Perceptual Artifacts Localization for Inpainting

Lingzhi Zhang[1], Yuqian Zhou[2], Connelly Barnes[2], Sohrab Amirghodsi[2], Zhe Lin[2], Eli Shechtman[2], and Jianbo Shi[1]

[1] University of Pennsylvania
[2] Adobe Research

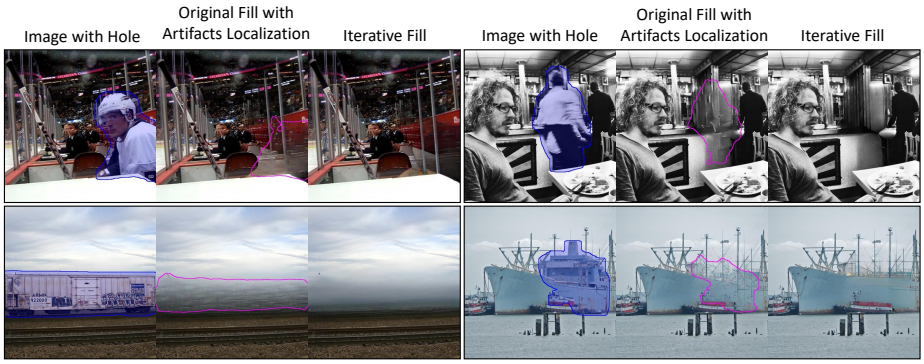

Fig. 1. Visual examples to show that our segmentation network can reliably localize the perceptual artifacts region, as indicated by the pink boundary in the second columns. Given the artifacts localization, we enable the inpainting model [40] iteratively fill on the artifacts region to obtain better inpainting quality, as shown in the third columns.

Abstract. Image inpainting is an essential task for multiple practical applications like object removal and image editing. Deep GAN-based models greatly improve the inpainting performance in structures and textures within the hole, but might also generate unexpected artifacts like broken structures or color blobs. Users perceive these artifacts to judge the effectiveness of inpainting models, and retouch these imperfect areas to inpaint again in a typical retouching workflow. Inspired by this workflow, we propose a new learning task of automatic segmentation of inpainting perceptual artifacts, and apply the model for inpainting model evaluation and iterative refinement. Specifically, we first construct a new inpainting artifacts dataset by manually annotating perceptual artifacts in the results of state-of-the-art inpainting models. Then we train advanced segmentation networks on this dataset to reliably localize inpainting artifacts within inpainted images. Second, we propose a new interpretable evaluation metric called Perceptual Artifact Ratio (PAR), which is the ratio of objectionable inpainted regions to the entire inpainted area. PAR demonstrates a strong correlation

with real user preference. Finally, we further apply the generated masks for iterative image inpainting by combining our approach with multiple recent inpainting methods. Extensive experiments demonstrate the consistent decrease of artifact regions and inpainting quality improvement across the different methods. Dataset and code are available at: https://github.com/owenzlz/PAL4Inpaint

1 Introduction

Deep GAN-based image synthesis methods have been continuously improving image inpainting performance [64,73,30,4,36,68,66,74,41] for practical applications like object removal and image editing. Due to the ill-posed nature of image inpainting tasks, when encountering large holes or complex structures [41] within the hole, image inpainting becomes extremely challenging. Along with almost all state-of-the-art algorithms, inpainting artifacts tend to appear in the generated images. Those artifacts mostly include broken structures or color bleeding in the traditional patch synthesis methods [2], imperfect structures like disconnected or distorted lines, GAN-based generation artifacts or color blobs. In typical retouching workflows, users tend to judge the inpainting performance by those artifacts, and fix them by drawing masks on those regions and re-runing the automatic inpainting tools. Therefore, localizing and segmenting those artifacts is intuitively and naturally beneficial for inpainting algorithm evaluation and performance improvement.

Intuitively, finding more or larger artifacts within the hole area indicates a worse inpainting performance. Traditionally, image inpainting is regarded as an image reconstruction and restoration problem, and commonly-used metrics like PSNR, MSE and LPIPS [69] are utilized to compare the inpainted result to the original image in terms of content or pixel similarity. However, in many cases, image inpainting is used for foreground object removal [9,10]. Users prefer a visually plausible background generation rather than a faithful foreground reconstruction. Other quantitative metrics like Frechet Incept Distance (FID) [17,35] and Paired/Unpaired Inception Discriminative Score (P/U-IDS) [74] are computed on the entire images over large evaluation datasets. We are lacking in an intuitive metric which is more interpretable, operates on localized hole regions, and supports single result evaluation. Therefore, an automatic and reliable artifacts segmentation network may fill the gap.

In practical inpainting applications, users may choose to manually fix those artifacts by re-masking perceptually bad regions and re-running the models. Intuitively, after a couple of iterations, inpainting results are expected to be largely improved compared with the initial ones. Iterative hole filling has been studied in deep learning pipelines [66,34,13], and is shown to outperform one-pass inpainting. But the masks used in each iteration are either unreliable ones [66] learned with image reconstruction loss or predefined eroded masks [34,13]. Hence, an automatic artifacts segmentation network can effectively detect the perceptual artifacts in each iteration, and make the iterative filling run in a more efficient and effective way.

Although these inpainting artifacts are easily identifiable by humans, very few studies [66] have developed models to automatically detect and localize these artifacts in inpainting results. Researchers have studied identifying manipulated or synthesized images [50,8,48,71,32,50,5], edited image regions [5], or the entire inpainted image regions [23,53,22]. However, automatic localization of those artifacts within the inpainted holes was seldom discussed. This is mainly because a representative and well-organized dataset consisting of image inpainting results and artifact annotations is not yet available. Using the knowledge and expertise of professional photographers, deep networks can learn to efficiently detect and segment these artifacts.

In this paper, inspired by a typical user workflow when using inpainting tools, we assume that automatic perceptual artifacts segmentation for image inpainting will potentially benefit algorithm evaluation and boost inpainting performance. To verify our hypothesis, we collect inpainting results generated by multiple state-of-the-art deep inpainting models and annotate pixel-wise artifacts with a team of human professionals, and benchmark the dataset using advanced segmentation networks. Our proposed artifacts localization network outputs a binary mask highlighting the artifacts region. This mask can be used to: (1) compute the occupation ratio over the hole mask to evaluate and compare different inpainting algorithms on single test image without ground truth, and (2) achieve iterative filling to progressively improve inpainting performance. In summary, our contributions are in three folds:

- We study the importance of a novel task, inpainting artifacts segmentation. Given its strengths in inpainting evaluation and result refinement, we construct a dataset consisting of 4,795 inpainting results with per-pixel perceptual artifacts annotations. We further benchmark the dataset using multiple segmentation network structures and analyze the human subjective factors in detail. Extensive experiments demonstrate its robustness on state-of-the-art inpainting models.
- We present the Perceptual Artifact Ratio (PAR) calculated from the artifact area detected inside the hole. PAR is an interpretable, intuitive, simple yet effective evaluation metric for comparing inpainting algorithms on a single image without ground truth. Our metric makes it possible to automatically evaluate object removal performance. Our user study also shows that PAR correlates more strongly with real user preferences than other metrics.
- We applied the artifacts segmentation network to iterative filling pipeline. After each iteration, we visualize that the detected artifact regions are consistently shrinking for all the tested inpainting models, and the results are refined with better structures and colors. Another user study suggests that iterative filling using our proposed artifacts masks will likely not degrade the inpainting performance and in many cases improve it.

2 Related Work

2.1 Image Inpainting

Classical image inpainting methods include diffusion-based methods [3,1] that propagate information from the boundary inwards to fill the hole, and patch-based methods [2] that search for the reference region to fill the hole. On the rise of deep learning, researchers proposed deep models to improve the inpainting performance from diverse angles, such as attention mechanism [19][62][29][55][28][63][45][41], loss function and discriminator design [19][58][59][62][65], progressive [13][66][25] [24][24][67] or multiscale [57][60][65][52] architectures, use of intermediate guide respresentation [62][33][38][39][26][56][47][51][12], and multimodal plausible outputs [75][64][73][14][30][4][36][68]. Among these works, ProFill [66], CoMod-GAN [74], and LaMa [41] are the most recent leading models. ProFill [66] proposed to implicitly learn a confidence map that guides the generator to iteratively fill the hole, as well as a attention-guided refinement module to upsample the output. CoMod-GAN [74] leveraged the StyleGAN architecture [20] to conditionally synthesize filled region, where their filled content could be creative and not necessarily existed in the context. Finally, LaMa [40] integrated the fast Fourier convolution [6] to effectively capture global contextual information, and set the new state-of-the-arts. The goal of our work is to detect and localize the perceptual artifacts in the filled images independent of the inpainting models, and thus is in an orthogonal direction to these previous inpainting works.

2.2 Image Inpainting Quality Assessment

There are two types of commonly used metrics for image inpainting. The first quantifies the performance for a whole dataset of generated images. These metrics include Frechet Incept Distance (FID) [17][35] and Paired/Unpaired Inception Discriminative Score (P/U-IDS) [74], which measure the distance between the distribution of generated and real images using the deep Inception features [42]. For single image quality assessment, previous works often treat inpainting as a reconstruction task and thus compare the filled image with the original image using the reconstruction metrics, such as MSE, SSIM, PSNR or LPIPS [69]. This is reasonable only when holes are sampled on the background region. When holes largely overlap with or totally cover a foreground object, the current models would mostly fill the hole using background pixels, which is totally irrelevant to the original content. In these cases, the filled region of object removal could look natural and realistic, but could be totally different from the original object. Thus, reconstruction metric would no longer be a proper metric. Other potential assessments for measuring single image inpainting quality of object removal are No-Reference Image Quality Assessment (NR-IQA) methods [11][40][43][61][70][21]. Although previous inpainting works have rarely used NR-IQA metrics, we tried out two recent methods Hyper-IQA [40] and MUISQ [21], and found that MUISQ [21] has a relatively reasonable correlation with human perception compared to the other existing metrics for measureing object

removal inpainting quality. In this work, we aim to use the area size of localized inpainting artifacts as a no-reference metric to measure the quality of hard case inpainting, which is object removal. Experimental studies in section 5 show that our proposed metric outperforms both reconstruction-based metrics and existing NR-IQA in terms of correlation with human perception.

2.3 Detecting Artifacts in Generated Images

Other related works include detecting the generated/fake images and localizing the manipulated region in the image. One line of works [8][48][71][32][50][5] have studied training a binary classifier to classify the generated images, and Wang et al. [50] has shown surprising generalization on diverse and unseen model outputs by detecting the common artifacts in CNNs/GANs. Chai et al. [5] proposed patch-based classifier to localize the region that causes the fake image detectable. Another line of works [7][18][37][49] have proposed techniques to detect general image manipulations, such JPEG and resampling, other than GANs. In the image inpainting domain, Li et al. [23] first proposed to use the high-pass filter CNNs to detect the inpainting region given the filled image. Later, Wu et al. [53] and Li et al. [22] further improves the generalization of the mask detection to diverse inpainting models by proposing novel architecture or explicitly process high-frequency noise residual. Although all these works are related to us, a fundamental difference is that we aim to detect the perceptual artifacts that are judged by humans rather than simply detecting high-frequency noise/artifacts in the generated images. More specifically, in the inpainting context, our system detects the perceptual artifact region rather than the whole mask region, where perceptual artifact region is often a small subset of the mask. Thus, our work is essentially a different task from [23][53][22].

3 Dataset Labeling and Statistics

In order to train a system that can detect the perceptual artifacts in the inpainted images, we build a dataset that consists of 4,795 images with per-pixel perceptual artifacts labels from humans. We use three leading inpainting models ProFill [66], CoMod-GAN [74], and LaMa [41] to generate images to label. A labeling interface and a few examples of the labeled images are shown in the left and right of Fig.5, respectively. During labeling, we provide the users a filled image without showing the original image, and ask users to label regions with perceptual artifacts on their tablets. We intentionally do not include the original images in the interface, since otherwise users might have bias to compare everything with the original content in the hole. As we have discussed in section 2.2, the filled image could look natural and realistic, even though it's very different from the original image. We also put dilated bounding box around the hole region to help users more easily find the labeling region and focus on it. We intentionally do not indicate the hole mask in the image, so that the workers do not have any bias labeling around the hole boundary, and thus can purely make judgement based on the

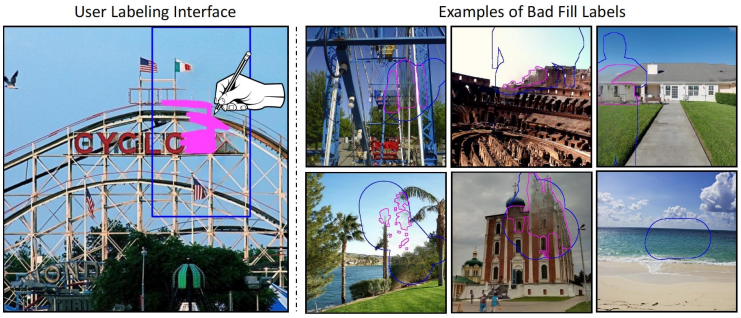

Fig. 2. The left is an illustration of interface that users label on the inpainted image, where the blue bounding box indicates the region the users should inspect. On the right, we show a few examples of bad fill labels from users, where the blue and pink boundaries indicate the holes and user labels, respectively.

perceptual quality. In this case, since workers do not know where the hole is, their labeling might go over the hole boundary. However, this is not an issue, as we simply intersect the hole masks with the human labels as a post-process. In addition, we provide a duplicate of the filled image in the interface, so that when workers can see the unbrushed filled image as reference during brushing.

From the labeling perspective, one fundamental challenge is that this task is highly subjective compared to traditional segmentation, since different people may have different opinions or standards to judge perceptual artifacts. Therefore, in order to standardize the labeling as much as possible, we recruit and train a professional team for this task. We use two rounds of checks to avoid "missed labeled" or "overly labeled" regions. In the first round, the professional workers cross check the results with each other. In the second round, a single human expert checks through all the labels. On average, approximately 10% of labels have been rectified during the checking process. In addition, in order to generate high-quality labels, we recruited five human experts with photography or design background in our team to label these images. Since these workers are heavy users of image editing tools, i.e. Photoshop, their labeling criterion could better reflect the common unsatisfactory/retouch regions in the hole filling process.

Among these 4,795 images, there are 832 images that have nearly perfect fills, and thus workers did not label anything on these images. Although these images do not have segmentation labels, adding them into training could effectively help the network avoid predicting false positives. In terms of size of the labeling region, we found that the averaged ratio of "perceptual artifacts region / hole mask region" is 29.67%. This number once again shows that detecting perceptual artifacts is fundamentally different from detecting the hole mask in [23][53][22]. We plan to release our dataset to the community for future research.

Models	IoU	Precision	Recall	Fscore
ResNet-50 backbone [16] + HRNet head [46]	41.35	58.45	58.56	58.51
Swin-B backbone [31] + Uper head [54]	44.20	63.01	59.69	61.30
ResNet-50 backbone [16] + PSPNet head [72]	46.04	59.78	66.71	63.05
- Perfect Filled Images	43.83	64.92	57.43	60.94
- Pretrained Weights	44.93	66.22	58.29	62.00
+ Hole Mask	45.96	66.07	60.16	62.98
+ Pseudo Pretraining	46.44	62.01	64.91	63.43
+ Pseudo Pretraining & Real Images	**46.77**	59.59	**68.49**	**63.73**
Human Subject A	45.60	**75.07**	53.73	62.64
Human Subject B	42.21	60.40	58.36	59.36
Human Subject C	36.85	61.47	47.93	53.86

Table 1. An ablation study of the segmentation model, and human performance.

4 Perceptual Artifacts Segmentation

In this section, we discuss the details of our segmentation model along with extensive ablation studies. During training, we used "8:1:1" ratio to randomly split the train/val/test set. In total, we have 3,836 training images, 480 validation images, and 479 test images. In each model training, we use the validation set to select the best checkpoint, and evaluate the performance on the test set. All of our models are trained and evaluated using the MMSegmentation codebase [1].

4.1 Ablation Studies

In the ablation study, we first tried out a few advance segmentation backbones/heads, such as HRNet [46] head, PSPNet [72] head, ResNet-50 backbone [16], and Swin Transformer [31] backbone, as shown in the top 3 rows (excluding the header) of Table 1. However, we do not observe obvious improvement when using the more complex backbones or heads for our task, after several trials of training comparison. We think a major potential reason is that our segmentation performance of the simpler backbone [16] and head [72] is nearly saturated given the highly subjective labels, and thus simply adding capacity or complexity of backbone does not improves much. This is discussed more in details when we compare with human performance in section 4.2. Thus, we chose ResNet-50 backbone [16] + PSPNet head [72] as our base network, due to its simplicity and efficiency. The rest of ablation studies all shared the same base network for fair comparison, and thus the results should be compared with 3^{rd} row.

Besides the network backbones, we also studied other aspects that might potentially affect the segmentation performance. As we mentioned in section 3,

[1] MMSegmentation github: https://github.com/open-mmlab/mmsegmentation

832 images in the labeled dataset have almost perfect fill and thus have no mask labeling, and thus we wonder whether having these images in the training would be helpful. In the 4^{th} row, we can see that the model trained without using these images indeed has worse performance, which concludes that adding perfect fills to training is important. All of our models starts training based on the checkpoints pretrained on ADE20K [76], and we show that the performance also decreases obviously without pretrained weights, as shown in the 5^{th} row. Another intuitive thing is to concatenate the hole mask in the input, as it could theoretically help the network quickly localize the potential artifacts region. However, as shown in the 6^{th} row, our experiments show that adding the mask into input channel does not actually boost the segmentation performance, and thus we decide not to use it for the simplicity purpose.

We also studied the possibility of generating pseudo labels on large scale unlabeled images for the pretraining purpose. Inspired by BoxInst [44], which used bounding box masks as weak supervision to train instance segmentation, we aim to find some similar "enlarged" masks covering the artifacts region as our pseudo labels. Initially, we tried using the hole mask as weak supervision, but realize that the network quickly overfits on the high-frequency artifacts on the hole boundary, which is not useful for our purpose. To this end, we used a pretrained artifacts segmentation network to generate artifacts mask regions on 100K unlabeled images, and then enlarged the segmented masks by some random dilation iterations to cover the perceptual artifacts region. The results in the 7^{th} row show that such pretraining strategy slight improves the performance. Finally, we also tried adding the same quantity of real images into training, where the masks are empty for these real images. The 8^{th} row shows that this is also useful to further boost the performance.

Fig. 3. Qualitative results of the predicted bad fill segmentation on six different inpainting model outputs. The pink and blue boundaries indicate the predicted bad fill region and the hole region, respectively. Please feel free to zoom in to see the details.

4.2 Analysis on Human Perceptual Judgement

As mentioned in section 3, the labeled dataset from our labeling team has been carefully checked and thus have relatively high quality. In order to better understand how subjective human judgements are and the human performance bound, we ask three more human subjects to label on the 479 test images. Then, we compare the labels from these three human subjects with the previous labels from the labeling team, which are shown in the last three rows of Table 1. Regarding the three workers' background, human subject A has worked on this task before but not these images, and human subjects B & C have never worked on this task before but are taught by the labeling team with a bunch of labeled examples. Thus, human subject A should theoretically have better understanding of the task as well as the labeling criterion of the labeling team, compared to the other two subjects. All of them have photography or design background.

Interestingly, the results show that our segmentation model reaches and even surpasses the best human subject on all metrics except for precision. This infers that our model actually learns a better understanding of averaged judgement criterion of the labeling team, compared to each individual human. On the other hand, these results also indicate that humans have very subjective opinions on the labeling the artifacts regions, as the quantitative scores deviate obviously from each other. A visual illustration of different people's labels on the same filled image is shown in Fig. 4, and we include more examples like this in the supplemental. Since our segmentation performance surpasses the human performance, this indicates that our segmentation model reaches to a near saturation point for this highly subjective segmentation task. This might also explain why more complex backbone [31] or other tweaks of data or training do not provide significant performance improvement, as we observed in the ablation study.

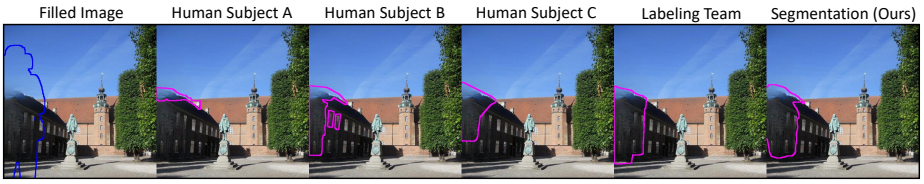

Fig. 4. A visual comparison between labels from multiple human subjects on the same filled image. Our segmentation result is shown in the last column.

5 Evaluating Inpainting Quality for Object Removal

5.1 Motivation

It has been widely discussed [62][63] that image inpainting lacks good evaluation metrics, especially for single image quality assessment. Previous works mostly

treat image inpainting as a kind of restoration task, so the reconstruction metrics, such as MSE, SSIM, PSNR, and LPIPS [69], are often used to quantify the similarity between the filled image and the original image. Thinking carefully, we realize that reconstruction metrics might reasonably measure inpainting performance only when the holes are not very large and on the background region. When the holes largely overlap with or cover the foreground objects, most inpainitng algorithms would fill the hole regions by using the background context, where the object is oftentimes completely removed from the image. In these scenarios, reconstruction metrics are no longer proper metrics to gauge the inpainting quality, since the filled region could be totally irrelevant to the original pixels inside the hole. As shown in Fig. 2, when removing the person from the image, output A is visually more plausible than output B, but somehow all the existing reconstruction metrics make opposite judgement. Embarrassingly, object removal is arguably the most frequently used applicable scenario for inpainting algorithms. Thus, it means we really lack good metric for assessing inpainting quality in this scenario. This motivates us to think if the perceptual artifacts localization could be used as a no-reference metric to evaluate inpainting quality in the object removal scenario.

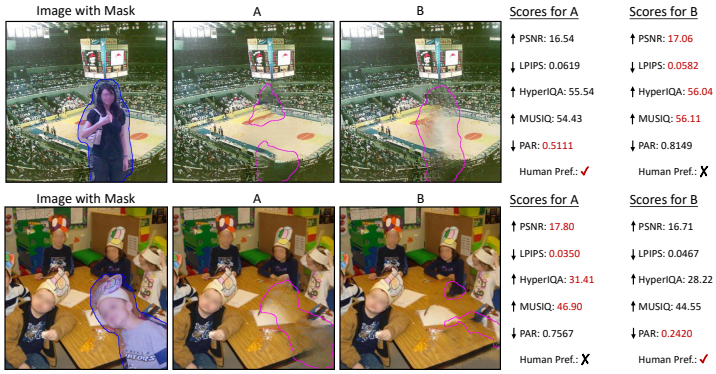

Fig. 5. A visual illustration of filled outputs by two inpainting models [66][33], with the corresponding metric scores. The red scores indicate the preferred choice according to each metric.

5.2 Metric Definition

Since our segmentation model could generalize reasonably well to diverse and unseen inpainting methods, we start to wonder whether the size of the detected artifacts region can be used as a metric to assess the inpainting quality. Basically, we assume that an image with good inpainting quality should have relatively smaller the perceptual artifacts region, and vice versa. We name this metric as

Perceptual Artifacts Ratio (PAR), which is the ratio of "size of the perceptual artifact region / size of the input hole". The metric computation procedure is that we first run the segmentation model on the filled image and then compute PAR for any filled images, without the need of using the original image. During the quality comparison between different inpainting models, we simply evaluate which inpainting ouptuts have smaller artifacts region among the comparisons.

5.3 Correlation with Human Perception

In order to evaluate how well our PAR metric correlates with human perception, we collected user preferences on the filled images between four pairs of inpainting methods. Among these user comparisons, two pairs of comparisons happen between two strong inpainting models, as shown in the first two rows of Table. 2, and another two pairs are between one strong and one relatively weak model, as shown in the bottom two rows of Table. 2. In each pair of methods comparison, we show users two filled images with randomized order, and ask users to pick the preferred image out of the two options. The user studies were conducted on Amazon Mechanical Turk (AMT), where we asked five users to vote on each image. Finally, we consider that one filled image is strongly preferred than the other, only if 4 out of 5 users reach an agreement. In this study, we only used the strongly preferred image pairs as human preference ground truth to reduce the noise as much as possible, where the number of strongly preferred cases are shown in the 2^{nd} column of Table 2. Since we are evaluating inpainting quality in the object removal scenarios, we use Mask R-CNN [15] pretrained on COCO [27] to generate object masks, and dilate three iterations with 5×5 kernel to increase the mask coverage on the object.

As shown in Table 2, out of 1,000 images for each pair of method comparison, we found that users reach strong agreement on a subset of images with quantity ranging from 321 to 718 shown in the 2^{nd} column. The reason why the number of strongly preferred cases of "LaMa vs. ProFill" are less than the others is that these two methods have relatively closer inpainting performance, which causes more disagreement. Other columns in Table 2 basically indicates the percentage of correct ranking from each metric, with respect to the human perceptual judgement. In this study, we compare with two reconstruction metrics PSNR and LPIPS [69], as well as two NR-IQA metrics Hyper-IQA [40] and MUSIQ [21]. Overall, the quantitative results show that our PAR metric outperforms all these existing metrics for assessing inpainting quality in object removal scenarios.

5.4 PAR Analysis with Hole Size and Scene Types

We claimed that inpainting artifacts mostly appear in larger holes and complex scene structures. Using our pretrained artifacts segmentation model, we also studied how PAR would change with respect to the hole size in two scenarios: man-made scenes and natural scenes. In the places2 testing dataset, we sampled man-made scenes from the categories, such as building, room, shop, stadium, studio, factory and so on. On the other hand, natural scenes are sampled from

Comparisons	No. Pairs	PSNR	LPIPS [69]	HyperIQA [40]	MUSIQ [21]	PAR (Ours)
LaMa vs. ProFill	321	56.70 %	62.31 %	39.97%	65.11%	**65.42 %**
LaMa vs. CoMod-GAN	367	48.77 %	48.77 %	51.50%	55.31%	**69.21 %**
ProFill vs. EdgeConnect	560	23.92 %	11.96 %	56.39%	49.62%	**79.82 %**
LaMa vs. EdgeConnect	718	44.71 %	43.45 %	35.71%	71.72%	**72.70 %**
Overall	1966	41.50%	38.55 %	45.24%	61.28%	**72.89 %**

Table 2. Quantitative results for measuring the correlation between different metrics and human perceptual judgement.

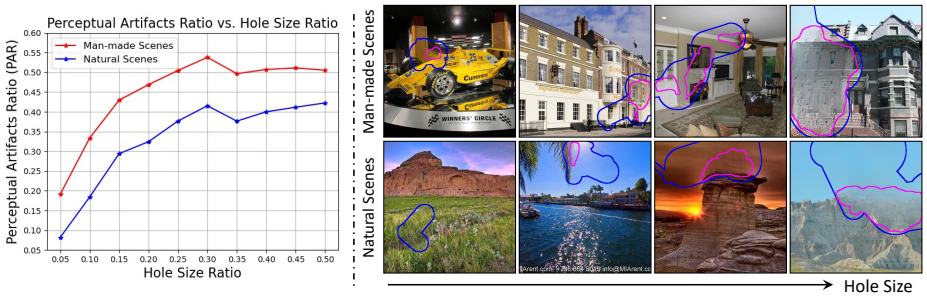

Fig. 6. Left: relationship between PAR and hole size for both man-made scenes and natural scenes. **Right**: some visual examples of segmented perceptual artifacts region (pink boundary) with varying hole (blue boundary) size. Inpainting models like LaMa produce more artifacts when the hole is larger and scenes are more complex.

categories, such as sky, land, mountain, forest, garden, pasture, beach, desert, and so on. We sampled 2,000 test images for both scenarios and randomly placed stroke holes of specific sizes on them. Then we run LaMa to fill the hole. The relationship between PAR and hole size for natural or man-made scenes is shown in Fig. 6. Our conclusions from the figure are: (1) As the hole size increases, LaMa has a higher possibility of generating inpainting artifacts. (2) Inpainting models like Lama struggles more to complete man-made structures than natural scenes. We believed that this rule applies to other inpainting algorithms as well.

6 Making Inpainting Models Iterative

Modern inpainting algorithms have shown consistent performance improvement over the last few years. However, when inpainting large holes, we often still observe that the inpainting models could often perfectly fill a partial region of the hole while generating obvious artifacts on the other regions. Given this observation, an intuitive idea is that: if the perceptual artifacts region can be reliably segmented out, can we enable the inpainting refill on the artifacts region? In this section, we discuss how we make the inpainting models iteratively fill on the artifacts region, and its effectiveness to improve the inpainting quality.

6.1 Iterative Fill Pipeline

In Fig. 7, we show an overview of our iterative fill pipeline. The input image with hole is first fed into an inpainting model to generate a filled image. Then, the filled image is fed into our perceptual artifacts segmentation model to detect the artifacts region, which are converted into the hole mask for the next iteration inpainting. We post-process the segmentation output of artifacts region by multiplying it with the original hole mask in an element-wise manner, so that we ensure not to change any pixels outside the original hole during iterative fill. Our iterative fill pipeline is extremely simple to integrate with and agnostic to all the inpainting models.

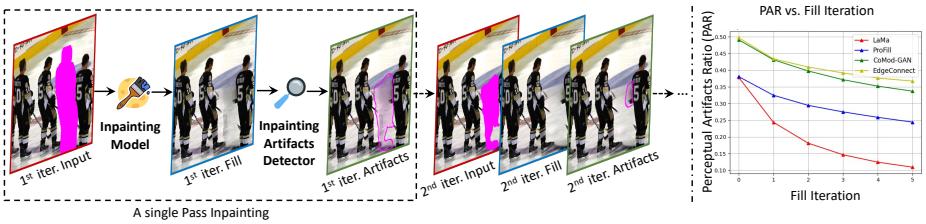

Fig. 7. Left: an overview pipeline of our iterative fill. **Right**: curves that show predicted perceptual artifacts ratio consistently decreases over the fill iteration for all inpainting models.

6.2 Performance Improvement by Iterative Fill

We evaluate the performance of the iterative fill from two aspects. First, we compute the size of the detected artifacts region or PAR over the fill iterations, as shown in the right of Fig. 7. We observe that the detected artifacts region consistently decreases as more iterative fill happens. This indicates that our iterative fill indeed improves the filled image quality, such that less perceptual artifacts are detected. Here, we show up to 5^{th} iterative fill in the main paper, and put analysis on more iterations of refill in the supplemental. We also conducted a user study that shows whether users think the 5^{th} iteratively filled image are better, same, or worse than the original fill for four inpainting models. As shown in Table 3, our iterative fill pipeline improves approximately 30% of images compared to the original fill, and rarely make the filled images worse, especially for the best model LaMa [40]. This implies that our system can be safely integrated into these inpainting models to boost inpainting quality. All of our user studies are conducted on AMT. In each inpainting method, we uniformly sampled 500 images from the testset, which result in 2,000 images in total. We asked 20 turkers to carefully check on each image and averaged the preference. We do not use the traditional metrics to quantify the performance between original fill and iterative fill, since we found that these metric scores between them are too close

and sometimes random, which does not reflect much information. We have more discussion on this in the supplemental.

We show visual comparison between original fill and 5^{th} iterative fill in Fig. 8. We observe that iterative fill could oftentimes help the inpainting models refine both structure and texture in many cases. However, due to the limitation of the inpainting algorithms themselves, the predicted perceptual artifact regions would not always reach to zero and thus would still leave some artifacts in the image.

Models	Preferred Original Fill	Same	Preferred Iterative Fill
EdgeConnect	53 (10.6%)	258 (51.6%)	189 (37.8%)
CoMod-GAN	45 (9.0%)	334 (66.8%)	121(24.2%)
ProFill	14 (2.8%)	337 (67.4%)	149 (29.8%)
LaMa	9 (1.8%)	341 (68.2%)	150 (30.0%)

Table 3. A user study to show the comparison between original fill and the 5^{th} refill.

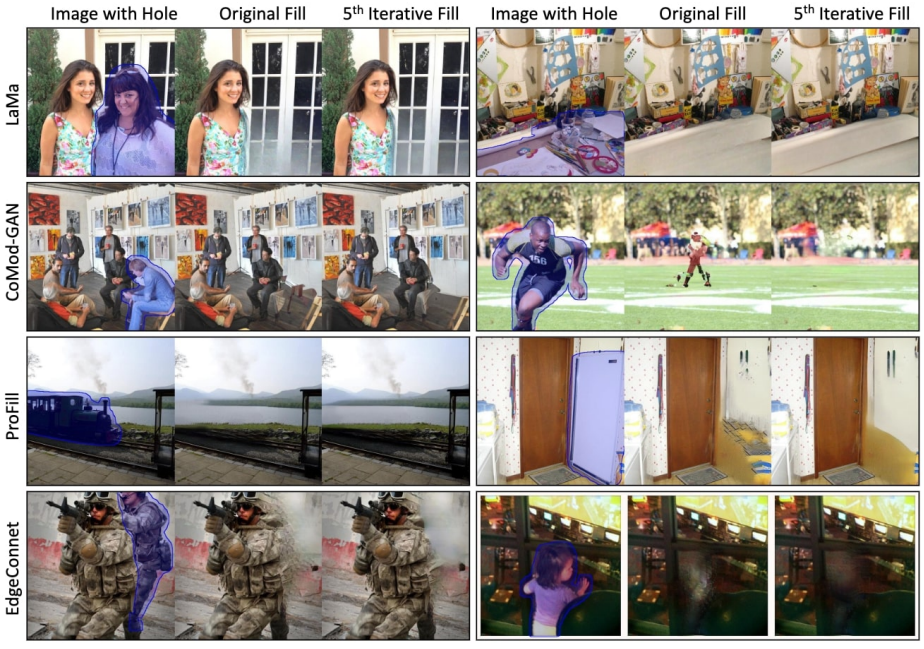

Fig. 8. Qualitative comparison between the original fill and the 5^{th} iterative fill.

References

1. Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. IEEE TIP (2001) 4
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009) 2, 4
3. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: SIGGRAPH (2000) 4
4. Cai, W., Wei, Z.: Piigan: generative adversarial networks for pluralistic image inpainting. IEEE Access **8**, 48451–48463 (2020) 2, 4
5. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? understanding properties that generalize. In: European Conference on Computer Vision. pp. 103–120. Springer (2020) 3, 5
6. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. Advances in Neural Information Processing Systems **33** (2020) 4
7. Cozzolino, D., Poggi, G., Verdoliva, L.: Splicebuster: A new blind image splicing detector. In: 2015 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6. IEEE (2015) 5
8. Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., Verdoliva, L.: Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510 (2018) 3, 5
9. Criminisi, A., Perez, P., Toyama, K.: Object removal by exemplar-based inpainting. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. vol. 2, pp. II–II. IEEE (2003) 2
10. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on image processing **13**(9), 1200–1212 (2004) 2
11. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3677–3686 (2020) 4
12. Guo, X., Yang, H., Huang, D.: Image inpainting via conditional texture and structure dual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14134–14143 (2021) 4
13. Guo, Z., Chen, Z., Yu, T., Chen, J., Liu, S.: Progressive image inpainting with full-resolution residual network. In: Proceedings of the 27th acm international conference on multimedia. pp. 2496–2504 (2019) 2, 4
14. Han, X., Wu, Z., Huang, W., Scott, M.R., Davis, L.S.: Finet: Compatible and diverse fashion image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4481–4491 (2019) 4
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 11
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 7
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 2, 4
18. Huh, M., Liu, A., Owens, A., Efros, A.A.: Fighting fake news: Image splice detection via learned self-consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 101–117 (2018) 5

19. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics (ToG) **36**(4), 1–14 (2017) 4
20. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020) 4
21. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021) 4, 11, 12
22. Li, A., Ke, Q., Ma, X., Weng, H., Zong, Z., Xue, F., Zhang, R.: Noise doesn't lie: Towards universal detection of deep inpainting. arXiv preprint arXiv:2106.01532 (2021) 3, 5, 6
23. Li, H., Huang, J.: Localization of deep inpainting using high-pass fully convolutional network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8301–8310 (2019) 3, 5, 6
24. Li, J., He, F., Zhang, L., Du, B., Tao, D.: Progressive reconstruction of visual structure for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5962–5971 (2019) 4
25. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7760–7768 (2020) 4
26. Liao, L., Xiao, J., Wang, Z., Lin, C.W., Satoh, S.: Guidance and evaluation: Semantic-aware image inpainting for mixed scenes. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. pp. 683–700. Springer (2020) 4
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 11
28. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 85–100 (2018) 4
29. Liu, H., Jiang, B., Xiao, Y., Yang, C.: Coherent semantic attention for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4170–4179 (2019) 4
30. Liu, H., Wan, Z., Huang, W., Song, Y., Han, X., Liao, J.: Pd-gan: Probabilistic diverse gan for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9371–9381 (2021) 2, 4
31. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021) 7, 9
32. Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L.: Detection of gan-generated fake images over social networks. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 384–389. IEEE (2018) 3, 5
33. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019) 4, 10
34. Oh, S.W., Lee, S., Lee, J.Y., Kim, S.J.: Onion-peel networks for deep video completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4403–4412 (2019) 2

35. Parmar, G., Zhang, R., Zhu, J.Y.: On buggy resizing libraries and surprising subtleties in fid calculation. arXiv preprint arXiv:2104.11222 (2021) 2, 4
36. Peng, J., Liu, D., Xu, S., Li, H.: Generating diverse structure for image inpainting with hierarchical vq-vae. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10775–10784 (2021) 2, 4
37. Rao, Y., Ni, J.: A deep learning approach to detection of splicing and copy-move forgeries in images. In: 2016 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6. IEEE (2016) 5
38. Ren, Y., Yu, X., Zhang, R., Li, T.H., Liu, S., Li, G.: Structureflow: Image inpainting via structure-aware appearance flow. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 181–190 (2019) 4
39. Song, Y., Yang, C., Shen, Y., Wang, P., Huang, Q., Kuo, C.C.J.: Spg-net: Segmentation prediction and guidance network for image inpainting. arXiv preprint arXiv:1805.03356 (2018) 4
40. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3667–3676 (2020) 1, 4, 11, 12, 14
41. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv:2109.07161 (2021) 2, 4, 5
42. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016) 4
43. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE transactions on image processing **27**(8), 3998–4011 (2018) 4
44. Tian, Z., Shen, C., Wang, X., Chen, H.: Boxinst: High-performance instance segmentation with box annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5443–5452 (2021) 8
45. Wan, Z., Zhang, J., Chen, D., Liao, J.: High-fidelity pluralistic image completion with transformers. arXiv preprint arXiv:2103.14031 (2021) 4
46. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020) 7
47. Wang, N., Ma, S., Li, J., Zhang, Y., Zhang, L.: Multistage attention network for image inpainting. Pattern Recognition **106**, 107448 (2020) 4
48. Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., Liu, Y.: Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. arXiv preprint arXiv:1909.06122 (2019) 3, 5
49. Wang, S.Y., Wang, O., Owens, A., Zhang, R., Efros, A.A.: Detecting photoshopped faces by scripting photoshop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10072–10081 (2019) 5
50. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8695–8704 (2020) 3, 5
51. Wang, T., Ouyang, H., Chen, Q.: Image inpainting with external-internal learning and monochromic bottleneck. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5120–5129 (2021) 4

52. Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J.: Image inpainting via generative multi-column convolutional neural networks. arXiv preprint arXiv:1810.08771 (2018) 4
53. Wu, H., Zhou, J.: Giid-net: Generalizable image inpainting detection network. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 3867–3871. IEEE (2021) 3, 5, 6
54. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 418–434 (2018) 7
55. Xie, C., Liu, S., Li, C., Cheng, M.M., Zuo, W., Liu, X., Wen, S., Ding, E.: Image inpainting with learnable bidirectional attention maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8858–8867 (2019) 4
56. Xiong, W., Yu, J., Lin, Z., Yang, J., Lu, X., Barnes, C., Luo, J.: Foreground-aware image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5840–5848 (2019) 4
57. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6721–6729 (2017) 4
58. Yang, J., Qi, Z., Shi, Y.: Learning to incorporate structure knowledge for image inpainting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12605–12612 (2020) 4
59. Yeh, R.A., Chen, C., Yian Lim, T., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5485–5493 (2017) 4
60. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: 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 (2020) 4
61. Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., Bovik, A.: From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3575–3585 (2020) 4
62. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5505–5514 (2018) 4, 10
63. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4471–4480 (2019) 4, 10
64. Yu, Y., Zhan, F., Wu, R., Pan, J., Cui, K., Lu, S., Ma, F., Xie, X., Miao, C.: Diverse image inpainting with bidirectional and autoregressive transformers. arXiv preprint arXiv:2104.12335 (2021) 2, 4
65. Zeng, Y., Fu, J., Chao, H., Guo, B.: Learning pyramid-context encoder network for high-quality image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1486–1494 (2019) 4
66. Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., Lu, H.: High-resolution image inpainting with iterative confidence feedback and guided upsampling. In: European Conference on Computer Vision. pp. 1–17. Springer (2020) 2, 3, 4, 5, 10
67. Zhang, H., Hu, Z., Luo, C., Zuo, W., Wang, M.: Semantic image inpainting with progressive generative networks. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 1939–1947 (2018) 4

68. Zhang, L., Wang, J., Shi, J.: Multimodal image outpainting with regularized normalized diversification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3433–3442 (2020) 2, 4
69. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 2, 4, 10, 11, 12
70. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology **30**(1), 36–47 (2018) 4
71. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6. IEEE (2019) 3, 5
72. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017) 7
73. Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., Lu, D.: Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5741–5750 (2020) 2, 4
74. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. arXiv preprint arXiv:2103.10428 (2021) 2, 4, 5
75. Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1438–1447 (2019) 4
76. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017) 8