Salient Object-Aware Background Generation using Text-Guided Diffusion Models

Amir Erfan Eshratifar[1], João V. B. Soares[1], Kapil Thadani[1],
Shaunak Mishra[2*], Mikhail Kuznetsov[2*], Yueh-Ning Ku[3*], and Paloma de Juan[1]

[1]Yahoo Research [2]Amazon [3]ByteDance

Abstract

Generating background scenes for salient objects plays a crucial role across various domains including creative design and e-commerce, as it enhances the presentation and context of subjects by integrating them into tailored environments. Background generation can be framed as a task of text-conditioned outpainting, where the goal is to extend image content beyond a salient object's boundaries on a blank background. Although popular diffusion models for text-guided inpainting can also be used for outpainting by mask inversion, they are trained to fill in missing parts of an image rather than to place an object into a scene. Consequently, when used for background creation, inpainting models frequently extend the salient object's boundaries and thereby change the object's identity, which is a phenomenon we call "object expansion." This paper introduces a model for adapting inpainting diffusion models to the salient object outpainting task using Stable Diffusion and ControlNet architectures. We present a series of qualitative and quantitative results across models and datasets, including a newly proposed metric to measure object expansion that does not require any human labeling. Compared to Stable Diffusion 2.0 Inpainting, our proposed approach reduces object expansion by 3.6× on average with no degradation in standard visual metrics across multiple datasets.

1. Introduction

Image outpainting, also known as image extrapolation or extension, has been a longstanding challenge within computer vision. Prior image outpainting techniques relied upon retrieval and stitching methods using image patches, or learning-based methods [9, 17, 21, 45, 47, 57]. The new wave of generative image models [34–36] has been adapted to also solve the outpainting task, representing a breakthrough in image quality and adding controllability via text prompts and other control inputs. Our work focuses specifically on *salient object outpainting*, an outpainting problem that involves generating a natural and coherent background for a salient object, while optionally conditioned on a text prompt, as shown in Figure 1.

Given an object, humans can readily imagine its empirical context by relating objects to their context in daily life while also being able to imagine them in unconventional settings, such as a swan in a bedroom, as depicted in Figure 1. There are many potential applications for the salient object outpainting problem we study here, such as generating backgrounds for products in online advertising, filmmaking, creative design, and augmented reality. Object outpainting is much more challenging than usual image completion tasks like inpainting and outpainting for two reasons: (i) the object and background contents may not be related to each other, (ii) to generate a background constrained by the salient object, the model needs to understand the correlations within the scene at a semantic level.

Recently, diffusion models [14, 39] such as Latent Diffusion Models [35], unCLIP [34], and Imagen [36] have shown outstanding results in text-to-image generation. An early approach for adapting them to the inpainting task consisted of replacing the random noise in the fixed portion of the image with a noisy version of itself during the diffusion reverse process [26]; however, the model's inability to observe the global context during sampling led to unsatisfactory samples [29]. GLIDE [29] and Stable Inpainting (SI) [35] improved upon this by using the masked image as extra conditioning information to the reverse diffusion process. They trained their models using randomly generated masks to specify what portion of the image to inpaint; however, the masks were randomly placed and had circular, square, or highly irregular shapes which are rarely ever seen in real-world inpainting scenarios. To better mimic real-world inpainting masks, later works [53, 59] propose masks that follow object shapes obtained from various segmentation datasets. To ensure that salient objects are not masked out, they subtract the portion of some masks that correspond to salient objects in the training image. However, these masks can be from any object, resulting in small

* Work done at Yahoo Research

Figure 1. Examples of outpainting a salient object (leftmost column) using the Stable Inpainting 2.0 (SI2) model (columns 2, 4, 6 from left) and using our proposed model (columns 3, 5, 7 from left). The images in each paired column (2 & 3, 4 & 5, 6 & 7) are generated using the same seed and prompt, but one uses SI2, and the other uses our model. Objects are often expanded using the SI2 model, which may catastrophically change the object's identity. For example, the legs of the tables are expanded in the first two rows; in the third row, a bench is transformed into a bed; in the last row, a swan is blended into a rock and a bed.

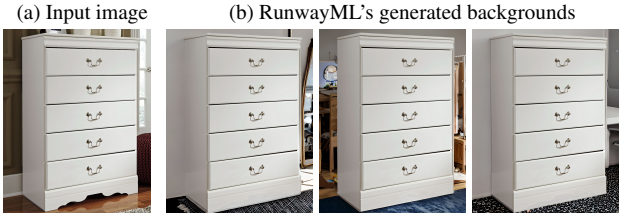

Figure 2. Significant object expansion is seen at the bottom of the white dresser with RunwayML's Background Remix, a popular commercial tool. These examples are generated with the prompt of "a modern room."

masks relative to the image size. Thus, despite these improvements, inpainting models are primarily trained to fill in missing parts of an image rather than synthesize complete backgrounds conditioned on salient objects.

In practice, we observe that when such inpainting models are used for background generation, they often ignore the salient object's original boundaries and modify or recharacterize the object, as shown in Figure 1 for the Stable Inpainting 2.0 model (SI2) [35, 41]. We call this phenomenon *object expansion*. As shown in Figure 2, even popular commercial tools for background generation are prone to this limitation. To quantify object expansion, we propose an automated metric that avoids the need for any human labeling. We also propose a solution to object expansion using ControlNet [55] to maintain object boundaries. ControlNet was introduced to control large text-to-image models with extra input conditions like edge maps, segmentation maps, key points, etc. We utilize the mask of the salient object as a new input condition to address expansion. Although ControlNet was initially designed for standard text-to-image diffusion models, we modify its architecture to be compatible with diffusion-based inpainting.

Though our approach could also be applied to background generation for non-salient objects, we train our

model specifically for salient object outpainting for two reasons. First, we can use readily available manually annotated salient object detection datasets to train our salient object outpainting model. If we were to tackle background generation for non-salient objects, we would have to train our model with panoptic or instance segmentation datasets, whose masks, unfortunately, are not pixel-perfect, containing noisy labels that may not help reduce the object expansion problem. Second, a notable application for background generation is e-commerce, where personalized and aesthetically pleasant backgrounds can be generated for products that would be salient objects in the final images.

Our main contributions are:

- A novel study of diffusion-based inpainting models applied to salient object-aware background generation.
- A characterization of the object expansion problem when inpainting models are applied for background generation, as well as a measure for quantifying it.
- An architecture based on ControlNet for adapting diffusion-based inpainting models to salient object-aware background generation.*
- Extensive experimental evaluation, comparing our proposed approach to prior work across various metrics and demonstrating its effectiveness in addressing object expansion. Compared to a state-of-the-art baseline, our proposed approach reduces the object expansion by 3.6× on average with no degradation in standard visual metrics across multiple datasets.

2. Related Work

2.1. Diffusion Models

Diffusion models [14, 39] are a class of generative models that learn the data distribution by learning to invert a Markov noising process. They have gained widespread attention recently due to their training stability and superior performance in image synthesis compared to prior approaches such as generative adversarial networks (GANs). Given a clean image x_0, the diffusion process adds noise to the image at each step t, obtaining a set of noisy images x_t. Then, a model is trained to recover the clean image x_0 from x_t in the backward process. Diffusion models have produced appealing results on different tasks, e.g., unconditional image generation [14, 16, 39, 42], text-to-image generation [33–36], video generation [15], image inpainting [1, 2, 26, 29], image translation [27, 46, 58], and image editing [6, 10, 18].

2.2. Text-guided Image Inpainting

Leveraging the recent triumph of diffusion-based text-to-image models, a natural transition from text-to-image creation to text-guided inpainting involves running diffusion with a standard synthesis model, but at each step replacing the portion of the image being generated that is outside the mask with a noised version of the input image. In practice, this approach does not properly condition on the input image, leading to incongruent generations. GLIDE [29] effectively addresses this issue by using the masked image and mask as direct conditioning inputs to the diffusion model. Blended Diffusion [1] promotes the alignment of the final output with the text prompt through the use of a CLIP-based score [32]. The Repaint method [26] resamples during each retrograde step, yet lacks support for text input. PaintByWord [3] creates an alliance between a large-scale generative adversarial network (GAN) and a complete-text image recovery network, facilitating multi-modal image editing; however, the GAN structure restricts specific modifications to regions indicated by the mask. TDANet [54] introduces a dual attention mechanism that uses text features related to the masked area by contrasting the text with the original and noised image. SmartBrush [49] proposes a diffusion-based model for completing a missing region with an object using text and shape guidance. None of the prior arts study the task of background generation for salient objects using diffusion models.

3. Salient Object Outpainting

Here, we introduce our proposed model architecture for salient object outpainting. We use Stable Inpainting 2.0 (SI2) as a base model and add the ControlNet model on top to adapt it to the salient object outpainting task. We explain each component of our model in the following subsections.

3.1. Stable Inpainting

The training of Stable Diffusion (SD) [35], a text-to-image diffusion model, involved billions of images. The main component of the model is a denoising U-Net, which itself consists of an encoder, a middle block, and a decoder, with skip connections between the encoder and decoder blocks. The U-Net is composed of 25 blocks, divided into 12 symmetric blocks for each encoder and decoder, plus a middle block. There are 25 blocks in total, with 8 being downsampling or up-sampling convolution layers, and the remaining 17 being main blocks containing two transformer layers and four residual layers each. Each transformer layer contains several cross-attention and/or self-attention mechanisms. CLIP is the source of text embeddings while sinusoidal positional encoding is used for diffusion time steps.

SD has been shown to be a competent and versatile text-to-image generative model. Stable Diffusion `v2-base` [40], referred to as SD2, was initially trained for 550k steps at 256×256 pixel resolution on a subset of LAION-5B [37] with aesthetic score of 4.5 or higher. SD2 differs from previous versions due to its subsequent training on a dataset with at least 512 x 512 pixel resolution, resulting in more detailed and visually appealing images. Stable Diffusion `v2-inpainting` [41], referred to here as SI2, is built on top of SD2 and trained for an additional 200k steps. The training process incorporates the mask-generation approach introduced in LaMa [42], while adding

*The code and model checkpoints will be available at https://github.com/yahoo/photo-background-generation.

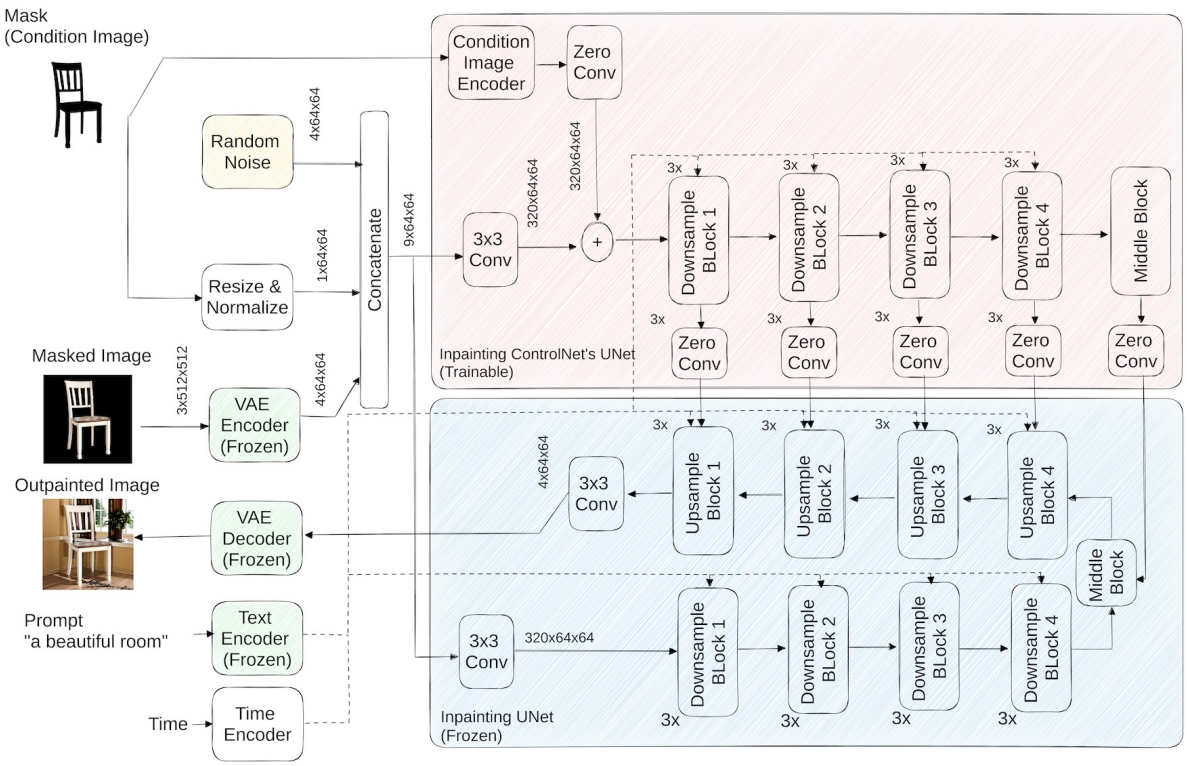

Figure 3. The proposed architecture for salient object outpainting. The original ControlNet architecture only works with text-to-image Stable Diffusion. To make it compatible with text-to-image Stable Inpainting we modified the ControlNet's U-Net architecture to take two extra inputs: 1) mask and 2) masked image. The blue region denotes the frozen Stable Inpainting's U-Net model. The red region includes replicas of the encoder layers of the blue region. The zero convolution outputs from the red region modulate the outputs of decoder layers in the blue region. Initially, during training, the modulation has no effect on the output as the weights of the convolution layer are initialized to zero. Gradually, during training, the nuances of the task of background generation for salient objects will be encoded in modulated values.

the latent VAE representations of the masked image as conditioning inputs. We select SI2 as the base model in this work because SI2 already has outpainting capabilities and provides a better initialization compared to SD2.

3.2. ControlNet for Stable Inpainting

ControlNet [55] is a neural network architecture to control the output of existing text-to-image diffusion models by enabling them to support additional input conditions. We adapt ControlNet's architecture to text-to-image *inpainting* diffusion models as they provide a good initialization point for outpainting tasks. This adaptation requires adding extra inputs to ControlNet: (i) a masked image, which contains the pixel values of the salient object; (ii) a binary mask in which **1**s are the pixels to fill in and **0**s are the pixels to keep from the salient object in the masked image. Figure 3 shows the architecture of the proposed salient object outpainting. We employ the ControlNet architecture on top of the Stable Inpainting 2.0 (SI2) model.

To enable computationally efficient training, SD applies a pre-processing technique akin to VQ-GAN [8] where the entire collection of 512×512 images is transformed into smaller $(64 \times 64 \times 4)$ *latent images*. To match the convolution size, it is necessary to convert image-based conditions to a $64 \times 64 \times 4$ feature space in the ControlNet architecture. The image-space condition is encoded into feature maps with a tiny neural network comprising of four convolution layers. The network uses 4×4 kernels and 2×2 strides, ReLU activations, and channel dimensions of 16, 32, 64, and 128 (respectively for each of the four convolution layers) and is initialized with Gaussian weights. This network is trained jointly with the ControlNet model and later passed to the U-Net model. Training the ControlNet is computationally efficient as the original weights of the UNet are locked, and only the gradients from the UNet's decoder are required so we need not compute gradients for the original UNet's encoder. As only the encoder layers are copied to ControlNet, the cost of running the decoder layers is avoided. Specifically, in the forward pass, we must operate a lightweight mask encoder, two U-Net encoders, one U-Net decoder, and a few zero convolution layers. We observe a 33% increase in runtime and a 25% increase in GPU memory consumption on V100 GPUs when fine-tuning the whole SI2's U-Net, relative to training the ControlNet.

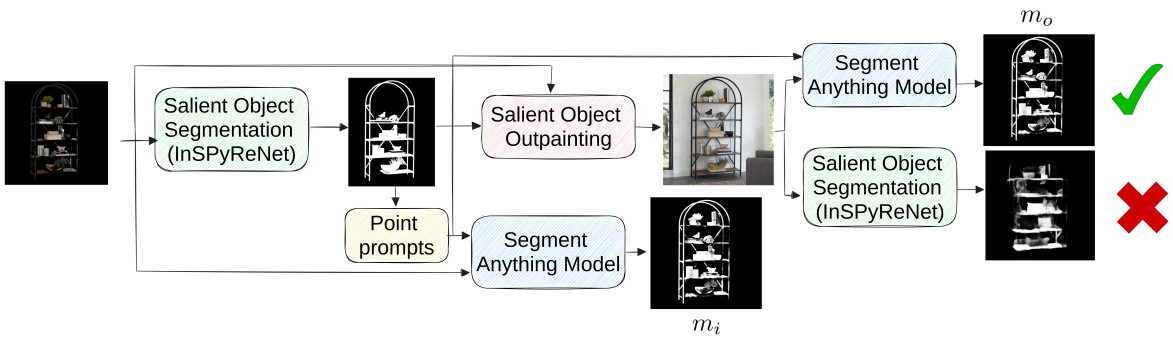

Figure 4. Pipeline for computing salient object masks of the original image (m_i) and the outpainted image (m_o) to measure object expansion. We found that existing salient object segmentation (SOS) models underperform on synthetic images, but the Segment Anything Model (SAM) works robustly. Therefore, we (i) obtain the salient object mask of the original image using the SOS model, (ii) sample random points from the mask, and (iii) pass sampled point coordinates as the input point prompt to SAM to obtain the salient object mask m_o. We obtain a new salient mask from SAM for the original image (m_i) as well for an apples-to-apples comparison with m_o.

In the ControlNet's U-Net, the downsampling blocks (i.e., encoder) and middle block are copied from SI2 and initialized using the same weights. The U-Net in the SI2 model takes two additional inputs besides the random noise $(4 \times 64 \times 64)$ by concatenation across the channel dimension: (i) binarized mask $(1 \times 64 \times 64)$, and (ii) encoded masked image $(4 \times 64 \times 64)$. The resulting latent input $(9 \times 64 \times 64)$ is then passed to a 3×3 convolution layer which outputs a tensor of size $320 \times 64 \times 64$. The input to the ControlNet's U-Net is the encoded condition image (i.e., salient mask) which is encoded using a convolutional encoder followed by a *zero convolution* layer.

The ControlNet uses several zero convolution layers to modify the U-Net decoder outputs gradually. Mathematically, we are given the feature map $x \in R^{h \times w \times c}$ with $\{h, w, c\}$ being height, width, and channel numbers, a U-Net encoder block $E(.; \Theta_e)$ with a set of parameters Θ_e, and a U-Net decoder block $D(.; \Theta_d)$ with a set of parameters Θ_d. We denote the zero convolution operation as $\mathcal{Z}(.; \Theta_z)$. The structure of the ControlNet we are using is defined by:

$$y = D(d; \Theta_d) + \mathcal{Z}\left(E(e; \Theta_e), \Theta_z\right) \quad (1)$$

where y becomes the output of a decoder layer modulated by the ControlNet structure. As the parameters of a zero convolution layer are initialized as zeros, in the first gradient descent step, we have $\mathcal{Z}(x; \Theta_z) = \mathbf{0}$, which means the original output of the decoder layer does not change. As a result, all the inputs and outputs of both trainable and frozen copies of the U-Net model are not changed, as if the ControlNet did not exist. When the ControlNet structure is applied to some layers before any gradient descent step, it will not influence the intermediate features.

Training. Image diffusion models learn to progressively denoise images to generate samples. The denoising process can happen in pixel space or a *latent* space encoded from training data. SD uses latent images as the training domain. Given an image (or latent image) x_0, diffusion algorithms progressively add noise to the image and produce a noisy image x_t, with $1 \leq t \leq T$ being the number of timesteps for which the noise is added. When t is large enough, the image approximates pure noise. Given a set of conditions including timestep t, text prompts c_t, as well as a task-specific condition (i.e., the salient mask) c_f, image diffusion algorithms learn a network ϵ_θ to predict the noise, ϵ_t, added to the noisy image x_t with

$$\mathcal{L} = \mathbb{E}_{t \sim [1,T], x_0, c_t, c_f, \epsilon_t} \left\| \epsilon_t - \epsilon_\theta(x_t, c_t, c_f, t) \right\|_2^2 \quad (2)$$

where $\mathcal{L}$ is the overall learning objective of the entire diffusion model which can be directly used in fine-tuning as well. As we are applying classifier-free guidance (CFG) [13], we randomly drop 10% of text guidance during training to not drift away from the learned unconditional image generation. Text dropping also facilitates ControlNet's capability to recognize semantic contents from the salient mask.

3.3. Measuring Object Expansion

The primary limitation of text-guided diffusion models for salient object outpainting tasks is their inability to preserve object boundaries. To assess a method for handling the issue of object expansion, we need a way to measure the error quantitatively. To avoid the need for expensive human labeling, we initially attempted to use salient object segmentation (SOS) models to generate salient masks for both the input (the salient object on a blank background) and output (outpainted) images. The SOS models we experimented with [19, 31] were observed to have extremely poor performance on outpainted images, which we hypothesize is likely due to a distribution shift. However, we note that the Segment Anything Model (SAM) [20] is immune to this issue on outpainted images. While SAM is not an SOS model, it can take a set of positive and negative points as a prompt to segment the objects represented by the positive points, while avoiding those represented by negative points.

Dataset	Model	FID ↓	LPIPS ↓	CLIP Score ↑	Obj. Similarity ↑	Obj. Expansion ↓
ImageNet-1k	Blended Diffusion	31.63	0.41	24.95	0.49	0.21
	GLIDE	26.35	**0.28**	24.82	0.62	0.18
	Stable Diffusion 2.0	16.90	0.38	**27.46**	0.56	0.15
	Stable Inpainting 2.0	10.56	0.34	27.21	0.63	0.12
	SI2 + ControlNet (ours)	**8.56**	0.32	26.34	**0.69**	**0.04**
ABO	Blended Diffusion	30.13	0.36	25.70	0.75	0.25
	GLIDE	25.67	**0.19**	26.17	0.80	0.26
	Stable Diffusion 2.0	9.58	0.31	**28.45**	0.72	0.18
	Stable Inpainting 2.0	9.31	0.28	28.10	0.80	0.10
	SI2 + ControlNet (ours)	**5.93**	0.27	27.74	**0.83**	**0.04**
COCO	Blended Diffusion	30.88	0.43	23.82	0.40	0.21
	GLIDE	25.96	0.37	24.40	0.48	0.13
	Stable Diffusion 2.0	18.89	0.42	**27.51**	0.47	0.17
	Stable Inpainting 2.0	11.35	0.38	27.25	0.52	0.12
	SI2 + ControlNet (ours)	**9.38**	**0.36**	26.37	**0.57**	**0.04**
DAVIS	Blended Diffusion	29.98	0.48	22.14	0.52	0.12
	GLIDE	24.78	0.40	24.02	0.54	0.07
	Stable Diffusion 2.0	20.69	0.44	**28.14**	0.56	0.16
	Stable Inpainting 2.0	11.77	0.39	28.10	0.64	0.06
	SI2 + ControlNet (ours)	**8.70**	**0.37**	27.62	**0.69**	**0.01**
Pascal	Blended Diffusion	30.10	0.45	24.33	0.48	0.12
	GLIDE	26.95	**0.30**	24.58	0.55	0.14
	Stable Diffusion 2.0	18.83	0.40	**27.41**	0.50	0.14
	Stable Inpainting 2.0	11.26	0.36	27.30	0.56	0.10
	SI2 + ControlNet (ours)	**8.28**	0.34	26.39	**0.59**	**0.03**

Table 1. Evaluation results of text-guided salient object outpainting. Our proposed approach (SI2 + ControlNet) reduces object expansion relative to SI2 by 3.6× on average, while also surpassing SI2 on the visual metrics (FID, LPIPS).

We randomly pick 10 positive and negative points from the salient mask of the original image obtained using the SOS model InSPyReNet [19]. The positive (negative) points are inside (outside) the mask. Then, the outpainted image and point prompts are passed to the SAM model to segment the salient object and produce a mask m_o. We also obtain a salient mask of the input object-only image m_i using the same process to enable an apples-to-apples comparison between masks. Figure 4 illustrates the pipeline for obtaining these salient masks.

Given the masks m_o and m_i, a natural measure of object expansion E can be defined as:

$$E = \text{AREA}(m_o) - \text{AREA}(m_i) \tag{3}$$

with AREA expressed as a percentage of the image. Because our outpainting models never shrink the salient object, the salient mask area in the outpainted image would ideally always be larger than the salient mask area in the original image, i.e., $\text{AREA}(m_i) \leq \text{AREA}(m_o)$. However, as segmentation models are prone to error, all pixels in m_i may not be included in m_o, leading to an underestimate of the magnitude of the expansion. To account for this, we modify the outpainted mask to include m_i and instead propose the following measure for object expansion:

$$E = \text{AREA}(m_o \cup m_i) - \text{AREA}(m_i) \tag{4}$$

With this score, an upper bound exists for the object expansion based on the size of the salient object. The larger the salient object in the original image, the lower the upper bound of expansion.

4. Experiments

We use the following salient object segmentation datasets as training data, with 56k images in total: CSSD [50], ECSSD [38], DIS5k [31], DUTS [44], DUT-OMRON [51], HRSOD [52], MSRA-10k [4], MSRA-B [43], and XPIE [48]. As addressed in Section 4.2.1, training only on salient object datasets can reduce the diversity of generated backgrounds; for this reason, we also include the training partition of COCO [24], which has 118K images. The salient masks for COCO were generated using the state-of-the-art InSPyReNet [19] salient object segmentation model. To train text-guided diffusion models, we need image captions. We use ground truth captions for COCO and obtain the captions for the salient objects datasets using BLIP-2 [22]. The proposed architecture is trained on 8 NVIDIA V100 GPUs

	Model	Prompted Background		
		Empty	Likely	Unlikely
FID ↓	SI2	12.37	11.40	18.58
	Ours	**6.44**	**7.12**	**10.50**
LPIPS ↓	SI2	**0.32**	0.33	0.36
	Ours	**0.32**	**0.30**	**0.33**
CLIP Score ↑	SI2	-	**25.89**	**29.01**
	Ours	-	25.69	27.79
Obj. Sim. ↑	SI2	0.59	0.70	0.59
	Ours	**0.65**	**0.72**	**0.62**
Obj. Exp. ↓	SI2	0.117	0.104	0.102
	Ours	**0.038**	**0.041**	**0.044**

Table 2. Comparison of SI2 and our model given different types of prompts: (i) empty, (ii) a likely setting for the object, and (iii) an unlikely setting. Metrics are averaged over all evaluation datasets.

Figure 5. Controlling the strength of ControlNet using the adjustable weight w at inference time. With $w = 0.0$, objects can expand freely. Setting $w = 1.0$ aggressively prevents expansion.

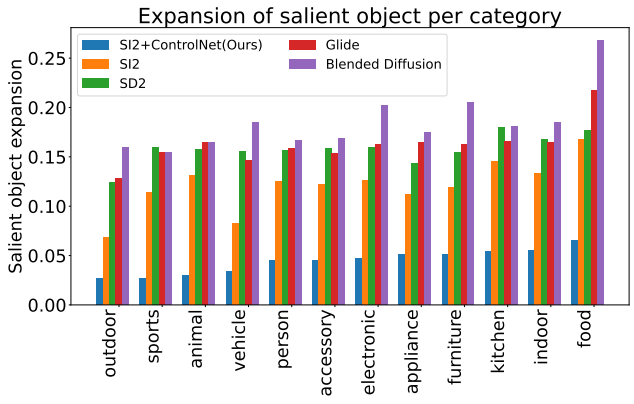

Figure 6. Comparison of salient object expansion across 12 COCO supercategories. The worst expansion scores for each model are observed in indoor settings with fine details.

for 300k iterations with the AdamW [25] optimizer using a learning rate of $5e^{-5}$ and batch size of 15 per GPU.

4.1. Experimental Procedure

We choose the state-of-the-art image inpainting methods with available code as our baselines: Blended Diffusion [1], GLIDE [29], Stable Diffusion [35], and Stable Inpainting [35]. Stable Diffusion, Stable Inpainting, and our method support image generation for 512×512 sizes, but because Blended Diffusion only supports an image size of 256×256, we resize all results to 256×256 for fair comparison. We compare these techniques using:

1. **Fréchet Inception Distance (FID)** [12] which evaluates perceptual quality by measuring the distribution distance between the synthesized images and real images. A portion of ImageNet [7] is used as the reference dataset.
2. **Perceptual Image Patch Similarity (LPIPS)** [56] which evaluates the diversity of generated backgrounds by computing the average LPIPS score between pairs of outpainted images for the same salient object image.
3. **CLIP Score** [11] which measures the alignment between the text prompt and generated images as the cosine distance between their embeddings using `CLIP-ViT-L/14`.
4. **Object Similarity** measures how much the salient object identity is *conceptually* preserved after background generation. This is computed as the cosine distance between the embeddings of the outpainted image and input object-only image using BLIP-2.
5. **Object Expansion** quantifies the degree of expansion of the salient object in pixel space, as described in Section 3.3.

We conduct evaluations on five datasets: ImageNet [7], Amazon Berkeley Objects (ABO) [5], the validation split of COCO [24], DAVIS [52], and Pascal [23]. DAVIS and PASCAL already have ground truth salient masks, but we obtain the salient object masks of ImageNet, ABO, and COCO images using InSPyReNet [19] and discard images in which the salient object occupies less than 5% of the image area.

4.2. Results

The detailed results are presented in Table 1. Our method reduces object expansion by 3.6× on average compared to the state-of-the-art SI2. The SI2 model has been trained on the LAION [37] dataset, which includes billions of web images; however, the web image data may contain non-realistic images such as collages, cartoons, images of text, etc. As we train this model on real image datasets, we obtain improved FID and LPIPS scores across standard datasets, which contain images that also tend to be more realistic.

After GLIDE, which generates the most diverse backgrounds, our model ranks second in LPIPS by a small margin. However, GLIDE generations perform poorly under FID and CLIP Score and show significant object expansion. SD2 achieves the highest alignment between text prompt and generation—as measured by CLIP Score—because the generated background is less constrained by the salient object, thus giving the model more freedom to follow the prompt accurately. Our model slightly degrades the CLIP Score of SI2, which may be attributed to the distribution of our training images (67% COCO) and reliance on BLIP-2 synthetic captions for the salient object datasets in our training corpus. Because these captions can be short and noisy, they may contribute to a decreased adherence to the input text prompt by the trained model. However, our architecture allows controlling the strength of ControlNet at inference time, using an adjustable weight ranging from 0 (no

Model	Training Dataset	Train LPIPS ↓	FID ↓	LPIPS ↓	CLIP Score ↑	Obj. Sim. ↑	Obj. Exp. ↓
SD2 + ControlNet	SODs	0.41	13.12	0.40	23.91	0.57	0.16
	SODs + COCO	0.31	11.93	0.37	24.80	0.60	0.13
SI2 + ControlNet	SODs	0.41	9.51	0.42	25.66	0.63	0.06
	SODs + COCO	**0.31**	**8.17**	**0.33**	**26.89**	**0.68**	**0.03**

Table 3. Comparison of training the ControlNet U-Net initialized with SD2 and SI2 using two training sets: (a) only the salient object datasets (SODs), and (b) SODs plus the COCO training split with segmentation-derived salient object masks. Our results demonstrate that performance measures improve significantly due to (i) initializing with the SI architecture and weights compared to SD, and (ii) adding COCO data to the training set, even without ground truth masks for salient objects.

ControlNet) to 1 (full-scale ControlNet). As demonstrated in Figure 5, using this feature, one can adjust the amount of control from the ControlNet to different desired levels.

Our approach achieves the highest Object Similarity score, demonstrating that the identity of the salient object is better preserved when expansion is explicitly controlled. A dramatic improvement is seen in the Object Expansion measure from Section 3.3, with a 3.6× decrease over SI2, which is ranked second. This improvement can be attributed to both the model architecture and the training data, which effectively address the task of salient object outpainting.

4.2.1 Ablation Studies

Role of text prompts. To study the effect of text prompts on the outpainted images, we evaluate our model and SI2 using different types of prompts in Table 2, including an empty prompt as well as prompts describing likely and unlikely settings for the salient objects. For example, a chair is likely to be found in a room but unlikely to be found in the sky. This is done by using BLIP-2 to caption the salient object image and produce a salient object caption σ, prompting OpenAI's GPT-4 [30] with: *"You are a creative and professional photo editor. Question: What is a very/least likely scene for the object described in triple parentheses to be found in? (((σ))). Answer: The object is very/least likely to be found in"* and then using the API response as the text prompt for outpainting. The results in Table 2 show that FID drops significantly for outpainted images with unlikely backgrounds, while object identity via the Object Similarity score is preserved the most in likely settings. Prompting with implausible backgrounds also leads to a slight decrease in the diversity of the generated backgrounds as shown through LPIPS, but a large increase in prompt alignment via CLIP Score. We hypothesize that when the object and the prompt are unrelated, the foreground and background become independent during the diffusion process, making them easily distinguishable under these measures. Finally, object expansion does not appear sensitive to the background's naturalness, and our proposed model reduces expansion robustly across different prompt types.

Object expansion across categories. In Figure 6, we plot salient object expansion across twelve COCO supercategories. We observe that the ordering of supercategories by the expansion score is fairly similar across the benchmarked models. The highest expansion scores for each model are seen in indoor settings, which tend to contain many fine details and salient objects with less defined dimensions, such as FOOD, KITCHEN, and FURNITURE. Similarly, the lowest expansion scores occur in outdoor scenes like SPORTS and ANIMAL where objects contrast well with the background.

Effectiveness of inpainting models. The original ControlNet architecture was proposed for controlling text-to-image models; however, we have adapted it here to work with text-guided inpainting models. As shown in Table 3, object expansion with SD2 + ControlNet (a text-to-image model) is higher than that of SI2 + ControlNet (inpainting) because inpainting models already can infill missing image regions, whereas the ControlNet needs to learn this ability from scratch for text-to-image models. SD2 (SI2)'s UNet was used in both the frozen stack and the initialization for the ControlNet encoder stack in the SD2 (SI2) + ControlNet solution.

Effectiveness of expanding the training set. We observed that the background diversity of the salient object datasets is lower than in-the-wild datasets such as COCO. This is also indicated in Table 3 by the LPIPS score for real training images. We added the training split of COCO to our training corpus to improve the diversity of our generated backgrounds. As there is no ground truth segmentation for salient objects in COCO, we generated synthetic salient masks for this data using InSPyReNet [19]. The results show that including COCO data in training, even with segmentation-derived masks, significantly improves visual and expansion metrics performance.

5. Conclusions and Future Work

In this paper, we presented an approach based on diffusion models for generating backgrounds for salient objects without altering their boundaries, as preserving the identity of objects is necessary in applications such as design and e-commerce. We identified the problem of object expansion and provided a measure to capture it. We leave generating backgrounds for non-salient objects as future work because it may require high-quality instance or panoptic segmentation masks. Additionally, future work can explore alternatives to ControlNet such as the T2I-adapter [28]—which modulates the U-Net encoder rather than the decoder—or novel combinations of control architectures for the task of object-aware background generation.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18187–18197, 2021. 3, 7

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42: 1 – 11, 2022. 3

[3] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *ArXiv*, abs/2103.10951, 2021. 3

[4] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015. 6

[5] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. Vicente, T. Dideriksen, H. Arora, M. Guillaumin, and J. Malik. ABO: Dataset and benchmarks for real-world 3d object understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21094–21104, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 7

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *ArXiv*, abs/2210.11427, 2022. 3

[7] 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, 2009. 7

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 4

[9] Dongsheng Guo, Hongzhi Liu, Haoru Zhao, Yunhao Cheng, Qingwei Song, Zhaorui Gu, Haiyong Zheng, and Bing Zheng. Spiral generative network for image extrapolation. In *Computer Vision – ECCV 2020*, pages 701–717, Cham, 2020. Springer International Publishing. 1

[10] Amir Hertz, Ron Mokady, Jay M. Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *ArXiv*, abs/2208.01626, 2022. 3

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 7

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

[13] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. 5

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 1, 3

[15] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *ArXiv*, abs/2210.02303, 2022. 3

[16] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 3

[17] Biliana Kaneva, Josef Sivic, Antonio Torralba, Shai Avidan, and William T. Freeman. Infinite images: Creating and exploring a large photorealistic virtual space. *Proceedings of the IEEE*, 98(8):1391–1407, 2010. 1

[18] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Hui-Tang Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *ArXiv*, abs/2210.09276, 2022. 3

[19] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. Revisiting image pyramid structure for high resolution salient object detection. In *Computer Vision – ACCV 2022: 16th Asian Conference on Computer Vision, Macao, China, December 4–8, 2022, Proceedings, Part VII*, page 257–273, Berlin, Heidelberg, 2023. Springer-Verlag. 5, 6, 7, 8

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5

[21] Dilip Krishnan, Piotr Teterwak, Aaron Sarna, Aaron Maschinot, Ce Liu, David Belanger, and William Freeman. Boundless: Generative adversarial networks for image extension. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10520–10529, 2019. 1

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. 6

[23] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7

[24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 6, 7

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 7

[26] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11451–11461, 2022. 1, 3

[27] 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. 3

[28] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. 8

[29] 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. In *International Conference on Machine Learning*, 2021. 1, 3, 7

[30] OpenAI. GPT-4 technical report, 2023. 8

[31] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022. 5, 6

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 3

[33] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *ArXiv*, abs/2204.06125, 2022. 1

[35] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 1, 2, 3, 7

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. 1, 3

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models, 2022. 3, 7

[38] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended CSSD. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4): 717–729, 2016. 6

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, Lille, France, 2015. PMLR. 1, 3

[40] Stability.AI. Stable diffusion 2 base model card. https://huggingface.co/stabilityai/stable-diffusion-2-base, 2022. Accessed: 2024-03-27. 3

[41] Stability.AI. Stable diffusion inpainting 2.0 model card. https://huggingface.co/stabilityai/stable-diffusion-2-inpainting, 2022. Accessed: 2024-03-27. 2, 3

[42] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor S. Lempitsky. Resolution-robust large mask inpainting with Fourier convolutions. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3172–3182, 2021. 3

[43] Jingdong Wang, Huaizu Jiang, Zejian Yuan, Ming-Ming Cheng, Xiaowei Hu, and Nanning Zheng. Salient object detection: A discriminative regional feature integration approach. *International Journal of Computer Vision*, 123(2): 251–268, 2017. 6

[44] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 6

[45] Miao Wang, Yu-Kun Lai, Yuan Liang, Ralph R. Martin, and Shi-Min Hu. BiggerPicture: Data-driven image extrapolation using graph matching. *ACM Trans. Graph.*, 33(6), 2014. 1

[46] 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. *ArXiv*, abs/2205.12952, 2022. 3

[47] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1399–1408, 2019. 1

[48] Changqun Xia, Jia Li, Xiaowu Chen, Anlin Zheng, and Yu Zhang. What is and what is not a salient object? Learning salient object detector by ensembling linear exemplar regressors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4399–4407, 2017. 6

[49] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. SmartBrush: Text and shape guided object inpainting with diffusion model. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22428–22437, 2022. 3

[50] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1162, USA, 2013. IEEE Computer Society. 6

[51] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3166–3173. IEEE, 2013. 6

[52] Yi Zeng, Pingping Zhang, Zhe Lin, Jianming Zhang, and Huchuan Lu. Towards high-resolution salient object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7233–7242, 2019. 6, 7

[53] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *Computer Vision – ECCV 2020*, Cham, 2020. Springer International Publishing. 1

[54] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. Text-guided neural image inpainting. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 3

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 4

[56] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, Los Alamitos, CA, USA, 2018. IEEE Computer Society. 7

[57] Yinda Zhang, Jianxiong Xiao, James Hays, and Ping Tan. Framebreak: Dramatic image extrapolation by guided shift-maps. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1171–1178, 2013. 1

[58] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *ArXiv*, abs/2207.06635, 2022. 3

[59] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Image inpainting with cascaded modulation GAN and object-aware training. In *Computer Vision – ECCV 2022*, pages 277–296, Cham, 2022. Springer Nature Switzerland. 1