Computer Vision and Image Understanding 203 (2021) 103147

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A comprehensive review of past and present image inpainting methods

Jireh Jam [a], Connah Kendrick [a], Kevin Walker [b], Vincent Drouard [b], Jison Gee-Sern Hsu [c], Moi Hoon Yap [a,*]

[a] Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK
[b] Image Metrics Ltd, Manchester, UK
[c] National Taiwan University of Science and Technology, Taiwan

ARTICLE INFO

Communicated by Nikos Paragios

Keywords:
Image inpainting
Restoration
Texture synthesis
Convolutional neural network
Generative adversarial networks

ABSTRACT

Images can be described as visual representations or likeness of something (person or object) which can be reproduced or captured, e.g. a hand drawing, photographic material. However, for images on photographic material, images can have defects at the point of captured, become damaged, or degrade over time. Historically, these were restored by hand to maintain image quality using a process known as inpainting. The advent of the digital age has seen the rapid shift image storage technologies, from hard-copies to digitalised units in a less burdensome manner with the application of digital tools. This paper presents a comprehensive review of image inpainting methods over the past decade and the commonly used performance metrics and datasets. To increase the clarity of our review, we use a hierarchical representation for the past state-of-the-art traditional methods and the present state-of-the-art deep learning methods. For traditional methods, we divide the techniques into five sub-categories, i.e. Exemplar-based texture synthesis, Exemplar-based structure synthesis, Diffusion-based methods, Sparse representation methods and Hybrid methods. Then we review the deep learning methods, i.e. Convolutional Neural Networks and Generative Adversarial Networks. We detail the strengths and weaknesses of each to provide new insights in the field. To address the challenges raised from our findings, we outline some potential future works.

1. Introduction

Image inpainting originated from an ancient technique performed by artists to restore damaged paintings or photographs with small defects such as scratches, cracks, dust and spots to maintain its quality to as close to the original as possible. Fig. 1 shows inpainting performed by hand.

The evolution of computers in the 20th century, its frequent daily use and the development of digital tools with image manipulation capability, has encouraged users to appreciate image editing, e.g. restoration, and the application of on-screen visual display and special effects to images. As a result image inpainting (henceforth inpainting) has become a state-of-the-art restoration technique. In a computer vision and graphics context, inpainting is a method that interpolates neighbouring pixels to reconstruct damaged, or defective, portions of an image without any noticeable change on the restored regions when visually compared with the rest of the image. These damaged portions/areas of an image are a set of unconnected pixels surrounded by a set of known adjacent pixels. During the reconstruction of disconnected pixels, the inpainting method uses known-information to fill unknown regions (disconnected pixels).

In this regard, Efros and Bertalmio are considered the pioneers (Paragios et al., 2006; Bertozzi et al., 2006; Wang et al., 2010) in this field and for advancing the research in texture synthesis and pixel interpolation respectively. In 1999, Efros and Leung (1999) proposed an advanced computational interpolation of pixels using Markov modelling. This novel concept is based on self-similarity to estimate a pixel value at the centre of a patch to synthesis a texture. This approach using image patches for texture synthesis, has largely influenced the success in developing image processing algorithms (Buades et al., 2005; Tauber et al., 2007; Buyssens et al., 2015). The method is a non-parametric approach to image synthesis, using an exemplar image as a source, and where pixel values are selected one pixel at a time. In this process, the chosen pixel merges and blends-in with the neighbourhood of the already synthesised output image.

In 2000, Bertalmio et al. (2000) pioneered the introduction of a novel geometry-attentive approach for the interpolation of pixels on images. This novel method is based on Partial Differential Equations (PDE) and diffusion as a technique to propagate local features from surrounding regions into the damaged areas. PDE use isophotes (level lines with the same intensity on the surrounding area), e.g. Fig. 2 shows the use of PDE for inpainting. The white mask regions denote the

* Corresponding author.
E-mail address: m.yap@mmu.ac.uk (M.H. Yap).

https://doi.org/10.1016/j.cviu.2020.103147
Received 29 May 2020; Received in revised form 11 November 2020; Accepted 17 November 2020
Available online 18 November 2020

Fig. 1. Hand inpainting performed by an artist. Image courtesy of Thottam (2015).

Fig. 2. Inpainting on CelebA-Dataset image using the traditional method of inpainting by Bertalmio et al. (2000). Techniques under traditional inpainting methods are limited in terms of mask size, accuracy and sometimes efficiency.

part or region to be filled-in, and the rest of the image is the source of propagated features. However, this technique is limited to small masks (unknown) regions. Then, in 2001, Efros and Freeman (2001) introduced a stitching technique, known as quilting, that synthesised a smaller patch of an image to a more substantial textured outcome of the same texture and structure of the initial image. This technique performs texture transfer on an initial seed (texture) through different stages, as shown in Fig. 3.

These pioneering works (Efros and Leung, 1999; Bertalmio et al., 2000), then considered to be state-of-the-art, caught the attention of the community (Wei and Levoy, 2000; Bornard et al., 2002; Shen et al., 2003; Jia and Tang, 2003; Drori et al., 2003; Criminisi et al., 2004; Shih and Chang, 2005; Cho et al., 2008; Zhuang et al., 2009; Kwok and Wang, 2009; Xu and Sun, 2010; Goyal et al., 2010; Kwok et al., 2010; Cao et al., 2011; Darabi et al., 2012; He and Sun, 2014; Batool and Chellappa, 2014; Akl et al., 2018; Abbad et al., 2018; Elharrouss et al., 2019b,a; Sridevi and Kumar, 2019; Liu et al., 2019a; Jin et al., 2018; Mo and Zhou, 2019) to further the research of these, "traditional" inpainting methods. Although these methods are reviewed in other literature, e.g. by Guillemot and Le Meur (2014) and Qureshi et al. (2017), their scope is limited to traditional methods only. Yet, despite the advancements of these methods in the last decade, inpainting continues to remain a very challenging problem in computer vision. The purpose of this review is to bridge the gap in the previous literature (Guillemot and Le Meur, 2014; Qureshi et al., 2017) and to include

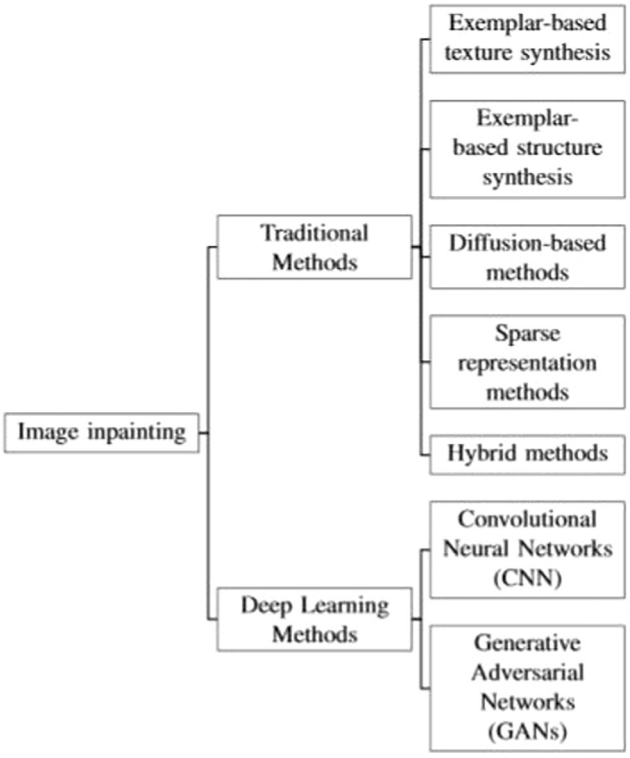

Fig. 4. Hierarchical representation of image inpainting techniques in two main categories: Traditional Methods (the past) and Deep Learning Methods (the present). There are five sub-categories for Traditional Methods: Exemplar-based texture synthesis, Exemplar-based structure synthesis Diffusion-based methods, Sparse representation methods and Hybrid methods. Deep Learning Methods are sub-categorised into CNNs and GANs.

traditional and deep learning methods for the state-of-the-art algorithms. It should be noted that the few techniques reviewed here are just a handful of selected methods from inpainting techniques already in use. These methods are selected and summarised under different categories to illustrate the transition from traditional to the now state-of-the-art deep learning methods. Included as a summary under each category are the challenges and limitations of these techniques. We also consider the importance of datasets and include the frequently used performance metrics, as well as the performance evaluations useful for inpainting methods. Fig. 4 shows a hierarchical representation of the various categories of inpainting in their respective groups. We organise our review as: Section 2 presents traditional inpainting methods under different categories; Section 3 reviews deep learning methods; Section 4 summarises the popular datasets for inpainting; Section 5 describes the common performance metrics for inpainting; Section 6 discusses our findings; and in Section 7, we draw our conclusion and suggestions for future work.

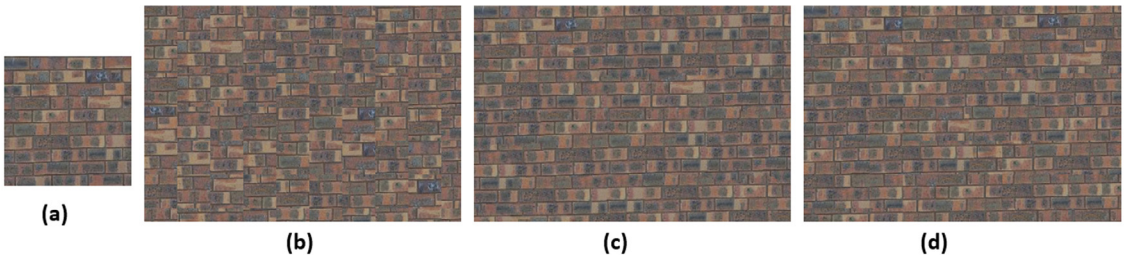

Fig. 3. Image synthesis by Efros and Leung (1999) (a) Original texture (b) Synthesised texture from random patches. (c) Results of pixel-difference computed by the Sum of Square Differences (SSD) (d) A fully synthesised texture output with seam carving restoring similar visual appearance. Image courtesy of Kuppig (2015).

2. Traditional image inpainting techniques

Since the evolution of digital technology, computer vision has experienced enormous research in transformations on images such as image-stitching (Szeliski et al., 1997), morphing (Gomes et al., 1999), image swapping (Chen et al., 2014), registration (Maes et al., 1997), denoising (Buades et al., 2005) and inpainting (Efros and Leung, 1999). Image inpainting experienced enormous amounts of research with considerable attention in the last few years, as researchers try to develop algorithms that are robust with less computational complexity. Various optimisation techniques are proposed to enhance the capability of these algorithms to handle more complex image structures. Because images are a visual representation in texture and structure, the image properties (patterns, corners, edges and changes in brightness) affect the performance of an inpainting algorithm. To understand the concept of inpainting in its fullness, we define texture and structure in terms of image composition.

A **texture** is a visual pattern on an infinite 2-D plane with a stationary distribution at some scale (Efros and Leung, 1999). This pattern refers to the feel (smooth, rough) of the image surface. Textures are either regular (repeated texels) or stochastic (imprecise texels) and can be synthesised based on the assumption that the sample is large and uniform with known statistics of regular patterns (Raad and Galerne, 2017). A geometric texture of an image is the entire representation as a texture based on statistical details of which a small patch is sufficiently a representative (Lai et al., 2005). In textural inpainting, the available data considered for the inpainting task are exemplar textures. Textural inpainting uses statistical knowledge of patterns due to its stationary distribution of missing regions and known parts of the image, commonly modelled by Markov Random Fields (MRF) (Efros and Leung, 1999).

The **structure** of an image is a visual object constructed by distinct parts (global contour information) of the image texture (Bertalmio et al., 2003). The geometric structure of an image is a representation of composition and structure. During inpainting, the geometric structure has a low dimensionality representation in subspace. That is, the coordinates of the inpainted region are exact representations of the subspace and do not exceed its dimension. This is because it must satisfy the coordinate vertices of the image representation before decomposition to yield an approximate representation of the parent structure. With this technique, the target region does not exceed the parent structure, and the outcome is a good representation of the global context. In structural inpainting, taking account the nature of the smoothness in the missing regions and the boundary conditions is a precondition and which uses either isotropic diffusion or anisotropic diffusion to propagate boundary data in the isotropic direction (Bertalmio et al., 2000). The main categories of traditional methods of inpainting put forward in this review are as follows:

- Exemplar-based texture synthesis methods
- Exemplar-based structure synthesis methods
- Diffusion-based methods
- Sparse representation methods
- Hybrid methods

2.1. Exemplar-based texture synthesis

Exemplar-based texture synthesis methods are based on distance measuring tools and aim to generate new visually similar texture images from an input source whilst not being an exact copy of the input.

As already mentioned, Efros and Leung (1999) proposed pioneering method laid the groundwork for exemplar-based texture synthesis. This method uses MRF modelling to locate pixel distribution and a new texture is formed in the unknown target region by querying existing texture to find blocks of similar pixels. This modelling process captures all neighbouring pixels to grow a new texture by synthesising the initial seed one pixel at a time. The process is iterative and uses a patch with known pixel values from a known patch of the previous step. The limitations are discontinuities, unwanted growings that do not respect statistics of a texture, thus causing the texture not to have uniformity.

Efros and Freeman (2001) also used samples from a textured seed to form a similar patch with different dimension via quilting. This technique densely samples square patches from the initial seed, to assemble a single row of pixels in an order that forms the final image. To achieve this, the next patch to quilt into the image comes from a set of candidate patches. This method uses SSD to compute scores between the patches obtained from the patch overlap region to the left and above the patch in the quilted image. The limitation of this method is the enforcement of random patch selection, which may misalign with the rest of the texture and eventually form cascades on subsequent patch alignments. Also, due to disruptive coarse textures, smoothness is not often achieved because the patch size does not always complement the texture coarseness.

Le Meur and Guillemot (2012) used non-parametric patch sampling (Efros and Leung, 1999) to synthesise a coarse version as an input low resolution image for inpainting. The proposed solution is to use K-Nearest Neighbour (KNN), K-coherence candidate SSD and the Battacharya distance (Bugeau et al., 2010) metrics for priority selection of matches at different scales across a multi-resolution of selected patterns from the input low-resolution image. This allows the inpainting process to be less sensitive to noise and work with more enhanced oriented structures in the image. The authors use KNN to perform inpainting at coarse level, and apply single-image super-resolution to recover high-frequency details of the missing area from the inpainted low-resolution image. Therefore this technique reduces noise sensitivity and computational complexity allowing the algorithm to focus on the extraction of dominant orientations of textural image structures. This method can handle inpainting by filling in missing areas using different parameter settings to influence the patch size to better handle textures. A limitation is the speed during inpainting and quality of the resultant image, which highly depends on the selection approach for high-resolution patches from the dictionary based on the user (parameter setting).

In summary, exemplar-based texture synthesis methods are well-known to produce similar texture as the resultant image. These methods also preserve local textures and are capable of synthesising discontinuous textures. Exemplar-based texture synthesis exploit image statistics and assign a priority pixel based on surrounding similarity to inpaint a region on an image. Exemplar-based texture synthesis can produce a texture with perceptual similarity to the input sample. Methods in this category can synthesis small textures and rearrange neighbouring pixels in a consistent way, however, they can grow meaningless textures and verbatim copies. Another limitation of these methods is that the synthesised texture can have an unnatural feel due to limited samples in the data region caused by limited pattern similarity. Also noted is its inability to yield high quality results for highly structured examples.

2.2. Exemplar-based structure synthesis

Methods use statistical constraints to sample texture patch-wise instead of pixel-by-pixel, thus resulting in faster synthesis and realistic results with regular patterns. Under this category, image inpainting is structure consistent with the resultant image generated randomly with statistical constraints. Lou et al. (2018) described exemplar-based structural synthesis as a search in the source image for a cluster of similar pixel patches to fill up missing pixels.

Criminisi et al. (2004) proposed a filling order approach that is isophote-oriented. This technique is dependent on priority filling order of pixel values on structure continuation which favours the action of filling joints in the direction of incoming structures. The process starts by assigning pixels at the edge of the target region as priority pixels. Texture synthesis is performed during stage two of the process by replicating information from a source region in blocks based on the priority

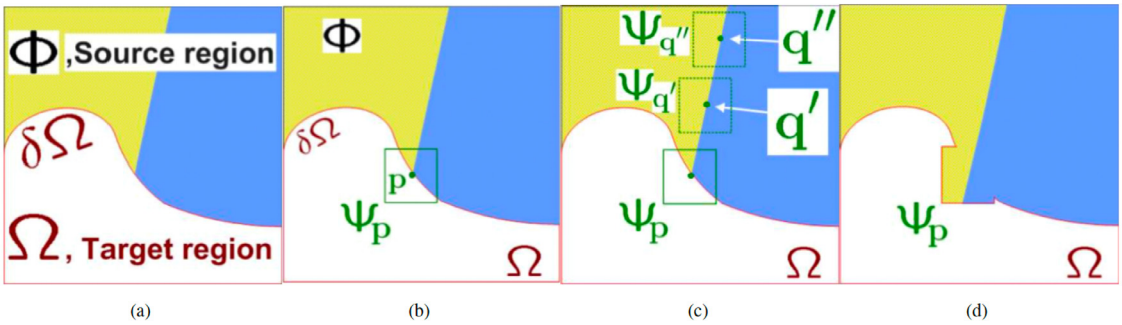

Fig. 5. Image inpainting process in Exemplar-based inpainting task adopted from Criminisi et al. (2004). (**a**) The original image with contours showing source and target regions. (**b**) shows the chosen filled patch based on high pixel priority. (**c**) shows the most likely candidates for filling the patch. (**d**) shows best matching patch selected from candidate patch and copied to its occupied position.

value that was initially assigned to each central pixel. Fig. 5 illustrates the propagation of linear structures during the inpainting process from (a) to (d). Experimental observations shows adaptations to changes in structure due to its isophote preserving properties while propagating best match. The algorithm can handle both texture and structure during inpainting. However, it cannot handle curved structures, and its high dependence on priority pixel value may cause accidental priority pixel dropping, yielding visual inconsistencies on the inpainted region.

Barnes et al. (2009) motivated by Ashikhmin (2001), Sun et al. (2005) and Simakov et al. (2008), used Nearest Neighbour Fields (NNF) and proposed a fast randomised algorithm (PatchMatch). This method finds patches via random sampling with the help of prior information of random fields using NNF defined on a possible patch location. The process is iterative, and uses pixel propagation based on natural coherence where the randomly selected candidate uses adjacent pixels to improve a nearest neighbour search for new candidates. The advantage of this method offers is the fine-scale control to output pixels with the desired colour and restores both structure and texture simultaneously. It avoids dense computation in finding patch similarity because it applies arbitrary distance metrics which enabled local interactions as a constraint for completion. Furthermore, it performs well on images with texture and structure and large missing regions but limited in performance as it incurs additional memory overhead to store current best distance.

Ružić and Pižurica (2014) proposed to use textural descriptors to model and facilitate the search for candidate patches. This method splits the image according to context, thereby restricting the search for candidate patches to matching context. With the use of MRF as a prior to encode knowledge about consistency of neighbouring patches, the selection of candidate patches is accelerated based on contextual features with improved performance. This technique adaptively selects patches with more than half of the missing pixels in top-down procedure based on homogeneity in context. Thus, applying this technique, is an advantage because fixed patch sizes can be used even when missing pixels are not dominant. However, despite the improved performance it still faces challenges when shifting patches into unknown regions on images with complex scenes due to failures in translations associated with MRF-based methods (Yang et al., 2016).

Wang et al. (2017) used space varying update strategy powered by Fast Fourier transform for full image search. The base technique uses a standard deviation-based patch matching criterion and confidence term, that evaluates the spatial distribution of patches to measure the amount of reliable information surrounding the priority point against a known-priority point. This technique reduces the fast dropping effect seen in Criminisi et al. (2004) and takes the distribution of patch differences into account. By eliminating fast priority estimation, a full search is achieved for better and agile matching results leading to improved visual consistency. The fill-in approach propagates linear structures surrounding the damaged area into the hole. These linear structures impose image constraints that influence the performance of exemplar-based texture synthesis efficiently and qualitatively. Also, this algorithm imposes a practical matching criterion of the region and priority function with high confidence in pixel and structure information. This method performs well in inpainting, but experiences discontinuities based on the coefficient value applied to adjust the weight on the standard deviation during the inpainting task.

Liu et al. (2019a) proposed to use multi-resolution for priority patch selection on high resolution images to complete an inpainting task. This method (Liu et al., 2019a) uses similar patch selection to compute multiple candidate patches based on colour, gradients and boundaries. In this technique, the authors assumed that high-resolution images are susceptible to high-frequency information (complex textures and noise) when extracting information around edges. The technique uses Structure Similarity Index Measure (SSIM) to select reasonable candidate patches and graph cut technique when filling the target region with the selected patch. To select a suitable candidate patch, the SSIM value of the known region is calculated to aid in selecting the best candidate patch. Using graph cut technology introduces smoothing, thus eliminating blockiness associated on the inpainted regions. The use of colour, gradients and boundary terms blends patches well with improved inpainting effect. However, to obtain best results, the algorithm relies on the sample patch size to be manually selected, which may vary from user to user, thus may yield poor results.

In summary, methods in this exemplar-based structure synthesis category use similar patches from a known neighbourhood to recover the texture and structure of a missing region. This is based on learned pixel similarity by sampling texture sample from known parts of the image. The source patch is consistent with the geometric structure to fill in targeted regions during an inpainting task (Buyssens et al., 2015). Furthermore, methods in this category overwrite missing pixels with corresponding pixels from patch to shrink the hole and update priorities. Also, limitations associated with speed, accuracy of texture (meaningless growing), and accurate propagation of linear structures are handled. Some limitations are lack of reasonable results when attempting to synthesis regions with no existence of similar pixels. Other limitations noted are failures in curved structures and depth ambiguity.

2.3. Diffusion-based methods

Methods in this category propagate image content smoothly from boundary regions into the interior of the missing region.

Bertalmio et al. (2000) proposed to use isophotes and diffusion process (Laplacian) to propagate pixels automatically. This enables a simultaneous fill of missing regions in any direction with no limitation of the region to be inpainted. To achieve this, a smoothness estimator is introduced during computation and the propagated information is along the isophotes direction. Also, a time-varying estimator along the isophote direction determines the spatial change based on a discretisation gradient vector. This algorithm is efficient on images with small

cracks due to anisotropic diffusion, but leads to a blurring effect with slightly bigger mask regions. It is limited in the reconstruction of large textured regions or images with multiple damaged areas.

Richard and Chang (2001) introduced a fast approach that uses a diffusion kernel (gaussian) with considerations based on the tolerance of blur areas by human vision on regions with high contrast edges. The algorithm is mainly user specified as it involves repeated convolutions with a gaussian. This process involves computing a weighted average of neighbouring pixels when convolving an image with the gaussian kernel, which is equivalent to isotropic diffusion. That is, using a linear heat equation (isotropic diffusion) as diffusion barriers (two pixel-wide line segment) to handle edge re-connection. However, the results introduce some blur without user-entered diffusion barriers due to the low pass linear filtering suppressing high frequencies. Also, if the mask used is not exact on the region to be inpainted, false information propagates into the inpainted area. This algorithm only works well on filling locally small missing areas.

Tschumperlé (2006) proposed a trace-based PDE as a regularisation technique on multi-valued (multiple colour channels) images. This method is tensor-driven PDE based on heat flow constraints on integral curves to preserve curvature on images. By using this technique, isophote diffusion is minimised in all directions, yielding a low pass filter that suppresses high frequencies on the image. This process allows smoothing with edge preservation on curved edges. The capability of PDE to control geometrical features usually supports variational principles (Faugeras and Keriven, 2002; Chan and Shen, 2005). Hence the use of the gaussian kernel on tensors to define orientation and strength of diffusion will fail to preserve curved structures. This method shows high performance on images with narrow damaged regions and small occlusions. The poor performance of this method is observed when filling large areas as it often results to image blur if the target region is large.

Daribo and Pesquet-Popescu (2010) motivated by Criminisi et al. (2004), proposed to use a depth aided texture (depth map) that considers background pixels as high priority over foreground pixels on an image during an inpainting task. This method (Daribo and Pesquet-Popescu, 2010) uses the depth map for image-based rendering to fill holes caused by disocclusion. Also, the distance measure of the depth minimises patch search with the same depth level. The depth map is smooth and disoccluded regions are easily verified. A smoothing strategy on the depth map correct deformities in disoccluded regions by correcting edges. This solves the problem of disocclusion occurrences, edge inconsistencies and overly smoothness during the process. However, this smoothing process is adaptive according to surrounding scene structures. Local smoothing of depth maps and edge correction are simultaneous during the inpainting process. Due to texture-less nature of depth maps, the assumption of a "virtual" image plane is where the depth map projects to perform the hole filling. This inpainting algorithm performs well on video inpainting tasks. However, this method lacks spatial and temporal stability and sometimes lead to more errors on the foreground depth map. This algorithm requires an iterative implementation of numerical methods, which has slow rendering speed, thus making it less robust on still images.

Le Meur et al. (2011) used exemplar-based with PDE technique to compute patch priority on structure tensors to fill in missing regions. The main objective to employ PDE is to propagate information in the direction of isophote lines to continue geometric and photometric information as described by Bertalmio et al. (2000). Li et al. (2017a) used PDE combined with smoothness constraints as regularisation technique to propagate local information. These constraints force the algorithm to follow directions given by local structure and are regularised iteratively, thus resulting in a sequence of continuous smooth image. Hence information showing a local pixel on an image contour propagates smoothness along contour direction and not across boundaries, thus addressing the limitations of previous methods (Bertalmio et al., 2000, 2001, 2003). Pixels located on uniform surfaces will spread smoothness in all direction. Based on this finding, using isophotes (line of constant intensity) alongside PDEs leads to an inpainted image with a continuous evolution in structure. It has shown great success on textured images with scratches/smaller gaps but poor results on large gap regions. The disadvantage of these methods is computational cost due to slow or prolonged arrival of structures at border regions.

Sridevi and Kumar (2019) proposed to use fractional-order derivative (integer-order derivative) with Discrete Fourier Transform (DFT) for inpainting task. The authors (Sridevi and Kumar, 2019) used this method to achieve a good trade-off between the restored region and edge preservation, and also because DFT are easy to implement. Using fractional order derivative, pixel level on the whole image is considered instead of just considering neighbouring pixel values. The technique employs fractional-order nonlinear diffusion model (difference curvature driven Chen et al., 2010) to handle gap regions and fractional-order variational model for denoising and de-blurring of the image. The advantage of this method is its ability to preserve edges during an image restoration task. Also, it is effective in eliminating noise and blur without affecting the edges. The disadvantage of this model is that it relies on user interaction for manual selection of fractional order, which may lead to poor results on the inpainted region.

In summary, methods in this category push good pixels from boundary into the gap region, filling in altered pixels to set a colour similar to or the same as the source region. They are suitable for inpainting with scratches, straight lines curves and edges. However they turn to blur large textured regions due to prolonged arrival of pixels to fill in gap regions. Therefore it is not well suited for textured images with large gaps/regions. Also, the iterative implementation of numerical methods that will eventually render it slow. Therefore this algorithm is not robust on still images.

2.4. Sparse representation methods

The sparse-based methods assume that images contain natural signals that admit a sparse decomposition over a redundant dictionary leading to efficient algorithms that can handle sources of such data (Mairal et al., 2008).

Inspired by Shih et al. (2003) and Chang and Shih (2008) proposed to use colour space in facial images to correct facial images which are overly-exposed in digital photography. This method uses multi-resolution (Shih et al., 2003), to consider level details in a suitable colour space for layer separation and fusion layer during an inpainting task. This process exploits the characteristics and level-by-level features of the image and segments the skin region on facial image. The multi-resolution technique uses mean colour (average of a group of pixel colours) and neighbouring pixels to perform an inpainting task based on the percentage of damaged pixels. A sliding window, based technique is utilised to select bright spots (reflection artefacts) on the face. Based on this method, there is evidence that a considerable pixel variation contains detailed shapes of an image which supports the claim by Shih et al. (2003), for a multi-resolution inpainting strategy. However, this algorithm is highly effective on facial images with high accuracy, but cannot generalise well on other images. It is used to correct facial images with too much light exposure in digital photography. Although, this algorithm is highly effective and accurate on facial images, it has not been tested on natural scene images.

Kawai et al. (2008) proposed an approach that examines two sample textures by considering the change in brightness and spatial locality of texture patterns. Energy minimisation is the key technique to this method because initial values are assigned to missing regions and targeted regions completed by minimising the energy function. This technique also checks the data region for image patterns, and uses a sliding window to match fixed pixels on that region. The preceding step uses a central pixel of the window overlapping the expanded area and the missing region fills-in with the reference pixel inwardly during

inpainting. The energy function is the weighted SSD representing the pattern similarity in conjunction with the change in brightness and the similarity difference representing the spatial locality. This method allows a change in intensity with spatial locality as a constraint during inpainting, but shows poor performance when single weight coefficient is used.

Shen et al. (2009) proposed to use the sparse representation of image signals over a redundant dictionary with the assumption that the image is thinly distributed on the basis of wavelets. This method relies on discrete cosine transform to build a redundant dictionary of patch observations. The inpainting task performs sequential computation iteratively over these sparse representations, completing every uncompleted patch at the boundary of the target region. In this method, the user is allowed to specify the area to be inpainted, which eliminates the problem of finding the corresponding input signal to the corrupted area of the known pixel specified by the user. The image is inpainted inwardly from the boundary of the targeted region with the priority pixel given the most probable chance. At each iteration, the pixel closest to the target filling region, has the maximum priority since the patch is the current iteration centred at the boundary of the target region. Overall, the algorithm recovers incomplete image signals with each signal corresponding to a patch, and the target region is filled based on the patches for each sparse patch representation.

He et al. (2018) used a dual-phase algorithm (Thiele's rational interpolation function and Newton–Theile's function) for adaptive inpainting. This method uses continued fractions to update pixel intensity during the reconstruction of damaged portions based on the surrounding pixel information of known regions along the target region. That is, if the damaged pixel points are vertical, the selected points for interpolation of pixels are in the horizontal direction. The masked image is scanned line by line to locate and adopt information of known pixel points to perform interpolation of pixel intensity. The first phase repairs the damage and the second phase refines the restored image to closely resemble the original image. This second phase updates the intensity of all the damaged pixels in the reconstructed image by using the masked image to locate damaged positions originally corresponding to the previous damaged repaired regions. Whilst this method performs well, its limitation is that the damaged pixels need to be in the vertical direction.

In summary methods in this category assume that the known and unknown regions share similar sparse representation. Also, another assumption is the images are represented in sparse linear combination as a complete dictionary which can be adaptively updated to target inferred pixels. Methods in this category help to improve the visual quality of the image.

2.5. Hybrid methods

The continued success of exemplar-based texture synthesis and exemplar-based structure synthesis methods in inpainting tasks, has motivated researchers to explore the combined capabilities of these methods. Where some are suitable for small gaps, or structures with curvatures and edges, others work best with texture restoration. For this reason, combining two methods to handle images with composite structures and texture has become an area of great interest in inpainting.

Bertalmio et al. (2003) used PDE to determine synthesis ordering (Bertalmio et al., 2000) and texture synthesis by Efros and Leung (1999) to recover geometrical structures and small textured regions. This method decomposes an image into texture and structure layers for inpainting. After decomposition, the energy function for texture synthesis is applied to the texture layer and diffusion-based method for inpainting applied to the structure layer. This method breaks down the inpainting into two. It uses the diffusion based technique in Bertalmio et al. (2000) in the structure layer and adds synthesised textures derived by using Efros and Leung (1999) method to the in-filled region. The energy function utilised in patch stitching minimises the seam area during the inpainting task. The energy function measures self-similarity, coherence and diffusion with each measure having a role during the corresponding mapping of pixels. The similarity between the patch of central filling pixel and the known pixel of the source region of the image is computed by self-similarity measure. A discrete Laplacian equation calculates the corresponding map of the image region to be inpainted for diffusion. This method overcomes the limitation by diffusion-based methods which causes overly smooth outcomes. The method is computationally expensive and fails in some cases of large missing structures.

Allène and Paragios (2006) used a combination of variational (Efros and Freeman, 2001) and statistical (Bertalmio et al., 2000) methods to propose a graph-based approach based on the concept of progressive stitching. This technique uses MRF-based cost function to find similarity of the existing image and patches that are needed as well as the measure of the boundary to the inpainted region. The concept of progressive stitching consists of a selection of pixel values, corresponding to selected patches from possible image patches that best approximates the original data. The technique of Allène and Paragios (2006) uses MRF-based cost function to find similarity of the existing image and patches that are needed as well as the measure of the boundary to the inpainted region. To achieve this, they introduced constraints to minimise the distance between chosen patches and existing data to capture properties of the area surrounding the missing content. With these constraints, multiple candidate patches are then superimposed over the target region, forging a multi-source label solvable by graphs to perform inpainting. However, because discontinuities are created during the stitching of selected patches, local smoothness on the final texture is not always achieved. This method also incurs a computational cost due to the particular attention required for selecting the right patches, to best approximate the corresponding pixels.

Zhang and Dai (2012) proposed an algorithm that decomposes an image into structure and texture using wavelet transform with the aim being to capture the image texture and structures without loss of information with wavelet transform. In structural propagation, the patches copied are specified by missing structures of the unknown regions and are in the same direction of similar curves in the known area. The use of curvature driven diffusion also applies to structural reconstruction, while exemplar-based texture synthesis is the fill-in process during textural reconstruction. The two inpainted components are combined into a plausible outcome with similarity compared to the input image. However, this method is computationally demanding for large fill regions.

Ghorai et al. (2018) proposed to use patch selection and refinement method based on joint filtering alongside a modified MRF to enhance optimal patch assignment to perform an inpainting task. This technique uses subspace clustering to select target patches from boundary regions into groups, which are refined via joint patch filtering to capture patterns and remove artefacts. The selected patches are targeted in sequential order from the interior regions based on neighbouring patches from candidate patches along the boundary regions. The subgroup of similar patches are merged into larger groups, alongside ensuring deselection of any patch that is too different from the boundary patch. However, despite a faster patch selection in reduced search space compared to global patch selection by other methods, the limitation is in the cost of group formation and grouping of each target patches. Quantitative evaluations using PSNR showed the algorithm's best at 29.7 db for regular-shaped mask and 24.23 db for irregular-shaped mask.

In summary, methods in this category can remove text overlays, fill in regions with complex textures and structures. It is noted to handle discontinuity in boundary regions and blur, and produce images with local coherency in visual quality. Despite the excellent results, methods in this category still perform poorly in some disocclusion and object removal task. Also, there is no guarantee in convergence and they cannot be applied to error concealment applications due to time constraints (computationally expensive).

Table 1
Analysis of the state-of-the-art traditional methods on image inpainting.

Category	Advantage(s)	Disadvantage(s)	Prior(s)	Application(s)
Exemplar-based texture synthesis (Efros and Leung, 1999; Efros and Freeman, 2001; Simakov et al., 2008; Le Meur and Guillemot, 2012)	Preserves artefacts, No occurrence of blur.	Failures in the reconstruction of large textured regions or images with multiple damaged areas.	Smoothness.	Image restoration, editing, disocclusion. & concealment
	Preserves both structural and textural information, Used in wireless transmission for lost block re-transmissions.	Can lead to repetitive patterns.		
Exemplar-based structure synthesis (Bertalmio et al., 2000; Barnes et al., 2009; Ružić and Pižurica, 2014; Liu et al., 2019a)	Performs better during inpainting of large textured regions. Inpainted task is copy/paste fashion with almost verbatim copies.	It is time consuming and exhaustive for methods in this category to generate different candidate patches. Can mix several verbatim copies with distinctive overlaps.	Priority assignment. Best patch selection.	Image restoration, editing, disocclusion. & concealment.
	Restores texture, structure and colour.	Incurs additional memory overhead to store the current best distance.		
Diffusion-based methods (Bertalmio et al., 2000, 2001; Shen and Chan, 2002; Shen et al., 2003; Liu et al., 2013; Guillemot and Le Meur, 2014; Li et al., 2017a)	It generates good results when filling in small or gap regions. Preserves edge information. Suitable for completing lines and curves. Does not produce verbatim copies in the synthesised textured region. Maintains the structure of the inpainted region.	Fails to inpaint large textured regions, resulting in blurry artefacts on image.	Smoothness	Image restoration
Sparse representation Methods (Shih et al., 2003; Mairal et al., 2008; Chang and Shih, 2008; Kawai et al., 2008; Shen et al., 2009; He et al., 2018)	Inpaints facial images with high exposure to light. Allows change in light intensity.	May not work well on natural scene images. On image reconstruction, the damages pixels must be in the vertical direction.	Colour space, self-similarity & sparsity.	Image restoration
Hybrid inpainting (Zhang and Dai, 2012; Guillemot and Le Meur, 2014; Wang et al., 2017)	Preserves edge and restores smoothness. Impressive results on the linear structure of the image improve the speed.	Computational complexity with no guarantee in convergence.	Smoothness, Similarity & Sparsity.	Image restoration, editing, disocclusion. & concealment

2.6. Summary

Some of the inpainting methods described above are summarised in Table 1. These have shown great success in nearest neighbour searching, i.e. using patches or pixels to synthesise images. However, the challenge to inpainting is maintaining a realistic structure and texture in the output image. For example, traditional methods attempt to fill in missing pixels using image patches from existing regions or use the diffusion mechanism to propagate pixels into a hole region from high pixel similarity areas. Whilst, these methods can propagate vivid textures for background-inpainting, they often failed to capture high-level semantics, yielding non-realistic images with repetitive patterns. Moreover, these methods do not yield plausible outcomes for inpainting tasks on complex mask regions such as a face or objects with non-repetitive structures. Generally, and despite their success in generating high-frequency seamless textures, they continue to fail in the generation of structures that are globally consistent. However, the advent of the generative neural network, inpainting algorithms can be taught to learn meaningful and high-level semantics which have proven to generate coherent structures for missing regions. This is discussed in the next section.

3. Deep learning methods

In more recent research, the use of Convolutional Neural Network (CNN) (Gatys et al., 2015b; Li and Wand, 2016; Ulyanov et al., 2016) and Generative Adversarial Network (GAN) (Pathak et al., 2016; Yang et al., 2017; Yeh et al., 2017) have become the state-of-the-art methods used to perform image inpainting task. These methods use CNN as a feature extraction method through the process of convolution to capture abstractions. The use of CNN combined with adversarial training (Goodfellow et al., 2014) has produced excellent results on inpainting task, with perceptual similarity to the original image. It is advantageous to use CNN combined with GAN because a CNN has an encoder that is used as a feature extractor to capture high dimensional data abstractions; and a decoder that reconstructs these learnable features in an end-to-end fashion, while the GAN enhances the sharpness of the image (Pathak et al., 2016).

The performance of algorithms in this category depends on the datasets used. Datasets are valuable in research as it consists of ground-truth images and information (Roh and Lee, 2007). Different types of datasets have been used to train and test image inpainting algorithms. These datasets are made up of images with diverse textural and structural information, which mainly test the robustness of the algorithms to learn different image features. Table 2 shows a summary of the deep learning methods, models, description, the loss function and datasets used for evaluation. Before diving into the datasets used for the evaluation, we discuss the popular deep learning methods for image inpainting.

3.1. Convolutional neural networks

Jain and Seung (2009) pioneered the use of CNN to inpainting, by framing the computational task within a statistical framework of regression rather than density estimation. The authors used an image

Table 2
Summary of reviewed literature on deep learning methods in image inpainting.

Method	Model	Description	Loss function	Dataset
Jain and Seung (2009)	CNN	Auto encoder. Formulated for image denoising and extended to image inpainting.	Reconstruction loss.	In-house 100 greyscale images
Xie et al. (2012)	CNN	Sparse coding and deep neural networks as a denoising auto-encoder and extended to image inpainting.	Reconstruction loss.	In-house images
Pathak et al. (2016)	GAN	Encoder–decoder architecture as the generator and Discriminator network.	ℓ_2 reconstruction, Adversarial loss.	Paris Street View, ImageNet, PASCAL VOC2007
Iizuka et al. (2017)	GAN	Encoder–decoder generator with a refinement network based on dilated convolutions combined with Global and local discriminators.	weighted ℓ_2 Adversarial loss	Places2, ImageNet, CMP FAcade.
Yang et al. (2017)	GAN	Uses style transfer network and context-encoder to ensure Multi-scale neural patch synthesis to preserve contextual structure with local patch similarity on images with high-frequency details.	Adversarial loss, ℓ_2-based texture loss computed with features from VGG19.	Paris Street View and ImageNet.
Yeh et al. (2017)	GAN	Searches latent space encodings assisted by spatial attention mechanism to reconstruct the original image.	Weighted context ℓ_1 based loss and Adversarial loss.	CelebA, SVHN, Standford Cars.
Yu et al. (2018)	GAN	A two stage model network (encoder–decoder) that uses cosine similarity assisted by contextual attention layers, redesigned to use dilated convolutions	Reconstruction loss and two Wasserstein GAN losses.	CelebA, CelebA-HQ, DTD, ImageNEt and Places2.
Liu et al. (2018a)	GAN	A U-NET architecture, that uses partial convolutions instead of normal convolutions.	Perceptual loss, style loss, adversarial loss.	CelebA, CelebA-HQ, Places2 and ImageNet.
Yan et al. (2018)	GAN	Uses a U-NET architecture to introduction Shift connection layer to transfer fine texture details.	Guidance loss, ℓ_1 and adversarial loss.	Paris Street View and Places2.
Wang et al. (2018)	GAN	A Laplacian pyramid GAN, reconstruction and residual learning in generator.	VGG-Feature loss, Adversarial loss.	CelebA and Paris Street View.
Huang et al. (2019)	GAN	Introduced padding and pooling operations in an Encoder–decoder, to avoid edge disappearance. The model also uses a mini-batch discriminator for realistic photo completion.	ℓ_2-based SSIM loss and Adversarial loss.	In-house dataset containing 2015 images.
Zeng et al. (2019)	GAN	A U-NET architecture that uses a cross-layer attention and pyramid filling mechanism.	ℓ_2 and Adversarial loss.	Facade, DTD, CelebA-HQ and Places2.
Li et al. (2019)	GAN	U-NET, visual reconstruction layers, residual block, partial convolutions , patch discriminator.	Perceptual loss, style loss, adversarial loss.	Places2, Paris Street View and CelebA.
Yu et al. (2019)	GAN	Gated convolutions, contextual attention layer and SN-PatchGAN	Pixel-wise reconstruction loss (ℓ_1) and adversarial loss.	Places2
Liu et al. (2019b)	GAN	A two-stage network that uses Coherent semantic attention layer that preserves the spatial structure of the image within the refinement network.	Consistency loss (ℓ_2), feature loss (VGG) and adversarial loss.	CelebA, Places2 and Paris Street View.
Jam et al. (2020a)	GAN	An encoder–decoder network that uses a reverse masking mechanism to enforce prediction only on missing pixel regions with preserved realism on the unmasked regions, to improve the quality of the inpainted image.	Reversed-mask loss (ℓ_2-base), feature loss (VGG) and adversarial loss (WGAN).	CelebA-HQ, Places2 and Paris Street View.
Zhao et al. (2020)	GAN	A trio-network network that uses joint probability distribution combined with a cross-semantic attention layer	conditional constraint loss (ℓ_2-base), feature loss (VGG) using ℓ_1-base and adversarial loss (WGAN).	CelebA-HQ, Places2 and Paris Street View.
Zhou et al. (2020)	GAN	A U-Net architecture that uses a dual spatial attention layer	ℓ_1, feature loss (VGG) and adversarial loss (PatchGAN).	Flickr-Faces

denoising task which was formulated with parameter learning for back propagation. The image denoising thus becomes the learning problem in the CNN with noise integrated clean images during training. However, this method is restricted to greyscale (one colour channel) images, removal of "salt and pepper" noise, whilst also requiring substantial computation cost. The method of Jain and Seung (2009) was improved by Xie et al. (2012), who proposed combining sparse coding and deep neural networks as a denoising auto-encoder, to handle inpainting of

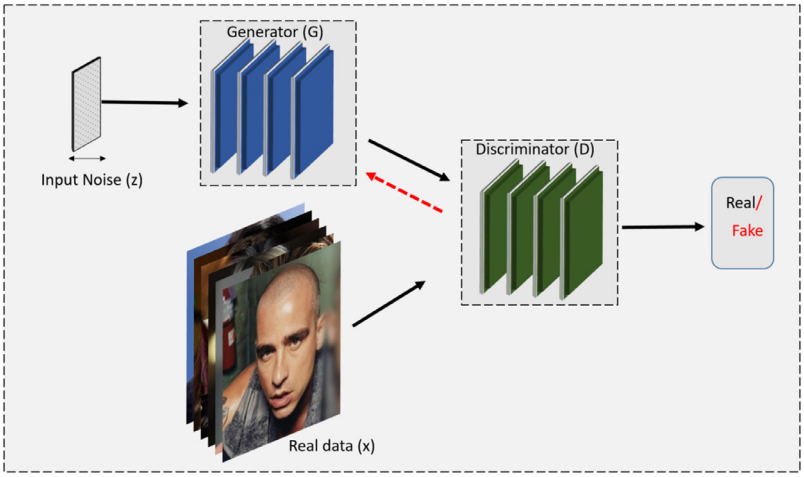

Fig. 6. An example GAN block. The generator input (z) is sampled from a random noise vector and the Discriminator input (x) is sampled from real data distribution.

inconsistent localities of corrupted pixels. The use of sparse coding and deep neural network overcame the limitation of computational cost, and eliminated noisy pixels supplied to the algorithm as the regions required for inpainting. However, the use of stacked sparse denoising auto-encoders strongly relying on supervised training and can only handle images with small denoising tasks, such as the reconstruction of images with controlled procedural pixel corruption.

3.2. Generative adversarial network

GANs refers to a two model framework of unsupervised learning algorithms that estimate adversarial processing. Inpainting methods using the GAN process aim at generating a conditional image for high-level recognition, based on low pixel synthesis formulated into a convolutional network (encoder–decoder). The trained adversarial network enhances coherency between generated and original pixels. For example, Fig. 6 shows the GAN framework for estimating generative models via adversarial networks. GANs were first proposed by Goodfellow et al. (2014), as a two model network; the generative and discriminative model. The generative model (generator) of the neural network captures data as a random input noise and transforms into a fake image, intending it to look like the real image from the training set. The discriminative model (discriminator) tries to distinguish this generated image from the training set, by estimating the probability that it came from the training set rather than from the generative model. Eq. (1) shows optimisation of the loss during the combined training of the discriminator (D) and generator (G) network, where **z** is sampled from a prior distribution p_z and **x** is the sample from p_{data} distribution. G maps the random vector **z** and D discriminates between images generated from G and real images **x** sampled from p_z (Amos, 2016).

$$\min_G \max_D V(D,G) = E_{x\sim P_{data}(x)}[log(D(x))] + E_{z\sim P_z(z)}[log(1-D(G(z)))] \quad (1)$$

However, the difficulty during the training process increases uncertainty, and the generator can improve, leading to a vanishing gradient of the discriminator, thus making it difficult to converge. Fig. 7 shows an inpainting by deep learning method.

Pathak et al. (2016), pioneered adversarial training (Goodfellow et al., 2014), an end-to-end network based on CNN (Hinton and Salakhutdinov, 2006; Bengio et al., 2009; Krizhevsky et al., 2012), to predict the missing content of an arbitrary image region conditioned on its surroundings with realistic output. The authors (Pathak et al., 2016) considered an encoder–decoder backbone (Fig. 8) as part of their design, and implemented a channel-wise fully connected

Fig. 7. Inpainting task on CelebA-HQ Dataset (Karras et al., 2017) shows the performance of deep learning methods. The slightly thicker mask obtained from Nvidia Mask Dataset (Liu et al., 2018a).

layer. This network captures semantic visual structures aided by ℓ_2 reconstruction and adversarial loss. The ℓ_2 loss is capture the overall structure of the missing region with regards to context and coherency, but tends to average the multiple modes in prediction. Introducing adversarial loss (Goodfellow et al., 2014) enables the network to predict more realistically by picking particular patterns from the distribution. Overall, performance evaluations using PSNR show that adversarial training with reconstruction loss, yields higher PSNR value of 18.58 with Doersch et al. (2012) compared to 14.70 & 12.79 obtained by Hays and Efros (2007) model, suggesting a more accurate pixel inference of missing content to the entire image. It is, however, limited to small image sizes of low resolution due to training with regards to ℓ_2 loss. It also lacks spatial support with more substantial inputs, and often produces images with considerable amounts of implausible results that are overly-smooth (blurry) and which lack edge preservation. Furthermore, the discriminator focuses on the missing region and does not take into account the global context of the image. Thus, this method cannot guarantee structural cohesion nor a harmonious texture between the inpainted region and the image context.

Iizuka et al. (2017), motivated by Pathak et al. (2016), proposed the use of dilated convolutions (Yu and Koltun, 2015), within a network as the encoder–decoder backbone (Fig. 8), combined with two discriminators. Dilated convolutions increase the input area of each layer without loss of resolution or parameter accretion. The use of dilated convolutions increases the receptive fields for neurons at the output, thus replacing the channel-wise fully connected layer in Pathak et al. (2016). The global discriminator for assessing the entire coherency of the reconstructed image and a local discriminator assesses the area of

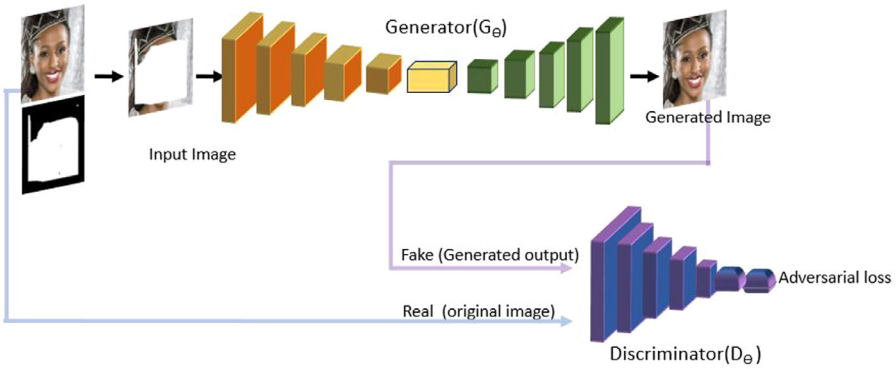

Fig. 8. An overview of encoder–decoder architecture; a backbone used by the state-of-the-art inpainting networks. For example, an encoder–decoder framework used by Jam et al. (2020a) with ongoing adversarial loss.

the completed region to ensure consistency within the entire image. The input to the global discriminator is a 256 × 256 resolution image, while the input to the local discriminator is 128 × 128 on the centre of the completed region. The authors used subjective evaluation reporting 77% approval rating on naturalness of inpainted images against the state-of-the-art (Barnes et al., 2009; Pathak et al., 2016) and 96.5% for ground-truth images. However, this method fails to capture long-ranged textural information (limited to image completion with mask at border region) and relies heavily on post processing using fast matching (Telea, 2004) and Poisson blending (Pérez et al., 2003).

Yang et al. (2017) used style transfer (Johnson et al., 2016; Ulyanov et al., 2016; Li and Wand, 2016) to propose a multi-scale neural patch synthesis approach combined with adversarial loss. Yang et al. (2017) used the context-encoder (Pathak et al., 2016) to capture image content, texture and to preserve contextual structures, thus producing images with high-frequency detail. The style transfer network ensures the context-encoder predicts the global content with the local patch similarity of the predicted region. A texture network, pre-trained on image classification, takes the output image of the prediction network as an input. The architecture uses a local texture loss function computed VGG19 (Simonyan and Zisserman, 2014), pre-trained on ImageNet and a "holistic" loss based on ℓ_2. The joint loss function is ℓ_2-norm and the texture term (loss function) computed on extracted feature maps from feature layers (relu3 and relu4) of the VGG19 block. Quantitative results based on Paris Street View (Doersch et al., 2012) shown in Table 4, show that inpainted images performed better compared to the state-of-the-art (Pathak et al., 2016; Goldman et al., 2014). However, the proposed algorithm suffers some limitations with the content and texture networks failing to guarantee correct image structure, blurry images whilst being computationally expensive with high-resolution images (taking longer to inpaint). Another limitation is the difficulty in hallucinating suitable texture for larger irregular mask regions since it is designed for rectangular holes.

Yeh et al. (2017) used the model architecture from Radford et al. (2015) to train a generative model that searches for encodings of the corrupted image in latent space to recover the lost area based on the surrounding image features as reference; thus, with this encoding, the generator reconstructs the original image. This algorithm uses context loss (ℓ_1-norm) and adversarial loss information to search for closest encoding on the trained generative model regardless of the structure (mask) of the missing content. The context loss is a "weighted context" considered close to the corrupted pixel region while the prior loss corrects unrealistic images. Iterative optimisation of the objective function is through back-propagation in combination with prior and context losses. The algorithm scored high PSNR values (22.8, 33.0 & 18.9) compared to PSNR (20.6, 24.1 & 16.1) using Pathak et al. (2016) across the various datasets for this method as shown in Table 2. However, despite excellent performance, the algorithm struggles with misalignment on images and some failures in finding "closest" encoding to the corrupted image in latent space. This may be as a result of the difficulty in training GANs, which can lead to poor data distribution capture increasing its inability to handle high resolution or complex scene images.

Yu et al. (2018) introduced a coarse-to-fine network and used the contextual attention module by Iizuka et al. (2017) to design their network. The authors (Yu et al., 2018) considered an encoder–decoder backbone to design the coarse and refinement networks as a two-stage (coarse-to-fine) network. This model uses cosine similarity to learn the relationship between background and foreground feature patches. The contextual attention module is redesigned to use dilated convolutions and the model optimised using a reconstruction loss and two Wasserstein GAN losses by Arjovsky et al. (2017) and Gulrajani et al. (2017). The two-stage model produces a roughly restored intermediate image with filled predictions and refines this result using a refinement network designed with dilated convolutions. The contextual attention module includes spatial propagation layers to encourage spatial coherency and fuse attention scores for more realistic outcomes. The contextual layers refine the image and alienate the idea of Poisson blending in Iizuka et al. (2017). The final reconstruction performing convolutions on foreground patches and background patches relies on the attention score for each pixel value, obtained using Softmax. These are then propagated channel-wise to reconstruct layer. The quantitative performance are reported on rectangular mask only and are shown in Table 4. However, despite great results, it also lacks fine textural details and inconsistencies with the background pixel-wise on high resolution images.

Liu et al. (2018a) used the U-Net (Ronneberger et al., 2015b; Isola et al., 2017) and replaced convolutions with partial convolutional in inpainting task. The U-Net backbone considered by Liu et al. (2018a), implemented with partial convolutional operations that have automatic mask update steps. The re-normalised masked-convolutions operations focus on valid pixels, followed by an automatic mask generation to the next layer as a forward pass. The loss functions used to handle pixel-reconstruction accuracy of the hole region are, ℓ_1 loss, perceptual loss (Gatys et al., 2015a) and style-loss. The mask updating step is non-learnable and with a fixed convolutional layer with a kernel size that matches that of the partial convolutional operation with weight initialised to 1 and no bias layer. The partial convolution layer, with automatic mask-update mechanism, undergoes a sufficient number of continuous updates to remove any masking on the unmask value in return for accurate feature maps. The comparative evaluation against the state-of-the-art Barnes et al. (2009), Iizuka et al. (2017), Yu et al. (2018), using a binary mask of hole-to-image area ratio of [0.5,0.6]. The performance evaluation of this method (Liu et al., 2018b) by the authors compared to the state-of-the-art showed high PSNR and SSIM values for all mask sizes. This method suffers from the reliance on initial hole values, that causes the algorithm to produce images that

lack plausible output texture. Also, it struggles with sparsely structured images and binary masks with larger hole-to-image-ratio. This is because neurons with receptive fields cover valid or invalid pixels at different spatial locations. The invalid pixels disappear following the rule-based mask layer by layer leading to some missing information in deeper layers that may be needed to synthesis pixels in mask regions.

In a different study, Yan et al. (2018) used the U-Net (Ronneberger et al., 2015b) (as shown in Fig. 9) to introduce a shift-connection layer combined with guidance loss to inpaint images using deep feature rearrangement. The shift-connection layer introduced in a U-Net backbone to handles images with sharp structures and fine texture details. The technique concatenates the encoder feature of the first convolutional layer to serve as an estimator of the missing parts on the last decoder layer after the fully connected layer. This approach uses a guidance loss function, ℓ_1 loss and adversarial loss to obtain photo-realistic textures. The guidance loss implemented based on the shift-connection layer, uses SSD on concatenated features of first convolutional layer of encoder and features of last convolutional layer of decoder. The end recovery used the encoded features to approximate the missing portion based on the ground-truth. Overall, the performance evaluation scored high values quantitatively and are shown in Table 4. However, this method may experience poor performance due to the parameter value of the guidance loss chosen to perform the shift operation. A limitation is that a smaller parameter value may lead to a more extensive feature map size that will increase computational time to 400 ms per image. Also, a more substantial parameter value may lead to a more modest feature map, leading to a loss in image detail information. Thus, the best trade-off reported a computational time of 80 ms per image, compared to 40 ms per image which results to a generated image with less texture and coarse details. Although their shift-connection implementation of the U-Net structure has shown excellent results, it struggles in terms of efficiency and computational speed due to network parameters that do not make it suitable for most applications (see Fig. 9).

Wang et al. (2019) introduced a Laplacian-pyramid based GAN to inpainting. The authors (Wang et al., 2019) used a backbone as Fig. 8 with additional implementation. They used a modified ResNet block by He et al. (2016), with the aim to propagate high-frequency details from the surrounding to predict precise missing information while eliminating colour discrepancy. The modified ResNet block, implemented with dilated convolutions, implies a larger receptive field with batch normalisation layers and rectified linear unit for speed convergence. The Wang et al. (2019) introduced a combined representation learning with reconstruction and residual learning in the generator network to extract predicted missing regions and therefore combine features of low middle level of fine layers. The generator model captures the image content and compresses it to a latent representation. These are feature extraction progressively predict missing regions while the residual learning phase learns the difference between visible colour similarities of predicted pixels and surrounding pixels. The loss function uses trained-VGG model to extract features and uses feature space, combined with pixel-wise and adversarial loss to learn photo-realistic images, therefore maintaining the natural artefacts of the original image while completing the missing region. The masks used are rectangular and of size 128 × 128, and are randomly positioned on the input image. The Wang et al. (2019) method achieved the best performance scores compared to Pathak et al. (2016), Iizuka et al. (2017), Li et al. (2017b), Yu et al. (2018), Liu et al. (2015a) and Yeh et al. (2017). This model performed particularly well with regularly shaped binary mask, but was not experimented with irregularly shaped mask. Also, the model output had colour discrepancies, hence the model cannot be generalised to natural scene inpainting.

Huang et al. (2019), motivated by Goodfellow et al. (2014), Mirza and Osindero (2014) and Ronneberger et al. (2015a), studied the network structure of the encoder–decoder and introduced padding and pooling operations to avoid edge disappearance. The proposed completion network uses adversarial training with a new loss function based on SSIM and ℓ_2 loss. SSIM loss works as an authentication mechanism on the reconstructed image to improve the structure and texture. This method also introduced the use of a mini-batch discriminator to optimise training, thus increasing diversity of the generated sample. This loss enables photo-realistic images which are further judged by the adversarial network to obtain the output as close as possible to the original image. The in-house dataset contains 2015 images for authentic street images by Huang et al. (2019). The images are 256 × 256 with a variety of situations such as foggy, rainy, day and night. The mask is rectangular, with various sizes randomly generated and applied to the image. The randomly generated binary masks are applied to the images and randomly shuffled with 80% used for training and 20% for testing. The data distribution is similar for both training, and test sets for the filling task handled in various situations. Both qualitative and quantitative evaluations carried out on the in-house dataset shows this algorithm performs with good results, scoring ℓ_2 (8.99), PSNR (39.63) and SSIM (0.97) values, compared to Pathak et al. (2016) (11.02, 37.36 & 0.95).

Zeng et al. (2019) used the U-Net (Ronneberger et al., 2015b), a backbone as shown in Fig. 9, to design a network that learns high-level semantic features from region affinity to fill in missing regions in a pyramid fashion. The authors introduced a cross-layer attention and pyramid filling mechanisms in each layer, referred to as Attention Transfer Network (ATN) with each layer being derived from region affinity between patches. The ATN transfers relevant features outside missing regions and makes use of softmax and cosine similarity between patches inside/outside missing regions extracted from patches. With softmax applied, the attention scores obtained are used as the valid pixel to fill in the missing region. A multi-scale decoder using dilated convolutions at different rates acts as a refiner during the filling-in of missing regions. The loss function used is ℓ_1 loss combined with GAN loss (Goodfellow et al., 2014) for realistic images. Masks sizes of 32 × 32, 64 × 64 and 128 × 128 are used for the evaluation of this method. Overall, and based on the qualitative evaluations by the authors, high quality images are obtained with smaller non-border size masks. Also included in the evaluation analysis are random mask used for visual comparison . The overall performance using non-border mask sizes of 128 × 128 show good results compared with the state of the art (Barnes et al., 2009; Iizuka et al., 2017; Liu et al., 2018a; Yang et al., 2017) as shown in Table 4. Also, the authors do not show detailed results for images with border mask regions.

Li et al. (2019), proposed the use of visual features and structures to restore or inpaint missing parts of an image. This model uses the U-Net (Ronneberger et al., 2015b), implemented with a visual reconstruction layer combined with partial convolutions (Liu et al., 2018a), and a bottleneck residual block. The encoder uses upper bound additional visual reconstruction layers to estimate the edges of the missing structure before passing these to partial convolution layers. Within the decoder are lower-bound additional visual reconstruction and convolution feature reconstruction layers. The visual reconstruction layers entangle the reconstruction of visual structures and features of an image . The masks regions are progressively filled-in with meaningful content based on the reconstructed edges and the input image. The use of Patch-GAN discriminator (Isola et al., 2017) with slight adjustments to include spectral normalisation controls the generalisation error. The network is end-to-end with detailed generation of restored missing structure assisted by adversarial loss combined with loss functions from Liu et al. (2018a). It should be noted that parameter fine-tuning is required before training the network. Across all hole-to-image ratios used during the evaluation, the network had a slight edge in performance compared with the state-of-the-art (Nazeri et al., 2019; Liu et al., 2018a). We report results for 10%–30% hole-to-image ratio on Table 4. However, it is time consuming for the visual reconstruction layers to learn structural parts, thus increasing the time to filter out unwanted structures not needed for image reconstruction.

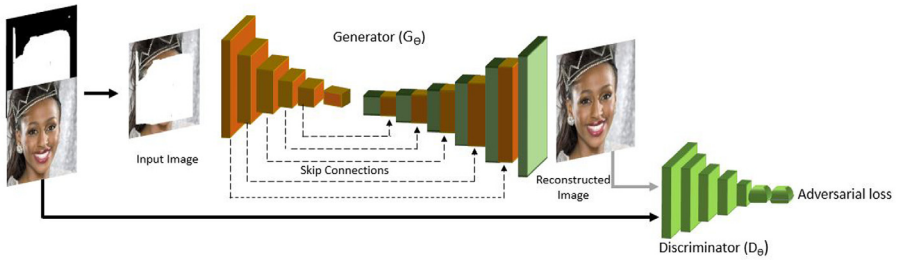

Fig. 9. An overview of U-Net combined with a discriminator without any additional implementation (for example attention layer) as described in some state-of-the-art methods. This U-Net is similar to Yan et al. (2018) without the shift-connection layer.

Yu et al. (2019) used gated convolutions combined with contextual attention layer and Spectral-normalised Markovian Discriminator (SN-PatchGAN) for inpainting task. The backbone of the network is an encoder–decoder stacked with gated convolutions, contextual attention layers and a refinement unit of dilated convolutions. Gated convolutions allows the network to learn soft mask from input data. Within these convolutions are processes that learn features from input data progressively for each channel. At each spatial location the prediction of missing pixels are conditioned on the valid pixels in the input image. During the process, each gating block produces two outputs that go through different activation mechanisms producing gating values and learned features. The contextual attention layer enables the network to capture long-ranged features from distant spatial locations. This is assisted by dilated convolution blocks used as refining mechanisms within the network. The discriminator as part of the combined network, outputs a 3D-shape feature based on three inputs (image, mask and guidance channel). Based on the report by Yu et al. (2019), the performance of this algorithm as shown in Table 4 is better with free-form mask than rectangular mask, though the sizes of the masks are not detailed in their report. However, gated convolutions will perform better with free-form than rectangular masks, which limits the algorithms performance to generalise larger masks or large images with hole-to-image ratio. Additionally, with gated convolutions, correlation between valid features is not guaranteed and may lead to colour discrepancies on the completed image. Furthermore, this network is computationally expensive due to the gating within convolutions and the three stage network which has to be trained end-to-end.

Liu et al. (2019b) proposed the use of semantic relevance between hole and non-hole regions for effective prediction of the hole, aided by a contextual structure preserving mechanism known as Coherent Semantic Attention Layer (CSAL). The design is a two stage network based on encoder–decoder backbone as rough network and the U-Net (Ronneberger et al., 2015b) as the refinement network, combined with two discriminators. The CSAL is an embedding within the refinement block of the two-step U-Net architecture, implemented to elevate the quality of reconstructed images. A proposed consistency loss is used, combined with feature patch and patch discriminator (Isola et al., 2017) to improve stability during training and maintain the natural statistics of the image details. The consistency loss computes the error between corresponding CSAL layers of the encoder–decoder block and VGG features. The feature and patch discriminator combined introduces ongoing adversarial training, based on the relativistic average adversarial loss (Jolicoeur-Martineau, 2018). The patch discriminator evaluates the pixels values on the final output compared to the input image. Introduced within the consistency loss is the reconstruction loss (ℓ_1), as a constraint to assist the model in learning meaningful parameters that can approximate the ground-truth image. Note, these results shown in Table 4 are based on a hole-to-image ratio of 10%–20%. However, despite the high performance of this algorithm, the CSAL may fail due to the nature of the network. That is, if the network is too deep or too shallow, loss of information may occur leading to increased time overhead.

Yi et al. (2020) modified gated convolutions (Yu et al., 2019) into light-weight model and introduced high-frequency residuals to generate rich and detailed textures for high resolution images. To achieve this, the authors used a coarse network based on an encoder–decoder backbone and a refinement network, implemented with a contextual residual aggregation mechanism. Motivated by the size of images captured by mobile phone cameras, the authors proposed a contextual residual aggregation mechanism that borrows contexts, features and residuals. The scores computed are between patches inside/outside the missing region within a specific region that has high affinity of similar patches. Gated convolutions are modified into depth-separable, pixel-wise and single-channel variants. Depth-separable uses depth-wise convolutions followed by 1×1 convolutions as a gating mechanism. Pixel-wise uses 1×1 convolutions as gating mechanism. Single-channel broadcast a single mask to all channels as a hard mask, similar to partial convolutions (Liu et al., 2018a). The network is a two-stage network that uses single-channel for all layers in a coarse-network and depth-separable or pixel-wise in the refinement network. The loss function is a reconstruction loss based on ℓ_1 in the generator and adversarial loss. The qualitative results gives a high-visual quality of the images compared to the state-of-the-art (Iizuka et al., 2017; Yu et al., 2018, 2019; Zeng et al., 2019; Liu et al., 2018a). In quantitative analysis, there is not a significant effect in the measurement compared to the state-of-the-art. From the table of result presented by the authors, we observe four performance evaluation metrics and in addition a time factor as a measure to rate the algorithms performance. Different resolutions on the places2 (Zhou et al., 2017) dataset were compared against the state-of-the-art, with higher resolution images of size 1024×1024 the algorithm performed overall best with a time difference of −696 ms. The results on Table 4 shows the results for 1024×1024 image sizes. The algorithm induced a large effect on high resolution images, proving its ability to generate high-quality contents for missing regions such images. However, with the poor performance in low resolution images, this area of studies still remain a challenge.

Zheng et al. (2019) proposed to use a two stage probabilistic distribution framework, combined with an attention layer (short+long term-based), both using GANs for image inpainting task. The first network uses a Variational Autoencoders-based model to reconstruct an image based on prior distribution of missing parts given the ground-truth. The second network uses the same model (encoder–decoder backbone) with a conditional completion based on information obtained from the first network to predict the missing regions based on the visible pixels. The short and long term layers are used to improve appearance and consistency by measuring the distance between related features of both encoders and decoders of the two networks. However, both frameworks sample from a probabilistic distribution of the masked image with ground-truth visible pixels, and the complement of the masked image with ground-truth missing regions (the reverse of the masked image). To achieve this, a conditional variational autoencoder is employed to estimate the parametric distribution in latent space, where sampling is possible. Therefore, a lower bound conditional log-likelihood is the probability of observed training data given the deep network parameters that generate the missing data.

To optimise both networks, two ℓ_1-Norm based reconstruction loss of which one is geared towards reconstructing the entire image and the other focused on valid (visible) pixels combined with adversarial loss are employed. For evaluations, quantitative comparisons with the state-of-the-art Iizuka et al. (2017) and Yu et al. (2018) showed the model's superiority based on ℓ_1 (12.91), PSNR (20.10), Total Variational (TV) Loss (12.18). For IS (Sohn et al., 2015), the model's performance rate was 24.90 conducted on 20000 test images from ImageNet (Russakovsky et al., 2015b) using 128 × 128 centre binary mask. However, the authors state that evaluations were carried out based on a selection of samples since the goal was not to achieve a single solution.

Li et al. (2020) proposed a recurrent learning approach, where feature maps are inferred in shared recurrent units. The approach uses partial convolutions (Liu et al., 2018a) within a U-Net backbone to identify target regions and use the output as input to an encoder–decoder generator with skip connections and no discriminator. The mask updating mechanism within partial convolutions is exploited during each recurrence as a prerequisite to identify the target regions for subsequent recurrences. Within this network, encoded features undergo a series of recurrence to maximise inference capability to obtain high-quality features during an inpainting task. This means that the hole regions shrink with each recurrence until a high-quality feature is achieved. The mean of the various feature outputs of the network is the final output the decoder. The authors also proposed an attention layer that uses prior knowledge of background pixels to assist the model to obtain best patches at different occurrences that are consistent with the predicted regions and the image. The authors used perceptual and style loss (Johnson et al., 2016) formulated using feature maps from i^{ith} pooling layer extracted from VGG-16 network. Other loss functions used calculate the ℓ_1 of unmasked and masked regions respectively as valid and hole loss. Quantitative evaluations were conducted compared to the state-of-the-art approaches (Zheng et al., 2019; Liu et al., 2018a; Yu et al., 2019; Nazeri et al., 2019; Li et al., 2019) and are shown in Table 4. The limitation of this method is that some boundary artefacts may occur due to inconsistencies with feature maps posing as shadow-like regions during feature merging process.

Zhou et al. (2020) used the U-Net backbone architecture to learn facial textures at multiple scales with help of seven discriminators. The proposed method uses a Dual Spatial Attention (DSA), that learns correlations between facial textures based on two inputs (masked-image and ground-truth image) to obtain attention scores for foreground and background pixels for reconstruction. The attention layer is applied to multiple layers within the decoder, with foreground attention scores from softmax layer, acting as direct supervision to the inpainted regions. Within this layer, the masked regions are the foreground and the unmasked region is the background. The DSA works has foreground–background cross-attention and foreground self-attention units within its module. The first unit uses the mask to segment the input feature into foreground and background features and uses 1 × 1 convolutions to rebuilt the original foreground features based on correlations with background features. The second unit is similar, without the foreground features taken into consideration. The attention maps are learned from the ground-truth to ensure high quality filling of missing regions during training. Four discriminators ensure realistic features of the left-eye, right-eye, nose and mouth. The other discriminators are the global and local discriminators that look ensure consistency on the entire image and local masked region. The authors used facial landmarks to locate the eyes, nose and mouth. These locations are cropped using a mask with fixed size corresponding to the landmarks. This model uses the ℓ_1 and perceptual loss (Johnson et al., 2016) to optimise the generator combined with ongoing adversarial loss based on the PatchGAN discriminator (Isola et al., 2017). Segments of the face (eye, nose, mouth) are cropped and each passed through a discriminator to authenticate its generated features. Qualitative and Quantitative evaluations compared the effectiveness of the model with the state-of-the-art (Barnes et al., 2009; Yu et al., 2018; Zheng et al., 2019; Zeng et al., 2019; Yu et al., 2019). Quantitatively, the ℓ_1, PSNR, SSIM and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) were used and shown in Table 4. The advantage of this network is that it uses ground-truth as a direct supervision to obtain high fidelity features for the masked regions on the input masked-image. The limitation of this model is that if learned attention is insufficient or not accurate, poor quality filling will result in the generated image due to unsuitable features filling in the missing regions.

Zhao et al. (2020) used three network modules: a conditional encoder module, a manifold projection module and a generation module combined with cross semantic attention for image inpainting. The authors (Zhao et al., 2020) used a jointed probability distribution analysis to come up with a hypothesis to solve the inpainting task. That is, given that a set of reconstructed images generated from a set of masked-images is expressed as conditional probability distribution, then a set of masked-images is expressed as marginal probability distribution, then the training data is a joint probability distribution. This means that in an image inpainting task, finding the conditional probability distribution depends on the marginal probability distribution and joint probability distribution. Therefore borrowing information from the ground-truth (training data) by traversing an image completion space is in a sense using marginal and joint probability distribution to obtain conditional probability distribution. The architecture is based on encoder–decoder backbone, but designed to have dual encoders with different inputs, where one branch takes an instance image (ground-truth) and the other a masked-image to perform a one-to-one mapping in the same low-dimensional space in order to reconstruct an image. Within this network, are a set of instance images (ground-truth) corresponding to the masked-image used for guidance during training. The network uses cross-space translation to learn one-to-one mappings between the instance image and the masked-image. Therefore, the two spaces (instance and conditional completion) are associated in one latent space by one-to-one mapping, where the instance images corresponding to the mapped restored images have the same representation in low dimensional space. To optimise the network, a conditional constraint loss handles appearance and perceptual features extracted from VGG16 (Johnson et al., 2016) using the ℓ_1 as base. Both appearance and feature loss use the instance and masked image expressed as a function of the network and the mask. Other losses used are the KL divergence, reconstruction and ongoing adversarial loss. The cross semantic attention layer uses 1 × 1 convolutions to transform feature maps obtained by instance and masked images to evaluate cross attention before adding them to feed the decoder. Comparative evaluations were carried out using the baseline models Pathak et al. (2016), Yu et al. (2018), Liu et al. (2019b), Ren et al. (2019), Song et al. (2018), Yan et al. (2018), Sohn et al. (2015), Zhu et al. (2017) and Zheng et al. (2019). Quantitatively, the performance on 1000 CelebA-HQ images using centre mask of size 128 × 128 were better than the state-of-the-art and are shown in Table 4. The limitation of this network is that there is a possibility of suffering from mode collapse (i.e poor diversity in generated images) during training if trained in an unsupervised manner.

In summary, the use of deep learning methods in inpainting produces plausible results when compared to the original image. However, the limitation thus far by the state-of-the-art is the failure in reporting appropriate analysis of results. So far, the best evaluation and analysis of results is by Liu et al. (2018a), Nazeri et al. (2019), Xie et al. (2019) and Li et al. (2019). These authors provide details of the various hole-to-image ratios used for qualitative and quantitative evaluation. For example, Liu et al. (2018a) report details of non-border mask and mask at border regions across evaluation metrics compared with the state-of-the-art. This provides the reader with a true picture of the performance of the algorithm. However, with the details of various mask sizes and mask regions provided by this author (Liu et al., 2018a), the training dataset is not publicly available. Also, on a subjective evaluation, the authors used a wider audience compared to Li et al.

Fig. 10. Examples of binary masks from Nvidia Mask Dataset (Angelova et al., 2005).

Fig. 11. Mask Dataset (Iskakov, 2018).

Fig. 12. Caltech Dataset (Angelova et al., 2005).

Fig. 13. Places2 Dataset (Zhou et al., 2018).

Fig. 14. Paris street view Dataset (Doersch et al., 2012).

(2019), and Yu et al. (2019) to judge the quality of the images. Additionally, randomised missing data (random mask) is more difficult to learn compared to missing data in a central region of an image. The difficulty for an algorithm to capture semantic information for images with masks at border regions and preserve edges still remain challenging. For example, Liu et al. (2018a) and Yu et al. (2019) introduced algorithms that uses masks to infer missing pixels compared to all other methods. In terms of quantitative evaluations, report by Yeh et al. (2017) point out that quantitative results do not have a true representation for different methods.

4. Datasets

With the wider use of deep learning in present inpainting research, the data and the masks are two essential components to train and evaluate the performance of the methods. The following discuss some popular mask datasets and image datasets in image inpainting.

4.1. Nvidia mask dataset

The Nvidia Mask dataset proposed by Liu et al. (2018a), Fig. 10 has six categories of masks with different hole-to-image ratios. This dataset contains 55,116 training masks and 24,866 testing mask, and of 512 × 512 resolution.

4.2. Quick draw irregular mask dataset

The quick draw mask dataset proposed by Iskakov (2018), Fig. 11 contains 50,000 train and 10,000 test masks. The samples are of size 512 × 512 resolution and used for image inpainting task (Iskakov, 2018; Jam et al., 2020b).

4.3. Caltech faces

Caltech Faces (Angelova et al., 2005), Fig. 12 is a sample from the Caltech dataset, containing 450 face images, from 27 different people, at 896 × 592 resolution in JPEG format under different lighting conditions, expressions and background.

4.4. Places2

Places2 is designed following the principles of Human visual system (HVS) (Zhou et al., 2017, 2018) containing images of diverse scenery used for high-level visual understanding task. Consisting of more than 10 million images and containing more than 400 unique scene categories, it has 5000 to 30,000 training images consistent with real-world occurrences. It is used for learning in-depth scene features using CNN for various scene recognition task, e.g. Fig. 13.

4.5. Paris street view

Doersch et al. (2012) developed the Paris Street View dataset from Google Street View (Gronat et al., 2011; Anguelov et al., 2010) to examine which specific algorithms would work on a computational geographic task, and therefore enable automatic location of geoinformation features for a particular place or city. The images are distinctive and geographically informative, being based on a variety of architectural correspondences and geospatial scales (summarised appearance on one specific scale) of different cities from around the world. Two perspectives of images of 936 × 537 resolution are scraped automatically from a dense sampling of panoramas (Gronat et al., 2011). Approximately 10,000 images per city were downloaded from 12 cities across the world, with a focus on Paris and suburban areas. A sample of the dataset is shown in Fig. 14.

4.6. CelebA

CelebFaces Attributes Dataset (CelebA) is a collection of 202,599 facial images of celebrities (Liu et al., 2018b, 2015b) containing 10,177 identities, five landmark locations and each with 40 binary attribute annotations cropped to size 178 × 218 resolution as of 2015. This dataset

Table 3
Datasets used in Image Inpainting.

Datasets	Total images	Purpose	Resolution
Nvidia Mask (Liu et al., 2018a)	79,982	masks	512 × 512
Quick draw mask (Iskakov, 2018)	60,000	masks	512 × 512
Caltech Faces (Angelova et al., 2005)	450	Various	896 × 592
Places2 (Zhou et al., 2018)	30,000	Urban	Variable
Paris Street View (Doersch et al., 2012)	14,900	Urban	936 × 537
CelebA (Liu et al., 2015a, 2018b)	202,599	Various	178 × 218
CelebA-HQ (Karras et al., 2017)	30,000	Inpainting	1024 × 1024
ImageNet (Russakovsky et al., 2015b)	>8 Million	Classification	Variable
PASCAL VOC (Everingham et al., 2010)	14,974	Classification	Variable

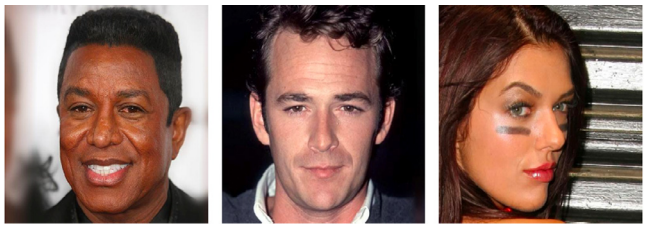

Fig. 15. CelebA-HQ dataset (Karras et al., 2017).

Fig. 16. ImageNet dataset (Krizhevsky et al., 2012).

makes it an appropriate test set for facial image synthesis (Shen et al., 2018) since there are considerable pose variations and background clutter associated with the database alongside a broad diversity and rich annotations.

4.7. CelebA-HQ

The CelebA-HQ dataset, developed by Karras et al. (2017) is developed from the CelebA dataset consisting of 30,000 high-quality images of 1024 × 1024, 512 × 512 and 128 × 128 resolution. The original image resolution in the CelebA dataset varied from 43 × 55 to 6732 × 8984 with various backgrounds and processed by different image quality measures to ensure the image is on the central region (Karras et al., 2017). To obtain the high-quality images for the CelebA-HQ, each JPEG image was processed using two pre-trained neural networks. The authors then used the model proposed by Mao et al. (2016) to remove JPEG image artefacts which was combined with an adversarially-trained 4x super-resolution network for high-resolution images similar to that in Ledig et al. (2017). Padding and filtering were applied to extend the dimension of the images. The authors then used facial landmark annotations included in the original CelebA dataset to orientate and crop the images. The 202,599 subjects in the dataset were processed and analysed, resulting in the best 1024 × 1024 resolution image, and sorted to estimate the best quality images to select 30,000 images. To obtain the rest of the image sizes, we used a resizing tool by Karras et al. (2017), e.g. Fig. 15 shows a sample from the dataset.

4.8. ImageNet

The ImageNet Large scale visual recognition challenge (ILSVRC) has collated millions of images classifying hundreds of different object categories (Russakovsky et al., 2015b,a). It is large ground-truth annotated dataset of images put together for object recognition, detection and classification for the comparison of state-of-the-art algorithms for computer vision accuracy with human accuracy. It contains over 10,000 categories with more than 8 million images of variable resolution (Krizhevsky et al., 2012), e.g. Fig. 16 show samples from three classes.

Fig. 17. PASCAL VOC dataset (Everingham et al., 2010).

4.9. PASCAL Visual Object Classification (PASCAL VOC)

The PASCAL VOC visual object classes consists of two components: a publicly available and an annual competition datasets (PASCAL VOC2005, PASCAL VOC2007, PASCAL VOC20012). Established in 2005, it provides a standardised dataset closest to ILSVRC for object detection, image classification, object segmentation, person layout and action classification (Everingham et al., 2010) for the annual competition. As of 2010, the PASCAL VOC dataset has a total of 19,737 for 20 object categories organised into train, validation and test sets; Fig. 17 shows a sample taken from three different categories of this dataset.

In summary, the challenges to developing applications rely on the mathematical equations, optimisation parameters and dataset used to test the robustness. Table 3 shows a summary of the popular datasets used by researchers for the evaluation of inpainting algorithms.

5. Performance metrics for image inpainting algorithms

Image inpainting algorithms generate images which are distorted or show changes in appearance. To evaluate the performance of these algorithms, different performance metrics are used to quantify the generated images. Methods, based on the highly developed HVS, have mostly used qualitative questionnaire evaluation to extract structural context without need for a large dataset, making this both time consuming and costly. However, some authors use both qualitative and quantitative performance metrics with most commonly used being ℓ_1 (Mean Absolute Error), ℓ_2 (Mean Square Error), Peak signal to Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM). These tools measure the perception of an image to quantify the quality of the error between the distorted pixels of the reconstructed image and the

Table 4
Summary of quantitative results of some deep learning methods for image inpainting on Places2 (Zhou et al., 2017, 2018), CelebA-HQ (Karras et al., 2017) and Paris Street View (Doersch et al., 2012) datasets. The performance evaluation vary from method to method and are approximated to 2 decimal places. The results included are for distortions (image-to-mask ratio) between 10%–20% on image sizes 256 × 256.

Method	Mask type /	Image	Places2						CelebA-HQ						Paris street view					
	Image-to-mask ratio	Size	MAE[a]	MSE[a]	FID[a]	PSNR[b]	SSIM[b]	IS [b]	MAE[a]	MSE[a]	FID[a]	PSNR[b]	SSIM[b]	IS [b]	MAE[a]	MSE[a]	FID[a]	PSNR[b]	SSIM[b]	IS [b]
Pathak et al. (2016)	Square		–	–	–	–	–	–	–	–	–	–	–	–	0.10	0.23	–	17.59	–	–
Yang et al. (2017)	Square (64 × 64)	128 × 128	–	–	–	–	–	–	–	–	–	–	–	–	10.01	2.21	–	18.00	–	–
Yu et al. (2018)	Irregular (10%–20%)	256 × 256	8.6	2.1	–	18.91	–	–	–	–	–	–	–	–						
Liu et al. (2018a)	Irregular (1%–10%)	256 × 256	0.49	–	–	33.75	0.94	0.05	–	–	–	–	–	–	–	–	–	–	–	–
	Irregular (10%–20%)	256 × 256	1.18	–	–	27.71	0.86	0.16	–	–	–	–	–	–	–	–	–	–	–	–
	Irregular (20%–30%)	256 × 256	2.07	–	–	24.54	0.77	0.44	–	–	–	–	–	–	–	–	–	–	–	–
	Irregular (30%–40%)	256 × 256	3.19	–	–	22.01	0.68	0.95	–	–	–	–	–	–	–	–	–	–	–	–
	Irregular (40%–50%)	256 × 256	4.37	–	–	20.34	0.53	1.88	–	–	–	–	–	–	–	–	–	–	–	–
	Irregular (50%–60%)	256 × 256	6.45	–	–	18.21	0.46	3.60	–	–	–	–	–	–	–	–	–	–	–	–
Yan et al. (2018)	Irregular (10%–20%)	256 × 256	–	–	–	–	–	–	–	–	–	–	–	–	–	0.02	–	26.51	0.90	–
Wang et al. (2019)	Square (128 × 128)	256 × 256	–	–	–	–	–	–	–	–	–	23.45	0.86	–	–	–	–	–	–	–
Zeng et al. (2019)	Square (128 × 128)	256 × 256	9.94	–	15.19	–	–	50.51		–	–	–	–	–						
Li et al. (2019)	Irregular (10%–20%)	256 × 256	0.012	–	–	28.87	0.95	–	–	–	–	–	–	–	–	–	–	–	–	–
	Irregular (20%–30%)	256 × 256	0.022	–	–	25.66	0.91	–	–	–	–	–	–	–	–	–	–	–	–	–
	Irregular (30%–40%)	256 × 256	0.033	–	–	23.46	0.86	–	–	–	–	–	–	–	–	–	–	–	–	–
	Irregular (40%–50%)	256 × 256	0.046	–	–	21.74	0.79	–	–	–	–	–	–	–	–	–	–	–	–	–
	Irregular (50%–60%)	256 × 256	0.068	–	–	19.51	0.67	–	–	–	–	–	–	–	–	–	–	–	–	–
Yu et al. (2019)	Irregular (10%–20%)	512 × 512	9.1	1.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Square	512 × 512	8.6	2.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Liu et al. (2019b)	Irregular (10%–20%)	256 × 256	–	–	–	–	–	–	0.72	0.04	–	34.69	0.98	–	–	–	–	–	–	–
	Irregular (20%–30%)	256 × 256	–	–	–	–	–	–	0.94	0.07	–	32.58	0.98	–	–	–	–	–	–	–
	Irregular (30%–40%)	256 × 256	–	–	–	–	–	–	2.18	0.37	–	25.32	0.92	–	–	–	–	–	–	–
	Irregular (40%–50%)	256 × 256	–	–	–	–	–	–	2.85	0.44	–	24.14	0.88	–	–	–	–	–	–	–
	Square (32 × 32)	256 × 256	0.01	–	–	27.75	0.93		1.83	0.27	–	26.54	0.93	–	–	–	–	–	–	–
Li et al. (2020)	Irregular (10%–20%)	256 × 256	0.014	–	–	27.75	0.93	–	0.007	–	–	33.56	0.98	–	0.011	–	–	31.71	0.95	–
	Irregular (30%–40%)	256 × 256	0.038	–	–	22.63	0.81	–	0.02	–	–	27.76	0.93	–	0.027	–	–	26.44	0.86	–
	Irregular (50%–60%)	256 × 256	0.076	–	–	18.92	0.59	–	0.047	–	–	22.88	0.81	–	0.054	–	–	22.40	0.68	–
Zhou et al. (2020)	Square (128 × 128)	256 × 256	–	–	–	–	–	–	1.46	–	–	26.36	0.91	–	–	–	–	–	–	–
Yi et al. (2020)	Square (128 × 128)	512 × 512	5.43	–	4.89	–	–	17.72	–	–	–	–	–	–	–	–	–	–	–	–
	Square (128 × 128)	1024 × 1024	5.43	–	4.89	–	–	17.72	–	–	–	–	–	–	–	–	–	–	–	–
	Square (128 × 128)	2048 × 2048	5.49	–	4.89	–	–	17.85	–	–	–	–	–	–	–	–	–	–	–	–
	Square (128 × 128)	4096 × 4096	5.50	–	4.890	–	–	17.81	–	–	–	–	–	–	–	–	–	–	–	–
Jam et al. (2020a)	Irregular (10%–60%)	256 × 256	0.27	–	4.47	39.66	0.93	–	0.31	–	3.09	40.40	0.94	–	0.33	–	17.64	39.55	0.91	–
Zhao et al. (2020)	Irregular (10%–20%)	256 × 256	–	–	–	–	–	–	1.51	–	–	26.38	0.88	3.01	–	–	–	–	–	–

[a]Lower is better.
[b]Higher is better.

(original) reference image. Other evaluation metrics (Visual information fidelity (Sheikh and Bovik, 2006), universal quality index (Wang and Bovik, 2002), inception score (Salimans et al., 2016), Multi-scale SSIM (Wang et al., 2003), Frechet inception distance (Heusel et al., 2017) and LPIPS (Zhang et al., 2018)) have been reported in literature, however, we focused mainly on the most used ones in this review. The quantitative measure, or score, of the generated image need change only by a few pixels to validate the effectiveness of an algorithm. Given the ground-truth image and inpainted image, ℓ_1 is the total value of the absolute difference between the pixel values of the predicted image and the actual pixel values of the ground-truth image.

$$\ell_1(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{2}$$

In Eq. (2), ℓ_1 gives an overview of the average error for a predicted image. A low computed ℓ_1 indicates that the quality of the image is good (Losson et al., 2010). In Eq. (3), ℓ_2 averages the squared intensity difference between the reference image and the reconstructed image (Haccius and Herfet, 2017).

$$\ell_2(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{3}$$

However, the error in Eq. (3) may not match the perceived visual quality of the image.

PSNR is the ratio of the maximum possible value (power signal) to the power of distorting noise that affects the representation of quality based on two images (reconstructed/original) of the same kind.

$$PSNR = 20log_{10}\frac{(MAX_I)^2}{\sqrt{MSE}} \tag{4}$$

Eq. (4) computes PSNR (dB), well known for assessing the quality of noisy images (Hore and Ziou, 2010), and an approximate value of 48 dB is considered good (Avcibaş et al., 2002). The higher the PSNR value, the better the quality of the reconstructed/generated image. The SSIM (Wang et al., 2004) has become a good correlator for quality perception that discounts aspects of an image not important to the HVS. The SSIM models three factors (loss of correlation, luminance distortion and contrast distortion) of two images based on neighbouring and corresponding pixels. Given the input signals (x,y), SSIM computes the combination of luminance, contrast and structure to output a similarity measure expressed in Eq. (5);

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x{}^2 + \mu_y{}^2 + C_1)(\sigma_x{}^2 + \sigma_y{}^2 + C_2)} \tag{5}$$

where C_1 and C_2 are constants. This method assesses the quality of the image based on the degradation of structural information of the reconstructed image. However, the above quality measures quantify the image whereas subjective assessments depends on the HVS to extract the structural information of the image. This method is, again, both time-consuming and costly.

6. Discussion

In this review, we found that inpainting is remains an important, yet challenging, research area in computer vision. We observed that traditional approaches (Bertalmio et al., 2003; Allène and Paragios, 2006; Zhang and Dai, 2012; He et al., 2018; Ghorai et al., 2018) can handle textural and structural target regions, and are suitable for disocclusion or object removal. We noted that inpainting methods in this category (traditional inpainting methods) use various techniques e.g. Bertalmio et al. (2000), Barnes et al. (2009), Li et al. (2017a), Liu et al. (2013) and have shown exceptional performance in linear structures using diffusion. However, we noted that some methods e.g. Efros and Freeman (2001) and Simakov et al. (2008) in this category perform poorly on mask regions along curves and edges. In contrast, some methods e.g. Bertalmio et al. (2000) and Barnes et al. (2009) handled such regions with plausible outcomes, but are slower and can only perform inpainting on single images. A number of these (Bertalmio et al., 2000; Richard and Chang, 2001; Tschumperlé, 2006) methods produce good results in propagating linear structure using diffusion, but also introduced blur into the target region, making them unsuitable for highly textured images with large missing regions. These methods whilst showing excellent performance have a shortcoming of preserving image realism.

Datasets, e.g. Doersch et al. (2012), Liu et al. (2015a) and Karras et al. (2017), which often contain thousands or millions of images, are crucial to deep learning methods and enable an algorithm to learn features and complete target regions to produce semantically plausible outcomes. The wide diversity in the complexity of structural and textural information of images has a significant impact on the results. To test the robustness of an algorithm, a complex dataset is a requirement since it provides a model with multitude of patterns (sophisticated features) to learn and output plausible satisfactory results. Furthermore, deep learning methods e.g Pathak et al. (2016), Iizuka et al. (2017), Liu et al. (2018a), Yu et al. (2018, 2019) and Huang et al. (2019) have produced plausible outcomes in inpainting when compared to the original image, due to their ability to extract features in an end-to-end fashion. However, we noted that, the weakness encountered is the difficulty in reproducibility across papers. Generally, most authors do not appear to share their code and therefore meaningful comparison is both made difficult and is not progressive. A fair comparison between the deep learning techniques is challenging due to the lack of a dedicated and standardised benchmarking system. Such a system would allow newer models to be compared against an accepted set of models acting as the baseline. We also noted that not all proposed algorithms using similar parameters use the same dataset for training and testing. Also, most algorithms do not disclose parameter fine-tuning, data-preprocessing steps and complexity of the model, and for codes made publicly available, not all are complete. This definitely leads to poor evaluation of different codes if all the information is not available, which may or may not be robust, making it harder to obtain a progressive trajectory in research direction. Additionally, hardware is a hindering/limiting factor in the progress of newly proposed algorithms in this field. However, we note that a good inpainting algorithm should have qualities which, in addition to robustness, also embody high performance with low computational cost.

Computational analysis is a key factor that needs addressing in detail to support the quality of an inpainting algorithm. This will aid the reader to understand the quality of different methods based on some factors such as; running time, training speed, inference time and training time as mention, also highlighted by Elharrouss et al. (2019b). To this effect, it will be advantageous for future works to consider the type of mask used (difficulty in terms of image-to-hole ratio), the dataset, the image sizes and the type of GPU machine used as reported by Liu et al. (2019b) to perform computational analysis. Table 5 shows a summary of methods that have some form of record with regards to computational analysis. This is not in full detail because it does not have all the relevant information that can determine the best algorithm. However, it will be important to provide the full specifications of the machine, measure in terms of the number of epochs, batch size, training data, and provide information on dataset preprocessing. Other methods (Pathak et al., 2016; Yan et al., 2018; Liu et al., 2018a) have reported some training time. Another analysis measure to consider is the time taken to compute the metric score based on a single image resolution. Despite the great success of deep learning methods, there are still drawbacks in terms of computational complexity and failures in preserving image realism.

7. Conclusion and future work

Image inpainting, from traditional to deep learning methods, has achieved immense, and continued, success. We have reviewed a range

Table 5
Summary of the type of GPU and record on experimental details for evaluation computational time by some deep learning image inpainting methods. Note that the record on this table has been extracted from the proposed method and are not based on our evaluation.

Method	Type of GPU & computational analysis	Image resolution	Batch size	Dataset
Pathak et al. (2016)	The training time took 100,000 iterations is 14 h.	256 × 256	–	Paris street view
Iizuka et al. (2017)	The training took 500,000 iterations on a single machine with four K80 GPU took two months. Also,further evaluations using GeForce TITAN X GPU reports a drop in computational time to 0.141 s per 512 × 512 image. 0.141 s per image	512 × 512	96	ImageNet
Yu et al. (2018)	Initial reported time was 11,520 GPU hrs on K80. Improved training time to 120GPU hrs using GTX 1080TI. The full model runs at 0.2 s per frame	512 × 512	–	Places2
Yang et al. (2017)	The time taken to fill in 256 × 256 hole on an image size of 512 × 512 using a TITAN X GPU which is slower compared to Pathak et al. Pathak et al. (2016). It takes 1 min for a single image as reported by the authors.	512 × 512	–	CelebA
Liu et al. (2018a)	The network inference time is 0.029 s using NVIDIA V100 GPU (16 GB), regardless on the mask-to-image ratio on the image.	512 × 512	6	CelebA-HQ
Yan et al. (2018)	IT takes one day on a TITAN X Pascal i.e. 24 h for 30 epochs.	256 × 256	1	Paris Street View
Huang et al. (2019)	GeForce GTX 1070Ti	256 × 256	4	In-house road images. Huang et al. (2019)
Zeng et al. (2019)	The model runs at 0.19 s per frame on a TITAN V GPU.	256 × 256	–	CelebA-HQ
Li et al. (2019)	The training time is 3 days using RTX 2080TI 11G GPU for CelebA and two weeks for Places2.	256 × 256	5	CelebA Places2
Yu et al. (2019)	During testing, it takes 0.21 s per image using NVIDIA(R) Tesla(R) V100 GPU.	512 × 512	–	CelebA-HQ
Liu et al. (2019b)	It takes 9, 5 and 2 days for Places2, CelebA, and Paris Street View using NVIDIA 1080TI GPU(11GB). Overall inference time is 0.82 s per image.	256 × 256	1	Places2 CelebA Paris Street View.
Yi et al. (2020)	Trained using two NVIDIA GTX 1080TI GPU.	512 × 512	8	CelebA-HQ. Places2.
Zheng et al. (2019)	–	256 × 256	1	CelebA
Zhao et al. (2020)	It takes 500,000 iterations to train the model.	256 × 256	8	CelebA-HQ
Zhou et al. (2020)	It takes 4 days to train using NVIDIA TITAN Xp (12 GB).	256 × 256	16	CelebA-HQ

of methods from the perspective of the algorithm (its development and how it is used) for inpainting tasks, datasets, performance evaluation and limitations of the methods. We note the poor(er) performance of traditional methods on images with more extensive binary mask and facial images due to complexity in features on the image. Remedying this limitation, deep learning methods have developed to become state-of-the-art, showing great success on images containing intricate patterns, but also with shortcomings in computational complexity, failures in edge (due to mask size) and image realism preservation.

Research on image inpainting using deep learning has witnessed good progress in recent years. Many datasets and the masks region for inpainting were generated but lack of standardisation. We outline some potential works to advance the field:

- **Datasets**. To track and determine which model is the state-of-the-art in inpainting, and curtail the propagation of weak baselines into research, standardised training and testing datasets of images and masks, should be established.

- **Algorithms**. To enable reproducibility of the work, we encourage the transparency of the experiments by reporting the number of epochs and parameters for each method. This will allow future work to benchmarking against the baseline models. Current algorithms train on specific dataset and only work on the data with similar nature. Future work should explore a generalisable algorithms that can work on any image type.

- **Necessity of standardised performance metrics**. There are no standardised metrics to evaluate the performance of the algorithms. We recommend future research to report ℓ_1, ℓ_2, PSNR and FID as these metrics reflect different aspects of the performance.

- **Human Visual System**. The human visual system should be taken into consideration in more subjective evaluations to evaluate the perceptual quality of inpainted regions. For example, this can be in the form of Mean opinion score or measuring direct gaze information by direct bearing and the subjective rating of the observer.

We have noted that compared to traditional methods, deep learning methods have proven extremely powerful. For the future, we expect research efforts to continue to increase in this area of study. We therefore call on the community for more rigorous and improved practices with respect to reproducibility, evaluation and reduced complexity in computational cost.

CRediT authorship contribution statement

Jireh Jam: Conceptualization, Review, Investigation, Conduct comparison, Review datasets, Writing - original draft. **Connah Kendrick:** Supervision, Writing - review & editing. **Kevin Walker:** Conceptualization, Funding acquisition, Supervision, Writing - review the manuscript. **Vincent Drouard:** Collaboration, Identify the latest related work, Writing - review & editing. **Jison Gee-Sern Hsu:** Supervision, Writing - review & editing. **Moi Hoon Yap:** Conceptualization, Funding acquisition, Supervision, Writing - review, Structuring, Editing and handling the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank The Royal Society, UK (Grant number: IF160006 and INF/PHD/180007) and Kim's English Corner (https://kimsenglishcorner.com) for proofreading. All authors involved in writing, reviewing and editing the manuscript. All the authors have read and approved this version of the manuscript.

References

Abbad, A., Elharrouss, O., Abbad, K., Tairi, H., 2018. Application of meemd in post-processing of dimensionality reduction methods for face recognition. Iet Biometrics 8 (1), 59–68.

Akl, A., Yaacoub, C., Donias, M., Da Costa, J.-P., Germain, C., 2018. A survey of exemplar-based texture synthesis methods. Comput. Vis. Image Underst. 172, 12–24.

Allène, C., Paragios, N., 2006. Image renaissance using discrete optimization. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, Vol. 3. IEEE, pp. 631–634.

Amos, B., Image Completion with Deep Learning in TensorFlow. http://bamos.github.io/2016/08/09/deep-completion. Accessed: [Insert date here].

Angelova, A., Abu-Mostafam, Y., Perona, P., 2005. Pruning training sets for learning of object categories. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1. IEEE, pp. 494–501.

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J., 2010. Google street view: Capturing the world at street level. Computer 43 (6), 32–38.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein gan. arXiv preprint arXiv:1701.07875.

Ashikhmin, M., 2001. Synthesizing natural textures. In: Proceedings of the 2001 Symposium on Interactive 3D Graphics. Citeseer, pp. 217–226.

Avcibaş, I., Sankur, B., Sayood, K., 2002. Statistical evaluation of image quality measures. J. Electron. Imaging 11 (2), 206–223.

Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B., 2009. Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. (ToG) 28 (3), 24.

Batool, N., Chellappa, R., 2014. Detection and inpainting of facial wrinkles using texture orientation fields and Markov random field modeling. IEEE Trans. Image Process. 23 (9), 3773–3788.

Bengio, Y., et al., 2009. Learning deep architectures for ai. Found. Trends Mach. Learn. 2 (1), 1–127.

Bertalmio, M., Bertozzi, A.L., Sapiro, G., 2001. Navier-stokes, fluid dynamics, and image and video inpainting. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, Vol. 1. IEEE, p. I.

Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C., 2000. Image inpainting. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. ACM Press/Addison-Wesley Publishing Co., pp. 417–424.

Bertalmio, M., Vese, L., Sapiro, G., Osher, S., 2003. Simultaneous structure and texture image inpainting. IEEE Trans. Image Process. 12 (8), 882–889.

Bertozzi, A.L., Esedoglu, S., Gillette, A., 2006. Inpainting of binary images using the cahn–hilliard equation. IEEE Trans. Image Process. 16 (1), 285–291.

Bornard, R., Lecan, E., Laborelli, L., Chenot, J.-H., 2002. Missing data correction in still images and image sequences. In: Proceedings of the Tenth ACM International Conference on Multimedia. pp. 355–361.

Buades, A., Coll, B., Morel, J.-M., 2005. A non-local algorithm for image denoising. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 2. IEEE, pp. 60–65.

Bugeau, A., Bertalmío, M., Caselles, V., Sapiro, G., 2010. A comprehensive framework for image inpainting. IEEE Trans. Image Process. 19 (10), 2634–2645.

Buyssens, P., Daisy, M., Tschumperlé, D., Lézoray, O., 2015. Exemplar-based inpainting: Technical review and new heuristics for better geometric reconstructions. IEEE Trans. Image Process. 24 (6), 1809–1824.

Cao, F., Gousseau, Y., Masnou, S., Pérez, P., 2011. Geometrically guided exemplar-based inpainting. SIAM J. Imaging Sci. 4 (4), 1143–1179.

Chan, T.F., Shen, J.J., 2005. Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods, Vol. 94. Siam.

Chang, R.-C., Shih, T.K., 2008. Multilayer inpainting on digitalized artworks. J. Inf. Sci. Eng. 24 (4), 1241–1255.

Chen, Q., Montesinos, P., Sun, Q.S., Heng, P.A., et al., 2010. Adaptive total variation denoising based on difference curvature. Image Vis. Comput. 28 (3), 298–306.

Chen, J.-x., Zhu, Z.-l., Fu, C., Yu, H., 2014. A fast image encryption scheme with a novel pixel swapping-based confusion approach. Nonlinear Dynam. 77 (4), 1191–1207.

Cho, T.S., Butman, M., Avidan, S., Freeman, W.T., 2008. The patch transform and its applications to image editing. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, pp. 1–8.

Criminisi, A., Pérez, P., Toyama, K., 2004. Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. 13 (9), 1200–1212.

Darabi, S., Shechtman, E., Barnes, C., Goldman, D.B., Sen, P., 2012. Image melding: combining inconsistent images using patch-based synthesis. ACM Transactions on graphics (TOG) 31 (4), 1–10.

Daribo, I., Pesquet-Popescu, B., 2010. Depth-aided image inpainting for novel view synthesis. In: Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on. IEEE, pp. 167–170.

Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A., 2012. What makes paris look like paris?. ACM Trans. Graph. 31 (4).

Drori, I., Cohen-Or, D., Yeshurun, H., 2003. Fragment-based image completion. In: ACM Transactions on Graphics (TOG). ACM, pp. 303–312.

Efros, A.A., Freeman, W.T., 2001. Image quilting for texture synthesis and transfer. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. ACM, pp. 341–346.

Efros, A.A., Leung, T.K., 1999. Texture synthesis by non-parametric sampling. In: Iccv. IEEE, p. 1033.

Elharrouss, O., Al-Maadeed, N., Al-Maadeed, S., 2019a. Video summarization based on motion detection for surveillance systems. In: 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, pp. 366–371.

Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Akbari, Y., 2019b. Image inpainting: A review. Neural Process. Lett. 1–22.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88 (2), 303–338.

Faugeras, O., Keriven, R., 2002. Variational Principles, Surface Evolution, PDE's, Level Set Methods and the Stereo Problem. IEEE.

Gatys, L.A., Ecker, A.S., Bethge, M., 2015a. A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

Gatys, L.A., Ecker, A.S., Bethge, M., 2015b. Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. arXiv preprint arXiv:1505.07376 12.

Ghorai, M., Mandal, S., Chanda, B., 2018. A group-based image inpainting using patch refinement in mrf framework. IEEE Trans. Image Process. 27 (2), 556–567.

Goldman, D., Shechtman, E., Barnes, C., Belaunde, I., Chien, J., 2014. Content-aware fill. Accessed on 25.

Gomes, J., Darsa, L., Costa, B., Velho, L., 1999. Warping & Morphing of Graphical Objects. Morgan Kaufmann.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680.

Goyal, P., Diwakar, S., et al., 2010. Fast and enhanced algorithm for exemplar based image inpainting. In: Image and Video Technology (PSIVT), 2010 Fourth Pacific-Rim Symposium on. IEEE, pp. 325–330.

Gronat, P., Havlena, M., Sivic, J., Pajdla, T., 2011. Building Streetview Datasets for Place Recognition and City Reconstruction. Research Reports of CMP, Czech Technical University in Prague.

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

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. pp. 5767–5777.

Haccius, C., Herfet, T., 2017. Computer vision performance and image quality metrics: areciprocal relation. Computer Vision Performance and Image Quality Metrics-A Reciprocal Relation 1, 27–37.

Hays, J., Efros, A.A., 2007. Scene completion using millions of photographs. ACM Trans. Graph. 26 (3), 4.

He, K., Sun, J., 2014. Image completion approaches using the statistics of similar patches. IEEE Trans. Pattern Anal. Mach. Intell. 36 (12), 2423–2435.

He, L., Xing, Y., Xia, K., Tan, J., 2018. An adaptive image inpainting method based on continued fractions interpolation. Discrete Dyn. Nat. Soc. 2018.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. pp. 6626–6637.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.

Hore, A., Ziou, D., 2010. Image quality metrics: Psnr vs. ssim. In: 2010 20th International Conference on Pattern Recognition. IEEE, pp. 2366–2369.

Huang, Y., Wang, M., Qian, Y., Lin, S., Yang, X., 2019. Image completion based on gans with a new loss function. In: Journal of Physics: Conference Series. IOP Publishing, 012030.

Iizuka, S., Simo-Serra, E., Ishikawa, H., 2017. Globally and locally consistent image completion. ACM Trans. Graph. 36 (4), 107.

Iskakov, K., 2018. Semi-parametric image inpainting. arXiv preprint arXiv:1807.02855.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. arXiv preprint.

Jain, V., Seung, S., 2009. Natural image denoising with convolutional networks. In: Advances in Neural Information Processing Systems. pp. 769–776.

Jam, J., Kendrick, C., Drouard, V., Walker, K., Hsu, G.-S., Yap, M.H., 2020a. R-mnet: A perceptual adversarial network for image inpainting. arXiv preprint arXiv:2008.04621.

Jam, J., Kendrick, C., Drouard, V., Walker, K., Hsu, G.-S., Yap, M.H., 2020b. Symmetric skip connection wasserstein gan for high-resolution facial image inpainting. arXiv preprint arXiv:2001.03725.

Jia, J., Tang, C.-K., 2003. Image repairing: Robust image synthesis by adaptive nd tensor voting. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, Vol. 1. IEEE, p. I.

Jin, X., Su, Y., Zou, L., Wang, Y., Jing, P., Wang, Z.J., 2018. Sparsity-based image inpainting detection via canonical correlation analysis with low-rank constraints. IEEE Access 6, 49967–49978.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. Springer, pp. 694–711.

Jolicoeur-Martineau, A., 2018. The relativistic discriminator: a key element missing from standard GAN. arXiv preprint arXiv:1807.00734.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

Kawai, N., Sato, T., Yokoya, N., 2008. Image inpainting considering brightness change and spatial locality of textures.. In: VISAPP (1). Citeseer, pp. 66–73.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105.

Kuppig, R., 2015. Image Quilting for Texture Synthesis and Transfer. CS129-CS-Brown Edu, http://cs.brown.edu/courses/cs129/results/proj4/rkuppig/.

Kwok, T.-H., Sheung, H., Wang, C.C., 2010. Fast query for exemplar-based image completion. IEEE Trans. Image Process. 19 (12), 3106–3115.

Kwok, T.-H., Wang, C.C., 2009. Interactive image inpainting using dct based exemplar matching. In: International Symposium on Visual Computing. Springer, pp. 709–718.

Lai, Y.-K., Hu, S.-M., Gu, D., Martin, R.R., 2005. Geometric texture synthesis and transfer via geometry images. In: Proceedings of the 2005 ACM Symposium on Solid and Physical Modeling. pp. 15–26.

Le Meur, O., Gautier, J., Guillemot, C., 2011. Examplar-based inpainting based on local geometry. In: Image Processing (ICIP), 2011 18th IEEE International Conference on. IEEE, pp. 3401–3404.

Le Meur, O., Guillemot, C., 2012. Super-resolution-based inpainting. In: European Conference on Computer Vision. Springer, pp. 554–567.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4681–4690.

Li, J., He, F., Zhang, L., Du, B., Tao, D., 2019. Progressive reconstruction of visual structure for image inpainting. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5962–5971.

Li, Y., Liu, S., Yang, J., Yang, M.-H., 2017b. Generative face completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 33911–3919.

Li, H., Luo, W., Huang, J., 2017a. Localization of diffusion-based inpainting in digital images. IEEE Trans. Inf. Forensics Secur. 12 (12), 3050–3064.

Li, C., Wand, M., 2016. Combining markov random fields and convolutional neural networks for image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2479–2486.

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

Liu, H., Bi, X., Lu, G., Wang, W., 2019a. Exemplar-based image inpainting with multi-resolution information and the graph cut technique. IEEE Access 7, 101641–101657.

Liu, H., Jiang, B., Xiao, Y., Yang, C., 2019b. Coherent semantic attention for image inpainting. arXiv preprint arXiv:1905.12384.

Liu, Z., Luo, P., Wang, X., Tang, X., 2015a. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3730–3738.

Liu, Z., Luo, P., Wang, X., Tang, X., 2015b. Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV).

Liu, Z., Luo, P., Wang, X., Tang, X., 2018b. Large-scale celebfaces attributes (celeba) dataset. Retrieved August 15, 2018.

Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., Catanzaro, B., 2018a. Image inpainting for irregular holes using partial convolutions. arXiv preprint arXiv:1804.07723.

Liu, S., Wang, Y., Wang, J., Wang, H., Zhang, J., Pan, C., 2013. Kinect depth restoration via energy minimization with tv 21 regularization. In: Image Processing (ICIP), 2013 20th IEEE International Conference on. IEEE, p. 724.

Losson, O., Macaire, L., Yang, Y., 2010. Comparison of color demosaicing methods. In: Advances in Imaging and Electron Physics, Vol. 162. Elsevier, pp. 173–265.

Lou, S., Fan, Q., Chen, F., Wang, C., Li, J., 2018. Preliminary investigation on single remote sensing image inpainting through a modified gan. In: 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS). IEEE, pp. 1–6.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imaging 16 (2), 187–198.

Mairal, J., Elad, M., Sapiro, G., 2008. Sparse representation for color image restoration. IEEE Trans. Image Process. 17 (1), 53–69.

Mao, X.-J., Shen, C., Yang, Y.-B., 2016. Image restoration using convolutional auto-encoders with symmetric skip connections. arXiv preprint arXiv:1606.08921.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Mo, J., Zhou, Y., 2019. The research of image inpainting algorithm using self-adaptive group structure and sparse representation. Cluster Comput. 22 (3), 7593–7601.

Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M., 2019. Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212.

Paragios, N., Chen, Y., Faugeras, O.D., 2006. Handbook of Mathematical Models in Computer Vision. Springer Science & Business Media.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2536–2544.

Pérez, P., Gangnet, M., Blake, A., 2003. Poisson image editing. ACM Trans. Graph. 22 (3), 313–318.

Qureshi, M.A., Deriche, M., Beghdadi, A., Amin, A., 2017. A critical survey of state-of-the-art image inpainting quality assessment metrics. J. Vis. Commun. Image Represent. 49, 177–191.

Raad, L., Galerne, B., 2017. Efros and freeman image quilting algorithm for texture synthesis. Image Process. On Line 7, 1–22.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Ren, Y., Yu, X., Zhang, R., Li, T.H., Liu, S., Li, G., 2019. Structureflow: Image inpainting via structure-aware appearance flow. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 181–190.

Richard, M.M.O.B.B., Chang, M.Y.-S., 2001. Fast digital image inpainting. In: Appeared in the Proceedings of the International Conference on Visualization, Imaging and Image Processing (VIIP 2001), Marbella, Spain. pp. 106–107.

Roh, M.-C., Lee, S.-W., 2007. Performance analysis of face recognition algorithms on Korean face database. Int. J. Pattern Recognit. Artif. Intell. 21 (06), 1017–1033.

Ronneberger, O., Fischer, P., Brox, T.-n., 2015a. Convolutional networks for biomedical image segmentation. In: Paper Presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Ronneberger, O., Fischer, P., Brox, T., 2015b. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015a. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) 115 (3), 211–252. http://dx.doi.org/10.1007/s11263-015-0816-y.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015b. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Ružić, T., Pižurica, A., 2014. Context-aware patch-based image inpainting using Markov random field modeling. IEEE Trans. Image Process. 24 (1), 444–456.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. In: Advances in Neural Information Processing Systems. pp. 2234–2242.

Sheikh, H.R., Bovik, A.C., 2006. Image information and visual quality. IEEE Trans. Image Process. 15 (2), 430–444.

Shen, J., Chan, T.F., 2002. Mathematical models for local nontexture inpaintings. SIAM J. Appl. Math. 62 (3), 1019–1043.

Shen, B., Hu, W., Zhang, Y., Zhang, Y.-J., 2009. Image inpainting via sparse representation. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, pp. 697–700.

Shen, J., Kang, S.H., Chan, T.F., 2003. Euler's elastica and curvature-based inpainting. SIAM J. Appl. Math. 63 (2), 564–592.

Shen, Y., Luo, P., Yan, J., Wang, X., Tang, X., 2018. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 821–830.

Shih, T.K., Chang, R.-C., 2005. Digital inpainting-survey and multilayer image inpainting algorithms. In: Information Technology and Applications, 2005. ICITA 2005. Third International Conference on, Vol. 1. IEEE, pp. 15–24.

Shih, T.K., Lu, L.-C., Wang, Y.-H., Chang, R.-C., 2003. Multi-resolution image inpainting. In: 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698), Vol. 1. IEEE, pp. I–485.

Simakov, D., Caspi, Y., Shechtman, E., Irani, M., 2008. Summarizing visual data using bidirectional similarity. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, pp. 1–8.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sohn, K., Lee, H., Yan, X., 2015. Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems. pp. 3483–3491.

Song, Y., Yang, C., Shen, Y., Wang, P., Huang, Q., Kuo, C.-C.J., 2018. Spg-net: Segmentation prediction and guidance network for image inpainting. arXiv preprint arXiv:1805.03356.

Sridevi, G., Kumar, S.S., 2019. Image inpainting based on fractional-order nonlinear diffusion for image reconstruction. Circuits Systems Signal Process. 38 (8), 3802–3817.

Sun, J., Yuan, L., Jia, J., Shum, H.-Y., 2005. Image completion with structure propagation. In: ACM Transactions on Graphics (ToG). ACM, pp. 861–868.

Szeliski, R., Shum, H.-Y., Shum, H.-Y., Shum, H.-Y., 1997. Creating full view panoramic image mosaics and environment maps. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques. ACM Press/Addison-Wesley Publishing Co., pp. 251–258.

Tauber, Z., Li, Z.-N., Drew, M.S., 2007. Review and preview: Disocclusion by inpainting for image-based rendering. IEEE Trans. Syst. Man Cybern. C 37 (4), 527–540.

Telea, A., 2004. An image inpainting technique based on the fast marching method. J. Graph. Tools 9 (1), 23–34.

Thottam, I., 2015. The Cost of Conservation and Restoration. Art Business News, http://artbusinessnews.com/2015/12/the-cost-of-conservation-and-restoration/.

Tschumperlé, D., 2006. Fast anisotropic smoothing of multi-valued images using curvature-preserving pde's. Int. J. Comput. Vis. 68 (1), 65–82.

Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S., 2016. Texture networks: Feed-forward synthesis of textures and stylized images. In: ICML. pp. 1349–1357.

Wang, Z., Bovik, A.C., 2002. A universal image quality index. IEEE Signal Process. Lett. 9 (3), 81–84.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wang, Q., Fan, H., Sun, G., Cong, Y., Tang, Y., 2019. Laplacian pyramid adversarial network for face completion. Pattern Recognit. 88, 493–505.

Wang, H., Jiang, L., Liang, R., Li, X.-X., 2017. Exemplar-based image inpainting using structure consistent patch matching. Neurocomputing 269, 90–96.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8798–8807.

Wang, Z., Simoncelli, E.P., Bovik, A.C., 2003. Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Vol. 2. Ieee, pp. 1398–1402.

Wang, M., Yan, B., Gharavi, H., 2010. Pyramid model based down-sampling for image inpainting. In: 2010 IEEE International Conference on Image Processing. IEEE, pp. 429–432.

Wei, L.-Y., Levoy, M., 2000. Fast texture synthesis using tree-structured vector quantization. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. ACM Press/Addison-Wesley Publishing Co., pp. 479–488.

Xie, C., Liu, S., Li, C., Cheng, M.-M., Zuo, W., Liu, X., Wen, S., Ding, E., 2019. Image inpainting with learnable bidirectional attention maps. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8858–8867.

Xie, J., Xu, L., Chen, E., 2012. Image denoising and inpainting with deep neural networks. In: Advances in Neural Information Processing Systems. pp. 341–349.

Xu, Z., Sun, J., 2010. Image inpainting by patch propagation using patch sparsity. IEEE Trans. Image Process. 19 (5), 1153–1165.

Yan, Z., Li, X., Li, M., Zuo, W., Shan, S., 2018. Shift-net: Image inpainting via deep feature rearrangement. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 1–17.

Yang, S., Liu, J., Song, S., Li, M., Quo, Z., 2016. Structure-guided image completion via regularity statistics. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1711–1715.

Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H., 2017. High-resolution image inpainting using multi-scale neural patch synthesis. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 3.

Yeh, R.A., Chen, C., Lim, T.-Y., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N., 2017. Semantic image inpainting with deep generative models. In: CVPR. p. 4.

Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z., 2020. Contextual residual aggregation for ultra high-resolution image inpainting, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7508–7517.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018. Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5505–5514.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2019. Free-form image inpainting with gated convolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4471–4480.

Zeng, Y., Fu, J., Chao, H., Guo, B., 2019. Learning pyramid-context encoder network for high-quality image inpainting. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1486–1494.

Zhang, H., Dai, S., 2012. Image inpainting based on wavelet decomposition. Procedia Eng. 29, 3674–3678.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595.

Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., Lu, D., 2020. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5741–5750.

Zheng, C., Cham, T.-J., Cai, J., 2019. Pluralistic image completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1438–1447.

Zhou, T., Ding, C., Lin, S., Wang, X., Tao, D., 2020. Learning oracle attention for high-fidelity face completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7680–7689.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40 (6), 1452–1464.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2018. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40 (6), 1452–1464.

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E., 2017. Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems. pp. 465–476.

Zhuang, Y.-t., Wang, Y.-s., Shih, T.K., Tang, N.C., 2009. Patch-guided facial image inpainting by shape propagation. J. Zhejiang Univ.-Sci. A 10 (2), 232–238.