Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19102-1

1232: HUMAN-CENTRIC MULTIMEDIA ANALYSIS

Image de-photobombing benchmark

Vatsa S. Patel[1] · Kunal Agrawal[1] · Samah S. Baraheem[1] · Amira Yousif[1] · Tam V. Nguyen[1]

Received: 11 May 2023 / Revised: 22 February 2024 / Accepted: 27 March 2024

Abstract
Removing photobombing elements from images is a challenging task that requires sophisticated image inpainting techniques. Despite the availability of various methods, their effectiveness depends on the complexity of the image and the nature of the distracting element. To address this issue, we conducted a benchmark study to evaluate 10 state-of-the-art photobombing removal methods on a dataset of over 300 images. Our study focused on identifying the most effective image inpainting techniques for removing unwanted regions from images. We annotated the photobombed regions that require removal and evaluated the performance of each method using peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and Fréchet inception distance (FID). The results show that image inpainting techniques can effectively remove photobombing elements, but more robust and accurate methods are needed to handle various image complexities. Our benchmarking study provides a valuable resource for researchers and practitioners to select the most suitable method for their specific photobombing removal task.

Keywords Photobombing removal · Image de-photobombing · Image inpainting · Benchmark · Metrics · Deep learning

1 Introduction

"A photograph is not just an image, it is a moment frozen in time." - Anonymous

Indeed, photos freeze for a moment in time whereas photobombs make the unseen seen. In today's world, photobombing has become a more common occurrence, with people often doing it intentionally for fun or comedic effect. However, this does not negate the importance of preserving the intended composition of a photograph, especially when it holds deep personal or emotional significance. In conclusion, removing photobombing from a photograph is crucial to ensure that the intended message and emotion of the image are not lost or compromised. In today's digital age, with smartphones and social media,

✉ Tam V. Nguyen
tamnguyen@udayton.edu

[1] Department of Computer Science, University of Dayton, Dayton, USA

photography has become an integral part of our lives. It allows us to capture and share our experiences with the world, connecting us to people and places in once unimaginable ways. However, the rise of photobombing has introduced a new challenge to this process, threatening to spoil even our most treasured memories. With the increasing popularity of group selfies, for example, it's become easier for unwanted elements to creep into the frame, making it difficult to capture the perfect shot. As such, it's important to be mindful of our surroundings when taking photos and to respect the privacy and wishes of others. While photobombing can be amusing in some instances, it can also be a source of frustration and disappointment for those trying to create lasting memories. By being considerate and thoughtful in our photography, we can ensure that our memories remain beautiful and untainted.

It may come as a surprise, but photobombing dates back to 1853 [1], when the first known incident occurred. During a photograph, an unidentified person unexpectedly inserted themselves into the frame, despite not being part of the intended shot. Photographs have always been a means of capturing significant moments and preserving them for posterity. They possess the ability to capture fleeting moments and evoke emotions. Some moments are irreplaceable and can only be remembered through the aid of a photograph. However, if a photobomber spoils an image, it becomes crucial to remove them from the picture, so that the memory can be relieved without any hindrance.

Image inpainting is a powerful technique used to replace unwanted portions of an image with visually plausible content. Traditionally, this task requires expertise and consumes considerable time. However, modern image inpainting algorithms can be employed to achieve the same result with greater efficiency. For example, when removing photobombs from an image, tools such as Adobe Photoshop can be used. Alternatively, sophisticated image inpainting algorithms can be utilized to transform the original image and remove the undesired region, resulting in a more aesthetically pleasing output. Figure 1 provides a clear example of such an image inpainting. In this paper, we have explored the results of image inpainting algorithms, which to the best of our knowledge, are the first ones to address this interesting yet challenging problem of benchmarking the images with the undesired region removed. Our methodology begins with collecting photobombed images from various online sources, followed by generating masks by annotating the unwanted areas in the collected images. The generated masks are then used as inputs to image inpainting algorithms, together with the photobombed images. Finally, the results are compared with the ground truth images edited by a professional photoshopper. By benchmarking the performance of image inpainting algorithms, we aim to provide a deeper understanding of their strengths

Fig. 1 Input and output of photobombing removal process. From left to right: **a** the original photobombed image, **b** the mask of the unwanted regions, **c** the photobombing removal image

and limitations. Ultimately, this research can help identify the most suitable algorithms for specific image inpainting tasks, making the process more efficient and accessible for a broader range of users. Finally, we use Frechet Inception Distance (*FID*) [2], Structural Similarity Index (*SSIM*) [3], and Peak to Noise Signal Ratio (*PSNR*) [4, 5] in the evaluation stage.

Our earlier version was published in the ISVC 2022 conference [37]. In this journal version, our contributions are threefold.

- We further doubled the benchmark data to 300 images with more semantic classes.
- We investigate more state-of-the-art image inpainting methods for the benchmark.
- We provide in-depth analysis for the benchmark.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the related work. Section 3 presents the DPD-300 dataset and methods used in our benchmarking study. In Section 4, we present the extensive evaluation results obtained from the benchmark. Finally, in Section 5, we conclude our paper and outline future research directions.

2 Related work

Image inpainting is an essential image processing task that involves the restoration of damaged or missing regions in an image. Over the years, various approaches have been proposed to tackle this problem, ranging from classical fluid dynamics concepts to texture synthesis and depth-based algorithms. The effectiveness and efficiency of these methods vary, and the choice of the method depends on specific application requirements and available resources.

2.1 Traditional methods

Various approaches to image inpainting have been proposed in the literature. Bertalmio et al. [10] introduced an automated technique based on classical fluid dynamics concepts. Their method propagates isophote lines into the area to be painted, aiming to maintain image continuity by matching gradient vectors at the inpainting region's edge. Although effective, this approach can be computationally intensive. Criminisi et al. [6] presented a best-first technique that combines texture synthesis and inpainting to propagate texture and structure information efficiently. This method achieves computational efficiency by using a block-based sampling process and propagating confidence in synthetic pixel values in a way comparable to information propagation in inpainting. The approach is accurate and efficient, producing high-quality results. Seyedsaeid et al. [29] introduced Depth-Wise Image Inpainting that combines exemplar-based and depth-based algorithms to fill holes in digital images. It uses depth information and a database of multi-views to find the order of objects in the target image and the best patches to fill the holes. The method is capable of filling holes with multiple objects in a proper order. However, it is limited to replacing only marked areas with proper patches and may lead to inaccurate object retrieval and alignment on smooth surfaces. Overall, these approaches offer different strategies for achieving image inpainting with varying degrees of computational efficiency and accuracy. The choice of method will depend on the specific application requirements and available resources, such

as computational power and time constraints. Bornemann et al. [8] propose a non-iterative method based on the analysis of stable first-order transport equations. Their approach uses the fast-marching method to traverse the inpainting domain, conveying picture data in a coherent direction securely calculated by the structure tensor. The approach alternates between diffusion and directional transport based on a measure of coherence strength, leading to fast and accurate results. Telea [9] presents a fast-matching method-based algorithm that uses a mathematical boundary model to inpaint missing sections, followed by the fast-marching method. This approach achieves accurate and efficient results but is limited to smooth surfaces and may lead to inaccurate object retrieval and alignment.

2.2 Deep learning-based methods

The field of computer vision has witnessed tremendous advancements in recent years, with a particular focus on practical applications such as image inpainting and automatic photo cropping. These tasks have become increasingly important in fields such as photography, image editing, and video processing. Jiahui et al. [11, 12] propose a novel approach for generative image inpainting, which involves filling in missing or damaged areas of an image. The authors use gated convolutions to learn dynamic feature selection mechanisms and can handle free-form masks. Traditional GANs for rectangular masks do not work well for irregular masks, which led the authors to introduce SN-PatchGAN, a patch-based GAN loss function that significantly improves image inpainting quality. The SN-PatchGAN method uses a spectral-normalized discriminator on dense image patches, allowing it to complete images with complex and irregular masks. In addition to image inpainting, Zhou et al. [30] present an automatic photo cropping system that aims to produce visually pleasing images by removing distractions and focusing on the main subject. The framework combines aesthetic cues and learned representations from a convolutional autoencoder and PCA. This approach outperforms existing automatic cropping methods and provides a significant improvement over previous technique. The system has the potential to revolutionize the way images are processed and edited, and it demonstrates the power of deep learning in computer vision applications. Together, the approaches proposed by Jiahui et al. and Zhou et al. represent significant advances in the field of computer vision. They showcase the potential of deep learning techniques and demonstrate the importance of developing new methods to address practical challenges in image processing. In this paper Hassanpour et al. [31] introduces E2F-GAN, a GAN-based model for reconstructing faces using the periocular region. It includes a coarse and refinement module and a dataset called E2Fdb was used for evaluation. Results show that E2F-GAN outperforms previous methods and generates realistic images while preserving identity. The implementation of the proposed approach is available on GitHub for reproducible research. As the demand for high-quality images continues to grow, research in this area is likely to remain a high priority for computer vision researchers in the coming years.

3 Image de-photobombing benchmark

3.1 DPD-300 dataset

Photobombed images, logical binary masks, and ground truth images are required to carry out a thorough benchmark. From a variety of online sources, including Facebook [14],

Gettyimages [15], Bored Panda [16], Adobe Stock [17], Shutterstock [18], and Pinterest [19], we carefully and manually collect photobombed images. In total, 300 photobombed images are collected for the De-Photobombing Dataset dubbed **DPD-300**. For each photobombed image in our dataset, the photobombed regions are annotated through the Freehand object of the "Region of Interest'' function present in Matlab [20]. This step is very important to generate the binary logical masks. A single photobombed image is iterated based on how many photobombed elements are present in the image until all elements in the image are masked. The result of this step is a logical binary image. Depending on how many photobombed elements should be removed from a single image, the average time needed was about 50 to 60 s to annotate a single image. Following this, the photobombed images along with their corresponding logical masks are fed as inputs into different inpainting methods to remove the unwanted and distracting items.

In our DPD-300 dataset, one of the main challenges is acquiring ground truth images for comparing various algorithms. To achieve this, we hired a skilled professional photo editor to remove the undesirable or photobombed area(s) from each collected photobombed image. The altered images are considered as the ground truth in our benchmark.

To show the difference between our previous work [37] and our current extension (this work) in terms of the collected dataset, we illustrate the word clouds of DPD-150 and DPD-300, respectively in Fig. 2. We further show an example of both datasets in Fig. 3, where DPD-150 contains only humans as photobombed elements; however, DPD-300 consists of humans and/or objects as photobombed elements.

3.2 Benchmarking methods

The photobombed image and its corresponding logical mask are used as inputs for various inpainting approaches. Figure 4 shows an example of the inputs used by inpainting methods. The outcome is an inpainted image.

(a)

Fig. 2 Word clouds of the dataset. **a** ISVC 2022 benchmark, while (**b**) DPD-300 (this work)

(b)

Fig. 2 (continued)

Fig. 3 The dataset comparison between (left) DPD-150 [37] which only contains humans as photobombed elements, and (right) DPD-300 which contains humans and/or objects as photobombed elements

For benchmarking, various image inpainting approaches are utilized. More specifically, Exemplar-Based Image Inpainting (EBII) [6, 7], Coherence Transport (CT) [8], Fast Marching (FM) [9], Fluid Dynamics (FD) [10], DeepFill [34], High Resolution Inpainting using GAN (HiFill) [36], Region Normalization (RN) [35], Tfill – coarse [32], and Tfill – refined [32] are leveraged to decide which approach is best. Figure 5 shows an overview of de-photobombing incorporating different image inpainting techniques. The photobombed image and its respective logical mask are input into different image inpainting techniques, as seen in the figure, to eliminate the undesired and distracting parts and

Fig. 4 Sample inputs to image inpainting approaches. **a** illustrates the photobombed images, while (**b**) shows the respective logical masks of distracting and undesirable regions

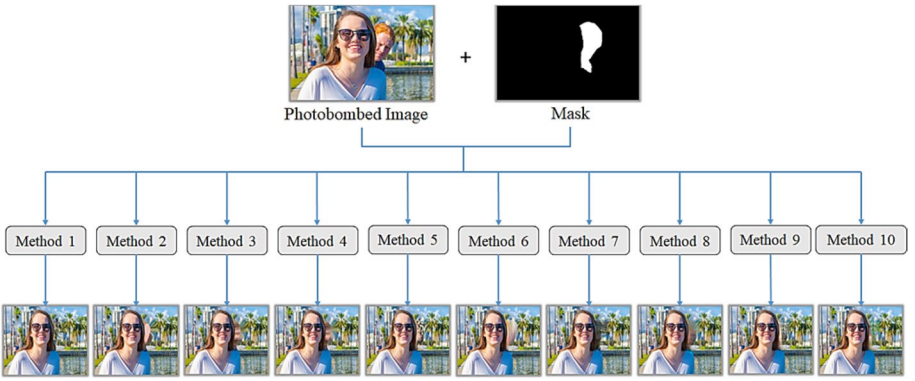

Fig. 5 The general overview of our proposed image de-photobombing benchmark that leverages various image inpainting methods

reconstruct the image. Particularly, all inpainting methods' source code or executable files are obtained from already-existing sources [11–13, 21–26]. In order to generate the de-photobombing images, we finally ran all the ten aforementioned methods separately on the testing set. The outcomes are compared to the ground truth using several metrics, which will be introduced in more detail in the next section.

4 Experiments

4.1 Evaluation Metrics

The photobombing removal images are compared to the ground truth images, produced by the invited professional photoshoppers, to benchmark the results of all inpainting approaches. To evaluate the effectiveness of photobombing removal images, several performance metrics are used. Specifically, Fréchet inception distance (FID) [2], Structural Similarity Index (SSIM) [3], and Peak to Noise Signal Ratio (PSNR) [4, 5] are used to compare the reconstructed images to the ground truth.

The first evaluation metric is FID [2]. FID [2] calculates the difference between the ground truth distribution and the reconstructed image distribution depending on the extracted features, as in (1).

$$FID(x,y) = d^2 = ||mu_x - mu_y||^2 + Tr\left(c_x + c_y - 2*\right)\sqrt{c_x * c_y} \tag{1}$$

where mu_x and mu_y refer to the feature-wise mean of the ground truth and reconstructed image, and c_x and c_y indicate the covariance matrix of the feature vectors of ground truth and reconstructed image. Tr is the trace linear algebra operation. The second evaluation metric is SSIM [3]. It computes the similarity between the ground truth and reconstructed image through extracting three primary features from the provided images which are luminance, contrast, and structure. These three features are taken into account while comparing the two images. SSIM is calculated as follows:

$$SSIM(x,y) = [l(x,y)]^{\alpha}.[c(x,y)]^{\beta}.[s(x,y)]^{\gamma} \tag{2}$$

where x and y are the ground truth and the respective reconstructed image. l, c, and s are the luminance, contrast, and structure, respectively. α, β, and γ indicate the related significance of each metric. In the meantime, PSNR [4, 5] measures the ratio between the signal's maximum possible power and the distorting noise power. The mathematical illustration is represented in Eq. (3).

$$PSNR = 20log_{10}\left(\frac{MAX_f}{\sqrt{MSE}}\right) \tag{3}$$

where Mean Squared Error (MSE) is computed as in Eq. (4).

$$MSE = \frac{1}{mn}\sum_{0}^{m-1}\sum_{0}^{n-1}||f(i,j) - g(i,j)||^2 \tag{4}$$

, where f and g refer to the data matrix of the ground truth and the respective reconstructed image, respectively. While m is the number of pixels in the image's rows, and i is the row's index, n is the number of pixels in the image's columns, and j indicates the column's index.

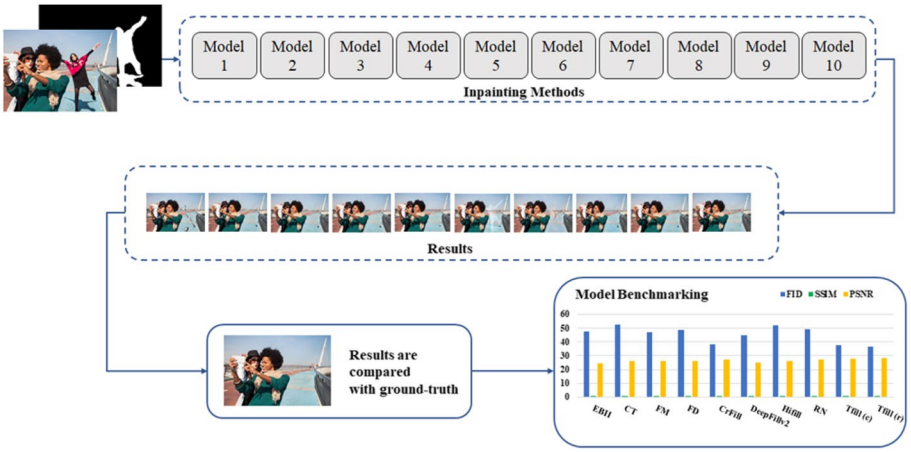

Fig. 6 The flowchart of the image de-photobombing benchmark in terms of FID [2], SSIM [3, 28], and PSNR [4, 5]

Table 1 The performance of various image inpainting approaches in our benchmark with regards to FID [2], SSIM [3, 28], and PSNR [4, 5]

Benchmarking Methods	FID ↓	SSIM ↑	PSNR ↑
EBII [6, 7]	47.60	0.923	24.57
CT [8]	52.36	0.936	25.85
FM [9]	46.96	0.933	26.22
FD [10]	48.77	0.933	26.07
CrFill [33]	38.16	0.920	26.88
DeepFill v2 [34]	44.92	0.915	24.92
HiFill [36]	52.10	0.915	25.99
RN [35]	48.99	0.915	26.99
Tfill – coarse [32]	37.68	0.930	27.75
Tfill – refined [32]	**36.48**	**0.939**	**28.09**

The best results are marked with boldface

4.2 Experimental Results

Several metrics are computed to assess how well various inpainting approaches remove the photobombed regions while maintaining the overall quality. Figure 6, Tables 1 and 2 show the performance of different approaches on our benchmark.

As can be seen from Table 1, Tfill – refined [32] achieves the best performance with 36.48 and 28.09 for FID [2] and PSNR [4, 5], respectively. Also, Tfill – refined [32] achieves the best performance in terms of SSIM [3]. This indicates that while Tfill is good at producing images with high pixel-level similarity to the original images, it is also effective in preserving the original structure of the images. Note that SSIM measures the structural similarity between two images by comparing their luminance, contrast, and structure. Therefore, the methods that perform better in SSIM tend to preserve the edges, textures, and other structural features of the original image.

Meanwhile, CT [8] achieves the second-best SSIM score with 0.9364, indicating that the generated images are visually similar to the original images. CT [8] can preserve the

Table 2 Performance evaluation of image inpainting methods on Human and Non-Human images using FID [2], SSIM [3, 28], and PSNR [4, 5] scores

Benchmarking Methods	Human			Non—Human		
	FID ↓	SSIM ↑	PSNR ↑	FID ↓	SSIM ↑	PSNR ↑
EBII [6, 7]	75.61	0.854	19.82	120.14	0.895	21.99
CT [8]	73.44	0.904	20.93	**112.46**	0.916	22.46
FM [9]	70.63	0.910	21.02	118.17	0.913	22.34
FD [10]	69.11	0.910	21.16	114.16	0.913	22.44
CrFill [33]	70.68	0.895	20.99	117.46	0.907	22.40
DeepFill v2 [34]	126.85	0.790	18.76	189.71	0.626	17.93
HiFill [36]	421.27	0.193	10.03	391.08	0.190	10.70
RN [35]	73.96	0.864	20.68	122.64	0.908	22.45
Tfill – coarse [32]	**64.26**	0.913	21.34	117.55	0.926	22.69
Tfill – refined [32]	64.63	**0.916**	**21.59**	115.24	**0.930**	**23.03**

The best results are marked with boldface

structural similarity between the reconstructed image and the original image to a greater extent than the other methods evaluated in the benchmark. However, it obtains the worst FID score among the other methods, which suggests that the generated images have significant differences in terms of image distribution compared to the original images. This is because CT [8] focuses on maintaining the global structure of the image and may not generate accurate texture information. On the other hand, HiFill [36] has the worst SSIM score, which may be due to its hierarchical inpainting process that can lead to inconsistencies and artifacts in the generated images. Meanwhile, EBII [6, 7] has the worst PSNR score among the other methods, which may be because it relies on a deep generative model to fill in the missing regions but may not be as effective in preserving the high-frequency details of the image.

Furthermore, we analyze the performance for human and non-human photobombing categories. As shown in Table 2, for human-related photobombing images, CT [8], FM [9], FD [10], and CrFill [33] demonstrate relatively competitive results, with FID scores ranging from 69.11 to 73.44, indicating decent similarity between generated and ground truth images. Notably, Tfill–refined [32] stands out with the lowest FID score of 64.63, suggesting superior performance in maintaining similarity with original human images. Regarding SSIM, Tfill–refined achieves the highest score of 0.916, indicating strong structural similarity, closely followed by Tfill–coarse [32] and FD [10]. Additionally, in terms of PSNR, Tfill–refined leads with a score of 21.59, closely followed by Tfill–coarse [32], FD [10], and CT [8]. Conversely, HiFill [36] exhibits notably poor performance across all metrics for human images, with extremely high FID, low SSIM, and PSNR scores, indicating significant challenges in accurately inpainting human images.

In contrast, when considering non-human category images, similar trends emerge with Tfill–refined [32] showcasing the most promising performance across all metrics. It achieves the lowest FID score of 115.24, indicating strong similarity with the original non-human images. Furthermore, Tfill–refined [32] also achieves the highest SSIM (0.930) and PSNR (23.03) scores among all methods, signifying its effectiveness in preserving structural details and achieving high-fidelity inpainted non-human images. Notably, other methods such as Tfill – coarse [32], FD [10], and CrFill [33] also demonstrate competitive

performance with relatively low FID scores and high SSIM and PSNR scores. On the other hand, HiFill [36] again exhibits the poorest performance among all methods, underscoring its challenges in effectively inpainting non-human images. Overall, the evaluation underscores the importance of methodological advancements, particularly demonstrated by Tfill–refined [32], in achieving superior inpainting results for both human and non-human category images.

The de-photobombing results of various inpainting methods used in this paper are visualized in Fig. 7. In this paper, we sought to evaluate the performance of various inpainting methods with respect to the percentage of the region to be painted. The results, shown in Fig. 8, provide insights into which methods are most effective for different percentages of missing data. Figure 8(i) displays the FID scores of the different inpainting algorithms at different mask percentages. Notably, all algorithms performed better when the percentage of missing data was between 0 and 10, as evidenced by the smaller FID values. The trend observed in Fig. 8(i) is that the FID score is directly proportional to the percentage of missing data. As the percentage increases, the FID score generally increases as well, indicating that the inpainting methods struggle more to generate realistic samples with higher amounts of missing data. Figure 8(ii) shows the performance of the inpainting methods in terms of the SSIM score at different mask

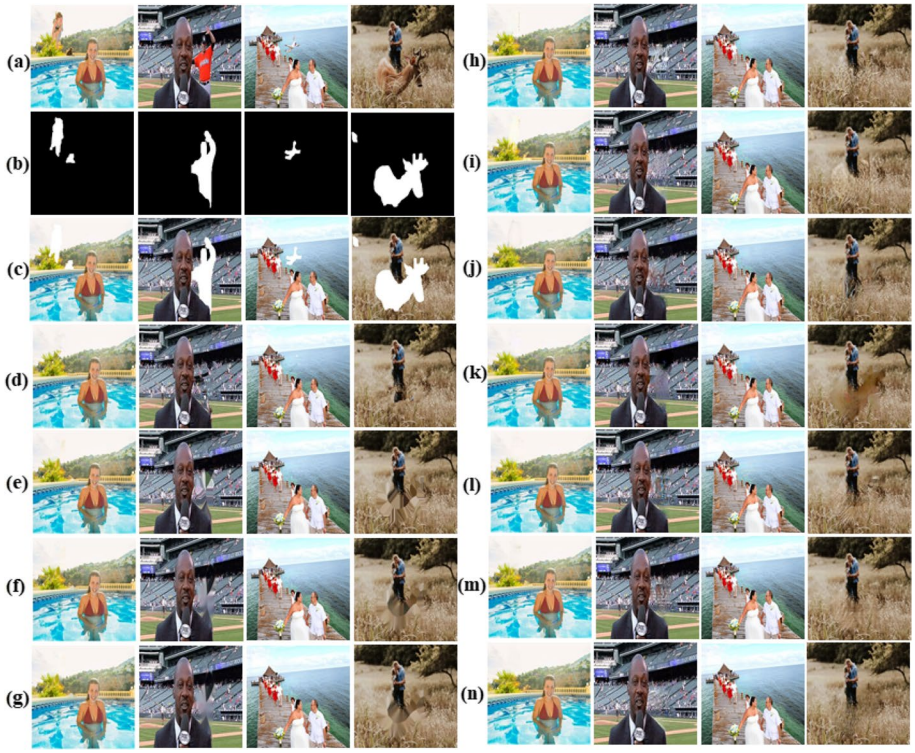

Fig. 7 Samples of various inpainting techniques' outputs in our proposed framework to remove distracting and unwanted regions. **a** Photobombed images, **b** Annotated masks, **c** Image with mask **d** EBII [6, 7], **e** CT [8], **f** FM [9], **g** FD [10], **h** CrFill [33], **i** DeepFill v2 [34], **j** HiFill [36], **k** RN [35], **l** Tfill – coarse [32], **m** Tfill – refined [32] and finally **n** the ground truth images. Please see the color pdf with 400% zoom

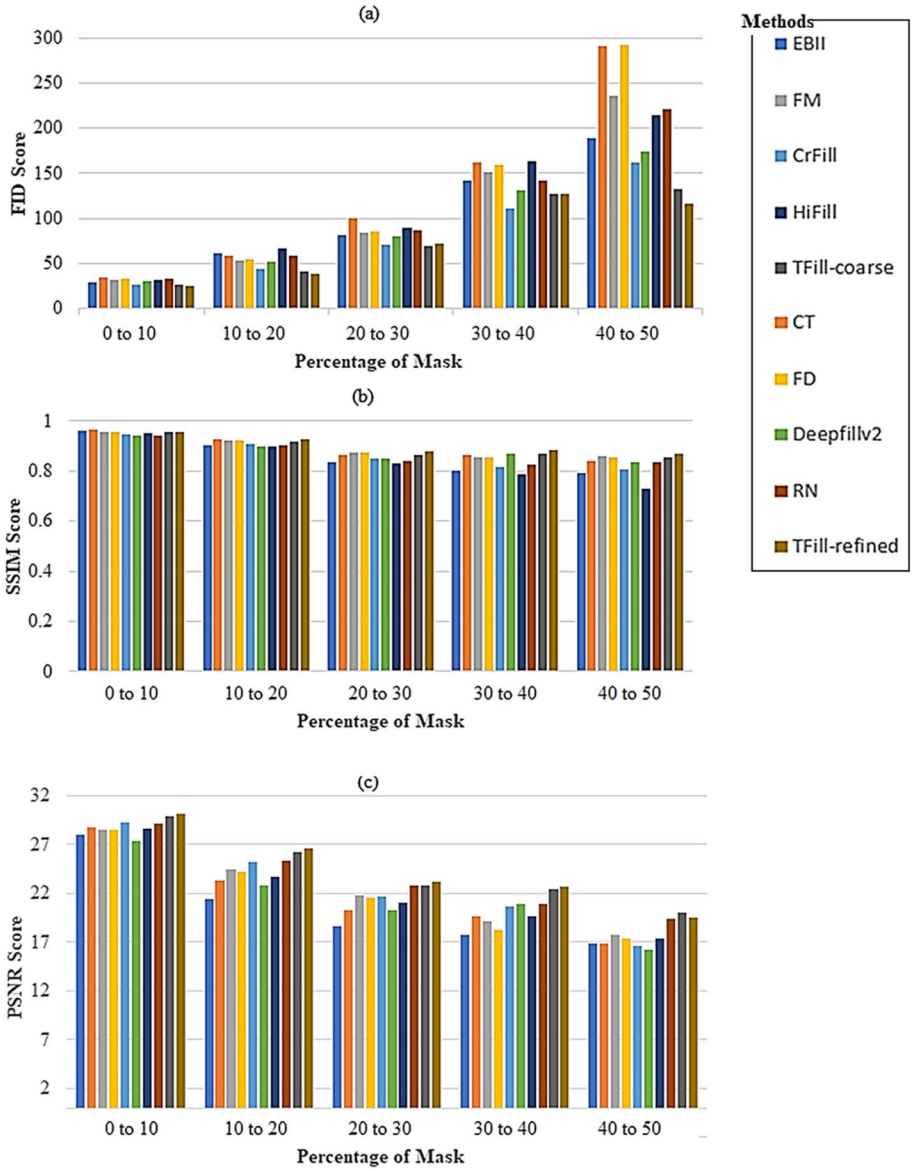

Fig. 8 Performance of various inpainting techniques with the percentage of mask utilized for inpainting. **a** FID score [2] with the percentage of masks, **b** SSIM score [3, 28] with the percentage of masks, and **c** PSNR score [4, 5] with the percentage of masks

percentages. The trend observed in Fig. 8(ii) is that the SSIM score is inversely proportional to the percentage of missing data. As the percentage increases, the SSIM score decreases, indicating that the inpainting methods generate samples that are less similar to the ground truth with higher amounts of missing data. Finally, Fig. 8(iii) displays the PSNR scores of the different inpainting algorithms at different mask percentages.

The trend observed in Fig. 8(iii) is that the PSNR score is inversely proportional to the percentage of missing data, with higher percentages of missing data resulting in lower PSNR scores. These insights can inform researchers and practitioners on which inpainting methods to use depending on the specific context of their missing data problem.

Figure 9 demonstrates some failure cases of various inpainting approaches in the framework of eliminating unwanted regions from images. In particular, the images in (a) (i) and (a) (ii) show examples where all methods failed to remove the photobombing regions perfectly. In (a) (i), all methods used the subject to generate the masked region due to the high percentage of subject presence in the image and its adjacency to the masked region. However, this resulted in incomplete removal of the unwanted region. Furthermore, in (a) (iii), all methods showed poor performance in generating the texture of the letters in the image, indicating a limitation of these methods in handling complex textures. The results suggest that while existing inpainting methods have shown impressive performance in many cases, they still face challenges when it comes to handling complex images and textures.

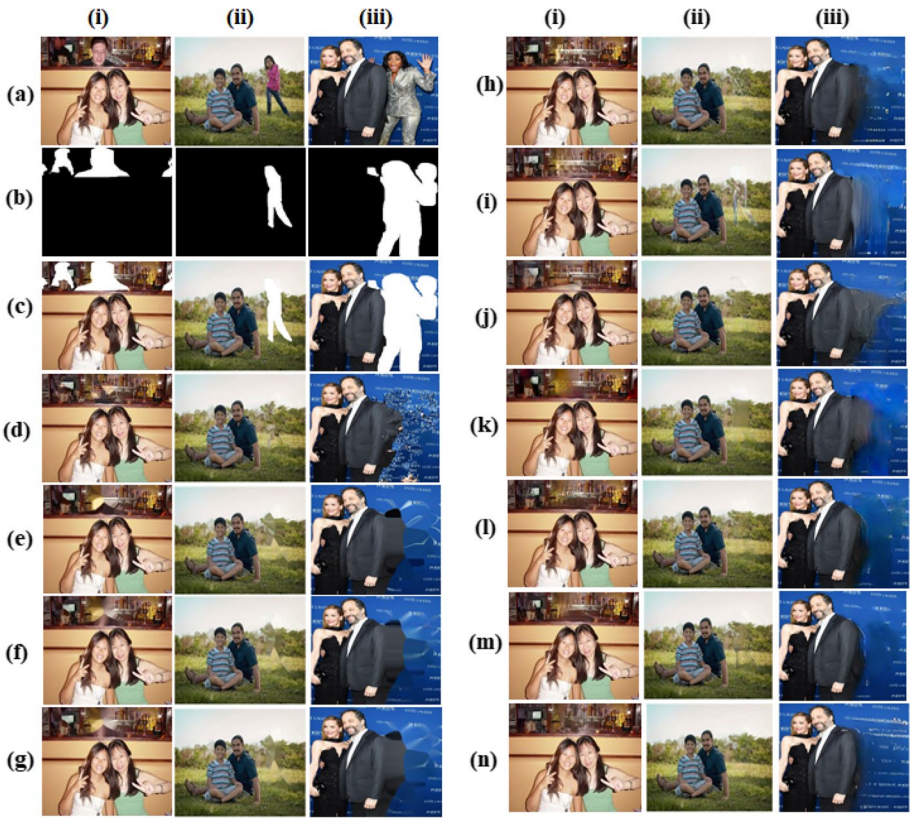

Fig. 9 An example of failure cases of various inpainting approaches in our benchmark to eliminate unwanted regions. **a** Photobombed images, **b** Annotated masks, **c** Images with mask **d** EBII [6, 7], **e** CT [8], **f** FM [9], **g** FD [10], **h** CrFill [33], **i** DeepFill v2 [34], **j** HiFill [36], **k** RN [35], **l** Tfill – coarse [32], **m** Tfill – refined [32] and finally **n** the ground truth images. Please see the color pdf with 400% zoom

Indeed, the failure cases presented in Fig. 9 highlight the need for further research and development of more robust inpainting methods that can handle various image complexities and achieve more accurate and complete removal of unwanted regions from images.

5 Conclusion and future works

Image De-Photobombing is a challenging task that involves removing undesired items or individuals from images. The procedure commences by identifying the areas within the images that have been photobombed, whether by humans or non-humans, and then manually creating annotations in the form of masks for these unwanted regions. To eliminate these regions and reconstruct the images, various image inpainting techniques are employed. The results are then evaluated using different metrics to compare them with the corresponding ground truth images, which are produced by manually removing the unwanted regions. Creating the DPD-300 dataset, which includes photobombed images, logical binary masks, and ground truth images, was a difficult and time-consuming task. This is because removing distracting and unwanted regions from images is a tedious and painful process that requires a lot of effort. As far as we know, this dataset is unique and not available elsewhere. It can be used to evaluate and improve the effectiveness and accuracy of different image inpainting approaches.

To date, we have conducted extensive experiments using various image inpainting techniques on the DPD-300 dataset. The results have shown that these techniques are effective in removing photobombing and producing aesthetically pleasing images. In addition, the experiment with human/non-human categories will shed light on the future improvement of algorithms for photobombing removal. In the future, we plan to investigate additional image inpainting techniques to enhance the benchmarking of the dataset. We would also like to explore one possible approach to address the difficulty and time consumption in creating the dataset for photobombing removal is to explore alternative methods of collecting image pairs. Further research could investigate the feasibility of this approach and evaluate its effectiveness compared to the manual image recovery process. Additionally, exploring alternative methods for masking photobombed regions could also be a potential avenue for future work. The DPD-300 dataset will be publicly available along with this publication.

Funding This research was supported by the National Science Foundation (NSF) under Grant 2025234.

Data availability Data will be made available on reasonable request.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Edwards P (2015) This 1853 image might show the first photobomb, Vox. [Online]. Available: https://www.vox.com/2015/9/25/93977-33/first-photobomb. Accessed 12 Mar 2023]
2. Keeling M (2022) Frechet Inception Distance," MATLAB Central File Exchange. [Online]. Available: https://www.mathworks.com/matlabcentral/file-exchange/79071-frechet-inception-distance. Accessed 12 Mar 2023
3. Zhou W, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
4. Hore A, Ziou D (2010) Image quality metrics: PSNR vs. SSIM. In: 2010 20th international conference on pattern recognition. IEEE, pp 2366–2369
5. Fardo FA, Conforto VH, de Oliveira FC, Rodrigues PS (2009) A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms. IEEE Trans Image Process 18(5):969–976
6. Criminisi A, Perez P, Toyama K (2004) Region filling and object removal by exemplar-based image inpainting. IEEE Trans Image Process 13(9):1200–1212
7. Le Meur O, Ebdelli M, Guillemot C (2013) Hierarchical super-resolution-based-inpainting. IEEE Trans Image Process 22(10):3779–3790
8. Bornemann F, März T (2007) Fast image inpainting based on coherence transport. J Math Imaging Vis 28:259–278
9. Telea A (2004) An image inpainting technique based on the fast marching method. J Graph Tools 9(1):23–34
10. Bertalmio M, Bertozzi AL, Sapiro G (2001) Navier-stokes, fluid dynamics, and image and video inpainting. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. IEEE. CVPR 2001, vol 1, pp I–I
11. TFill: https://github.com/lyndonzheng/TFill. Accessed 19 Jan 2023
12. CR-Fill: https://github.com/zengxianyu/crfill. Accessed 19 Jan 2023
13. DeepFillv2: https://github.com/csqiangwen/DeepFillv2_Pytorch. Accessed 19 Jan 2023
14. Fridman J. Facebook Page, https://www.facebook.com/jamesfridmanpage/photos. Last retrieved on March 1 2023
15. Getty Images, Photobombing Images, https://www.gettyimages.com/photos/photobombing. Last retrieved on March 1 2023
16. Borepanda, Photobomb Images, https://www.boredpanda.com/?s=photobomb&utm_source=google&utm_medium=organic&utm_campaign=organic. Last retrieved on March 1, 2023
17. Adobe stock, photobomb, https://stock.adobe.com/search?k=photobomb. Last retrieved on March 1, 2023
18. Shutterstock, photobombing, https://www.shutterstock.com/search/photob-ombing. Last retrieved on March 1, 2023
19. Pinterest, photobomb https://www.pinterest.com/agasca11/photobomb/. Last retrieved on March 1, 2023
20. Region of Interest (ROI) Creation Overview. MATLAB documentation. [Online]. Available: https://www.mathworks.com/help/images/roi-creation-overview.html. Accessed 1 Mar 2023
21. MathWorks. inpaintExemplar function. MATLAB Documentation, https://www.mathworks.com/help/images/ref/inpaintexemplar.html. Last retrieved on Mar 1, 2023
22. Elad M (2019) Inpaint coherent, Version 1.1.1", MathWorks, Natick, MA. [Online]. Available: https://www.mathworks.com/help/images/ref/inpain-tcoherent.html. Accessed 1 Mar 2023
23. OpenCV. Image inpainting, OpenCV documentation, [Online]. Available: https://docs.opencv.org/3.4/df/d3d/tutorial_py_inpainting.html. Accessed 1 March 2023
24. RN: https://github.com/geekyutao/RN. Accessed 19 Jan 2023
25. 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
26. HiFill: https://github.com/Atlas200dk/sample-imageinpainting-HiFill. Accessed 19 Jan 2023
27. Peak signal-to-noise ratio (PSNR). MathWorks, [Online]. Available: https://www.mathworks.com/help/images/ref/psnr.html. Accessed 1 Mar 2023
28. Structural similarity (SSIM) index function, MathWorks, [Online]. Available: https://www.mathworks.com/help/images/ref/ssim.html.Accessed. Accessed 1 Mar 2023
29. Seyedsaeid M, Nagabhushan P (2015) Object removal by depth-wise image inpainting. SIViP 9:427–436

30. Shan N, Tan DS, Denekew MS, Chen Y-Y, Cheng W-H, Hua K-L (2020) Photobomb defusal expert: automatically remove distracting people from photos. IEEE Trans Emerg Top Comput Intell 4(5):717–727
31. Hassanpour A et al (2022) E2F-GAN: Eyes-to-face inpainting via edge-aware coarse-to-fine GANs. IEEE Access 10:32406–32417
32. Zheng C, Cham T-J, Cai J, Phung D (2021) Bridging global context interactions for high-fidelity image completion. arXiv [cs.CV], pp 11512–11522
33. Zeng YY, Lin Z, Lu H, Patel VM (2021) CR-fill: Generative image inpainting with auxiliary contexutal reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 14164–14173
34. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang T (2019) Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4471–4480
35. Yu T, Guo Z, Jin X, Wu S, Chen Z, Li W, Zhang Z, Liu S (2020) Region normalization forimage inpainting. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, no 7, pp 12733–12740
36. 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
37. Prakya SPK, Sainath MMV, Patel VS, Baraheem SS, Nguyen TV (2022) Photobombing removal benchmarking. In: International Symposium on Visual Computing. Springer Nature, Cham, pp 55–66
38. Flickr, "Photobomb," https://www.flickr.com/search/?text=photobomb. Accessed on 12 Mar 2023