MaskSketch: Unpaired Structure-guided Masked Image Generation

Dina Bashkirova[1*], José Lezama[2] , Kihyuk Sohn[2] , Kate Saenko[1,3] , Irfan Essa[2,4]
[1]Boston University, [2]Google Research, [3]MIT-IBM Watson AI Lab, [4]Georgia Institute of Technology

Abstract

Recent conditional image generation methods produce images of remarkable diversity, fidelity and realism. However, the majority of these methods allow conditioning only on labels or text prompts, which limits their level of control over the generation result. In this paper, we introduce MaskSketch, an image generation method that allows spatial conditioning of the generation result using a guiding sketch as an extra conditioning signal during sampling. MaskSketch utilizes a pre-trained masked generative transformer, requiring no model training or paired supervision, and works with input sketches of different levels of abstraction. We show that intermediate self-attention maps of a masked generative transformer encode important structural information of the input image, such as scene layout and object shape, and we propose a novel sampling method based on this observation to enable structure-guided generation. Our results show that MaskSketch achieves high image realism and fidelity to the guiding structure. Evaluated on standard benchmark datasets, MaskSketch outperforms state-of-the-art methods for sketch-to-image translation, as well as unpaired image-to-image translation approaches. The code can be found on our project website: https://masksketch.github.io/

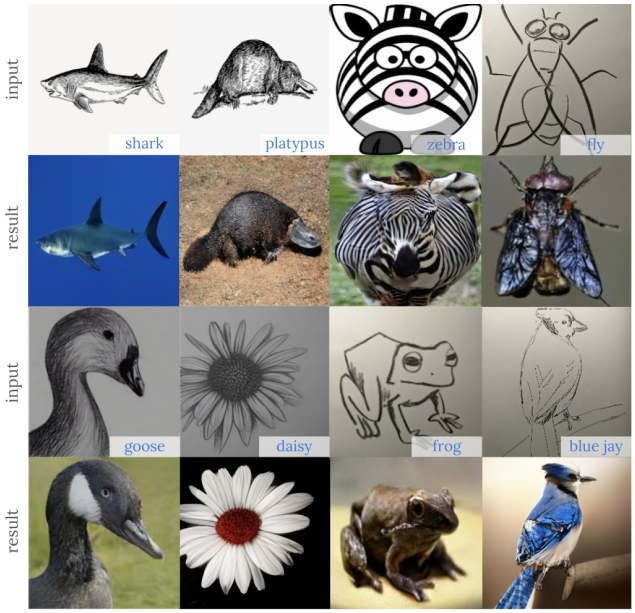

Figure 1. Given an input sketch and class label, MaskSketch samples realistic images that follow the given structure. MaskSketch works on sketches of various degree of abstraction by leveraging a pre-trained masked image generator [4], while not requiring model finetuning or pairwise supervision.

1. Introduction

Recent Image generation methods achieved remarkable success, allowing diverse and photorealistic image synthesis [4, 11, 44, 46]. The majority of state-of-the-art generative models allow conditioning with class labels [2, 4, 11, 13] or text prompts [40, 41, 44, 46]. However, some applications require a more fine-grained control over the spatial composition of the generation result. While methods conditioned with segmentation maps [14] or strokes [34] achieve some spatial control over the generated image, sketching allows a more fine-grained specification of the target spatial layout, which makes it desirable for many creative applications.

In this paper, we propose MaskSketch, a method for conditional image synthesis that uses sketch guidance to define the desired structure, and a pre-trained state-of-the-art *masked generative transformer*, MaskGIT [4], to leverage a strong generative prior. We demonstrate the capability of MaskSketch to generate realistic images of a given structure for sketch-to-photo image translation. Sketch-to-photo [5, 20, 32] is one of the most challenging applications of structure-conditional generation due to the large domain gap between sketches and natural images. MaskSketch achieves a balance between realism and structure fidelity. Our experiments show that MaskSketch outperforms state-of-the-art sketch-to-photo [20] and unpaired image translation methods [6, 25, 37], according to standard metrics for image generation [23] and user preference studies.

In MaskSketch, we formulate a structure similarity constraint based on the observation that the intermediate self-attention maps of a masked generative transformer [4] encode rich structural information (see Fig. 2). We use this

*Work done during an internship at Google.

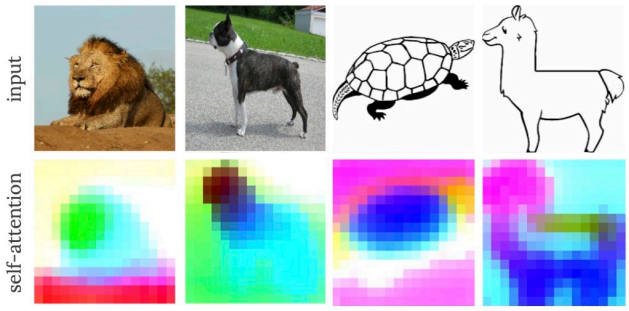

Figure 2. Self-attention maps (PCA) of the intermediate layers of a pre-trained masked generative transformer [4] encode information about the spatial layout of the input. Notably, they are robust to the domain shift between natural images (**left**) and sketches (**right**).

structure similarity constraint to guide the generated image towards the desired spatial layout [22, 48]. Our study shows that the proposed attention-based structure similarity objective is robust to the domain shift occurring in sketch-to-photo translation. The proposed structure-based sampling method leverages a pre-trained image generator, and does not require model finetuning or sketch-photo paired data. Moreover, it is significantly faster than other methods that exploit self-attention maps for guided image generation [48]. Figure 1 shows the translation results produced by our method on sketches of various levels of abstraction.

The limitations of existing sketch-to-photo translation methods [5, 20, 32] come from having to learn both an implicit natural domain prior and the mapping that aligns sketches to natural images, for which the domain gap is severe. MaskSketch, on the other hand, uses the strong generative prior of a pre-trained generative transformer, which allows highly realistic generation. In addition, MaskSketch uses the domain-invariant self-attention maps for structure conditioning, allowing its use on sketches of a wide range of abstraction levels.

Our contributions can be summarized as follows:

- We show that the self-attention maps of a masked generative transformer encode important structural information and are robust to the domain shift between images and sketches.
- We propose a sampling method based on self-attention similarity, balancing the structural guidance of an input sketch and the natural image prior.
- We demonstrate that the proposed sampling approach, MaskSketch, outperforms state-of-the-art methods in unpaired sketch-to-photo translation.
- To the best of our knowledge, MaskSketch is the first method for sketch-to-photo translation in the existing literature that produces photorealistic results requiring only class label supervision.

2. Related Work

While there is a vast volume of literature on image generative models thanks to recent progress ranging from generative adversarial networks [2, 18, 27] generative transformers [4, 13, 54] and diffusion models [11, 35, 40, 46], in this section, we focus on reviewing image-conditioned image generation, also known as image translation.

Supervised image conditional generation Sketch-to-photo image translation is a special case of image-conditional image generation. Early conditional image generation methods were based on generative adversarial networks. For example, pix2pix [26] conditioned the generation result by minimizing the patchwise distance between the ground truth and the generated image; SPADE [38] and OASIS [47] used spatially-adaptive instance normalization to condition generation on a segmentation map; CoCosNet [55], CoCosNet V2 [57] warped the reference image using a correlation matrix between the image and the given segmentation map. Similarly to MaskSketch, Make-a-Scene and NUWA [14, 52] are designed to condition generation on semantic segmentation and text prompts with a VQ-based transformer. While these methods allow spatial conditioning, they are inapplicable for sketch-to-photo due to the lack of ground truth paired data, domain gap between sketches and segmentation maps and lack of efficient methods that extract semantic segmentation from sketches.

Unsupervised image-conditional generation In unsupervised image-conditioned translation, the ground truth input and translation pairs are not available for training. For example, CycleGAN [58] used a cycle reconstruction loss to ensure a semantically consistent translation, UNIT [31], MUNIT [25], and StarGANv2 [8] disentangled domain-specific and shared information between the source and target image domains by mapping them to a shared latent embedding space. PSP [42] used StyleGAN [46] inversion along with style mixing for segmentation- and edge-guided translation. SDEdit [33] uses a diffusion model to translate the input strokes or segmentation maps to natural images.

The closest work to ours in this line may be SpliceVIT [48], which uses self-attention key self-similarity extracted from the discriminative ViT (Dino [3]) to represent the structure of an input image. As pointed in [48], SpliceVIT works only in case when both the input and expected output images come from the same domain, which makes it inapplicable for sketch-to-photo translation. VQ-I2I [6] is another work on unsupervised image translation that leverages the generative power of a VQ-GAN [13]-based generative transformer. Unlike MaskSketch, VQ-I2I uses the embedding reconstruction loss for controllable generation.

Recently, [15, 28, 45] also demonstrated how to leverage pre-trained image generators for image synthesis based on novel conditioning inputs. These methods allow to replicate

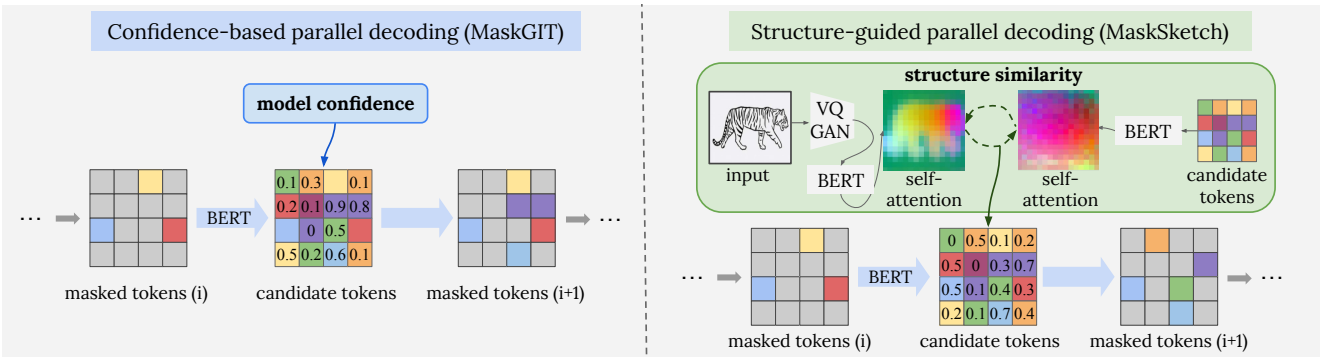

Figure 3. **Left:** confidence-based token rejection (masking) in MaskGIT. **Right:** structure-based token rejection. Confidence-based rejection masks out the least 'likely' tokens, while structure-based rejection masks the tokens with the highest structural distance (Eq. 1) w.r.t. the input sketch.

a given object or subject in the generated image. However, they do not allow to finely specify the spatial layout of the generated image as MaskSketch and typically require some degree of fine-tuning. Prompt-to-prompt tuning [22] uses the attention features to perform spatially aligned prompt-conditional generation. General image-conditional methods show remarkable results when the source and target domains are visually similar, e.g., translating horses to zebras, performing artistic style transfer, etc, however, they tend to struggle on the more challenging sketch-to-photo translation task. MaskSketch shows a promising alternative for transferring the spatial composition from sketches since it requires no paired data or model training, thanks to leveraging a powerful pre-trained generator.

Sketch-to-photo translation The sketch-to-photo application received attention in recent years thanks to the advancement in the field of image generation. For example, SketchyGAN [5] proposes an a GAN-based approach based on edge-preserving image augmentations, ContextualGAN [32] leverages conditional GAN along with joint image-sketch representation, iSketchNFill [17] uses a gating mechanism to condition output images on the class label and an MUNIT-based generator to synthesize images with diverse appearance. Photosketcher [12] uses sketch-based image retrieval to compose a real image. PITI [51] is pre-trained with ground truth edge maps and semantic segmentation maps to learn a domain invariant semantic representation. The state-of-the-art supervised method CoGS [20] learns the structure embeddings using the ground truth image-sketch pairs in the vector-quantized space of a VQ-GAN [13]. In contrast, MaskSketch does not rely on paired data for training, which allows it to use sketches of different abstraction levels, as well as real photos.

3. Method

In this section, we describe the main components in MaskSketch that introduce sketch-guided spatial control to a conditional masked image generator. We first review masked image generation in Section 3.1. Then we introduce the two main components of MaskSketch, a structure similarity distance in Section 3.2, and structure-guided parallel sampling in Section 3.3. Finally, we discuss how to balance the trade-off between structure fidelity and generation realism in Section 3.4.

3.1. Background: Masked Image Generation

Masked image generation is a state-of-the-art approach for efficient generation [4, 19, 29], combining the strengths of masked token modeling [10] and non-autoregressive sampling [16]. It encodes images as discrete sequences of visual tokens using a VQ-GAN encoder [13], and then trains a bi-directional transformer (BERT [10]) to model natural image distribution in the discrete token sequence space. Generation is performed iteratively, where significant gains in efficiency are obtained by using parallel sampling instead of auto-regressive sampling. MaskGIT [4] starts from a blank canvas where all visual tokens are masked. At each sampling iteration, all the missing tokens are sampled in parallel, and a rejection criteria is used, where the tokens with low model likelihood are masked and will be re-predicted in the next refinement iteration. See Figure 3 (left) for an illustration of a single MaskGIT decoding step. MaskSketch extends the parallel sampling of MaskGIT to sample images that follow the structure determined by an input image (Fig. 3, right), as described in the following sections.

3.2. Structure Similarity via Attention Maps

We consider two images to be structurally similar when their self-similarity maps are close to each other. MaskSketch leverages the self-similarity encoded in the self-attention maps of a masked generative transformer (Section 3.1) to define structural distance. One key observation in our work is that a class-conditional MaskGIT trained on ImageNet shows a high degree of domain invariance in its attention maps and is able to capture the self-similarity in out-of-distribution domains such as sketches (Fig. 2).

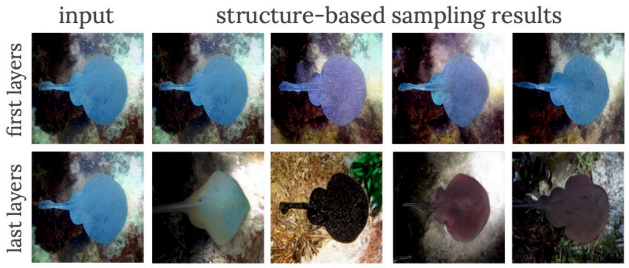

Figure 4. Structure-guided sampling using attention maps from the first three layers of a masked generative transformer results in a nearly perfect reconstruction of the input, whereas using the final layers $(16, 18, 20)$, out of 24, yields realistic images with a similar structure but highly diverse appearance. (*Best viewed in color.*)

Formally, we define a structural distance based on a comparison of self-attention maps. Let $\mathbb{Z}$ be the indices representing the VQ-GAN [13] dictionary of vector-quantized image tokens. Let $\boldsymbol{x} \in \mathbb{Z}^N$ be sequence of N discrete tokens obtained using a VQ-GAN encoding an input image in the vector-quantized space. Given an input image $\boldsymbol{x}$ and a generated image $\boldsymbol{y}$, let $A^\ell(\boldsymbol{x}) \in [0,1]^{N\times N}$ be the transformer self-attention map at layer ℓ. Each row in $A^\ell(\boldsymbol{x})$ represents the attention weights of each token with respect to all tokens, normalized with a softmax function. We define the structural distance between the i^{th} tokens of images $\boldsymbol{x}$ and $\boldsymbol{y}$ across layers $\mathcal{L}$ as:

$$d_S^i(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\ell \in \mathcal{L}} d_J\left(A_i^\ell(\boldsymbol{x}), A_i^\ell(\boldsymbol{y})\right), \tag{1}$$

where d_J is the Jeffrey's divergence:

$$d_J(\boldsymbol{u}, \boldsymbol{v}) = \frac{KL\left(\boldsymbol{u} \| \boldsymbol{v}\right) + KL\left(\boldsymbol{v} \| \boldsymbol{u}\right)}{2}. \tag{2}$$

Intuitively, the image regions represented by the i^{th} tokens of $\boldsymbol{x}$ and $\boldsymbol{y}$ are structurally similar if their distributions of attention self-similarities are close to each other.

3.3. Structure-guided Parallel Decoding

MaskSketch adapts the parallel sampling of MaskGIT to take into account the structural similarity between the output and the reference input sketch. More precisely, the token rejection criteria in each decoding iteration is modified to also reject the sampled tokens that have low self-similarity score (1). The proposed structure-guided decoding strategy can also be seen as a greedy optimization technique that balances minimizing the structural distance and following the model's image prior.

While MaskGIT sampling rejects token candidates with the lowest likelihood by masking them at the end of each decoding iteration, MaskSketch creates an additional mask that rejects tokens based on the structural similarity to the input sketch (see Fig. 3). At the end of one decoding iteration, we compute the logical OR between the confidence-based and the structure-based masks to optimize both realism and structure similarity (Sec. 3.4). The pseudocode of our algorithm is described in Algorithm 1. It relies on the function `sample_mask`, which takes as input a vector of structure similarity scores $\boldsymbol{s}^s$ and the number of masked elements k, and samples a mask by Gumbel top-k using $\boldsymbol{s}^s$ to mask the tokens with the highest structure distance.

The choice of layers $\mathcal{L}$ selected for computing the structure similarity significantly impacts the generation results. Our experiments show that sampling with the attention maps extracted from the first layers results in nearly identical reconstruction of the given input image, whereas minimizing the structure distance based on the last layers results in images of diverse appearance that are spatially aligned with the input image at a high level, as shown in Fig. 4.

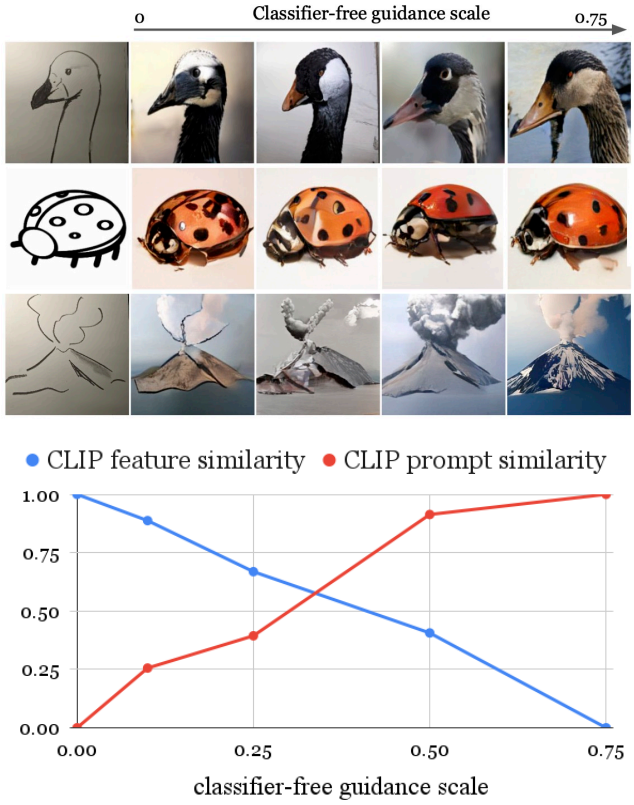

Figure 5. **Top:** Traversal of the realism-fidelity trade-off by varying the classifier-free guidance scale. **Bottom:** Increasing the guidance scale leads to higher realism and a high CLIP similarity to the prompt "photo of a c", where c is the class name, at the cost of lower structure fidelity and lower CLIP feature similarity.

3.4. Structure fidelity vs realism trade-off

One of the biggest challenges in sketch-to-real translation is the immense domain gap between the source and target domains. The input sketches and natural images differ

significantly not only in appearance, but also in the distribution of shapes and spatial composition. Due to the domain gap, optimization based solely on the structure distance often results in structurally similar but unrealistic images. To overcome this issue, we propose a combined masking approach that optimizes both structure fidelity and realism.

To navigate this trade-off, we use a parameter $\lambda_s \in [0, 1]$ to determine the proportion of tokens masked according to the structure similarity scores and those masked according to the model confidence or likelihood scores. Given an overall masking rate schedule function $\gamma(t)$ at step t, the structure-based mask rate is computed as $\lambda_s\gamma(t)$, whereas the confidence-based mask rate is $(1 - \lambda_s)\gamma(t)$. Two independent masks, $\boldsymbol{m}_t^s$ and $\boldsymbol{m}_t^c$, are computed for the structure-based and confidence-based scores, respectively. The final mask at iteration t is then computed as the logical OR between $\boldsymbol{m}_t^s$ and $\boldsymbol{m}_t^c$. Please refer to Fig. 9 in the Appendix for an ablation study on the parameter λ_s.

Classifier-free Guidance To further improve the level of realism in the translation result, we use classifier-free guidance [14, 24, 35] when computing the model likelihood scores. Specifically, for a given sequence of sampled tokens $\bar{\boldsymbol{y}}$ and input image $\boldsymbol{x}$, we use the pre-trained generator G to compute the per-token logits $\log p(\bar{\boldsymbol{y}}(i)|\boldsymbol{x}, c)$ conditioned on the correct class c and logits conditioned on a random class r: $\log p(\bar{\boldsymbol{y}}(i)|\boldsymbol{x}, r)$, and calibrate the final confidence-based scores as follows:

$$\begin{aligned}\boldsymbol{s}^c(i) = &\log p(\bar{\boldsymbol{y}}(i)|\boldsymbol{x}, c) \\ &- \beta \left(\log p(\bar{\boldsymbol{y}}(i)|\boldsymbol{x}, c) - \log p(\bar{\boldsymbol{y}}(i)|\boldsymbol{x}, r)\right) \end{aligned} \quad (3)$$

where β is the classifier-free guidance scale. Figure 5 shows how varying β affects the fidelity-realism trade-off.

Global CLIP-based rejection sampling Minimization of the structure similarity distance in the space of visual tokens is a discrete optimization problem that cannot be efficiently tackled via continuous optimization methods such as gradient descent. Moreover, since MaskGIT was trained to minimize a different objective, such a greedy optimization process requires more iterations than regular sampling. To increase the stability of the proposed method, we improve the overall fidelity by producing multiple translation samples for a given sketch with different random seeds and guidance scales β, and selecting the image that yields the highest structure fidelity and realism according to a CLIP-based score. Inspired by the recent success in photo-to-sketch mapping with CLIP [39] domain-invariant representations [49], we use the L_1 distance between features of a CLIP encoder $\text{CLIP}_s(\boldsymbol{x}, \boldsymbol{y})$ of the input image $\boldsymbol{x}$ of class c and generation result $\boldsymbol{y}$ to estimate the structure similarity. We also use the CLIP similarity score $\text{CLIP}_r(c, \boldsymbol{y})$ between the translated image and the corresponding prompt $\text{prompt}(c) =$ 'photo of a c' to assess the realism for each generated example, more details can be found in Appendix F. We normalize the scores across R trials and keep the result with the highest overall quality score:

$$\boldsymbol{y}_{final} = \underset{\boldsymbol{y}\in\{\boldsymbol{y}_1 \dots \boldsymbol{y}_R\}}{\text{argmax}} \; (1-\text{CLIP}_s(\boldsymbol{x}, \boldsymbol{y}))^2 \, \text{CLIP}_r(c, \boldsymbol{y}) \quad (4)$$

Tab. 5 in the Appendix shows how the proposed CLIP-based selection approach improves the overall generation result as the number of sampling trials increases.

Algorithm 1 MaskSketch sampling

Input: Pre-trained BERT generator G, structure and confidence masking schedule function $\gamma(t)$, structure-based sampling ratio λ_s, input sketch $\boldsymbol{x}$, layer(s) ℓ.
Output: Generated image encoding $\boldsymbol{y}_0$

$A^\ell(\boldsymbol{x}) \leftarrow \texttt{attn\_map}(G, \boldsymbol{x}, \ell)$
Initialize $\boldsymbol{y}_T$
for $t = T - 1 \dots 0$ **do**
  $\bar{\boldsymbol{y}}_t,\ \boldsymbol{s}_t^c = G(\boldsymbol{y}_{t+1})$
  $A^\ell(\bar{\boldsymbol{y}}_t) = \texttt{attn\_map}(G, \bar{\boldsymbol{y}}_t, \ell)$
  $\boldsymbol{s}_t^s \leftarrow \left\{d_J\left(A_i^\ell(\bar{\boldsymbol{y}}_t),\ A_i^\ell(\boldsymbol{x})\right)\right\}_{i=1\dots N}$
  $\boldsymbol{m}_t^s = \texttt{sample\_mask}(\boldsymbol{s}_t^s, \lfloor \lambda_s\gamma(t) \cdot N \rfloor)$
  $\boldsymbol{m}_t^c = \texttt{sample\_mask}(\boldsymbol{s}_t^c, \lfloor (1 - \lambda_s)\gamma(t) \cdot N \rfloor)$
  $\boldsymbol{m}_t = \boldsymbol{m}_t^s \ \vee \ \boldsymbol{m}_t^c$
  $\boldsymbol{y}_t = \bar{\boldsymbol{y}}_t \odot \boldsymbol{m}_t$
end for

4. Experiments

Experimental Setup In all experiments, we used a class-conditional MaskGIT model pretrained on the ImageNet 2012 [9] dataset with the output resolution 256×256. We used layers $1, 3, 16, 20, 21$ and 22 to formulate the structure preservation objective. We validated this choice on 100 random sketches considering structure preservation and realism. In our experiments, for each input sketch, we sample the images four times ($R = 4$) with different classifier-free guidance scales (i.e. $\beta \in \{0.0, 0.05, 0.1, 0.25\}$), and select the one that maximizes the CLIP-based objective in Eq. (4). We use a linear decay mask rate schedule in all experiments, starting from $\gamma(T) = 0.95$ and stopping sampling at the mask rate $\gamma(0) = 0.25$, which results in higher realism and reduces artifacts associated with structure-based sampling. To further increase realism, we postprocess the generated samples with Token-Critic refinement [29], which adds 32 sampling iterations. See Appendix D for more details. We generate each image using $T = 500$ sampling iterations, and the overall sampling time for a batch of 8 images is on average 750 seconds on a single TPUv4, including four trials and the CLIP-based evaluation.

Baselines We consider well-established *unpaired* image-to-image translation methods as baselines. Specifically, we

used CUT [37], MUNIT [25] and VQI2I [6] in our comparisons. We note that for sketch-to-photo translation, methods that use ground truth attribute information to translate from one attribute to another, e.g. StarGAN [7, 8], fail to minimize the gap between sketch and real domains [20]. CUT [37] uses a contrastive objective to ensure structural similarity between the corresponding patches of the input image and the translation result. MUNIT [25] is a GAN-based model that uses latent embedding reconstruction losses to disentangle appearance from structure. VQI2I [6] uses a vector-quantized GAN to encode images into sequences of tokens representing the structure and appearance of the input images, and uses embedding reconstruction losses to enforce the disentanglement of the structure. Since these methods are not class-conditional, we trained them on each class separately. We report the average result across the examples of all classes as well as the results of training on the entire datasets.

Although MaskSketch does not utilize paired data, we also consider as baseline the state-of-the-art *paired* sketch-to-photo method CoGS [20]. We note that VQI2I, CoGS and MUNIT allow diverse sampling with an additional appearance image or vector as input, whereas MaskSketch samples diverse results by varying the random seed.

Datasets For qualitative evaluation of MaskSketch, we propose OpenSketches, a novel dataset made of 200 openly licensed sketches. OpenSketches contains real sketches drawn with pencil and paper, as well as digital ske'tches. Furthermore, to mimick realistic, highly detailed sketches, we utilized the open source implementation of Stable Diffusion [44] to generate input examples. All the sketches shown in this manuscript are from OpenSketches.

For quantitative evaluation, we considered two datasets: ImageNet-Sketch [50], a dataset of 50 real sketches of 1000 classes of ImageNet-2012 [9] and the Pseudosketches dataset [20], consisting of pairs of ground truth real images and their corresponding automatically extracted edge maps from 125 classes from the ImageNet21K [43] dataset. We present qualitative results for these datasets in Appendix C[1]. In our quantiative experiments, we report the results on two versions of the datasets: 1) *full*: using all examples from each of the datasets, and 2) *10-class*: using the 10 classes that are reported to result in the highest-quality translation results in CoGS [20]: "songbird", "pizza", "volcano", "zebra", "castle", "door", "shark", "mushroom", "cup", "lion". The 10-class subsets of Pseudosketches and ImageNet-Sketch consist of 1,749 and 508 examples respectively, whereas the full datasets consist of 113,370 examples and 52,888 examples, respectively. For the 10-class subsets, we trained unpaired image translation baselines that are not class-conditional on each class separately and reported the aggregated results over all 10 classes for a fair comparison with the class-conditional MaskSketch. For the full version of the datasets, we train the baseline methods on all classes without class conditioning. Since ImageNet-Sketch does not provide ground truth paired data, it is impossible to train CoGS [20] on this dataset, therefore we use the model trained on Pseudosketches for both datasets.

[1] Not shown in the main manuscript due to copyright concerns.

Metrics Quantitative evaluation of sketch-to-photo translation consists of two aspects: evaluation of realism of the generation results, and evaluation of structure fidelity with respect to the input sketch. To estimate realism, we use the FID score [23]. To assess generation diversity, we used the LPIPS-based diversity score [36], which computes the average LPIPS [56] distance between the generated examples. For a fair comparison and due to the limited number of samples in the 10-class subsets, we report the FID and LPIPS results over 10,000 examples generated with different 'appearance' inputs with the baseline methods CoGS, MUNIT and VQI2I, and with different seeds for MaskSketch. For CUT, we diversify the generated set with augmentations. The FID score is computed with respect to the images from ImageNet [9] for the ImageNet-Sketch experiments, and with respect to the ground truth Pseudosketches images for the Pseudosketches experiments.

To provide additional quantitative evaluation of structure preservation quality and realism, we also report the two CLIP-based metrics defined in Sec. 3.4: image feature distance and prompt similarity score. The CLIP feature distance metric is more appropriate for the evaluation of structure preservation quality than the edge-based metrics [20] since the CLIP features are more invariant to the domain gap as shown in the recent works on image-to-sketch translation [49]. We note that these metrics are identical to the CLIP-based rejection sampling in Sec. 3, and we include the quantitative results without CLIP-based sampling in Appendix C.1.

User Preference Studies Quantitative evaluation of structure fidelity is challenging due to the distribution shift between the shapes of real objects and abstract sketches and outlines. To complement the quantitative results, we performed user preference studies. We asked users the question: *"Given the task of converting the sketch shown on the left into a realistic photo, which result do you prefer?"*. Users were asked to pick one result among the five compared methods (CoGS, MUNIT, VQI2I, CUT and MaskSketch) according to their preference. We collected three preference evaluations for each example in the 10-class ImageNet-Sketch and Pseudosketches datasets. Finally, we counted only unanimous votes to guarantee statistical significance. Please see Appendix E for more details.

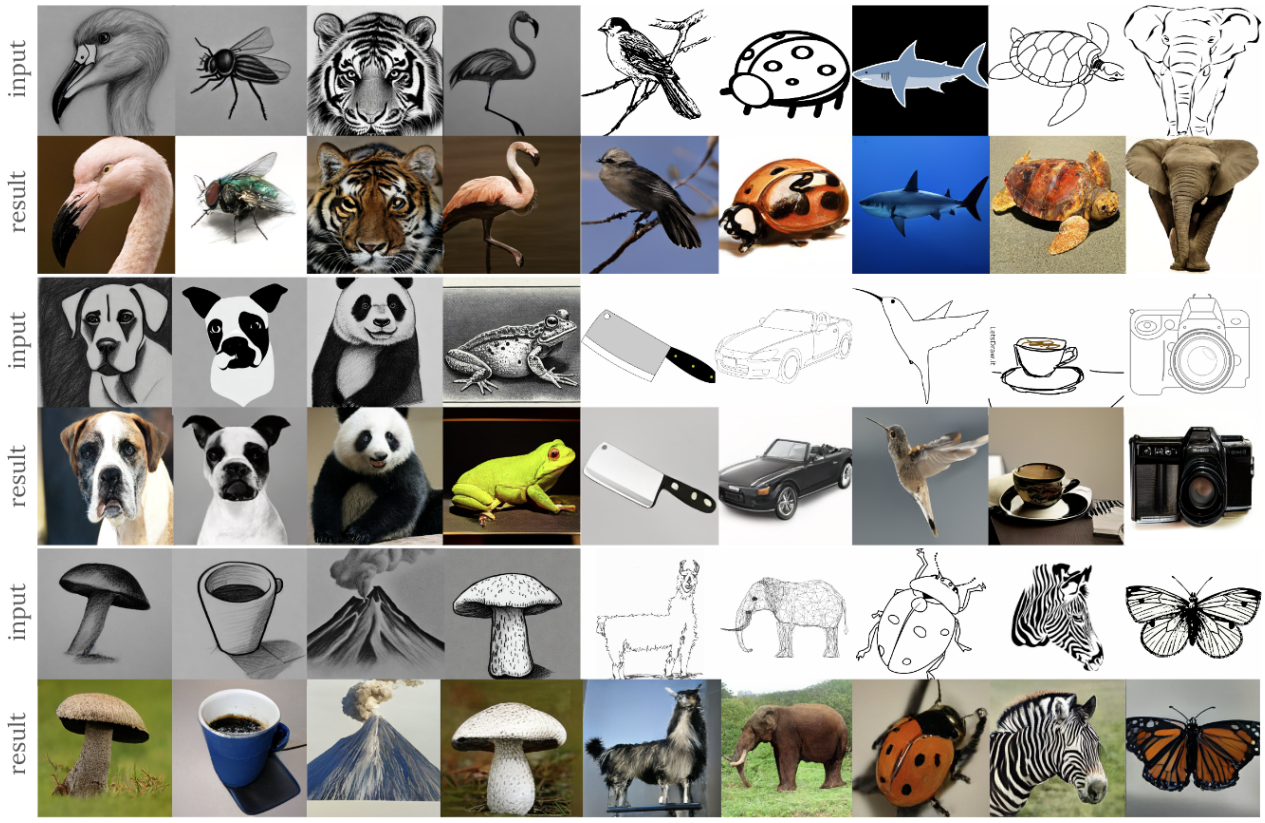

Figure 6. Example translations by MaskSketch on the OpenSketches dataset. The model takes as input a sketch and a class label.

5. Results

Quantitative Results In Tab. 1, we report the quantitative evaluation results on the 10-class subsets of ImageNet-Sketch and Pseudosketch. Additionally, we report the FID and LPIPS diversity scores over the full ImageNet-Sketch and Pseudosketch datasets in Tab. 2. The results on the 10-class subsets indicate an advantage of MaskSketch in terms of realism and diversity, with a two-fold decrease in the FID score compared to the baseline MUNIT on both ImageNet-Sketch and Pseudosketch datasets. In our experiments, MaskSketch outperformed the baselines on the entire ImageNet-Sketch dataset of real sketches, including the fully-supervised CoGS on the Pseudosketches. Notably, general image translation methods, such as MUNIT and CUT, outperform the fully-supervised CoGS in a class-supervised setup. As seen from the FID and LPIPS results, VQI2I struggles to generalize on the relatively small ImageNet-Sketch dataset that contains only 50 examples in each class, mainly due to mode collapse.

Qualitative Results Qualitative comparison shows that the baseline image translation methods, including the supervised CoGS, capture the overall layout and outlines of the input sketch but sometimes fail to produce realistic results. For instance, the GAN-based architectures, namely CUT and MUNIT, produce structurally similar results by practically recoloring the input sketch, which results in a sub-par realism, especially on the more abstract sketches. In our experiments, the VQ-GAN-based VQI2I model failed to learn the correspondences between hand-drawn sketches from ImageNet-Sketch and images from the real photo domain due to a limited number of examples in ImageNet-Sketch, therefore we observe a severe mode collapse on most classes. The fully-supervised CoGS sometimes failed to produce realistic and semantically meaningful results, especially on the hand-drawn sketches. MaskSketch achieved a good balance between realism and structure fidelity on the majority of sketches from ImageNet-Sketch. However, MaskSketch struggled to preserve structure on some examples from Pseudosketches due to the extreme complexity of the extracted edge maps.

To further refine the resulting token sequence to produce a more realistic image, we use token-critic refinement as proposed by Lezama et al. [29]. Token-critic approach uses a discriminator to detect and refine the tokens that decrease the probability of the overall sequence.

	supervision	FID ↓	LPIPS ↑	CLIP prompt ↑	CLIP feat. ↓	User preference ↑
ImageNet-Sketch						
MUNIT	class	68.65	0.58	52.10	30.27	10.70%
CUT	class	77.74	0.68	65.59	28.05	19.78%
VQ-I2I	class	181.77	0.32	53.76	31.05	0%
CoGS	class + pairs	97.31	0.64	56.62	29.52	8.55%
MaskSketch(ours)	class	**33.24**	**0.78**	**67.10**	**26.63**	**59.35%**
Pseudosketches						
MUNIT	class	93.23	0.69	41.91	27.65	23.08%
CUT	class	112.11	0.42	45.67	27.62	25.0%
VQ-I2I	class	169.1	0.77	34.5	28.47	0.64%
CoGS	class + pairs	102.66	0.68	34.79	27.52	14.10%
MaskSketch (ours)	class	**56.55**	**0.78**	**59.48**	**25.60**	**35.25%**

Table 1. Sketch-to-photo translation performance on ImageNet-Sketch (**top**) and Pseudosketches (**bottom**) 10-classes subsets.

	ImageNet-Sketch		Pseudosketches	
	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
MUNIT	113.45	0.74	121.69	0.71
CUT	161.33	0.74	163.82	0.70
VQI2I	131.70	0.72	135.47	0.71
CoGS (sup.)	85.09	0.72	49.31	0.71
MaskSketch	**23.89**	**0.77**	**46.44**	**0.78**

Table 2. Comparison on the full ImageNet-Sketch (**left**) and Pseudosketches-validation (**right**) datasets.

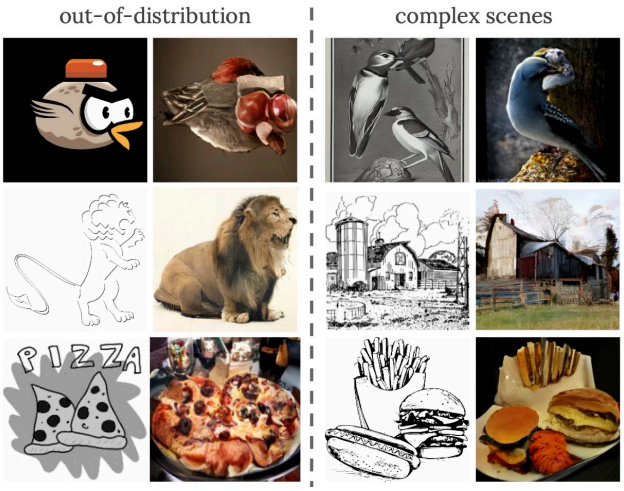

Figure 7. Failure cases of MaskSketch on the hand-drawn sketch examples: out-of-distribution composition and shapes, complex scenes containing multiple objects.

Limitations The main limitation of MaskSketch is computational efficiency. To achieve a successful optimization of the structural constraint, MaskSketch requires significantly more sampling iterations than the regular MaskGIT. Furthermore, to improve the stability of results it was necessary to apply a multiple trials rejection scheme. Two other important limitations for MaskSketch are the coarse granularity of the attention maps in a transformer, and the flexibility of the prior model, in our case an ImageNet-pretrained MaskGIT. Figure 7 illustrates the common failure cases of our method: out-of-distribution scene composition scarcely or not represented in the training set of MaskGIT, multiple objects forming an unrealistic scene, as well as the complex scenes with multiple foreground and background objects.

6. Conclusion

We proposed MaskSketch, a sketch-guided image generation method that allows control over the spatial layout of the generation result. MaskSketch achieves high realism and structure preservation without pairwise supervision, does not require model finetuning and works on sketches of various levels of abstraction. We show that the self-attention maps of the intermediate layers of a masked generative transformer encode important structural information of the input image and are sufficiently domain-invariant, which allows their use in a structure similarity constraint. Our experimental results show that the proposed attention-based sampling approach outperforms state-of-the-art sketch-to-photo and general image translation methods in terms of both realism and structure fidelity.

Acknowledgements We thank Tali Dekel, Huiwen Chang, Lu Jiang, and David Salesin for their insightful advice and guidance. This work was done during an internship at Google Research.

References

[1] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. 12

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1, 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 1, 2, 3

[5] 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, 3

[6] Yu-Jie Chen, Shin-I Cheng, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Vector quantized image-to-image translation. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 6

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 6

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 6

[9] 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. 5, 6

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 2

[12] 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. 3

[13] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 1, 2, 3, 4

[14] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 1, 2, 5

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[16] Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019. 3

[17] 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/CVF International Conference on Computer Vision*, pages 1171–1180, 2019. 3

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[19] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. *arXiv preprint arXiv:2111.14822*, 2021. 3

[20] Cusuh Ham, Gemma Canet Tarres, Tu Bui, James Hays, Zhe Lin, and John Collomosse. Cogs: Controllable generation and search from sketch and style. *European Conference on Computer Vision*, 2022. 1, 2, 3, 6

[21] Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2020. 12

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 3

[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. *Advances in neural information processing systems*, 30, 2017. 1, 6

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5

[25] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 1, 2, 6

[26] 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. 2

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[28] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2

[29] José Lezama, Huiwen Chang, Lu Jiang, , and Irfan Essa. Improved masked image generation with token-critic. *European Conference on Computer Vision*, 2022. 3, 5, 7, 13

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 14

[31] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017. 2

[32] 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, 3

[33] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2

[34] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 1

[35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 5

[36] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021. 6

[37] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020. 1, 6

[38] 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*, 2019. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[41] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[42] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 2

[43] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 6

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 6

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2

[47] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020. 2

[48] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. *arXiv preprint arXiv:2201.00424*, 2022. 2

[49] Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching, 2022. 5, 6

[50] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 6

[51] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. In *arXiv*, 2022. 3, 14

[52] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. N\" uwa: Visual synthesis pre-training for neural visual world creation. *arXiv preprint arXiv:2111.12417*, 2021. 2

[53] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 14

[54] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[55] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 2

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[57] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11465–11475, 2021. 2

[58] 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. 2