Self-Supervised Sketch-to-Image Synthesis

Bingchen Liu[12], **Yizhe Zhu**[2], **Kunpeng Song**[12], **Ahmed Elgammal**[12]
[1] Playform - Artrendex Inc., USA
[2] Department of Computer Science, Rutgers University
{bingchen.liu, yizhe.zhu, kungpeng.song}@rutgers.edu, elgammal@artrendex.com

Abstract

Imagining a colored realistic image from an arbitrary drawn sketch is one of human capabilities that we eager machines to mimic. Unlike previous methods that either require the sketch-image pairs or utilize low-quantity detected edges as sketches, we study the exemplar-based sketch-to-image (s2i) synthesis task in a self-supervised learning manner, eliminating the necessity of the paired sketch data. To this end, we first propose an unsupervised method to efficiently synthesize line-sketches for general RGB-only datasets. With the synthetic paired-data, we then present a self-supervised Auto-Encoder (AE) to decouple the content/style features from sketches and RGB-images, and synthesize images both content-faithful to the sketches and style-consistent to the RGB-images. While prior works employ either the cycle-consistence loss or dedicated attentional modules to enforce the content/style fidelity, we show AE's superior performance with pure self-supervisions. To further improve the synthesis quality in high resolution, we also leverage an adversarial network to refine the details of synthetic images. Extensive experiments on 1024^2 resolution demonstrate a new state-of-art-art performance of the proposed model on CelebA-HQ and Wiki-Art datasets. Moreover, with the proposed sketch generator, the model shows a promising performance on style mixing and style transfer, which require synthesized images being both style-consistent and semantically meaningful.

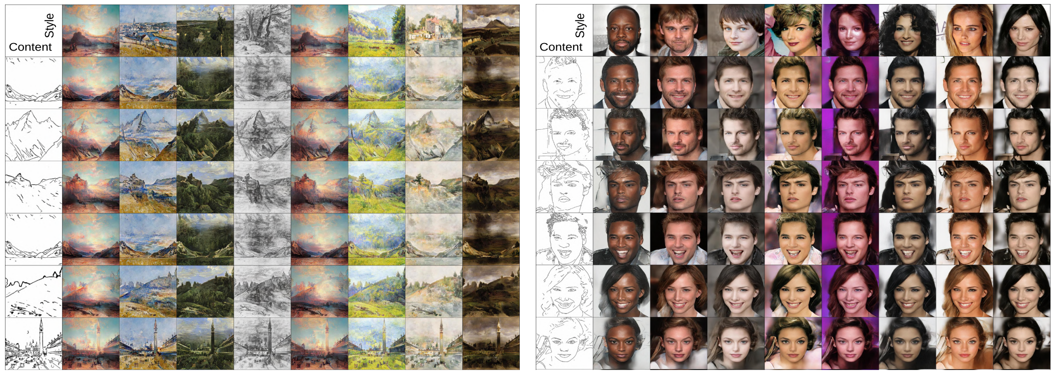
Introduction

Exemplar-based sketch-to-image (s2i) synthesis has received active studies recently (Liu, Yu, and Yu 2019; Zhang et al. 2020; Lee et al. 2020b; Liu, Song, and Elgammal 2020) for its great potential in assisting human creative works (Elgammal et al. 2017; Kim et al. 2018; Elgammal et al. 2018). Given a referential image that defines the style, an s2i model synthesizes an image from an input sketch with consistent coloring and textures to the reference style image. A high-quality s2i model can help reduce repetitive works in animation, filming, and video game story-boarding. It can also help in sketch-based image recognition and retrieval. Moreover, since the model generates images that are style-consistent to the referential images, it has great potential in style-transfer and style harmonization, therefore impacting the human artistic creation processes.

Sketch-to-image synthesis is one important task under the image-to-image (i2i) translation (Isola et al. 2017; Liu, Breuel, and Kautz 2017; Zhu et al. 2017; Kim et al. 2019) category, which benefits a lot from recent year's advances in generative models (Kingma and Welling 2013; Goodfellow et al. 2014). Unlike general i2i tasks, exemplar-based s2i is challenging in several aspects: 1) The sketch domain contains limited information to synthesize images with rich content; especially, real-world sketches have lines that are randomly deformed and differ a lot from the edges in the desired RGB-images. 2) The referential style image usually has a big content difference to the sketch, to avoid content-interference from the style image, the model has to disentangle the content and style information from both inputs effectively. 3) Datasets with paired sketches and RGB-images are rare, even for unpaired sketches that are in the same content domain as the RGB dataset are hard to collect.

Existing works mostly derive their customized attention modules (Vaswani et al. 2017; Zhang et al. 2019), which learn to map the style cues from the referential image to the spatial locations in the sketch, to tackle the first two challenges, and leverage a cycle-consistent (Zhu et al. 2017) or back-tracing (Liu, Breuel, and Kautz 2017) framework to enforce the style and content faithfulness to the respective inputs. However, the derived attention modules and the required supporting models for consistency-checking significantly increase the training cost and limit them to work on low resolution (256^2) images. Moreover, due to the lack of training data, previous methods either work around edge-maps rather than free-hand sketches or on datasets with limited samples, restricting their practicality on image domains with more complicated style and content variance.

Aiming to break the bottleneck on datasets with reliable matched sketches and RGB-images, we propose a dedicated image domain-transfer (Gatys et al. 2016; Huang et al. 2017) model. The model synthesizes multiple paired free-hand sketches for each image in large RGB datasets. Benefit from the paired data, we then show that a simple Auto-encoder (AE) (Kramer 1991; Vincent et al. 2010) equipped with self-supervision (Feng, Xu, and Tao 2019; Kolesnikov, Zhai, and Beyer 2019; He et al. 2020) exhibits exceptional performance in disentangling the content and style information and synthesizing faithful images. As a result, we abandon commonly-used strategies such as cycle-consistent loss



(a) Synthesis on artistic landscapes (b) Synthesis on human portraits

Figure 1: Exemplar-based sketch-to-image synthesis from our model on varied image domains in 1024^2 resolution.

and attention mechanisms. It makes our model neat with less computation cost while having a superior performance at 1024^2 resolution.

In summary, our contributions in this work are:

- We propose a line-sketch generator for generic RGB-datasets, which produces multiple sketches for one image.
- We introduce an efficient self-supervised auto-encoder for the exemplar-based s2i task, with a momentum-based mutual information minimization loss to better decouple the content and style information.
- We present two technique designs in improving DMI (Liu, Song, and Elgammal 2020) and AdaIN (Huang et al. 2017), for a better synthesis performance.
- We show that our method is capable of handling both the high-resolution s2i task and the style-transfer task with a promising semantics-infer ability.

Related Work

Basics Auto-encoder (Kramer 1991; Vincent et al. 2010) (AE) is a classic model that has been widely applied in image-related tasks. Once trained, the decoder in AE becomes a generative model which can synthesize images from a lower-dimensional feature space. Apart from AE, Generative Adversarial Network (GAN) (Goodfellow et al. 2014) significantly boosts the performance in image synthesis tasks. GAN involves a competition between a generator G and a discriminator D, where G and D iteratively improves each other via adversarial training.

Sketch to image synthesis Recent s2i methods can be divided into two categories by the training scheme they based on 1) Pix2pix-based methods (Isola et al. 2017) which is a conditional-GAN (Mirza and Osindero 2014) while G is in the form of an encoder-decoder, and paired data is required to train G as an AE; 2) CycleGAN-based methods (Zhu et al. 2017) that accept unpaired data but require two GANs to learn the transformations back and forth.

Representing Pix2pix-based models includes AutoPainter (Liu et al. 2017), ScribblerGAN (Sangkloy et al. 2017), and SketchyGAN (Chen and Hays 2018). However, none of them have a delicate control to synthesis via exemplar-images. Sketch2art (Liu, Song, and Elgammal 2020) addresses style-consistency to a referential image, but requires an extra encoder for style feature extraction. Zhang et al. and Lee et al. propose reference-based module (RBNet) and cross-domain correspondence module (CoCosNet) respectively, both leverage an attention map to relocate the style cues to the sketch, to enable the exemplar-based synthesis.

Early successors of CycleGAN includes UNIT (Liu, Breuel, and Kautz 2017), which employs an extra pair of encoders to model an assumed domain-invariant feature space. MUNIT (Huang et al. 2018; Lee et al. 2018) further achieves multi-modal image translation. U-GAT-IT (Kim et al. 2019) is a recent exemplar-based model which includes an attention module to align the visual features from the content and style inputs. Furthermore, US2P (Liu, Yu, and Yu 2019) is the latest work that dedicates to s2i, which first translates between sketch and grey-scale images via a CycleGAN, then leverages a separate model for exemplar-based coloration.

Different from both categories, only an simple auto-encoder is applied in our model. We show that an AE, with self-supervision methods including data-augmenting and self-contrastive learning, is sufficient to get remarkable content inference and style translation.

Sketch Synthesis for Any Image Dataset

Few of the publicly available RGB-image datasets have paired sketches, and generating realistic line-sketches for them is challenging. Edge-detection methods (Canny 1986; Xie and Tu 2015) can be leveraged to mimic the "paired sketches"; however, such methods lack authenticity. Moreover, the lack of generalization ability on edge detection methods can lead to missing or distracting lines. There are dedicated deep learning models on synthesizing sketches (Chen et al. 2018; Li et al. 2019; Yu et al. 2020), but most of them focus on pencil sketches with domain-specific tweaks (e.g., only works for faces). Instead, we are interested in sketches of simple lines (Simo-Serra et al. 2018) that one can quickly draw, and should be realistic with random shape deformations (lines that are neither straight nor continuous).

We consider the sketch synthesis as an image domain

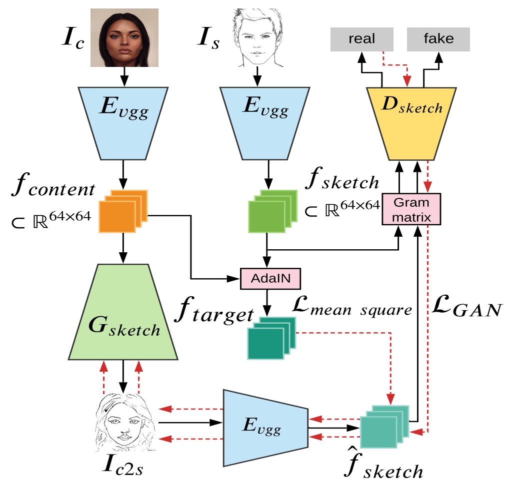


Figure 2: Illustration of our TOM. Dashed arrows in red indicate the gradient flow to train the sketch generator.

transfer problem, where the RGB-image domain R is mapped to the line-sketch domain S. Accordingly, we propose a GAN-based domain transfer model called TOM, short for "Train Once and get Multiple transfers". To produce multiple paired sketches for each image in R, we design an online feature-matching scheme, and to make TOM neat and efficient, we adopt a single-direction model which we empirically found performing well enough for our sketch generation purpose. We will show that the model is 1) fast and effective to train on R with varied domains, such as faces, art paintings, and fashion apparels, 2) so data-efficient that only a few line-sketches (not even need to be in an associated domain to R) are sufficient to serve as S.

TOM consists of three modules: a pre-trained VGG (Simonyan and Zisserman 2014) E that is fixed, a sketch Generator G_{sketch}, and a Discriminator D_{sketch}. We have:

$$f_{content} = E(I_c), \qquad I_c \sim \mathcal{R}; \tag{1}$$

$$f_{sketch} = E(I_s), \qquad I_s \sim \mathcal{S}; \tag{2}$$

$$\hat{f}_{sketch} = E(I_{c2s}), \qquad I_{c2s} = G(f_{content}); \tag{3}$$

$$f_{target} = \sigma(f_{sketch}) \cdot \mathrm{IN}(f_{content}) + \mu(f_{sketch}), \tag{4}$$

where IN is instance normalization (Ulyanov et al. 2016), and f_{target} is produced via adaptive IN (AdaIN (Huang et al. 2017)) which possesses the content information of I_c while having the feature statistics of I_s. As shown in Figure 2, D_{sketch} is trained to distinguish the feature statistics from real sketches I_s and generated sketches I_{c2s}. G_{sketch} is trained to synthesis sketches $I_{c2s} = G(E(I_c))$ for an RGB image I_c. The objectives of TOM are:

$$\begin{aligned}\mathcal{L}_{D_{sketch}} = &-\mathbb{E}[log(D_{sketch}(Gram(f_{sketch})))]\\ &-\mathbb{E}[log(1 - D_{sketch}(Gram(\hat{f}_{sketch})))],\end{aligned} \tag{5}$$

$$\begin{aligned}\mathcal{L}_{G_{sketch}} = &-\mathbb{E}[log(D_{sketch}(Gram(\hat{f}_{sketch})))]\\ &+\mathbb{E}[\|f_{target} - \hat{f}_{sketch}\|^2],\end{aligned} \tag{6}$$

where $Gram$ is gram matrix (Gatys et al. 2016) which computes the spatial-wise covariance for a feature-map. The objectives for G_{sketch} are two-fold. Firstly, the discriminative loss in Eq.6 makes sure that I_{c2s} is realistic with random deformations and stroke styles, and enables G_{sketch} to generalize well on all images from $\mathcal{R}$. Secondly, the mean-square loss in Eq.6 ensures the content consistency of I_{c2s} to I_c.

Importantly, we randomly match a batch of RGB-images I_c and real sketches I_s during training. Therefore, f_{target} is created in an online fashion and is always changing for the same I_c. In other words, for the same I_c, Eq.6 trains G_{sketch} to generate a sketch towards a new "sketch style" in every new training iteration. Combined with such an online feature-matching training strategy, we leverage the randomness from the SGD optimizer (Robbins and Monro 1951) to sample the weights of G_{sketch} as checkpoints after it is observed to output good quality I_{c2s}. As a result, we can generate multiple sketches for one image according to the multiple checkpoints, which can substantially improve our primary sketch-to-image model's robustness.

Style-guided Sketch to Image Synthesis

We consider two main challenges in the style-guided sketch to image synthesis: 1) the style and content disentanglement, 2) the quality of the final synthesized image. We show that with our designed self-supervised signals, an Auto-Encoder (AE) can hallucinate rich content from a sparse line-sketch while assigning semantically appropriate styles from a referential image. After the AE training, we employ a GAN to revise the outputs from AE for a higher synthesis quality.

Self-supervised Auto-encoder

Our AE consists of two separate encoders: 1) a style encoder E_{style} that takes in an RGB-image I^t_{rgb} to generate a style vector $f_{style} \subseteq \mathbb{R}^{512}$, 2) a content encoder $E_{content}$ which takes in a sketch I^t_{skt} and extracts a content feature-map $f_{content} \subseteq \mathbb{R}^{512\times 8\times 8}$. The extracted features from both sides are then taken by a decoder G_1 to produce a reconstructed RGB-image I^{ae}_g. Note that the whole training process for our AE is on paired data after we synthesize multiple sketches for each image in the RGB-dataset using TOM.

Translation-Invariant Style Encoder To let E_{style} extracts translation-invariant style information, thus approach a content-invariant property, we augment the input images by four image translation methods: cropping, horizontal-flipping, rotating, and scaling. During training, the four translations are randomly configured and combined, then applied on the original image I_{rgb} to get I^t_{rgb}. Samples of I^t_{rgb} drawn from an I_{rgb} are shown on the top-left portion of Figure 3, which E_{style} takes one as input each time. We consider I^t_{rgb} now possesses a different content with its style not changed, so we have an reconstruction loss between the decoded image I^{ae}_g and the original I_{rgb}.

To strengthen the content-invariant property on f_{style}, a triplet loss is also leveraged to encourage the cosine similarity on f_{style} to be high between the translations of the same image, and low between different images:

$$\mathcal{L}^s_{tri} = \max(cos(f^t_s, f^{org}_s) - cos(f^t_s, f^{neg}_s) + \alpha, 0), \tag{7}$$

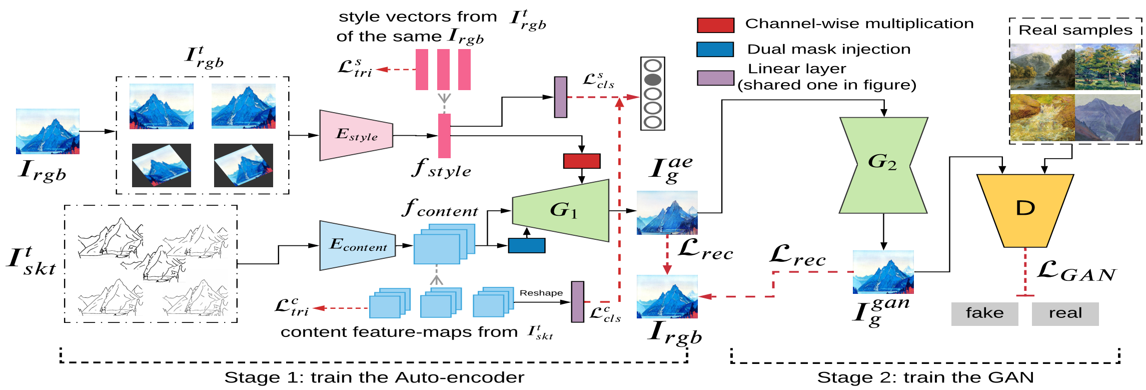


Figure 3: Overview of the proposed model.

where α is the margin, f^t_s and f^{org}_s are feature vectors from the same image, and f^{neg}_s is from a different random image. The translations on I_{rgb} enforces E_{style} to extract style features from an content-invariant perspective. It guides our AE learn to map the styles by the semantic meanings of each region, rather than the absolute pixel locations in the image.

Momentum mutual-information minimization A vanilla AE usually produces overly smooth images, making it hard for the style encoder to extract style features such as unique colors and fine-grained textures. Moreover, the decoder may rely on the content encoder to recover the styles by memorizing those unique content-to-style relations.

Inspired by momentum contrastive loss (He et al. 2020), we propose a momentum mutual-information minimization objective to make sure E_{style} gets the most style information, and decouples the style-content relation on $E_{content}$. Specifically, a group of augmented images translated from the same image are treated as one unique class, and E_{style} associated with an auxiliary classifier is trained to classify them. To distinguish different images, E_{style} is enforced to capture as much unique style cues from each image as possible. Formally, E_{style} is trained using cross-entropy loss:

$$\mathcal{L}^s_{cls} = -\log(\frac{\exp(E^{cls}_{style}(f_{style})[label])}{\sum_j \exp(E^{cls}_{style}(f_{style})[j])}), \tag{8}$$

where $E^{cls}_{style}(\cdot)$, implemented as one linear layer, yields the class prediction vector and $label$ is the assigned ground truth class for I_{sty}.

While E_{style} is predicting the style classes, we can further decouple the correspondence between f_{style} and $f_{content}$ by implicitly minimizing their mutual-information:

$$\mathrm{MI}(f_{style}, f_{content}) = \mathrm{H}(f_{style}) - \mathrm{H}(f_{style}|f_{content})$$

where H refers to entropy. Since $\mathrm{H}(f_{style})$ can be considered as a constant, we only consider $\mathrm{H}(f_{style}|f_{content})$ and encourage that style information can hardly be predicted based on $f_{content}$. In practice, we make the probability of each style class given $f_{content}$ equal to the same value. The objective is formulized as:

$$\mathcal{L}^c_{cls} = \|\mathrm{softmax}(E^{cls}_{style}(f_{content})) - v\|^2, \tag{9}$$

where v is a vector with each entry having the same value $\frac{1}{k}$ (k is the number of classes). Note that we use average-pooling to reshape $f_{content}$ to match f_{style}. Eq.9 forces $f_{content}$ to be classified into none of the style classes, thus helps removing the correlations between $f_{content}$ and f_{style}.

"Generative" Content Encoder Edge-map to image synthesis possesses a substantial pixel alignment property between the edges from the input and the desired generated image. Instead, realistic sketches exhibit more uncertainty and deformation, thus requires the model to hallucinate the appropriate contents from misaligned sketch-lines. We strengthen the content feature extraction power of $E_{content}$ with a self-supervision manner using data augmenting.

Firstly, we already gain multiple synthesised sketches for each image from TOM (with varied line straightness, boldness and composition). Secondly, we further transform each sketch by masking out random small regions, to make the lines dis-continue. An example set of I^t_{skt} can be find in Figure 3. Finally, we employ a triplet loss to make sure all the sketches paired to the same I_{rgb} have similar feature-maps:

$$\mathcal{L}^c_{tri} = \max(d(f^t_c, f^{pos}_c) - d(f^t_c, f^{neg}_c) + \beta, 0), \tag{10}$$

where $d(,)$ is the mean-squared distance, β is the margin, f^t_c and f^{pos}_c are features from the sketches that correspond to the same I_{rgb}, and f^{neg}_c is from one randomly mismatched sketch. Such self-supervision process makes $E_{content}$ more robust to the changes on the sketches, and enables it to infer a more accurate and completed contents from sketches with distorted and discontinued lines.

Feature-space Dual Mask Injection DMI is proposed in Sketch2art (Liu, Song, and Elgammal 2020) for a better content faithfulness of the generation to the input sketches. It uses the sketch-lines to separate two areas (object contours and plain fields) from a feature-map and shifts the feature values via two learnable affine transformations. However, DMI assumes the sketches aligns well to the ground truth RGB-images, which is not practical and ideal. Instead of the raw sketches, we propose to use $f_{content}$ to perform a per-channel DMI, as $f_{content}$ contains more robust content information that is hallucinated by $E_{content}$.

Simplified Adaptive Instance Normalization AdaIN is an effective style transfer module (Huang et al. 2017):

$$f'_c = \underset{c\times h\times w}{\overset{1}{\mathrm{IN}}(f_c)} \overset{2}{\times} \underset{c\times 1\times 1}{\sigma(f_s)} \overset{3}{+} \underset{c\times 1\times 1}{\mu(f_s)}, \tag{11}$$

where IN is instance normalization, μ and σ are the instance-wise mean and std. In spite of AdaIN's success on style transfer, its instance normalization (operation-1 in Eq.11) usually causes droplet effects to models that are trained on large corpus of images (Karras et al. 2020). To resolve the problem, we only preserve the **channel-wise multiplication** part (operation-2 in Eq.11) in AdaIN, and abandon the IN and addition (operation-1 and 3 in Eq.11). Such simplification turns out working great in our model.

All objectives Figure 3 stage-1 shows the overview of our AE. Via the proposed self-supervision training strategies, our encoders extract the disentangled features $f_{content}$ and f_{style}, and the decoder G_1 takes $f_{content}$ via DMI and applies f_{style} via channel-wise multiplication to synthesis a reconstructed image. The summed objective for our AE is:

$$\begin{aligned}\mathcal{L}_{ae} = \ & \mathbb{E}[\|G_1(E_s(I_{rgb}), E_c(I_{skt})) - I_{rgb}\|^2] \\ & + \mathcal{L}^c_{tri} + \mathcal{L}^s_{tri} + \mathcal{L}^s_{cls} + \mathcal{L}^c_{cls},\end{aligned} \tag{12}$$

where the first part in Eq.12 computes the mean-square reconstruction loss between I_{ae} and I_{rgb}. Please refer to the appendix for more discussions on why we choose AE over variational AE (Kingma and Welling 2013), and the implementation details on the revised DMI and simplified AdaIN.

Revised Synthesis via Adversarial Training

Once our AE is trained, we fix it and train a GAN to revise AE's output for a better synthesis quality. As shown in Figure 3 stage-2, our Generator G_2 has a encoder-decoder structure, which takes I_g^{ae} from G_1 as input and generates our final output I_g^{gan}. The final results of our model on unpaired testing data can be found in Figure 4, where G_1 already gets good style features and composites rich content, while G_2 revises the images to be much more refined.

Same as our AE, only paired sketch and image data are used during the training. We do not randomly mismatch the sketches to images, nor do we apply any extra guidance on D. In sum, the objectives to train our GAN are:

$$\begin{aligned}\mathcal{L}_D = & -\mathbb{E}[min(0, -1 + D(I_{sty}))] \\ & -\mathbb{E}[min(0, -1 - D(G_2(I_g^{ae}))],\end{aligned} \tag{13}$$

$$\begin{aligned}\mathcal{L}_{G_2} = & -\mathbb{E}[D(G_2(I_g^{ae}))] \\ & + \lambda\mathbb{E}[\|G_2(I_g^{ae}) - I_{sty}\|^2],\end{aligned} \tag{14}$$

which we employ the hinge version of the adversarial loss (Lim and Ye 2017; Tran, Ranganath, and Blei 2017), and λ is the weight for the reconstruction term which we set to 10 for all datasets. Please refer to the appendix for more details.

Experiments

Datasets We evaluate our model on two datasets, CelebA-HQ (Liu et al. 2015; Lee et al. 2020a) and WikiArt [1].

[1]https://www.wikiart.org/

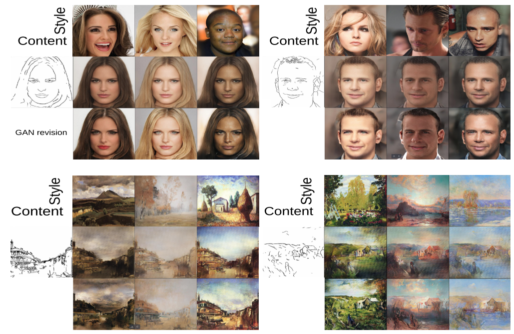


Figure 4: In each panel, the second row shows the images from AE, and the third row shows the GAN revisions.

- CelebA-HQ contains 30000 portrait images of celebrities worldwide, with a certain amount of visual style variance. We train our model on 1024^2 resolution on randomly selected 15000 images and test on the rest images.
- We collect 15000 high-quality art paintings from WikiArt, which covers 27 major art styles from over 1000 artists. We train on 11000 of the images on 1024^2 resolution and test on the rest images.

Synthesis Sketches via TOM

To train TOM, we find it sufficient to collect 20 sketches in the wild as $\mathcal{S}$. Moreover, the collected sketches work well for both the CelebA and WikiArt datasets. The whole training process takes only 20 minutes on one RTX-2080 GPU. We save ten checkpoints of G_{sketch} to generate ten different sketches for an RGB-image. Figure 5-(a) shows the sketches generated from TOM. Among various checkpoints, we get sketches with diverse drawing styles, e.g., line boldness, line straightness, and stroke type. Moreover, while providing the desired sketch variations, it maintains a decent synthesis quality across all checkpoints. In comparison, edge detection methods are less consistent among the datasets.

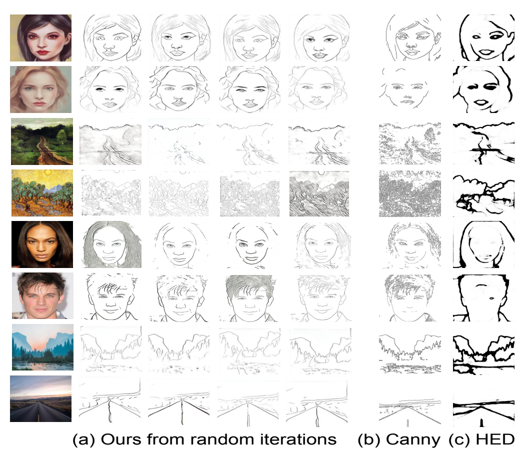


Figure 5: Synthesises from TOM. TOM generalizes well across multiple image domains, from photo-realistic to artistic, and from human portrait to nature landscape.

Quantitative Evaluations

Quantitative metrics We use three metrics: 1) Fréchet Inception Distance (FID) (Heusel et al. 2017) is used to measure the overall semantic realism of the synthesized images. We randomly mismatch the sketches to the RGB-images and generate 40000 samples to compute the FID score to the real testing images. 2) Style relevance (SR) (Zhang et al. 2020) leverages the distance of low-level perceptual features to measure the consistency of color and texture. It checks the model's style consistence to the inputs and reflects the model's content/style disentangle performance. 3) Learned perceptual similarity (LPIPS) (Zhang et al. 2018) provides a perceptual distance between two images; we use it to report the reconstruction quality of our Auto-encoder on paired sketch and style image input.

Comparison to baselines We compare our model to the latest state-of-the-art methods mentioned in Section : RBNet (CVPR-2020), Sketch2art (SIGGRAPH-2020-RT-Live), CocosNet (CVPR-2020), and SPADE (CVPR-2019). Results from earlier methods, including Pix2pixHD, MUNIT, and SketchyGAN, are also presented. Some models are adopted for exemplar-based synthesis to make a fair comparison and are trained on edge-maps as they originally proposed on. Instead, we train our model on synthesized sketches, which are more practical but arguably harder. We report the author's scores provided from the official figures, which, if not available, we try to train the models if the official code is published.

	CelebA-HQ		WikiArt	
	FID ↓	SR ↑	FID ↓	SR ↑
Pix2pixHD	62.7	0.910	172.6	0.842
MUNIT	56.8	0.911	202.8	0.856
SPADE	31.5	0.941	N/A	N/A
CocosNet	14.3	0.967	N/A	N/A
SketchyGAN	N/A	N/A	96.3	0.843
RBNet	47.1	N/A	N/A	N/A
Sketch2art	28.7	0.958	84.2	0.897
Ours AE	25.9	0.959	74.2	0.902
Ours AE+GAN	**13.6**	**0.972**	**32.6**	**0.924**

Table 1: Quantitative comparison to existing methods, bold indicates the best score.

As shown in Table 1, our model outperforms all competitors, and by a large margin on WikiArt. The self-supervised AE does a great job in translating the style features, while the GAN further boosts the overall synthesis quality.

CelebA	FID ↓	LPIPS ↓
Vanilla AE	44.3	18.7
AE + $\mathcal{L}^{c}_{tri}$	34.8	15.8
AE + $\mathcal{L}^{s}_{tri}$	35.7	16.3
AE + $\mathcal{L}^{s}_{cls}$	36.4	16.4
AE + $\mathcal{L}^{s}_{cls}$ + $\mathcal{L}^{c}_{cls}$	34.7	15.2
AE + all	25.9	11.7

Table 2: Benchmarks on the self-supervised objectives.

Objectives Ablation To evaluate the performance of AE, we compute FID from unpaired data to show its generalize ability, and compute LPIPS from paired data to show the reconstruction performance. Table 2 presents the contribution of each individual self-supervision objective. Compared to a vanilla AE with only reconstruction loss, each objective can independently boost the performance. Figure 6 better demonstrates the model behavior during training. We can see that the data-augmenting objectives $\mathcal{L}^{c/s}_{tri}$ make the biggest difference in the synthesis quality of the AE. Moreover, the contrastive objectives $\mathcal{L}^{c/s}_{cls}$ cooperates well with $\mathcal{L}^{c/s}_{tri}$ and further improves the scores.

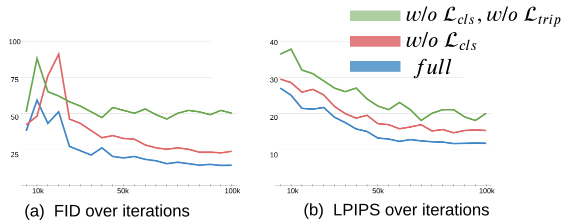


Figure 6: Model performance on CelebA during training.

Qualitative Analysis

A general sketch-to-image synthesis result of our model can be found in Figure 1. We select the style images that have a significant content difference to the sketches, to demonstrate the content/style disentangle ability of our model. Figure 1-(a) shows the result on WikiArt, which in a few examples, we still observe the "content-interference from style image" issue, such as row.2-col.2 and row.7-col.3. Instead, on CelebA, as shown in Figure 1-(b), the model disentangles better even for rare style images such as col.4 and 5. This is expected as CelebA is a much simpler dataset in terms of content variance, whereas WikiArt contains much more diverse shapes and compositions.

Synthesis by mixing multiple style images Via feeding structurally abnormal style images to the model, we demonstrate the model's superior ability on 1) capturing style cues from multiple style images at once; 2) imposing the captured styles to the sketch in a semantically meaningfully manner. Figure 7 shows the synthesis comparison between our model and CocosNet on CelebA. We cut and stitch two or four images into one, and use the resulting image as the referential style. Our model harmonizes different face patches into unified style features, resulting in consistent hair color, skin tone, and textures. In contrast, CocosNet exhibits a patch-to-patch mapping between the input and output, yielding unrealistic color isolation on the synthesized images. Moreover, the color consistency of the style image on CocosNet is severely downgraded on mixed images, while our model summarizes a "mixture style" from all patches.

Synthesis on out-domain images To demonstrate the generalization ability of our model, we use images from a different semantic domain than what the model is trained on as style images or sketches. In figure 8-(a) and (b), we use

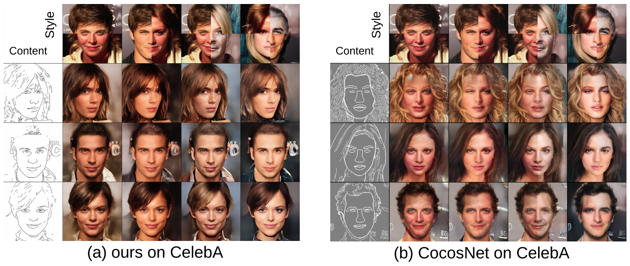


Figure 7: Synthesis by mixing multiple style images.

style images from photo-realistic nature scenes on our model trained on Wikiart. In figure 8-(a), the sketches are also from photo-realistic images, as shown in col.1, which we synthesize via TOM (note that TOM is also only trained on Wikiart). Although the out-domain images are different in texture, the model still gets accurate colors and compositions from the inputs. In figure 8-(b) row 2 and 4, the buildings are adequately colored, showing an excellent semantic inference ability of our model. Interestingly, an artistic texture is automatically applied to all the generated images, reflecting what the model has learned from the WikiArt corpus.

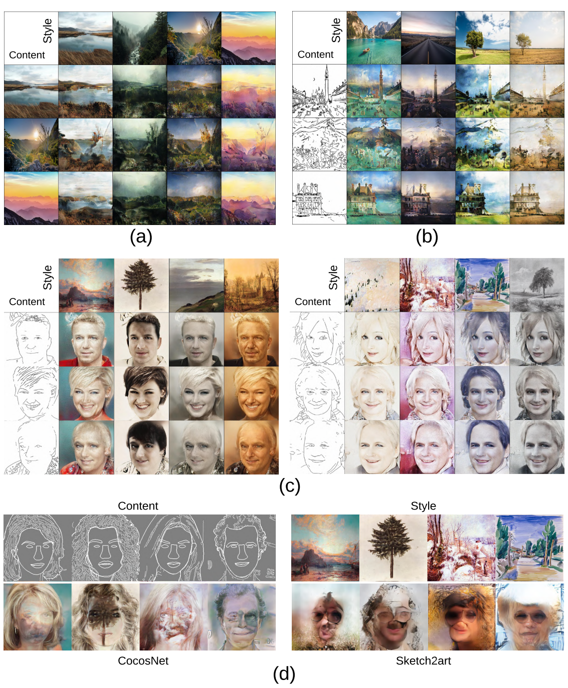


Figure 8: Synthesis from out-domain style images.

In figure 8-(c), we use the art paintings as style images for the model trained on CelebA. All the faces are correctly generated and not interfered with the contents from the style images, showing our model's excellent job in hallucinating the contents from the input sketches. Amazingly, the model knows to apply the colors to the content follow proper semantics. Note how the hair, clothes, and backgrounds are separately and consistently colored. In contrast, figure 8-

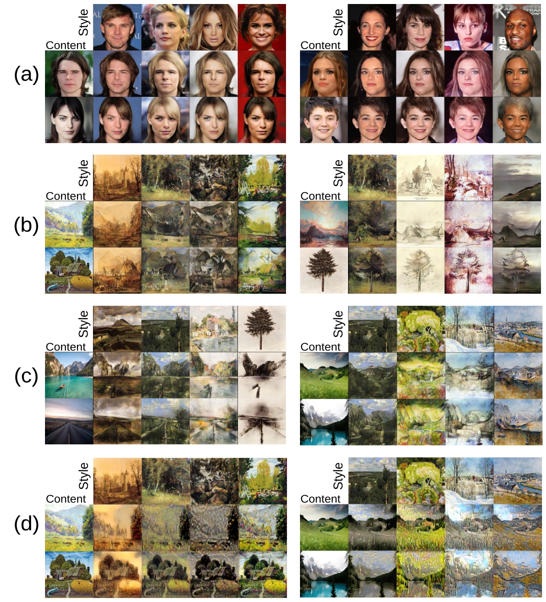


Figure 9: Works as a feed-forward style-transfer model.

(d) shows how the other models suffer from generalizing on out-domain style images. CocosNet exhibits sever content-interference issue from the style images, and Sketch2art can hardly synthesis a meaningful face.

Work as a style-transfer model Combined with TOM, our model possesses competitive style-transfer ability. We first convert a content image into sketch-lines, then colorize it according to the style image. In this progress, our model can apply the style cues to different objects in the content image in a semantically appropriate manner. As shown in figure 9-(a) and (b), our model can easily transfer the styles between in-domain images on face and art. Figure 9-(c) further shows how the model performs on out-domain images. In contrast, Figure 9-(d) shows the result of the traditional Neural Style Transfer (NST) method from Gatys et al., which an undesired texture covers the whole image in most cases. Note that we do not intend to compete with NST methods. Instead, our model provides a new perspective on the style transfer task towards a more semantic-aware direction.

Conclusion

In this paper, we present a self-supervised model for the exemplar-based sketch to image synthesis. Without computationally-expensive modules and objectives, our model shows outstanding performance on 1024^2 resolution. With the mechanisms (self-supervisions) in this model orthogonal to existing image-to-image translation methods, even more performance boosts are foreseeable with proper tweaking and integration. Moreover, the extraordinary generalization performance on out-domain images showing a robust content and style inference ability of our model, which yields a promising performance on style-mixing and style-transferring, and reveals a new road for future studies on these intriguing applications.

Bibliography

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* (6): 679–698.

Chen, C.; Liu, W.; Tan, X.; Wong, Kwan-Yee K, B.; C, D.; and E. 2018. Semi-supervised learning for face sketch synthesis in the wild. In *Asian Conference on Computer Vision*, 216–231. Springer.

Chen, W.; and Hays, J. 2018. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9416–9425.

Elgammal, A.; Liu, B.; Elhoseiny, M.; and Mazzone, M. 2017. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068* .

Elgammal, A.; Mazzone, M.; Liu, B.; Kim, D.; and Elhoseiny, M. 2018. The shape of art history in the eyes of the machine. *arXiv preprint arXiv:1801.07729* .

Feng, Z.; Xu, C.; and Tao, D. 2019. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10364–10374.

Gatys, L. A.; Ecker, A. S.; Bethge, Matthias, B.; and C. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

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

Huang, X.; Belongie, Serge, B.; C, D.; and E. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 172–189.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

Kim, D.; Liu, B.; Elgammal, A.; and Mazzone, M. 2018. Finding principal semantics of style in art. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, 156–163. IEEE.

Kim, J.; Kim, M.; Kang, H.; and Lee, K. 2019. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830* .

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Kolesnikov, A.; Zhai, X.; and Beyer, L. 2019. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1920–1929.

Kramer, M. A. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* 37(2): 233–243.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020a. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, 35–51.

Lee, J.; Kim, E.; Lee, Y.; Kim, D.; Chang, J.; and Choo, J. 2020b. Reference-Based Sketch Image Colorization Using Augmented-Self Reference and Dense Semantic Correspondence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Y.; Fang, C.; Hertzmann, A.; Shechtman, E.; and Yang, M.-H. 2019. Im2pencil: Controllable pencil illustration from photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1525–1534.

Lim, J. H.; and Ye, J. C. 2017. Geometric gan. *arXiv preprint arXiv:1705.02894* .

Liu, B.; Song, K.; and Elgammal, A. 2020. Sketch-to-Art: Synthesizing Stylized Art Images From Sketches. *arXiv preprint arXiv:2002.12888* .

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, 700–708.

Liu, R.; Yu, Q.; and Yu, S. 2019. An unpaired sketch-to-photo translation model. *arXiv preprint arXiv:1909.08313* .

Liu, Y.; Qin, Z.; Luo, Z.; and Wang, H. 2017. Auto-painter: Cartoon image generation from sketch by using conditional generative adversarial networks. *arXiv preprint arXiv:1705.01908* .

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* .

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics* 400–407.

Sangkloy, P.; Lu, J.; Fang, C.; Yu, F.; and Hays, J. 2017. Scribbler: Controlling deep image synthesis with sketch and color. In *CVPR*, 5400–5409.

Simo-Serra, E.; Iizuka, S.; Ishikawa, Hiroshi, B.; C, D.; and E. 2018. Mastering sketching: adversarial augmentation for structured prediction. *ACM Transactions on Graphics (TOG)* 37(1): 1–13.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Tran, D.; Ranganath, R.; and Blei, D. M. 2017. Deep and hierarchical implicit models. *arXiv preprint arXiv:1702.08896* 7(3): 13.

Ulyanov, D.; Vedaldi, A.; Lempitsky, Victor, B.; C, D.; and E. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* .

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A.; and Bottou, L. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11(12).

Xie, S.; and Tu, Z. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, 1395–1403.

Yu, J.; Xu, X.; Gao, F.; Shi, S.; Wang, M.; Tao, D.; Huang, Qingming, B.; C, D.; and E. 2020. Toward Realistic Face Photo-Sketch Synthesis via Composition-Aided GANs. *IEEE Transactions on Cybernetics* .

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2019. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, 7354–7363.

Zhang, P.; Zhang, B.; Chen, D.; Yuan, L.; Wen, Fang, B.; C, D.; and E. 2020. Cross-domain Correspondence Learning for Exemplar-based Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5143–5153.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.