

Differential Diffusion: Giving Each Pixel Its Strength

Eran Levin
Tel Aviv University
eran.levin@cs.tau.ac.il

Ohad Fried
Reichman University
ofried@runi.ac.il

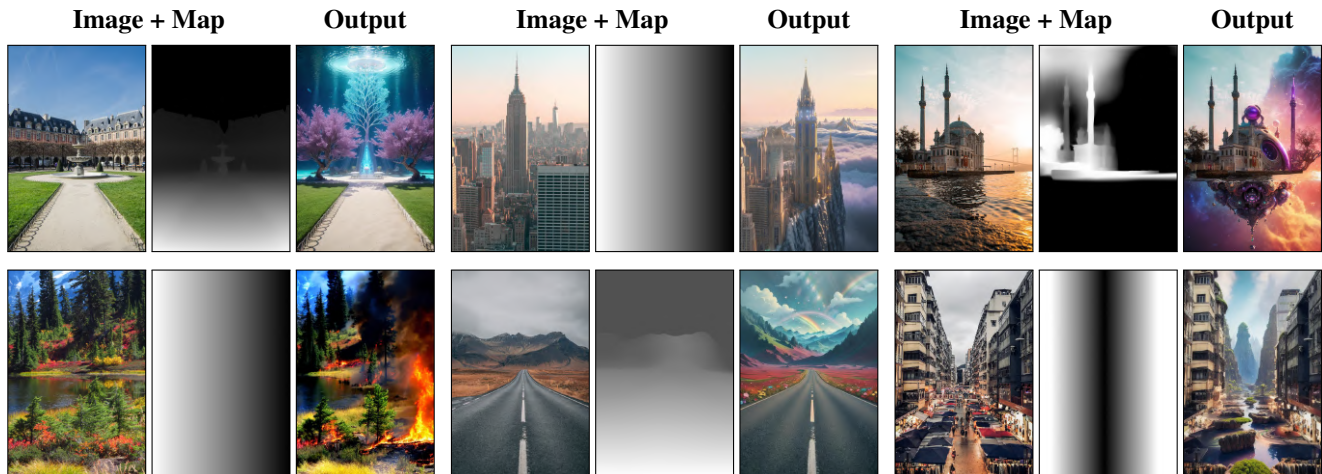


Figure 1. Our method changes different regions of an image in different amounts, according to a given map and text prompt. This controllability allows us to reproduce gradual processes (e.g., fire, bottom left) and to seamlessly blend between varying edit strengths. Edits, left to right: “tree of life under the sea”, “palace above the clouds”, “3d depth outer space nebulae background”, “fire”, “whimsical illustration of a rainbow”, “fantasy art” (full prompts in the supplemental).

Abstract

Diffusion models have revolutionized image generation and editing, producing state-of-the-art results in conditioned and unconditioned image synthesis. While current techniques enable user control over the degree of change in an image edit, the controllability is limited to global changes over an entire edited region. This paper introduces a novel framework that enables customization of the amount of change per pixel or per image region. Our framework can be integrated into any existing diffusion model, enhancing it with this capability. Such granular control on the quantity of change opens up a diverse array of new editing capabilities, such as control of the extent to which individual objects are modified, or the ability to introduce gradual spatial changes. Furthermore, we showcase the framework’s effectiveness in soft-inpainting—the completion of portions of an image while subtly adjusting the surrounding areas to ensure seamless integration. Additionally, we introduce a new tool for exploring the effects of different change quantities. Our framework operates solely during inference, requiring no model training or fine-tuning. We demonstrate

our method with the current open state-of-the-art models, and validate it via both quantitative and qualitative comparisons, and a user study. Our code will be available at: <https://github.com/exx8/differential-diffusion>.

1. Introduction

Recently, diffusion models have taken the lead in image generation [16], offering a robust method to generate and edit high quality images [46, 48]. Typical editing methods allow specifying only one change quantity (which is usually termed “strength”), changing the entire image uniformly. Inpainting methods, at most, allow partitioning the picture into an unchanged and a changed region according to a single selected strength. In this paper, we introduce a major advancement of editing with diffusion models—a framework that allows the user to change an arbitrary number of regions of the picture by different strengths efficiently and simultaneously, thus generalizing some of the edit tasks present today. This is primarily accomplished by the insight that, by selectively

modifying various regions at different timesteps during the diffusion’s inference process, we can control the fidelity to the original image on a spatial basis. Our framework does *not* require any optimization process such as fine-tuning or training. It enables greater flexibility and finer-grained editing of the image, allowing use cases that have not been attainable by any previous methods, such as editing a region with a continuous range of strength values.

Consider for example, introducing a wildfire into a wooded area of a photo. As fire is a continuous phenomenon, we would not want to make abrupt and complete transformations such as replacing all the trees with burnt stumps. Instead, we would like to introduce different amounts of fire into different regions on the photo, in a controllable manner (Figure 1 bottom-right).

We will show that even for simple edits (with as little as three regions), controlling the amount of change of each region opens up a wide range of unique edits (e.g., Figure 9). In addition, we demonstrate that we can apply soft-inpainting that exhibits superior blending and visual quality compared to conventional blending methods. Furthermore, we demonstrate our method’s capability to generate a tool we call “Strength Fan”, which enables the analysis of the effects of different strength levels for a given prompt and input image.

1.1. Contributions

We introduce several key observations (Section 3.2) that allow us to create an efficient inference process (Section 3.3). We demonstrate several applications of our method (Section 4), and evaluate it against baselines and other methods (Sections 5.1 and 5.2). Our main contributions are:

- We define a new concept termed “change map” that generalizes the “mask” concept of image editing, and create a framework that implements it (Section 3).
- We extend the framework to apply better soft-inpainting than was previously possible (Section 4.1).
- We introduce a tool to visualize the effects of different strength values (Section 4.2).
- We devise a new evaluation procedure (Section 5.2.1) and metrics (Section 5.2.2) to compare different techniques according to their adherence to a change map.

Our framework does not require training or fine-tuning, and it only adds minimal memory overhead to the inference process (Section 5.4).

2. Related Work

We first review methods that generate complete images from text (Section 2.1), and then describe editing methods using either text (Section 2.2) or masks (Section 2.3) as input.

2.1. Text-based Image Synthesis

In recent years, research on text-based image generation has become increasingly prevalent. Early models such as

DRAW [18] and alignDRAW [33], while impressive, tend to produce blurry outputs. Many papers [3, 28, 44, 49, 56, 59, 60] propose GAN-based [17] solutions, which produce higher quality output, but they often lack coherence—they sometimes miss hierarchical structures, which lead complex objects to be blurry [44]. Recently, diffusion models [20, 50] have emerged as a leading solution for image generation [16, 35], and for text-to-image synthesis [58]. PromptPaint [14], MultiDiffusion [10], and SpaText [8] offer methods to synthesize images based on user-provided prompts and shapes. Recently, a number of models that support text-based image synthesis were released [4, 38, 43, 46]. Our method is an enhancement of editing methods rather than a method for producing whole new images.

2.2. Text-based Editing

Several approaches take an image as input and use text to guide the editing process. DiffusionCLIP [23] uses domain-specific diffusion models, DDIM inversion, and fine-tuning, for image editing. “More Control for Free!” [29] suggests a framework for semantic image generation that allows guiding it simultaneously with a text and an image. InstructPix2Pix [11] enables the user to edit pictures with human-friendly instructions. Prompt-to-prompt [19] presents “real image editing” which utilizes inversion, extracting text-guided masks from attention layers. DiffEdit [15] introduces a diffusion-based text-guided photo-editing method, which extracts masks guided by a reference prompt through the comparison of noise outputs, followed by a modified DDIM denoising. Other method exists, addressing various challenges in the realm of video and photo content editing [9, 13, 22, 36, 54, 55, 64]. While the methods above yield impressive outcomes, we claim that masks offer irreplaceable controllability that surpasses textual prompts alone. This distinction is amplified by our paper, which introduces strength spatial control.

2.3. Mask-based Editing

Lately, multiple methods which support mask-based editing emerged [1, 2, 24, 31, 39, 53, 57]. Next, we describe diffusion models that support mask-based editing, followed by techniques for mask-based editing that are applicable to diffusion models in general.

2.3.1 Diffusion Models with Native Support

Palette [48] introduces a diffusion-based model that solves four tasks including inpainting—the ability to discard and re-synthesize parts of an image in a photo-realistic way. In contrast to our solution, it always discards the selected content completely and does not grant the users the capability to guide content completion through text-based instructions. Recently, Stable Diffusion [46]—a deep learning diffu-

sion model which supports inpainting and I2I, gained attention from the research community, industry, and the public, in part because the researchers chose to freely release the model’s checkpoints¹. In recent months, more diffusion models that support inpainting and I2I have been released, such as Kandinsky [43], DeepFloyd IF [4], and Stable Diffusion XL [38]. All are compatible with our method (Section 3.4). All these models offer guided inpainting, taking as input a binary edit mask, a guiding text, and a strength parameter. However, they do not allow specifying different strengths in the same transformation. All but DeepFloyd IF require special fine-tuning for the inpaint task as opposed to our solution which does not require fine-tuning or training.

2.3.2 Other Approaches

Blended Diffusion [6] allows the user to edit specific regions of a picture according to a text. Blended Latent Diffusion [7] expands this approach to latent-based diffusion models. All these solutions offer minimal control on the strength applied during the edit. They allow only manipulations that are equivalent to editing by a change map with two or fewer distinct values. SDEdit [34] is a diffusion-based method for editing images. The paper introduces the “realism-faithfulness trade-off”, which is now referred to as “strength”. It demonstrates that diffusion models can be used for editing images by starting the diffusion process at a later stage than usual with matching noise. In addition, SDEdit offers stroke-based picture editing. Unlike our method, it lacks text-guided capabilities, and the strength is uniform for the entire edited region. Our method generalizes most of the approaches described by enabling local strength control, converting it from a scalar into a matrix, while allowing text-guided editing.

3. Method

Given an image, a mono-channel change map representing the desired change amount of each pixel, and a text prompt, our goal is to edit the image to produce a high-quality result that satisfies the desired change and adheres to the prompt.

3.1. Preliminaries

Diffusion models are deep learning models that have been inspired by thermodynamics [50]. In computer vision context, they are usually trained to gradually denoise images, that have been corrupted by a random Gaussian noise. As explained in Denoising Diffusion Probabilistic Models [20], typically, the image-to-image translation process (**The Inference Process**) begins with an image with added Gaussian noise, then in an iterative process, the noise is gradually

¹We hope that other researchers (and companies) will choose to do the same. Our code is available at <https://github.com/exx8/differential-diffusion>.

Algorithm 1 Differential Image to Image Diffusion

Input x (image to edit), k (number of steps), μ (change map with values between 0 and 1), p (prompt)

Output \hat{x}

```

1: procedure INFERENCE( $x, k, \mu, p$ )
2:    $z_{init} = \text{ldm\_encode}(x)$ 
3:    $\mu_s = \text{down\_sample}(\mu)$ 
4:    $z'_k = \text{add\_noise}(z_{init}, k)$ 
5:    $z_k = \text{denoise}(z'_k, p, k)$ 
6:   for  $t = k - 1$  to 0 do
7:      $z'_t = \text{add\_noise}(z_{init}, t)$ 
8:      $mask = \mu_s \odot \frac{k-t}{k}$ 
9:      $z_t^{mix} = z_{t+1} \odot mask + z'_t \odot (1 - mask)$ 
10:     $z_t = \text{denoise}(z_t^{mix}, p, t)$ 
11:   end for
12:    $\hat{x} = \text{ldm\_decode}(z_0)$ 
13:   return  $\hat{x}$ 
14: end procedure

```

We denote \odot , \ominus as element-wise less-than and element-wise multiplication, respectively. \odot returns a tensor of 1s and 0s.

removed. This inference process creates a series of images (**Intermediate Images**), where each is the result of the denoising operation of the previous one (**The Inference Chain**). The **Prompt** is a text input that controls the content of the generated picture. The **Strength** is a quantity that determines the amount of change applied by the edit. For this paper, a **Mask** is defined as a map comprised of only two distinct values, as used in traditional inpainting and localized editing. Extending the concepts of strengths and masks, in this paper we introduce the **Change Map**, a matrix of the same dimensions as the original input image, describing the strength of the edit to be applied at each location.

In Latent Diffusion Models [45], the diffusion process occurs in a dimensionally reduced latent space, which the original image is translated to by a **Latent Encoder** at the beginning of the inference process. Therefore, in these models, the intermediate images are represented as their latent version. At the end of the process, a **Latent Decoder** translates the latent output into an image.

In this paper, we present *Differential Diffusion* — an inference-time enhancement of image-to-image diffusion models that adds the ability to control the amount of change applied to each image region according to a change map. Our method decomposes the map into a series of nested masks that are applied iteratively, such that each region begins the inference at a different timestep according to the masks.

3.2. Observations

Our method is based on the following observations:

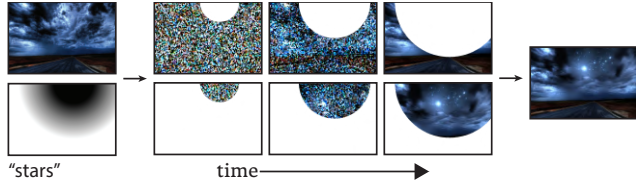


Figure 2. **Breakdown of Algorithm 1 Line 9 over time.** Top: $z'_t \odot (1 - \text{mask})$, the regions copied from a noised version of the input. Bottom: $z_{t+1} \odot \text{mask}$, the residue regions copied from the previous U-Net output. Observe how the change map determines the inference process—the darker the region, the earlier it is copied from the residue.

1. **The Suffix Principle:** For every complete image-to-image inference chain Σ , every suffix σ is also an inference chain. Let N, n be the number of timesteps of Σ, σ , the noise levels in σ 's intermediate images match those of an inference chain, with a strength of $\frac{n}{N}$.
2. **Overridability:** Regions in the intermediate images can be overridden by external content with the same distribution, and influence the generated image without breaking the inference process.
3. **Locality:** All the latent encoders that were examined [38, 43, 46] generally encode pixels to the same relative positions, which is the ratio between their positions to the dimensions of the picture.

3.3. Algorithm

In our framework, we change the inference algorithm of the diffusion process (Algorithm 1). First, the original image is encoded to the latent space (z_{init}), and the map is downsampled to the latent space spatial dimensions. Due to *locality*, the down-sampled map (μ_s) aligns with the positions of the latent pixels in the latent tensor. The denoising loop is changed for each time step t as follows. We first noise the encoded original image according to the current timestep (z'_t). Then, we calculate a mask of all points which are lower than the threshold $\frac{k-t}{k}$ in the change map. $\frac{k-t}{k}$ is the value of the mask complement of the strength. Therefore, the values of the mask determine the last timestep where each region is overridden, controlling the amount of change of the region, due to *the suffix principle*.

The masks are nested (Figure 2), therefore some regions are copied from the noised original image multiple times, in contrast to copying each region once according to its strength. Because the diffusion model has not been trained on intermediate images with holes, this mimics the distribution which it has been trained on, and gives it an advance knowledge of the content of lighter regions. We conducted an experiment that affirms these advantages in Figure 3—we changed Line 8 in Algorithm 1 to $\text{mask} = \frac{k-t}{k} \odot \mu_s \odot \frac{k-(t+1)}{k}$, which means for each timestep, only the regions that match the exact timestep were copied. Next, we copy all the se-

lected regions in the mask from the previous timestep. The rest is copied from the noised version of the original image (Figure 4). This is possible due to *Overridability*. Finally, the U-Net denoises the result (z_t^{mix}). After the loop, z_0 is decoded to the pixel space, yielding the result (\hat{x}).

3.3.1 Optimization: Skipping

We observed that the inference process can be optimized for change maps devoid of small values. In this scenario, we can significantly enhance efficiency by skipping steps related to these small values and avoiding their calculation, as their values will be overridden by the injections of subsequent steps. It can be implemented by adding before Line 4 in Algorithm 1: $L = \lfloor (1 - \min(\mu_s)) \cdot k \rfloor$ and replace the references for k to L in Lines 4 to 6 only. Refer to the supplementary materials for the explicit algorithm.

3.3.2 Technical Details

Model: We have used the checkpoint 512-base-ema.ckpt of Stable Diffusion 2.1, unless otherwise stated.

Prompts: For most experiments, we used a simple description of the edit. For others, we found it beneficial to expand the prompt by taking the input image, reversing it into a prompt via a Clip interrogator that uses both BLIP [26] and CLIP [40], and then adding the desired edit to the prompt.

Change Maps: Our method does not assume anything about the source of the change maps. We show results using various sources: (1) Segment-Anything [25] and Language-Segment-Anything, mainly for discrete change maps. (2) MiDaS [42] mainly for continuous change maps; from a single depth map, many change maps can be created by simple histogram transformations. (3) Manually drawn change maps. **Other settings:** Unless stated otherwise, we used 100 inference steps for each experiment.

3.4. Extension For Different Diffusion Models

We generalize our framework for Stable Diffusion XL [38], Kandinsky [43], and DeepFloyd IF [4]. See Figure 5 and the supplementary materials for more information.

4. Applications

Our method is the first to allow users to edit images with an arbitrary number of strengths (Figure 1). This introduces fine grained control over the impact of the edit prompts on different image regions (Figure 9). It also allows continuous edit transitions (e.g., Figures 3, 5 and 10), a feature that has been found to be visually appealing in our user study (Section 5.3). Besides our main application of guiding an edit with a change map, our method also supports soft-inpainting, and allows for a novel visualization tool we call a strength fan, as we detail next.



Figure 3. **Ablation of nested masks.** Our result is more complex, blends better with the scene, and less blurry. Note the difference in transitions (1st row: the sharp transition in the wall) and placements (2nd row: the building is inside the lake). The seed is fixed for each row. Prompts: “a fine art painting”, “a city skyline...”.

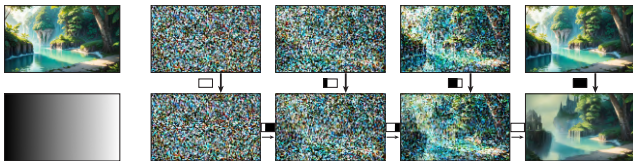


Figure 4. **illustration to the inference process.** Top: z_t^i - the original image noised to the current timestep, bottom: the intermediate images that the diffusion model denoises. The masks near the arrow represent the regions that were copied from each picture. Follow the arrows to discern the influence of the origins on the output image, and observe the correlation with the decomposed masks and the change map. The prompt is “Gothic painting”.

4.1. Soft-Inpainting

Inpainting with no softening can sometimes appear unnaturally superimposed due to noticeable differences in style and lighting compared to the surrounding background. “Soft-Inpainting” is the process of completing parts of a picture while gently modifying the surrounding regions to guarantee smooth blending. We extend our framework to support Soft-Inpainting, by allowing the user to input a binary *mask* and a scalar value to determine transition softness. We Gaussian blur the mask according to the softness parameter, turning it into a *change map*, that is then processed by our standard framework. Our method achieves superior soft-inpaint editing compared to previous methods (Figure 6).

4.2. Strength Fan

When editing images with diffusion models, finding the perfect balance between preserving certain elements and altering others can be challenging. Often, choosing the right strength for an edit can be non-intuitive, especially as the optimal setting varies across input prompts and images. To simplify this process, we propose a new visualization tool called “Strength Fan”. This fan is a modified image created by dividing it into columns, with each column undergoing editing at a different strength level. This allows users to observe multiple strength settings simultaneously, thereby simplifying the task of comparing and tuning edit strengths (Figure 7). Our framework is uniquely suited to produce

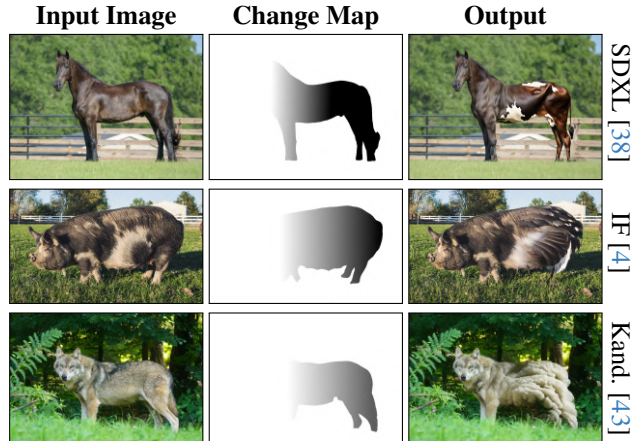


Figure 5. **Our method with different diffusion models.** We applied our framework to several diffusion models: SDXL [38], DeepFloyd IF [4], and Kandinsky [43], demonstrating its generality. Prompts: “cow”, “feathers”, “sheepskin”.

strength fans, as we only need to invoke our inference chain *once*, with a change map of rectangles corresponding to the various strengths under examination.

5. Evaluation

We compare our method to baselines (Section 5.1) and to other methods (Section 5.2), evaluate our method via a user study (Section 5.3), and report memory consumption and running time (Section 5.4).

5.1. Comparison to Baselines

We compare our inference algorithm to four baseline alternatives, founded on different ideas. Visual results are in Figure 8.

We start with several iterative alternatives. **Composition:** Since most existing diffusion methods accept a *single* strength parameter (e.g., SDEdit [34]), a straightforward method to support a varying change map is to decompose the map into K different masks (we use $K = 100$), each with a single value, and applying them iteratively. The image degrades rapidly (after applying five masks), leaving a meaningless image at the end. The deterioration is mostly caused by the recurrent transition via the latent encoder of Stable-Diffusion. **Tiling:** We try to circumvent the previous limitation by avoiding the cumulative degradation. As before, we decompose the change map into 100 masks, and then apply a series of inpaints of the masks successively. But instead of using all the pixels of the output image, after each inpainting we copy the pixels that lie outside the current mask from the previous step (for the first step we copy them from the original image). This guarantees that each region will only traverse once through the latent space. Here, the model has difficulties inpainting narrow masks, which is

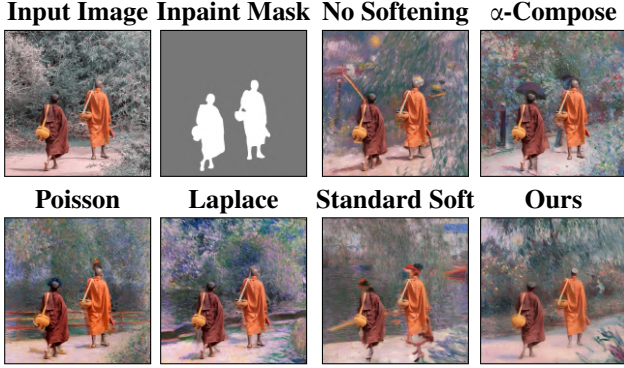


Figure 6. **Soft-inpainting.** We compare our approach to no softening, α -compositing, Poisson-based [37] and Laplace-based [12] compositing, and standard soft-inpainting (as implemented in Stable Diffusion web UI [5]). For α -compositing, Poisson-based and Laplace-based methods, we blend the original image with a regular inpaint result using a Gaussian blurred version of the inpaint mask. In all other methods, artifacts appear in the transition area, and the unchanged region looks pasted. For standard softening, even the inner parts of the figures are corrupted. Our method produces a more natural blend. Prompt: “Impressionist”.

inherent in tiling and every process that decomposes the map into fine-grained masks, leading to pixels being replaced with gray stripes. **Five Tiles:** We attempted to overcome the limitations of tiling by increasing the size of the tiles. In this alternative, we grouped the map values to five distinct bins. We then apply tiling as before, this time with only 5 masks. This method generates reasonable visual content by avoiding the use of narrow mask segments. However, the method usually fails to semantically change the regions except for the darkest tile, thus not fulfilling the desired task. All iterative alternatives exhibit inefficient running times, sometimes tens of times longer than our method, depending on the number of applied masks.

Masked Noise: It can be hypothesized that spatially adjusting the added noise’s magnitude can produce similar results to our method. We show it cannot. In this alternative, we multiplied the added noise by the change map. The model soon converges to a random single-color image. This can be anticipated, as the model was originally trained to handle a specific noise distribution, which this technique alters.

5.2. Comparison to Other Methods

We demonstrate the effectiveness of our change maps over text-guided-only methods: InstructPix2Pix [11] and DiffEdit [15] in Figure 9, and over mask-based methods: Stable Diffusion 2’s Text-Guided Inpainting [52], and Blended Latent Diffusion [7] in Figure 10.

For quantitative comparison, we begin by demonstrating how to measure edit strength spatially given input-output pairs (Section 5.2.1). We then use this technique to establish metrics for assessing change map adherence (Section 5.2.2), and use them to compare our method to Stable Diffusion 2



Figure 7. **Strength Fans.** Our Strength Fans allow users to visually compare and tune edit strengths. Prompts: “post-apocalyptic”, “Cubist painting”. Extended version in the supplemental materials.

Inpaint and Blended Latent Diffusion (Section 5.2.3).

5.2.1 Edit Strength Measurement

We propose the following procedure for spatially measuring edit strength, providing a means to evaluate the applied map during an edit. Given a change map M , and an input-output image pair, we first compute the LPIPS [63] perceptual similarity map [62] between input and output. For robustness, we repeat this procedure with 1,000 different input-output pairs and average the results. This process is referred to as “biased measurement”. For every change map M , the edit strength measurement map (E_M) for a method is defined to be the biased measurement of M subtracted by the biased measurement of a black map (full change), which eliminates spatial bias in the perceptual similarity map.

5.2.2 Adherence Metrics: CAM and DAM

Our objective is to measure the extent to which each method adheres to change maps and to establish metrics for assessing adherence. As far as we know, we are the first to quantify change map adherence. We suggest two metrics: Correlation Adherence Metric (CAM) and Distance Adherence Metric (DAM). In general, CAM tends to focus on high-level features; for example, CAM will usually assign a poor similarity score when comparing maps of different shapes. DAM, on the other hand, focuses on lower-level features; for instance, DAM will usually assign a lower similarity score to maps that differ by regions in which their brightness has been changed. As a result, each metric offers a distinctive view of adherence quality. Refer to the supplementary material for an example of their differences.

Let M be a change map. CAM is defined as: $CAM(M, E_M) = \rho(M, E_M)$, where ρ is the Pearson correlation coefficient calculated element-wise. DAM is defined as: $DAM(M, E_M) = \min_{(a,b) \in \mathbb{R}^2} \|M - aE_M + b\|_F$, where $\|\cdot\|_F$ is the Frobenius norm. We introduced the parameters a and b because LPIPS has a magnitude and additive bias depending on the change map and method used. Pearson correlation coefficients are invariant under positive multiplication and addition.

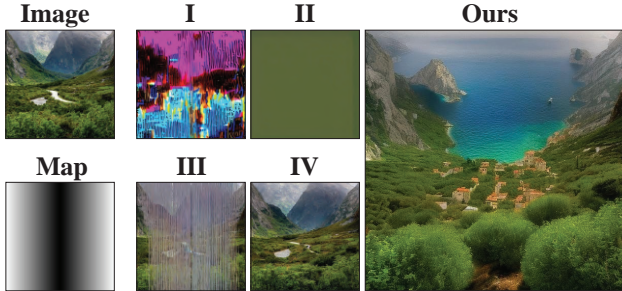


Figure 8. **Comparison to baselines.** All baselines fail, even for this simple unstructured scene and a smooth map. “Composition” (I) and “Masked Noise” (II) both fail to create a meaningful image, while “Tiling” (III) does not produce an edit related to the prompt and corrupts the image. “Five Tiles” (IV) stands out as the most successful among the baselines. Albeit, the edit is primarily noticeable in the darkest tile. Prompt: “Mediterranean Sea”.

5.2.3 Experimental Setup and Results

We sample 1,000 images randomly from ImageNet’s validation split [47], and use them to measure 3 different maps, with an empty prompt. As we are the first method to allow change maps with an arbitrary number of strengths, other methods support only two strengths. For a fair comparison, we tested other methods using the best performing binarized version of the maps (by sweeping over threshold values and reporting the best result). Our results produce the most accurate maps (Figure 11), and our method obtains the best numerical scores (Table 1).

5.3. User Study

We conducted a user study that included STEM students and volunteers from social media platforms to assess the usability of our method. Each participant received 30 questions at random from a pool of 69 questions. Each question was answered on average by 8 participants. The participants were able to skip questions if they were undecided. The study had three parts: **1. Image-to-Map Matching:** Participants identified the applied map out of three options for ten input-output image pairs. This part tests the ability of untrained users to estimate the impact of different change maps on the output and their intuitiveness. **2. Method Comparison:** Participants ranked three methods (our method, blended-latent-diffusion, Stable-Diffusion’s inpaint) based on adherence to change maps and visual quality for given input images and maps. **3. Text-Guided Editing Assessment:** Participants chose between two images, one edited with text guidance and the other without, for triplets consisting of input images, change maps, and prompts. This part tests whether the changes which are applied by our framework adhere to the textual prompt. Throughout all stages of the study, both the images and prompts were selected at random from the InstructPix2Pix [11] dataset. Images smaller than

Metric	Method	Pattern		
		Gradient	Shapes	Triangles
CAM (↑)	Ours	0.97	0.81	0.93
	BLD [7]	0.92	0.68	0.83
	SD Inpaint [52]	0.93	0.65	0.82
DAM (↓)	Ours	19.41	52.2	35.75
	BLD [7]	29.05	65.65	53.95
	SD Inpaint [52]	28.69	67.84	54.44

Table 1. Comparison between different methods’ performance across different patterns by utilizing CAM and DAM metrics.

256 × 256 pixels were excluded, as they are unsuitable for all the techniques under examination. For the second and third parts, the change maps were automatically created using MiDaS [42]. For evaluation, we determined results based on the majority vote. As shown in Table 2, users mostly chose the applied change maps, and also chose the text-guided images on most of the edits. The user mostly preferred our framework over alternative approaches, as shown in Table 3. Examples of the questions can be found in the supplementary materials.

Criteria	Match	P-Value
Map Matching	80.43%	1.31×10^{-5}
Text-Guidance	92.11%	3.64×10^{-4}

Table 2. Users are able to identify which change map was used for a given edit, and to determine that a text-guided edit is indeed closer to the text than a non-guided edit. This shows that the use of text-guided change maps leads to perceptible results.

5.4. Memory Consumption & Inference Time

We measured the inference memory consumption of Stable Diffusion’s img2img [45] with and without our framework. The overhead of using our framework is less than 3MB (0.07%). In addition, we measured our model with and without skipping. For high-fidelity maps, skipping can reduce 89% of the inference time. See the supplemental materials for more details.

6. Limitations and Future Work

One limitation of our method is in users’ ability to anticipate the result of a given edit. Some users already struggle to predict the effects of strengths in existing diffusion models, and our method further complicates the input-output relation by allowing users to select multiple values for each map. While this is far from solved, we alleviate the issue by introducing our strength fans.

In the future, our algorithm can be further optimized by calculating all z_i s and mask decompositions in advance and

Criteria	BLD [7]	SD2.1 [52]	Ours	P-Value
Map adherence	10%	32%	58%	0.0164
Visual Quality	10%	35%	55%	0.037

Table 3. The preferred method by users according to map adherence and visual quality. The results are statistically significant.

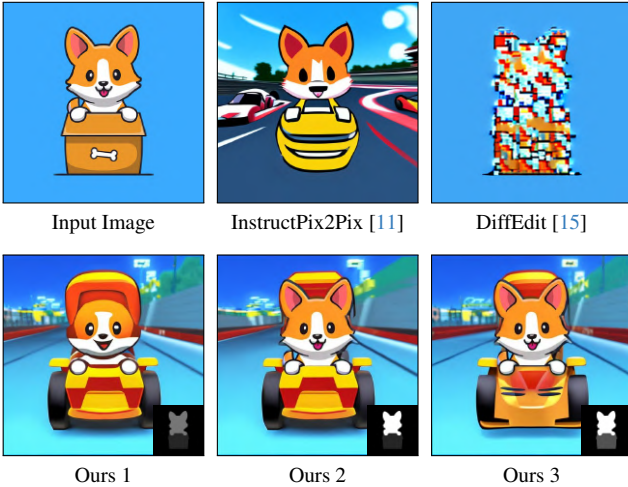


Figure 9. **Comparison to InstructPix2Pix [11] and DiffEdit [15].** The prompts are “[in a] race car video game”, and the same seed. Our change map controllability allows the user to experiment with a variety of different edits. The user can specify how faithful the dog will be to the original image, and how much the box and background will change. DiffEdit’s reference prompt is “dog”; the technique mistakenly include the surrounding box and struggle to generate a high-quality result. InstructPix2Pix changes the entire image uniformly. The supplementary materials contain more examples. Change maps in bottom-right inset.

in parallel, as they are independent of the diffusion process. Another research direction is on different methods to automatically produce change maps. In this paper, the user creates the change map, or uses depth or segmentation maps. We believe that other algorithms for map generation can extend our framework to solve various editing tasks. Lastly, we have shown Strength Fan as a tool for exploring the editing space. We believe that many other tools can be developed by using different map patterns and with automatically generated change maps.

7. Conclusion

In this paper, we presented a new framework that enables region-wise control over the strength of the image-to-image translation process. Our framework does not require optimization or training, has minimal overhead, and can be used with diverse images and prompts. Our work expands the scalar strength parameter into a more flexible 2D array, and we hope other researchers will advance this further.



Figure 10. **Comparison to Blended Latent Diffusion [7] and Stable Diffusion 2.1’s Inpaint [52].** 1st row: “glasses with ice cubes”, all other methods fail to blend the ice with the juice, as such a blend should be gradual. 2nd row: “Photograph titled Villa Garden by the photographer Bi Wei Liang Tronolone”. Because masks contain only two distinct values, it is not possible to differ the rear, the front, and the house. Therefore, the house is absent in the other methods. 3rd row: “Night court mountains, watercolor painting”, to achieve the watercolor style, a high strength is required for the skies. For other methods this destroys the structure of the mountain completely, and the intended similarity between the edited picture to the original is lost. Change maps are presented in the bottom-left inset.

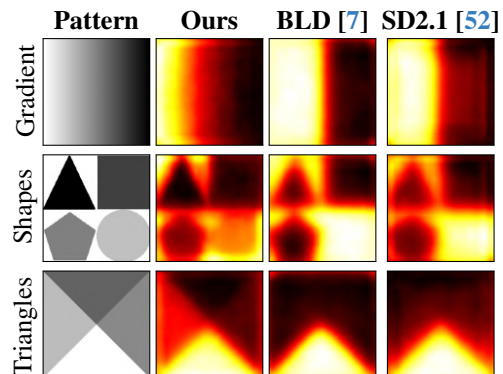


Figure 11. **Edit strength measurement maps.** We evaluate our method, Blended Latent Diffusion [7], and Stable Diffusion 2.1’s inpaint [52]. Comparing the measured map to the original, we see that the other methods are able to express only two strengths, even when the original change map contains more.

Acknowledgments

We thank Almog Friedlander, Yuval Ran-Milo, Shahr Sarfaty and Eyal Michaeli for proofreading the paper. This work was supported in part by the Israel Science Foundation (grant No. 1574/21), and by the Tel Aviv University Center for AI and Data Science (TAD).

References

- [1] Johannes Ackermann and Minjun Li. High-resolution image editing via multi-stage blended diffusion, 2022. 2
- [2] Alex Andonian, Sabrina Osmany, Audrey Cui, YeonHwan Park, Ali Jahanian, Antonio Torralba, and David Bau. Paint by word, 2023. 2
- [3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223. PMLR, 2017. 2
- [4] DeepFloyd Lab at StabilityAI. DeepFloyd IF: a novel state-of-the-art open-source text-to-image model with a high degree of photorealism and language understanding. <https://www.deepfloyd.ai/deepfloyd-if>, 2023. Retrieved on 2023-11-08. 2, 3, 4, 5, 13
- [5] AUTOMATIC1111. Stable Diffusion Web UI. <https://github.com/AUTOMATIC1111/stable-diffusion-webui>, 2022. 6, 14
- [6] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 3
- [7] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, 42(4):1–11, 2023. 3, 6, 7, 8, 14
- [8] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Sptext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18370–18380, 2023. 2
- [9] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing, 2022. 2
- [10] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 2
- [11] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 2, 6, 7, 8, 13
- [12] Peter J. Burt and Edward H. Adelson. A multiresolution spline with application to image mosaics. *ACM Trans. Graph.*, 2(4): 217–236, 1983. 6, 14
- [13] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion, 2023. 2
- [14] John Joon Young Chung and Eytan Adar. Promptpaint: Steering text-to-image generation through paint medium-like interactions. *arXiv preprint arXiv:2308.05184*, 2023. 2
- [15] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022. 2, 6, 8
- [16] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 1, 2
- [17] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2
- [18] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation, 2015. 2
- [19] 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
- [20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2, 3, 19
- [21] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. 19
- [22] 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, 2023. 2
- [23] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation, 2022. 2
- [24] Soo Ye Kim, Kfir Aberman, Nori Kanazawa, Rahul Garg, Neal Wadhwa, Huiwen Chang, Nikhil Karnad, Munchurl Kim, and Orly Liba. Zoom-to-inpaint: Image inpainting with high-frequency details. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 477–487, 2022. 2
- [25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 4
- [26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 4
- [27] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds, 2022. 19
- [28] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+gan space optimization, 2021. 2
- [29] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance, 2022. 2
- [30] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps, 2022. 19
- [31] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, 2022. 2

- [32] mage. Mage space: An ai image generation platform, 2023. 40
- [33] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention, 2016. 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, 2022. 3, 5
- [35] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 2
- [36] 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, 2022. 2
- [37] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318, 2003. 6, 14
- [38] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis, 2023. 2, 3, 4, 5
- [39] Weize Quan, Ruisong Zhang, Yong Zhang, Zhifeng Li, Jue Wang, and Dong-Ming Yan. Image inpainting with local and global refinement. *IEEE Transactions on Image Processing*, 31:2405–2420, 2022. 2
- [40] 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, 2021. 4
- [41] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. 15
- [42] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 4, 7, 13, 15
- [43] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion, 2023. 2, 3, 4, 5
- [44] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis, 2016. 2
- [45] 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. 3, 7
- [46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 2, 4, 12
- [47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 7
- [48] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2022. 1, 2
- [49] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis, 2023. 2
- [50] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. 2, 3
- [51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 19
- [52] Stability AI Ltd. Stability.ai blog: Stable diffusion v2 release. <https://stability.ai/blog/stable-diffusion-v2-release>, 2023. Accessed: April 25, 2023. 6, 7, 8, 14
- [53] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2149–2159, 2022. 2
- [54] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions, 2023. 2
- [55] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model, 2022. 2
- [56] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks, 2017. 2
- [57] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting, 2023. 2
- [58] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey, 2023. 2
- [59] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017. 2
- [60] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation, 2022. 2
- [61] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator, 2023. 19
- [62] Richard Zhang. The unreasonable effectiveness of deep features as a perceptual metric github repository, 2018. 6
- [63] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 6, 12

[64] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models, 2022. [2](#)

Differential Diffusion: Giving Each Pixel Its Strength

Supplementary Material

We present additional results (Figures 12 to 14, 16, 18, 19, 21, 24 and 27), extend the discussion on experiments from the main paper (Figures 22 and 23), provide additional experiments that show the usability of our framework (Figure 26), and specify additional technical details for reproducibility (Section 13).

We provide further details on the “skipping” optimization in Section 8. We provide an extended discussion on strength fans in Section 9. We provide additional comparisons for soft-inpainting in Figure 12. In figure 10 in the main paper we show a comparison between our method and other mask-based methods. In Figure 13 we show that these results have not been caused by outlier seeds. We show that for different seeds, our method produces better results. We show more examples of the control that change maps introduce in Figures 14 to 16.

In Section 5.1 of the main paper we discuss the limitations of traversing multiple times across the latent space. We show the degradation in the image quality in Figure 17. We provide additional comparisons to the baselines described in this section in Figure 18. In Section 5.2 of the main paper we use 1,000 pairs of images to measure the applied change map. We show an ablation to this amplification in Figure 19, where we compare it to a single pair measurement, and also show the spatial bias of the LPIPS similarity maps [63].

We show that our algorithm is compatible with many samplers in Figure 20. We show more examples of edits in Figure 21. In the main paper, we have outlined the structure of the user study; in Figure 22 we provide representative questions from the study.

8. Running Time of Skipping

We measured our model with and without skipping, using gradient maps with different minimum values. We average results over 100 different inference processes. In the main paper, we detail the modification of the original algorithm to integrate the skipping optimization. We write the explicit skipping algorithm in Algorithm 2, and we visualize the runtime as a function of the minimum value in the change map in Figure 23. A clear linear correlation exists between these two variables. Our measurements (Table 4) demonstrate the utility of the skipping optimization, by saving up to 89 percent of the running time.

9. Strength Fan

The main paper demonstrates the ability of our framework to create a “strength fan”, a tool to estimate the effects of different strength values for given prompts and seeds. We

believe that with the adoption of the tool, tuning the strength parameter when using diffusion models will become more prevalent. We envision that this tool will be integrated into existing software solutions, where users will specify the number of strengths they wish to explore simultaneously. We show several strength fans in Figure 24. Some users may opt to iteratively tune the strength by using fans with different magnitudes and offsets, as we demonstrate in Figure 25.

10. CAM versus DAM properties

In the main paper, CAM and DAM consistently ranked methods in the same order, leading to the tempting assumption that the metrics give similar information. However, this assumption is incorrect. In Figure 26 we present a counterexample: comparing map 1 (the circle) to the two other maps using CAM and DAM. As observed, while CAM indicates a stronger similarity to the halved circle, DAM indicates a higher similarity with the inverted circle. In the case of the inverted map, CAM yields the minimal similarity score possible, while DAM provides the maximal similarity score. The differences between the metrics are also notable when comparing the triangle map to the shaded circle one. By just changing the shape, the CAM drops to almost 0, while DAM retains a similar score.

11. Extensions For Various Diffusion Models

In the main paper, we mostly present our framework applied to Stable Diffusion 2.1 [46]. We generalized our framework to other diffusion models.

11.1. Stable Diffusion XL

The latest version of Stable Diffusion, XL, introduces several modifications, most of which are unrelated to our method. However, it offers a mode of ensemble of experts for inferring high-quality images. In this ensemble, a “base model” is responsible for handling higher timesteps, while a “refiner model” takes care of the rest. The split ratio s is a user-defined parameter. Therefore, our inference process is similarly split. In our algorithm, for t smaller than sk , the denoising (in line 10) is performed by the base model. Otherwise, denoising is conducted by the refiner.

11.2. Kandinsky

Kandinsky, despite its differences from Stable Diffusion’s models, seamlessly fits into our algorithm without requiring any special adaptations.

Algorithm 2 Differential Image to Image Diffusion With Skipping

Input x (image to edit), k (number of steps), μ (change map with values between 0 and 1), p (prompt)

Output \hat{x}

```
1: procedure INFERENCE( $x, k, \mu, p$ )
2:    $z_{init} = \text{l dm\_encode}(x)$ 
3:    $\mu_s = \text{down\_sample}(\mu)$ 
4:    $L = \lfloor (1 - \min(\mu_s)) \cdot k \rfloor$ 
5:    $z'_L = \text{add\_noise}(z_{init}, L)$ 
6:    $z_L = \text{denoise}(z'_L, p, L)$ 
7:   for  $t = L - 1$  to 0 do
8:      $z'_t = \text{add\_noise}(z_{init}, t)$ 
9:      $mask = \mu_s \odot \frac{k-t}{k}$ 
10:     $z_t^{mix} = z_{t+1} \odot mask + z'_t \odot (1 - mask)$ 
11:     $z_t = \text{denoise}(z_t^{mix}, p, t)$ 
12:   end for
13:    $\hat{x} = \text{l dm\_decode}(z_0)$ 
14:   return  $\hat{x}$ 
15: end procedure
```

We denote \odot , \ominus as element-wise less-than and element-wise multiplication, respectively. \odot returns a tensor of 1s and 0s.

11.3. DeepFloyd IF

We adapt DeepFloyd’s [4] cascading paradigm, by applying our algorithm for the first two (of three) stages of inference. We suggest preserving the original inference process for the last stage, as this stage primarily handles super-resolution.

12. Automatic Edits Creation

In the main paper we mostly assume that the change map inputs to the framework are handily chosen and crafted by the users. In this section, we show that some procedures can be employed to automate those choices. In Section 12.1.1 we use the input images and the prompts from the Instruct-Pix2Pix dataset [11]. For change maps, we first explore three simple fixed change maps. Surprisingly, even simple patterns seem to produce interesting results. We then explore depth maps generated by MiDaS [42] with no further intervention in Section 12.1.2. The continuous nature of these change maps produces successful and unique edits, despite being chosen arbitrarily.

Min Value	Running Time (sec)	Delta (sec)	Delta (%)
no skipping	8.0	—	—
0.1	7.14	-0.86	-10.74%
0.2	6.36	-1.64	-20.55%
0.3	5.61	-2.39	-29.90%
0.4	4.81	-3.19	-39.87%
0.5	4.02	-3.98	-49.70%
0.6	3.24	-4.76	-59.51%
0.7	2.45	-5.55	-69.37%
0.8	1.67	-6.33	-79.17%
0.9	0.88	-7.12	-88.99%

Table 4. The averaged running time as a function of the minimum value of the map, with delta compared to no skipping. As can be seen, skipping can boost the algorithm’s performance dramatically.

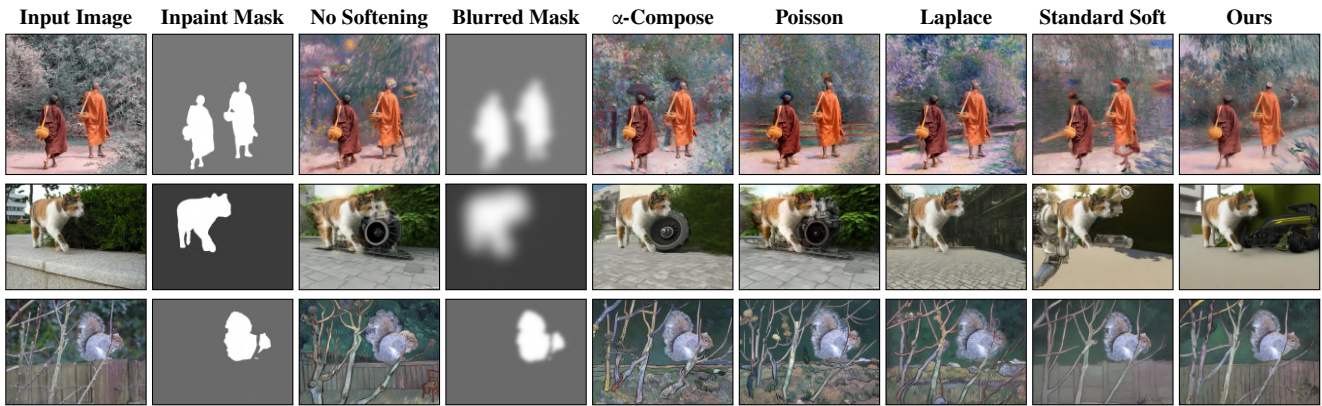


Figure 12. Extended version of Figure 6 from the main paper. We compare our approach to no softening, α -compositing, Poisson-based [37] and Laplace-based [12] compositing, and standard soft-inpainting (as implemented in Stable Diffusion web UI [5]). For α -compositing, Poisson-based and Laplace-based methods, we blend the original image with a regular inpaint result using a Gaussian blurred version of the inpaint mask. Our method produces a much more natural blend. The prompts are “Impressionist”, “Real Engine”, “Van Gogh”. The blurring radii are: 64px, 40px, 10px.

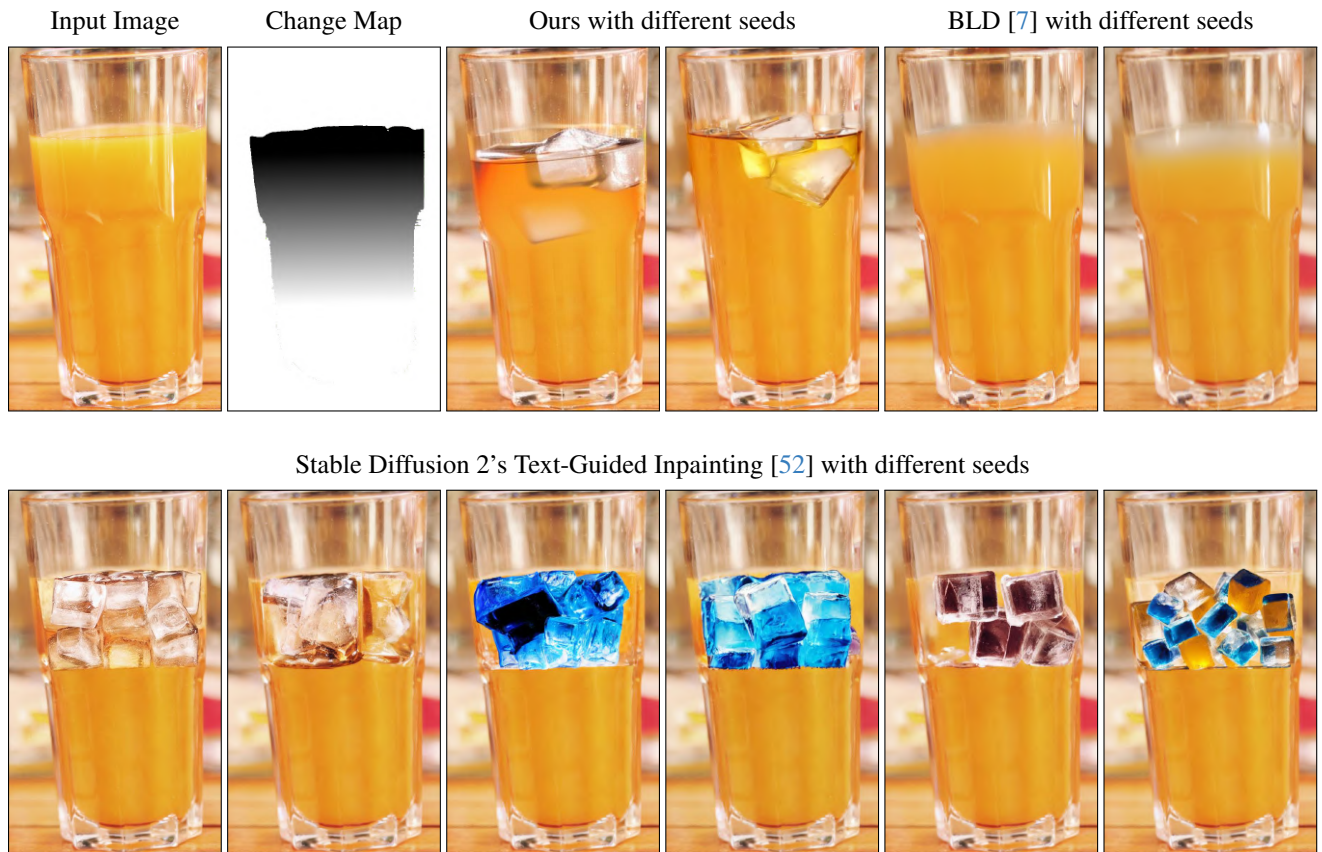


Figure 13. **Comparison to other methods with different seeds.** The prompt for all results is “glass with ice cubes”. Blended Latent Diffusion fails to generate ice-cubes, and Stable Diffusion cannot properly blend between the added cubes and the juice.

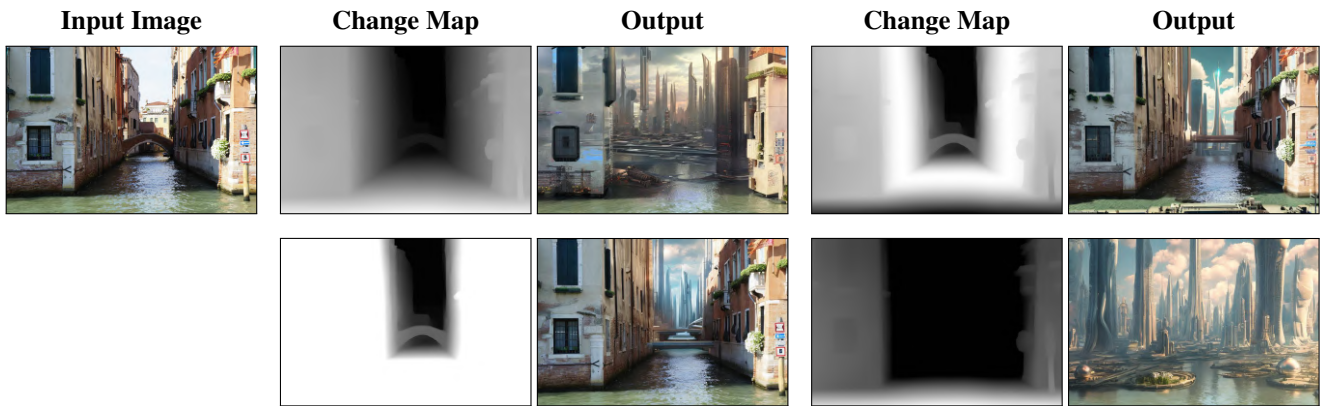
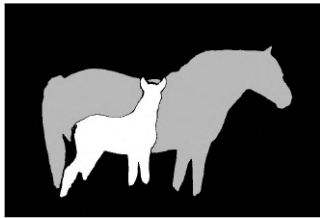


Figure 14. **Various change maps created from the same depth map.** As described in the main paper, a MiDaS [41, 42] depth map can be used to create many different change maps via simple image transformations. Prompt: “a futuristic city with tall buildings and a lot of traffic in the foreground and a cloudy sky in the background, a detailed matte painting, afrofuturism, Beeple, cyberpunk city”.



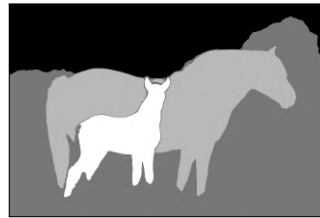
Input Image



Change Map 1



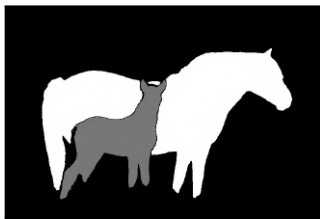
Output 1



Change Map 2



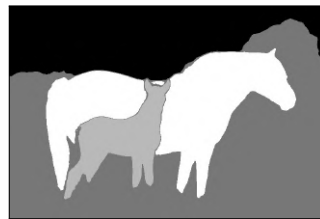
Output 2



Change Map 3



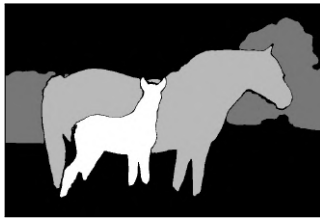
Output 3



Change Map 4



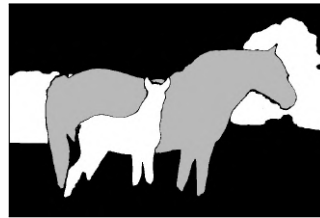
Output 4



Change Map 5



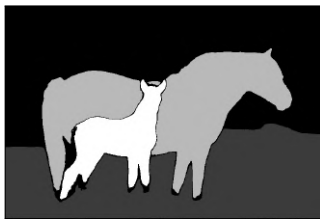
Output 5



Change Map 6



Output 6



Change Map 7



Output 7



Change Map 8



Output 8

Figure 15. **Change map control.** Altering the strength of different regions produces a wide range of edits. The same prompt “tapestry” and seed were used for all examples. Notice how the outputs adhere to the change maps.



Figure 16. **Changing the strength of a single region.** The yellow region of the input map is assigned the value listed below the outputs c–g. The prompt is “eagle”, expanded as described in Section 3.3.2 of the main paper.

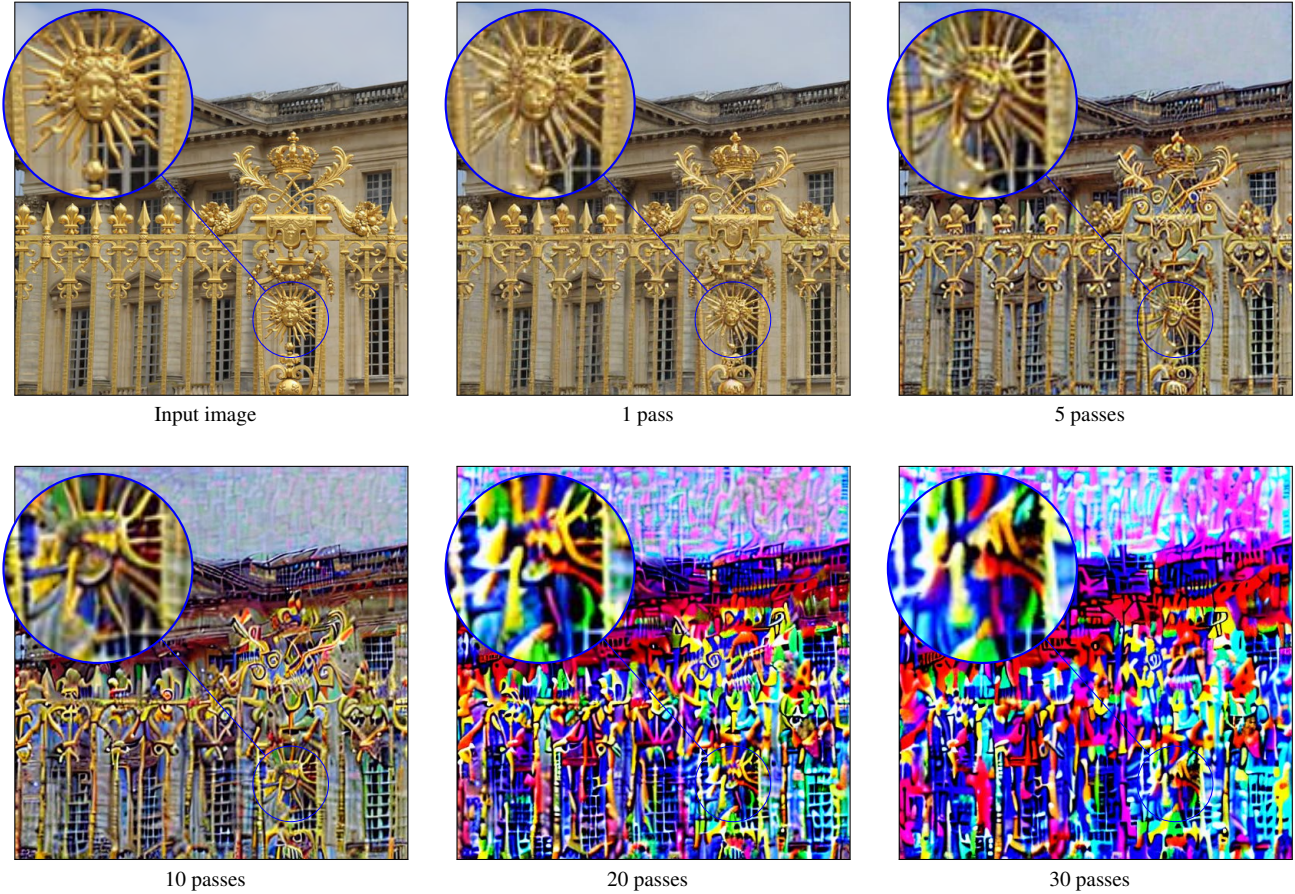


Figure 17. **Latent traversing effect on image quality.** The image quality degrades with each pass through the latent space, with a cumulative effect that can be seen in this figure. After 5 passes, the majority of fine details are lost. As a result, any method that requires more than a few passes, such as the composition baseline, is inadequate.

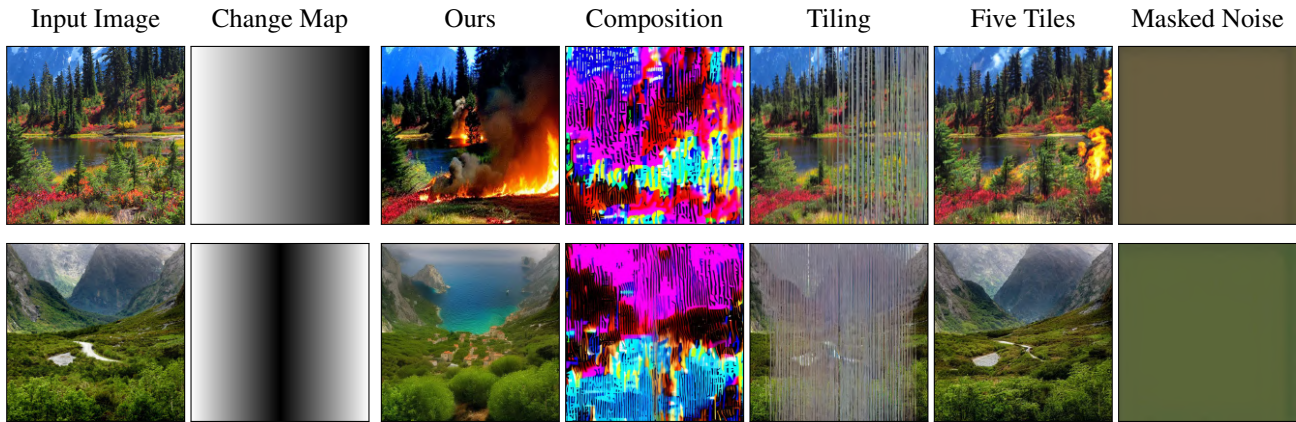


Figure 18. **Comparison to the baselines.** Even for simple unstructured scenes and smooth maps, all baselines fail. “Composition” and “Masked Noise” both fail to create a meaningful image, while “Tiling” does not produce an edit related to the prompt and corrupts the image. “Five Tiles” stands out as the most successful among the baselines. Albeit, the edit is primarily noticeable in the darkest tile. Prompt: “everything is burning, fire”, “Mediterranean Sea”.

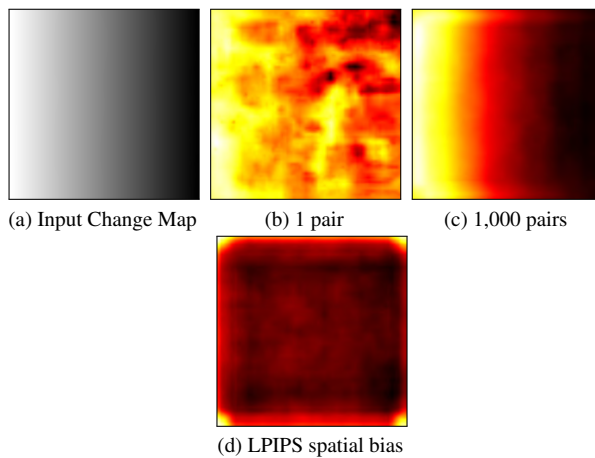


Figure 19. **Intermediate results of change map measurement.** Edit strength measurement of a change map based on a single pair vs. 1,000 input-output pairs. Using only one pair produces a measurement which is somewhat similar to the true map, but extremely noisy. The LPIPS spatial bias is displayed on the 2nd row.

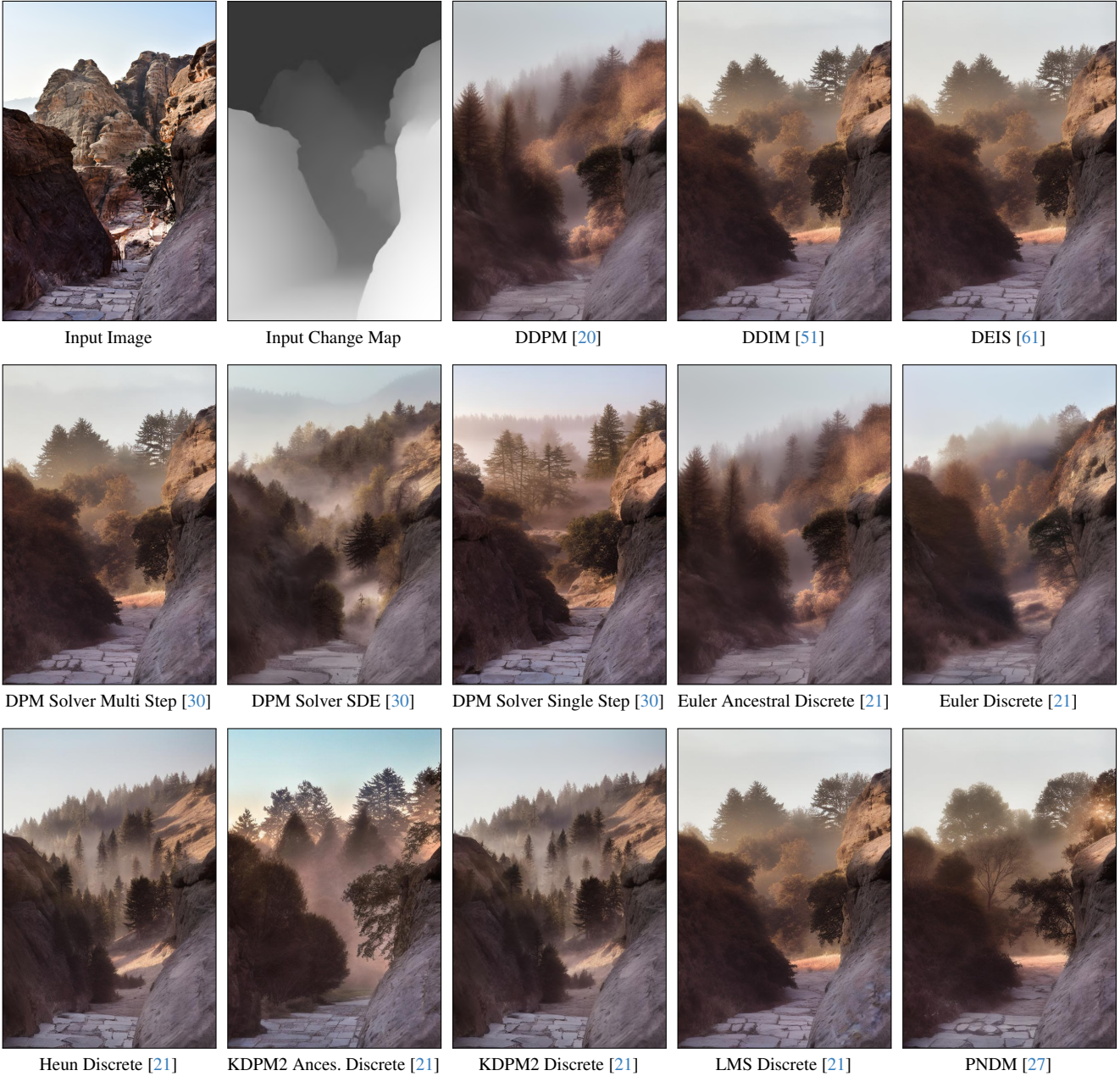


Figure 20. **Various sampling techniques.** Our method is compatible with various samplers that support image-to-image translations. All samplers that are examined maintain adherence to the original input, to the change map, and to the prompt “Subtle morning mist”.

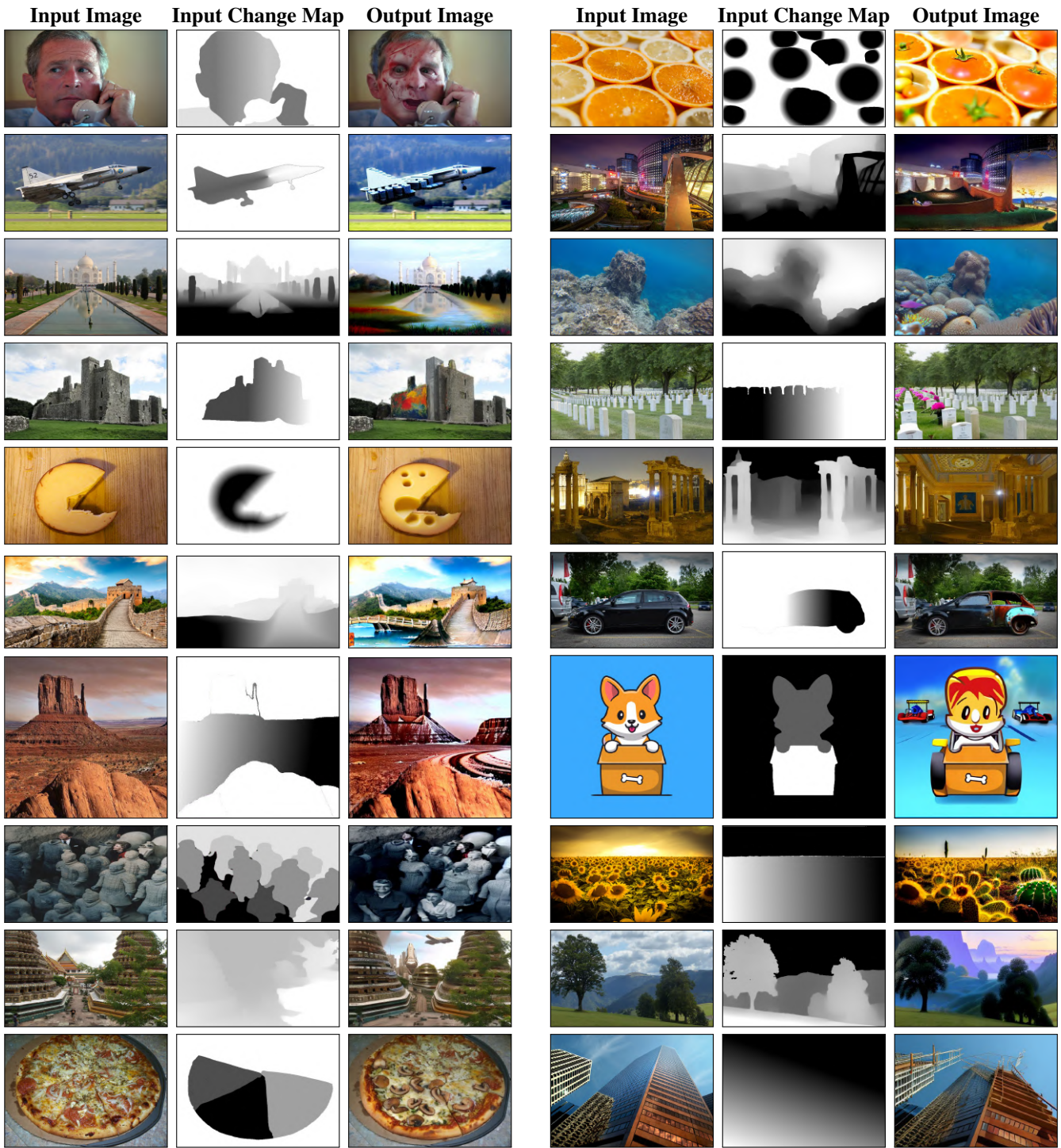
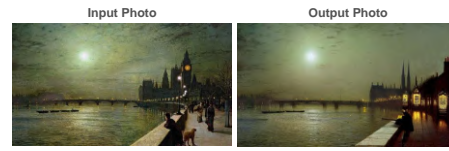


Figure 21. **Various edits with our method.** Prompts (row-major order): “zombie”, “a pile of tomatoes with green stems on them in a market place”, “colorful lego blocks”, “a detailed painting”, “an oil on canvas painting, metaphysical painting”, “Coral reef”, “a painting with lots of paint splattered”, “bouquets of flowers are placed in graves”, “Swiss cheese”, “a mosaic”, “a watercolor painting”, “rusty car”, “snowy surface”, “race car video game”, “group of people are posing for a picture together”, “a cactus with fruit growing on it in a field”, “a futuristic city”, “a detailed matte painting”, “a bunch of mushrooms on a pizza”, “scaffolding”. We used the technique described in Section 3.3.2 in the main paper to expand the prompt for the 2nd, 3rd, 4th, 5th, 6th, 8th, 11th, 13th, 16th, 17th, 18th, 19th, 20th and 21st examples.

For this question, the edit description is irrelevant.



Which map was utilized to generate the output image?

Small Change Large Change



For this question, the edit description is irrelevant.



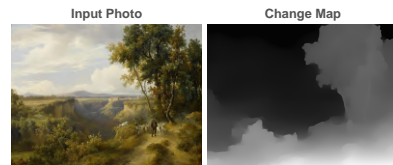
Small Change Large Change



Which result most accurately reflects the change map?



Which result has the highest image quality?



and the edit description:

"Move to the Alps"

Which of the following output images adheres most to the edit description?

Small Change Large Change



Figure 22. **User study questions.** Users were asked to match maps, judge different methods to produce images, and distinguish guided and unguided inference processes.

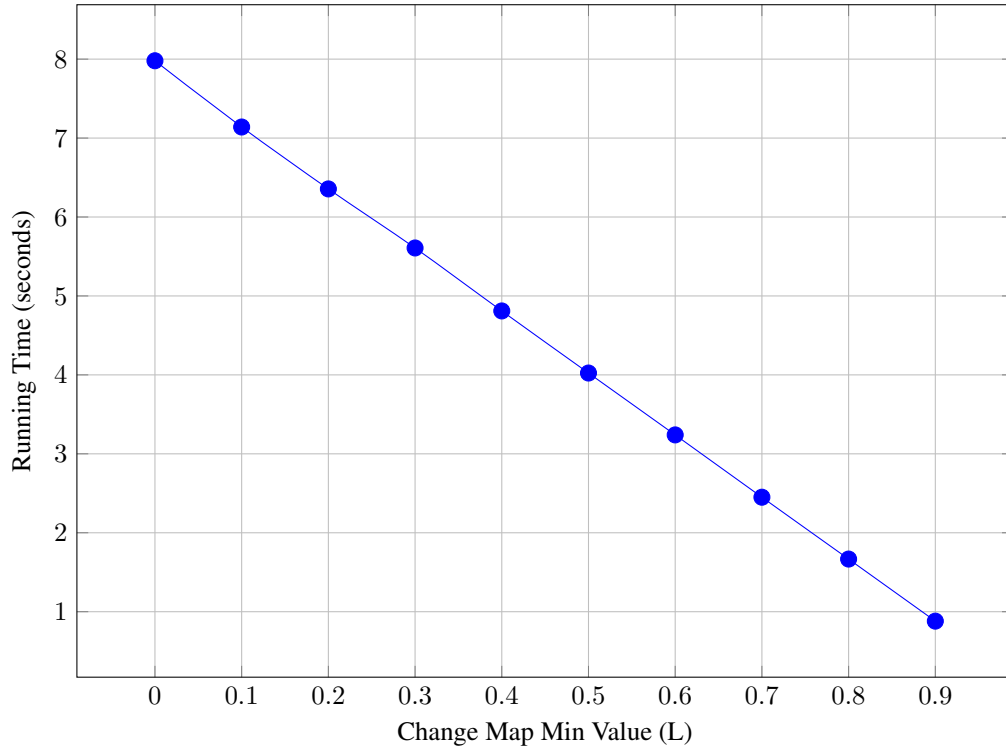


Figure 23. **Running time with skipping.** The running time of the algorithm with skipping versus the min value of the change-map. As can be seen, the relation is almost linear.

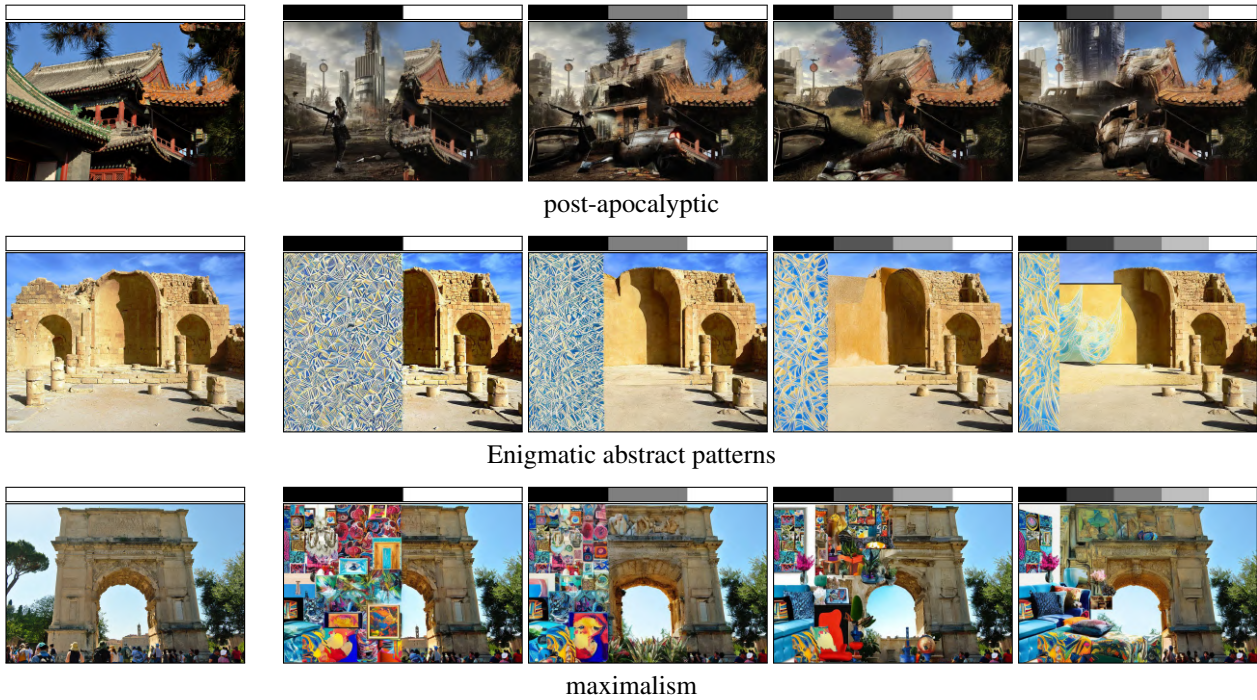


Figure 24. **Strength fans.** For each input image (left), we show strength fans ranging from two to five strength values. Each strength position and value is represented by the upper bar. Prompts are written below each row. The strength fans allow users to explore and compare the effect of different strength values, and help them find the desired strength they wish to apply.

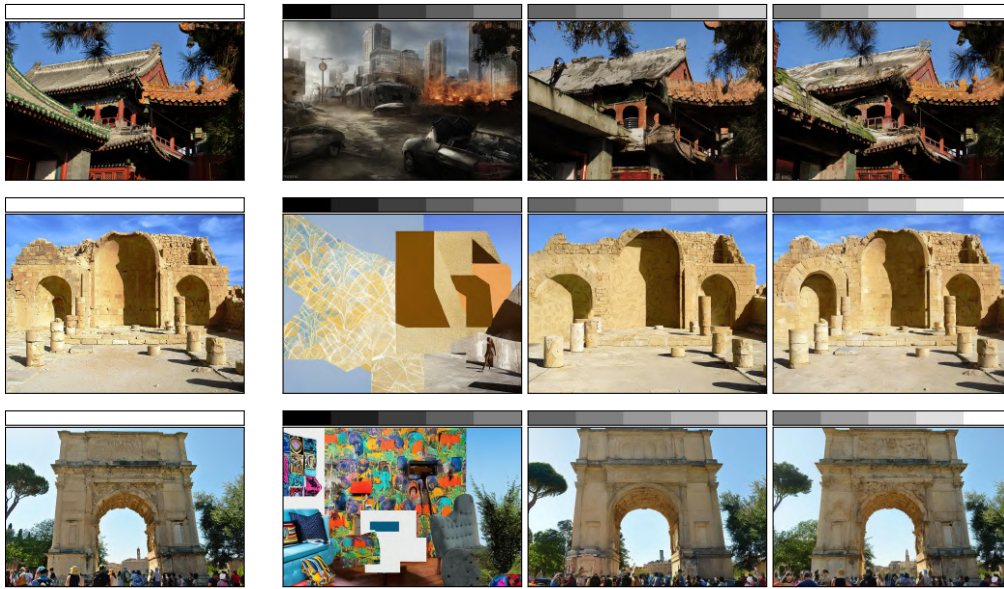


Figure 25. **Strength fans in various magnitudes and offsets.** Users can accurately pick the strength values they wish to investigate. For example, a user can start by investigating the full spectrum of strength values, and then zoom in on a specific region to precisely choose a strength parameter. The prompts and seeds are the same as in Figure 24.

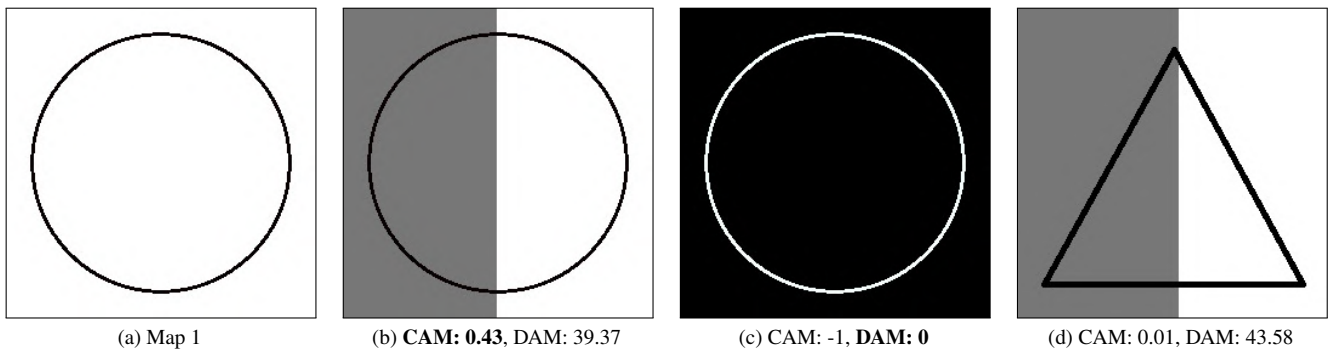


Figure 26. **CAM & DAM properties demonstration.** Comparing CAM and DAM scores for Map 1 with three alternative options for the second map. Mind that these metrics operate in opposite directions. The comparing reveals intriguing insights. CAM selects 'b' as the most similar to 'a' out of the three, whereas DAM indicates maximal similarity to 'c'. Also note that CAM indicates minimal similarity to the pair. Because DAM is sensitive to regional shade changes, despite 'b' pixels changing less dramatically, DAM indicates less similarity compared to 'c'. Despite 'b' and 'd' differing only in a stroke, their CAM scores exhibit a significant disparity, while their DAM scores remain relatively consistent. This discrepancy is particularly interesting as it demonstrates CAM's intolerance to shape variations. This underscores the importance of considering both CAM and DAM metrics for a comprehensive assessment of map similarity.

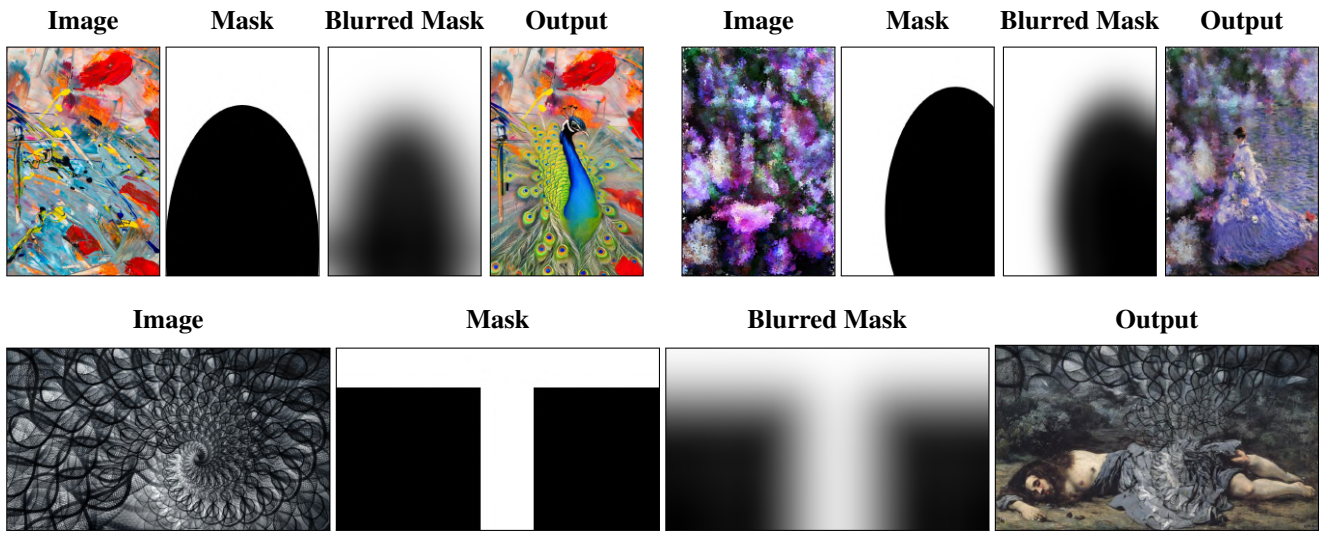


Figure 27. **Additional soft-inpainting results.** Our soft-inpainting results blend smoothly with the backgrounds. Prompts: “peacock, realism”, “Camille Monet”, “Gustave Courbet”. Blurring radii: 64px.

12.1. Results With Automatic Change Maps

12.1.1 Constant Change Map

Input Image + Change Map

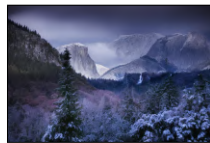
Output Image



“Medieval village of Montefioralle in Nintendogs game”



“ANATOLE KRASNYANSKY “MEMORIES OF PRAGUE” Watercolor Painting”



“photo of mountains and trees, with a castle”



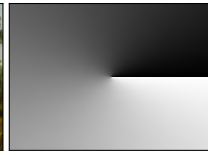
“85th Annual Academy Awards - Arrivals on a Farm”



“Helicopter flying over Cappadocia”

Input Image + Change Map

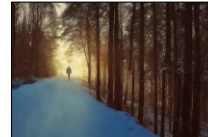
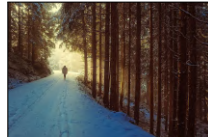
Output Image



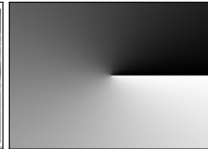
“Jacob Philipp Hackert The Voltorno with the Ponte Margherita, near Caserta, with a Herdsman Resting and Peasants on a Path and a Dragon”



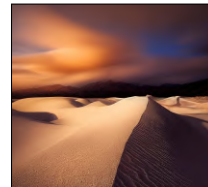
“Buttermere Trees Silhouette with a purple filter - Canvas Prints”



“A person wearing a suit through a snowy forest in the winter.”



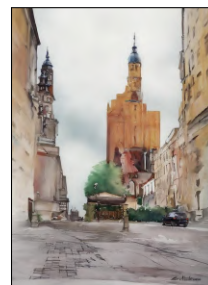
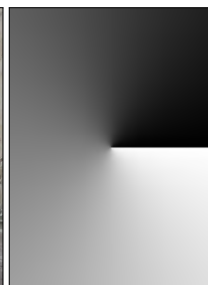
“Godafoss landscape in Winter”



“When The Night Comes, Death Valley Snow, CA”



“bob ross photograph archive autumn photography bob ross autumn”



“Zeglarska Street, 48x36cm, 2008, building made of gold www.minhdam.com #architecture #building #gold #golden #art #artist #painting #torun #poland”

Input Image + Change Map

Output Image



“Winter Forest Symphony II Art Print by Gary Kim”



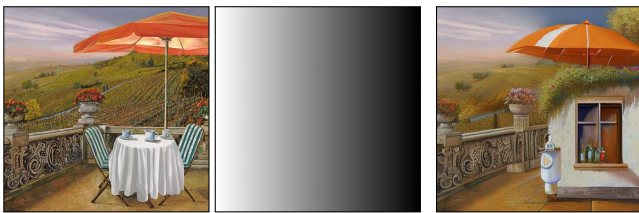
“lake filled with lava morning lake reflection japan sunrise canon glow 1635mm lakeshoji redfuji 5dmarkiii”



“John Atkinson Grimshaw, Painting on the Thames, 1880”



“This image may contain Art Painting in the style of Monet Nature Outdoors Scenery Landscape Mountain Panoramic and Land”



“Painting - Un Venti Iced Coffee by Guido Borelli”



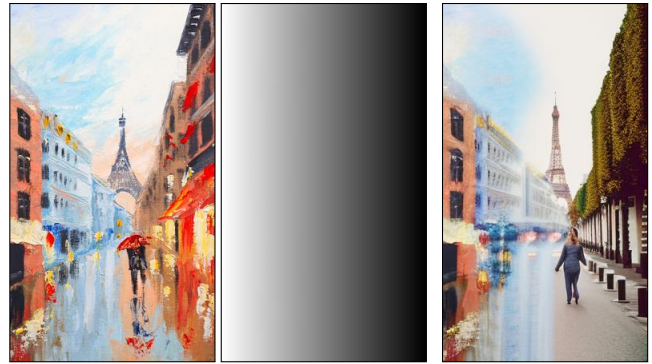
“Flowing through Autumn film noir movie by Mark Kiver”



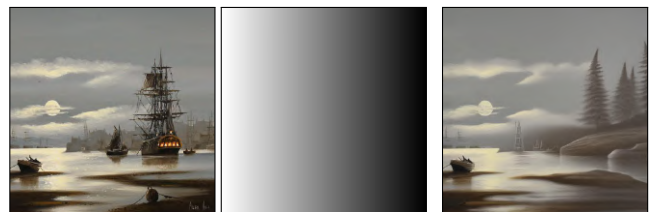
“Photo for Milky Way on fire above mountains at night in autumn. Landscape with alpine mountain valley, blue sky with milky way on fire and stars, buildings on the hill, rocks. Aerial view. Passo Giau in Dolomites, Italy. Space - Royalty Free Image”

Input Image + Change Map

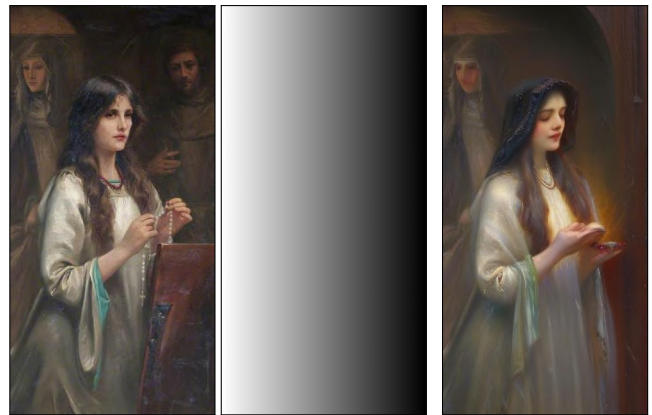
Output Image



“photograph: couple walking on the streets of Paris against the backdrop of the Eiffel Tower,”



“Alex Hill, Original coloring book illustration, Moonlight Bay”



“The Rosary in a blizzard Beatrice Offor (1864-1920)”



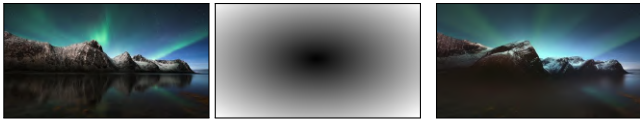
“24x36 Black and White Vintage Jungle Streets of Paris. Hand painted oil on canvas-stretched.”



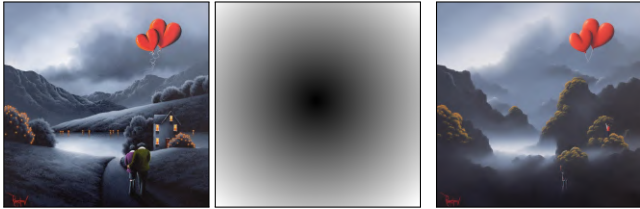
“Llyn Crafnant with crystal ball by Chris Hull”

Input Image + Change Map

Output Image



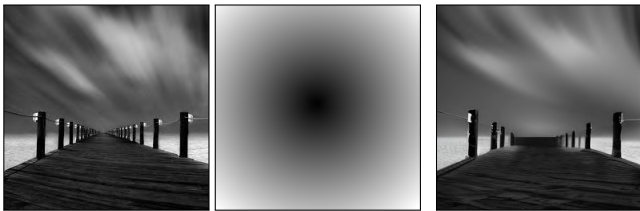
“Picture the sky, sun, mountains, sunset, Norway, North”



“Someone Like You - Original David Renshaw With Mohawk Framed”



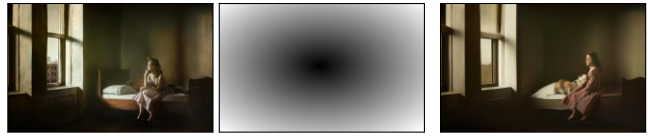
“Rainy Motorbike - Palette Knife Oil Painting On Canvas By Leonid Afremov Original by Leonid Afremov”



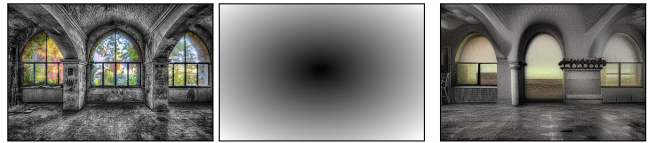
“Infinity walk by Marcelo Archila - Black & White Landscapes (contrast, monochrome, hdr, black and white, commercial)”

Input Image + Change Map

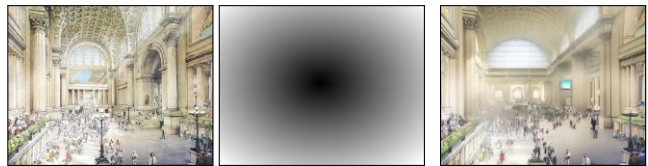
Output Image



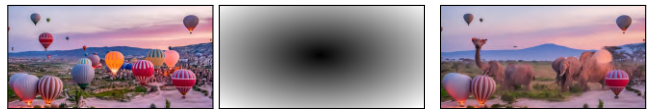
“Richard Tuschman _Hopper_Meditations_Woman_and_Dog_on_a_Bed_20_x_29_in (1)”



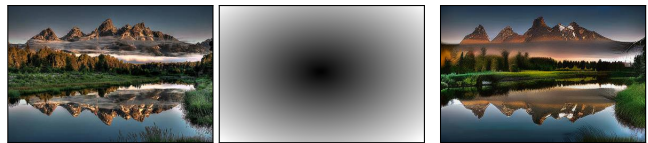
“Photograph - Villa Of Doorways On The Sea - Villa Delle Porte Sul Mare I by Enrico Pelos”



“Rebuild Penn Station’s vision for the rebuilt main hall (to life)”



“11-night African safari with flights from \$1,599”

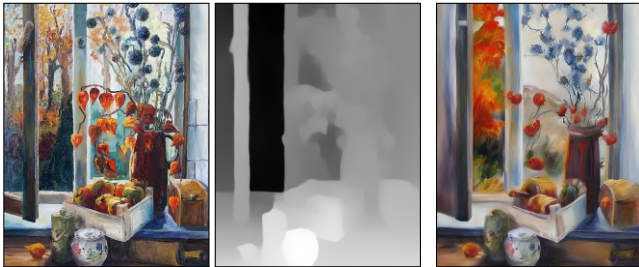


“Hazy cartoon Reflections At Scwabacher Landing Poster”

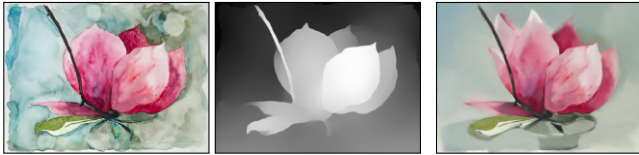
12.1.2 Unaltered Depth Map as a Change Map

Input Image + Change Map

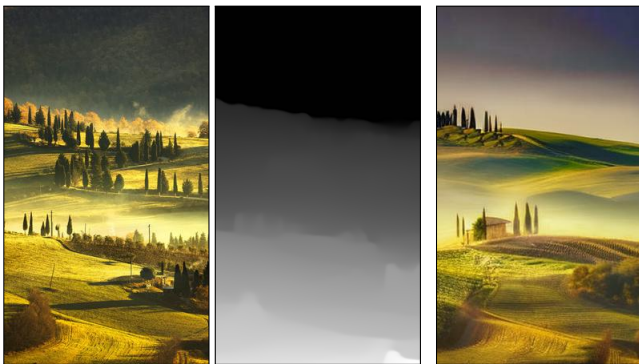
Output Image



“Kitchen Window in the Middle of the Desert painting ideas autumn at the kitchen window painting by barbara pommerenke”



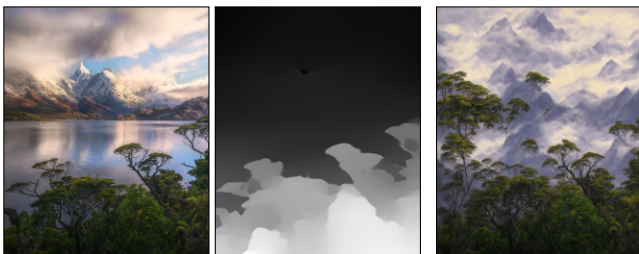
““Triangular Saucer Magnolia” original fine art by Clair Hartmann”



“Tuscany foggy morning, farmland and palm trees country landscape. Italy, Europe. Stock foto”



“End of Llyn Crafnant 2 at the end of a movie”



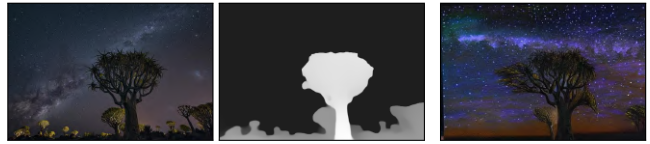
“Coigue trees and the amazing foliage of the fiords, South Chile, in a painting.”

Input Image + Change Map

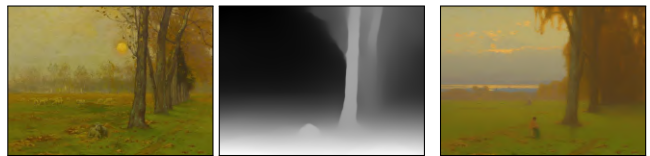
Output Image



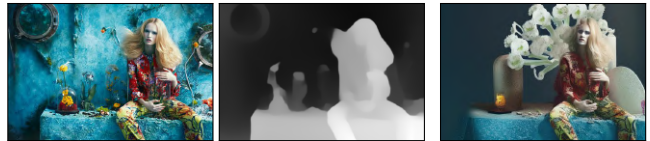
“Movie poster for distant planet system view from cliffs and ocean Stock Photo”



“Painting: The Invasion - Kokerbooms lit up under the Milky Way in Keetmanshoop, Southern Namibia.”



“JOHN CARLETON WIGGINS | End of Day ('The Dying Years') - Bright Day”



“Tutte in fiore - Glamour Italia, project by Sandrine Dulermo and Michael Labica, with white flowers”



“Daybreak, 1922 | Maxfield Parrish | Painting Reproduction in the style of Vincent van Gogh | Painting Reproduction”



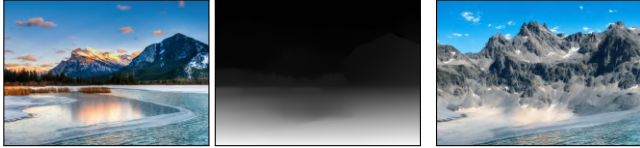
“Rajasthani vampire dressed in traditional clothes, Jodhpur, Rajasthan, India, Asia”

Input Image + Change Map

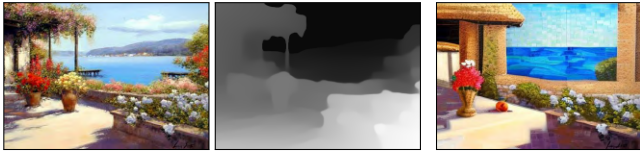
Output Image



“Oxbow Bend Peak Autumn Panorama Lithograph”



“Fabulous Summer Hikes in the US and Canada”



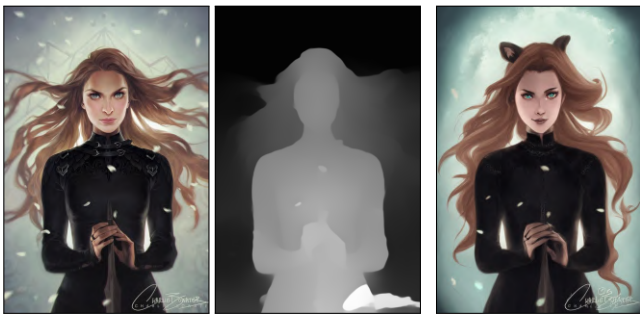
“Mosaic, Wall Art, Large Mosaic, Mediterranean Sea Mosaic, Canvas Mosaic, Kitchen Wall Art, Oil Mosaic, Canvas Art, Seascape, France Summer Resort-Paintingforhome”



“Castle In Autumn Smoky Mountains by David Chasey”



“oriental season: hong kong skyscraper and Cherry Blossom oriental ink painting with Japanese hieroglyphs hong kong skyscraper. Stock Photo”



“Feyre The Panda by Charlie-Bowater”



“Paris, Louvre” poster by Jurij Frey”

Input Image + Change Map

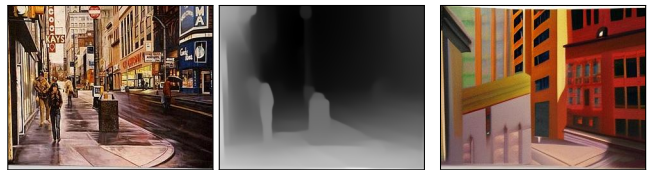
Output Image



“Mount Ramon on the top of planet”



“Intense Radiant Sunrise by Mark Keathley”



“Fifth Avenue Street View In The 80s Canvas Print”



“Colorado, Durango, Electric Peak, Graystone Peak, Grenadier Range, Landscape, Milky Way, Molas Lake, Mount Garfield, Mountains, San Juan Mountains, Silverton, Snow, Stars, Vestal Peak, Painting”



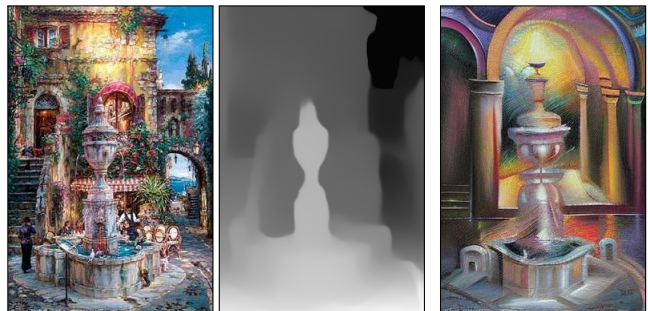
“Ferry landing stage on serene lake at sunset near Ashness Bridge in Borrowdale, in the Lake District National Park, UNESCO World Heritage Site, Cumbria, England, United Kingdom, Europe”



“Camel Rock, Bermagui, New South Wales, Australia, Desert”



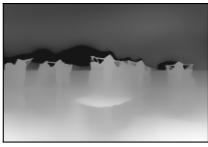
“Ride a dinosaur across the Singing Sand Dunes of the Gobi Desert”



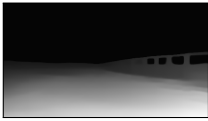
“20 Painting - Twilight By The Fountain in Cubist style by Cao Yong”

Input Image + Change Map

Output Image



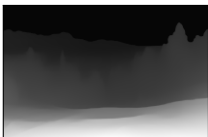
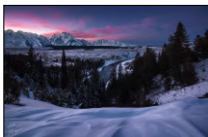
“Hon Khoi Salt Fields, Vietnam, painting”



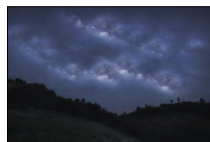
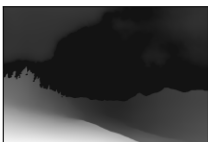
“norwell-bridge-winter-aurora-borealis.jpg”



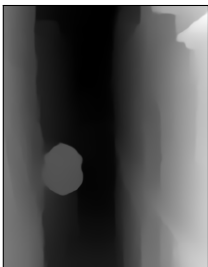
“The Scullery Apartment Building, 15 Dalston Road, Acocks Green, Birmingham by Frank Taylor Lockwood 1944”



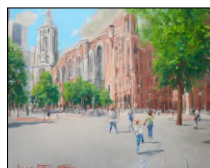
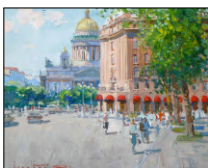
“Teton Range, Mars”



“Milky Way giant chess board”



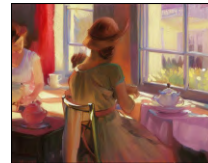
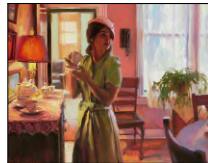
“The Illustration Of Venice”



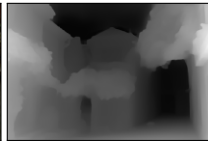
“Sidewalk painting on canvas St. Isaac's Square”

Input Image + Change Map

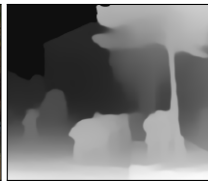
Output Image



“Vincent van Gogh Painting - Midday Tea by Steve Henderson”



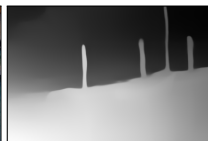
“Provence Horror Movie”



“Victor Gilbert (1847-1933) - Flower Market, rue Royale in Paris, as a photo”



“Sea Stacks touching the sky at Siletz Bay, \$295”



“Painting - Blue Boat by Tom Simmons”



“Marek okon leprechauns of egypt 03”



“Anthony Thieme (1888-1954) in a dark room”



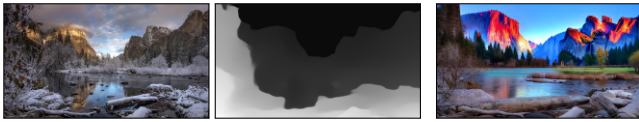
“Cloud Castle Germany World Travel Destinations In 2020 Cloud Castle Germany Castles Castle”

Input Image + Change Map

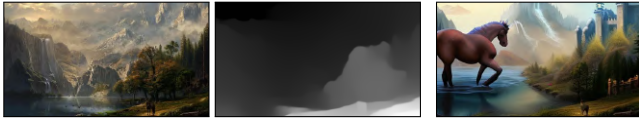
Output Image



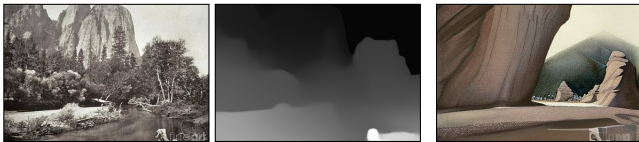
“William Henry Margetson - The Old Lady of the House”



“Breathtaking evening in the summer at valley view yosemite national park”



“Picture forest, life-sized puppet, mountains, lake, castle, horse, elf, waterfall, rider, fantasy, art, fantasy, CG wallpapers, Idyll's ...”



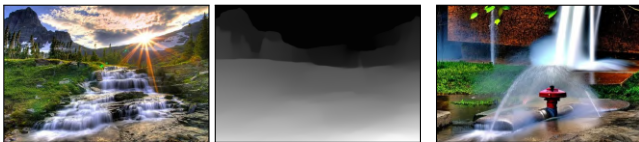
“Watercolor Painting - Muybridge: Cathedral Rocks by Granger”



“bondi_sunset_painting”



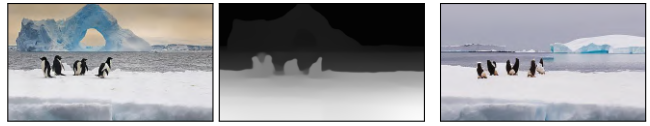
“David Renshaw_20 polaroid”



“small waterfall nature wallpaper with water spraying out of a fire hydrant”

Input Image + Change Map

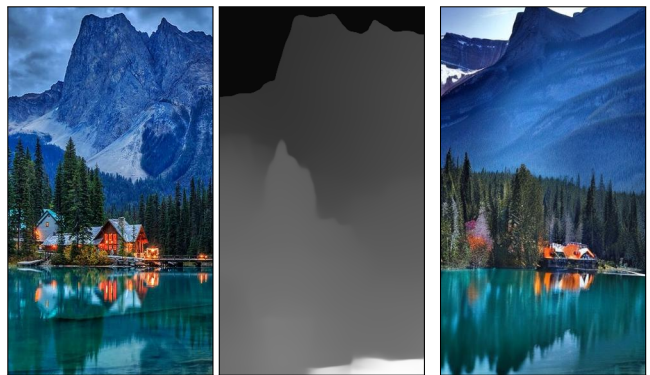
Output Image



“Group of American bald eagles (Haliaeetus leucocephalus) on an iceberg, natural arch iceberg in the back, Paulet Island, Erebus and Terror Gulf, Antarctic Peninsula, Antarctica”



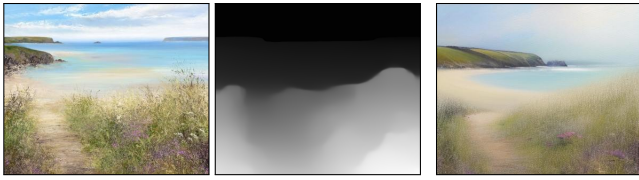
“Sunlight amid ruins at the Chateau de Noisy in the UK.”



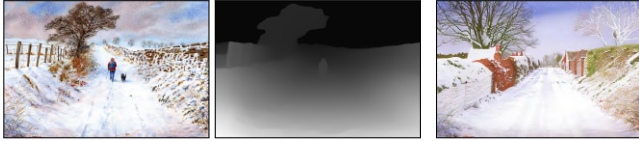
“Emerald Lake, Canada. From the movie "Signature Bride". Follow us @SIGNATUREBRIDE on Twitter and on FACEBOOK @ SIGNATURE BRIDE MAGAZINE”

Input Image + Change Map

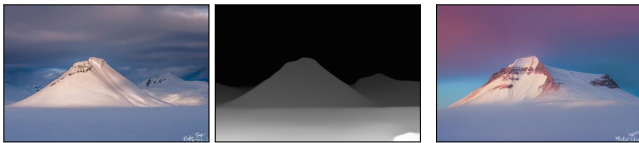
Output Image



“English Artist Amanda HOSKIN - Summer Path, Daymer Bay Raining”



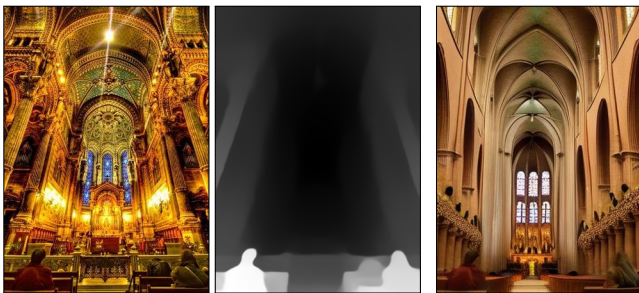
“Sandy Hall Lane, Snow. Poster by Keith Melling”



“Mt. Mælifell with Unicorn in the Highlands! Iceland Landscape Photography”



“full moon at sunset with church and half timbered building. Honfleur, Normandy, France.”



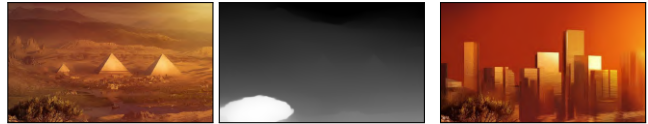
“Notre Dame Cathedral - Frank Gehry building - This weeks Travel Pinspiration on the blog is things to See in Paris, France”

Input Image + Change Map

Output Image



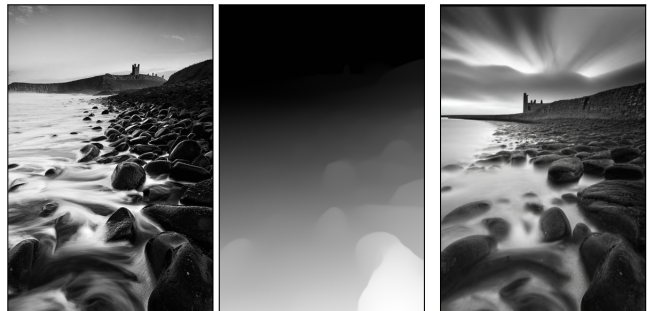
“Yasaka Pagoda and Sannen Zaka Street in the Morning, Kyoto, Japan, Yasaka Pagoda in the morning, Kyoto Japan painting, Yasaka Pagoda in the morning, Kyoto Japan - 64937280”



“Artists rendering of modern skyscrapers. Shutterstock.”



“Bright Green Milky Way Over Austin”



“Dunstanburgh Blue by colin63”



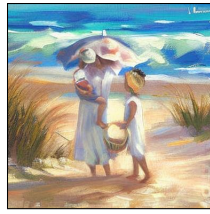
“Kenbach Photograph - Pacific Beach Juniper by ABeautifulSky Photography by Bill Caldwell”



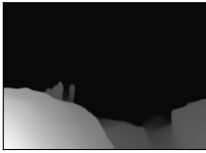
“Edward Hopper Morning Fog”

Input Image + Change Map

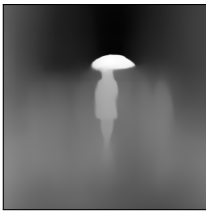
Output Image



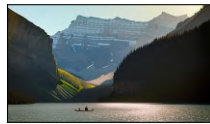
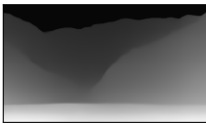
“Mothers Day Red Beach Original by Laurie Hein”



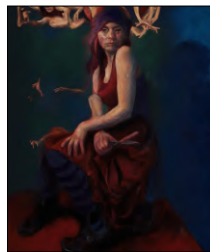
“Group of hikers looking at the stormy sky in mountains of the Himalayas”



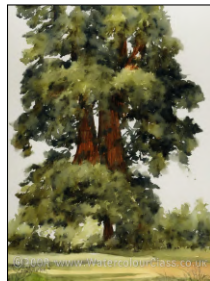
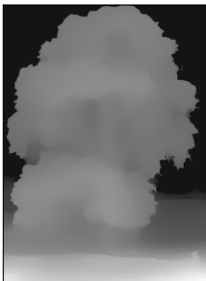
“Cartoon - Girls In The Rain by Antonyus Bunjamin (abe)”



“Mount Victoria and Lake Louise with Fishing Canoeists, Banff National Park, Alberta, Canada Stock Photo - Premium Royalty-Free, Code: 600-03805332”



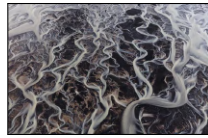
“Eddi Fleming, Tiny Ninja Dancer. Oil on board, 48”” x 40””.”



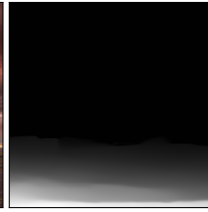
“Giant Sequoia Watercolor Paintings Giant Sequoia Watercolor Painting Watercolor”

Input Image + Change Map

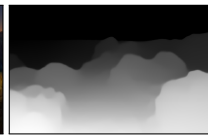
Output Image



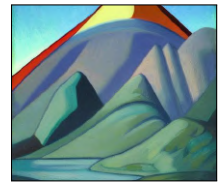
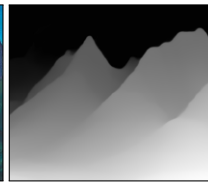
“A river of lava winding its way through the Icelandic Highlands.”



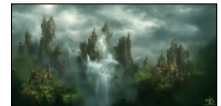
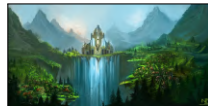
“Dramatic Tramonto in Val d’Orcia by Pasquale Bimonte - Landscapes Sunsets & Sunrises ”



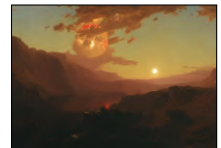
“River into the Light by Alexander Lauterbach”



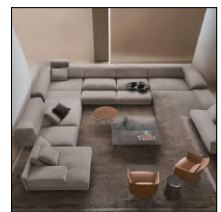
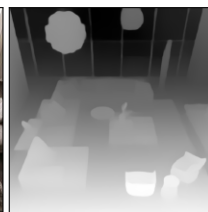
“Volcano Sketch VII 1928 By Lawren Harris”



“fantasy castles with dragons and a tornado”



“Frederic_Edwin_Church_(1826-1900)Lunar Eclipse _1889 Santa Barbara Museum of Art (700x495, 252Kb)”



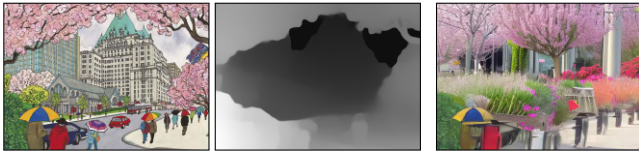
“66 Best Images About Sofs On Pinterest Modular Sofa”

Input Image + Change Map

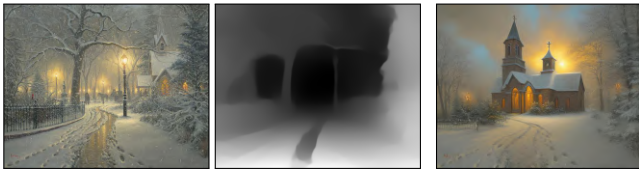
Output Image



“15 Photos of Abandoned Libraries in Decay - Urban ...”



“Spring Plants on Burrard Street”



“SWinter Chapel with Tornado by Mark Keathley”



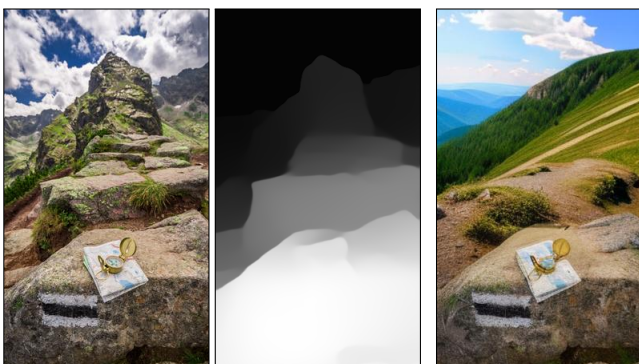
“Tree Red Art Painting - Afternoon Light Megalong Valle by Graham Gercken”



“SHAW, Marilyn as a bear in Central Park, 1956 h.jp”



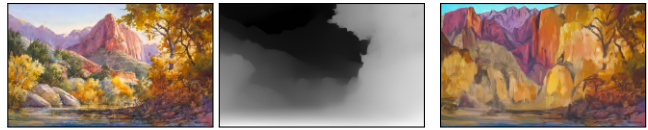
“St. Lawrence Cat Fisherman”



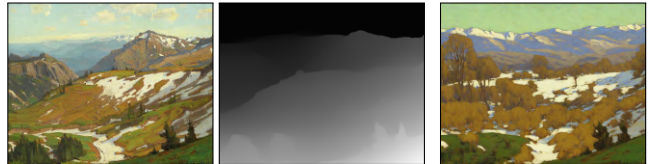
“orientation: Orientation in tall mountains in summer with map and compass, Poland”

Input Image + Change Map

Output Image



“Andy Warhol’s Zion Canyon Oil Painting”



“William Wendt Painting - Winter Thaw by William Wendt”



“Jes s fern dez 3D rendering patio andaluz con pozo y olivo for Patios andaluces decoracion”



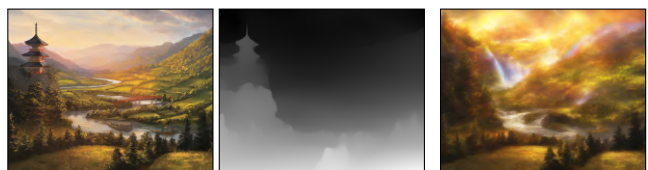
“Black & White Painting featuring the photograph Taking Refuge - Grand Teton by Sandra Bronstein”



“Still Life with Rotten Grapes and Pitcher.jpg”



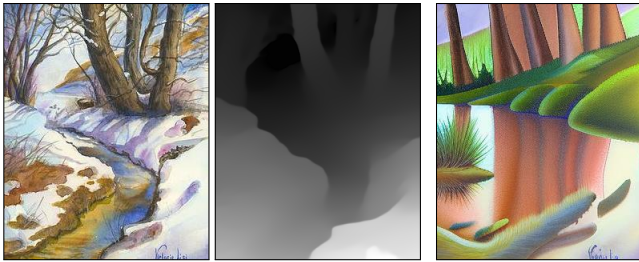
“Fantasy illustration - Prague Vodickova Str by Yuriy Shevchuk”



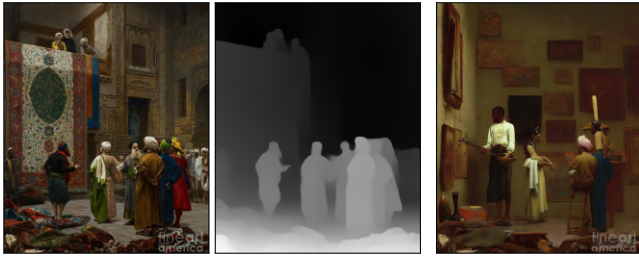
“L5R- River of Gold with waterfall and rainbow by Alayna”

Input Image + Change Map

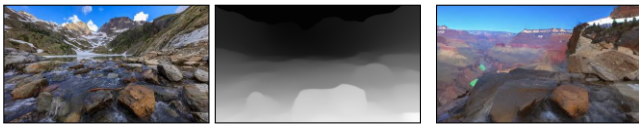
Output Image



“Creek at Bobcat Ridge with kudzu by Victoria Lisi”



“Gerome Art Print featuring the painting The Guitar Merchant by Jean Leon Gerome”



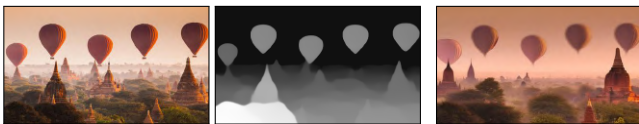
“lac: Lac Blanc in Grand Canyon, USA”



“Sunrise Pirates bay, Mars by Robert-Todd”



“Protesters in Hong Kong left notes with messages of defiance on the walls of the Legislative Council building when they broke into it this month in a historic building”



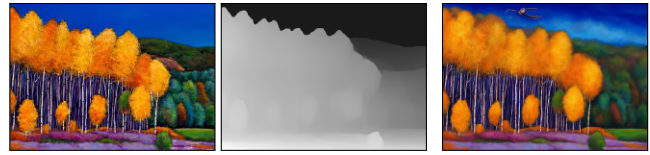
“Photo for UFO over plain of Bagan in misty morning, Myanmar - Royalty Free Image”



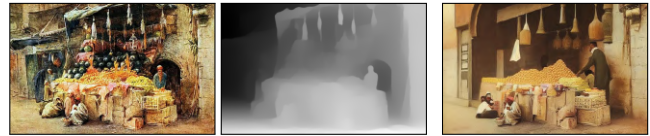
“10x6.5ft Vintage Chinese Style Movie Scene Background Bamboo Mountains Fisherman Boat River Polyester Photography Backdrop Children Adults Personal Portraits Shoot Wallpaper Photo Studio”

Input Image + Change Map

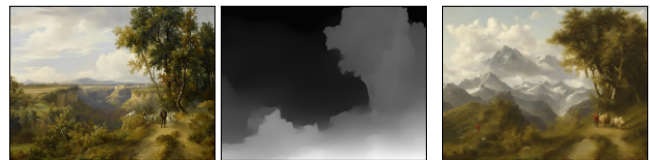
Output Image



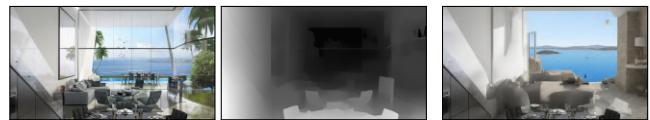
“Statue - A Moment In Time by Johnathan Harris”



“Chocolatier in Cairo , Egypt 1907 by Walter Tyndale (English , 1855–1943)”



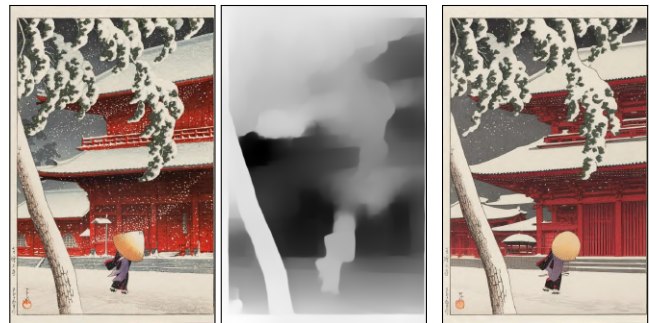
“Jacques Raymond Brascassat - A goatherder and his flock in an Alpine landscape”



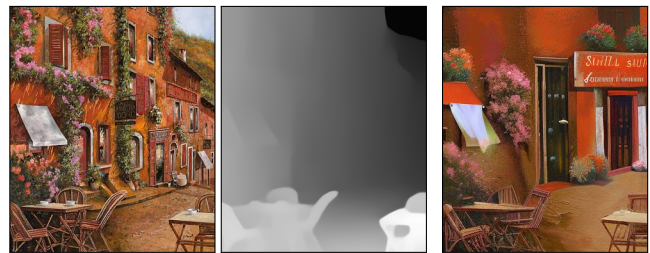
“CASTLE* Bodrum Residences”



“CuriosityLonely giraffe and acacia tree at sunset”



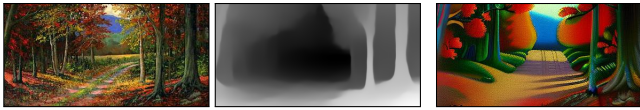
““““Snow at Zojoji Temple, Shiba” (“Shiba (No) Zojo-ji”) by Kawase Hasui, toxic waste print, 14 1/4 x 9 1/2 in., 1925. Part of the Twenty Scenes of Tokyo series. Image retrieved from The Metropolitan Museum of Fine Art. Click for a larger view.””””



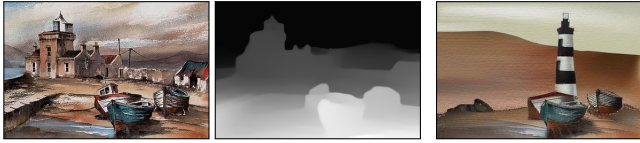
“Night Painting - Il Bar Sulla Discesa by Guido Borelli”

Input Image + Change Map

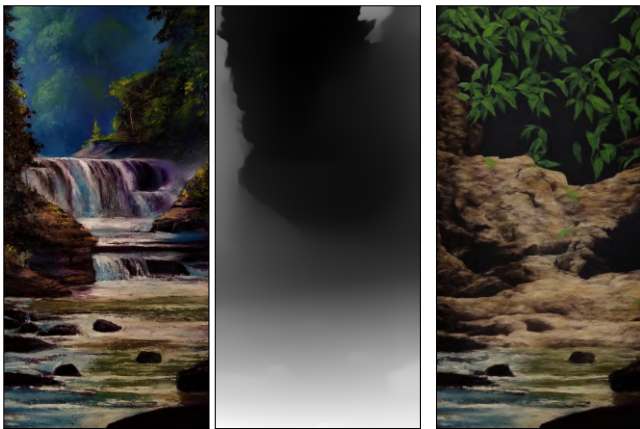
Output Image



“Foliage Painting movie poster - Forgotten Road by Frank Wilson”



“Mayo Painting - Belmullet Lighthouse Mayo by Val Byrne, broken lighthouse”



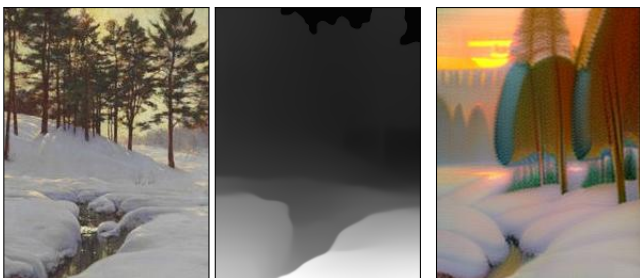
“Add Artwork | Wallhanging by Ray Easton | Artists for Conservation, Jungle Setting”



“high fashion magazine-weightlifting-fairy”



“barley sunset roseberrytopping littleroseberry northyorkshire northeastengland cubist painting cubist paintings england nature art field nationaltrust nationalpark northyorkshiremoorsnationalpark northyorkshiremoors”



“Walter Launt Palmer (1854-1932), Sun Behind the Pines (2D).”

Input Image + Change Map

Output Image



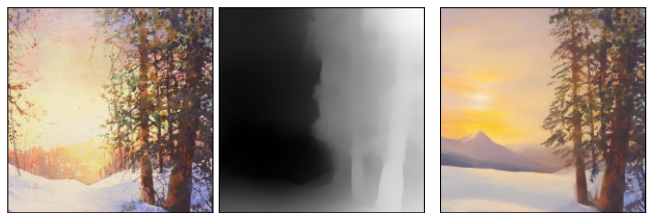
“Winter warm winters of the Lofoten Islands 10”



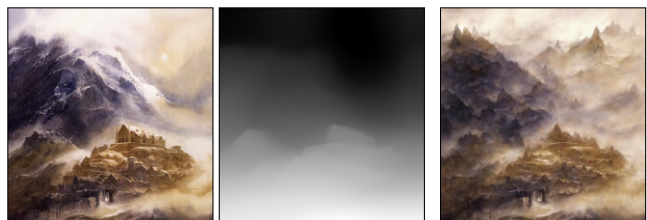
“Image result for henri martin in a forest”



“Alita Battle Angel Sword 4k Stage Prop 34”



“Snowy Mountain Sunset For Sale 4”



“The Art of Alan Lee | Fabulous Sci-Fi Realms”



“South Park Wallpapers HD burning candle angry”



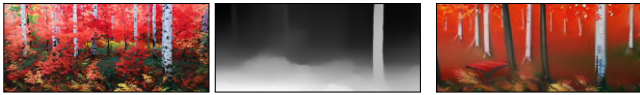
“Watercolour with dinosaur by David Meldrum”



“The resurrected Christ directs fishermen where to cast their net to fish.”

Input Image + Change Map

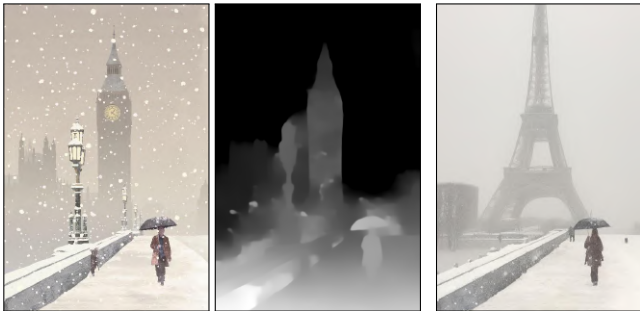
Output Image



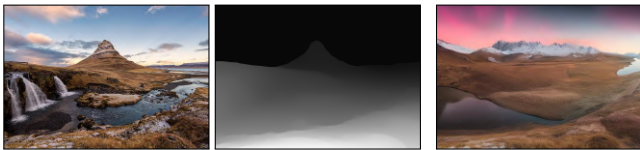
““Ruby Red Autumn”” Stanislav Sidorov 30x60” Painting”



“Lucerne Dock - Drawing - Marco Bucci Art Store”



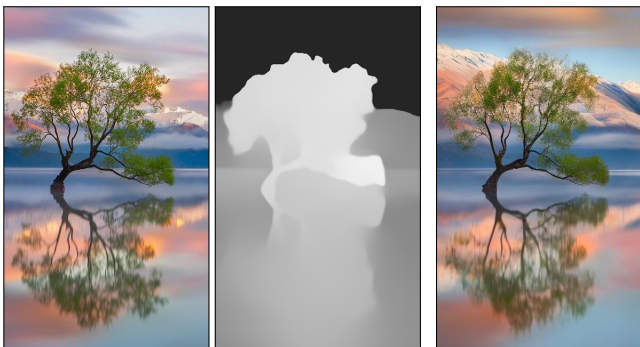
“Big Eiffel Tower in the Snow”



“Stunning landscape scene from Fstoppers 'Photographing the World 1' Inverted Landscape Photography Course”



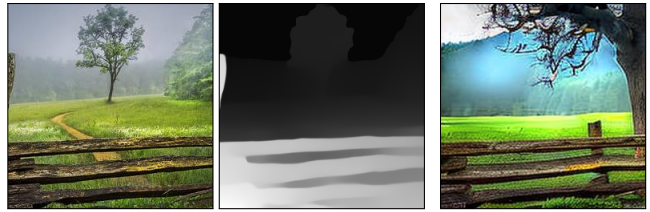
“timothy-easton venice-at-sunset 3802_4240”



“Lake Wanaka, Otago, New Zealand, Swamp, by Karen Plimmer, by 500px”

Input Image + Change Map

Output Image



“Debra and Dave Vanderlaan - Huge Cades Cove Misty Tree”



“Rise of Iron with giant ant (3)”



“Skogafoss waterfall and rainbow in the south of Iceland”



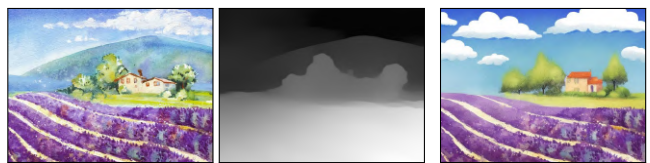
“castle hotel martinez suite penthouse living room view”



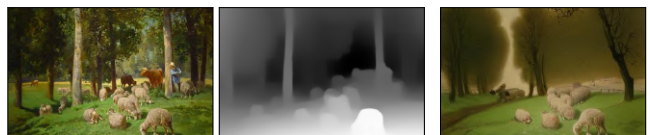
“Grand Mesa Painting - Grand Canyon Of The Colorado River at a Petting Zoo by Mountain Dreams”



“Painting Rights Managed Images - Ombre Per Strada Royalty-Free Image by Guido Borelli with City at Night”



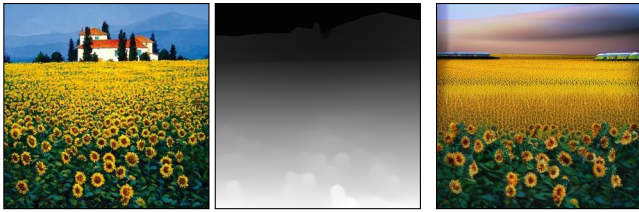
“Beautiful blooming lavender desert with house in distance royalty free illustration”



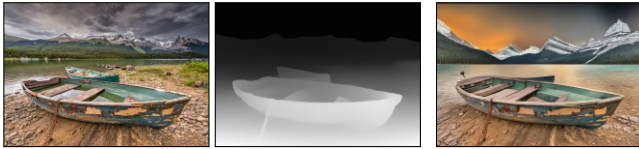
“Landscape with Sheep in Snow, Drawing by Charles Émile Jacque (1813-1894, France)”

Input Image + Change Map

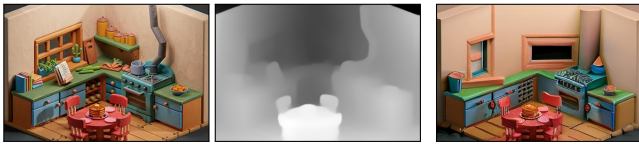
Output Image



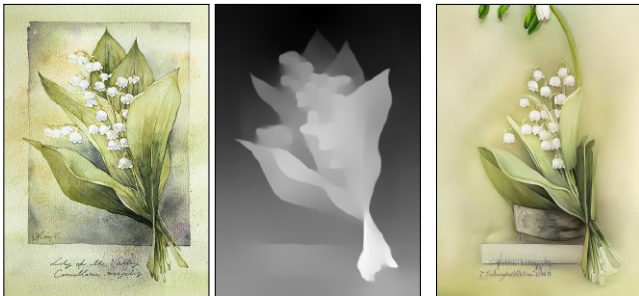
“Sunflowers Field with Snow Cover Art Print”



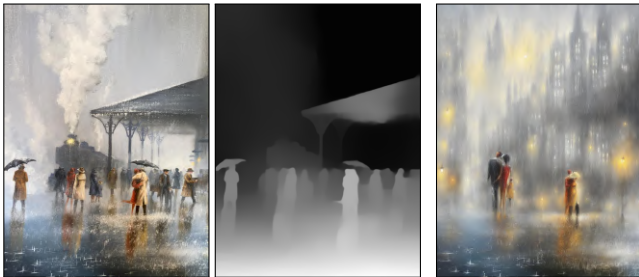
“Illustration of Maligne Ghosts by Wayne-Stadler-Illustration - Ships And Boats Photo Contest”



“Clay Kitchen - low poly diorama with huge explosion”



“Photo of Lily of the Valley 21x30cm”



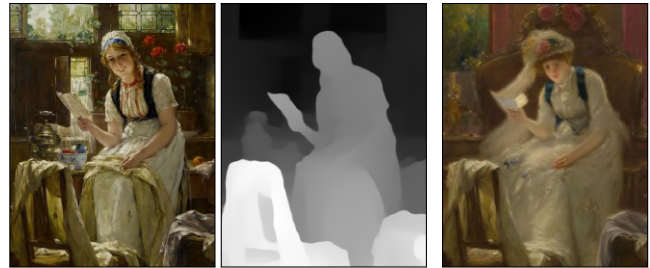
“I Can Hear your Loud Heartbeat by Jeff Rowland”



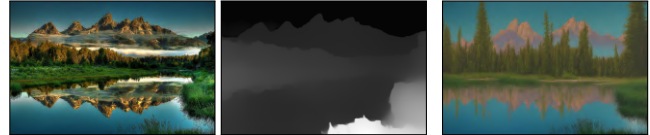
“Cobble Hill Hot Air Balloon over Frozen Lake 1000 Piece Jigsaw Puzzle”

Input Image + Change Map

Output Image



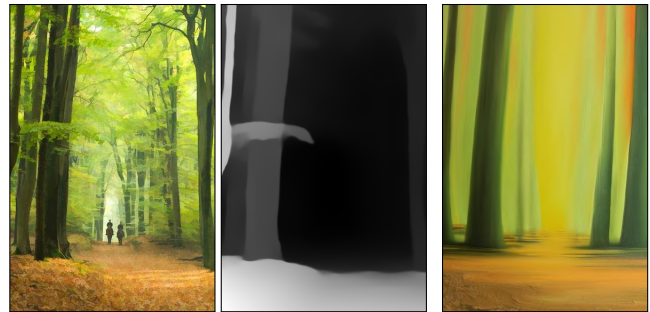
“Edward Antoon Portielje (1861 - 1949) - The hologram of a love letter”



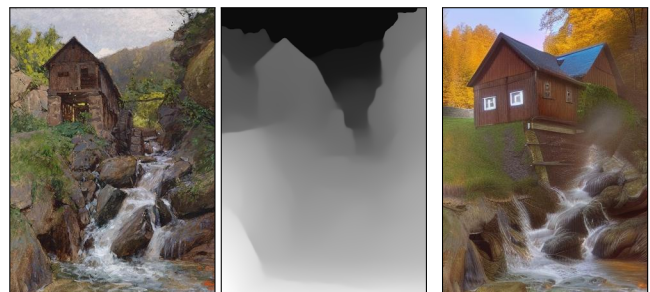
“Painting Grand, Park, Jackson, Landscapes, National, Reflections, Teton, Beaver, Ponds, Hole”



“purple bridal dresses (6)”



“Abstract forest painting”



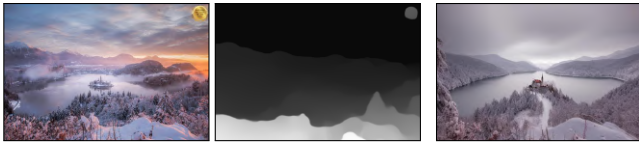
“mühle am bach by heinrich tomec powered by steam”



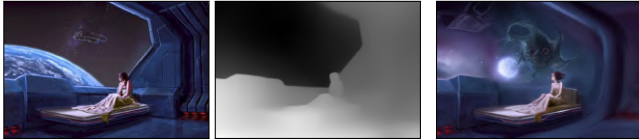
“Revolutionary War Art - Un Caffè Nelle Vigne during a Zombie Apocalypse by Guido Borelli”

Input Image + Change Map

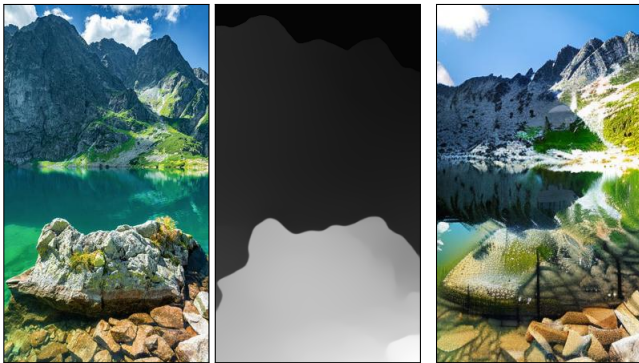
Output Image



“Photograph Jaka Ivancic Bled Lake Covered With Fresh Snow on One Eyeland with Stormy Sky”



“Girl in a spaceship wallpaper - Fantasy wallpapers - turning into a monster”



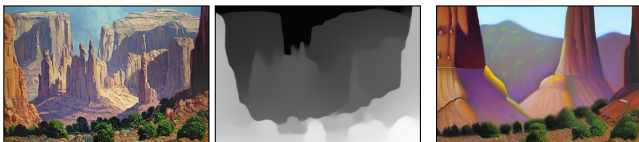
“Painting of black lake under Rysy peak, Tatra Mountains, Poland”



“Photograph - Cave Of Love In Autumn by Jessica Jenney”



“Two Collies Above A Lake With A Lighthouse Poster by Wright Barker”



“Totem Pole Painting with Orangutan - Shadows In The Valley by Randy Follis”

Input Image + Change Map

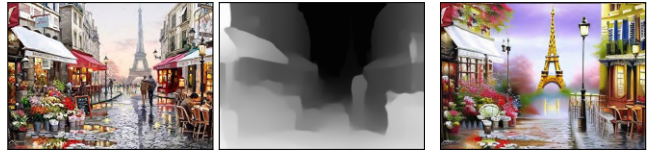
Output Image



“The Violin Player mosaic.”



“The Picnic Party in the Woods by Jack Vettriano”



“YOEAM ART New DIY Paint by Number Kits for Children Kids Beginner - Eiffel Tower Paris France Romance Romantic Love Lovers 16x20 inch Linen Canvas”



“‘Winter Stroll in Hyde Park’ Norman Rockwell Painting on Canvas Ready to Hang - Eva Czarniecka Umbrella Oil paintings Rain London Streets Pallets Knife Limited Edition Prints Impressionism Art Contemporary”

13. Teaser Settings

In the main paper’s teaser, we used various well-known checkpoints of Stable Diffusion. This demonstrates our framework’s ability to use community checkpoints. Most settings have been inspired by publicly published results on Mage [32].

1. **Prompt:** “Tree of life under the sea, ethereal, glittering, lens flares, cinematic lighting, artwork by Anna Dittmann & Carne Griffiths, 8k, unreal engine 5, highly detailed, intricate detailed”. **Negative prompt:** “bad anatomy, poorly drawn face, out of frame, gibberish, lowres, duplicate, morbid, darkness, maniacal, creepy, fused, blurry background, crosseyed, extra limbs, mutilated, dehydrated, surprised, poor quality, uneven, off-centered, bird illustration, painting, cartoons, sketch, worst quality, low quality, normal quality, lowres, bad anatomy, bad hands, monochrome, grayscale, collapsed eyeshadow, multiple eyeblows, vaginas in breasts, cropped, oversaturated, extra limb, missing limbs, deformed hands, long neck, long body, imperfect, bad hands, signature, watermark, username, artist name, conjoined fingers, deformed fingers, ugly eyes, imperfect eyes, skewed eyes, unnatural face, unnatural body, error, two body, two faces”. **Checkpoint:** AbsoluteReality.
2. **Prompt:** “yellow-white-blue-purple-golden glowing mountains palace above the clouds, magical reality, high definition, 32K, dynamic lights, cinematic surrounding, intricate, natural lighting, ray tracing, bloom, extreme hdr, Octane render, unreal engine, 16K hyper realism, character design, hyper detailed, volumetric lighting, hdr, shining, vibrant, photo realism, Canon EOS 7D, Canon EF 70-200mm f/2.8L IS, vibrant colors, beautiful picture quality, breathtaking scene, focused”. **Negative prompt:** “Extra limbs, extra fingers, long neck, deformed, More than one nipple per breast, pointy nose, asian, japanese, chinese, lowres, disfigured, ostentatious, ugly, oversaturated, grain, low resolution, disfigured, blurry, bad anatomy, disfigured, poorly drawn face, mutant, mutated, extra limb, ugly, poorly drawn hands, missing limbs, blurred, floating limbs, disjointed limbs, deformed hands, blurred, out of focus, long neck, long body, ugly, disgusting, bad drawing, childish, cut off cropped, distorted, imperfect, surreal, bad hands, text, error, extra digit, fewer digits, cropped , worst quality, low quality, normal quality, jpeg artifacts, signature, watermark, username, blurry, artist name, Lots of hands, extra limbs, extra fingers, conjoined fingers, deformed fingers, old, ugly eyes, imperfect eyes, skewed eyes , unnatural face, stiff face, stiff body, unbalanced body, unnatural body, lacking body, details are not clear, details are sticky, details are low, distorted details, ugly hands, imperfect hands, mutated hands and fingers, long body , mutation, poorly drawn bad hands, fused hand, missing hand, disappearing arms, disappearing thigh, dis-

appearing calf, disappearing legs, ui, missing fingers, text, letters, illustration, painting, cartoons, sketch, worst quality, low quality, normal quality, lowres, bad anatomy, bad hands, monochrome, grayscale, collapsed eyeshadow, multiple eyeblows, vaginas in breasts, cropped, oversaturated, extra limb, missing limbs, deformed hands, long neck, long body, imperfect, bad hands, signature, watermark, username, artist name, conjoined fingers, deformed fingers, ugly eyes, imperfect eyes, skewed eyes, unnatural face, unnatural body, error, two body, two faces, hat, hats, pirate hat, headgear, helmet captains hat”. **Checkpoint:** DucHaiten GODofSIMP.

3. **Prompt:** “Fully In frame, 3D, centered, colorful, complete, poofy, single alien, fantasy Surreal Tree, 3d depth outer space nebulae background, 3D Art nouveau curvilinear lines, centered, no circles, no frame, dark atmosphere, curvilinear clouds, matte painting, deep color, fantastical, 3D intricate detail, complementary colors, 8k, cgsociety, artstation, hyperrealistic, cinematic, ultra hd, 4k, 8k, highly detailed cinematic global lighting, octane render, unreal engine 5”. **Negative prompt:** “Multiple trees, stacked trees, trees on top of trees, illustrations, split image, 2d, painting, cartoons, sketch, worst quality, low quality, normal quality, low res, monochrome, grayscale, error, bad image, bad photo illustration, 2d, painting, cartoons, sketch, worst quality, low quality, normal quality, low res, monochrome, greyscale, signature, watermark, error, bad image, bad photo black and white, monochrome, comic, text, error, cropped, letterbox, jpeg artifacts, signature, watermark, username, artist name, censored, worst quality, low quality, anime, digital illustration, 3d rendering, comic panel, scanlation, multiple views, lowres, ostentatious, ugly, oversaturated, grain, bad drawing, childish, cropped , worst quality, low quality, normal quality, jpeg artifacts, signature, watermark, username, blurry, details are not clear, details are sticky, details are low, distorted details, mutation, poorly drawn 2d, painting, cartoons, sketch, worst quality, low quality, normal quality, low res, monochrome, grayscale, cropped, oversaturated, signature, watermark, username, artist , error, bad image, bad photo deformed, distorted, disfigured, poorly drawn, bad anatomy, wrong anatomy, extra limb, missing limb, floating limbs, mutated hands and fingers, disconnected limbs, mutation, mutated, ugly, disgusting, blurry, amputation illustration, painting, cartoons, sketch, worst quality, low quality, normal quality, lowres, bad anatomy, bad hands, monochrome, grayscale, collapsed eyeshadow, multiple eyeblows, vaginas in breasts, cropped, oversaturated, extra limb, missing limbs, deformed hands, long neck, long

- body, imperfect, bad hands, signature, watermark, username, artist name, conjoined fingers, deformed fingers, ugly eyes, imperfect eyes, skewed eyes, unnatural face, unnatural body, error, two body, two faces Poorly drawn face, poorly drawn hands, poorly drawn weapons, unnatural pose, blank background, boring background, render, unreal engine”. **Checkpoint:** “AbsoluteReality”.
4. **Prompt:** “everything is burning, fire”. **Negative prompt:** “bad anatomy, bad proportions, blurry, cloned face, cropped, deformed, dehydrated, disfigured, duplicate, error, fused fingers, worst quality”. **Checkpoint:** “Stable-Diffusion v1.5”.
 5. **Prompt:** “A highly detailed alien landscape, alien buildings, weird colors, strange plants, xenomorphs, slime, oozing, HDR, 4k, volumetric lights, fantasy art, digital painting, beautiful, colorful, serene, intricate, eldritch, Nvidia ray tracing, Imax, slow shutter speed.”. **Negative prompt:** “cartoon, 3d, disfigured, bad art, deformed, extra limbs, close up, b&w, wierd colors, blurry, duplicate, morbid, mutilated, [out of frame], extra fingers, mutated hands, poorly drawn hands, poorly drawn face, mutation, deformed, ugly, blurry, bad anatomy, bad proportions, extra limbs, cloned face, disfigured, out of frame, ugly, extra limbs, bad anatomy, gross proportions, malformed limbs, missing arms, missing legs, extra arms, extra legs, mutated hands, fused fingers, too many fingers, long neck, Photoshop, video game, ugly, tiling, poorly drawn hands, poorly drawn feet, poorly drawn face, out of frame, mutation, mutated, extra limbs, extra legs, extra arms, disfigured, deformed, cross-eye, body out of frame, blurry, bad art, bad anatomy, double face, double face realistic, semi-realistic, cgi, 3d, render, sketch, cartoon, drawing, anime, cropped, worst quality, low quality, jpeg artifacts, ugly, duplicate, morbid, mutilated, out of frame, extra fingers, mutated hands, poorly drawn hands, poorly drawn face, mutation, deformed, blurry, dehydrated, bad anatomy, bad proportions, extra limbs, cloned face, disfigured, gross proportions, malformed limbs, missing arms, missing legs, extra arms, extra legs, fused fingers, too many fingers, long neck”. **Checkpoint:** Realistic Vision V2.
 6. **Prompt:** “no humans, scenery, flower A whimsical illustration of a rainbow connecting two vibrant and cheerful worlds, signifying the connection between people who find humor in life’s moments”. **Negative prompt:** “blur haze child, loli, paintings, sketches, worst quality, low quality, normal quality, lowres, normal quality, monochrome, grayscale, skin spots, acnes, skin blemishes, age spot, glans, mutated hands, poorly drawn hands, blurry, bad anatomy, extra limbs, lowers, bad hands, missing fingers, extra digit, bad hands, missing fingers worst quality, low quality, sketch, watermark, text copyright signature, cut out, cgi, hands, Two bodies, Two

heads, doll, extra nipples, bad anatomy, blurry, fuzzy, extra arms, extra fingers, poorly drawn hands, disfigured, tiling, deformed, mutated, out of frame, cloned face illustration, painting, cartoons, sketch, worst quality, low quality, normal quality, lowres, bad anatomy, bad hands, monochrome, grayscale, collapsed eyeshadow, multiple eyeblows, vaginas in breasts, cropped, oversaturated, extra limb, missing limbs, deformed hands, long neck, long body, imperfect, bad hands, signature, watermark, username, artist name, conjoined fingers, deformed fingers, ugly eyes, imperfect eyes, skewed eyes, unnatural face, unnatural body, error illustration, 3d, 2d, painting, cartoons, sketch, worst quality, low quality, normal quality, lowres, bad anatomy, bad hands, vaginas in breasts, monochrome, grayscale, collapsed eyeshadow, multiple eyeblows, cropped, oversaturated, extra limb, missing limbs, deformed hands, long neck, long body, imperfect, bad hands, signature, watermark, username, artist name, conjoined fingers, deformed fingers, ugly eyes, imperfect eyes, skewed eyes, unnatural face, unnatural body, error, bad image, bad photo worst quality, low quality, normal quality, lowres, bad anatomy, bad hands, vaginas in breasts, monochrome, grayscale, collapsed eyeshadow, multiple eyeblows, cropped, oversaturated, extra limb, missing limbs, deformed hands, long neck, long body, imperfect, bad hands, signature, watermark, username, artist name, conjoined fingers, deformed fingers, ugly eyes, imperfect eyes, skewed eyes, unnatural face, unnatural body, error, painting by bad-artist sketch, worst quality, low quality, normal quality, lowres, bad anatomy, bad hands, monochrome, grayscale, collapsed eyeshadow, multiple eyeblows, vaginas in breasts, cropped, oversaturated, extra limb, missing limbs, deformed hands, long neck, long body, imperfect, bad hands, signature, watermark, username, artist name, conjoined fingers, deformed fingers, ugly eyes, imperfect eyes, skewed eyes, unnatural face, unnatural body, error”. **Checkpoint:** DucHaiten StyleLikeMe.