
Discriminative Candidate Selection for Image Inpainting Applications to the Fine Arts

Lucia Cipolina-Kun¹ Sergio M. Papadakis¹ Simone Caenazzo²

Abstract

Within the field of Cultural Heritage, image inpainting is a conservation process that fills in missing or damaged parts of an artwork to present a complete image. Multi-modal diffusion models have brought photo-realistic results on image inpainting where content can be generated by using descriptive text prompts. However, these models fail to produce content consistent with a particular painter's artistic style and period, being unsuitable for the reconstruction of fine arts and requiring laborious expert judgement. Moreover, generative models produce many plausible outputs for a given prompt. This work presents a methodology to improve the inpainting of fine art by automating the selection process of inpainted candidates. We propose a discriminator model that processes the output of inpainting models and assigns a probability that indicates the likelihood that the restored image belongs to a certain painter.

1. Introduction

Inpainting techniques are used to restore or complete missing or damaged sections of a painting. The aim of these techniques is to continue the artwork inferring what could have been in the place of the missing region such that the restorative work passes unnoticed. The traditional inpainting techniques rely upon the interpolation of the neighboring pixels of a masked region in order to obtain a smooth continuation within the existing and new parts. With the recent development of Machine Learning techniques, new inpainting paradigms have been made available to Cultural Heritage restorers.

A myriad of techniques have been long developed in the field of computer vision for the inpainting of images (Jam-

¹ML Collective ²Riskcare, Ltd, London. Correspondence to: Lucia Cipolina-Kun <lucia.kun@bristol.ac.uk>.

et al., 2021). The prolific production of inpainting models released by the computer vision community offers different inpainting solutions that can accommodate different use cases. However, all these models share a similar calibration methodology, training on thousand of examples until convergence to the desired model behaviour. In particular, among all the proposed models, diffusion models such as GLIDE (Nichol et al., 2021) have attained success due to its impressive photo-realistic results on image generation as well as inpainting. In this work, we use GLIDE as an example of multi-modal generative modelling since its code has been open sourced.

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, as 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* CLIP (Radford et al., 2021) 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. The text gradients guide the model into generating the image that best resembles the text, on the basis of a highest cosine similarity criterion between the image and the text. This prompt allows for virtually infinite possibilities in the number of outputs generated, without having the inconvenience of fine-tuning large models, as is the case with traditional GAN-based models.

Like any generative model, GLIDE has the possibility of generating a wide and diverse number of samples (i.e. inpainted images) from the trained distribution. In a theoretical sense, for any single given text prompt and masked image, GLIDE produces an infinite supply of inpainting options. An example of the diversity of GLIDE's output is shown in Figure 1. The Figure shows a rectified version of M.C. Escher's Print Gallery which will be used as a use-case example (de Smit & Lenstra Jr, 2003) since it contains a wide blank area that requires inpainting. 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

¹.


Figure 1. Examples of images generated by GLIDE for the caption “*a gallery with arches wooden windows and arcades and floors with tiles Escher*” along their Escher-likeness. The white area in the red boxed image is the masked region.

Diffusion models are known for their great variability and diversity for image generation (Kawar et al., 2022), which creates the problem of image selection. One option is to manually classify GLIDE’s inpainting output via human experts; however, this process is laborious and not scalable. The objective of our work is to have a pre-trained classifier aid in the image selection process. Once the classifier is trained to match the style of a certain artist (or a certain school of painting), it can be used to automate the selection of outputs. This allows the restorer to narrow down the number of images to choose from and pick only among the best pre-classified ones. Eventually, the classifier can be used by a wider audience of non-experts.

In order to aid the selection process among the (potentially infinite) samples generated by the model, we propose to append a trained discriminator/classifier as a head to the GLIDE architecture. We used a discriminator architecture similar to (Dhariwal & Nichol, 2021) but instead of adding the discriminator *inside* the diffusion model, we placed it as the final unit in the architecture. This allows us to have a discriminator tool for *any* diffusion model, regardless of whether the code is available to the public or not.

The aim of our model is to help in selecting the inpainted image that best corresponds to the input text-prompt as well as, best matches the inpainted image in an artistic sense. By a match in *artistic sense*, we refer to content that:

- Harmonizes well with both the content and style of the artwork being inpainted; and

¹The depicted inpainting alternatives provided by GLIDE for the white masked region were deemed uncanny by human experts

- Complies with additional desiderata like matching with the artist’s style, their historical period, and can in fact pass as an original to connoisseurs in the artistic domain.

Contributions In summary, our contributions are as follows:

1. We propose a method to aid the selection of inpainted images based on an automatic classification of images.
2. The method is based on a discriminator trained from a dataset of artistic images related to the use-case.
3. Our method can be applied regardless of the white-box or black-box access to the generator’s code.

The proposed method is summarized in Figure 2 below.

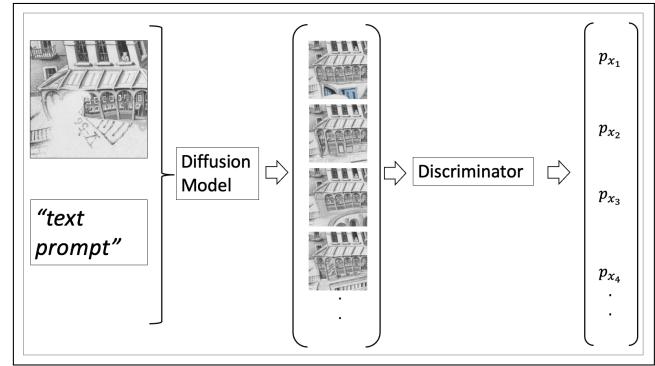


Figure 2. Proposed model architecture. The input to the model is a masked image to inpaint and a prompt to guide the inpainted content. The diffusion model generates N number of inpainted alternatives and the discriminator assigns a probability to each indicating the likelihood of belonging to a certain artist.

2. The Discriminator Module

The objective is to aid the inpainting process in selecting the best output among the many produced by GLIDE. By *best* we mean, the inpainted image that best matches the inpainted region in content and artistic style, as explained before. For this, we propose to train a discriminator network that assigns a probability score to each inpainted image, which accounts for the likelihood that the image is an original production of a given artist. Next we describe the discriminator’s architecture and the training dataset.

2.1. Discriminator Architecture

The discriminator role is to act as a classifier. We used the same architecture as in DCGANs (Radford et al., 2015). The classifier is composed of:

- Five convolutional layers, with Leaky ReLU activation functions and increasing window sizes from 64 up to 512 pixels;
- A final softmax layer for classification.

2.2. Training Dataset

The discriminator requires a definition of "good/valid" and "bad/unvalid" outputs. In our case, we used a combination of Escher images and expert knowledge to create a sample of 400 images classified into a balanced sample of "Escher/non-Escher" categories. This dataset was passed to train the discriminator, which is relevant for our particular use-case. In a more general case, the discriminator can be trained from a batch of images of a specific painter, which can be retrieved from Wiki-art² and used as a dataset. Additionally, style-transfer techniques (Gatys et al., 2015) can be used to augment the dataset with images that resemble the particular style of a painter.

3. Results

Once a batch of different inpainting options has been produced by GLIDE, it is passed to the discriminator, which assigns a probability score to each image indicating the likelihood that it corresponds to an original Escher.

The aim is that the images produced will have a better visual adherence with the artist and also less variance among the results. In practical terms, this results in increased similarities between the images selected by the discriminator, and increased similarities with respect to the style of the painter upon which the discriminator was trained.

We have tested the results of the discriminator classification in a quantitative and qualitative fashion, as described below.

3.1. Qualitative Assessment of the Discriminator

In Figure 3, we observe that the classifier guidance is capable of producing images with less variance among them, if compared against the outputs in Figure 1. Also, the objects produced are more aligned with the context of the image, i.e. there are no uncanny objects or artifacts in the selected images. Additionally, as expected, images with high score have a better alignment with the style of Escher's body of work.

3.2. Quantitative Assessment of the Discriminator

To evaluate the proposed model's performance in classifying images according to a desired artistic style, in this case, M.C. Escher's work, we conducted a human study. Similar to the one performed by the authors of GLIDE, we asked

Masked image	100%	99%
90%	80%	100%

Figure 3. Examples of inpainting options generated by GLIDE alongside their Escher-Likelihood by the discriminator. The caption used is "*a gallery with arches wooden windows and arcades and floors with tiles Escher*".

ten human evaluators to judge the six inpainted images displayed on Figure 4 below.

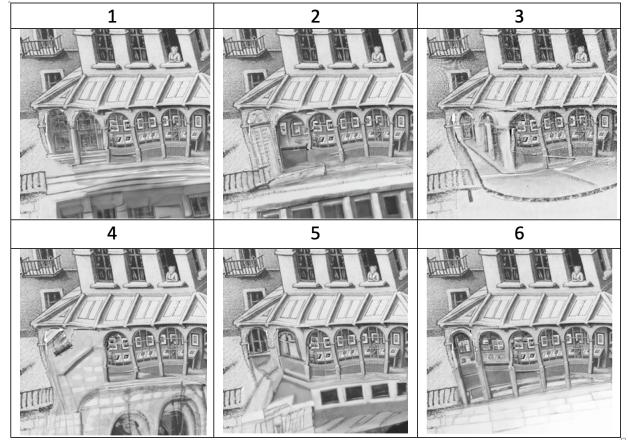


Figure 4. GLIDE outputs for a given masked image and prompt used for the human study. Participants were asked to assign a probability to each image of being a real Escher.

Human evaluators were asked two questions:

1. Assign a probability to each image representing the likelihood that it was painted by M.C. Escher
2. Assign a score to your own knowledge of M.C. Escher's work from 1-6. This is, how much of an Escher connoisseur you are.

The assigned probabilities were weighted by the connoisseur

²<https://www.wikiart.org/>

score to produce a final mean probability accounting for the expertise of each human evaluator. As follows.

$$score_i = \frac{w_j \cdot p_{x_{i,j}}}{\sum_{j=1}^{10} w_j} \quad (1)$$

where $i \in (1, 6)$ is the index for the evaluated image, $j \in (1, 10)$ is the index for the human evaluator, w_j is the connoisseur score auto-assigned by the evaluator, and $p_{x_{i,j}}$ is the Escher-likeness assigned by individual j to image i .

Lastly, we compared the likelihood assigned by the evaluators against the discriminator's probability. Results are shown in Table 1

Table 1. Probability scores assigned by human evaluators and the discriminator model for each inpainted image

Image Idx	Human	Model
1	0.33	0.0
2	0.34	0.0
3	0.56	0.88
4	0.35	0.0
5	0.33	0.0
6	0.37	0.98

We can see that the human evaluators assigned a high score to images 3 and 6, similar to the discriminator. However, in a much lower score. It is worth noticing that the median of the connoisseur score for the human evaluators is 2.5 meaning that the evaluator's self-assigned knowledge of the works of M.C Escher is below average. In the case of evaluators with high knowledge of the painter, the score is closer to the discriminator's score. Full results are shown in Appendix A.

4. Conclusions

We have presented a simple yet effective model that can aid in the automatic classification of inpainted images by generative models. The discriminator model is particularly useful in the context of diffusion inpainting models as they present high variability and diversity of results, usually presenting artifacts and uncanny objects that do not correspond with the specific artist or artwork being restored. The model can be trained via expert knowledge, as in our case, but additionally on simple *Wiki-Art* image stocks. Additionally, the discriminator does not need white-box access to the generator's code and can be used in proprietary models such as OpenAI's DALLE (Ramesh et al., 2021) series.

By presenting qualitative and quantitative results, we have shown that the discriminator can match human experts in

the image selection process. This creates a tool that helps to narrow down the number of inpainted images to select from, easing the restoration process.

5. Future Work

As future work, one can try to add the discriminator's gradients inside the generative step of the model, for models where white-box access is available, such as GLIDE. This is an alternative solution, similar to the one proposed by (Dhariwal & Nichol, 2021) that narrows down the amount of images generated, making the end-to-end process more efficient.

An extra layer of model guidance can be obtained if the discriminator is combined with prompt engineering. The diffusion model can be treated as a black box and the text prompt can be the free variable to modify in a way that the probability of the discriminator is maximized for all images generated. This would require the development of a prompt solver and moreover, the guidance that one can obtain with the prompt is limited, since GLIDE nor DALLE2 (Ramesh et al., 2022) were trained to understand positional directives such as "on top of" or "to the right of" etc.

Acknowledgements

The authors would like to thank the anonymous reviewers and Rosanne Liu, Pablo Samuel Castro, Cris Luengo and Yash Sharma for the useful comments.

References

- de Smit, B. and Lenstra Jr, H. W. The Mathematical Structure of Escher's Print Gallery. *Notices of the AMS*, 50(5): 446–451, 2003.
- Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis, 2021. URL <https://arxiv.org/abs/2105.05233>.
- Gatys, L. A., Ecker, A. S., and Bethge, M. A Neural Algorithm of Artistic style. *CoRR*, 2015. URL <http://arxiv.org/abs/1508.06576>.
- Jam, J., Kendrick, C., Walker, K., Drouard, V., Hsu, J. G.-S., and Yap, M. H. A comprehensive review of past and present image inpainting methods. *Computer Vision and Image Understanding*, 203:103147, 2021. ISSN 1077-3142. doi: <https://doi.org/10.1016/j.cviu.2020.103147>. URL <https://www.sciencedirect.com/science/article/pii/S1077314220301661>.
- Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models, 2022. URL <https://arxiv.org/abs/2201.11793>.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. 2021. doi: <https://doi.org/10.48550/arXiv.2112.10741>.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. 2015. doi: 10.48550/ARXIV.1511.06434. URL <https://arxiv.org/abs/1511.06434>.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL <https://arxiv.org/abs/2103.00020>.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation, 2021. URL <https://arxiv.org/abs/2102.12092>.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents, 2022. URL <https://arxiv.org/abs/2204.06125>.

A. Appendix

A.1. Additional Information on the Discriminator

The results of the discriminator training, expressed as the confusion matrix of the discriminator, are shown in Figure 5.

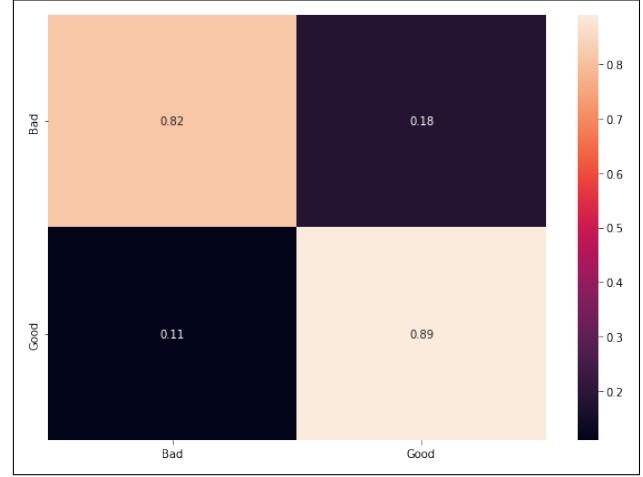


Figure 5. Confusion matrix of the trained discriminator

A.2. Additional Results on the Human Evaluation Experiment

Figure 6 presents the results of the ten human evaluators over the six evaluated images. Results are ordered by the connoisseur score.

		Probability of being an Escher						
		Image Id	1	2	3	4	5	6
Human reviewer Id	Connoisseur score	Human reviewer scores						Discriminator score
		1	5%	5%	1%	10%	5%	5%
2	1	20%	20%	60%	20%	80%	90%	0
3	2	5%	5%	50%	5%	5%	15%	0
4	2	75%	75%	25%	75%	75%	25%	88%
5	2	40%	60%	20%	60%	40%	50%	0
6	3	50%	50%	15%	50%	50%	66%	0
7	3	10%	40%	30%	50%	8%	5%	0
8	4	50%	40%	75%	30%	15%	25%	0
9	5	60%	50%	90%	20%	80%	80%	0
10	6	0%	0%	80%	30%	0%	15%	0
Human weighted average score		34%	37%	54%	37%	34%	38%	
Discriminator score		0	0	88%	0	0	98%	

Figure 6. Discriminator vs. human reviewer comparison

A.3. Supplementary Images

This section presents additional inpainted images generated by GLIDE. Each image contains its corresponding likelihood score assigned by the discriminator as well as the human expert label into "good/bad" with respect to consistency style with M.C. Escher's work.

The aim is to show the reader the variety of inpainting options that GLIDE provides for a given text and masked region. For all the images shown in Figure 7, we used the GLIDE parameters, the same mask and the same prompt *a gallery with arches wooden windows and arcades and floors with tiles Escher*.

Masked image	10%	0%
80%	90%	10%
Masked image	0%	100%

Figure 7. Examples of inpainting options generated by GLIDE alongside their Escher-Likelihood score by the discriminator.