Perceptual Artifacts Localization for Image Synthesis Tasks

Lingzhi Zhang★♠ [1,2]   Zhengjie Xu★ [2]   Connelly Barnes[1]   Yuqian Zhou[1]   Qing Liu[1]

He Zhang[1]   Sohrab Amirghodsi[1]   Zhe Lin[1]   Eli Shechtman[1]   Jianbo Shi[2]
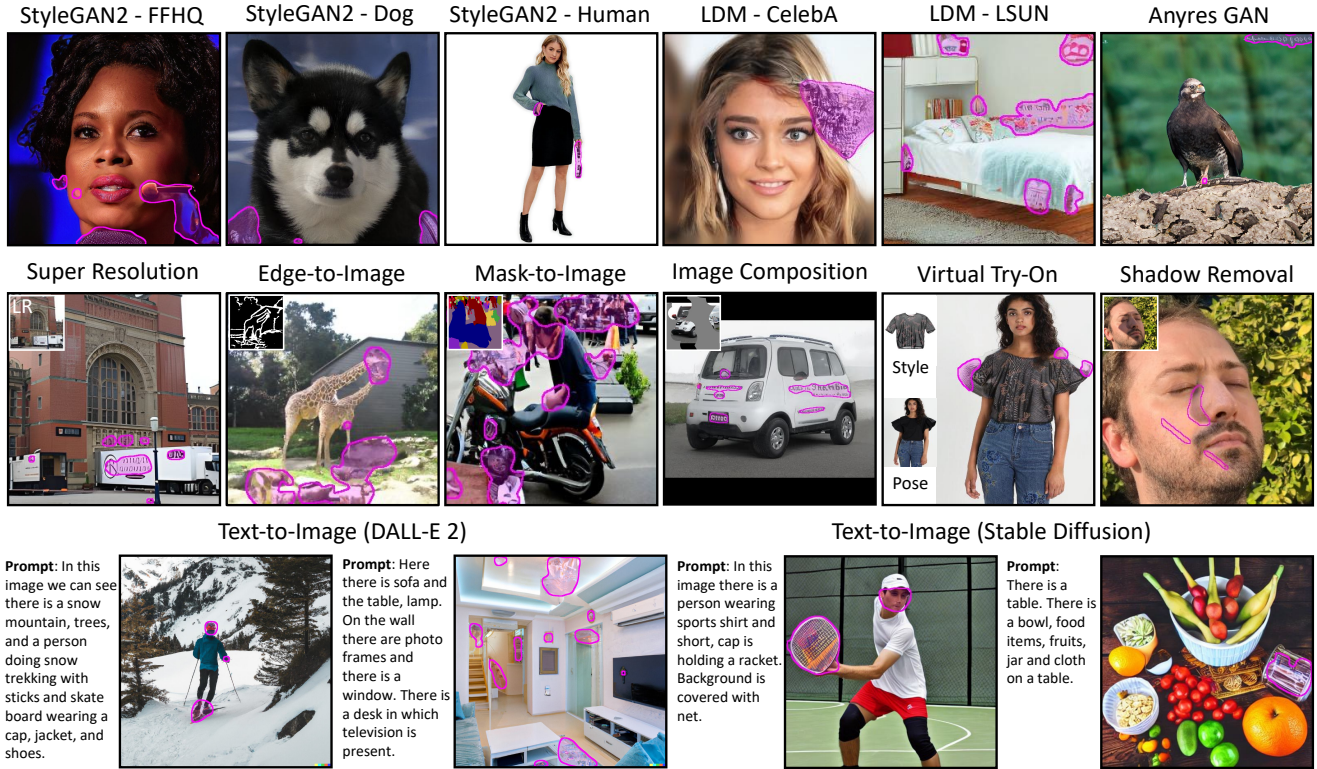
[1]Adobe Inc.   [2]University of Pennsylvania

Figure 1: The visualization of predicted perceptual artifacts localization on ten image synthesis tasks. The first row contains unconditionally generated images from StyleGAN2 [19], Latent Diffusion Model (LDM) [43], and Anyres GAN [8]. The second row shows the results on types of conditional generated images, including super-resolution with Real-ESRGAN [57], edge-to-image with PITI [55], mask-to-image with PITI [55], image latent composition [9], virtual try-on [15], and portrait shadow removal [66]. The conditional inputs are placed at the top left of the images. In the last row, we show predictions on the text-to-image outputs from DALL-E 2 [41] and Stable Diffusion [43].

Abstract

Recent advancements in deep generative models have facilitated the creation of photo-realistic images across various tasks. However, these generated images often exhibit perceptual artifacts in specific regions, necessitating manual correction. In this study, we present a comprehensive empirical examination of Perceptual Artifacts Localization (PAL) spanning diverse image synthesis endeavors. We introduce a novel dataset comprising 10,168 generated images, each annotated with per-pixel perceptual artifact labels across ten synthesis tasks. A segmentation model, trained on our proposed dataset, effectively localizes artifacts across a range of tasks. Additionally, we illustrate

---

★ indicates equal contribution. ♠ work done when Lingzhi is a graduate student at University of Pennsylvania.

its proficiency in adapting to previously unseen models using minimal training samples. We further propose an innovative zoom-in inpainting pipeline that seamlessly rectifies perceptual artifacts in the generated images. Through our experimental analyses, we elucidate several practical downstream applications, such as automated artifact rectification, non-referential image quality evaluation, and abnormal region detection in images. The dataset and code are released here: https://owenlz.github.io/PAL4VST

1. Introduction

Generative models have made significant progress in a myriad of image synthesis tasks, including unconditional generation [5, 21, 19, 17, 12], image inpainting [69, 51, 32, 24, 70, 64], image-to-image translation [38, 42, 50, 45, 55], and text-to-image synthesis [13, 63, 36, 41, 43, 46, 2], among others. However, even cutting-edge models occasionally generate implausible content or display unpleasant artifacts in specific regions of the image, which we refer to as perceptual artifacts. These artifacts are easily detectable by the human eye. Therefore, in typical image editing processes, users often retouch generated images, masking and re-editing these regions to achieve perfection.

The manual retouching of perceptual artifacts is time-consuming and iterative. Such artifacts also pose challenges for generative models in achieving full automation in image synthesis, editing, or batch processing without human oversight. These challenges drive our exploration into the feasibility of training AI oracle models to identify and segment these perceptual artifacts. A successful implementation would present users with an automatically delineated mask of potential artifact areas, eliminating manual masking. Moreover, we could offer users the option to deploy established editing techniques, like inpainting, to these detected regions, thereby enhancing the automation of the retouching process.

Technically, the ideal goal is to generate a flawless image in a single pass. However, today's leading large-scale diffusion models often struggle to capture intricate details like subtle facial features, hands, and other object-specific nuances. While integrating more training data or using weighted loss might appear as potential solutions to these issues, they could compromise image quality in broader contexts. Until we achieve perfect single-pass outputs, automating the localization and refinement of perceptual artifacts stands as a promising direction to improve image synthesis quality.

To meet this objective, we've amassed a dataset of generated images, complemented with per-pixel artifact segmentation labels across a range of synthesis tasks. Using this dataset, we trained a segmentation model adept at localizing perceptual artifacts across various tasks. Our pretrained ar-

tifact detector showcases its versatility across multiple new models, adapting with enhanced accuracy even with limited training samples.

In conjunction with our artifact detection, we also unveil several practical applications. The foremost of these is the automatic refinement of artifacts in generated images using inpainting. However, it's observed that leading diffusion inpainting models, like DALL-E [41] and Stable Diffusion [43], sometimes falter in generating high-fidelity object details, such as facial features. We hypothesize this may stem from an unsuitable inpainting context. Consequently, we introduce a zoom-in inpainting pipeline, presenting a more apt input context before inpainting. This simple approach effectively mitigates challenges tied to object detail generation, without necessitating model training or alterations.

Our primary contributions include:

- A novel high-quality dataset comprising 10,168 images with per-pixel artifact annotations from humans, spanning ten diverse image synthesis tasks.
- A segmentation model adept at localizing perceptual artifacts across multiple synthesis tasks. Our pretrained model exhibits a rapid adaptation capability to new techniques with minimal training examples.
- An novel zoom-in inpainting pipeline for the automated refinement of intricate details in generated images.
- Demonstrated applications of our artifact detector, which include: 1). automatic artifact refinement; 2). reference-free image quality evaluation; and 3). anomaly detection in natural images.

We will release the dataset and the code.

2. Related Work

Detecting Generated Images. As generated images become increasingly photo-realistic, numerous studies [68, 27, 1, 4, 10, 34, 44, 7, 35, 3] have sought to automatically detect machine-generated images for forensic purposes. Given that new state-of-the-art generative models emerge frequently, an essential question arises: Can we train a model that generalizes to entirely unseen generative models? This query is explored in several works [11, 54, 67, 61]. Notably, several of these studies [16, 54, 26] have discerned that high-frequency details in both generated and real images can serve as valuable indicators, enabling the development of classifiers that can distinguish between fake and real images up to a certain extent. For example, Wang et al. [54] demonstrated that classifiers, when trained on images generated by a specific GAN, can detect the majority of fake images stemming from other unencountered generative models. Chai et al. [7] devised a patch-based classifier with limited receptive fields, shedding light on the regions of fake images that are more transparently discerned.
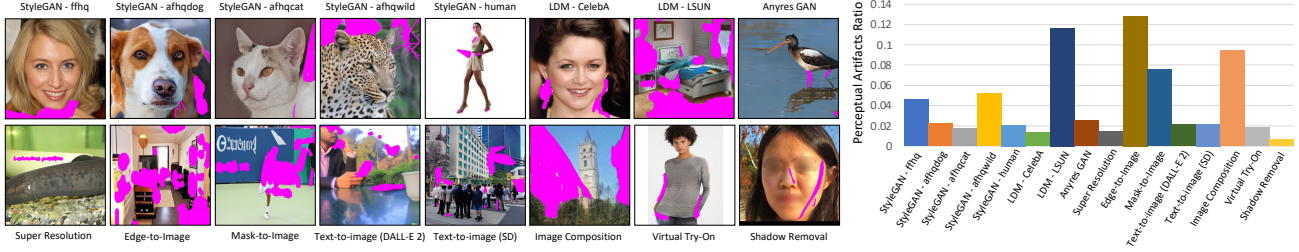
2

Figure 2: The left figure shows the raw labels, where the image order (left to right, and top to bottom) follows the order in the histogram. The histogram demonstrates the Perceptual Artifacts Ratio computed from human labels for different tasks and domains.
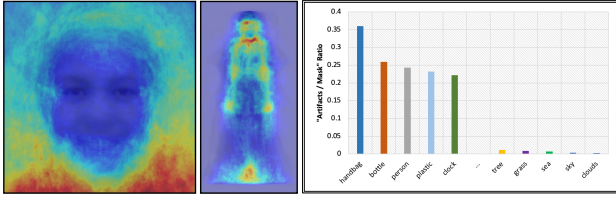

Figure 3: Visualization of the distribution of perceptual artifacts: On the left, for the StyleGAN2-generated [19] images, we observe that perceptual artifacts predominantly concentrate below the chin and around the neck region in facial images, and around the keypoint regions in full-body human images. On the right, we display the five COCO-stuff [6] semantic classes that exhibit the highest and lowest amounts of artifacts for in-the-wild generated images.

More recently, Ojha and Li et al. [37] identified that an earlier classifier [53] had an inclination to overfit to high-frequency noise present in GANs, leading it to erroneously categorize all diffusion-generated images as real. To rectify this, they incorporated frozen pretrained CLIP-ViT features [40], deploying techniques like nearest neighbors or linear probing to strengthen the model's generalization capabilities. Bringing it closer to the context of our work, techniques like Grad-CAM [47] can be employed to visualize the 'active' regions in a real-vs-fake classifier, explicitly highlighting the areas indicative of 'fakeness' from the *perspective of artificial neural networks*.

Localizing Editing Regions. Beyond simply classifying images as real or generated, numerous research efforts have sought to localize the edited regions within the generated or edited images. For example, Wang et al. [53] exploited Photoshop's scripting capabilities to automatically generate edited images. They then effectively trained a model to predict manipulated facial areas in photos. In the field of image inpainting, several studies [58, 23, 56] have demonstrated that high-frequency noise can be sufficiently informative to precisely segment inpainted areas, especially

when training a model alongside the ground truth inpainting mask. While these studies bear some resemblance to our work, their main objective is to pinpoint systematic inconsistencies in generated images to distinguish fakes. In contrast, our research is centered on detecting and segmenting artifact areas that are noticeable to *human perception*. We argue that localizing perceptual artifacts at a fine-grained level, rather than the entire edited region, can pave the way for enhanced generated quality. Notably, Zhang et al. [65] focused on perceptual artifacts in inpainting tasks. Our methodology expands upon this, aiming at various synthesis tasks and their potential downstream applications.

Improving Synthesis Quality During Inference. Several previous works have introduced techniques to enhance image synthesis quality during inference time [28, 5, 12], and our research aligns with this domain. For instance, the truncation trick, initially introduced in BigGAN [5], limits latent code sampling to a constrained space using a specific threshold. This method has been observed to result in improved visual quality of individual samples. Classifier guidance, as detailed in [12] for diffusion-based models, suggests utilizing the gradient of a pretrained classifier to steer the diffusion process in image generation. This ensures that the generated images are accurately recognized by the classifier. Contrasting with these methods, our approach specifically targets the detection and refinement of perceptual artifact regions in images while striving to retain as much of the original content as possible. This differs from the 'hard' sample rejection seen in the truncation trick or the overarching image synthesis guidance provided by classifier guidance. Furthermore, our method neither relies on hyperparameters, such as the truncation threshold, nor requires gradient computation.

3. Methods

3.1. Data Collection and Statistics

We collect our dataset by running inference using the pretrained models from ten different synthesis tasks. Within each task, we might run more than one model or checkpoint

if the model, i,e. StyleGAN2 [19], are trained on multiple domains. Overall, we collect 10,168 images with the per-pixel artifacts segmentation labels by human experts. Each image takes roughly one minute to label by a human expert, and thus, the entire dataset cost ∼170 hours of labor. We split the dataset into a train/test/val set divided as 80%/10%/10%, respectively.

Statistics on Tasks. The area of perceptual artifacts region highly depends on the complexity of image content, the nature of the task, and the performance of the synthesis model. As shown in the left of Figure 2, we can see that some generated images receive larger marked artifacts regions than the others. We compute the Perceptual Artifacts Ratio (PAR), which is simply the labeled artifacts region divided by the image area, to quantify the levels of artifacts for each task. As shown in the histogram of Figure 2, we can see that the tasks like Edge-to-Image [55], Mask-to-Image [55], LDM-LSUN [43], and Image Composition [9] have obviously larger PAR than the others, since the models are generating relatively complex in-the-wild visual content.

Statistics on Content. Perceptual artifacts are more prevalent in certain object categories or semantic parts of images for two primary reasons. Firstly, visual content with significant variations is inherently challenging to generate. Secondly, human perceptual judgments tend to be more sensitive to specific regions. With these considerations, we embarked on a quantitative exploration of the distribution of perceptual artifacts in generated images. As depicted on the left of Figure 3, we calculated the average PAR heatmap for both face and human images. The heatmap for faces reveals that artifact-prone areas primarily fall around and beneath the chin. This is a region where StyleGAN2 [19] often endeavors to generate content with high variance—like microphones, necklaces, and clothing—but struggles to maintain high fidelity. The heatmap for human images points out that artifacts predominantly arise around specific human keypoints, such as the head, neck, hands, and feet—areas to which human perception is especially attuned. Regarding in-the-wild generated images, as seen on the right side of Figure 3, we discern that "object" regions typically exhibit a higher PAR compared to "stuff" regions.

3.2. Segmenting Perceptual Artifacts

Training Segmentation Models. We formulate the localization of perceptual artifacts as a binary semantic segmentation problem. To train a single unified model for detecting generic perceptual artifacts, we use data collected from all tasks. During training, we adopt random cropping and horizontal flipping to augment the dataset. Our model is implemented with a Swin-T [29] backbone, where UperNet [52] serves as the main head and FCN [30] as the auxiliary head. We train the model using a cross entropy loss and optimize it with the AdamW [31] optimizer, with a learning rate of 6×10^{-5}, betas of (0.9, 0.999), and weight decay of 0.01. The models are initialized using the pretrained weights from ADE20K [72]. We observe that it generally takes less than 20,000 iterations or *less than five hours* to converge on 8 NVIDIA A100 GPUs. Note that we do not focus on the the architecture design in this work.

Efficient Adaption to Unseen Models. Generalizing to unseen domains is challenging for deep networks, yet it is necessary for practical usage. Therefore, we also explore how our pretrained artifacts detector could generalize to totally unseen generative models. In the experiment, we find that our pretrained model can detect a reasonable amount of perceptual artifacts in the unseen models. Furthermore, we find that fine-tuning our pretrained model with as few as ten images can quickly and effectively improve segmentation performance. The results are discussed in section 4.2. Fine-tuning converges within 2,000 iterations, which would take *less than 30 minutes*. Labeling 10 images from an unseen method would take *approximately 10 minutes*, making this approach practical. Therefore, our pretrained model allows for fast adaptation to unseen models with roughly one hour of effort, making our work applicable and easily scalable in the future.

3.3. Framing the Inpainting Context by Zoom-In

An important application of our artifacts detector is to automatically fix the perceptual artifacts region in the generated images. To construct this pipeline, a simple approach involves sequentially stacking the pretrained artifacts detector (F) and an inpainting model (G). During inference, we first feed the generated image (I_g) into the artifacts detector (F) to segment the artifacts region. Next, we consider the segmented artifacts mask with slight dilation as the inpainting mask for I_g and process it through the inpainting model (G). The resulting artifacts-corrected image is denoted as I_f, and the overall pipeline can be expressed as $I_f = G(I_g, \phi(F(I_g)))$, where ϕ represents dilation.

Although naive inpainting can effectively fix many perceptual artifacts, we have observed systematic errors in recent diffusion-based inpainting models [43, 41] when generating specific object details such as faces and hands. However, are these models genuinely incapable of generating such details? We conjecture that one potential cause of this error might be incorrect inpainting context. For instance, we have noticed that image generation generally has higher fidelity when human faces or hands are relatively prominent in the image. This could be due to two underlying reasons: 1) photographs or portraits of humans are typically large and centered in images, resulting in dataset bias; and 2) the loss on large object regions tends to be relatively greater than that of small objects, providing stronger feedback to the model during training.

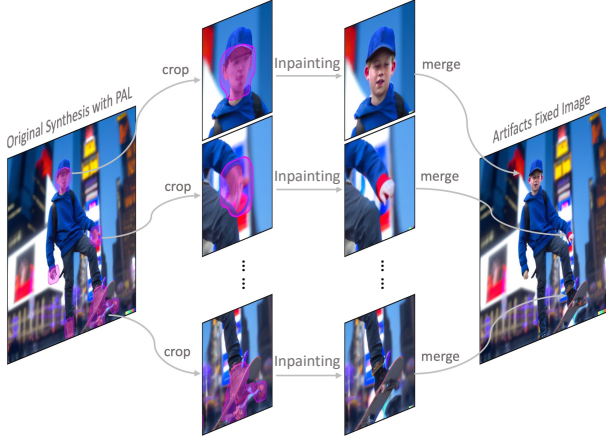Motivated by this realization, we introduce a zoom-in in-

Figure 4: A zoom-in inpainting pipeline refines perceptual artifacts. Starting with the generated image that has predicted perceptual artifacts, we first crop around the artifact regions, using connected components as a guide. We then inpaint these artifact regions within each cropped area and ultimately composite them back into the full image.

painting pipeline that properly frames the input inpainting context before generating the output. As illustrated in Figure 4, we first conduct connected component analysis on the predicted artifacts segmentation mask, then crop around each component of artifacts, perform inpainting on these zoomed-in patches, and finally merge the inpainted patches back into the full image. We empirically set the patch size to be 50% larger than the length of the longest axis of the artifacts mask within each connected component. Remarkably, this simple design significantly enhances artifacts refinement on object details *without modifying the inpainting models*, as demonstrated in section 4.5.

4. Experiments

4.1. Performance on Diverse Image Synthesis Tasks

Our main goal is to develop artifact detectors that can effectively perform on a wide range of generic image syn-

thesis tasks and detect various types of artifacts. In order to achieve this, our first step is to gain a deeper understanding of how existing relevant methods would perform on this particular task. To begin with, we investigate the CNNgenerates [55] classifier, which is specifically designed to distinguish between generated and real images in the context of generic deep generative models. To visualize the gradient activation of the model, we employ Grad-CAM [47], which could reveal regions in the image that are deemed as "fake" by the network. Nonetheless, our findings demonstrate a significant disparity between the model's interpretation and human perception of what constitutes "fake" or artifacts, as presented in the 1st row of Table 1. In addition, we leverage the Patch Forensics model [7] to calculate the "fake" regions based on the patch-based classifier. The results demonstrate that the model's prediction also significantly deviates from human perception, as shown in 2nd row of Table 1. Related to our work, PAL4Inpaint [65] focuses on developing a perceptual artifacts localization method for the inpainting task. However, we observe that a model trained exclusively on inpainting images struggles to generalize to other tasks, as shown in 3rd of Table 1.

The aforementioned observations from previous studies underscore the compelling need for a diverse dataset encompassing multiple tasks and domains to train a generalized artifacts detector. In light of this, we collect a fine-grained labeled dataset spanning ten image synthesis tasks. Subsequently, we train specialized models for each task, as well as a single unified model for all tasks, as shown in the last two rows of Table 1. The unified model confers a memory-efficient advantage from the deployment perspective and performs comparably to the specialist models, with the exception of portrait shadow removal (PSR) task. We believe that this discrepancy may be attributed to the dissimilarity between the artifacts in PSR task and those in other tasks. Nevertheless, our specialist models and unified model both demonstrate significant superior performance in contrast to existing methods.

Methods	StyleGAN2	LDMs	AnyRes	SR	Inpaint	E2I	M2I	T2I	Comp.	VTON	PSR
CNNgenerates [54] + Grad-CAM [47]	4.38	2.43	1.39	0.86	3.54	0.95	0.48	0.51	7.13	2.56	0.0
Patch Forensics [7]	3.81	9.08	8.76	1.34	5.35	14.19	10.54	2.71	9.63	2.14	0.66
PAL4Inpaint [65]	6.12	0.98	1.03	0.81	42.07	0.86	0.31	0.51	14.42	15.94	0.0
Specialist Model (Ours)	37.85	**35.39**	9.15	**14.41**	**42.07**	45.56	35.01	**21.79**	25.31	37.44	**21.33**
Unified Model (Ours)	**38.53**	30.86	**34.74**	11.92	41.81	**46.01**	**39.37**	19.65	**29.53**	**38.07**	5.10

Table 1: Quantitative mIoU ($\uparrow$) evaluation of perceptual artifacts segmentation on 10 image synthesis tasks. We use the following brevity to indicate the tasks: LDMs $\rightarrow$ Latent Diffusion Models [43], SR $\rightarrow$ Super Resolution [57], E2I $\rightarrow$ Edge-to-Image [55], M2I $\rightarrow$ Mask-to-Image [55], T2I $\rightarrow$ Text-to-Image [43, 41], Comp. $\rightarrow$ Image Composition [9], VTON $\rightarrow$ Virtual Try-On [15], PSR $\rightarrow$ Portrait Shadow Removal [66], and finally U.M. $\rightarrow$ Unified Model.

5

4.2. Performance on Unseen Methods

Given the rapidly evolving landscape of generative image models, with novel models emerging on a monthly basis, an ideal artifacts detector should be able to effectively function or swiftly adapt to these untested methods. To evaluate the performance on the unseen methods, we collected additional 500 generated images with labels from two previously unseen GAN-based models and three diffusion-based models. Note that our dataset includes images from Style-GAN2 and Stable Diffusion v1.4, but excludes any images from StyleGAN3 and Stable Diffusion v2.0 (SD2). The other three models are BlobGAN [14], Verstile Diffusion [60] and Diffusion Transformer (DiT) [39], which do not have any counterparts in our training dataset.

We first visualize what the previous and our methods detect as visual artifacts in the generated images from unseen methods. To this ends, we compute the raw heatmap out-

Methods	StyleGAN3 [17]	BlobGAN [14]	SD2 [43]	VD [60]	DiT [39]
CNNgen [54] + [47]	2.30	3.67	0.12	0.57	1.42
Patch Forensics [7]	11.43	5.96	3.08	2.97	3.18
PAL4Inpaint [65]	5.18	13.0	0.85	0.63	1.15
Ours	**46.45**	25.39	6.75	5.92	16.46
Ours w/ Fine-tuning	40.81	**33.33**	**11.04**	**22.18**	**31.76**

Table 2: Quantitative mIoU (↑) evaluation of the binary artifacts segmentation on five unseen models. We use the following brevity to indicate the tasks: CNNgen → CNNgenerates [54], SD2 → Stable Diffusion v2.0 [43], VD → Verstile Diffusion [60], DiT → Diffusion Transformer [39].

puts and compare them in Figure 5. For quantitative comparison in Table 2, the previous methods exhibit poor performance in segmenting the perceptual artifacts in these unseen images. In contrast, our unified model trained on the newly proposed dataset demonstrates reasonable general-
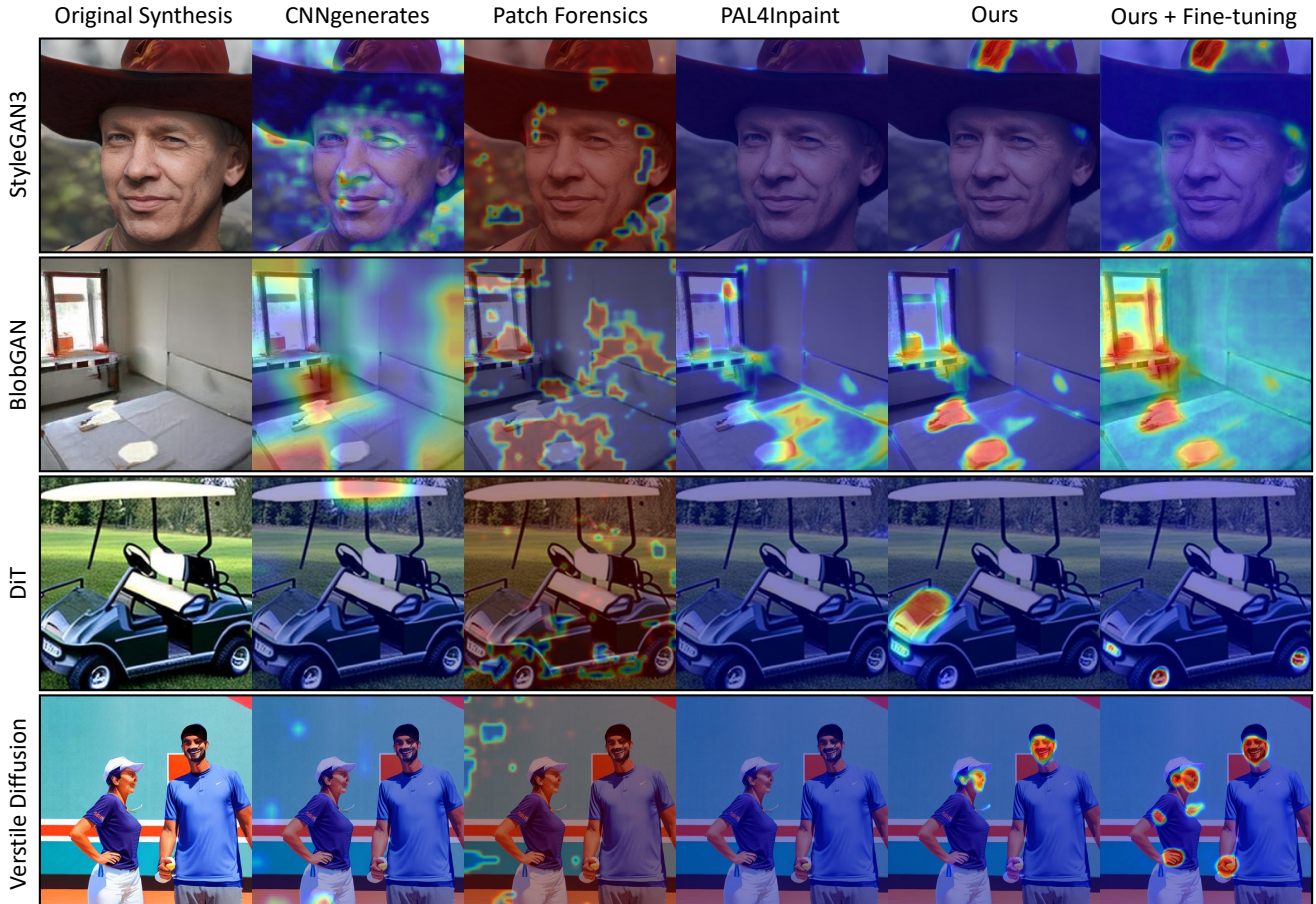

Figure 5: Qualitative comparison of artifacts localization on several unseen methods. We use Grad-CAM [47] to visualize the gradient maps of CNNgenerates [54], and use the pretrained checkpoints from Patch Forensics [7] and PAL4Inpaint [65] to directly compute the heatmap. The results demonstrate that our approaches exhibit a much stronger correlation with human judgement in detecting perceptual artifacts. **Please zoom in the first column to check the perceptual artifacts.**
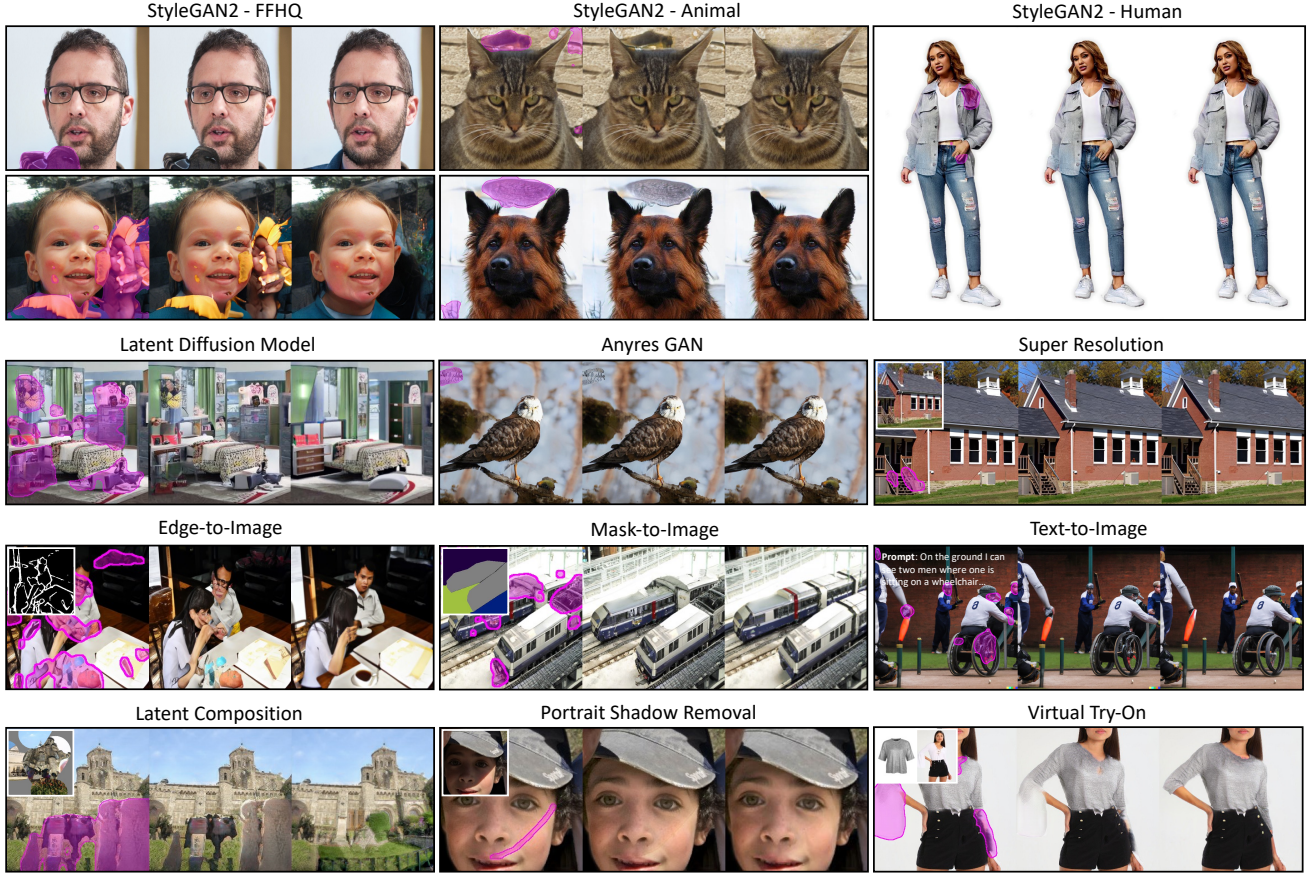
Figure 6: Qualitative results of automatic artifacts fixing on ten image synthesis tasks. In each grid, the left is the overlaid artifacts segmentation, the middle is the original generated image, and the right is the refined image.
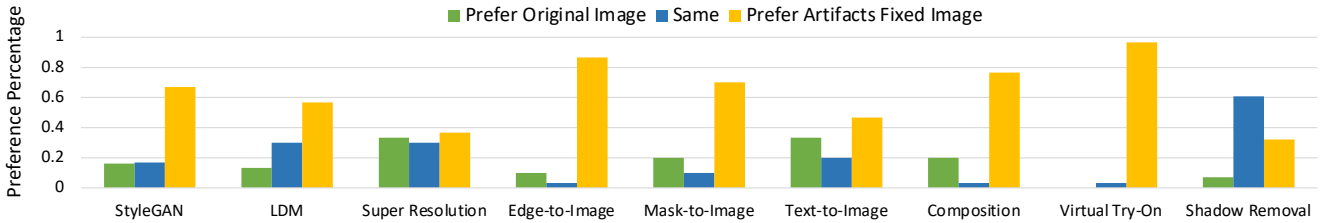

Figure 7: User study to evaluate whether the curated images are better, similar, or worse than the original generated images for diverse synthesis tasks. For each task, we sample 30 images and ask at least five users to vote for each image.

ization ability to these unseen models. Furthermore, fine-tuning our model with a minimum of 10 examples results in effective performance improvement.

4.3. Automatically Fixing Artifacts

An essential downstream application of perceptual artifacts localization is the automatic correction or refinement of artifacts in generated images, as discussed in Section 3.3.

In this section, we provide both qualitative (Figure 6) and quantitative demonstrations of how perceptual artifacts segmentation is capable of effectively correcting a significant portion of the perceptual artifacts in diverse synthesis tasks.

To quantitatively measure the artifacts fixing performance, we conduct a user study to assess whether the artifacts corrected images are better, similar, or worse than the original generated images. For each task, we randomly
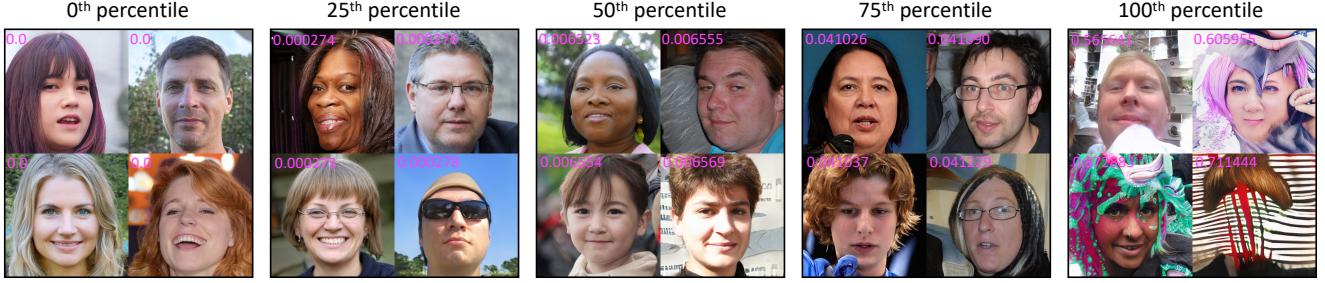
Figure 8: Using Perceptual Artifacts Ratio (PAR) as a metric to rank 4,000 synthesized face image by StyleGAN2. We show samples at different percentile from the PAR ranking, where the actual PAR score is written in pink at the top left corner of each image.
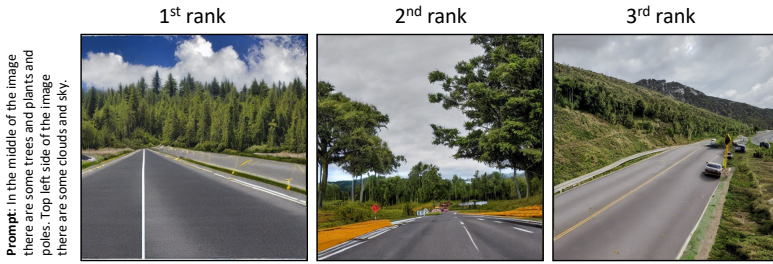

Figure 9: Using Perceptual Artifacts Ratio (PAR) to rank the multimodal outputs of text-to-image synthesis by Stable Diffusion.

Tasks	StyleGAN2 [19]	Stable Diffusion [43]
Random Chance	50.0%	50.0%
HyperIQA [49]	58.51% (+8.51%)	43.30% (-6.70%)
MUSIQ [20]	61.63% (+11.63%)	48.45% (-1.55%)
PAR (Ours)	74.47% (+24.47%)	63.92% (+13.92%)

Table 3: A quantitative study of user agreement with metrics to rank the visual quality between pairs of images in unconditional StyleGAN2 [19] sampling and multimodal text-to-image synthesis with Stable Diffusion [43]. We show comparison with two state-of-the-arts non-referenced IQA methods [43, 20].

select approximately 30 images and solicit feedback from at least five Amazon Turk workers to vote on each pair of images. As depicted in Figure 7, the user preferences demonstrate that our artifacts correction pipeline significantly enhances the visual quality for most of the tasks, and rarely degrades the quality of the generated images. Although we believe that the user study provides a more accurate reflection of perceptual improvement, we also calculate the FID-CLIP [22] of 5,000 Stable Diffusion (SD) [43] text2image images, which improves from 15.98 to 13.28 (+16.9%) after our refinement with SD inpainting [43].

4.4. Perceptual Artifacts as A Quality Metric

Evaluating the quality of generated images remains an ongoing area of research. Among various types of image quality assessment (IQA), no-reference IQA is the most challenging, as there are no reference ground truth images for comparison. In this study, we demonstrate that our artifacts detector can be leveraged to compute a no-reference IQA metric referred to as Perceptual Artifacts Ratio (PAR), which is calculated as the ratio of the perceptual artifacts region to the entire image area. Essentially, a larger PAR value indicates more perceptual artifacts in the image and, consequently, lower visual quality.

We demonstrate the application of the PAR metric to rank unconditional and multimodal image samples. In Figure 8, we showcase how PAR can rank thousands Style-GAN2 face images, where smaller PAR values indicate fewer artifacts and, thus, better visual quality. We extract four images at percentiles of 0th, 25th, 50th, 75th, and 100th, where 0th and 100th percentiles correspond to the least and largest PAR, respectively. Our results indicate that the visual quality gradually deteriorates with increasing percentiles, which is consistent with human perception. Additionally, we demonstrate that the PAR score can help rank the visual quality of multimodal text-to-image outputs generated by Stable Diffusion [43], as illustrated in Figure 9.

Furthermore, we conduct a user study to evaluate user preference agreement with the no-reference IQA metrics for ranking around 100 pairs of images for both StyleGAN2 [19] and Stable Diffusion [41]. The results presented in Table 3 indicate that our PAR metric outperforms two state-of-the-art methods in terms of ranking image quality. In practical use cases, the PAR score can facilitate automatic ranking or filtering of a large batch of candidate images.

4.5. Effect of Zoomed-in Inpainting Context

In section 3.3, we discussed how diffusion-based inpainting models tend to produce better outputs for object details when zoomed-in context is provided. In this section, we aim

8

to quantitatively evaluate the impact of zoom-in inpainting on artifacts refinement performance. It is widely known that diffusion models, such as Stable Diffusion [43] or DALL-E 2 [41], often struggle with generating realistic human faces and hands. Therefore, we investigate how zoom-in inpainting can aid in refining these challenging cases.
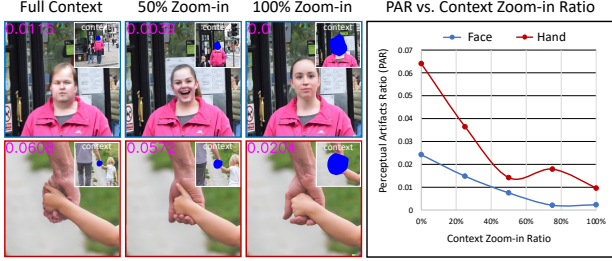

Figure 10: The relationship between perceptual artifacts ratio (PAR) and context zoom-in. In each image of left examples, top right is the input context for inpainting.

As shown in Figure 10, the two examples on the left illustrate that gradually zooming-in the input context leads to the generation of more realistic face and hand pixels. To further support our findings, we employed the PAR score to quantify the relationship between zoomed-in scale and image quality. Our results demonstrate that providing gradually zoomed-in context consistently enhances the quality of generated face and hand pixels over a set of 100 images, as shown on the right of Figure 10. Hence, we can observe that our simple zoom-in inpainting pipeline effectively addresses the artifacts present in these object details.

4.6. Abnormal Detection on Real Images

Since our artifacts detector demonstrates reasonable perceptual artifacts prediction ability on photo-realistic generated images, we are curious if the model would detect anything unusual in real images. As expected, the majority of the real images, which do not contain any artificially generated content, receive no prediction from the artifacts detector. Interestingly, for a small portion of images that receive some prediction, we observe that the predicted perceptual artifacts tend to be on abnormal objects, distractors, or blurry/fuzzy regions, as shown in several examples in Figure 11. For instance, in an FFHQ [18] face image, the artifacts detector finds an "abnormal" object that appears to be a tattooed arm. For other in-the-wild images sampled from commonly used datasets [71, 25, 62], the artifacts detector model also detects artifacts such as watermark text, fuzzy or distractor bedroom corners, and tennis rackets with motion blur, as shown in the right three images of Figure 11.

Figure 11: Inference on real natural images, where the predicted abnormal regions are indicated by the pink contour.

5. Conclusion

In this paper, we present a comprehensive empirical study on perceptual artifacts localization for image synthesis tasks. Firstly, we collected a high-quality dataset comprising 10,168 images with per-pixel artifact labels. Subsequently, we trained a segmentation model to accurately locate the artifacts for ten diverse synthesis tasks, and demonstrated that our pre-trained model can efficiently adapt to unseen methods. Utilizing our learned artifact detector, we explore three downstream applications: 1) automatically refining artifacts in the generated images; 2) evaluating image quality without reference; and 3) detecting abnormal objects in natural images. To address the issue of diffusion-based inpainting generating incorrect content in object details such as faces and hands, we propose a simple zoom-in inpainting pipeline that effectively mitigates this problem.

Future Directions. We view our dataset and benchmark models as foundational for future research in this domain. We identify three primary avenues for advancement: 1) crafting specialized architecture or loss functions for enhanced perceptual artifact segmentation; 2) creating task-specific modules to achieve more granular refinement for each task, moving beyond mere inpainting; and 3) broadening the dataset to encompass varied individual preferences concerning perceptual artifacts in generated images.

Practical Impact. To the best of our knowledge, our research on the localization of perceptual artifacts in generic image synthesis is pioneering. Manually retouching these perceptual artifacts can be both tedious and vexing, especially for professionals. Our automated approach stands to greatly enhance productivity and alleviate this significant burden. Furthermore, our no-reference perceptual artifacts metric can assist users in sifting through and selecting quality content from multimodal generation candidates, such as text-to-image applications [43, 41], potentially boosting user satisfaction. In summary, we hope that our contributions might offer valuable insights and tools for the evolution of practical image editing software in the coming years.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018. 2

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[3] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, and Benedetta Tondi. Cnn detection of gan-generated face images based on cross-band co-occurrences analysis. In *2020 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2020. 2

[4] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, pages 5–10, 2016. 2

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2, 3

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 3

[7] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European conference on computer vision*, pages 103–120. Springer, 2020. 2, 5, 6

[8] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. *arXiv preprint arXiv:2204.07156*, 2022. 1, 14, 18

[9] Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in gans. *arXiv preprint arXiv:2103.10426*, 2021. 1, 4, 5, 20

[10] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *Proceedings of the 5th ACM workshop on information hiding and multimedia security*, pages 159–164, 2017. 2

[11] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 2

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2, 3

[13] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 2

[14] Dave Epstein, Taesung Park, Richard Zhang, Eli Shechtman, and Alexei A Efros. Blobgan: Spatially disentangled scene representations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 616–635. Springer, 2022. 6

[15] Benjamin Fele, Ajda Lampe, Peter Peer, and Vitomir Struc. C-vton: Context-driven image-based virtual try-on network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3144–3153, 2022. 1, 5, 19

[16] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 2

[17] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2, 6

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 9, 14, 21

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1, 2, 3, 4, 8, 13, 16

[20] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 8

[21] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 2

[22] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022. 8

[23] Ang Li, Qiuhong Ke, Xingjun Ma, Haiqin Weng, Zhiyuan Zong, Feng Xue, and Rui Zhang. Noise doesn't lie: towards universal detection of deep inpainting. *arXiv preprint arXiv:2106.01532*, 2021. 3

[24] 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. 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 9

[26] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, pages 95–110. Springer, 2022. 2

[27] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 772–781, June 2021. 2

[28] Yuejiang Liu, Parth Kothari, and Alexandre Alahi. Collaborative sampling in generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4948–4956, 2020. 3

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[32] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[33] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 14

[34] Huaxiao Mo, Bolin Chen, and Weiqi Luo. Fake faces identification via convolutional neural network. In *Proceedings of the 6th ACM workshop on information hiding and multimedia security*, pages 43–47, 2018. 2

[35] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, Amit K Roy-Chowdhury, and BS Manjunath. Detecting gan generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*, 2019. 2

[36] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[37] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 3

[38] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 6

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2, 4, 5, 8, 9, 13, 14, 19, 21, 22

[42] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 2

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 4, 5, 6, 8, 9, 13, 14, 17, 21

[44] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 2

[45] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[47] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3, 5, 6

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 14

[49] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 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*, pages 3667–3676, 2020. 8

[50] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022. 2

[51] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor

Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 2, 14, 21

[52] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 4

[53] Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10072–10081, 2019. 3

[54] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 2, 5, 6

[55] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 1, 2, 4, 5, 18

[56] Xinyi Wang, Shaozhang Niu, and He Wang. Image inpainting detection based on multi-task deep learning network. *IETE Technical Review*, 38(1):149–157, 2021. 3

[57] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 1, 5, 17

[58] Haiwei Wu and Jiantao Zhou. Iid-net: Image inpainting detection network via neural architecture search and attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1172–1185, 2021. 3

[59] Li Xu, Qiong Yan, Yang Xia, and Jiaya Jia. Structure extraction from texture via relative total variation. *ACM transactions on graphics (TOG)*, 31(6):1–10, 2012. 14

[60] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022. 6

[61] Xinsheng Xuan, Bo Peng, Wei Wang, and Jing Dong. On the generalization of gan image forensics. In *Chinese conference on biometric recognition*, pages 134–141. Springer, 2019. 2

[62] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 9

[63] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[64] Lingzhi Zhang, Connelly Barnes, Kevin Wampler, Sohrab Amirghodsi, Eli Shechtman, Zhe Lin, and Jianbo Shi. Inpainting at modern camera resolution by guided patchmatch with auto-curation. In *European Conference on Computer Vision*, pages 51–67. Springer, 2022. 2, 21

[65] Lingzhi Zhang, Yuqian Zhou, Connelly Barnes, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for inpainting. *arXiv preprint arXiv:2208.03357*, 2022. 3, 5, 6, 13

[66] Xuaner Zhang, Jonathan T Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E Jacobs. Portrait shadow manipulation. *ACM Transactions on Graphics (TOG)*, 39(4):78–1, 2020. 1, 5, 20

[67] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019. 2

[68] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2185–2194, June 2021. 2

[69] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 2, 14, 21

[70] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Image inpainting with cascaded modulation gan and object-aware training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 277–296. Springer Nature Switzerland Cham, 2022. 2, 14, 21

[71] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 9

[72] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 4

Supplementary Materials

A. Details on Data Labeling

We have discussed about the overall data labeling and statistics in the main paper. Here, we added more details regarding the data labeling.

Labeling Criterion Labeling perceptual artifacts is a highly subjective task, and therefore different workers may have varying opinions on which regions should be considered as 'artifacts'. We instruct the workers to keep a specific criterion in mind while labeling, which is to imagine that we have a perfect artifact-fixing model that can correct any marked region. Hence, if a worker believes that any region in the image can be enhanced or refined, they should mark those regions accordingly.

User Interface We use a labeling interface similar to the one used in [65], where we duplicate the generated image and stitch the copies side-by-side. During labeling, workers can identify perceptual artifacts on the right side of the image and refer to the 'unmarked' image on the left side as a reference.

Visualization of Labels We show more visualization of the perceptual artifacts labels in Fig 13.

B. Implementation Details

Training Details of PAL Models We implement our PAL model using the Swin-L as the backbone, UperNet as the head with a loss weight of 1.0, and a FCN auxiliary head with a loss weight of 0.4. During training, we use random crop with max cutout ratio of 0.75, and random flip with a probability of 0.5. Our code implementation is based on MMSegmentation *.

Code Resources for Data Generation We use the official github repos to generate the images for each synthesis tasks. The StyleGAN 2 images from domain ffhq, afhq-dog, afhqcat, and afhqwild are generated using the official NVIDIA StyleGAN repo †. For StyleGAN 2 human images, we use the StyleHuman repo ‡. For unconditional generation with Latent-Diffusion Model (LDM), we use the code from §. We generate Anyres GAN images using ¶, and super resolution with Real-ESRGAN ‖. For Edge-to-Image and Mask-to-Image, we use the same diffusion-based model PITI ** but different checkpoints. For DALL-E 2 text-to-image synthesis and inpainting, we use the OpenAI API

††. For Stable Diffusion, we use the v1.4 checkpoint from ‡‡ for text-to-image synthesis, and v1.5 checkpoint from §§ for inpainting. We use official latent composition repo ¶¶ for image composition synthesis. Finally, we directly use the synthesized images from repo *** for virtual try-on task and repo ††† for portrait shadow removal. For other inpainting models used in artifacts fixing pipeline, we use official LaMa github repo ‡‡‡, and official CoMod-GAN github repo §§§.

Prompt for Text-based Inpainting Text-based diffusion inpainting requires additional text prompt besides the image and mask inputs. In this section, we discuss how we decide the fixed text prompts for each type of generated images. Generally, we use "a person's face" as the text prompt for all facial images generated by StyleGAN2 [19] and LDM [43], and use "a person" for all human images in StyleGAN2 human and virtual try-on task. For LDM LSUN bedroom images, we just use "bedroom" as the text prompt. For the rest of in-the-wild images, we use "photograph of a beautiful empty scene, highest quality settings" as the fixed text prompt, which is the default option used in Stable Diffusion inpainting.

Selecting Multimodal Outputs As text-based inpainting models, i.e. DALL-E 2 [41], have multimodal outputs, we select the final output image based on the Perceptual Artifacts Ratio (PAR), which has some correlation with human judgement as described in section 5 of the main paper. Specifically, suppose we have N multimodal outputs, we denote the candidate images as I_i, where $i = 1, ..., N$. We compute the PAR scores for each image, which is denoted as $PAR(I_i)$. The finally selected output image is determined by $\mathrm{argmin}_i PAR(I_i)$.

C. Statistical Analysis of User Study

We conduct user studies to evaluate whether the artifacts fixed images are better, same, or worse the original generated images. We perform statistical hypothesis testing using a null hypothesis that the mean of preferences is zero, where the preference is -1 if the original image was preferred, 0 if no preference, and +1 if the artifacts-fixed image was preferred. We use a one sample permutation t test with 10^6 permutations. If we combine all user votes into a single list, the null hypothesis is rejected with $p = 0$. If we run a test per task, using Holm-Bonferroni correction and a familywise error rate of 0.05, we find the null hypothesis is

---

*mmsegmentation: https://github.com/open-mmlab/mmsegmentation

†stylegan: https://github.com/NVlabs/stylegan3

‡StyleGAN-Human: https://github.com/stylegan-human/StyleGAN-Human

§latent-diffusion: https://github.com/CompVis/latent-diffusion

¶anyres-gan: https://github.com/chail/anyres-gan

‖Real-ESRGAN: https://github.com/xinntao/Real-ESRGAN

**PITI: https://github.com/PITI-Synthesis/PITI

††dalle-api: https://openai.com/api/

‡‡stable-diffusion-v1.4: https://github.com/CompVis/stable-diffusion

§§stable-diffusion-v1.5: https://github.com/runwayml/stable-diffusion

¶¶latent-composition: https://github.com/chail/latent-composition

***c-vton: https://github.com/benquick123/C-VTON

†††portrait-shadow-manipulation: https://github.com/google/portrait-shadow-manipulation

‡‡‡lama: https://github.com/saic-mdal/lama

§§§co-mod-gan: https://github.com/zsyzzsoft/co-mod-gan

rejected for every task except super-res, text-to-image, and shadow removal. This indicates that for 6 out 10 tasks and for the combination of all user votes across tasks, there is a significant preference, which per our data is the artifacts fixed image.

D. More Qualitative Results

In this section, we show more visualization results.

D.1. PAL and Artifacts Fixed Results

We show more qualitative results of perceptual artifacts segmentation and artifacts fixed results for ten synthesis tasks. These visual results are shown in Fig 14 - 23. In each example, first image is the generated image with perceptual artifacts localization (PAL), which is indicated by the pink mask. The second image is the original generated image, and the third is the corresponding artifacts fixed image using the predicted PAL. We put the original and artifacts refined images side-by-side for more direct visual comparison.

D.2. The Choices of Inpainting Models

In this paper, we mainly use CoMod-GAN [69], LaMa [51], and DALL-E 2 inpainting [41] in our artifacts fixing pipeline, as discussed in section 4 in the paper. In this section, we show ablation studies on how different inpainting models can be used to fix the perceptual artifacts in different cases. As shown in Figure 25, for face inpainting, CoMod-GAN trained on FFHQ [18] face dataset produce more realistic results than the CM-GAN [70], and has similar performance to DALL-E 2 inpainting. Since CoMod-GAN has faster inference speed than DALL-E 2 by a order of magnitude, we choose CoMod-GAN for general face inpainting cases. For other in-the-wild inpainting cases, as shown in Figure 24, we observe that GAN-based models LaMa and CM-GAN have reasonably good performance on the relatively easy cases, such as the first two rows. However, when the images are under perspective (3rd row) or involve object completion (4th and 5th rows), diffusion-based models generally produce much better results. Within diffusion-based models, DALL-E 2 produce much more realistic details than Stable Diffusion inpainting [43] with v1.5 checkpoint. Therefore, we use LaMa for the easy background inpainting in tasks like Anyres GAN [8], and DALL-E 2 for the rest of tasks with complex scene or object completion.

D.3. Zoom-in Effect on Inpainting

In the main paper, we discuss that diffusion-based models, i.e. DALL-E 2 [41], systematically struggles to generate high-fidelity object details, such as faces and hands. Here, we show more qualitative results. Inspired by this insight, we further propose a 'zoom-in' inpainting pipeline that can fix the perceptual artifacts in the object detail level.

As show in Figure 27, we can see that this zoom-in inpainting pipeline can significantly refine the object details and outperform naively inpainting using the full images and masks. More detailed comparisons on hands and faces are illustrated in Figure 26. In this work, we use the fixed text prompt for all the patches, but more tailored text prompt for the individual cropped patch should theoretically improve the visual quality, which we leave as future work.

E. SDEdit for Perceptual Artifacts Fixing

Using inpainting methods to fix the perceptual artifacts might not be ideal for certain synthesis tasks, since it could change too much of the original generated image identity. We also explore an alternative approach SDEdit [33], which enables stroke-based editing using a diffusion model generative prior DDIM [48]. In the implementation, we convert the pixels in the perceptual artifacts region into stroke painting by RTV smooth algorithm [59], and then run SDEdit to re-generate pixels in the artifacts region. As shown in Figure 12, SDEdit preserves more image identity with respect to the original generation, but underperforms DALL-E 2 inpainting [41] in terms of realism. SDEdit also has a hyperparameter that controls the tradeoff between realism and faithfulness (identity preservation), and this can be adjusted for different tasks. In this work, we showcase the usage of SDEdit with DDIM trained on LSUN Church dataset. To apply this in the wild, we might either need to re-train DDIM in larger diverse dataset or integrate SDEdit with other diffusion-based models, i.e. Stable Diffusion [43], and we leave this as future work.
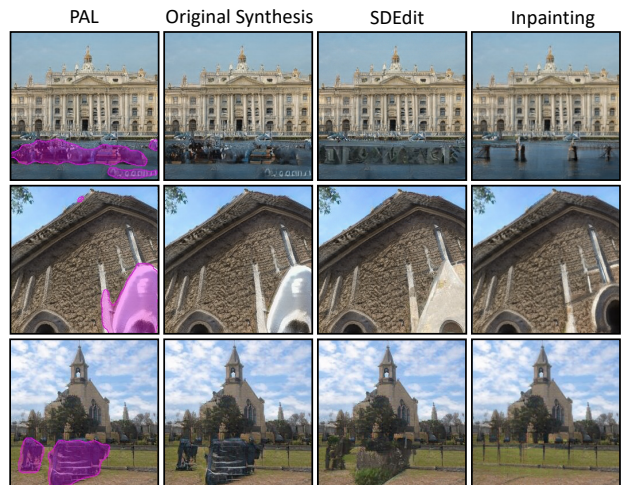

Figure 12: Qualitative comparison between SDEdit [33] and DALL-E 2 inpainting [41] for artifacts fixing. In general, we can see that SDEdit preserves more image identity (more similar to the original synthesis), while DALL-E 2 inpainting produces better realism.
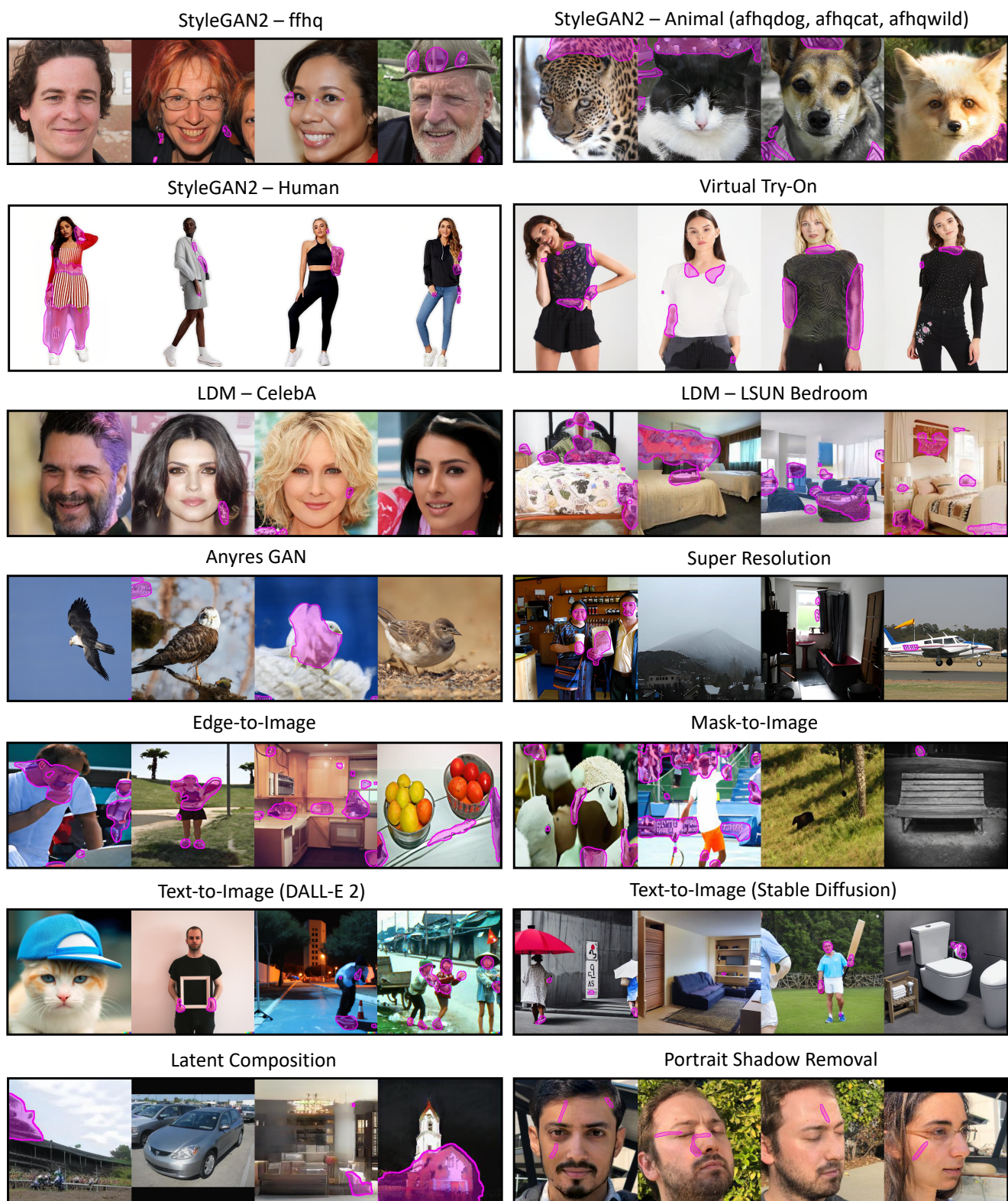
Figure 13: A sampled visualization of our labeled perceptual artifacts dataset in diverse synthesis tasks and domains. Note that if there is no mask in the image, it indicates that workers do not think there are any artifacts in the generated image.
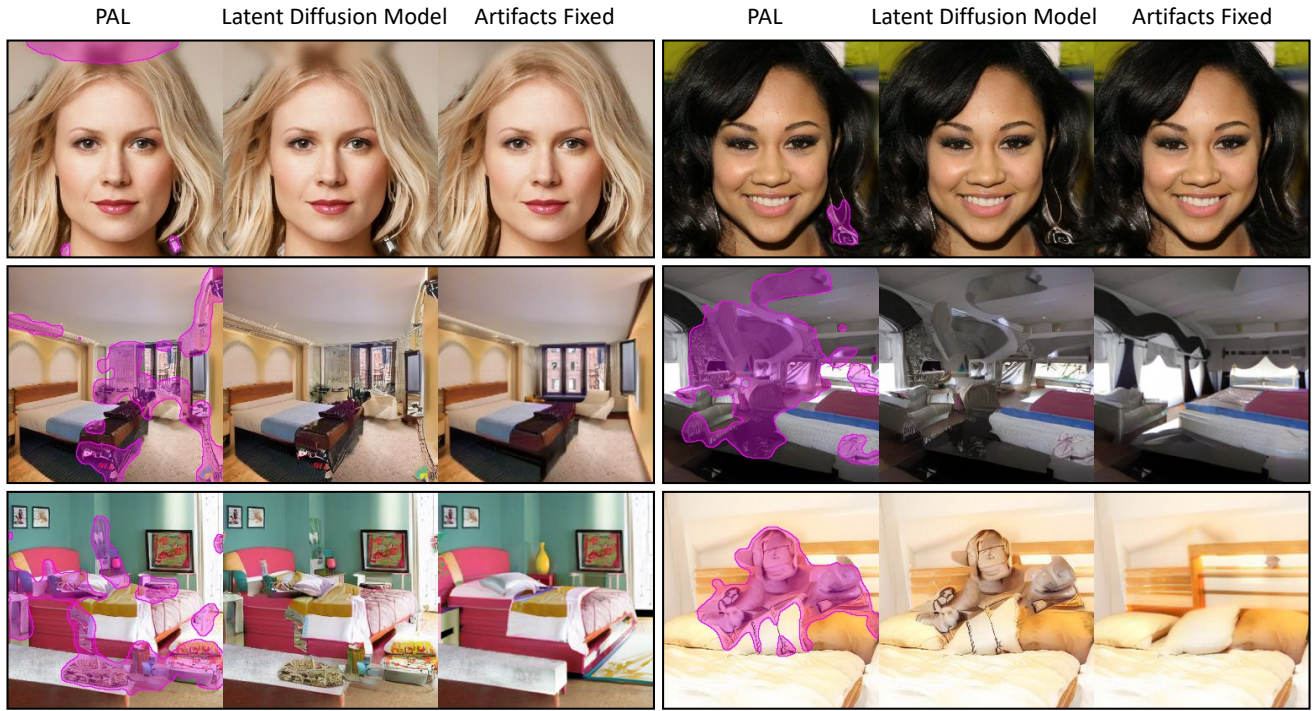
Figure 14: More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for StyleGAN [19]. **Left**: original generated image with PAL prediction. **middle**: original generated image. **right**: artifacts fixed/refined generated image.

Figure 15: More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Latent Diffusion Model [43]. **Left**: original generated image with PAL prediction. **middle**: original generated image. **right**: artifacts fixed/refined generated image.
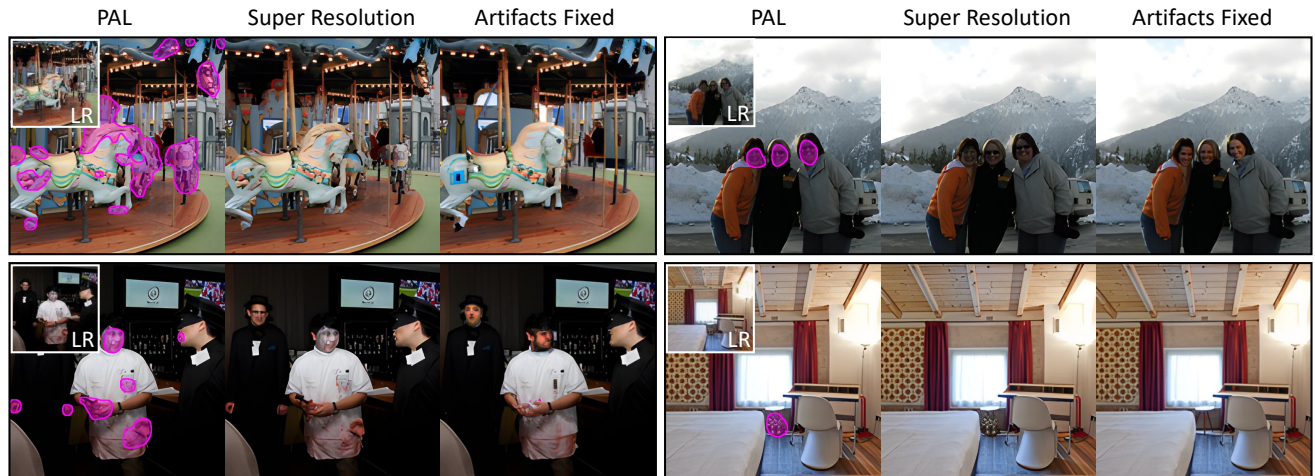

Figure 16: More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for super resolution with Real-ESRGAN [57].
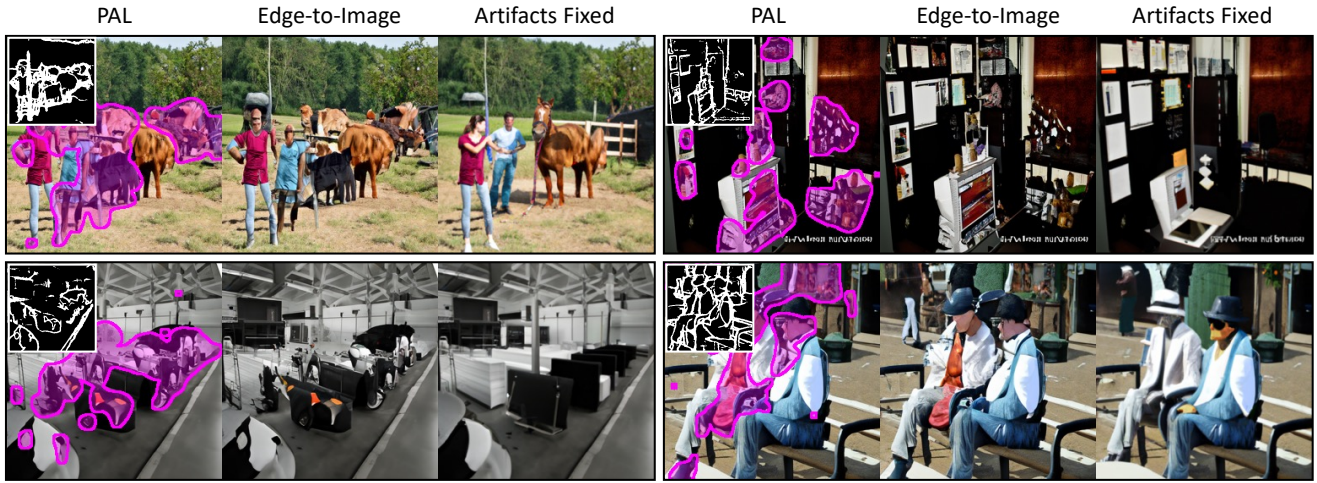
Figure 17: More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Edge-to-Image translation with PITI [55].
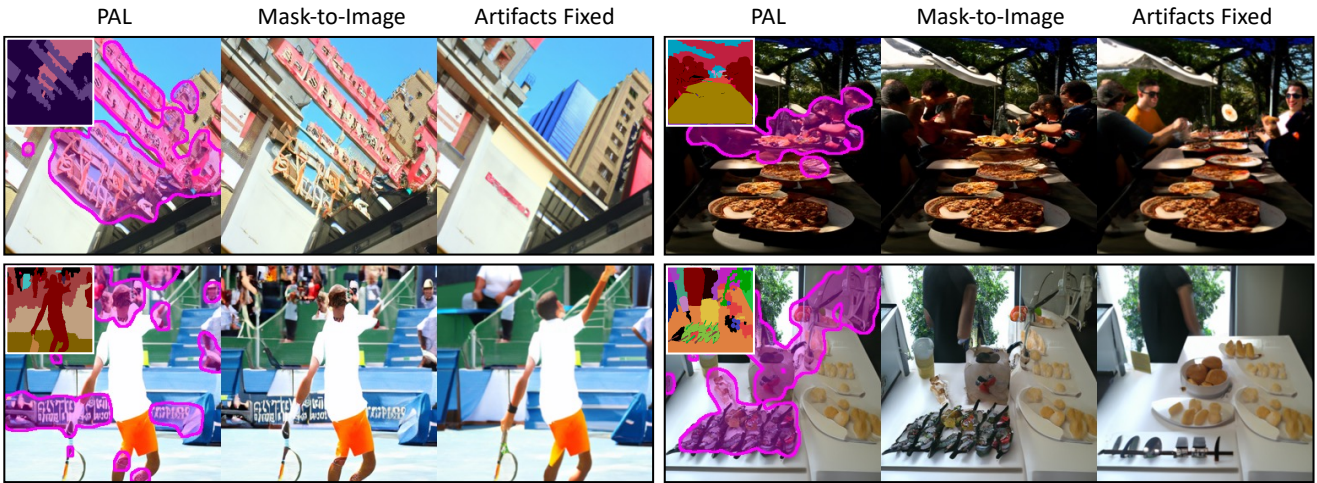

Figure 18: More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Mask-to-Image translation with PITI [55].
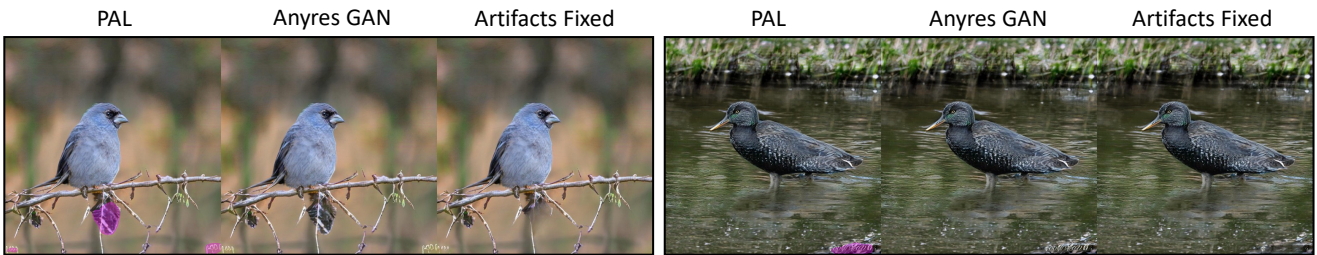

Figure 19: More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Anyres GAN [8].

PAL   Text-to-Image   Artifacts Fixed

PAL   Text-to-Image   Artifacts Fixed

Prompt: In this picture we can see a bus , number and a registration plate on it.

Prompt: In this picture we can see big image on a board, in front there are people in crane setting that image, top there is person watching that.

Prompt: Here there is sofa, table, and lamp. On the wall there are photo frames and there is a window, there is a desk in which television is present.

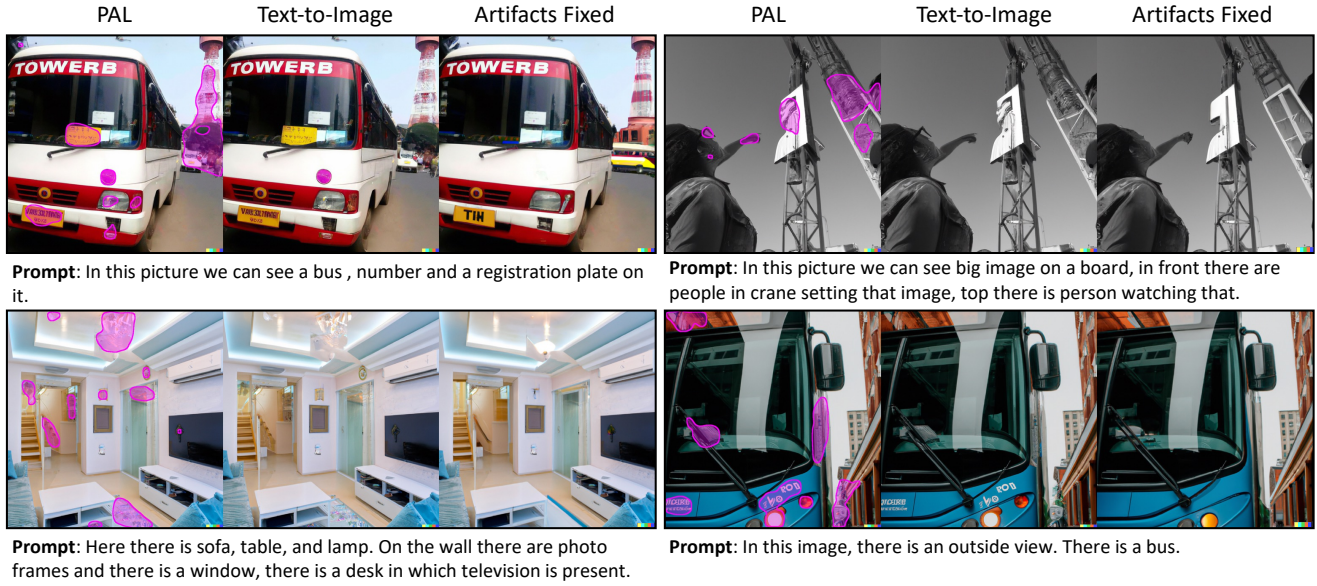**Prompt**: In this image, there is an outside view. There is a bus.

Figure 20: More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Text-to-Image synthesis with DALL-E 2 [41].

Input   PAL   VTON   Artifacts Fixed   Input   PAL   VTON   Artifacts Fixed

Figure 21: More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for virtual try-on with [15].
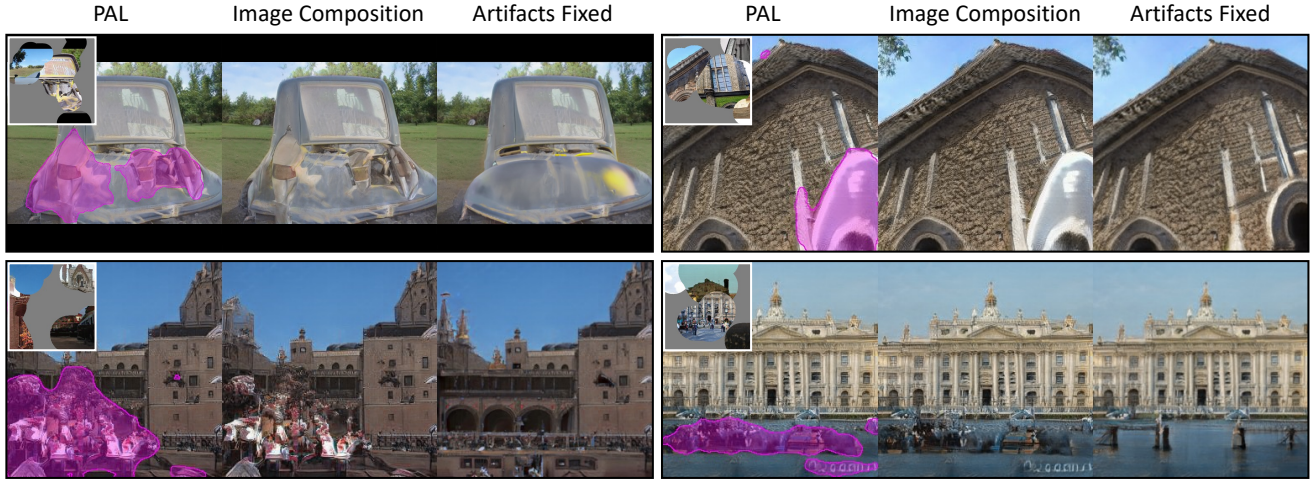
Figure 22: More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for latent composition [9].
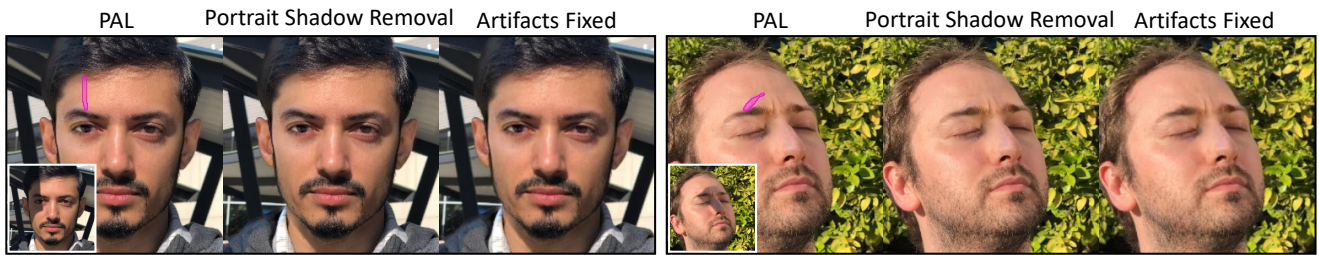

Figure 23: More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Portrait Shadow Removal [66]. Please *zoom in* to see the detailed comparisons.
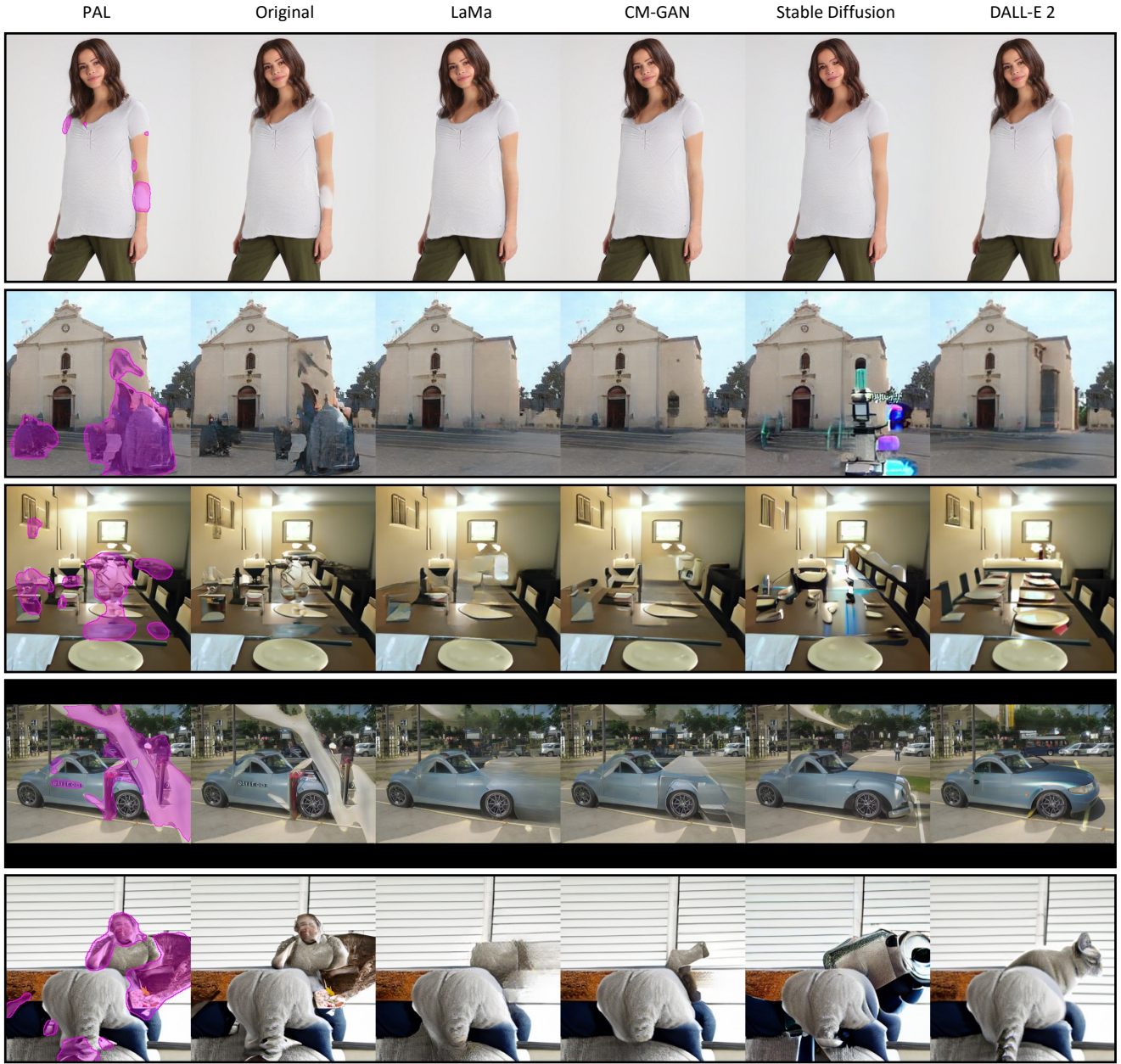
Figure 24: An ablation study on how four different state-of-the-arts inpainting models, including LaMa [51], CM-GAM [64], Stable Duffion [43], and DALL-E 2[41], could fix the perceptual artifacts in types of generated images using our PAL prediction as the inpainting masks.

Figure 25: An ablation study on how inpainting models work on face artifacts removal. Note that CM-GAN [70] and DALL-E 2 [41] are not tailored for face inpainting, while CoMod-GAN [69] is trained on the FFHQ [18] dataset for face inpainting specifically.
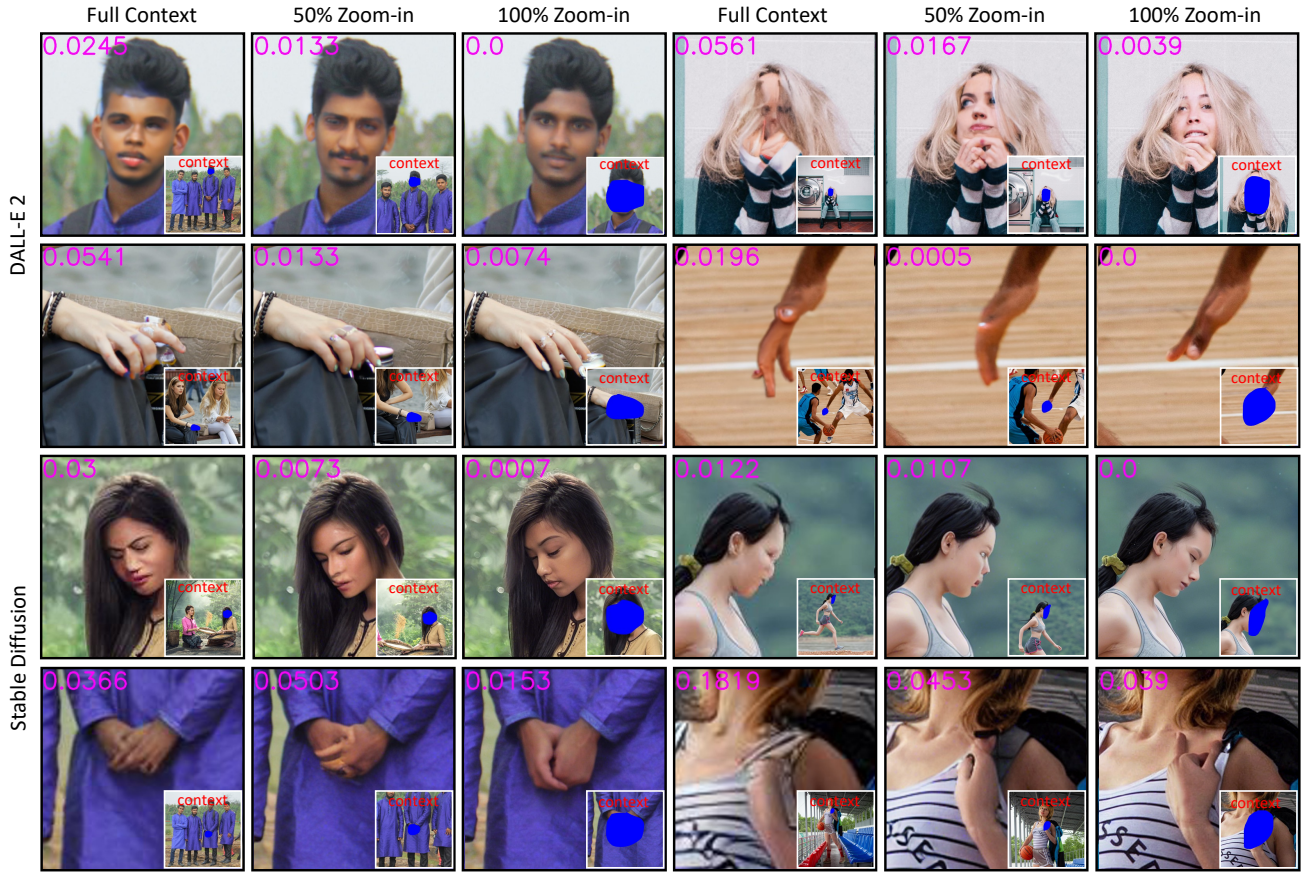
Figure 26: More qualitative results showing that DALL-E 2 inpainting [41] and Stable Diffusion [41] tend to generate less perceptual artifacts when zooming in around the object region, such as faces and hands. We show that our PAR scores, which are placed at the top left corner of the images, can be used to quantify this observation and confirm our insight.

Text Prompt: A boy at age of 10 on the skateboard in time square.

Text Prompt: A girl at age of 10 on the skateboard in time square.

Figure 27: Qualitative comparison between naive inpainting and zoom-in inpainting for fixing perceptual artifacts in text-to-image outputs. In the above examples, we use DALL-E 2 [41] for both text-to-image generation and inpainting. Naive inpainting could fix certain artifacts compared to the original synthesis, but still struggles to generate high-fidelity object details. In contrast, zoom-in inpainting pipeline produces much more realistic object details.