https://doi.org/10.2352/lim.2024.5.1.7
©2024 Society for Imaging Science and Technology

Auto-MAT: Image Denoising via Automatic In-painting

Abdullah Hayajneh,Texas A&M University, College Station, TX, a.hayajneh@tamu.edu
Erchin Serpedin, Texas A&M University, College Station, TX, eserpedin@tamu.edu
Mitchel Stotland, Sidra Medicine, Doha, Qatar, mstotland@sidra.org

Abstract

This paper introduces an innovative blind in-painting technique designed for image quality enhancement and noise removal. Employing Monte-Carlo simulations, the proposed method approximates the optimal mask necessary for automatic image in-painting. This involves the progressive construction of a noise removal mask, initially sampled randomly from a binomial distribution. A confidence map is iteratively generated, providing a pixel-wise indicator map that discerns whether a particular pixel resides within the dataset domain. Notably, the proposed method eliminates the manual creation of an image mask to eradicate noise, a process prone to additional time overhead, especially when noise is dispersed across the entire image. Furthermore, the proposed method simplifies the determination of pixels involved in the in-painting process, excluding normal pixels and thereby preserving the integrity of the original image content. Computer simulations demonstrate the efficacy of this method in removing various types of noise, including brush painting and random salt and pepper noise. The proposed technique successfully restores similarity between the original and normalized datasets, yielding a Binary Cross Entropy (BCE) of 0.69 and a Peak-Signal-to-Noise-Ratio (PSNR) of 20.069. With its versatile applications, this method proves beneficial in diverse industry and medical contexts.

Introduction

The challenge of image semantic restoration involves the task of deriving a clean, original image from a corrupted version. One approach to image restoration involves the use of image in-painting techniques, which involve replacing pblackeetermined image pixels with new content based on neighboring regions and the overall context of the image. Noteworthy progress has been made in addressing the image in-painting problem [12, 22, 31, 34]. However, employing manual image inpainting for the restoration of corrupted images presents its own set of difficulties. Consider a scenario where an effective inpainting method is available for the restoration of a corrupted image. The manual selection of each individual pixel becomes arduous and time-consuming when corrupted pixels are scatteblack throughout the entire image. Additionally, human judgment encounters challenges in determining the corruption status of specific pixels. To illustrate, anomalous human faces pose a difficulty in objectively distinguishing between normal and abnormal facial pixels, leading to varied recommendations from different individuals.

Blind inpainting [3, 15] enables the automatic generation of the mask necessary for the inpainting procedure. It can be useful in many applications including image restoration, data recovery, and medical imaging. There has been limited progress in this research field thus far. The proposed approaches leverage deep learning techniques for image inpainting, presuming that the pixels requiring inpainting are filled with constant values or Gaussian noise. Such assumptions facilitate pixel identification. However, for different types of contaminations, determining the underlying probability distribution of the noise becomes challenging. A recent inpainting model, VCNet [24], employs Generative Adversarial Networks (GANs) to establish a two-phase blind inpainting model. The initial phase utilizes a Convolutional Neural Network (CNN) to estimate the mask, while the second phase carries out the image inpainting process. Similar to prior methods, VCNet assumes that image contamination could originate from another dataset (another image), limiting the effectiveness of blind inpainting. To address these challenges, the sole assumption made is to regard the input image as originating from a particular probability distribution, while any noisy pixel contents are assumed to be drawn from an unspecified data distribution. In this study, we propose a simulation-based approach for estimating a mask that eliminates undesiblack areas in images. Undesirable areas are defined based on substantial differences observed after applying image inpainting to the input image using a proficiently trained inpainting model.

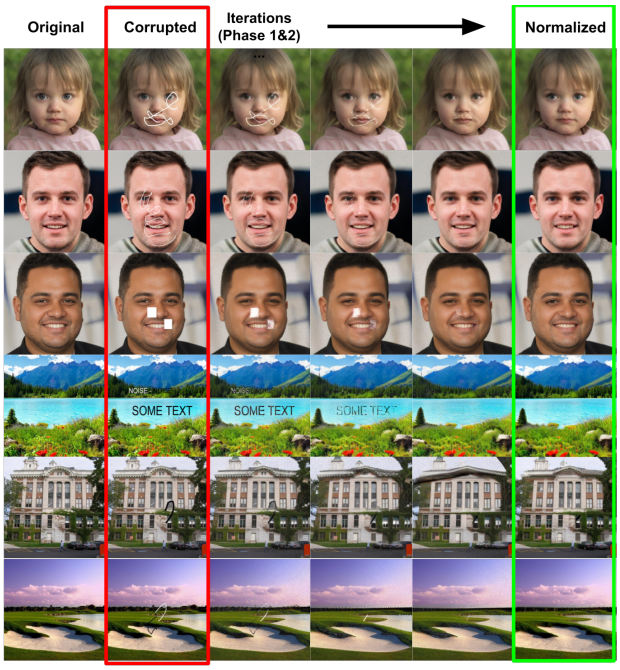

Figure 1. *Some examples demonstrating the effectiveness of the proposed method for noisy images.*

This paper introduces several key contributions:

- A pioneering blind inpainting model that utilizes a high-performing inpainting model to identify potential locations of out-of-distribution pixels.
- A Monte-Carlo simulation-based approach to generate a confidence heatmap on the input image, revealing the likelihood of a given pixel being normal or anomalous.
- The development of a thresholding technique to transform the confidence heatmap into a mask suitable for the image

inpainting process.

Related Work

Inpainting methods utilize internal/local image information and external (geometrical) rules to replace missing pixel blobs of images with new content [2, 5, 10, 11]. Regarding automatic inpainting, current research focuses on images corrupted by simple and small data structures such as text or thin strokes with constant intensity [3, 15]. For these scenarios, data interpolation from local regions is even possible at least for simple inpainting cases. In the case of substantial gaps or holes in images, recent approaches have successfully employed generative models [1, 16, 32, 23] for image inpainting. These generative models, which may be conditional, produce commendable results by leveraging image context and adhering to label constraints [29, 27, 26, 33].

Various inpainting approaches have been proposed, showcasing diverse strategies. For instance, Pathak et al. [18] constructed and trained an encoder-decoder model using reconstruction and adversarial losses. Iizuka et al. [8] utilized multi-discriminators for both global and local image regions, enhancing inpainting quality. Coarse-to-fine [29, 27] or multi-branch [25] networks were employed to achieve improved inpainting quality. Yang et al. [28] implemented an optimization-based approach, incorporating texture, semantics, and smoothness details in the loss function to enhance image inpainting.

Attention networks [28] were utilized for inpainting by learning how to incorporate best-fit background information into the foreground, while Wang et al. [25] introduced an implicit diversified Markov random fields (ID-MRF) loss during training to facilitate best-fit searches without extensive computation.

Additional enhancements consideblack image structures such as edges [17] and object-level representations [27] in a two-stage operation. Zheng et al. [33] introduced pluralistic inpainting results, and Sagong et al. [20] designed more efficient image generation methods. Alternative research directions exploblack convolution variants, like partial and gated convolutions, for obtaining highly detailed inpainting results directly [14, 30]. Some studies investigated masking operations applied to image representations beyond the input image. Moreover, automatic inpainting tasks, such as raindrop removal [19], relied on prior assumptions and feature statistics to detect whether a pixel is clean or requires inpainting.

Methods

Consider an original human face image with facial deformity, denoted as $x_{org} \in \mathbb{N}^{n\times m\times c}$, where n, m, and c represent its height, width, and the number of color channels, respectively. The primary research goal is to derive an optimal mask $\mathbf{M}^*$ that highlights potential out-of-distribution pixels. This mask is then utilized to generate a normalized version, x_{norm}, from the original image x_{org}. Let G represent a reliable inpainting method, implying that for a given image pixel $x_{org,i,j}$, if the pixel is normal, its brightness level is expected to be restoblack to its original value if tested for inpainting. This assumption hinges on the premise that the majority of image pixels are normal. If the inputs to G are x_{org} and mask M, then the output x_{norm} is expressed as:

$$x_{norm} = G(x_{org}, \mathbf{M}). \tag{1}$$

The overall workflow of operations is shown in Figure 2.

Initial Mask Estimation

Given that $\mathbf{M}$ is unknown, the proposed algorithm iteratively estimates the optimal mask $\mathbf{M}^*$ by starting from an initial

Algorithm 1 Automatic Mask Generation. Input: Face image x_{org}, MAT in-painting model G, in-painting probability pr, noise threshold $t(=1)$, $r = 1.0$ and in-painting iterations L. Output: Normalized image x_{norm} and optimal mask $\mathbf{M}^*$.

$\mathbf{M}^0_{mother} \leftarrow \mathbf{0}$ *# estimate the initial mask*
$D_c \leftarrow \mathbf{0}$
For $l = 1, \ldots, L$
  $\mathbf{M}_{rnd} \leftarrow \mathbf{Bin}(1, pr)^{n\times m}$
  $\mathbf{M}^l_{mother} \leftarrow \mathbf{M}^{l-1}_{mother} \cup \mathbf{M}_{rnd}$
  $x^l_{norm} \leftarrow G(x_{org}, \mathbf{M}^l_{mother})$
  $D_c \leftarrow D_c + ((x_{org} - x^l_{norm}) \odot (x_{org} - x^l_{norm}))^{\circ 2}$
  For each $x_{org,i,j}$ in x_{org}
    if $|x_{org,i,j} - x^l_{norm,i,j}| < t$ then
      $\mathbf{M}^l_{mother,i,j} \leftarrow 0$
  End For
End For
$H \leftarrow hist(D_c)$ *# thresholding the confidence map*
$\sigma \leftarrow std(H)$
$t_{opt} \leftarrow argmax(H) + r\sigma$
  For each $D_{c,i,j}$ in D_c
    if $D_{c,i,j} < t_{opt}$ then
      $\mathbf{M}^L_{mother,i,j} \leftarrow 0$
    else
      $\mathbf{M}^L_{mother,i,j} \leftarrow 1$
  End For
$\mathbf{M}^* \leftarrow \mathbf{M}^L_{mother}$
$x^*_{norm} \leftarrow G(x_{org}, \mathbf{M}^*)$
return $\mathbf{M}^*$, x^*_{norm}

random mask $\mathbf{M}_{rnd}$ sampled from a binomial distribution, and employs it to inpaint x_{org}. Subsequently, the resulting x^l_{norm} is compablack to x_{org} through pixel-wise subtraction. If the difference for a pixel is significant, it indicates an out-of-distribution pixel requiring replacement. The set of pixel locations needing replacement in x_{org} is stoblack in a distinct "mother mask" $\mathbf{M}_{mother}$ using the union operation. Subsequently, a new random mask is generated and combined with $\mathbf{M}_{mother}$ to identify potential pixels for replacement. This sequence of steps repeats for a total of L iterations. Concurrently with estimating $\mathbf{M}_{mother}$, a "cumulative difference map" D_c is computed by aggregating the difference map D between x_{org} and x^l_{norm} at each iteration. The cumulative difference map serves to construct a confidence map indicating the locations of pixels that require replacement.

Thresholding the Cumulative Difference Map

Once $\mathbf{M}_{mother}$ has been acquiblack, the subsequent phase involves enhancing it through the utilization of the cumulative difference map D_c. This map represents the variation level experienced by each pixel throughout the L inpainting iterations. Converting D_c into a final mask involves assigning a value of 1 to pixels exhibiting significant variation, signifying the necessity for inpainting. Pixels with minimal variation in D_c are assigned a value of 0, indicating that they are normal. To establish a threshold for the division of pixels into two groups, a histogram of pixel variation (H) is generated. The mean and standard deviation (σ) of the variation values are calculated, serving to determine the optimal threshold as follows:

$$t_{opt} = argmax(H) + r\sigma\ . \tag{2}$$

Figure 3 displays a histogram plot H belonging to a facial image example. In the final step, the new mask M_{mother} is applied to in-

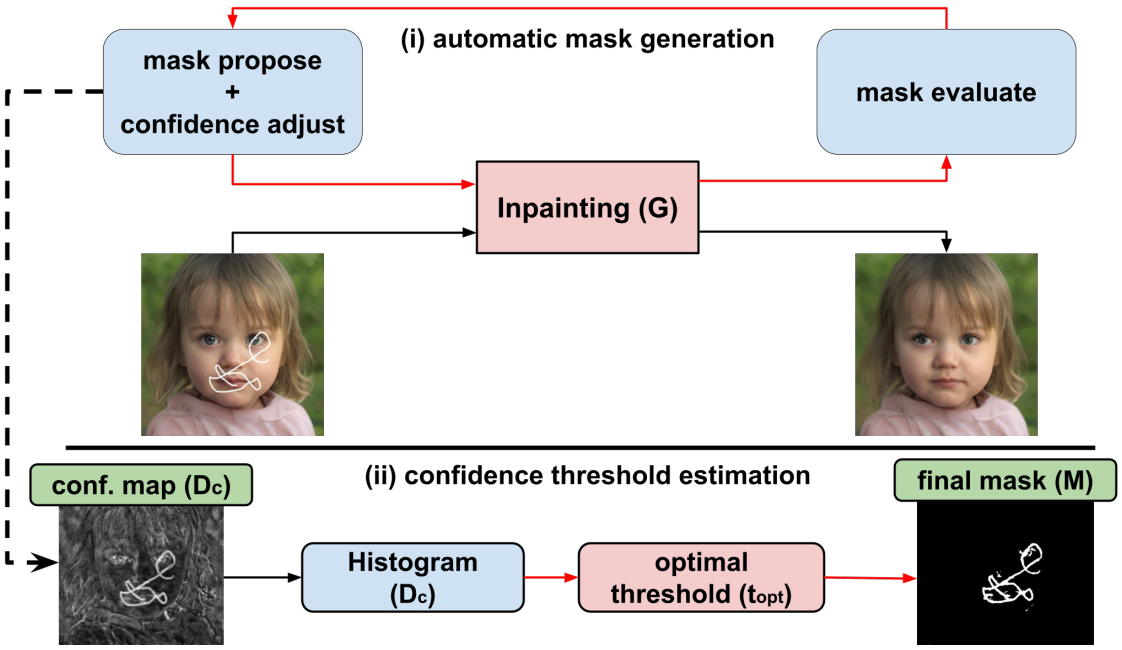

Figure 2. *Flowchart of the proposed blind inpainting operations.*

paint x_{org} in a single forward pass, resulting in the optimal x_{norm} with noise fully eliminated. The process for estimating the mask M is delineated in Algorithm 1. For this algorithm to offer the best performance, a good inpainting method has to be utilized. A good inpainting method is assumed to keep normal pixels unchanged.

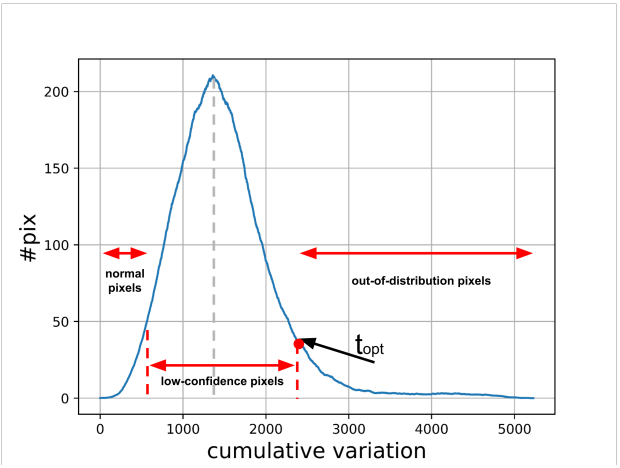

Figure 3. *The histogram of the confidence map (D_c) shows that the cumulative variation is concentrated in a specific and limited mid-range (500-2500). Most of the pixels experience variations with low confidence, meaning that these pixels vary during most of the processing iterations. Therefore, these pixels should not be inpainted. Only pixels with high inpainting confidence (> 2500) should be inpainted.*

Experiments

Two datasets were used to evaluate the performance of the proposed method. The first dataset consists of 150 facial images from the CelebA-HQ dataset. The second dataset contains 150 images of natural scenes obtained from the Places2 dataset. Both datasets were contaminated with a set of 150 masks containing synthetic random brush strokes with varying thicknesses (4 and 8 pixels). The proposed algorithm is tested to remove these random strokes from the original image x_{org}, and the result of removal is stoblack in x_{norm} along with its final estimated mask $\mathbf{M}$. The Mask Aware Transformer (MAT) [12] inpainting model was utilized in the proposed framework during the inpainting step in Algorithm 1. A network pretrained on the FFHQ dataset was employed to estimate the mask for the faces dataset, while another pretrained network on the Places2 dataset was utilized for the Places images. MAT showed a good inpainting performance on the FFHQ and Places2 datasets. The number of inpainting iterations L was set to be 50 to ensure most pixel combinations were tested for inpainting. To test for the automatic inpainting performance, the Ground Truth (GT) masks are compablack with the estimated ones for both datasets using the Binary Cross Entropy (BCE). Also, the Peak-Signal-to-Noise-Ratio (PSNR) and Structural Similarity Index (SSIM) between $\{x_{org}\}$ and $\{x_{norm}\}$ are calculated to evaluate the quality of the inpainted images and compablack to the quality of the GT datasets. Different values of success probability pr and inpainting threshold were also tested to obtain the best combination. The best-performing combination was used to compare the Auto-MAT against VCNet. An evaluation dataset is only requiblack to test if the parameters in the proposed method generalize well to eliminate the noise from the samples. This proves valuable in situations where abnormal samples are scarce within a dataset.

Results and Discussion

To compare the proposed method with other competing blind inpainting models, Table 1 shows the BCE, SSIM, and PSNR of the proposed algorithm and for VCNet on the masked FFHQ and Places2 datasets. The proposed algorithm outperforms VCNet in terms of mask estimation and image quality in both datasets. The results of the Auto-MAT show that it can perform well in both the mask estimation as well as the inpainting phases. However, VCNet did not succeed well in performing the blind inpainting on the images under analysis. More visual results are shown in Figure 4. VCNet's design may not generalize to different pixel contaminations, relying on specific assumptions about image noise during training.

Table 2 shows the blind inpainting performance of the masked FFHQ dataset using the BCE, SSIM, and PSNR metrics, under different combinations of hyperparameters. The results show that the best combination is when using the threshold $t = 8$ and $pr = 0.9$. This combination archives a BCE, SSIM,

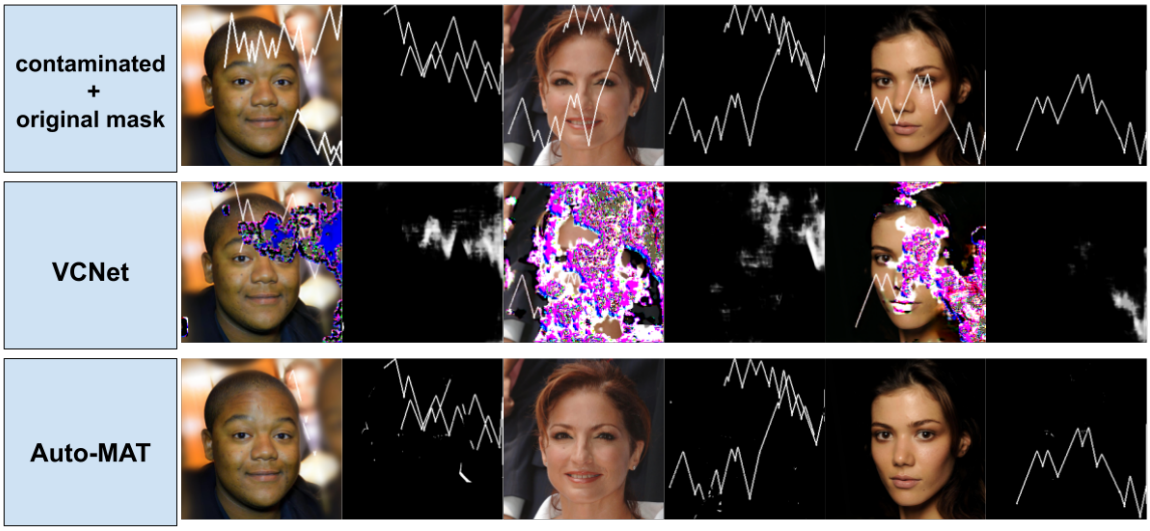

Figure 4. *Some examples to visually compare the performance of the proposed method and VCNet. As can be seen, Auto-MAT can restore the contamination area and replace it with normal pixel intensities using the context of the remaining image.*

Table 1: PSNR, SSIM, and BCE between the ground truth and output images and masks for the proposed method and VCNet.

Method	FFHQ		
	PSNR ↑	SSIM ↑	BCE↓
VCNet	10.490	0.2426	0.7152
Auto-MAT	**18.054**	**0.8679**	**0.6889**
	Places2		
	PSNR ↑	SSIM↑	BCE↓
VCNet	14.9524	0.2500	0.8321
Auto-MAT	**21.6945**	**0.9314**	**0.7184**

Table 2: Assessment for different combinations of pr and t of the proposed algorithm applied on the FFHQ dataset. The best combinations are highlighted in bold numbers.

	t	PSNR↑	SSIM↑	BCE ↓
	2	16.458	0.8693	0.6931
$pr = 0.7$	4	17.080	0.8633	0.7006
	8	19.084	0.8640	0.6959
	2	16.526	0.8689	0.6930
$pr = 0.8$	4	17.7839	0.8651	0.6923
	8	**20.069**	**0.8707**	0.6875
	2	16.800	0.8684	0.6906
$pr = 0.9$	4	18.054	0.8679	0.6889
	8	17.747	0.8690	**0.6873**

and PSNR of 0.69, 0.87, and 17.75, respectively. Figure 1 illustrates several examples of progressively eliminating undesiblack content from the input images utilizing the suggested method.

Raising the threshold t will increase the rejection inpainting probability. This forces the proposed method to keep most of the image unchanged as possible, when fixing the number of inpainting iterations L. However, raising the value of L enhances the frequency of testing a specific pixel for inpainting, thereby elevating the likelihood of inducing a pixel alteration. Additionally, the inpainting probability pr plays a direct role in determining the likelihood of testing mask pixels for inpainting, influencing the detection of out-of-distribution pixels across various sizes. pr determines the number of random pixels tested for inpainting at each inpainting iteration. If pr is high, less image context will be used to inpaint the missing regions, which may cause the normal content to be lost. Conversely, a small pr results in the algorithm taking longer to test all possible pixel combinations. These three hyperparameters are interconnected, requiring a delicate balance to achieve accurate mask estimation while retaining the majority of normal pixels unchanged, thereby preserving the original image content to the greatest extent possible.

Inadvertently removing normal pixels depends on factors like the effectiveness of the inpainting method. Hence, the confidence map (Dc) is introduced to distinguish normal from abrupt pixels, blackucing the chance of removing normal ones mistakenly. The proposed method takes about 5 seconds for 50 inpainting iterations using a modern GPU. In contrast, VCNet completes a single forward pass in approximately 100ms. However, our method outperforms VCNet despite its quicker execution. Utilizing advanced techniques to optimize L, pr, and t values could further enhance processing efficiency.

Conclusions

This study introduces an automated inpainting method that leverages high-performing inpainting models, eliminating the need for manual mask creation—an often challenging task in practical scenarios. The proposed method successfully estimated and eliminated brush strokes in both facial images and natural scenes, replacing these pixels with color intensities that appear normal. The approach employed pixel-wise subtraction to identify varying pixels, enhancing confidence in optimal mask estimation. In alternative experiments, one could explore other deep learning-based comparison methods for a more precise determination of semantically alteblack pixels during the algorithm's progression. Additionally, varying pixel sizes could be introduced to identify different sizes of image corruption. More experiments can be conducted to evaluate and refine the proposed method and to ensure the method's robustness in eliminating other types of image noise. This work can be applied to different anomaly detection applications in industry and medicine when the number of anomalous samples is very small compablack to the number of normal samples [7, 6].

Acknowledgement: This publication was made possible by NPRP13S-0127-200108 from the Qatar National Research Fund (QNRF). QNRF had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The institutional IRB approvals were granted.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, 28(3):24, 2009.

[3] Nian Cai, Zhenghang Su, Zhineng Lin, Han Wang, Zhijing Yang, and Bingo Wing-Kuen Ling. Blind inpainting using the fully convolutional neural network. *The Visual Computer*, 33:249–261, 2017.

[4] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004.

[5] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012.

[6] Abdullah Hayajneh, Erchin Serpedin, Mohammad Shaqfeh, Graeme Glass, and Mitchell A Stotland. Cleftgan: Adapting a style-based generative adversarial network to create images depicting cleft lip deformity. *arXiv preprint arXiv:2310.07969*, 2023.

[7] Abdullah Hayajneh, Mohammad Shaqfeh, Erchin Serpedin, and Mitchell A Stotland. Unsupervised anomaly appraisal of cleft faces using a stylegan2-based model adaptation technique. *Plos One*, 18(8):e0288228, 2023.

[8] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[9] Jiaya Jia and Chi-Keung Tang. Image repairing: Robust image synthesis by adaptive and tensor voting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, Proceedings*, volume 1, pages I–I. IEEE, 2003.

[10] Johannes Kopf, Wolf Kienzle, Steven Drucker, and Sing Bing Kang. Quality pblackiction for image completion. *ACM Transactions on Graphics (ToG)*, 31(6):1–8, 2012.

[11] Levin and Zomet. Learning how to inpaint from global image statistics. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 305–312. IEEE, 2003.

[12] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10758–10768, 2022.

[13] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3911–3919, 2017.

[14] Guilin Liu, Fitsum A blacka, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[15] Yang Liu, Jinshan Pan, and Zhixun Su. Deep blind image inpainting. In *Intelligence Science and Big Data Engineering. Visual Data Engineering: 9th International Conference, IScIDE 2019, Nanjing, China, October 17–20, 2019, Proceedings, Part I 9*, pages 128–141. Springer, 2019.

[16] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[17] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

[18] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[19] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2482–2491, 2018.

[20] Min-cheol Sagong, Yong-goo Shin, Seung-wook Kim, Seung Park, and Sung-jea Ko. Pepsi: Fast image inpainting with parallel decoding network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11360–11368, 2019.

[21] Jian Sun, Lu Yuan, Jiaya Jia, and Heung-Yeung Shum. Image completion with structure propagation. In *ACM SIGGRAPH 2005 Papers*, pages 861–868. 2005.

[22] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4692–4701, 2021.

[23] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[24] Yi Wang, Ying-Cong Chen, Xin Tao, and Jiaya Jia. Vcnet: A robust approach to blind image inpainting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 752–768. Springer, 2020.

[25] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[26] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8858–8867, 2019.

[27] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5840–5848, 2019.

[28] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6721–6729, 2017.

[29] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.

[30] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019.

[31] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 69–78, 2021.

[32] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019.

[33] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.

[34] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Tfill: Image completion via a transformer-based architecture. *arXiv preprint arXiv:2104.00845*, 2(3):6, 2021.