

An Approach for Objective Quality Assessment of Image Inpainting Results

Dylan Seychell

*Dept. of Computer and Communications Engineering
University of Malta
Msida, Malta
dylan.seychell@ieee.org*

Carl J. Debono

*Dept. of Computer and Communications Engineering
University of Malta
Msida, Malta
c.debono@ieee.org*

Abstract—Image Inpainting techniques are generally challenging to evaluate objectively due to the lack of comparative data, as a reference image of the new scene, does not exist.. This paper presents an approach that uses our newly released dataset specifically designed to allow objective evaluation of inpainting techniques. In this work we demonstrate how traditional inpainting techniques can be objectively evaluated and compared together with modern deep learning and adversarial approaches. We further demonstrate how an unsupervised technique compares better than deep learning approaches.

Index Terms—Inpainting, Dataset, RGBD, GANs, Computer Vision, Machine Learning Evaluation

I. INTRODUCTION

The evaluation of inpainting techniques is generally subjective due to the challenges involved in obtaining a ground truth that can be used to assess the quality of the result using objective metrics. The main contributions in the field of inpainting were evaluated using subjective methods that are based on user feedback. Approaches such as the Mean Objective Score (MOS) [1] have been used to evaluate inpainting techniques [2]. In other cases [3], results of inpainting techniques are presented without comparison with a quantifiable conclusion. When inpainting larger regions in an image, such as entire objects, the use of full-reference metrics such as Mean Square Error (MSE) or Peak Signal to Noise Ratio (PSNR) cannot be applied [4] unless there exists an identical image of the same scene without the inpainted object for comparison. This highlights the need for a dataset that is constructed in controlled conditions where new objects are added to a scene without variations to the conditions.

The in-depth discussion with users during subjective evaluation [2], brought to light the need for a more objective approach. This was one of the guiding principles in the development of the COTS¹ (Common Objects of a Traveling Scientist) dataset, presented in Section III-A, that was specifically designed to address this challenge in the inpainting evaluation. The procedure related to this objective approach is presented in Section III. Our previous work that used a monoscopic inpainting approach [2] is objectively evaluated

and compared to a modern deep learning approach [5] using the proposed method. The results are presented in Section IV.

II. INPAINTING TECHNIQUES

This section presents a brief survey of inpainting techniques. This is organized into two main parts. The first part presents the traditional inpainting approaches, including exemplar based inpainting. The second part presents modern deep learning approaches, including those using Generative Adversarial Networks (GANs).

A. Traditional Inpainting Approaches

Inpainting, or image completion, is the process of modifying an image in a low detectable way by filling the region of interest with texture that is known from another location within the image [6] or through a generative approach where the image is filled through techniques such as GANs that are explored in Section II-B1. Inpainting is a very important and popular function in graphics packages. It is used for image restoration, visual editing and also object removal [7] [8].

One of the most influential first inpainting approaches proposed by Bertalmio *et al.* [9] was based on the idea of diffusing information from around the target region into the missing gap. This approach used a variational method and partial differential equation (PDE). In a similar effort, Chan *et al.* [10] used a PDE and use the Euler-Lagrange equation to propagate the diffusion inwards towards the target. The change of information happens perpendicular to the edge of the mask towards the center of the region to be inpainted. The information used for this diffusion is based on lines of equal grey values, known as isophotes [8]. The information from extracted isophotes enable such techniques to efficiently preserve any structural information in the missing region. An improved fast marching approach was later proposed by Telea [11]. This approach is an improved method on the PDE techniques [8] where the computational overheads related to the propagation are removed. While this algorithm returns a result faster, its main disadvantage is the blurry effect that it leaves after the inpainting [8] as also visible in the examples of inpainting evaluation in this paper.

The concept of considering patches to solve the inpainting problem results from the work of Efros and Leung [12] that

¹The dataset is available for free as an open-source project on <http://cotsdataset.info>

works by recursively filling a gap inwards from the boundary of the empty region [13]. In this technique, the neighborhood of a pixel p is considered and similar pixels are selected using the sum of squared differences. This technique performs weakly when the inpainted region is considerably large [13].

The work of Criminisi *et al.* [14] builds upon this idea and improves it in two ways. The first improvement is that the filling order was changed from onion-peel to a priority scheme. Secondly, the entire patch is copied instead of taking single pixels.

B. Deep Learning Approaches

The main limitation of traditional inpainting approaches is the lack of semantic knowledge of the domain. Our work published in [2] also carries these limitations yet it was published before the techniques surveyed in this section. This is directly addressed by deep learning approaches that use models trained over a number of datasets. These approaches can be categorized into two: the use of a convolutional neural network (CNN) and the use of a generative adversarial network as surveyed below.

CNN approaches start by first filling the target region with placeholder values that is then passed to convolutional layers that would be already pre-trained with low level or medium level features [5]. This approach will first use content encoders [15] that feed the target region and then decode the feature space. This fills the target region with upsampled regions and the result is of a relatively low quality. Yang *et al.* [16] proposed an approach that uses the output of the content encoder as the input propagates texture information onto it from its trained model. The main limitations of these techniques are generally related to a lower quality output with diffusion of color or blurriness as witnessed in the traditional techniques.

Liu *et al.* [5] from NVIDIA present a technique that addresses the challenges of the other deep learning techniques. They employ the cost functions that are normally used in neural artistic style transfer as presented by Gatys *et al.* [17]. This approach uses two cost functions, one for the style and one for the content. The pre-trained network iteratively assesses the quality of the inpainting starting from a low quality result being fed as input to the cost functions. These cost functions would check the quality of the inpainted region until a certain threshold is reached and the network converges. The results of this paper state that a result is achieved in 0.029s. However, this is achieved when running the network on a NVIDIA V100 GPU [5]. There is also a significant effort of training required prior to being able to use the network. This paper [5] reports that the model was trained on 55,116 masks and tested on 24,866. Another limitation is that the image size was that of 512×512 pixels. An NVIDIA V100 GPU (16GB) with a batch size of 6 was used for training. This process took 10 days to train on ImageNet and Places2 while the CelebA-HQ took 3 days [5].

1) *Generative Adversarial Networks*: GANs are a framework composed of a pair of deep neural networks that are able

to generate new content [18]. The concept of GANs involves the training of two separate networks, the Generator G and the Discriminator D . Starting from random noise z , G starts to convert the content into more meaningful content, such as an image. On the other hand, for every output x from G , the discriminator D needs to output a probability of whether x is a real artefact or an output of G . The aim of the framework is to optimize $V_{\text{GAN}}(D, G)$ as a two-player minimax zero-sum game between the two networks [18] as outlined in Equation 1.

$$\min_G \max_D V_{\text{GAN}}(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]. \quad (1)$$

where $\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})}$ is the expectation of $\log D(\mathbf{x})$ with respect to $p_{\text{data}}(\mathbf{x})$ and where $\mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})}$ is the expectation of $p_{\mathbf{z}}(\mathbf{z})$ with respect to $\log(1 - D(G(\mathbf{z})))$.

Since 2017, GANs have also been applied to the problem of inpainting [19] [20] [21]. Given adequate training, GANs can be used to fill gaps in an image using a semantic context [21]. This matches the feedback given in our evaluation of our inpainting approach in [2] where users remarked about the importance of the preservation of visual structure when inpainting images. Other recent work [22] [23] [24] also builds upon such needs and returns a sharper inpainted region which is consistent within its neighborhood context.

While the quality of inpainting through GANs achieved very high levels, it still faces significant challenges in the general context of the inpainting problem. The main limitation is related to the training required, particularly on the training of the generator network [24]. The adversarial loss convergence can also be a challenge with specific training datasets [21]. This also implies the reliance on the datasets and therefore, situations not featuring clearly in datasets would result in the GAN inpainting approach being limited to a selection of domains. Moreover, the reliance on the training of deep neural networks has a direct implication on the time of training and the generation of results [23]. The current techniques surveyed focus on the quality of images produced by the frameworks, which is considerably good. The choice on whether GANs or traditional inpainting techniques are used depends solely on the trade-off required between quality and acquisition of results.

III. METHODOLOGY

A. COTS Dataset

COTS (Common Objects of a Traveling Scientist) Dataset, is a travel-themed dataset containing 120 different instances organized in a selection of scenes. The dataset has been made available online for free². The selected objects were configured in different scenes specifically designed to be useful in a variety of computer vision applications. The COTS dataset contains different instances of specific scenes with multiple objects. Every instance contains an object that was not present

²<http://cotsdataset.info>

in the previous scene as illustrated in Figure 1. An 8-bit depth map of the scene and ground truth binary image of every object for every scene is also available in the COTS dataset. Every scene has a uniform green background so that it can be eventually computationally replaced by any other background. This leaves the target objects together with their respective depth map and ground truth binary image the same while adding as much clutter in the background as required by the experiment. The background can also be replaced with different textures. This dataset was specifically designed to address the limitation in the current approaches of evaluating inpainting algorithms where these relied on subjective user evaluation. This paper demonstrates how the COTS dataset can be used to objectively evaluate inpainting algorithms and make results more comparable.


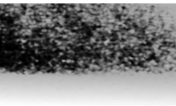




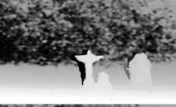




Instance	RGB	Depth	GT New Object
0			NA
1			
2			
3			

Fig. 1. Sample of a single incremental scene for the objective evaluation of inpainting applications. The rows represent different instances of the same scene with a single object being included in every scene. For every instance, one finds the RGB image together with its respective 8-bit depth image and the ground truth binary image for the new object being included.

B. Experimental Setup

The COTS dataset presented in Section III-A addresses the limitation of subjective evaluation by design. The second part of the dataset contains a selection of 23 scenes with multiple instances as demonstrated in Figure 1. Every instance includes a new object that was not present in the previous instance, without changing anything else in the scene. This allows for an evaluation process that enables the researcher to use an inpainting technique to remove an object from instance n and compare it with instance $n - 1$ that would not include the object being inpainted as demonstrated in Figure 1.

For further ease of use, the ground truth binary image for the newly included object is available for every instance. This can be beneficial since the evaluator can use this mask to guide the inpainting algorithm without the need of generating it through object detection and segmentation. An inpainting framework [2] was chosen to demonstrate how the COTS dataset can be used to evaluate an inpainting technique. The process is

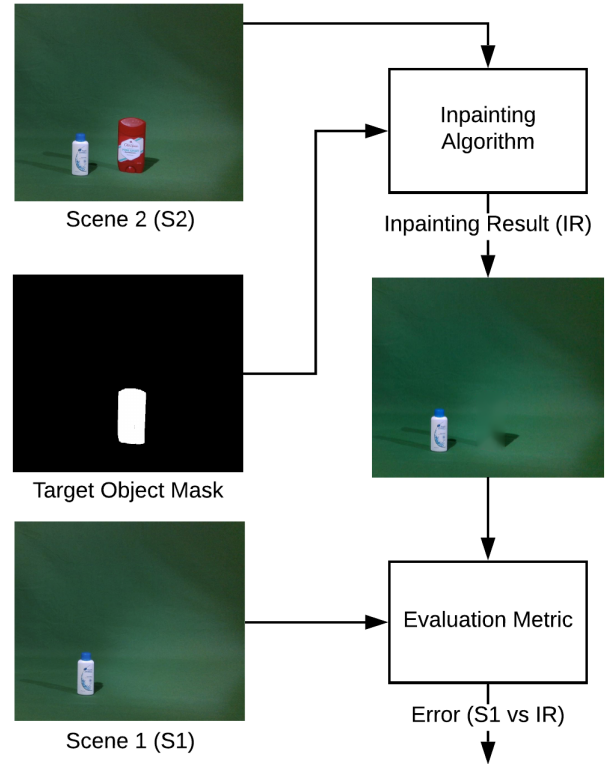


Fig. 2. The process of objectively evaluating the quality of inpainting using the COTS dataset.

presented in Figure 2. The red deodorant in Scene 2 (S2) was chosen as the target object to be inpainted. In this case, the Telea inpainting technique [11] was used within the above mentioned framework. The inpainting result can be evaluated against Scene 1 (S1) since it is S2 without the target object. Since the lighting conditions and setup were preserved in the dataset construction, S1 can be considered as the ground truth of S2 without the inpainted object. Evaluation metrics such as the MSE, PSNR or structural similarity index (SSIM) thus can be used for the comparison of S1 against the inpainting result. The dataset includes shadows so that future inpainting techniques that can trace the shadows of objects and inpaint them accordingly can also be evaluated using this dataset.

IV. EVALUATION

Six scenes from the COTS dataset were chosen for an objective comparative study of different inpainting techniques. Half of the scenes included occluded objects and the other half did not. The occluded scenes were the statues, shooter glasses and academic books scenes. The other scenes, namely the footwear, mugs and tech scenes did not include occluded objects.

The inpainting evaluation procedure visualized in Figure 2 was followed for each of the six scenes and the results are illustrated in Figure 3. The scenes marked as S2 contain an

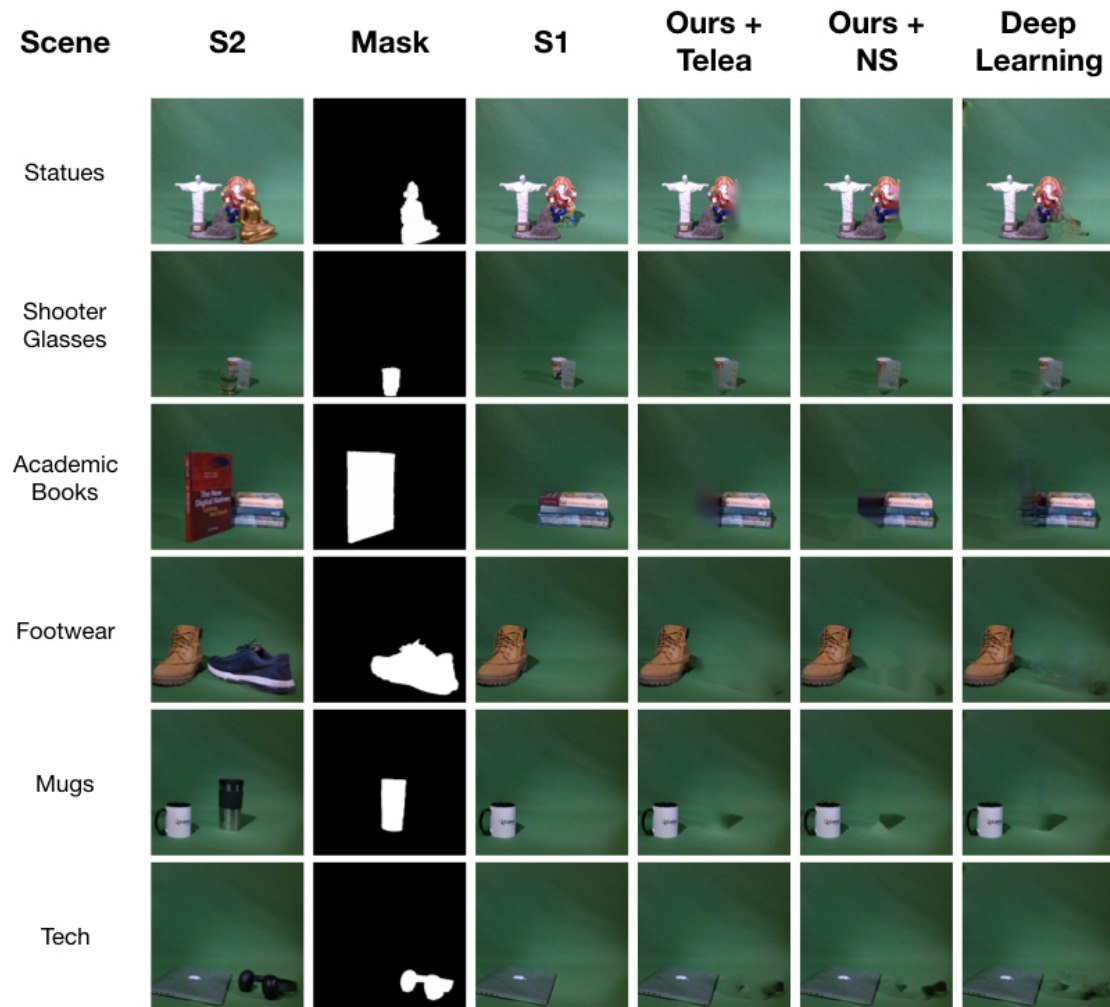


Fig. 3. A visual representation of the comparative result of the objective evaluation inpainting techniques. S2 is the original scene upon which the inpainting is carried out using the mask presented in the second column. S1 in the third column is the actual scene without the object represented by the mask, hence acting as ground truth of the inpainting. The last three columns are the results of different inpainting techniques. These are namely our technique with Teala's [11] and Bertalmio *et al.* [9] and the NVIDIA deep learning method of [5].

object that is represented by a binary mask. S1 is an actual instance of the scene without the object and this instance was a specific feature of the COTS dataset. This setup was used to compare three inpainting approaches. The first two use our approach presented in [2] first with Teala's [11] approach and then with Bertalmio *et al.* [9] approach. A deep learning approach was also used for comparison and NVIDIA's approach by Liu *et al.* [5] was used accordingly. The visual quality of the result in this situation where objects were placed in front of a plain background was very interesting. The dispersion based approaches gave a relatively blurred output and matched the result in previous work [2]. On the other hand, the deep learning approach gave a more crisp result when objects were occluded even though the quality of inpainting would still not score high marks in the subjective context. Moreover, the quality of inpainting when the inpainted object has a blank background had a visual quality comparable to the result of the traditional techniques. Deep learning techniques

TABLE I
THE RESULTS FROM THE COMPUTATION OF THE MSE OF THE INPAINTING TECHNIQUES IN RELATION TO MAXIMUM ERROR RETURNED WHEN THE S2 IS COMPARED TO S1.

		Mean Squared Error (MSE)			
		Occlusion	Ours + Telea	Ours + NS	Deep Learning
Statues	Yes		369.10	452.39	455.79
Shooters	Yes		57.20	68.17	72.11
Academic	Yes		384.76	488.48	484.78
Footwear	No		58.64	69.12	124.73
Mugs	No		79.31	101.61	108.91
Tech	No		112.46	153.91	142.79
					Max Error
					1139.27
					83.09
					1990.00
					1617.40
					407.76
					570.52

perform well when the inpainted object is surrounded by other objects or is in an outdoor environment due to the nature of the scenes used in the training data.

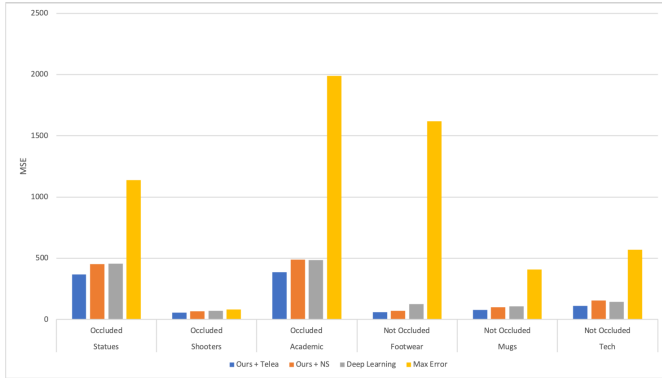


Fig. 4. A chart showing the quantitative results of the inpainting comparison. The y-axis represents the mean square error (MSE). The first 3 scenes include occluded items behind the target object and the other three do not. The Maximum error represents the error of the scene when S1 is compared with S2.

TABLE II
PERFORMANCE ANALYSIS OF THE INPAINTING TECHNIQUES IN RELATION TO THE PERCENTAGE AREA OF THE OBJECT TO THE ENTIRE SCENE. THE TIME DATA RELATED TO THE DEEP LEARNING TECHNIQUE IS AS SPECIFIED IN THE ORIGINAL PAPER BY LIU *et al.* IN [5].

	Time (Telea) [s]	Time (NS)[s]	Time (DL) [s]	Target Object Size [%]
Statues	3.63	2.83	0.03	6.4%
Shooters	1.51	1.06	0.03	2.0%
Academic Books	8.67	6.94	0.03	20.5%
Footwear	7.00	5.65	0.03	14.7%
Mugs	3.20	2.49	0.03	5.4%
Tech	2.70	2.20	0.03	3.9%
Average	4.86	3.53	0.03	8.8%

Processor	CPU	CPU	GPU V100	-
Training	NA	NA	3 - 10 Days	-

A. Results

The MSE was used to compare the quality of the inpainted result for the three designated techniques against the ground truth S1. The results are illustrated in the chart in Figure 4. The maximum error occurs when S1 is compared with S2 and therefore is the comparison of the two true scenes with and without the object that needs to be inpainted. The inpainted results are compared to S1 so their error should be as far as possible from the maximum error, hence closer to S1. In general, the Telea inpainting performs the best when compared to the other techniques.

A performance evaluation of the techniques was also carried out and the results are presented in Table II. An instance of our approach using Telea and NS inpainting was implemented in Python 3.6 and OpenCV 3.0 as a proof of concept. Performance testing was carried out on a machine with a 2.6 GHz Intel Core i7 processor and 16 GB 2400 MHz DDR4 memory running a MacOSX Mojave 10.14.6 operating system. The deep learning technique by Liu *et al.* was accessed using the online instance provided by NVIDIA³ and the performance

³<https://www.nvidia.com/research/inpainting/>

evaluation and training data is as reported in the results of the same paper [5].

These results show that there is a direct relationship in performance between our method and the size of the object that needs to be inpainted. The average time for the Telea approach was 4.45s while the average time for NS was 3.52s running on the above mentioned CPU architecture. These techniques are unsupervised and the time also includes the generation of the masks. Once trained, the deep learning approach returns the results significantly faster. However, this puts aside the effort taken to train the model. The paper specifies that a NVIDIA GPU V100 16 GB was used and training took between 3 to 10 days depending on the dataset. This means that if our technique is redesigned to exploit the GPU architecture it can potentially improve in terms of performance while still not requiring any effort in training and therefore can be used in different situations. The quality of the empirical results presented in Figure 4 compared with the performance of the techniques show that the choice of architecture or technique should depend on the priority between time or quality of inpainting. Collectively, these show that in general our approach is more efficient than the deep learning approach.

V. CONCLUSION

This paper presented an objective method for the comparison of inpainting techniques. We briefly presented a survey of traditional and modern deep learning inpainting approaches. This paper also presented a specifically designed dataset that enables an objective approach for inpainting evaluation. The efficacy of this method was demonstrated when our previously published monoscopic inpainting approach [2] was compared to NVIDIA's inpainting approach that employs generative adversarial networks [5]. In this case, the results show that despite not involving any training, our approach performs slightly better than the approach using GANs that requires significant training effort on high-end hardware.

REFERENCES

- [1] "ITU-T BT500. Methodology for the subjective assessment of the quality of television pictures," Aug 2012.
- [2] D. Seychell and C. J. Debono, "Monoscopic inpainting approach using depth information," in *Proc. of the 18th IEEE Mediterranean Electrotechnical Conference*, 2016.
- [3] M. K. Nanduri and K. S. Venkatesh, "Segmentation directed inpainting," in *Proc. of the 2016 27th Irish Signals and Systems Conference (ISSC)*, pp. 1–6, June 2016.
- [4] T. T. Dang, A. Beghdadi, and M. C. Larabi, "Inpainted image quality assessment," in *Proc. of the European Workshop on Visual Information Processing (EUVIP)*, pp. 76–81, June 2013.
- [5] G. Liu, F. A. Reda, K. J. Shih, T. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," *Proc. of the 2018 European Conference on Computer Vision (ECCV)*, Sept 2018.
- [6] L. Wang, H. Jin, R. Yang, and M. Gong, "Stereoscopic inpainting: Joint color and depth completion from stereo images," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.
- [7] S. Shivanjani and R. Priyadharsini, "A survey on inpainting techniques," in *Proc. of the 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp. 2934–2937, March 2016.

- [8] S. Zarif, I. Faye, and D. Rohaya, "A comparative study of different image completion techniques," in *Proc. of the 2014 International Conference on Computer and Information Sciences (ICCOINS)*, pp. 1–6, June 2014.
- [9] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proc. of the 2001 IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 355–362, 2001.
- [10] T. F. Chan and J. Shen, "Nontexture inpainting by curvature-driven diffusions," *Journal of Visual Communication and Image Representation*, vol. 12, no. 4, pp. 436 – 449, 2001.
- [11] A. Telea, "An image inpainting technique based on the fast marching method.," *Journal of Graphics, GPU and Game Tools*, vol. 9, no. 1, pp. 23–34, 2004.
- [12] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proc. of the International Conference on Computer Vision (ICCV 99)*, ICCV '99, (Washington, DC, USA), pp. 1033–1040, IEEE Computer Society, 1999.
- [13] M. S. Bertalmio A., Caselles V. and S. G., "Inpainting," in *Computer Vision: A Reference Guide* (I. K., ed.), Springer, 2014.
- [14] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *Image Processing, IEEE Transactions on*, vol. 13, pp. 1200 –1212, Sept 2004.
- [15] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544, June 2016.
- [16] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4076–4084, July 2017.
- [17] L. Gatys, A. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv*, vol. abs/1508.06576, 2015.
- [18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 2672–2680, Curran Associates, Inc., 2014.
- [19] K. Zhu, X. Liu, and H. Yang, "A survey of generative adversarial networks," in *Proc. of the 2018 Chinese Automation Congress (CAC)*, pp. 2768–2773, Nov 2018.
- [20] L. Yuan, C. Ruan, H. Hu, and D. Chen, "Image inpainting based on patch-gans," *IEEE Access*, vol. 7, pp. 46411–46421, 2019.
- [21] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544, June 2016.
- [22] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, pp. 107:1–107:14, July 2017.
- [23] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4076–4084, July 2017.
- [24] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6882–6890, July 2017.