

Comparison of CoModGANs, LaMa and GLIDE for Art Inpainting Completing M.C Escher’s Print Gallery

Lucia Cipolina-Kun
ML Collective and
University of Bristol. UK
lucia.kun@bristol.ac.uk

Simone Caenazzo
Riskcare Ltd., London

Gaston Mazzei
University Paris-Saclay. CNRS
LISN, VENISE team, Orsay, France.

Abstract

Digital art restoration has benefited from inpainting models to correct the degradation or missing sections of a painting. This work compares three current state-of-the-art models for inpainting of large missing regions. We provide qualitative and quantitative comparison of the performance by CoModGANs, LaMa and GLIDE in inpainting of blurry and missing sections of images. We use Escher’s incomplete painting Print Gallery as our test study since it presents several of the challenges commonly present in restorative inpainting.

1. Introduction

Artworks and images are part of our cultural heritage, but have a tendency to deteriorate over time. Inpainting is a restoration technique that has been applied traditionally to restore or complete the missing or damaged sections in a way that the restorative work passes unnoticed. In cases where the missing region is of considerable size, this task becomes delicate as the aim is to fill-in the area with content that ensembles well with the painting, whilst also fitting the painter’s style and historical period.

With the recent development of Machine Learning techniques, new inpainting models are available to the Cultural Heritage restorers. However, at present only few models are developed specifically with artwork restoration in mind. The training of these models requires dataset of images in the counts of thousands, a laborious and resource-intensive task *per-se*. The traditional solution is to fine-tune these models and re-train them with images similar to the restored piece; this is typically also a challenge, as it can be difficult to provide large sets of examples of relevant artwork.

Some examples of inpainting models specifically developed for art reconstruction include the works of Gupta *et al.* [10] and Amiri and Messinger [2], which both propose models derived from computer vision inpainting and ex-

tended to the art domain. Note that in both works domain experts were used to evaluate model performance, as an acknowledgment of the specific difficulty of evaluating inpainting in the specific context of art.

Our work aims to extend the current literature and provide an evaluation on how current state-of-the-art inpainting models can be used in an art restoration context. We offer a qualitative and quantitative comparison of three models developed for the inpainting of large missing sections, namely CoModGANs [21], LaMa [19] and GLIDE [1]. It is worth noting how none of these models was developed specifically for art reconstruction; however, given their versatility and simplicity, our aim is to show the context on which each of them can be successfully used as a restorative tool.

In order to stress-test the models in a challenging territory, we selected M.C Escher’s lithography, *Print Gallery*, as a test case. This work contains an entire missing region at the center where different semantic contents blend, thus being an excellent test case for inpainting models. Additionally, we compare the performance of each model in other well-known artworks, like the *Ecce Homo* by Elias Garcia Martinez and Escher’s ”Bird-Fish” used to highlight the weak and strengths of each model under different settings.

2. Inpainting Methods for Large Regions

The focus of our comparison is Computer Vision models developed specifically for the inpainting of large missing regions. In such contexts, the unmasked regions typically provide little information to guide the model towards the right choice of content, thus presenting additional challenges. Additionally, the larger the output required from the model, the more evident effects like pixelation and content mismatches can be. The three models selected are currently the state-of-the-art models for large-mask inpainting tasks.

CoModGAN. The model *Large Scale Image Completion via Co-Modulated Generative Adversarial Networks* (CoModGANs), implements a modulation of the unconditional image vectors into the traditional Generative Adver-

serial Networks to generate content consistent with the image’s semantics. It is based on the *StyleGAN* family of models [14], which allows to control salient features or styles of an image. To enhance performance on large-mask inpainting tasks, the model was trained using randomized large masks over the training datasets. A limitation of the current model distribution is the relatively low resolution required for input images: the model was trained on images of 512x512 size, requiring any other input image to match such size (thus potentially lowering the resolution of the overall output) when using the model. The model was trained on *Places2* [22], *CelebA-HQ* [17] and *COCO-Stuff* [3] datasets, making it versatile for a wide type of objects.

LaMa. The model *Resolution-robust Large Mask Inpainting with Fourier Convolutions* was designed with large regions in mind as well. It is a simple deterministic Pix2Pix-like model [13] with segmentation-based perceptual loss and a ResNet-like architecture with fast Fourier convolutions instead of the *StyleGAN* logic. The strength of the model is to target regular patterns in an image to repeat them across the masked region. The results of the model largely depend on the presence of regularities on the area surrounding the masked region. For example, in images with tiles, bricks and windows *surrounding* the mask. As an advantage over the others, this model is able to work with a higher resolution of 2048x2048. The model was trained only on two datasets, *Places2* and *CelebA-HQ*.

GLIDE. The model *Guided Language-to-Image Diffusion for Generation and Editing* is a multimodal diffusion model with text guidance. Diffusion models work similarly to upsampling models: the generator net is trained by progressively adding noise to an image and the learning objective is to revert the noise process, generating a de-noised image back. An additional component is the *text-guided* module, which allows the user to guide the image generation process by inserting a text prompt that acts like an additional constraint to the model. This prompt allows for virtually infinite possibilities in the number of outputs generated, while also avoiding the inconvenience of fine-tuning large models, as is the case of CoModGAN and LaMa. Additional model parameters such as the *guidance scale* and *temperature* allow the user to control the mix of conditional and unconditional outputs. An ablation study of GLIDE’s parameters is presented on the supplemental material. Resolution-wise, the released version of GLIDE accepts image inputs as large as 6Kx6K pixels; however, it then down-samples inputs to 64x64 for memory optimization and on the last stage it up-samples them back to 256x256, which is its final output resolution. The upsampling process, together with its training are the key to producing its claimed photorealistic quality. The model version released by OpenAI was trained on a filtered dataset excluding human figures from the MS-

COCO [16] dataset for images the and CLIP’s dataset for text [7].

Model	Type	Input size	Output Size
CoModGANs	StyleGan	512x512	512x512
LaMa	Fourier Conv	2048x2048	2048x2048
GLIDE	Text guided diff	6000x6000	256x256

Table 1. Comparison of model type, input and output sizes across models.

3. M.C. Escher’s Print Gallery

The artwork chosen for the present model testing exercise is *Print Gallery* (original title: *Prententoonstelling*), made in 1956 by the Dutch artist M. C. Escher. Figure 1 presents the original lithography, portraying a man that observes a painting in a gallery; the painting, in turn, portrays a gallery in the waterfront of the Grand Harbour of Valletta in Malta.



Figure 1. M.C. Escher’s lithography *Print Gallery* (*Prententoonstelling*), 1956. Image from Wikipedia Commons.

Print Gallery is a peculiar work for several reasons:

- It features a so-called *Droste effect* (i.e. a roto-homothety): the man stands in a gallery which is eventually portrayed again in the painting he is observing, creating a theoretical infinite loop;
- The painting is embedded in a spiral-like structure, clearly evident in the twisting of buildings, columns and other elements;

- Lastly, Escher did not complete the center of the painting, instead only adding his signature in the resulting blank - to this day, there is no definitive explanation for his choice.

The seemingly incomplete nature of the painting is arguably the main reason for the notoriety of *Print Gallery* among the artistic and mathematical community. Due to its challenging nature, several mathematicians and artist have attempted to complete it [5, 11]. In [5], Lenstra and de Smit present a class of exponential (conformal) complex maps [4] that share a similar shape with the spiral-like structure in the original painting. Such maps provide a bridge between a normal, undistorted space and the twisted space of the painting. Their work provides the mathematical foundation that we leverage upon in this present paper, together with Machine Learning techniques, to complete the center of the *original Print Gallery*. In particular, we show how the conformal map formulation can be used to pave the way for Computer Vision techniques - the performance of which we aim to compare as the main objective of our work.

3.1. Unrolling From Warped to Straight

It is worth noting that any attempt to apply Computer Vision/inpainting techniques *directly* on the blank of *Print Gallery* is faced with two main complications:

1. The painting, as described in the previous section, features a significant amount of twisting and rotation - Machine Learning models are, in general, not equipped to deal with extreme transformations in the sample image, since they are not equivariant to rotations, scaling and generally warping of images [9];
2. The size of objects to be completed in the center is very small in relation to the rest of the painting, once again creating challenges for any model trying to understand the sample image context.

The exponential maps described in [5] provides a solution, in that the twisted space in *Print Gallery* can be deconstructed into eight *straightened pictures in the Euclidean space* - each of which features an incomplete area in the shape of a spiral. This set of eight pictures are individually inpainted to complete the center.

The two equations below provide the mappings to first translate *Print Gallery*'s warped space into the Euclidean space (obtaining the eight straight images) and then back from Euclidean to warped. Let $z = (x, y)$ be the coordinates of RGB pixels in the complex plane, with (x, y) being standard Cartesian coordinates, and $T(z) : \mathbb{C} \rightarrow \mathbb{C}$ be the following complex exponential map:

$$T(z) = \exp^{\alpha L n(z)} \quad (1)$$

In the specific case of *Print Gallery*, a suitable value of the constant is: $\alpha = \frac{2\pi i + L n(256)}{2\pi i}$ [5]. Note that Equation 1 maps the straight Euclidean space into an *approximation* for the twisted space featured in the original *Print Gallery*. In order to map the twisted space into the straight one, we define the inverse map T^{-1} as below:

$$T^{-1}(z) = \exp^{\frac{1}{\alpha} L n(z)} \quad (2)$$

Note that Equation 2 describes a one-to-many mapping, as it is in fact periodic, with period $4^4 = 256$. As a result of the periodicity of the map, we obtain a set of eight straight images from *Print Gallery*, (which are shown in Figure 2), each of which is in relation to the next one via a zoom factor of 4.

The eight straight images obtained are used in the model comparison exercise as follows. First apply Equation 2 to *Print Gallery*, obtaining 8 straight sample images, second apply the tested models on the straight sample images, aiming to complete the spiral-shaped blank region, and lastly evaluate the performance of the models on each of the eight straight images. Summary metrics of our testings together with qualitative examples are presented in the following sections.

4. Model Comparison

Assessing the quality of an image depends very much on its context and usage. In the case of digital art, while the technical correctness of a restoration is important, there is an increased importance on *subjective* qualities of the restoration. We evaluated the three models using a group of subjective criteria such as: artistic consonance with the rest of the lithography; adherence to the painter's style and adherence of any new content to the historical period depicted in the artwork. Additionally, we compared model outputs using objective metrics traditionally used for *no-reference* image quality assessment.

4.1. Qualitative Analysis

As mentioned in Section 3.1, we tested the three models on the inpainting of the straight images in Fig. 2. For CoModGAN we used the demo provided¹ with the *Places 2* dataset. For LaMa, we used the demo provided with the *high-quality* setting². For GLIDE we used the Colab demo with a *Guidance Scale* of 4 and the text prompt "*Print Gallery*"³.

Fig. 3 shows an example output for each of the three models. The top-left image shows the target masked image. Note that the mask is placed in the left-most border of the

¹<https://github.com/zsyzzsoft/co-mod-gan>

²<https://cleanup.pictures/>

³<https://github.com/openai/glide-text2im/blob/main/notebooks/inpaint.ipynb>

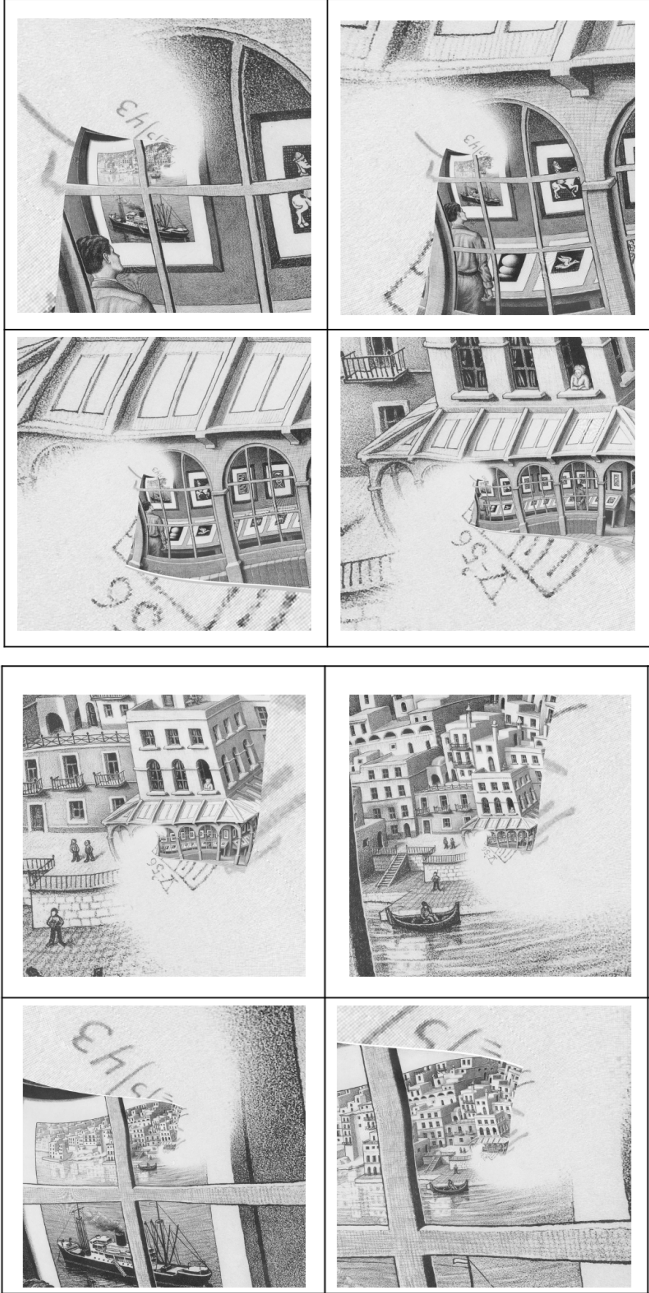


Figure 2. M.C. Escher’s lithography converted into eight straight images. The blank center appears as a white spiral on each image.

image, requiring the model to do inpainting as well as outpainting. This is an important observation, as CoModGAN and LaMa are models not natively suited for outpainting.

We now outline findings of the qualitative analysis in the form of conclusions.

Conclusion 1. The model output is significantly determined by the placement of the mask.

The three models evaluated are heavily dependant on the pixels surrounding the masked region. GLIDE and CoModGANs have a higher context awareness than LaMa. Besides the context, GLIDE is highly influenced by the prompt and other tunable parameters. An ablation study of GLIDE’s parameters is presented on the Appendix and the supplementary material.

Conclusion 2. GLIDE’s output is determined by the prompt, the seed and the guidance scale parameter, which determines the degree at which the prompt affects the output. For LaMa and CoModGANs, the only way to improve the output image is by performing costly fine-tuning.

Due to its multimodality, GLIDE can produce, in theory, an infinite number of outputs for the same mask, solely by changing the seed and the text prompt. This allows the user to rank the outputs or handpick the best inpainting solution for the context. The other models give a single output option per masked region, and thus, are more sensitive to the mask definition.

Conclusion 3. GLIDE is superior in outpainting (extrapolation) tasks when compared to LaMa and CoModGANs.

LaMa and CoModGANs are models developed for inpainting, this is, their output is primarily based on the information content read from the surrounding pixels of the mask. However, in outpainting, the mask extends beyond the borders of the image which leads the model with no surrounding information to work with. The images on the bottom show that LaMa and CoModGANs under-perform on outpainting tasks. This is in line with expectations, since none of them were developed specifically for outpainting.

Conclusion 4. GLIDE has a higher output variance, often producing uncanny objects.

Different from GAN models, GLIDE was not trained using a discriminator net, which is used to avoid the production of unrealistic artifacts. GLIDE on the other hand, is mostly text-guided, and as result, it produces a wide variance of surrealistic objects. In the artistic arena this diversity can be beneficial depending on the use case. The diversity of GLIDE’s output will be further analyzed on Sec. 4.3.

4.2. Detailed analysis of each model

This section analyses the results from each individual model in more detail. We present cases of both good performance and failures of each, with the aim of showing the aim is to show that each of these models specializes on different domains. To summarise, LaMa performs exceptionally well on image colors with well-defined patterns, CoModGANs is best suited for human faces and landscapes. As for GLIDE, while seems to be an all-terrain model, even

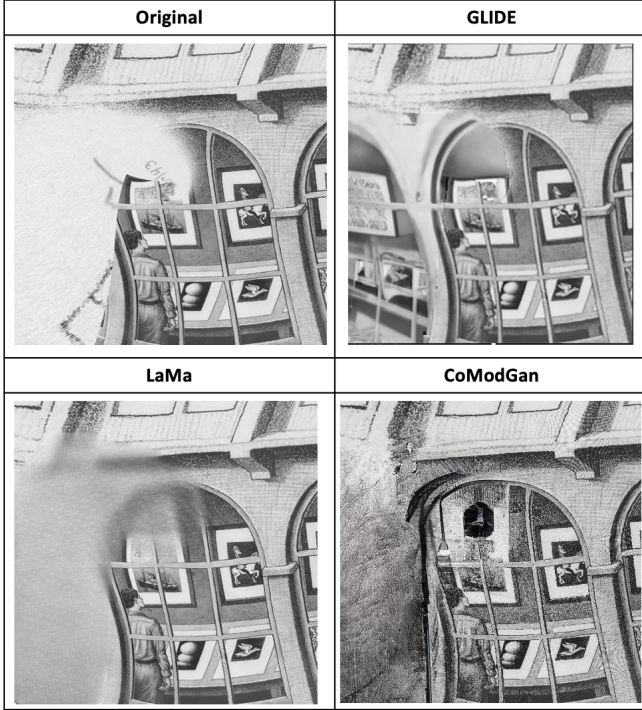


Figure 3. Example of model output for the same mask. The white area on the upper left image is the masked region.

capable of performing outpainting, its public release was filtered to not produce human figures. Additional comparison is presented in the supplementary material, where figures comparing the same failure cases across models are presented.

Fig. 4 below shows two examples of CoModGANs runs with the masked region boxed in red. The left image shows the limitations of the model on a simple outpainting task, where the natural expectation would have been for it to follow the color pattern. The image on the right shows instead a setting where the model performs very well as the model correctly learned to reproduce the buildings surrounding the mask. While the content generated is correct from a visual point of view, it is not in line with the painter’s style or the historical period of the painting, as the CoModGANs model has been trained on the modern (Places2) dataset. The way to shift the generation into a more suitable content is to fine-tune the net, which requires building a dataset of related artworks in the count of thousands, which is usually not available.

Fig. 5 below shows two examples of LaMa runs, again masked regions are shown boxed in red. The image on the left shows how the model fails on outpainting of images; in this particular case, the sample image presents a high degree of pixelation, making the recognition task harder. The output on the right shows a correct output, where the model

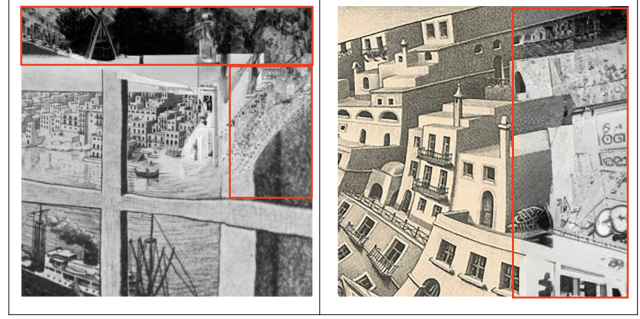


Figure 4. Example outputs of CoModGANs. The masked region is boxed in red. Note the graffiti painting produced by CoModGANs on the right image.

correctly identifies and mimics the pattern present in the surroundings of the mask. While the produced content is correct, the output still shows a certain degree of blurriness and pixelation.

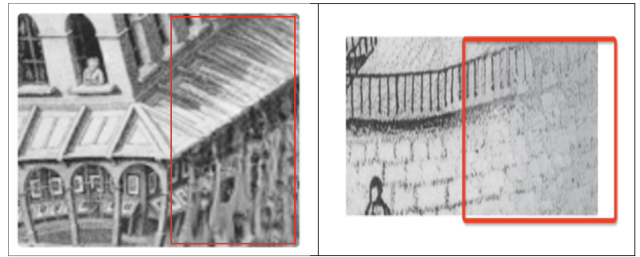


Figure 5. Example outputs from LaMa. Masked region boxed in red.

Fig. 6 below shows two failure cases of GLIDE. On the left image, the model tries to mimic the human figure and fails, producing additional inconsistent details. This is likely due to the fact that GLIDE’s training dataset does not contain humans, as a design choice. The image on the right shows a failure as a consequence of the model’s output variance which is further analysed on Sec. 4.3. We can see how the model produces unrealistic objects, which have no resemblance with a particular object on its training set. This could be explained by the fact that the model does not contain a discriminator network, as the output is only guided by the cosine similarity with the text prompt.

4.3. Analysis of GLIDE’s Output Diversity

As explained before, GLIDE’s distinctive feature is its multimodality, it takes as input a masked image with a text prompt and produces a (theoretically) infinite supply of inpainting options. This creates the problem of image selection; it is not clear a priori, how many batches of images are needed to find the best inpainting option and additionally, there is no selection metric provided with the model.

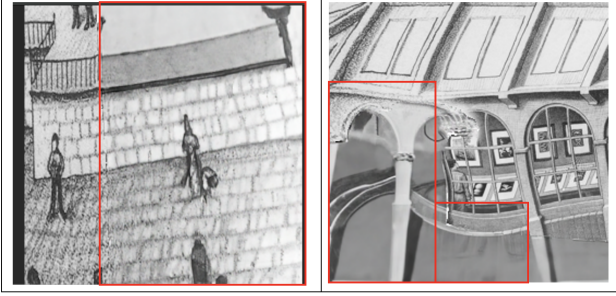


Figure 6. Example outputs from GLIDE. Masked sections boxed in red. Guiding Scale of 5

An example of the diversity of GLIDE’s output is shown below. We generated samples for the same mask, prompt and seed. We can see that the output is very dissimilar among the images selected and in a sense uncanny with the expectations for an Escher painting. Note that here we analyze dissimilarity over the content created, and not on image quality.

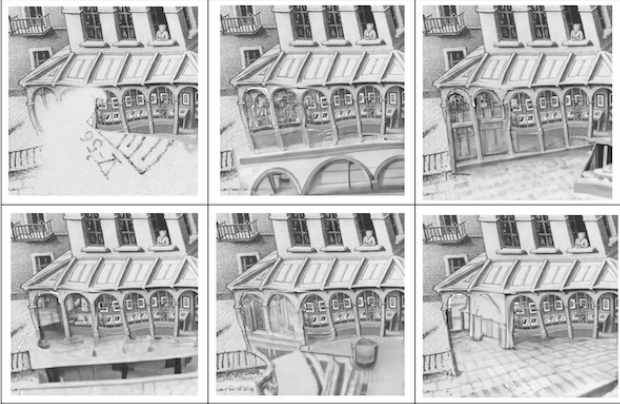


Figure 7. Examples of inpainted images generated by GLIDE for the caption “a gallery with arches wooden windows and arcades and floors with tiles” and a Guiding Scale of 5. The white area in the top-left image is the masked region.

To measure the diversity of the inpainted content created in an objective way, we calculated the CLIP score over 250 random samples of the top-left image in 7, using the same mask and prompt⁴. The CLIP score measures the cosine similarity between the text prompt and the output image [7], a higher text prompt means the content created resembles better the passed prompt. While the Coefficient of Variation of the CLIP score is only 3.62%, in visual terms, this variation translates into very significantly distinctive content. Additional outputs are shown on the supplementary material.

⁴The prompt used is “A man looks at a painting of Malta behind the windows of a gallery”

4.4. Analysis on Different Paintings

This section shows the performance of the models under alternative settings other than the eight straight images obtained from Print Gallery⁵. The main conclusion is that each of the analyzed models has been developed and trained for a specific use-case and there is no model that outperforms the others across the board, when it comes to qualitative assessment.

The Figure 8 shows a painting with clear color patterns where LaMa’s performance is the strongest as expected for a Fourier-based model. In fact, the only difference with the original image is the detail of the reconstruction of the eyes. GLIDE shows good results however, LaMa’s output is at 2048x2048 while GLIDE is only able to provide a quality of 256x256⁶.

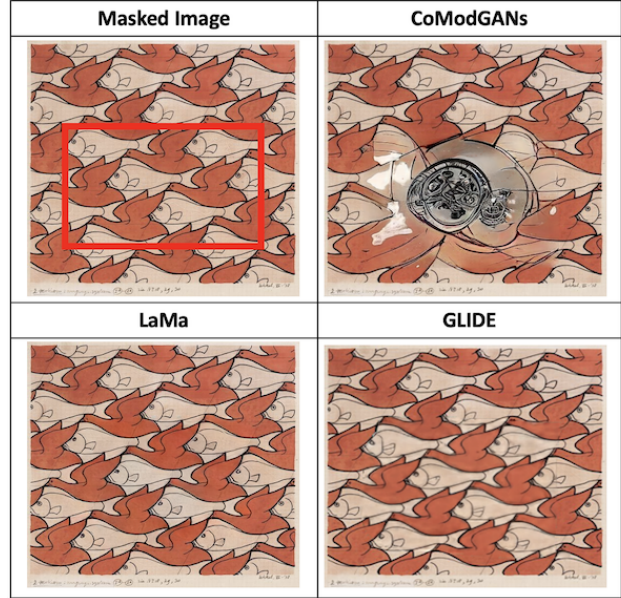


Figure 8. M.C. Escher’s Bird-Fish painting. 1938. Comparison of performance of the three models over a regular painting. The masked region is the entire square area delimited in red. Image reproduced under WikiMedia Commons.

The image on Figure 9 shows the limitations of GANS-based models on digital restoration. In particular CoModGANs is trying to blend the masked region with the neighboring colors, missing the context, as is a feature of the localized convolution of GANs. While LaMa succeeds on the face part, it fails on the lower part of the image. GLIDE’s outputs varies with the Guidance Scale parameter, however, none of the outputs is able to recognize the

⁵Additional examples and a longer analysis is presented on the Supplementary Material section.

⁶GLIDE’s prompt used is simply “pattern” and the Guidance scale is five. A low guidance scale helps the model to favor the image’s semantics over the text prompt

feature of a human face as its distributed version has been restricted to not produce humans ⁷.

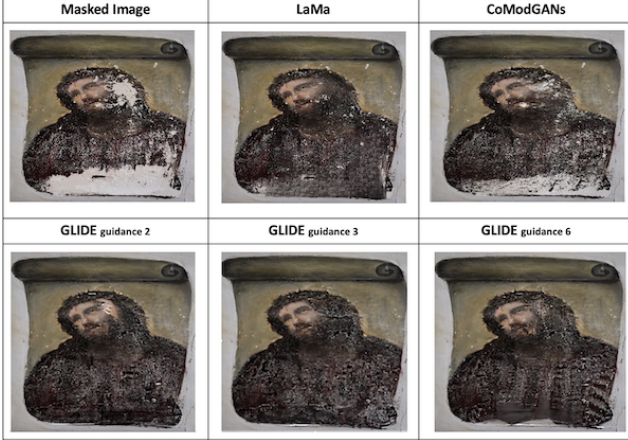


Figure 9. Sample outputs from different models on the Ecce Homo fresco by Elias Garcia Martinez. GLIDE’s outputs are presented for Guidance Scales of two, three and six. Image reproduced under WikiMedia Commons.

4.5. Quantitative Metrics

We used three different metrics to provide a quantitative comparison of the models’ outputs as shown in Tab. 2. The selected metrics are commonly used in the field of *no-reference* image quality assessment, where the quality of an image is determined without using any target image for comparison. In our case, each image was evaluated as a stand-alone output. The model *Koniq* produces a score by comparing the input image against the largest dataset of image quality up to date [12]. The model *BRISQUE* reports a score using a Support Vector Regression trained on an annotated image dataset with known distortions [18]; such dataset is, however, biased towards landscape pictures. Lastly, we used the *DOM* [15] model which gives a score based on the sharpness of gray images.

To obtain a diverse sample of images, we tested the models across the eight straight images in Figure 2, which contain large regions of inpainting and outpainting challenges. We created 50 different random masks on each model and used the same mask across models. The use of 50 masks is justified by an ANOVA test presented on the Appendix in Sec. 8.

Analysing Table 2 we can see that in all cases GLIDE shows a superior performance, except for the DOM score, which shows GLIDE almost matching with CoModGANs on sharpness ⁸. The good performance of GLIDE on the

⁷GLIDE prompt used is "a man staring like Jesus with shirt red and black stripes".

⁸GLIDE was run with a Guidance Score of 5 and Upsample Tempera-

Koniq and BRISQUE scores are in line with the recent literature showing that, in general, diffusion models beat GANs on image synthesis [6]. This result can be explained by several factors. First the upsampling module present on GLIDE’s acts similarly to a denoising feature creating a uniform density of pixels across an image.

Conclusion 5. GLIDE presents superior performance on blurriness and deformation while not on image sharpness. However its performance is dependent upon the parameter tuning.

Method	Koniq ↑	Brisque ↓	Dom ↑
CoModGANs	36.12	43.37	1.05
LaMa	38.76	42.38	1.10
GLIDE	41.61	7.94	1.04

Table 2. Average values for each metric. A higher Koniq score is better, a lower Brisque score is better and a higher DOM (edge sharpness) score is better.

5. Print Gallery Inpainting Result

Figure 10 below displays the result of Print Gallery completed by performing three steps. First we applied Eq. 2 obtaining the eight straight images in Figure 2, second we completed the missing region of each using GLIDE ⁹, and lastly we combined the eight straight images as in Eq. (1) to obtain back Print Gallery.

Figure 11 displays a zoom-in of the completed center. It is noticeable some mismatch between the boundaries of the warped straight images, this is due to the difference in Escher’s original lithography and the parametrized mappings applied in Eq. (2) and Eq. (1). To correct for this, a future direction is presented on Section 7. Note how the inpainted region is very small and rotated for any inpainting model to be used out of the box (i.e. without any fine tuning or passing to the Euclidean plane)¹⁰. Additionally, as a consequence of the one-to-many mapping in Eq. (2) the center presents an homothecy of Print Gallery itself, rotated by 157 degrees.

6. Conclusions

We have provided a quantitative and qualitative analysis of three of the current state-of-the-art models for inpainting on large masks. By using a particularly challenging setting, comprised of a mixture of inpainting and outpainting

ture of 0.997. The Supplemental material shows further analysis of GLIDE on the relationship between its parameters and the DOM score

⁹The parameters used and additional details of the completion process can be found on the supplementary material

¹⁰The supplementary material shows an alternative completions made by hand by professional artists

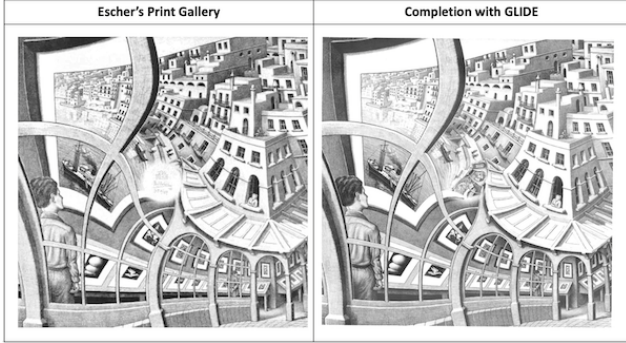


Figure 10. Comparison between original Print Gallery and our completion using GLIDE

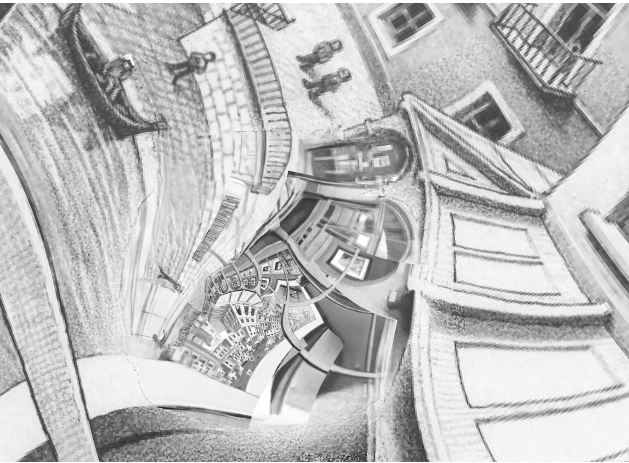


Figure 11. Detail of the completed center using GLIDE. In the center-bottom it shows the repetition of the original rotated by 157 degrees.

modalities over different images, we have obtained test-case results for each model’s strengths and weaknesses. GLIDE appears to be superior to LaMa and CoModGANs on outpainting tasks and it is benefited from an upsampling module obtaining photorealistic quality. Additionally, GLIDE provides the user with alternative completions for a given mask and prompt, which can be beneficial on artistic settings and allows one to calibrate the output result without the costly fine-tuning required by the other two methods. However, GLIDE’s output diversity can also lead to unrealistic outputs and thus, requires human discretion to select the best fit. We have shown how. According to expectations, LaMa was shown to be superior in pattern-replication tasks, and it has the best resolution output across all models. As for CoModGANs, similar to the family of StyleGANs models, it shows best performance on big masks over human faces and landscapes, since it was specifically trained on them, while GLIDE’s dataset filtered out human images.

7. Future Work

As mentioned in Section 3.1, the formulas in Eq. (2) and Eq. (1) have been used to translate the original Print Gallery lithography into eight straight images. This is, however, an imperfect process due to the natural differences between a hand-made process and any attempt to parametrize it with closed-form formulas. To address this difference, we propose to project Escher’s Print Gallery onto the conformal map space, for example using Thin Plate Splines (TPS) [8].

8. Appendix

To test for the statistical significance of the 50 means on table 2, we performed a one-way ANOVA test summarized below on Table 3. We can conclude that the average values presented on table are statistically different across all metrics, notwithstanding DOM which presents similar results for CoModGANs and GLIDE.

Method	Fvalue	Fcrit	RH0
Koniq	9.23	3.05	yes
Brisque	255.6	3.05	yes
DOM	74.17	3.05	yes

Table 3. Results of the ANOVA test performed over the mean results of the image quality metrics

9. Acknowledgements

We thank the anonymous reviewers for their incisive comments that were most useful in revising this paper. Additionally, the authors would like to thank Rosanne Liu, Cris Luengo, Roman Suvorov, Vaisakh M, Andrea Panizza, Alejandro Cabrera, Jim Schmitz, Pablo Samuel Castro, Meire Fortunato, Pablo Sprechmann, Rick Anderson, Niranjana Krishna and Vahid Yazdanpanah.

10. Supplementary Material

10.1. Further Qualitative Results

The panels 12,13,14 present additional comparison of the ‘failure’ cases across different models. It is clear that LaMa and CoModGans are not suited for outpainting tasks. GLIDE on the other hand, performs well across inpainting and outpainting demands but suffers from having the worst resolution output at only 256x256.

10.2. Additional Studies on GLIDE’s Parameters

The panel in 15,16,17 shows the summary results of ablation studies performed on GLIDE. We tested the effect of changing the Guidance Scale (which controls the relationship between the prompt and the generated image) and the

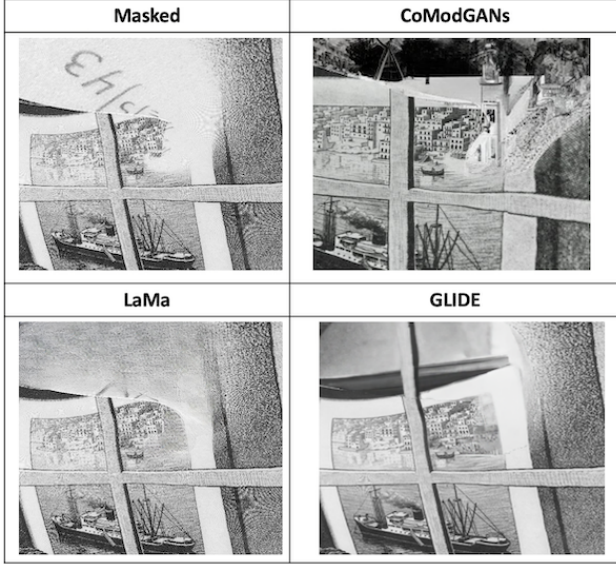


Figure 12. Comparison of outpainting image across different models. GLIDE caption is "window" and Guidance Scale of five.

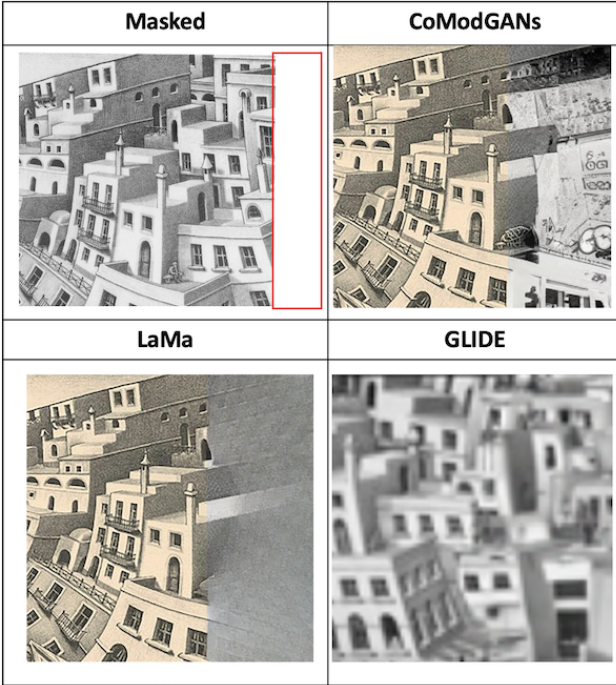


Figure 13. Comparison of outpainting image across different models. GLIDE caption is "buildings" and Guidance Scale of five.

Upsampling Temperature (which controls the degree of up-sampling) for the same image and same mask. The test was performed over 50 samples for each value ¹¹.

¹¹prompt: "a gallery with arches wooden windows and arcades floors with tiles"

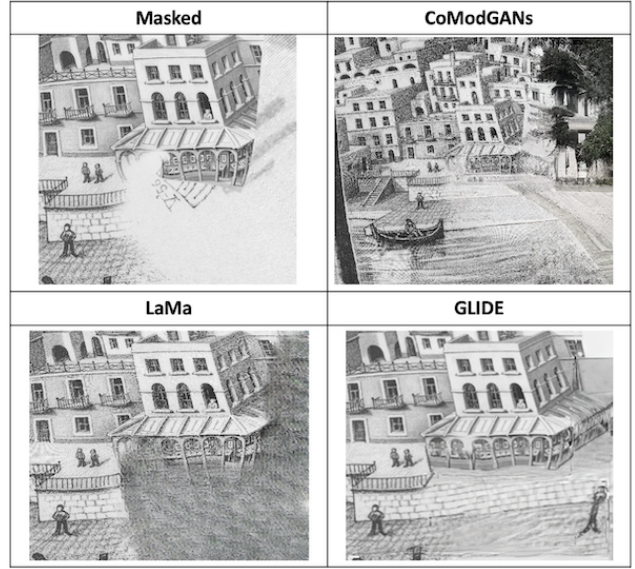


Figure 14. Comparison of outpainting/inpainting image across different models. GLIDE caption is "buildings" and Guidance Scale of five.

The results show a significant sensitivity of the model outputs to the parameters. As expected, a higher degree of upsampling improves results across all metrics. In fact, the recommended Upsampling Temperature is 0.997. The Guidance Scale controls the content, and thus, does not affect the sharpness or blurriness. In the case of BRISQUE, the sensitivity to the Guidance Scale might be explained by the degree of black color on the image, since this metric is sensitive to large black regions.

Next we analyse the effect of the prompt on the inpainted content. As mentioned, GLIDE is a text-guided model where the incidence of the text is controlled by the parameter "Guidance Scale". Under a low Guidance Scale, GLIDE produces content in consonance with the surrounding objects, which is ideal for art inpainting. On the contrary, a higher Guidance Scale gives more weight to the text prompt on the image generation. The effect of the Guidance Scale over the text guidance is shown on Figure 18 using a fixed seed and an Upsampling Temperature of 0.997.

10.3. Detailed Analysis of Inpainted Print Gallery

The panel 19 shows a side-by-side comparison of the eight images composing the straight version of Print Gallery with their inpainted counterparts. On each panel, the four images on the right are the inpainted results of the left images.

On panels 21 we show a comparative analysis of GLIDE's output under different Guidance Scales, everything else equal. Results indicate that a Guidance Scale

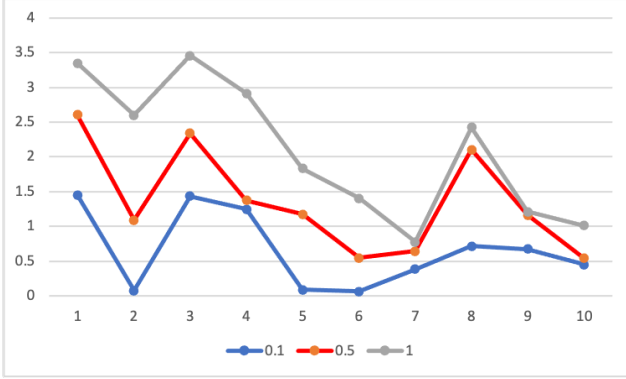


Figure 15. Plot of BRISQUE values. The three categories represent the Upsampling Temperature and the horizontal axis represent the Guidance Scale

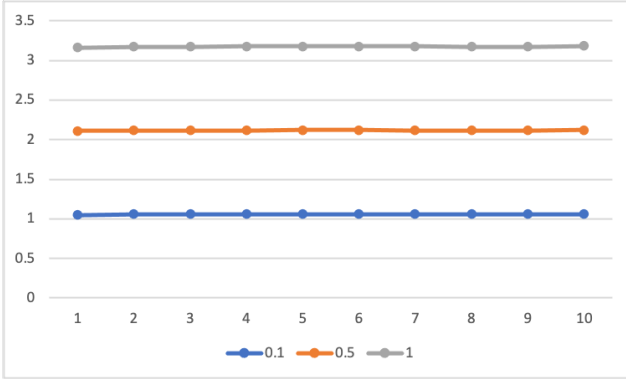


Figure 16. Plot of DOM values. The three categories represent the Upsampling Temperature and the horizontal axis represent the Guidance Scale

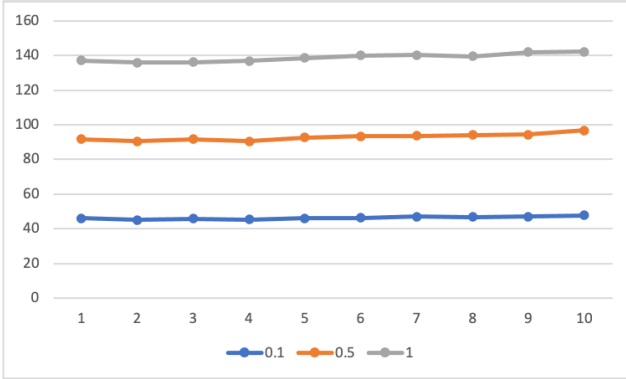


Figure 17. Plot of KONIQ values. The three categories represent the Upsampling Temperature and the horizontal axis represent the Guidance Scale

value of five results in the best object creation ¹². The

¹²Images by artist @litevex reproduced with permission

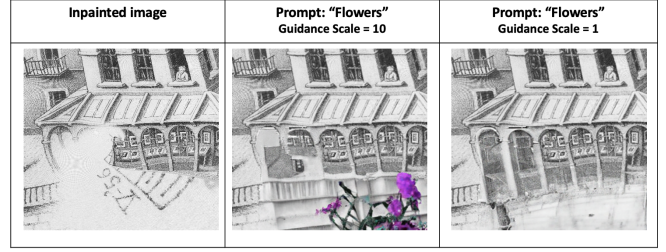


Figure 18. GLIDE's outputs for different Guidance Scales and same prompt.

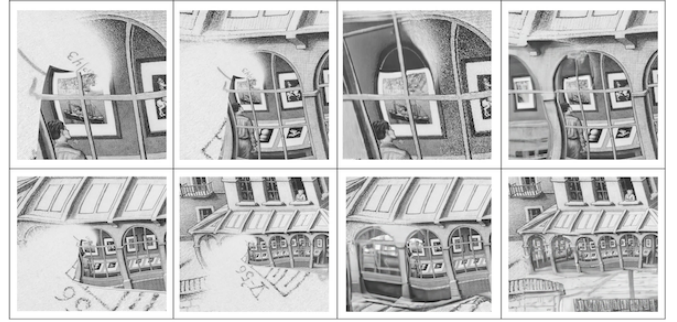


Figure 19. Inpainting of the first four images composing the blank center. GLIDE with Guidance Scale of five and prompt "Escher".

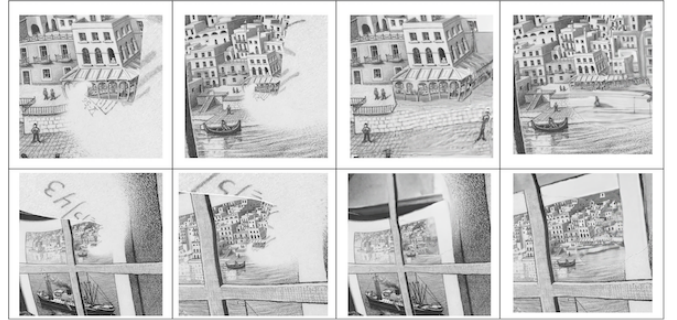


Figure 20. Inpainting of the second four images composing the blank center. GLIDE with Guidance Scale of five and prompt "Escher".

prompt used is: "a photograph of a teddy bear using a laptop 1080p 4k."

10.4. Analysis on Additional Paintings

This section provides further comparison of the model performance under different scenarios. First in Fig. 22 we provide a further analysis of the Ecce Homo by Elias Garcia Martinez from the main text. This fresco is a challenge since white patches from the degradation are visible all over the surface. In the case of CoModGANS, the model considers as a valid pattern the white areas appearing on the image's surface and tries to replicate them, resulting in a



Figure 21. Different guidance scale settings over the same prompt. "a photograph of a teddy bear using a laptop 1080p 4k".

poor inpainting performance. LaMa repeats this behaviour but in a lesser degree, as noted, this model provides the best output resolution among all. Additionally, it performs comparatively well on the face since it has been trained on the faces database Celeb-H. GLIDE, as expected, it is not able to recognize a human face by design. However, the upsampling module produces a nitid result compared to others.



Figure 22. Different outputs for the Ecce Homo by Elias Garcia Martinez. Image from Wiki Commons. Glide was used with a Guidance Scale of three and six, where indicated.

10.4.1 Automatic Generation of Prompts

The Fig. 23 is a work by Torres-Garcia "Composicion constructiva" (1932) [20]. The piece was burnt at the Brazilian

Museum of Modern art in 1981 and presents the traces of fire on the wood. This coloration creates a challenge for inpainting models as they deem the burnt area as a valid pattern to reproduce. We see that GLIDE is able to move away from the burnt colorization, either by ramping the Guidance Scale or by trimming the area with a clever prompt. For the left-bottom image, the prompt was generated using a image-to-text bot by the developers EleutherAI¹³, the generated prompt is "Paul Klee's rectangular piece of wood is a rare example of early Christian art.". The images with Guidance Scales two and 20 have used the simple prompt "patterns".

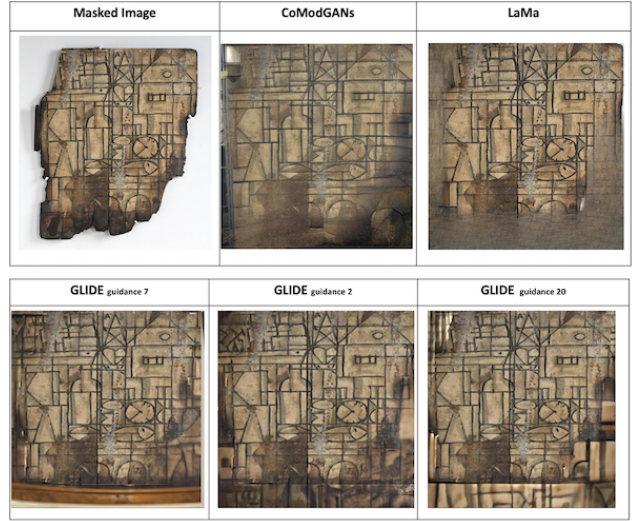


Figure 23. Inpainting outputs for the work of Torres-Garcia. Composicion constructiva (1932). GLIDE with different Guidance Scales and prompts generated by image-to-text bots. Image from Wiki Commons.

10.4.2 Crowd-sourcing of Prompts By Experts

The panels on 24 show Cezanne's unfinished "Turning Road" (1905), which has whole sections of the canvas bare. The inpainting of this work is more open-ended given the style of the painter. For this reason, all models present equivalent performance, to the casual eye. To generate GLIDE's prompt we did a crowd-sourcing experiment and relied on the expertise of ten visual artist from EleutherAI who suggested by consensus "the fog of the valley painting from last century". The two GLIDE images on the bottom were generated by using different seeds on the same prompt.

References

- [1] Nichol Alex, Dhariwal Prafulla, Ramesh Aditya, Shyam Pranav, Mishkin Pamela, McGrew Bob, Sutskever Ilya, and

¹³www.eleuther.ai

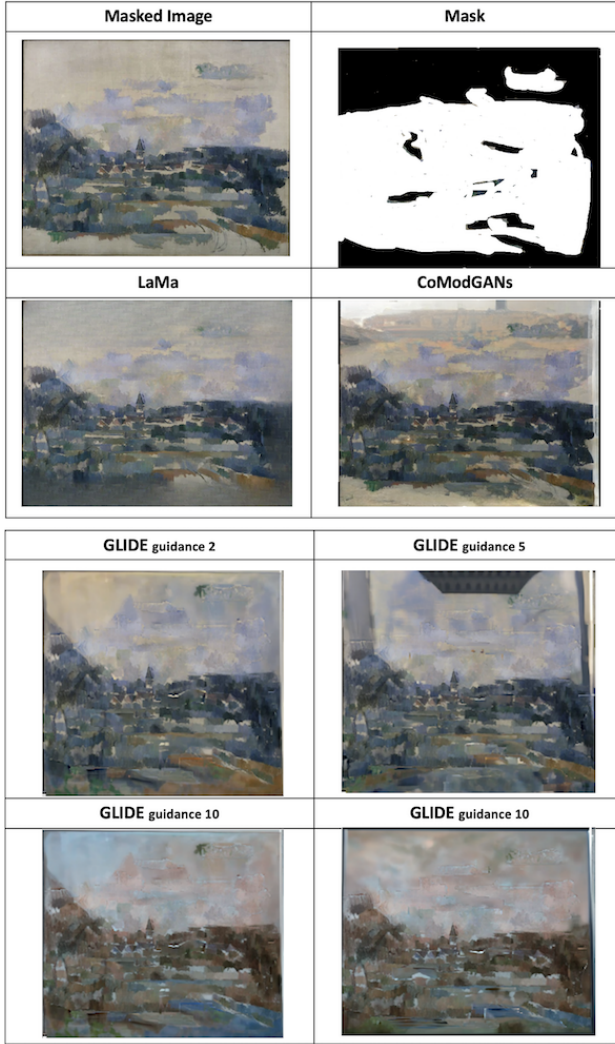


Figure 24. Comparison of inpainting models for Cezanne’s unfinished “Turning Road” (1905). GLIDE’s output is presented with several values for the Guidance Scale. Image from Wiki Commons.

chen Mark. Glide: Towards photorealistic image generation and editing with text-guided diffusion models.

- [2] M. Amiri and D Messinger. Virtual cleaning of works of art using deep convolutional neural networks.
- [3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context, 2016.
- [4] P. Ciarlet. *Handbook of Numerical Analysis*. North-Holland, 2000.
- [5] B. de Smit and H. W. Lenstra Jr. The Mathematical Structure of Escher’s Print Gallery. *Notices of the AMS*, 50(5):446–451, 2003.
- [6] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021.
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

- [8] Jean Duchon. ”Splines Minimizing Rotation-Invariant Semi-Norms in Sobolev Spaces”. In *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer Berlin Heidelberg, 1977.
- [9] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [10] Varun Gupta¹, Nitigya Sambyal, Akhil Sharma¹, and Praveen Kumar. Restoration of artwork using deep neural networks varun.
- [11] D. Hofstadter. *Gödel, Escher, Bach : an Eternal Golden Braid*. Basic Books, 20th anniversary edition, 1979.
- [12] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment.
- [13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.
- [14] T Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks.
- [15] Jayant Kumar, Francine Chen, and David Doermann. Sharpness estimation for document and scene images. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3292–3295, 2012.
- [16] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014.
- [17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [18] A. Mittal, A. Moorthy, and A. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21:4695–4708, 2012.
- [19] 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. *arXiv preprint arXiv:2109.07161*, 2021.
- [20] Joaquin Torres García. *Torres García. Obras Destruídas En El Incendio Del Museo De Arte Moderno De Río De Janeiro*. Fundación Torres García, 1981.
- [21] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large Scale Image Completion via Co-Modulated Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2021.
- [22] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.