An Analysis of Generative Methods for Multiple Image Inpainting

Coloma Ballester, Aurélie Bugeau, Samuel Hurault, Simone Parisotto, Patricia Vitoria

Abstract  Image inpainting refers to the restoration of an image with missing regions in a way that is not detectable by the observer. The inpainting regions can be of any size and shape. This is an ill-posed inverse problem that does not have a unique solution. In this work, we focus on learning-based image completion methods for multiple and diverse inpainting which goal is to provide a set of distinct solutions for a given damaged image. These methods capitalize on the probabilistic nature of certain generative models to sample various solutions that coherently restore the missing content. Along the chapter, we will analyze the underlying theory and analyze the recent proposals for multiple inpainting. To investigate the pros and cons of each method, we present quantitative and qualitative comparisons, on common datasets, regarding both the quality and the diversity of the set of inpainted solutions. Our analysis allows us to identify the most successful generative strategies in both inpainting quality and inpainting diversity. This task is closely related to the learning of an accurate probability distribution of images. Depending on the dataset in use, the challenges that entail the training of such a model will be discussed through the analysis.

Coloma Ballester  
University Pompeu Fabra, IPCV, DTIC, Spain e-mail: coloma.ballester@upf.edu

Aurélie Bugeau  
Université de Bordeaux, LaBRI, CNRS UMR 5800, France e-mail: aurelie.bugeau@labri.fr

Samuel Hurault  
Université de Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251, F-33400 Talence, France e-mail: samuel.hurault@math.u-bordeaux.fr

Simone Parisotto  
DAMTP, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, United Kingdom e-mail: sp751@cam.ac.uk

Patricia Vitoria  
University Pompeu Fabra, IPCV, DTIC, Spain e-mail: patricia.vitoria@upf.edu
Key words: Inverse Problems, Inpainting, Multiple Inpainting, Diverse Inpainting, Deep Learning, Generative Methods.

1 Introduction

Image inpainting, also called amodal completion or disocclusion in early days, is an active research area in many fields including applied mathematics and computer vision, with foundations in the Gestalt theory of shape perception. Inpainting relates to the virtual reconstruction of missing content in images in a way that is non-detectable by the observer [Bertalmío et al., 2000]. It is an ill-posed inverse problem that can have multiple plausible solutions. Indeed, the fact that the inpainted image is not unique can be understood both mathematically and also because the reconstruction quality is judged by independent humans. On top of that, it has a strong impact on many real-life applications, e.g., in medical imaging (sinograms [Tovey et al., 2019], CT scans [Chen et al., 2012]), 3D surface data (Biasutti et al., 2019), (Bevilacqua et al., 2017), (Hervieu et al., 2010), (Parisotto et al., 2020), art conservation (frescoes [Baatz et al., 2008], panel paintings [Ruzic et al., 2011] and manuscripts [Calatoni et al., 2018]), image compression (Peter and Weickert, 2015), camera artifact removal [Vitoria and Ballester, 2019] and the restoration of old movies and videos (Grossauer, 2006), (Newson et al., 2014), just to name a few.

State-of-the-art image inpainting methods have achieved amazing results regarding the complex work of filling large missing areas in an image. However, most of the methods generally attempt to generate one single result from a given image, ignoring many other plausible solutions. In this chapter, we focus on analyzing recent advances in the inpainting literature, concentrating on the learning-based approaches for multiple and diverse inpainting. The goal of those methods is to estimate multiple plausible inpainted solutions while being as much diverse as possible. Those methods mainly focus on the idea of exploiting image coherency at several levels along with the power of neural networks trained on large datasets of images. Unlike previous one-to-one methods, multiple image inpainting offers the advantage of exploring a large space of possible solutions. This procedure gives the capacity to the user to eventually choose the preferred fit under his/her judgment instead of leaving the task of singling out one solution to the algorithm itself.

This chapter is structured as follows. Section 2 provides a brief overview of both model-based and learning-based inpainting methods in the literature. Section 3 presents the underlying theory of several approaches for multiple and diverse inpainting together with a review of the most representative (to the best of our knowledge) state-of-the-art proposals using those particular strategies. Section 4 presents the evaluation metrics for both inpainting quality and diverse inpainting. The multiple inpainting results of the methods of Section 3 are presented and compared in Section 5 both quantitatively and qualitatively, on common datasets and masks, concerning three aspects: proximity to ground truth, perceptual quality, and inpainting diversity. Finally, Section 6 concludes the presented analysis.
An Analysis of Generative Methods for Multiple Image Inpainting

2 A Walk through the Image Inpainting Literature

In the literature, inpainting methods can fall under different categories, e.g. local vs. non-local depending on the ability to capture and exploit non-nearby content, or geometric vs. exemplar-based methods depending on the action on points or patches. For our purposes, it is more convenient to distinguish between learning- and model-based approaches, according to the usage or not of machine learning techniques. For extensive reviews of existing inpainting methods we refer the reader to the works in [Guillemot and Le Meur, 2014], [Schonlieb, 2015], [Buyssens et al., 2015], [Parisotto et al., 2022].

Model-based Inpainting

Model-based inpainting methods are designed to manipulate an image by exploiting its regularity and coherency features with an explicit model governing the inpainting workflow. One approach for restoring geometric image content is to locally propagate the intensity values and regularity of the image level lines inward the inpainting domain with curvature-driven [Nitzberg et al., 1993], [Masnou and Morel, 1998], [Ballester et al., 2001], [Chan and Shen, 2001], [Esedoglu and Shen, 2002], [Shen et al., 2003] and diffusion-based [Caselles et al., 1998], [Shen and Chan, 2002], [Tschumperle and Deriche, 2005] evolutionary partial differential equations (PDEs), possibly of fluid-dynamic nature [Bertalmio et al., 2000], [Bertalmio et al., 2001], [Tai et al., 2007] or with coherent transport mechanisms [Bornemann and Marz, 2007], also by invoking variational principles [Grossauer and Scherzer, 2003], [Bertozzi et al., 2007] and regularization (possibly of higher-order) priors [Papafitsoros and Schonlieb, 2013]. The filling-in of geometry, especially of small scratches and homogeneous content in small inpainting domains, is the most effective scenario of these methods, which perform poorly in the recovery of texture. Such issue is overcome by considering a patch (a group of neighboring points in the image domain) as the imaging atom containing the essential texture element. The variational formulation of dissimilarity metrics based on the estimation of a correspondence map between patches [Efros and Leung, 1999], [Bornard et al., 2002], [Demanet et al., 2003], [Criminisi et al., 2004], [Aujol et al., 2010] has led to the design of optimal copying-pasting strategies for inpainting large domains. However, these methods still fail, e.g. in the presence of different scale-space features. Thus, some researchers have exploited, also using a variational approach, the efficiency of PatchMatch [Barnes et al., 2009] in computing a probabilistic approximation of correspondence maps between patches to average the contribution of multiple source patches during the synthesis step. For example, [Arias et al., 2011], [Newson et al., 2014] use it in a non-local mean fashion [Wexler et al., 2004], to inpaint rescaled versions of the original image with results propagated from the coarser to the finer scale, [Cao et al., 2011] to guide the inpainting with geometric-sketches, [Sun et al., 2005] to guide structures or [Mansfield et al., 2011], [Eller and Fornasier, 2016], [Fedorov et al., 2016] to account for geometric transformations of patches. However,
these mathematical and numerical advances may result computationally expensive while suffering from the single-imaging source, and dependence on the initialization quality and the selection of associated parameters (e.g. the size of the patch). Thus, it seems natural to explore if image coherency, smoothness and self-similarity patterns can be further exploited by augmenting the dataset of source images and eventually synthesize multiple inpainting solutions: this is where diverse inpainting with deep learning generative approaches is a significant step forward.

One of the earliest model-based inpainting works dealing with multiple source images is [Kang et al., 2002], where salient landmarks are extracted in a scene under different perspectives and then synthesized by interpolation, guiding the imaging restoration. As said, model-based models are sensitive to initializations and chosen parameters: one way to diminish these drawbacks is to perform inpainting of the input image multiple times, by varying parameters like the patch size, the number of pyramid scales, initializations and inpainting methodologies. Thus, a final assembling step will produce an inpainted image which encodes locally the most coherent content [Hays and Efros, 2007], [Le Meur et al., 2013], [Kumar et al., 2016]. Still, the computational effort of estimating several solutions with different parameters and their fine-tuning is remarkable, leading to the need for a one-encompassing strategy that can locally adapt the synthesis step from multiple source images. This task can be solved with learning-based methods.

### Learning-based Methods

Learning-based methods address image inpainting by learning a mapping from a corrupted input to the estimated restoration by training on a large-scale dataset. Besides capturing local or non-local regularities and redundancy inside the image or the entire dataset, those methods also exploit high-level information inherent in the image itself, such as global regularities and patterns, or perceptual clues and semantics over the images.

Early learning-based methods tackled the problem as a blind inpainting problem [Ren et al., 2015], [Cai et al., 2015] by minimizing the distance between the predicted image and the ground truth. This type of methods behaved as an image denoising algorithm and was limited to tiny inpainting domains. To deal with bigger and more realistic inpainting regions, later approaches incorporated in the model the information provided by the mask, e.g. [Kohler et al., 2014], [Ren et al., 2015], [Pathak et al., 2016], [Lempitsky et al., 2018]. Also, several modifications to vanilla convolutions have been proposed to explicitly use the information of the mask, like Partial Convolutions [Liu et al., 2018] and Gated Convolutions [Yu et al., 2019], where the output of those layers only depends on non-corrupted points. Additionally, attempts to increase the receptive field without increasing the number of layers have been proposed with dilated convolutions [Iizuka et al., 2017], [Wang et al., 2018] and contextual attention [Yu et al., 2018], [Yu et al., 2019]. Learning to inpaint in a single step has shown to be a complex endeavor. Progressive learning approaches have also been introduced to split the learning into several steps: for instance, [Zhang et al.,...]
progressively fills the holes from outside to inside; similarly, [Guo et al., 2019], [Zeng et al., 2020], [Li et al., 2020] also learn how to update the inpainting mask for next iteration, and [Li et al., 2019] learns jointly structure and feature information.

To train the network, early approaches minimized some distance between the ground truth and the predicted image. But this approach takes into account just one of the several possible plausible solutions to the inpainting problem. Several approaches have been proposed to overcome this drawback. Some works use perceptual metrics based on generative adversarial networks (GANs) aiming to generate more perceptually realistic results [Pathak et al., 2016], [Yeh et al., 2017], [Iizuka et al., 2017], [Yu et al., 2018], [Vitoria et al., 2019], [Vitoria et al., 2020], [Dapogny et al., 2020], [Liu et al., 2019], [Lahiri et al., 2020]. Other works tackle the problem in the feature space by minimizing distances at feature space level [Fawzi et al., 2016], [Yang et al., 2017], [Vo et al., 2018] by using an additional pre-trained network, or by directly inpainting those features [Yan et al., 2018], [Zeng et al., 2019]. Also, two-steps approaches have been proposed. They are based on a first coarse inpainting [Yang et al., 2017], [Yu et al., 2018], [Liu et al., 2019], edge learning [Liao et al., 2018], [Nazeri et al., 2019], [Li et al., 2019] or structure prediction [Xiong et al., 2019], [Ren et al., 2019], and followed by a refinement step adding finer texture details. Furthermore, [Liu et al., 2020] aimed to ensure consistency between structure and texture generation. Another big problem of early deep learning methods is that deep models treat input images with limited resolution. While first approaches were able to deal with images of maximum size 64 x 64, the latest methods can deal with 1024 x 1024 resolution images by using, for example, a multi-scale approach [Yang et al., 2017], [Zeng et al., 2019], or even to 8K resolution by generating first a low-resolution solution and second its high-frequency residuals [Yi et al., 2020b].

Recent works (e.g. [Zheng et al., 2019], [Zhao et al., 2020b], [Cai and Wei, 2020], [Peng et al., 2021], [Van et al., 2021], [Liu et al., 2021]) deal with the ill-posed nature of the problem by allowing more than one possible plausible solution to a given image. They aim to generate multiple and diverse solutions by using deep probabilistic models based on variational autoencoders (VAEs), GANs, autoregressive models, transformers, or a combination of them. Note that those type of methods has been also used for real case applications such as diverse fashion image inpainting [Han et al., 2019] and Cosmic microwave background radiation (CMB) image inpainting [Yi et al., 2020a]. In this chapter, we will focus on the study of multiple image inpainting methods. More precisely, we will review, analyze and compare, theoretically as well as experimentally, the different approaches proposed on the literature to generate inpainting diversity.
3 How to Achieve Multiple and Diverse Inpainting Results?

In this section, we will describe the different tools and methods that successfully addressed multiple image inpainting. Later in Section 5, we will conduct a thorough experimental study comparing these methods visually and quantitatively.

As previously mentioned, image inpainting is an inverse problem with multiple plausible solutions. Generally, ill-posed problems are solved by incorporating some knowledge or priors into the solution. Mathematically, this is frequently done using a variational approach where a prior is added to a data-fidelity term to create an overall objective functional that is lastly optimized. The selected prior promotes the singling out of a particular solution. Traditionally, the incorporated priors were model-based, founded on properties of the expected solution.

More recently, data-driven proposals have emerged where the prior knowledge on the image distribution is implicitly or explicitly learned via neural networks optimization (we refer to the recent survey [Arridge et al., 2019] and references therein). Among them, generative methods have been used to learn the underlying geometric and semantic priors of a set of non-corrupted images. Indeed, generative methods aim to estimate the probability distribution of a large set, \( \mathcal{X} \), of data. In other words, any \( x \in \mathcal{X} \) is assumed to come from an underlying and unknown probability distribution \( P_X \) and the goal is to learn it from the data in \( \mathcal{X} \). Due to its capacity to produce several outcomes given a single output, some authors have proposed to address the multiplicity of solutions by leveraging the probabilistic nature of generative models.

Through the chapter, we will assume that \( X \) is a set of images. Images will be assumed to be functions defined on a bounded domain \( \Omega \subset \mathbb{R}^2 \) with values in \( \mathbb{R}^C \), with \( C = 1 \) for gray-level images and \( C = 3 \) for color images. With a slight abuse of notation, we will use the same notation to refer to the continuous setting, where \( \Omega \subset \mathbb{R}^2 \) is an infinite resolution image domain and \( x : \Omega \to \mathbb{R}^C \) represents a continuous image, and to the discrete setting where \( \Omega \) stands for a discrete domain given by a grid of \( H \times W \) pixels, \( H, W \in \mathbb{N} \), and \( x \) is a function defined on this discrete \( \Omega \) and with values in \( \mathbb{R}^C \). In the latter case, \( x \) is usually given in the form of a real-valued matrix of size \( H \times W \times C \) representing the image values.

In the context of image inpainting, the inpainting domain, denoted here by \( \Omega \), represents the region of the image domain where the image data is missing, and thus to be restored. Its complementary set, \( \Omega^c = \Omega \setminus \Omega \), represents the region of \( \Omega \) where the values of the image to be inpainted are known. The inpainting mask \( M \) will be defined as equal to 1 on the missing pixels of \( \Omega \), and equal to 0 on \( \Omega \setminus \Omega \).

The space \( \mathcal{X} \) of (complete) natural images is a high dimensional space and its distribution can be very complex. However, natural images contain local regularities, non-local self-similarities, global coherency and even semantic structure. This is one of the reasons that inspired the use of latent-based models. These models use latent variables \( z \in \mathcal{Z} \) in a lower dimension space \( \dim(\mathcal{Z}) \leq \dim(\mathcal{X}) \), associated with a probability distribution \( P_Z \). Generative latent-based models aim to learn a generative model \( G_\theta : \mathcal{Z} \to \mathcal{X} \), with parameters \( \theta \), mapping a latent variable \( z \) to an image \( x \). With a slight abuse of notation and if it is understood from the context, we will
forget about the \( \theta \) subindex and simply write \( G \). The main goal of this strategy is two-fold: (1) to be able to generate samples \( G(z) \) hoping that \( G(z) \in X \) for \( z \sim P_Z \) (or that \( P_G \) is close to \( P_X \), and (2) to use it for density estimation

\[
p_X(x) \approx p_G(x) = \int p(x|z)p_Z(z)dz,
\]

where \( p_G \) stands for the parametric density of \( P_G = G_*P_Z \), the pushforward measure of \( P_Z \) through \( G \) (defined in brief as \( G_*P_Z(B) = P_Z(\{z \in Z | G(z) \in B\}) \), for any \( B \) in the Borel \( \sigma \)-algebra associated to \( X \)). Let us notice that the likelihood \( p(x|z) \) depends on \( G \), and can be interpreted as a measure of how close \( G(z) \) is to \( x \).

Numerous strategies have been developed to parametrize \( G \) (or \( G_\theta \)) as a neural network model and to learn the appropriate parameters \( \theta \) by making \( P_G \) as close as possible to \( P_X \) for some probability distance \( d(P_G, P_X) \). Among these strategies we quote variational autoencoders, normalizing flows, generative adversarial networks, or autoregressive models.

The problem of image inpainting can also be naturally formulated in a probabilistic manner. Let \( y \) denote an observed incomplete image, which is unknown in \( O \). We are interested in modeling the conditional distribution, \( p(x|y) \), over the values of the variable \( x \) (corresponding to the complete image) conditioned on the value of the observed variable \( y \). As possibly many plausible images are consistent with the same input image \( y \), the distribution \( p(x|y) \) will likely be multimodal. Then, each of the multiple solutions can be generated by sampling from that distribution using a given sampling strategy. Thus, the goal is not only to obtain a generative model that minimizes \( d(P_G, P_X) \), where \( X_x \subset X \) is the set of possible solutions, but also to design a mechanism able to sample the conditional distribution \( p(x|y) \), i.e. for a given damaged incomplete image \( y \), output a set of plausible completions \( x \).

| Method       | VAE | Autoregressive | GAN | Transformers |
|--------------|-----|----------------|-----|--------------|
| PIC ([Zheng et al., 2019]) | ✓   | ✓              | ✓   |              |
| PiiGAN ([Cai and Wei, 2020]) | ✓   | ✓              |     |              |
| UCTGAN ([Zhao et al., 2020b]) | ✓   | ✓              |     |              |
| DSI-VQVAE ([Peng et al., 2021]) | ✓   | ✓              | ✓   |              |
| ICT ([Wan et al., 2021]) | ✓   | ✓              |     |              |
| PD-GAN ([Liu et al., 2021]) | ✓   |                |     |              |
| BAT ([Yu et al., 2021]) | ✓   | ✓              |     |              |
In this section, we will analyze the different families of generative models proposed in the literature to realize diverse image inpainting. We will in particular describe generative adversarial networks (GAN), variational autoencoders (VAE), autoregressive models and transformers. We will also detail the different objective losses proposed to train these networks. Finally, for each family of models, we will review several state-of-the-art diverse inpainting methods that relate to this model. Table 1 lists all the methods that will be reviewed in this section.

### 3.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a type of generative models that have received a lot of attention since the seminal work of [Goodfellow et al., 2014]. The GAN strategy is based on a game theory scenario between two networks, a generator network and a discriminator network, that are jointly trained competing against each other in the sense of a Nash equilibrium. The generator maps a vector from the latent space, \( z \sim \mathbb{P}_Z \), to the image space trying to trick the discriminator, while the discriminator receives either a generated or a real image and must distinguish between both. The parameters of the generator and the discriminator are learned jointly by optimizing a GAN objective by a min-max procedure. This procedure leads the probability distribution of the generated data to be as close as possible, for some distance, to the one of the real data. Several GAN variants have appeared. They mainly differ on the choice of the distance \( d(\mathbb{P}_1, \mathbb{P}_2) \) between two probability distributions \( \mathbb{P}_1 \) and \( \mathbb{P}_2 \). The first GAN by [Goodfellow et al., 2014] (also referred to as vanilla GAN) makes use of the Jensen-Shannon divergence, which is defined from the Kullback–Leibler divergence (KL), by

\[
d_{JS}(\mathbb{P}_1, \mathbb{P}_2) = \frac{1}{2} \left[ \text{KL} \left( \mathbb{P}_1 \parallel \frac{\mathbb{P}_1 + \mathbb{P}_2}{2} \right) + \text{KL} \left( \mathbb{P}_2 \parallel \frac{\mathbb{P}_1 + \mathbb{P}_2}{2} \right) \right],
\]

where the KL is defined, for discrete probability densities, as

\[
\text{KL}(\mathbb{P}_1, \mathbb{P}_2) = \sum_x \mathbb{P}_1(x) \log \left( \frac{\mathbb{P}_1(x)}{\mathbb{P}_2(x)} \right).
\]

and, for continuous densities, as

\[
\text{KL}(\mathbb{P}_1, \mathbb{P}_2) = \int_x \mathbb{P}_1(x) \log \left( \frac{\mathbb{P}_1(x)}{\mathbb{P}_2(x)} \right) dx.
\]

The Wasserstein GAN [Arjovsky et al., 2017] uses the Wasserstein-1 distance, given by

\[
\mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_2) = \inf_{\pi \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \mathbb{E}_{x,y \sim \pi} (\|x - y\|),
\]

where \( \Pi(\mathbb{P}_1, \mathbb{P}_2) \) is the set of measurable couplings between \( \mathbb{P}_1 \) and \( \mathbb{P}_2 \).
where $\Pi(P_1, P_2)$ is the set of all joint distributions $\pi$ whose marginals are respectively $P_1$ and $P_2$. By Kantorovitch-Rubenstein duality, the Wasserstein-1 distance can be computed as

$$W_1(P_1, P_2) = \sup_{D \in \mathcal{D}} \left( \mathbb{E}_{x \sim P_1} [D(x)] - \mathbb{E}_{y \sim P_2} [D(y)] \right),$$

where $\mathcal{D}$ denotes the set of 1-Lipschitz functions. In practice, the dual variable $D$ is parametrized by a neural network and it represents the so-called discriminator.

Both the generator and discriminator are jointly trained to solve

$$\min_G \sup_{D \in \mathcal{D}} \left( \mathbb{E}_{x \sim P_{X}} [D(x)] - \mathbb{E}_{y \sim P_{G}} [D(y)] \right),$$

in the case of the Wasserstein GAN, and

$$\min_G \max_D \mathbb{E}_{x \sim P_{X}} [\log D(x)] + \mathbb{E}_{y \sim P_{G}} [\log(1 - D(y))]$$

for the vanilla GAN. In (8), the discriminator $D$ is simply a classifier that tries to distinguish samples in the training set $X$ (real samples) from the generated samples $G(z)$ (fake samples) by designing a probability $D(x) \in [0, 1]$ for its likelihood to be from the same distribution as the samples in $X$.

GANs are sometimes referred to as implicit probabilistic models due to the fact that they are defined through a sampling procedure where the generator learns to generate new image samples. This is in contrast to variational autoencoders, autoregressive models, and methods that explicitly maximize the likelihood.

For the task of inpainting, several proposals set the problem as a conditioned one. The GAN approach is modified such that the input of the generator $G$ is both an incomplete image $y$ and a latent vector $z \sim P_Z$, and $G$ performs conditional image synthesis where the conditioning input is $y$. In the GAN-based works that we present in this section ([Cai and Wei, 2020, Liu et al., 2021]), the authors focus on multimodal conditioned generation where the goal is to generate multiple plausible output images for the same given incomplete image.

Finally, let us mention that in these works, and in general in some works described in this chapter, the used generative methods are combined with consistency losses that encourage the inpainted images to be close to the ground truth. Examples of those consistency losses include value and feature reconstruction losses, and perceptual losses. Nonetheless, multiple inpainting researchers acknowledge that it can be counterproductive to rely on consistency losses due to the fact that the ground truth is only one of the multiple solutions.

PiiGAN: Generative Adversarial Networks for Pluralistic Image Inpainting ([Cai and Wei, 2020])

One of the first methods that use GANs in order to generate pluralistic results is PiiGAN ([Cai and Wei, 2020]). PiiGAN is a deep generative model that incorporates
a style extractor that can extract the style features, in the form of a latent vector, from the ground truth image.

To be more precise, the network is composed of one generator and extractor network. The training follows two different paths.

First, given a ground truth image $x_{gt}$, the extractor is used to estimate the style feature $z_{gt}$. Once the style feature $z_{gt}$ is obtained, the ground truth image $x_{gt}$ is masked and concatenated with the computed style feature $z_{gt}$ and used as input of the generator network. The generator network will estimate the inpainted version $x$ of the masked ground truth image $y$. The estimated inpainted image $x_{out,1}$ is passed through the extractor network to estimate the corresponding style feature $z_{out,1}$. This path is supervised using the KL-divergence between the style features $z_{gt}$ and $z_{out,1}$.

In parallel, another path estimates inpainted images from masked images without ground truth. That is, masked images without ground truth $y_{raw}$ are fed to the generator with a random vector $z_{raw}$. An inpainted image $x_{out,2}$ is predicted followed by style feature prediction $z_{out,2}$. Additionally, they frame the inpainting of $y_{raw}$ in an adversarial approach equipped with a local (that focuses just in the inpainted area) and a global discriminator applied to the inpainted image $x_{raw}$. This path, is supervised using the $L^1$ norm of the difference between the style features $x_{raw}$ and $x_{out,2}$, and adversarial loss applied to the inpainted image $x_{out,2}$ based on the Wasserstein loss (7) with gradient penalty as defined in [Gulrajani et al., 2017].

The authors claim that their results are diverse and natural, especially for images with large missing areas. Figure 1 shows an overview of the algorithm pipeline.

**PD-GAN: Probabilistic Diverse GAN for Image Inpainting ([Liu et al., 2021])**

The authors of [Liu et al., 2021] propose a method to perform diverse image inpainting called PD-GAN. PD-GAN takes advantage of the benefits of GANs in generating diverse content from different random noise inputs. Figure 2 displays an overview of the algorithm pipeline. In contrast to the original vanilla GAN, in PD-GAN all the decoder deep features are modulated from coarse to fine by injecting prior information at each scale. This prior information is extracted from
an initially restored image at a coarser resolution together with the inpainting mask. For that purpose, they introduce a probabilistic diversity normalization (SPDNorm) module based on the Spatially-adaptive denormalization (SPADE) module proposed in [Park et al., 2019]. SPDNorm works by modeling the probability of generating a pixel conditioned on the context information. It allows more diversity towards the center of the inpainted hole and more deterministic content around the inpainting boundary.

The objective loss is a combination of several losses, including a diversity loss, a reconstruction loss, an adversarial loss, and a feature matching loss (difference in the output feature layers computed with the learned discriminator). In general, in the context of multiple image synthesis, diversity losses aim at ensuring that the different reconstructed images are diverse enough. In particular, the authors of PD-GAN [Liu et al., 2021] use the so-called perceptual diversity loss, defined as

$$L_{pd\text{iv}}(x_{out_a}, x_{out_b}) = \frac{1}{\sum_{l} \| M \odot (\Phi_l(x_{out_a}) - \Phi_l(x_{out_b})) \|_1 + \epsilon}.$$  \hspace{1cm} (9)

where $x_{out_a}$ and $x_{out_b}$ are two inpainted results, and $M$ the inpainting mask (with 1 values on the missing pixels and 0 elsewhere). The minimization of (9) favors the maximization of the perceptual distance of inpainted regions in $x_{out_a}$ and $x_{out_b}$. Notice that the non-masked pixels are not affected by this loss. A similar diversity loss was proposed in [Mao et al., 2019].

### 3.2 Variational Autoencoders and Conditional Variational Autoencoders

Variational Autoencoders (VAE) ([Kingma and Welling, 2013]) are generative models for which the considered distance between probability distributions is the Kullback-Leibler (KL) divergence. Minimizing $\text{KL}(P_{\hat{z}} \| P_{\mathcal{X}})$ with respect to the parameters $\theta$, is equivalent to the log likelihood maximization.
where, in the VAE context, $G_\theta$ is referred to as the decoder.

Let us first derive the vanilla VAE formulation in the general context of non-corrupted images $x \in X$. Using Bayes rule, the likelihood $p_{G_\theta}(x)$, for $x \sim \mathbb{P}_X$ and $z \sim \mathbb{P}_Z$, is given by

$$p_\theta(x) = \frac{p_\theta(x, z)}{p_\theta(z|x)} = \frac{p_\theta(x|z)p_z(z)}{p_\theta(z|x)} \tag{11}$$

where, to simplify notations, we have denoted $p_{G_\theta}$ simply by $p_\theta$. In order to bypass the intractability of the posterior $p_\theta(z|x)$, Variational Autoencoders introduce a second neural network, $q_\psi(z|x)$, to parametrize an approximation of the true posterior. This neural network is referred to as the encoder. Let us now derive the VAE objective function. Following [Kingma et al., 2019],

$$\log p_\theta(x) = \mathbb{E}_{q_\psi(z|x)} \left[ \log \frac{p_\theta(x, z)}{q_\psi(z|x)} \right] + \mathbb{E}_{q_\psi(z|x)} \left[ \log \frac{q_\psi(z|x)}{p_\theta(z|x)} \right] \tag{12}$$

$$= \mathbb{E}_{q_\psi(z|x)} \left[ \log p_\theta(x, z) - \log q_\psi(z|x) \right] + \text{KL} (q_\psi(z|x)||p_\theta(z|x)) \tag{13}$$

$$L_{\theta, \phi}(x) = \log p_\theta(x) - \text{KL} (q_\psi(z|x)||p_\theta(z|x)). \tag{14}$$

$L_{\theta, \phi}$ is the so-called Evidence Lower Bound (ELBO). By positivity of the KL, it verifies

$$L_{\theta, \phi}(x) = \log p_\theta(x) - \text{KL} (q_\psi(z|x)||p_\theta(z|x)) \leq \log p_\theta(x), \tag{15}$$

and $L_{\theta, \phi}(x) = \log p_\theta(x)$ if and only if $q_\psi(z|x)$ is equal to $p_\theta(z|x)$.

VAE training consists in maximizing $L_{\theta, \phi}$ in (15) with respect to the parameters $\{\theta, \psi\}$ of the encoder and of the decoder, simultaneously. The goal is to obtain a good approximation $q_\psi(z|x)$ of the true posterior $p_\theta(z|x)$ while maximizing the marginal likelihood $p_\theta(x)$.

The work [Sohn et al., 2015] extends VAEs by proposing Conditional Variational Autoencoders (CVAE). Their targeted distribution is the conditional distribution of $x$ given an input "conditional" variable $c$ and (10) becomes

$$\max_\theta \mathbb{E}_{x \sim \mathbb{P}_X} \log p_{G_\theta}(x|c). \tag{16}$$

CVAE loss is obtained with a similar argument as in (12)-(15) by maximizing the conditional log-likelihood which gives the variational lower bound of the conditional log-likelihood

$$\mathbb{E}_{q_\psi(z|x)} \left[ \log p_\theta(x|c, z) \right] - \text{KL} (q_\psi(z|x, c)||p_\theta(z|x)) \leq \log p_\theta(x|c). \tag{17}$$

Then, the idea of the deep conditional generative modeling is simple: given an observation (input) $x$, $z$ is drawn from a prior distribution $p_\theta(z|x)$. Then, the output is generated from the distribution $p_\theta(x|z, c)$. [Bao et al., 2017] combines a CVAE
with a GAN (CVAE-GAN) for fine-grained category image generation. Even if inpainting results are shown, the network is not trained explicitly for inpainting but for image generation conditioned on image labels.

In the context of multiple image inpainting, or more generally of multiple image restoration, a straightforward idea is to condition the generative model on the input degraded image $y$ and to generate multiple images $x$ sampling from $p_\theta(x|z, c = y)$. BicycleGAN [Zhu et al., 2017] uses this idea for diverse image-to-image translation. Their goal is to learn a bijective mapping between two image domains with a multi-modal conditional distribution. They combine CVAE-GAN with latent regressors and show that their method can produce both diverse and realistic results across various image-to-image translation problems. However, their method is not explicitly applied for image inpainting. Moreover, as observed by several authors (see, e.g. [Zheng et al., 2019], [Wan et al., 2021]), using standard conditional VAEs or CVAE-GAN for the specific task for image inpainting still leads to minimal diversity and quality. Several extension of these models have recently appeared for diverse image inpainting. They are presented below with more details.

Finally, let us notice that VAE model has been extended in [van den Oord et al., 2017], [Razavi et al., 2019] to Vector Quantized-Variational AutoEncoder (VQ-VAE) that uses vector quantization to model discrete latent variables. Such discretization is done to avoid posterior collapse. The quantization codebook is trained at the same time as the auto-encoder with an objective loss made of a reconstruction term and a regularization term that ensures that the embedding commits to the encoder outputs and respectively. The work [Razavi et al., 2019] is an hierarchical extension of [van den Oord et al., 2017]. In particular, the authors of [Razavi et al., 2019] show that, by only considering two levels of a multi-scale hierarchical organization of VQ-VAE [van den Oord et al., 2017], the information about image texture is disentangled from the information about the structure and geometry of the objects in an image. By combining the obtained hierarchical multi-scale latent data with an autoregressive model as prior (see Section 3.3 below), they show an improved ability for generating high-resolution images.

**Pluralistic Image Completion [Zheng et al., 2019]**

The work [Zheng et al., 2019] aims to estimate a probability distribution $p(x|y)$ from which to sample, where $y$ represents an incomplete image and $x$ one of its possible completions. They propose to use the Conditional Variational Autoencoder (CVAE) [Sohn et al., 2015] approach described above which estimates a parametric distribution over a latent space, as in equation (17), from which sampling is possible. However, in [Zheng et al., 2019] the authors observe that if they explicitly promote the inpainted output to be similar to the ground truth image (either by any error-based loss such as, for instance, the $L^1$ distance, or as the authors show, by maximizing $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x,y,z)]$ in (17) while $KL(q_\phi(z|x,y)||p_\theta(z|x))$ tends to zero), it results in a lack of diverse outputs. Alternatively, one could impose to fit the distribution of the training dataset by an adversarial approach including a
discriminator as described in Section 3.1. However, this approach is highly unstable. Instead, they propose a probabilistic framework with a dual pipeline composed of two paths. See a detailed pipeline in Figure 3. One is the reconstructive path which is a VAE-based model that utilizes the ground truth to get a prior distribution of missing parts, \( x|O \), with the variance on the latent variables’ prior depending on the hole area, and rebuild the exact same ground truth image from this distribution. The other is a generative path for which the conditional prior, based only on the visible regions, is coupled to the distribution obtained in the reconstructive path to generate multiple and diverse samples. Both parts are framed in an adversarial approach to fit the distribution of the training dataset. Accordingly, the whole training loss is a combination of three types of terms. First, they use the KL divergences between the mentioned distributions. Second, the appearance/error terms based on the L\(^1\) norm, where in the generative path it only has into account the visible pixels. And third and lastly, an adversarial discriminator-based term. It is based on the L\(^1\) difference among the discriminator features of the ground truth and the reconstructed image for the reconstructive path, and on the discriminator value on the generated image for the generative path. Additionally, to exploit the distant relation among the encoder and decoder, they use a modified self-attention layer that captures fine-grained features in the encoder and more semantic generative features in the decoder.

**Fig. 3** Overview of the PIC architecture of Pluralistic Image Completion [Zheng et al., 2019]. Figure from [Zheng et al., 2019].

**UCTGAN: Diverse Inpainting based on Unsupervised Cross-Space Translation [Zhao et al., 2020b]**

The authors of [Zhao et al., 2020b] aim to produce multiple and diverse solutions conditioned by an instance image that guides the reconstruction, again aiming to maximize the conditional log-likelihood involving the variational lower bound (17) on the training dataset. They call their proposal UCTGAN. The pipeline is presented in Figure 4. They use a two-encoders network that transforms the instance image and the corrupted image to a low-dimensional manifold space. A cross semantic attention
layer combines the information in both low-dimensional spaces. Consecutively, a generator is used to compute the conditional reconstructed image.

The objective loss is composed of four terms. First, a constraint loss in the uncorrupted pixels is applied by minimizing the $L^1$ norm of the difference both at pixel and feature level. Second, the KL divergence is used to project the low-dimensional manifold space of the instance image and masked image into a multivariate normal distribution space. Additionally, the $L^1$ norm of the difference in the low-dimensional manifold space of the instance image and the ground truth image is added. Finally, all the training is framed in an adversarial approach using the vanilla GAN (8), where the discriminator works in the image space.

**Generating Diverse Structure for Image Inpainting With Hierarchical VQ-VAE [Peng et al., 2021]**

The multiple inpainting proposal in [Peng et al., 2021] leverages three generative strategies, namely, variational autoencoders, generative adversarial methods, and autoregressive models. We first review below the main ideas of autoregressive models and then describe the proposal [Peng et al., 2021].

### 3.3 Autoregressive Models

In autoregressive models [Van Oord et al., 2016, Oord et al., 2016, Chen et al., 2018], the likelihood $p(y|x)$ is learned by choosing an order of the data variables $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X}$, frequently related to values on the $n$ pixels of an image, and exploiting the fact that the joint distribution can be decomposed as

---

**Fig. 4** Overview of the architecture of UCTGAN: Diverse Inpainting based on Unsupervised Cross-Space Translation [Zhao et al., 2020b]. Figure from [Zhao et al., 2020b].
\[ p(x) = p(x_1, x_2, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i|x_1, \ldots, x_{i-1}). \quad (18) \]

More generally, a similar decomposition to (18) can be obtained by splitting the set of variables in smaller disjoint subsets. In this case, and considering the variable order of \( x_1, x_2, \ldots, x_n \) to be represented by a directed and non-cyclic graph, one has

\[ p(x) = p(x_1, x_2, \ldots, x_n) = p(x_1) \prod_{i=2}^{m} p(x_i|S(x_i)), \quad (19) \]

where \( S(x_i) \) is the set of parent variables of variable \( i \), and \( m \leq n \).

Autoregressive models have been used to learn a probability distribution or a conditional distribution, for instance in the context of VAEs (cf. Section 3.2) to model the prior or the decoder, and also to tackle several problems in imaging such as image generation (e.g. [Razavi et al., 2019]), super resolution (e.g. [Dahl et al., 2017]), inpainting (e.g. [Peng et al., 2021]) or image colorization (e.g. [Zhao et al., 2020a, Guadarrama et al., 2017, Royer et al., 2017]), and also for other types of data such as audio and speech synthesis (e.g. [Oord et al., 2018]) or text (e.g. [Bowman et al., 2015]).

**Generating Diverse Structure for Image Inpainting With Hierarchical VQ-VAE [Peng et al., 2021]**

Inspired by the hierarchical vector quantized variational autoencoder (VQ-VAE) [Razavi et al., 2019] whose hierarchical architecture disentangles structural and textural information, the authors of [Peng et al., 2021] propose a two-stage pipeline (cf. Figure 5). As already pointed out by [Razavi et al., 2019], by using a two-step approach instead of directly computing the final inpainted image, they aim to generate richer structure and texture images than previous VAE-based methods that often produce a distorted structure or blurry textures.

The first stage of [Peng et al., 2021], known as diverse structure generator, generates multiple low-resolution results each of which has a different structure by sampling from a conditional autoregressive distribution. The second stage, known as texture generator, uses an encoder-decoder architecture with a structural attention module that refines each low-resolution result separately by augmenting texture. The structural information module facilitates the capture of distant correlations. They further reuse the VQ-VAE to calculate two feature losses, which help improve structure coherence and texture realism, respectively.

The authors first train the hierarchical VQ-VAE and, afterward, the diverse structure generator \( G_s \) depending on parameters \( \theta \) and the texture generator \( G_t \) depending on parameters \( \varphi \) are trained separately. These generators are later on used for inference. The structure generator \( G_s \) is constructed via a conditional autoregressive network for the distribution over structural features. In inference, it will generate different structural features via sampling. Its objective loss is defined as the negative log-likelihood
where \( y \) is the input image to be inpainted on the points of \( O \) where the hole mask \( M \) is equal to 1, \( P_X \) denotes the distribution of the training dataset, \( s_{gt} \) denote the vector quantized structural features of the ground truth at the coarser scale given by the hierarchical VQ-VAE, and \( \theta \) the parameters of \( G_s \).

Besides, the objective loss for the texture generator \( G_t \) is composed by: (i) the L\(^1\) norm comparing the inpainted solution to the ground truth at pixel level, (ii) an adversarial loss using the discriminator trained with the SN-PatchGAN hinge version [Yu et al., 2019] applied to the resulting image and, moreover, (iii) a structural feature loss \( L_{ts}(\varphi) \) and (iv) a textural feature loss \( L_{tt}(\varphi) \). These last two losses are defined similarly using a multi-class cross-entropy loss. In particular, the structural feature loss is defined as

\[
L_{ts}(\varphi) = - \sum_{k,j} \alpha_{k,j} \log \left( \text{softmax}(10 \, \delta_{k,j}) \right),
\]

where \( \delta_{k,j} \) denotes the truncated distance similarity score between the k-th feature vector of \( s_{comp} \) (computed from the inpainted image using the trained encoder) and the j-th prototype vector of the structural codebook of VQ-VAE. \( \alpha_{k,j} \) is an indicator of the prototype vector class. That is, \( \alpha_{k,j} = 1 \) when the k-th feature vector of \( s_{gt} \) belongs to the j-th class of the structural codebook; otherwise, \( \alpha_{k,j} = 0 \). The authors define the textural feature loss \( L_{tt}(\varphi) \) in an analogous way.

As mentioned, in Section 5 we will experimentally analyze this method. It will be denoted there as DSI-VQVAE.
3.4 Image Transformers

Self-attention-based architectures, in particular Transformers [Vaswani et al., 2017] are well explored architectures in Natural Language Processing (NLP). Transformers use a self-attention mechanism to model long range relationships between the elements of an input sequence (for instance, in a text) that has shown to be more efficient than Recurrent Neural Networks. They have achieved state-of-the-art results in several tasks not only in the field of NLP but also more recently for computer vision problems. The vanilla transformer [Vaswani et al., 2017] and its variants have been successfully applied in computer vision to, e.g. inpainting [Wan et al., 2021, Yu et al., 2021], object detection [Carion et al., 2020], image classification [Dosovitskiy et al., 2020], colorization [Kumar et al., 2021], super resolution [Yang et al., 2020].

Instead of using inductive local biases like CNNs, transformers in imaging aim to have a global receptive field. For this, the image is first transformed by, as in the most basic approach, flattening the spatial dimensions of the input feature map into a sequence of features of size $M \times N \times F$, where $M \times N$ represents the flattened spatial dimensions and $F$ the depth of a feature map. Then self-attention is applied over the extracted sequence. To ease the associated high computational cost, some authors substitute spatial pixels by patches. The attention mechanism looks at the input sequence and decides for each position which other parts of the sequence or image are important. More specifically, the transformers will transform the set of inputs, called tokens, using sequential blocks of multi-headed self-attention, which relate embedded inputs to each other. It is worth notice, that transformers will maintain the number of tokens throughout all computations. If tokens were related to pixels, each pixel would have a one-to-one correspondence with the output, thus, maintaining the spatial resolution of the original input image. Since transformers are set-to-set functions, they do not intrinsically retain the information of the spatial position for each individual token, thus the embedding is concatenated to a learnable position embedding to add the positional information to the representation.

One advantage of using a transformer for image restoration is that it naturally supports pluralistic outputs by directly optimizing the underlying data distribution. One drawback is the computational complexity that increases quadratically with the input length, thus making it difficult to directly synthesize high-resolution images.

High-Fidelity Pluralistic Image Completion with Transformers [Wan et al., 2021]

The authors of [Wan et al., 2021] exploit the benefits of both, transformers and CNNs. The use of transformers will enforce a global structural understanding and pluralism support in the inpainted region, at a coarse resolution. On the other hand, the use of CNNs will allow working with high-resolution images without a high computational cost due to its capacity of estimating local textures efficiently.

Concretely, in this work, image completion is performed in two steps as shown in Figure 6. In the first step, given a corrupted image, the authors use transformers to produce the probability distribution of structural appearance of complete images.
An Analysis of Generative Methods for Multiple Image Inpainting

given the incomplete one. Low-resolution results can be obtained by sampling from this distribution with diversities that recover pluralistic coherent image structures. In the second step, guided by the computed image structures together with the available pixels of the input image, another upsampling CNN model is used to render high fidelity textures for missing regions meanwhile ensuring coherence with neighboring pixels.

![Fig. 6 Overview of the architecture of High-Fidelity Pluralistic Image Completion with Transformers [Wan et al., 2021], referred to as ICT. Figure from [Wan et al., 2021].](image)

If $X_{\Pi}$ denotes the set of masked tokens $x_{\pi_k}$ (where $\Pi = \{\pi_1, \ldots, \pi_K\}$ denote their indexes), and $X_{-\Pi}$ denotes the set of unmasked tokens (corresponding to the visible regions), then the transformer is optimized by minimizing the negative log-likelihood of the masked tokens $x_{\pi_k}$, conditioned the visible regions $X_{-\Pi}$, that is,

$$L_{\text{MLM}}(\theta) = \mathbb{E}_{X} \frac{1}{K} \sum_{k=1}^{K} -\log p(x_{\pi_k} | X_{-\Pi}; \theta) \quad (22)$$

where $\theta$ contains the parameters of the transformer, and the subindex MLM stands for the masked language model which is similar to the one in BERT [Devlin et al., 2018]. One particularity of the ICT model is that each token attends simultaneously to all positions thanks to bi-directional attention. This enables the generated distribution to capture the full context, thus leading to a consistency between generated contents and unmasked region.

Once the transformer is trained, instead of directly sampling the entire set of masked positions which would lead to non-plausible results due to the independence property, they apply Gibbs sampling to iteratively sample tokens at different locations. To do so, in each iteration, a grid position is sampled from $p(x_{\pi_k} | X_{-\Pi}, X_{<\pi_k}, \theta)$ with the top-$K$ predicted elements, where $X_{<\pi_k}$ denotes the previous generated tokens.

The second step is to perform texture refinement at the original resolution using a CNN, which is optimized by minimizing the $L^1$ loss between the predicted image and the ground truth, together with an adversarial loss based on the vanilla GAN (cf. (8) in Section 3.1).
Diverse image inpainting with bidirectional and autoregressive transformers
[Yu et al., 2021]

This proposal exploits, as in [Wan et al., 2021], a two steps strategy where transformers will encode global structure understanding and high-level semantics at a first stage, followed by a CNN-based generation of additional texture. While [Wan et al., 2021] leverages bi-directional attention with the masked language model (MLM) as in BERT [Devlin et al., 2018], the authors of [Yu et al., 2021] propose BAT-Fill, that combines autoregressive models and bidirectional models (cf. Figure 7). The first transformer-based step estimates the distribution of inpainted low-resolution structures from which to sample, from an input damaged image, a set \( \{s_1, \ldots, s_J\} \) of plausible complete structures. Instead of only using a masked language model like BERT and [Wan et al., 2021] (see above), that use bidirectional contextual information but predicts each masked token separately and independently (which can result in inconsistency in the generated result), BAT-Fill incorporates autoregressive modeling (factorizing the predicted tokens with the product rule). The input sequence of tokens is sorted by first having the visible tokens (permuted) and then the missing tokens (with the original order). In this way, the autoregressive model starts at the position of the first missing pixel. The BAT training objective is given by

\[
\mathcal{L}_{\text{BAT}}(\theta) = \mathbb{E}_{X} \frac{1}{K} \sum_{k=1}^{K} \log p(x_{\pi_k} | X_{-\Pi}, M, X_{<\pi_k}; \theta). \tag{23}
\]

where we have used the same notations as in (22), namely, \( K \) is the length of masked tokens, \( X_{-\Pi} \) are all the unmasked tokens (corresponding to the visible regions). Finally, \( X_{<\pi_k} \) denote the previous predicted tokens, and \( M \) the masked positions.

Fig. 7 Overview of the BAT architecture of Diverse image inpainting with bidirectional and autoregressive transformers [Yu et al., 2021]. Figure from [Yu et al., 2021].
Finally, they construct a texture generator based on CNN-based synthesis, which is optimized by minimizing the $L^1$ loss between the predicted image and the ground truth together with an adversarial loss and a perceptual loss [Johnson et al., 2016].

In inference, each masked token is predicted bidirectionally and autoregressively. As in [Wan et al., 2021], they iteratively use top-$K$ sampling to randomly sample from the $K$ most likely next tokens.

4 From Single Image Evaluation Metrics to Diversity Evaluation

Currently, there is no consensus on automatic evaluation methods for single or diverse inpainting. As the problem is to recover a visually plausible image, performing quantitative evaluation is not trivial as the solution is not unique and the plausibility is a subjective term. Nevertheless, several evaluation metrics have been proposed through the years. We first detail those used for evaluating inpainting methods, one image at the time, before presenting the metric used as diversity scores.

Full reference metrics compare the ground truth image with the inpainted result. Famous measures in this category include $L^1$, $L^2$ distances, PSNR, or SSIM (Wang et al., 2004). These metrics analyze the ability of the model to reconstruct the original image content. Nevertheless, it is easy to demonstrate that they do not well characterize the realism of an image. Being close to the ground truth image does not ensure being realistic. Other perceptual metrics have been proposed and are supposed to be more consistent with human judgment. In particular, Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al., 2018b] has been demonstrated to correlate well with the human perceptual similarity. It relies on the observation that hidden activations in CNNs trained for image classification are indeed a space where distance can strongly correlate with human judgment. Precisely, LPIPS computes a weighted $L^2$ norm between deep features of pair of images:

$$\text{LPIPS}(x, x_{gt}) = \sum_{l} \sum_{i,j} \frac{1}{M_l N_l} \|w_l \odot (\Phi_l^x(i,j) - \Phi_l^{x_{gt}}(i,j))\|^2_2$$

where $x$ is the reconstructed image, $x_{gt}$ is the ground truth, $l$ is a layer number, $(i,j)$ a pixel, $w_l$ are weights for each features, and $\Phi_l^x$ and $\Phi_l^{x_{gt}} \in \mathbb{R}^{M_l \times N_l \times C_l}$ are features unit-normalized in the channel dimension. LPIPS has been used in the context of inpainting when generating one image (e.g. [Zheng et al., 2021]). In [Kettunen et al., 2019], it was shown that standard adversarial attack techniques can easily fool LPIPS. Therefore a slightly different metric called E-LPIPS (Ensemble LPIPS) is proposed by applying random simple image transformations and dropout. Nonetheless, up to our knowledge, it has never been used in the context of inpainting.

When, apart from the set of images, there is available corresponding image categories, other metrics, that are also supposed to be following human judgment, can be used. The inception score (IS) [Salimans et al., 2016] was designed to measure
how realistic the output from a GAN is. This score measures the variety of a set of generated images as well as the probability distribution of each image classification. This is done by comparing the class distribution of each image, which should have a low entropy, with the marginal distribution of the whole set, which should have high entropy:

\[
\text{IS}(G) = \exp \left( \mathbb{E}_{x \sim p_g} \text{KL} \left( p(y|x) || p(y) \right) \right)
\]

where \( p_g \) is the model distribution of the whole set given by the generative model \( G \), \( x \) an image sampled from \( p_g \), \( p(y|x) \) the conditional class distribution, \( \text{KL} \) the Kullback-Leibler divergence and \( p(y) \) the marginal class distribution. As detailed in [Barratt and Sharma, 2018], inception score has its own limitations: sensitivity to small changes in network weights, misleading results when used beyond the ImageNet dataset [Rosca et al., 2017], adversarial examples when used for model optimization. The IS score was adapted to diverse inpainting in [Zhao et al., 2020], leading to the Modified Inception Score (MIS). When performing inpainting, there is only one kind of image and so \( p(y) \) can be removed. The MIS is then defined as

\[
\text{MIS}(G) = \exp \left( \mathbb{E}_{x \sim p_g} \sum_i (p(y_i|x) \log p(y_i|x)) \right)
\]

where \( y_i \) of is the class label of the \( i^{th} \) generated sample. Another improvement of the IS is the Fréchet Inception Distance (FID) [Heusel et al., 2017] that compares the statistics of generated images to the ones of original images. FID uses the Inception pre-trained model to extract the feature vectors of real images and fake images, and compare their feature-wise means (\( \mu_r, \mu_f \)) and covariances (\( \Sigma_r, \Sigma_f \)):

\[
\text{FID} = ||\mu_r - \mu_f||^2 + \text{Tr}((\Sigma_r + \Sigma_f + 2(\Sigma_r \Sigma_f)^{1/2}))
\]

Fréchet Inception Distance has been widely used for validating single and diverse inpainting results in recent papers (e.g. [Peng et al., 2021], [Liu et al., 2021], [Yu et al., 2021]).

**Measuring diversity**

In the context of pluralistic inpainting, following the idea proposed for image-to-image translation in [Zhu et al., 2017], LPIPS has been used as a diversity score to measure how perceptually different the generated images are. [Cai and Wei, 2020], [Zhao et al., 2020b], [Liu et al., 2021]. The higher the LPIPS, the more diversity is present in the results. For instance, in [Cai and Wei, 2020], they compute the average distance between the 10000 pairs randomly generated from the 1000 center-masked image samples. LPIPS is computed on the full inpainting results and mask-region inpainting results, respectively.
Table 2 Generative methods for diverse inpainting: experimental conditions. Random regular and irregular masks are generated as in [Zheng et al., 2019].

| Method  | Input Size | Train Datasets         | Training Masks                      | Code |
|---------|------------|------------------------|-------------------------------------|------|
| PIC     | 256 × 256  | - Celeba-HQ            | - Regular (center 128×128 + random) | ✓    |
|         |            | - ImageNet             | - Irregular (random)                |      |
|         |            | - Paris                |                                     |      |
|         |            | - Places2              |                                     |      |
| PitGAN  | 128 × 128  | - CelebA               | - center 64 × 64                     | ✓    |
|         |            | - Mauflex              |                                     |      |
|         |            | - Agricultural Disease |                                     |      |
| UCTGAN  | 256 × 256  | - Celeba-HQ            | - Regular (center 128×128 + random) | ✗    |
|         |            | - ImageNet             | - Irregular (random)                |      |
|         |            | - Paris                |                                     |      |
|         |            | - Places2              |                                     |      |
| DSI-VQVAE | 256 × 256 | - Celeba-HQ            | - Regular (center 128×128 + random) | ✓    |
|         |            | - ImageNet             | - Irregular (random)                |      |
|         |            | - Places2              |                                     |      |
| ICT     | 256 × 256  | - FFHQ                 | - Irregular Pconv [Liu et al., 2018] | ✓    |
|         |            | - ImageNet             |                                     |      |
|         |            | - Places2              |                                     |      |
| PD-GAN  | 256 × 256  | - Celeba-HQ            | - Irregular Pconv [Liu et al., 2018] | ✗    |
|         |            | - Paris StreetView      |                                     |      |
|         |            | - Places2              |                                     |      |
| BAT     | 256 × 256  | - CelebA-HQ            | - Irregular Pconv [Liu et al., 2018] | ✓    |
|         |            | - Paris StreetView      |                                     |      |
|         |            | - Places2              |                                     |      |

5 Experimental Results

In this section, we present a quantitative and qualitative comparison of several existing methods for multiple image inpainting. We include an assessment of both the quality and the diversity of the inpainted solutions. All the results shown in this section are thanks to publicly available code together with pre-trained weights provided by the authors. In Table 2, we summarize, for all the methods previously reviewed, the conditions in which the experiments were conducted. Note that, among these methods, only PIC ([Zheng et al., 2019]), PitGAN ([Cai and Wei, 2020]), DSI-VQVAE ([Peng et al., 2021]), ICT ([Wan et al., 2021]) and BAT ([Yu et al., 2021]) provide source code and pre-trained models. In the rest of this section, we describe the experimental settings in Section 5.1 including datasets and used masks, quantitative results in Section 5.2 including proximity to ground truth, perceptual quality, and inpainting diversity; and finally a qualitative analysis is provided in Section 5.3.
5.1 Experimental Settings

Table 2 lists all the explained methods together with the training dataset and corresponding training masks. Aiming for a fair comparison, we compare and test the methods trained on the same training images i.e. the VAE-based model PIC ([Zheng et al., 2019]), the VQVAE-based model DSI-VQVAE ([Peng et al., 2021]) and the two transformer-based models ICT ([Wan et al., 2021]) and BAT ([Yu et al., 2021]). Notice that we do not analyze the performance of PiiGAN ([Cai and Wei, 2020]), as the training datasets and size images are different from the other methods.

Datasets

We evaluate the methods on the three datasets Celeba-HQ [Karras et al., 2018], Places2 [Zhou et al., 2017], and ImageNet [Russakovsky et al., 2015]. All the evaluated models take as input images of resolution 256 × 256. Due to the long inference time of DSI-VQVAE and ICT methods (see Table 7), quantitative experiments are made on 100 randomly selected images from each training dataset. For each kind of mask (see below) and for each image, we sample 25 different results.

For Celeba-HQ, the 1024 × 1024 resolution images are resized to 256 × 256. For Places2 and ImageNet, the compared methods were trained on 256 × 256 patches either by resizing the input images (PIC), either by cropping them, randomly (DSI), or to the center patch (BAT), either by both cropping and resizing (ICT). We will both consider center-cropped and resized versions of the input images to ensure a fair comparison among the trained models.

Note that ICT is not trained on Celeba-HQ but on the FFHQ face dataset [Karras et al., 2019]. FFHQ contains higher variation then Celeba-HQ in terms of age, ethnicity and image background. It also has a good coverage of accessories. Images from both datasets are however similarly aligned and cropped. Therefore, we still give the results of the ICT method tested on the Celeba-HQ dataset but the reader should remember this difference when analysing the results.

Inpainting Masks

We use the following type of masks: center, random regular, random irregular and irregular masks from [Liu et al., 2018] with several different proportions of hidden pixels. Figure 8 shows an example of each kind of mask. The random masks are generated once for each test image so that all the method are evaluated on the same degradation.

We would like to highlight that the methods PIC and DSI-VQVAE train a different model for regular and irregular holes. Testing on center or random regular mask is realized with the former model and testing on irregular masks with the latter. The transformer-based methods ICT and BAT only train on "irregular Pconv" holes given by [Liu et al., 2018]. Testing on each type of mask will be done with this unique model.
Fig. 8 Example for each kind of mask considered for evaluation. In grey the hidden pixels. From left to right: center, random regular, random irregular, and irregular Peconv masks from [Liao et al., 2018] with < 20%, [40%, 60%] and [40%, 60%] hidden pixels.

5.2 Quantitative Performance

We first analyse the numerical performance of each method. Table 3 shows quantitative results on Celeba-HQ dataset. Additionally, results on Places2 and ImageNet are respectively shown in Tables 4 and 5.

In Tables 4 and 5 we give our results obtained by center-cropping the images on Places2 and ImageNet, respectively. For fair comparison, we also give in Appendix (Tables 8 and 9), the results on these two datasets for resized images. The ICT method is runned, as proposed in the original paper, with its top-K parameter (cf. Section 3.4) set to 50. We investigate the influence of the top-K parameter in Table 6.

To measure inpainting quality, we take into account three factors: the similarity to the ground truth, the realism of inpainting outputs, and the diversity of those outputs. Definitions and details on the metrics for each factor can be found in Section 4. Note that contrary to [Zheng et al., 2019], we do not use here any discriminator score to select the best samples before evaluation.

In each table, the best and second-best results by column are in bold and underlined, respectively.

5.2.1 Proximity to Ground Truth

First, to measure the similarity between the inpainting results and the ground truth (GT), we use the following metrics: peak signal-to-noise ratio (PSNR), $L^1$ loss, and structural similarity (SSIM). For each input image to be inpainted, those metrics are averaged on the set of inpainted results.

Note that all the compared methods enforce somehow, in their training loss, similarity between the reconstructed image and the ground truth, either at pixel or at feature level. From the results in Tables 3, 4, 5, 8, and 9 we observe that, on all datasets, ICT and PIC obtained slightly lower scores than BAT and DSI-VQVAE in terms of GT similarity. A possible explanation for this performance gap is that, these two methods, contrary to the two others, consider a reconstruction loss only at image level and not at feature level. Being similar at feature levels encourages to generate images having similar low-level (pixels, contours, etc.) and higher-level semantics to the ground-truth.
Table 3 Quantitative comparison of four pluralistic image inpainting methods (PIC, DSI-VQVAE, ICT and BAT) on Celeba-HQ and for different kind of masks (central, random regular, random irregular and from [Liu et al., 2018]). Best and second best results by column are in bold and underlined, respectively.

| Mask     | Method         | Similarity to GT | Realism  | Diversity |
|----------|----------------|------------------|----------|-----------|
|          |                | PSNR↑ SSIM↑ L¹↓ | MIS↑ FID↓ LPIPS↑ |
| Irregular| DSI-VQVAE      | 35.49 0.968 1.41 | 0.0216 11.0 0.0081 |
| < 20%    | ICT            | 34.72 0.968 2.09 | 0.0200 9.84 0.0084 |
|          | BAT            | 36.25 0.974 1.20 | 0.0208 9.90 0.0056 |
| Irregular| DSI-VQVAE      | 27.36 0.888 4.06 | 0.0223 28.8 0.0357 |
| 20% - 40%| ICT            | 26.83 0.891 4.71 | 0.0189 26.7 0.0383 |
|          | BAT            | 27.28 0.900 3.85 | 0.0214 20.7 0.0269 |
| Irregular| DSI-VQVAE      | 22.53 0.770 9.01 | 0.0156 51.9 0.0916 |
| 40% - 60%| ICT            | 21.92 0.773 9.82 | 0.0153 50.7 0.0970 |
|          | BAT            | 22.35 0.787 8.91 | 0.0183 39.7 0.0731 |
| Central  | DSI-VQVAE      | 25.25 0.880 5.08 | 0.0210 21.7 0.0243 |
| 128 × 128| ICT            | 24.45 0.872 6.06 | 0.0170 27.3 0.0486 |
|          | BAT            | 25.10 0.882 5.21 | 0.0218 21.5 0.0365 |
| Random   | DSI-VQVAE      | 24.98 0.850 6.46 | 0.0200 30.5 0.0642 |
| regular  | ICT            | 24.51 0.852 7.24 | 0.0180 31.3 0.0665 |
|          | BAT            | 24.85 0.860 6.52 | 0.0209 24.6 0.0541 |
| Random   | DSI-VQVAE      | 24.27 0.785 7.56 | 0.0167 58.8 0.0744 |
| irregular| ICT            | 23.26 0.781 9.26 | 0.0148 52.2 0.0855 |
|          | BAT            | 24.36 0.810 7.13 | 0.0186 40.8 0.0495 |
| Average  | DSI-VQVAE      | 26.65 0.857 5.60 | 0.0195 33.8 0.0497 |
|          | ICT            | 25.95 0.855 6.53 | 0.0175 33.0 0.0575 |
|          | BAT            | 26.70 0.869 5.47 | 0.0203 26.2 0.0410 |

5.2.2 Perceptual Quality

Second, to measure realism in the outputs, we measure perceptual quality by using Modified Inception Score (MIS) and Fréchet Inception Distance (FID) metrics (defined by [26] and [27], respectively). These two metrics are computed directly on the whole sets of generated or ground truth images.

BAT, ICT, and DSI-VQVAE are the methods that provide the best scores on average on all datasets. On the opposite, PIC gives the worst results quantitatively and, as we will see later, also qualitatively. We argue that a possible reason for the superior performance of BAT, ICT, and DSI-VQVAE is that, with different strategies, they separate the tasks of texture and structure recovery. Each task is handled
Table 4: Quantitative comparison of four pluralistic image inpainting methods (PIC, DSI-VQVAE, ICT and BAT) on 256 × 256 center-cropped images from Places2, for different kind of masks (central, random regular, random irregular and from [Liu et al., 2018]).

| Mask      | Method      | Similarity to GT | Realism | Diversity |
|-----------|-------------|------------------|---------|-----------|
|           |             | PSNR↑ SSIM↑ L↑↓ | MIS↑    | FID↓      | LPIPS↑ |
| Irregular | PIC         | 30.48 0.937 2.02 0.0507 36.8 0.0050 |
| < 20%     | DSI-VQVAE   | 31.58 0.952 2.71 0.0482 19.3 0.0187 |
|           | ICT         | 29.86 0.943 3.64 0.0465 22.8 0.0198 |
|           | BAT         | 32.20 0.957 1.83 0.0463 14.2 0.0158 |
| Irregular | DSI-VQVAE   | 24.20 0.844 6.14 0.0438 63.6 0.0707 |
| 20% - 40% | ICT         | 23.08 0.831 8.05 0.0428 70.0 0.0769 |
|           | BAT         | 24.10 0.853 6.14 0.0423 53.2 0.0671 |
| Irregular | DSI-VQVAE   | 20.34 0.703 12.52 0.0398 110.2 0.1566 |
| 40% - 60% | ICT         | 19.49 0.686 14.66 0.0371 128.7 0.1668 |
|           | BAT         | 19.98 0.705 13.10 0.0364 107.0 0.1610 |
| Central   | DSI-VQVAE   | 20.98 0.812 9.00 0.0435 96.8 0.1080 |
| 128 × 128 | ICT         | 20.93 0.812 10.22 0.0476 92.2 0.1204 |
|           | BAT         | 21.20 0.822 8.76 0.0442 81.8 0.1190 |
| Random    | DSI-VQVAE   | 22.36 0.805 9.21 0.0412 75.8 0.1167 |
| regular   | ICT         | 21.75 0.796 10.77 0.0405 87.1 0.1237 |
|           | BAT         | 22.34 0.808 9.15 0.0436 76.6 0.1200 |
| Random    | DSI-VQVAE   | 20.86 0.658 12.80 0.0255 165.4 0.0979 |
| irregular | ICT         | 20.07 0.681 14.14 0.0333 131.9 0.1548 |
|           | BAT         | 20.85 0.708 12.00 0.0374 103.2 0.1454 |
| Average   | DSI-VQVAE   | 23.51 0.804 8.44 0.0418 76.9 0.1052 |
|           | ICT         | 22.53 0.792 10.25 0.0413 88.9 0.1107 |
|           | BAT         | 23.44 0.809 8.97 0.0417 72.7 0.1047 |

with a specific subnetwork, first reconstructing structures that then guide the texture recovery. From a more practical point of view, BAT and ICT use transformers for global structure understanding and high-level semantics at a coarse resolution, and CNNs for generating textures at the original resolution. DSI-VQVAE incorporates the multi-scale hierarchical organization of VQ-VAE where the information corresponding to the texture is disentangled from the one about structure and geometry. Accordingly, DSI-VQVAE incorporates two different generators respectively devoted to both levels (cf. Section 3). Although DSI-VQVAE and PIC are VAE-based methods, DSI-VQVAE has the advantage that first, at low resolution, it proposes diverse completions of structure inside the hole. These different structures then guide the completion of texture at high resolution. PIC does not have this global structure completion (at least, not explicitly). All in all, splitting the estimation of coarse and
Table 5: Quantitative comparison of three pluralistic image inpainting methods (PIC, DSI-VQVAE and ICT) on 256×256 center-cropped images from ImageNet, for different kind of masks (central, random regular, random irregular and from [Liu et al., 2018]).

| Mask                | Method      | Similarity to GT | Realism | Diversity |
|---------------------|-------------|------------------|---------|-----------|
|                     |             | PSNR ↑ | SSIM ↑ | L1 ↓ | MIS↑ | FID↓ | LPIPS↑ |             |
| Irregular < 20%     | PIC         | 30.33  | 0.941 | 2.02 | 0.2416 | 20.2 | 0.0036 |
|                     | DSI-VQVAE   | 30.44  | 0.946 | 2.38 | 0.2361 | 12.1 | 0.0199 |
|                     | ICT         | 29.23  | 0.940 | 3.98 | 0.2323 | 10.7 | 0.0185 |
| Irregular 20% – 40% | PIC         | 23.02  | 0.797 | 7.37 | 0.1709 | 83.7 | 0.0289 |
|                     | DSI-VQVAE   | 22.98  | 0.809 | 7.56 | 0.2015 | 53.4 | 0.0855 |
|                     | ICT         | 22.24  | 0.802 | 9.23 | 0.1970 | 24.9 | 0.0771 |
| Irregular 40% – 60% | PIC         | 18.33  | 0.623 | 16.34 | 0.0792 | 183.9 | 0.1269 |
|                     | DSI-VQVAE   | 18.92  | 0.651 | 14.82 | 0.1192 | 126.3 | 0.1907 |
|                     | ICT         | 18.52  | 0.646 | 16.41 | 0.1329 | 101.7 | 0.1700 |
| Central 128 x 128   | PIC         | 19.87  | 0.794 | 9.77 | 0.1591 | 95.8 | 0.1067 |
|                     | DSI-VQVAE   | 20.06  | 0.795 | 9.99 | 0.1754 | 85.6 | 0.1291 |
|                     | ICT         | 20.34  | 0.795 | 10.76 | 0.1753 | 73.8 | 0.1162 |
| Random regular      | PIC         | 19.81  | 0.737 | 13.24 | 0.0934 | 129.2 | 0.1027 |
|                     | DSI-VQVAE   | 20.54  | 0.756 | 11.52 | 0.1305 | 89.3 | 0.1540 |
|                     | ICT         | 20.32  | 0.752 | 12.76 | 0.1420 | 77.6 | 0.1360 |
| Random irregular    | PIC         | 19.51  | 0.598 | 14.78 | 0.0645 | 193.0 | 0.0982 |
|                     | DSI-VQVAE   | 19.81  | 0.636 | 14.04 | 0.1147 | 136.8 | 0.1757 |
|                     | ICT         | 19.02  | 0.628 | 16.08 | 0.1363 | 108.5 | 0.1574 |
| Average             | PIC         | 21.81  | 0.748 | 10.59 | 0.1347 | 117.6 | 0.0735 |
|                     | DSI-VQVAE   | 22.13  | 0.765 | 10.07 | 0.1629 | 83.9 | 0.1258 |
|                     | ICT         | 21.61  | 0.761 | 11.54 | 0.1693 | 66.2 | 0.1125 |

Table 6: Influence of the top-K parameter on the ICT results. Results obtained on Places2 dataset, with central mask.

| top-K | Similarity to GT | Realism | Diversity |
|-------|------------------|---------|-----------|
|       | PSNR ↑ | SSIM ↑ | L1 ↓ | MIS↑ | FID↓ | LPIPS↑ |
| 5     | 21.76  | 0.820 | 6.52 | 0.0510 | 87.6 | 0.0854 |
| 25    | 21.16  | 0.813 | 10.03 | 0.0495 | 90.2 | 0.1146 |
| 50    | 20.93  | 0.812 | 10.22 | 0.0476 | 92.2 | 0.1204 |

Note also that BAT is the method that achieves the best scores in terms of realism. Indeed, as explained before, autoregressive transformers have the ability to model longer dependencies across the image than CNN-based methods, which can be crucial for image inpainting. Note that BAT outperforms the other transformer-based method ICT, especially on irregular masks and large holes. As explained in Section 3.3, one can explain this difference by the fact that BAT was trained, not only with bidirectional attention but also with autoregressive sampling. Therefore,
it creates better consistency of the reconstructed structures, especially for large missing regions. The very good results of the DSI-VQVAE method also proves that autoregressive modeling is a performant strategy for realistic image inpainting.

Finally, one can observe the influence of the complexity of the training dataset on the performance. Notice that the underlying probability distribution of CelebA-HQ dataset is semantically less complex and diverse than the one of Places2 and ImageNet and, thus, training is more difficult in the latter cases. We hypothesize that this affects both inpainting quality and inpainting diversity. Regarding quality, for example, the average FID score on all the studied methods trained on CelebA is equal to 33.55, while in the case of Places2 and Imagenet is equal to 86.92 and 128.43, respectively. This gives us an idea of the difference in complexity for each particular dataset.

5.2.3 Inpainting diversity

To measure diversity, we rely on the LPIPS metric. The higher the LPIPS is, the more diverse are the outputs. For each generated sample, we compute the LPIPS distance with another sample randomly selected from the other 24 results from the same corrupted image. The reported LPIPS score corresponds to this distance averaged over the 2500 selected pairs.

First and foremost, from the range of LPIPS values on the different datasets, one can again observe the influence of the complexity of the training dataset. CelebA-HQ dataset is semantically more constraint and less complex than the one of Places2 and ImageNet, leading to lower diversity in the outputs. Indeed, the LPIPS is, in average, \( \approx 2 \) times smaller on CelebA-HQ than on Places2 or ImageNet. Similarly, as expected, all the methods create more diverse samples on larger holes than on smaller holes.

These observations argue for the existence of a trade-off between inpainting quality and inpainting diversity. The harder the inpainting problem gets (on a more complex dataset or for a larger hole), the more diverse outputs will be created. This trade-off, already highlighted in [Yu et al., 2021], also arises when parametrizing a method itself. We study in Table 6 the influence of the top-K parameter on the performance of the ICT algorithm. One can observe that using a smaller K creates outputs that are, on the one hand, closer to GT and more realistic but, on the other hand, less diverse.

PIC is the method giving the less diverse results on all datasets. One reason could be the aforementioned disentanglement of structure and texture of BAT, DSI-VQVAE and ICT. In practice, these three methods first attempt to produce a multiplicity of coherent structures and then fill each of the sampled structure with a deterministic texture generator. This divide-and-conquer approach makes easier the creation of diversity as it is only performed on low-resolution structures and not on the whole reconstructed output.

ICT slightly outperforms DSI-VQVAE and BAT in terms of LPIPS on the Celeba-HQ testing images. Remind that for this experiment ICT was trained on the more
diverse FFHQ dataset. This observation highlights again the influence of the training dataset on the capacity of the model to create diverse outputs.

5.3 Qualitative Performance

For qualitative comparison, we select for each method the 5 samples with the highest discriminator score out of the 25 generated samples. We use pretrained discriminators given by each of the models, i.e., for PIC, the discriminator of the generative pipeline, for DSI-VQVAE, the discriminator of the texture generation module, and for ICT and BAT, the discriminator of the upsampling module. We perform this comparison on a representative selection of testing images and masks. Figure 9, 10, 11, and 12 show some results on CelebA-HQ, Places2 and ImageNet datasets for the methods PIC, DSI-VQVAE, ICT and BAT. BAT does not provide weights for ImageNet. Remind that ICT was not trained on CelebA-HQ but on FFHQ. Additional visual results are also given in the Appendix.

At first glance, we observe that DSI-VQVAE, ICT and BAT provide more plausibly visual results than PIC. PIC struggles to recover information on less constrained datasets like Places2 and Imagenet, and creates strong artifacts when applied to large missing regions (see second examples in Figures 10 and 12). Among these methods, BAT and ICT propose the most realistic outputs. For instance, in Figure 9, PIC generates results that do not maintain the proportions and harmony of a face (see the second example). DSI-VQVAE neither has a full understanding of the image: for example, in the second example in Figure 9 and the third example in Figure 12, one eye is visible in the input image but the other is not. On the opposite, transformer-based methods are able to reconstruct a left eye similar to the right. This can be explained by the capability of transformers to have a global structure understanding and high-level semantics. More examples strengthening this observation are the first example of Figure 10 where the inpainting of the snow is sometimes no realistic, or all the ImageNet results Figure 12.

When images contain strong structures, like Figure 10 and 11, transformer-based methods again estimate more realistic reconstructions. This can be explained by the fact that they include previously predicted tokens in the training objective and thus, global consistency is imposed over the results. This consistency shall avoid problems in the center of big holes. In some situations, such as the middle example in Figure 10, the structure and texture disentanglement of DSI-VQVAE also provides good reconstructions.

Notice that, qualitatively speaking, although ICT and BAT were only trained on irregular masks, we do not observe a drop in performance while performing inpainting on regular masks. This shows the capacity of those methods to generalize to unseen type of mission regions.

In terms of diversity, transformer-based methods are visually more diverse. For example, in Figure 9, each transformer-based inpainted face corresponds to a different expression or different person, while in the case of DSI-VQVAE all generated
Output on 256 × 256 images masked with 128 × 128 center hole

Output on 256 × 256 images masked with random regular hole

Output on 256 × 256 images masked with random irregular hole

Fig. 9 Diverse inpainting output on 256× 256 images from Celeba dataset with center, random regular and random irregular masks. For each method, out of 25 generated samples, the 5 samples with highest discriminator score are displayed.
Fig. 10 Diverse inpainting output on 256×256 images from Places2 dataset with center and irregular masks with various proportion of hidden pixels. For each method, out of 25 generated samples, the 5 samples with highest discriminator score are displayed.
Fig. 11 Diverse inpainting output on 256×256 images from Places2 dataset with center and irregular masks with various proportion of hidden pixels. For each method, out of 25 generated samples, the 5 samples with highest discriminator score are displayed.
faces are very similar. Also, in Figure 11 even if one could imagine the result quite deterministic, ICT and BAT aims to propose multiple possibilities. Note the multiplicity of structures obtained by ICT compared to DSI-VQVAE and PIC in the chest of the dog or the skyline in Figure 12.

Regarding the difference across datasets, while methods trained on CelebA-HQ all obtain satisfactory results (Figure 9), results on Places2 (Figure 10) and ImageNet (Figure 12) are often not visually satisfactory. Also, as already noticed numerically, diversity is less visible on these two datasets. This is probably because the models have difficulties learning the underlying multimodal distribution of these complex

Fig. 12 Diverse inpainting output on 256×256 images from ImageNet dataset with center and irregular masks with various proportion of hidden pixels. For each method, out of 25 generated samples, the 5 samples with highest discriminator score are displayed.
and diverse datasets. This demonstrates the need for further research on the topic to be able to deal with real inpainting scenarios.

**Computational Time.** Despite image quality, an important aspect that should be considered when choosing an inpainting method is its inference time. In Table 7, for the four analyzed methods, we give the average runtime to sample one inpainting result from a central hole on a $256 \times 256$ input image. We run the experiments on a single P100 GPU. Despite showing lower inpainting quality or diversity (see before), PIC is tremendously faster to run than all the other methods (~100 times faster than DSI-VQVAE and ICT and ~ 50 times faster than BAT). While providing good results, inference time on autoregressive or transformer-based methods can be prohibitive for time-restricted applications.

**Table 7** Average runtime to sample one inpainting result on a single P100 GPU for the four compared methods. Experiments conducted for central masks

| Method   | Time (s) |
|----------|----------|
| PIC      | 0.4      |
| DSI-VQVAE| 55       |
| ICT      | 43       |
| BAT      | 21       |

6 Conclusions

In this chapter, we have tackled the question of whether generative methods are a suitable strategy to obtain multiple solutions to problems that do not have a unique solution. By focusing on the inpainting problem, we have reviewed the main generative models and recent learning-based image completion methods for multiple and diverse inpainting. We have compared the methods with available code and model weights on three public datasets. We have shown that the transformer-based method BAT (or BAT-Fill) and the VQ-VAE-based method DSI-VQVAE provide the best results in both inpainting quality and multiple inpainting diversity. This is true both quantitatively and qualitatively. Our analysis highlights that their advantageous results are due to their strategy that consists in, first, sampling multiple structures inside the missing regions, and second, generating textures at higher resolution in a deterministic way. The PIC method is however computationally way faster than the concurrence. Moreover, our analysis shows that the multiple inpainting problem is not solved yet. The difficulty of learning the probability distribution depending on the training dataset is also evident from our study. Therefore, we argue that most efforts should be made on improving and exploring new generative strategies to enhance both the quality and diversity of the solutions of this ill-posed inverse problem with multiple solutions. Finally, the computational burden of some of the
transformers-based or autoregressive methods is prohibitive for sampling a high number of solutions in reasonable time. We think that this limitation has been overviewed for the purpose of image quality but should be now primarily addressed.

Acknowledgements PV, CB and AB acknowledge the EU Horizon 2020 research and innovation programme NoMADS (Marie Skłodowska-Curie grant agreement No 777826). SP acknowledges the Leverhulme Trust Research Project Grant “Unveiling the invisible: Mathematics for Conservation in Arts and Humanities”. CB and PV also acknowledge partial support by MICINN/FEDER UE project, ref. PGC2018-098625-B-I00, and RED2018-102511-T. AB also acknowledges the French Research Agency through the PostProdLEAP project (ANR-19-CE23-0027-01). SH acknowledges the French ministry of research through a CDSN grant of ENS Paris-Saclay.

References

Arias et al., 2011. Arias, P., Facciolo, G., Caselles, V., and Sapiro, G. (2011). A variational framework for exemplar-based image inpainting. International Journal of Computer Vision, 93(3):319–347.

Arjovsky et al., 2017. Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In International conference on machine learning, pages 214–223. PMLR.

Arridge et al., 2019. Arridge, S., Maass, P., Öktem, O., and Schönlieb, C.-B. (2019). Solving inverse problems using data-driven models. Acta Numerica, 28:1–174.

Aujol et al., 2010. Aujol, J.-F., Ladjal, S., and Masnou, S. (2010). Exemplar-based inpainting from a variational point of view. SIAM Journal on Mathematical Analysis, 42(3):1246–1285.

Baatz et al., 2008. Baatz, W., Fornasier, M., Markowich, P. A., and bibiane Schönlieb, C. (2008). Inpainting of ancient austrian frescoes. In Conference Proceedings of Bridges, pages 150–156.

Ballester et al., 2001. Ballester, C., Bertalmío, M., Caselles, V., Sapiro, G., and Verdera, J. (2001). Filling-in by joint interpolation of vector fields and gray levels. IEEE Transactions on Image Processing, 10(8):1200–1211.

Bao et al., 2017. Bao, J., Chen, D., Wen, F., Li, H., and Hua, G. (2017). Cvae-gan: fine-grained image generation through asymmetric training. In Proceedings of the IEEE international conference on computer vision, pages 2745–2754.

Barratt et al., 2018. Barratt, S. and Sharma, R. (2018). A note on the inception score. arXiv preprint arXiv:1801.01973.

Bertalmío et al., 2001. Bertalmío, M., Bertozzi, A., and Sapiro, G. (2001). Navier-stokes, fluid dynamics, and image and video inpainting. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. IEEE Comput. Soc.

Bevilacqua et al., 2017. Bevilacqua, M., Aujol, J.-F., Biasutti, P., Brédif, M., and Bugeau, A. (2017). Joint inpainting of depth and reflectance with visibility estimation. ISPRS Journal of Photogrammetry and Remote Sensing, 125:16–32.

Biasutti et al., 2019. Biasutti, P., Aujol, J.-F., Brédif, M., and Bugeau, A. (2019). Diffusion and inpainting of reflectance and height LiDAR orthoimages. Computer Vision and Image Understanding, 179:31–40.
An Analysis of Generative Methods for Multiple Image Inpainting

Bornard et al., 2002. Bornard, R., Lecan, E., Laborelli, L., and Chenot, J.-H. (2002). Missing data correction in still images and image sequences. In Proceedings of the tenth ACM international conference on Multimedia - MULTIMEDIA '02. ACM Press.

Bornemann and März, 2007. Bornemann, F. and März, T. (2007). Fast image inpainting based on coherence transport. Journal of Mathematical Imaging and Vision, 28(3):259–278.

Bowman et al., 2015. Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2015). Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349.

Buysens et al., 2015. Buysens, P., Daisy, M., Tschumperle, D., and Lezoray, O. (2015). Exemplar-based inpainting: Technical review and new heuristics for better geometric reconstructions. IEEE Transactions on Image Processing, pages 1–1.

Cai et al., 2015. Cai, N., Su, Z., Lin, Z., Wang, H., Yang, Z., and Ling, B. W.-K. (2015). Blind inpainting using the fully convolutional neural network. The Visual Computer, 33(2):249–261.

Cai and Wei, 2020. Cai, W. and Wei, Z. (2020). Pigan: Generative adversarial networks for pluralistic image inpainting. IEEE Access, 8:48451–48463.

Calatroni et al., 2018. Calatroni, L., d’Autume, M., Hocking, R., Panayotova, S., Parisotto, S., Ricciardi, P., and Schönlieb, C.-B. (2018). Unveiling the invisible: mathematical methods for restoring and interpreting illuminated manuscripts. Heritage Science, 6(1).

Chan and Shen, 2001. Chan, T. F. and Shen, J. (2001). Nontexture inpainting by curvature-driven diffusions. Journal of Visual Communication and Image Representation, 12(4):436–449.

Chen et al., 2018. Chen, X., Mishra, N., Rohaninejad, M., and Abbeel, P. (2018). Pixelsnail: An improved autoregressive generative model. In International Conference on Machine Learning, pages 864–872. PMLR.

Chen et al., 2012. Chen, Y., Li, Y., Guo, H., Hu, Y., Luo, L., Yin, X., Gu, J., and Tounoulin, C. (2012). CT metal artifact reduction method based on improved image segmentation and sinogram inpainting. Mathematical Problems in Engineering, 2012:1–18.

Criminisi et al., 2004. Criminisi, A., Perez, P., and Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on Image Processing, 13(9):1200–1212.

Dahl et al., 2017. Dahl, R., Norouzi, M., and Shlens, J. (2017). Pixel recursive super resolution. In Proceedings of the IEEE international conference on computer vision, pages 5439–5448.

Dapogny et al., 2020. Dapogny, A., Cord, M., and Pérez, P. (2020). The missing data encoder: Cross-channel image completion with hide-and-seek adversarial network. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07):10688–10695.

Demanet et al., 2003. Demanet, L., Song, B., and Chan, T. (2003). Image inpainting by correspondence maps: a deterministic approach. Applied and Computational Mathematics, 1100:217–50.

Devlin et al., 2018. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dosovitskiy et al., 2020. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Efros and Leung, 1999. Efros, A. and Leung, T. (1999). Texture synthesis by non-parametric sampling. In Proceedings of the Seventh IEEE International Conference on Computer Vision. IEEE.

Eller and Fornasier, 2016. Eller, M. and Fornasier, M. (2016). Rotation invariance in exemplar-based image inpainting. In Variational Methods, pages 108–183. De Gruyter.

Esedoglu and Shen, 2002. Esedoglu, S. and Shen, J. (2002). Digital inpainting based on the mumford-shah-euler image model. European Journal of Applied Mathematics, 13(04).
Fawzi et al., 2016. Fawzi, A., Samulowitz, H., Turaga, D., and Frossard, P. (2016). Image inpainting through neural networks hallucinations. In *IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop*, pages 1–5. IEEE.

Fedorov et al., 2016. Fedorov, V., Arias, P., Facciolo, G., and Ballester, C. (2016). Affine invariant self-similarity for exemplar-based inpainting. In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications.

Goodfellow et al., 2014. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Grossauer, 2006. Grossauer, H. (2006). Inpainting of movies using optical flow. In *Mathematics in industry*, pages 151–162. Springer Berlin Heidelberg.

Grossauer and Scherzer, 2003. Grossauer, H. and Scherzer, O. (2003). Using the complex ginzburg-landau equation for digital inpainting in 2d and 3d. In *Scale Space Methods in Computer Vision*, pages 225–236. Springer Berlin Heidelberg.

Guadarrama et al., 2017. Guadarrama, S., Dahl, R., Bieber, D., Norouzi, M., Shlens, J., and Murphy, K. (2017). Pixcolor: Pixel recursive colorization. *arXiv preprint arXiv:1705.07208*.

Guillemot and Le Meur, 2014. Guillemot, C. and Le Meur, O. (2014). Image inpainting: Overview and recent advances. *IEEE Signal Processing Magazine*, 31(1):127–144.

Gulrajani et al., 2017. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*.

Guo et al., 2019. Guo, Z., Chen, Z., Yu, T., Chen, J., and Liu, S. (2019). Progressive image inpainting with full-resolution residual network. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM ’19, page 2496–2504, New York, NY, USA. Association for Computing Machinery (ACM).

Han et al., 2019. Han, X., Wu, Z., Huang, W., Scott, M. R., and Davis, L. S. (2019). Finet: Compatible and diverse fashion image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4481–4491.

Hays and Efros, 2007. Hays, J. and Efros, A. A. (2007). Scene completion using millions of photographs. *ACM Transactions on Graphics*, 26(3):87–94.

Hervieu et al., 2010. Hervieu, A., Papadakis, N., Bugeau, A., Gargallo, P., and Caselles, V. (2010). Stereoscopic image inpainting: distinct depth maps and images inpainting. In *2010 20th International Conference on Pattern Recognition*, pages 4101–4104. IEEE.

Heusel et al., 2017. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*.

Iizuka et al., 2017. Iizuka, S., Simo-Serra, E., and Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4).

Johnson et al., 2016. Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer.

Kang et al., 2002. Kang, S. H., Chan, T., and Soatto, S. (2002). Inpainting from multiple views. In *Proceedings. First International Symposium on 3D Data Processing Visualization and Transmission*. IEEE Comput. Soc.

Karras et al., 2018. Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*.

Karras et al., 2019. Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410.

Kettunen et al., 2019. Kettunen, M., Härkönen, E., and Lehtinen, J. (2019). E-lpips: robust perceptual image similarity via random transformation ensembles. *arXiv preprint arXiv:1906.03973*.

Kingma and Welling, 2013. Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. In *International Conference on Learning Representations*. 
An Analysis of Generative Methods for Multiple Image Inpainting

Kingma et al., 2019. Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Köhler et al., 2014. Köhler, R., Schulter, C., Schölkopf, B., and Harmeling, S. (2014). Mask-specific inpainting with deep neural networks. In Jiang, X., Hornerg, J., and Koch, R., editors, *Pattern Recognition*, pages 523–534, Cham. Springer International Publishing.

Kumar et al., 2021. Kumar, M., Weissenborn, D., and Kalchbrenner, N. (2021). Colorization transformer. arXiv preprint arXiv:2102.04432.

Kumar et al., 2016. Kumar, V., Mukherjee, J., and Mandal, S. K. D. (2016). Image inpainting through metric labeling via guided patch mixing. *IEEE Transactions on Image Processing*, 25(11):5212–5226.

Lahiri et al., 2020. Lahiri, A., Jain, A. K., Agrawal, S., Mitra, P., and Biswas, P. K. (2020). Prior Guided GAN Based Semantic Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13696–13705.

Le Meur et al., 2013. Le Meur, O., Ebdelli, M., and Guillemot, C. (2013). Hierarchical super-resolution-based inpainting. *IEEE Transactions on Image Processing*, 22(10):3779–3790.

Lempitsky et al., 2018. Lempitsky, V., Vedaldi, A., and Ulyanov, D. (2018). Deep image prior. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454. IEEE.

Li et al., 2019. Li, J., He, F., Zhang, L., Du, B., and Tao, D. (2019). Progressive reconstruction of visual structure for image inpainting. In *2019 IEEE/CVF International Conference on Computer Vision*. IEEE.

Li et al., 2020. Li, J., Wang, N., Zhang, L., Du, B., and Tao, D. (2020). Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7760–7768.

Liao et al., 2018. Liao, L., Hu, R., Xiao, J., and Wang, Z. (2018). Edge-aware context encoder for image inpainting. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3156–3160. IEEE.

Liu et al., 2018. Liu, H., Wan, Z., Huang, W., Song, Y., Han, X., and Liao, J. (2021). Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9371–9381. IEEE.

Masnou and Morel, 1998. Masnou, S. and Morel, J.-M. (1998). Level lines based disocclusion. In *Proceedings 1998 International Conference on Image Processing, ICIP98 (Cat. No.98CB36269)*. IEEE Comput. Soc.

Nazeri et al., 2019. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., and Ebrahimi, M. (2019). Edge-Connect: Generative image inpainting with adversarial edge learning. In *The IEEE International Conference on Computer Vision Workshops*.

Newson et al., 2014. Newson, A., Almansa, A., Fradet, M., Gousseau, Y., and Pérez, P. (2014). Video inpainting of complex scenes. *SIAM Journal on Imaging Sciences*, 7(4):1993–2019.

Nitzberg et al., 1993. Nitzberg, M., Mumford, D., and Shiota, T. (1993). *Filtering, Segmentation and Depth*. Springer Berlin Heidelberg.
Oord et al., 2018. Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., et al. (2018). Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR.

Oord et al., 2016. Oord, A. v. d., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., and Kavukcuoglu, K. (2016). Conditional image generation with pixelcnn decoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4797–4805.

Papafitsoros and Schönlieb, 2013. Papafitsoros, K. and Schönlieb, C. B. (2013). A combined first and second order variational approach for image reconstruction. *Journal of Mathematical Imaging and Vision*, 48(2):308–338.

Parisotto et al., 2020. Parisotto, S., Lellmann, J., Masnou, S., and Schönlieb, C.-B. (2020). Higher-order total directional variation: Imaging applications. *SIAM Journal on Imaging Sciences*, 13(4):2063–2104.

Parisotto et al., 2022. Parisotto, S., Vitoria, P., Ballester, C., Beauge, A., Reynolds, S., and Schönlieb, C.-B. (2022). *The Art of Inpainting - A Monograph on Mathematical Methods for the Virtual Restoration of Illuminated Manuscripts*. (submitted).

Park et al., 2019. Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346.

Pathak et al., 2016. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544. IEEE.

Peng et al., 2021. Peng, J., Liu, D., Xu, S., and Li, H. (2021). Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784.

Peter and Weickert, 2015. Peter, P. and Weickert, J. (2015). Compressing images with diffusion-and exemplar-based inpainting. In *Lecture Notes in Computer Science*, pages 154–165. Springer International Publishing.

Razavi et al., 2019. Razavi, A., van den Oord, A., and Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876.

Ren et al., 2015. Ren, J. S., Xu, L., Yan, Q., and Sun, W. (2015). Shepard convolutional neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 1 of *NIPS’15*, page 901–909. Cambridge, MA, USA. The MIT Press.

Ren et al., 2019. Ren, Y., Yu, X., Zhang, R., Li, T. H., Liu, S., and Li, G. (2019). Structure-Flow: Image inpainting via structure-aware appearance flow. In *2019 IEEE/CVF International Conference on Computer Vision*, pages 181–190. IEEE.

Rosca et al., 2017. Rosca, M., Lakshminarayanan, B., Warde-Farley, D., and Mohamed, S. (2017). Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*.

Royer et al., 2017. Royer, A., Kolesnikov, A., and Lampert, C. H. (2017). Probabilistic image colorization. *arXiv preprint arXiv:1705.04258*.

Russakovsky et al., 2015. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Ružič et al., 2011. Ružič, T., Cornelis, B., Platiša, L., Pižurica, A., Dooms, A., Philips, W., Martens, M., Mey, M. D., and Daubechies, I. (2011). Virtual restoration of the ghent altarpiece using crack detection and inpainting. In *Advanced Concepts for Intelligent Vision Systems*, pages 417–428. Springer Berlin Heidelberg.

Salimans et al., 2016. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *Advances in Neural Information Processing Systems*.

Schönlieb, 2015. Schönlieb, C.-B. (2015). *Partial Differential Equation Methods for Image Inpainting*. Cambridge University Press.
Shen and Chan, 2002. Shen, J. and Chan, T. F. (2002). Mathematical models for local nontexture inpaintings. *SIAM Journal on Applied Mathematics*, 62(3):1019–1043.

Shen et al., 2003. Shen, J., Kang, S. H., and Chan, T. F. (2003). Euler’s elastica and curvature-based inpainting. *SIAM Journal on Applied Mathematics*, 63(2):564–592.

Sohn et al., 2015. Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491.

Sun et al., 2005. Sun, J., Yuan, L., Jia, J., and Shum, H.-Y. (2005). Image completion with structure propagation. *ACM Transactions on Graphics*, 24(3):861–868.

Tai et al., 2007. Tai, X.-C., Osher, S., and Holm, R. (2007). Image inpainting using a TV-stokes equation. In *Image Processing Based on Partial Differential Equations*, pages 3–22. Springer Berlin Heidelberg.

Tovey et al., 2019. Tovey, R., Benning, M., Brune, C., Lagerwerf, M. J., Collins, S. M., Leary, R. K., Midgley, P. A., and Schönlieb, C.-B. (2019). Directional sinogram inpainting for limited angle tomography. *Inverse Problems*, 35(2):024004.

Tschumperle and Deriche, 2005. Tschumperle, D. and Deriche, R. (2005). Vector-valued image regularization with PDEs: a common framework for different applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):506–517.

van den Oord et al., 2017. van den Oord, A., Vinyals, O., and Kavukcuoglu, K. (2017). Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6309–6318.

Van Oord et al., 2016. Van Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. (2016). Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756. PMLR.

Vaswani et al., 2017. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vitoria and Ballester, 2019. Vitoria, P. and Ballester, C. (2019). Automatic flare spot artifact detection and removal in photographs. *Journal of Mathematical Imaging and Vision*, 61(4):515–533.

Vitoria et al., 2019. Vitoria, P., Sintes, J., and Ballester, C. (2019). Semantic image inpainting through improved Wasserstein generative adversarial networks. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 4 of VISAPP, pages 249–260. INSTICC, SciTePress.

Vitoria et al., 2020. Vitoria, P., Sintes, J., and Ballester, C. (2020). Semantic image completion through an adversarial strategy. In *Communications in Computer and Information Science*, pages 520–542. Springer International Publishing.

Vo et al., 2018. Vo, H. V., Duong, N. Q. K., and Pérez, P. (2018). Structural inpainting. In *2018 ACM Multimedia Conference on Multimedia Conference*, MM ’18, pages 1948–1956, New York, NY, USA. Association for Computing Machinery (ACM).

Wan et al., 2021. Wan, Z., Zhang, J., Chen, D., and Liao, J. (2021). High-fidelity pluralistic image completion with transformers. *arXiv preprint arXiv:2103.14031*.

Wang et al., 2018. Wang, Y., Tao, X., Qi, X., Shen, X., and Jia, J. (2018). Image inpainting via generative multi-column convolutional neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 331–340. Curran Associates, Inc.

Wexler et al., 2004. Wexler, Y., Shechtman, E., and Irani, M. (2004). Space-time video completion. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004*. IEEE.

Xiong et al., 2019. Xiong, W., Yu, J., Lin, Z., Yang, J., Lu, X., Barnes, C., and Luo, J. (2019). Foreground-aware image inpainting. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5840–5848. IEEE.

Yan et al., 2018. Yan, Z., Li, X., Li, M., Zuo, W., and Shan, S. (2018). Shift-net: Image inpainting via deep feature rearrangement. In *Computer Vision - ECCV 2018*, pages 3–19. Springer International Publishing.
Yang et al., 2017. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., and Li, H. (2017). High-resolution image inpainting using multi-scale neural patch synthesis. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, pages 6721–6729. IEEE.

Yang et al., 2020. Yang, F., Yang, H., Fu, J., Lu, H., and Guo, B. (2020). Learning texture transformer network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5791–5800.

Yeh et al., 2017. Yeh, R. A., Chen, C., Lim, T. Y., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. (2017). Semantic image inpainting with deep generative models. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, pages 5485–5493. IEEE.

Yi et al., 2020a. Yi, K., Guo, Y., Fan, Y., Hamann, J., and Wang, Y. G. (2020a). Cosmovae: Variational autoencoder for cmb image inpainting. arXiv preprint arXiv:2001.11651.

Yi et al., 2020b. Yi, Z., Tang, Q., Azizi, S., Jang, D., and Xu, Z. (2020b). Contextual residual aggregation for ultra high-resolution image inpainting. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7505–7514. IEEE.

Yu et al., 2019. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. (2019). Free-form image inpainting with gated convolution. In International Conference on Computer Vision, pages 4470–4479.

Yu et al., 2018. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5505–5514. IEEE.

Yu et al., 2021. Yu, Y., Zhan, F., Wu, R., Pan, J., Cui, K., Lu, S., Ma, F., Xie, X., and Miao, C. (2021). Diverse image inpainting with bidirectional and autoregressive transformers. arXiv preprint arXiv:2104.12335.

Zeng et al., 2019. Zeng, Y., Fu, J., Chao, H., and Guo, B. (2019). Learning pyramid-context encoder network for high-quality image inpainting. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1486–1494. IEEE.

Zeng et al., 2020. Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., and Lu, H. (2020). High-resolution image inpainting with iterative confidence feedback and guided upsampling. In European Conference on Computer Vision, pages 1–17. Springer.

Zhang et al., 2018a. Zhang, H., Hu, Z., Luo, C., Zuo, W., and Wang, M. (2018a). Semantic image inpainting with progressive generative networks. In 2018 ACM Multimedia Conference on Multimedia Conference, MM ’18, pages 1939–1947. ACM Press.

Zhang et al., 2018b. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018b). The unreasonable effectiveness of deep features as a perceptual metric. In Conference on Computer Vision and Pattern Recognition.

Zhao et al., 2020a. Zhao, J., Han, J., Shao, L., and Snoek, C. G. (2020a). Pixelated semantic colorization. International Journal of Computer Vision, 128(4):818–834.

Zhao et al., 2020b. Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., and Lu, D. (2020b). Uctigan: Diverse image inpainting based on unsupervised cross-space translation. In Conference on Computer Vision and Pattern Recognition, pages 5741–5750.

Zheng et al., 2019. Zheng, C., Cham, T.-J., and Cai, J. (2019). Pluralistic image completion. In Conference on Computer Vision and Pattern Recognition, pages 1438–1447.

Zheng et al., 2021. Zheng, C., Cham, T.-J., and Cai, J. (2021). Tfill: Image completion via a transformer-based architecture. arXiv preprint arXiv:2104.00845.

Zhou et al., 2017. Zhou, B., Lapadhriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence, 40(6):1452–1464.

Zhu et al., 2017. Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., and Shechtman, E. (2017). Multimodal image-to-image translation by enforcing bi-cycle consistency. In Advances in neural information processing systems, pages 465–476.
Appendix

Additional Quantitative Results

We provide in this section additional quantitative results on Places2 and ImageNet. Results from Tables 8 and 9 were conducted in the same conditions as Tables 4 and 5 but with 256 × 256 resized images instead of center-cropped images. Note that, in average, on both Places2 and ImageNet, the difference between methods is very similar when computed on resized or cropped images. The modification in aspect ratio due to the resize operation does not impede the results, even for models that were trained on “real” aspect ratios. The main reason for this is that the aspect ratio is not drastically changed when resizing Places2 and ImageNet images. Another explanation is that the training datasets are large enough and the models have enough capacity for being robust to such a transformation.

Table 8 Quantitative comparison of three pluralistic image inpainting methods (PIC, DSI-VQVAE, ICT) on 256 × 256 resized images from Places2

| Mask | Method   | Similarity to GT | Realism | Diversity |
|------|----------|------------------|---------|-----------|
|      |          | PSNR      SSIM↑  L1↓  FID↑  LPIPS↑     |
| Irregular | PIC     | 29.86 0.934  2.14 0.0489 32.3 0.0055 |
| < 20% | DSI-VQVAE | 30.64 0.948 2.30 0.0533 20.0 0.0214 |
|       | ICT     | 29.05 0.939  3.83 0.0450 23.0 0.0224 |
| Irregular | PIC     | 22.98 0.808  7.06 0.0394 91.0 0.0375 |
| [20%, 40%] | DSI-VQVAE | 23.04 0.832 6.92 0.0443 64.4 0.0789 |
|       | ICT     | 22.11 0.818  8.81 0.0423 72.8 0.0831 |
| Irregular | PIC     | 19.01 0.649  14.71 0.0273 144.2 0.1357 |
| [40%, 60%] | DSI-VQVAE | 19.15 0.684  13.90 0.0287 115.0 0.1700 |
|       | ICT     | 18.50 0.669  15.78 0.0330 127.4 0.1755 |
| Central | PIC     | 19.50 0.797  10.27 0.0335 104.5 0.1129 |
| 128 × 128 | DSI-VQVAE | 19.46 0.797 10.60 0.0387 94.6 0.1364 |
|       | ICT     | 19.42 0.796  11.72 0.0352 101.0 0.1284 |
| Random | PIC     | 20.80 0.773  10.95 0.0359 93.8 0.1152 |
| regular | DSI-VQVAE | 21.15 0.791 10.48 0.0426 79.0 0.1233 |
|       | ICT     | 21.03 0.787  11.51 0.0382 84.3 0.1239 |
| Random | PIC     | 19.91 0.640  13.85 0.0246 157.7 0.1023 |
| irregular | DSI-VQVAE | 20.05 0.682 12.98 0.0329 116.5 0.1539 |
|       | ICT     | 19.10 0.662  15.41 0.0285 131.4 0.1607 |
| Average | DSI-VQVAE | 22.25 0.789 9.53 0.0401 81.6 0.1140 |
|       | ICT     | 21.54 0.779  11.18 0.0370 90.0 0.1157 |
Table 9 Quantitative comparison of three pluralistic image inpainting methods (PIC, DSI-VQVAE, ICT) on 256 × 256 resized images from ImageNet.

| Mask             | Method     | Similarity to GT | Realism | Diversity |
|------------------|------------|------------------|---------|-----------|
|                  |            | PSNR↑ SSIM↑ L↑ ‖ | MIS↑ FID↑ | LPIPS↑    |
| Irregular < 20%  | PIC        | 31.37 0.944 1.82 | 0.1885 21.5 | 0.0028    |
|                  | DSI-VQVAE  | 31.83 0.952 2.08 | 0.1913 12.8 | 0.0175    |
|                  | ICT        | 30.21 0.946 3.41 | 0.2002 12.2 | 0.0203    |
| Irregular [20%, 40%] | PIC        | 23.13 0.807 6.91 | 0.1401 93.7 | 0.0323    |
|                  | DSI-VQVAE  | 23.45 0.825 6.72 | 0.1617 61.8 | 0.0790    |
|                  | ICT        | 22.36 0.817 8.34 | 0.1739 52.1 | 0.0810    |
| Irregular [40%, 60%] | PIC        | 18.39 0.636 15.84 | 0.0497 198.0 | 0.1314    |
|                  | DSI-VQVAE  | 18.95 0.672 14.14 | 0.0732 147.9 | 0.1901    |
|                  | ICT        | 18.34 0.663 15.85 | 0.0822 120.4 | 0.1764    |
| Central 128 × 128 | PIC        | 19.31 0.795 10.35 | 0.0583 153.9 | 0.1091    |
|                  | DSI-VQVAE  | 19.47 0.800 10.25 | 0.0700 177.1 | 0.1293    |
|                  | ICT        | 19.91 0.796 11.27 | 0.0790 120.3 | 0.1247    |
| Random regular   | PIC        | 19.63 0.745 13.13 | 0.0690 150.5 | 0.1071    |
|                  | DSI-VQVAE  | 20.13 0.769 11.59 | 0.1048 113.8 | 0.1457    |
|                  | ICT        | 20.13 0.766 12.56 | 0.1028 101.7 | 0.1376    |
| Random irregular | PIC        | 19.70 0.618 14.04 | 0.0457 194.6 | 0.1021    |
|                  | DSI-VQVAE  | 20.11 0.665 12.85 | 0.0642 155.9 | 0.1648    |
|                  | ICT        | 18.94 0.649 15.33 | 0.0859 131.2 | 0.1652    |
| Average          | PIC        | 21.92 0.758 10.35 | 0.0919 135.4 | 0.0949    |
|                  | DSI-VQVAE  | 22.32 0.781 9.61 | 0.1110 107.7 | 0.1211    |
|                  | ICT        | 21.64 0.773 11.13 | 0.1207 89.7 | 0.1175    |

Additional qualitative results

In Figures [3] and [4] we show additional inpainting visual results on Celeba-HQ and ImageNet datasets.
Fig. 13 Diverse inpainting output on 256×256 images from Celeba dataset with center, and irregular masks. For each method, out of 25 generated samples, the 5 samples with highest discriminator score are displayed.
Fig. 14 Diverse inpainting output on 256× 256 images from ImageNet dataset with center and irregular masks with different hidden proportions. For each method, out of 25 generated samples, the 5 samples with highest discriminator score are displayed.