56–66, 2011. 2

[18] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: Image generation from freehand scene sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5174–5183, 2020. 1, 2, 3, 5, 6, 7, 8

[19] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1171–1180, 2019. 1, 2, 3, 5, 6, 8

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 1, 2, 5

[21] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017. 1, 2, 5

[22] Xiaoguang Han, Chang Gao, and Yizhou Yu. Deepsketch2face: a deep learning based sketching system for 3d face and caricature modeling. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 1

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 5

[24] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 1

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 1

[26] Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1745–1753, 2019. 1

[27] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, 2018. 2

[28] Moritz Kampelmuhler and Axel Pinz. Synthesizing human-like sketches from natural images using a conditional convolutional decoder. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3203–3211, 2020. 6

[29] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019. 1

[30] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 5

[31] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 1

[32] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017. 1

[33] Mengtian Li, Zhe Lin, Radomir Mech, Ersin Yumer, and Deva Ramanan. Photo-sketching: Inferring contour drawings from images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1403–1412. IEEE, 2019. 1

[34] Yuhang Li, Xuejin Chen, Feng Wu, and Zheng-Jun Zha. Linestofacephoto: Face photo generation from lines with conditional self-attention generative adversarial networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2323–2331, 2019. 1, 2

[35] Yuhang Li, Xuejin Chen, Binxin Yang, Zihan Chen, Zhihua Cheng, and Zheng-Jun Zha. Deepfacepencil: Creating face images from freehand sketches. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 991–999, 2020. 1

[36] Yi Li, Timothy M. Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *In British Machine Vision Conference (BMVC)*, 2014. 1

[37] Fang Liu, Changqing Zou, Xiaoming Deng, Ran Zuo, Yu-Kun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. Scenesketcher: Fine-grained image retrieval with scene sketches. 2020. 1

[38] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2862–2871, 2017. 1

[39] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 1

[40] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10551–10560, 2019. 2

[41] Runtao Liu, Qian Yu, and Stella Yu. Unsupervised sketch-to-photo synthesis. *arXiv preprint arXiv:1909.08313*, 2019. 1, 2, 3

[42] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual gan. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 205–220, 2018. 1, 2

[43] Zhaoliang Lun, Matheus Gadelha, Evangelos Kalogerakis, Subhransu Maji, and Rui Wang. 3d shape reconstruction from sketches via multi-view convolutional networks. In *2017 International Conference on 3D Vision (3DV)*, pages 67–77. IEEE, 2017. 1

[44] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[45] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016. 2

[46] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, pages 2642–2651, 2017. 2

[47] Kyle Olszewski, Duygu Ceylan, Jun Xing, Jose Echevarria, Zhili Chen, Weikai Chen, and Hao Li. Intuitive, interactive beard and hair synthesis with generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7446–7456, 2020. 1

[48] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017. 1

[49] Kaiyue Pang, Da Li, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep factorised inverse-sketching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 36–52, 2018. 6

[50] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Generalising fine-grained sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 677–686, 2019. 1

[51] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 2

[52] Tiziano Portenier, Qiyang Hu, Attila Szabo, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: Deep sketch-based face image editing. *arXiv preprint arXiv:1804.08972*, 2018. 1

[53] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn

bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 1, 5

[54] Yuefan Shen, Changgeng Zhang, Hongbo Fu, Kun Zhou, and Youyi Zheng. Deepsketchhair: Deep sketch-based 3d hair modeling. *arXiv preprint arXiv:1908.07198*, 2019. 1

[55] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2107–2116, 2017. 5

[56] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5551–5560, 2017. 1, 3

[57] Song Tao and Jia Wang. Alleviation of gradient exploding in gans: Fake can be real. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2020. 6

[58] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 8

[59] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1883, 2015. 1

[60] Holger Winnemöller, Jan Eric Kyprianidis, and Sven C Olsen. Xdog: an extended difference-of-gaussians compendium including advanced image stylization. *Computers & Graphics*, 36(6):740–753, 2012. 3

[61] Jin Xie, Guoxian Dai, Fan Zhu, and Yi Fang. Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5068–5076, 2017. 1

[62] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. *arXiv preprint arXiv:2001.02890*, 2020. 1

[63] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 316–333, 2018. 1

[64] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2019. 6

[65] Qian Yu, Feng Liu, Yi-Zhe SonG, Tao Xiang, Timothy Hospedales, and Chen Change Loy. Sketch me that shoe. In *Computer Vision and Pattern Recognition*, 2016. 1, 3

[66] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 297–314, 2018. 1

[67] Ruofan Zhou and Sabine Susstrunk. Kernel modeling super-resolution on real low-resolution images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2433–2443, 2019. 5

[68] Fan Zhu, Jin Xie, and Yi Fang. Learning cross-domain neural networks for sketch-based 3d shape retrieval. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. 1

[69] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 1, 2, 5, 6, 7

[70] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 2

[71] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. Sketchyscene: Richly-annotated scene sketches. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 421–436, 2018. 1