Palette: Image-to-Image Diffusion Models

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet* and Mohammad Norouzi *
{sahariac, williamchan, davidfleet, mnorouzi}@google.com
Google Research

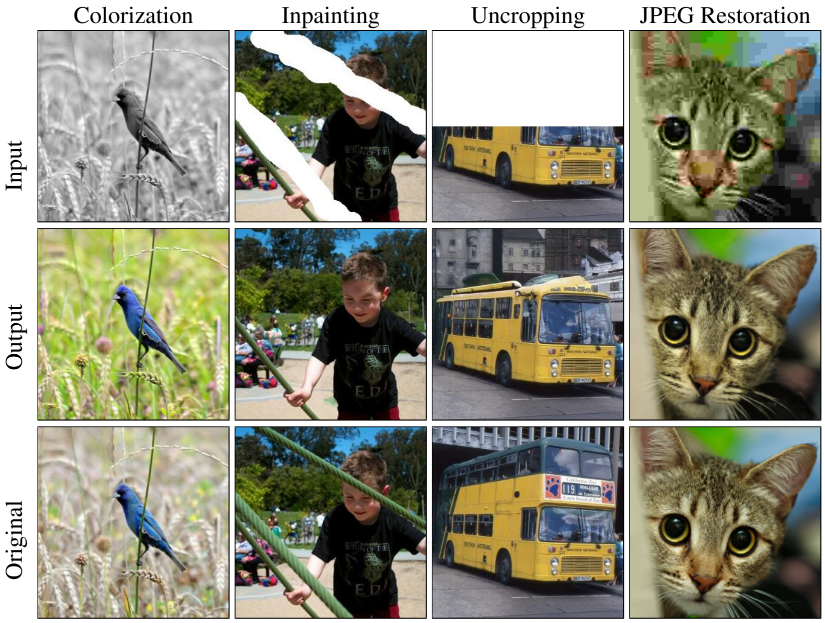

Figure 1: Palette models trained on four challenging image-to-image translation tasks are able to generate high-fidelity outputs. No task-specific customization, tuning, or auxiliary loss is used.

Abstract

We introduce Palette, a simple and general framework for image-to-image translation using conditional diffusion models. Palette models trained on four challenging image-to-image translation tasks (colorization, inpainting, uncropping, and JPEG restoration) outperform strong GAN and regression baselines and bridge the gap with natural images in terms of sample quality scores. This is accomplished without task-specific hyper-parameter tuning, architecture customization, or any auxiliary loss, demonstrating a desirable degree of generality and flexibility. We uncover the impact of an L_2 vs. L_1 loss in the denoising diffusion objective on sample diversity, and demonstrate the importance of self-attention through empirical architecture studies. Importantly, we advocate a unified evaluation protocol based on ImageNet, with human evaluation and sample quality scores (FID, Inception Score, Classification Accuracy of a pre-trained ResNet-50, and Perceptual Distance against original images). We expect this standardized evaluation protocol to play a critical role in advancing image-to-image translation research. Finally, we show that a generalist, multi-task Palette model performs as well or better than task-specific specialist counterparts. Check out https://bit.ly/palette-diffusion for more details.

*Equal Contribution

NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications

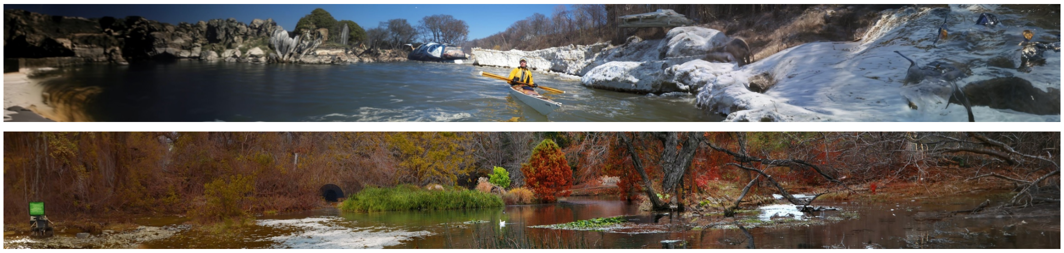

Figure 2: Palette panorama uncropping. Given the central 256×256 pixels, we extrapolate to the left and right in steps of 128 pixels (2×8 applications of 50% Palette uncropping), to generate the final 256×2304 panorama. See Fig. C.14, C.15 in the Appendix for more samples.

1 Introduction

Many problems in vision and image processing can be formulated as image-to-image translation. Examples include restoration tasks, like super-resolution, colorization, and inpainting, as well as pixel-level image understanding tasks, such as instance segmentation and the estimation of intrinsic images. Many such tasks, like those in Fig. 1, are complex inverse problems, where multiple output images are consistent with a single input. A natural approach to image-to-image translation is to learn the conditional distribution of output images given the input, using deep generative models that can capture multi-modal distributions in the high-dimensional space of images.

Generative Adversarial Networks (GANs) [Goodfellow et al., 2014, Radford et al., 2015] have emerged as the model family of choice for many image-to-image tasks [Isola et al., 2017a], as they are capable of generating high fidelity outputs, are broadly applicable, and support efficient sampling. Nevertheless, GANs can be challenging to train Arjovsky et al. [2017], Gulrajani et al. [2017], and often drop modes in the output distribution Metz et al. [2016], Ravuri and Vinyals [2019]. Autoregressive Models [van den Oord et al., 2016, Parmar et al., 2018], VAEs [Kingma and Welling, 2013, Vahdat and Kautz, 2020], and Normalizing Flows [Dinh et al., 2016, Kingma and Dhariwal, 2018] have also seen success in specific applications, but arguably, have not established the same level of sample quality and generality as GANs.

Diffusion and score-based models [Sohl-Dickstein et al., 2015, Song and Ermon, 2020, Ho et al., 2020] have received a surge of recent interest Cai et al. [2020], Song et al. [2021], Hoogeboom et al. [2021], Vahdat et al. [2021], Kingma et al. [2021], Austin et al. [2021], resulting in several key advances in modeling continuous data. On speech synthesis, diffusion models have achieved human evaluation scores on par with SoTA autoregressive models [Chen et al., 2021a,b, Kong et al., 2021]. On the class-conditional ImageNet generation challenge, they have outperformed the strongest GAN baselines in terms of FID scores [Dhariwal and Nichol, 2021, Ho et al., 2021]. On image super-resolution, they have delivered impressive face enhancement results, outperforming GANs [Saharia et al., 2021a]. Despite these results, diffusion models have not been applied to a broad family of tasks, and it is not clear whether they can rival GANs in offering a versatile and general solution to the problem of image-to-image translation.

This paper investigates the general applicability of *Palette*, our implementation of image-to-image diffusion models, to a suite of distinct and challenging tasks, namely colorization, inpainting, uncropping (*a.k.a.* outpainting or extrapolation), and JPEG restoration (*a.k.a.* JPEG artifact removal) (see Fig. 1, 2). Our empirical results suggest that Palette, with no task-specific architecture customization, nor changes to hyper-parameters or the loss, delivers high-fidelity outputs across all four tasks. It outperforms task-specific baselines and our own strong regression model, which uses an identical neural architecture. Importantly, we show that a single *generalist* Palette model, trained on colorization, inpainting and JPEG restoration, outperforms a task-specific JPEG restoration model and achieves competitive performance on the other tasks.

We study key elements of Palette, including the loss function and the neural net architecture. We find that while L_2 [Ho et al., 2020] and L_1 [Chen et al., 2021a] losses in the denoising objective yield similar sample-quality scores, L_2 leads to a higher degree of diversity in model samples, whereas L_1 [Chen et al., 2021a] produces more conservative outputs. We also find that removing self-attention layers from the U-Net architecture of Palette, to build a fully convolutional model, hurts performance.

Finally, we advocate a standardized evaluation protocol for inpainting, uncropping, and JPEG restoration based on ImageNet [Deng et al., 2009] and report sample quality scores for several baselines. We expect this benchmark to help advance image-to-image translation research. In summary:

1. We demonstrate the versatility and applicability of conditional diffusion models to image-to-image translation. Palette achieves SoTA in colorization and outperforms strong GAN and regression baselines in inpainting, uncropping, and JPEG restoration. Palette outputs bridge the gap with natural images in sample quality scores, without using any task-specific tuning or customization.
2. We recognize the impact of the L_1 *vs.* L_2 loss in the denoising objective of diffusion models on sample diversity. We also demonstrate the importance of self-attention for these translation tasks.
3. We propose a standardized evaluation protocol for image translation tasks based on ImageNet and report sample quality scores for multiple baselines including a strong regression baseline and prior work for each task.
4. We show that a multi-task Palette model performs as well or better than task-specific counterparts.

2 Related work

Our work is inspired by Pix2Pix [Isola et al., 2017a], which explored myriad of image-to-image translation tasks with GANs. Other GAN-based techniques have been proposed for various image-to-image problems such as unpaired translation [Zhu et al., 2017a], unsupervised cross-domain generation [Taigman et al., 2016], multi-domain translation [Choi et al., 2018], few shot translation [Liu et al., 2019] and many more. Nevertheless, existing GAN models are sometimes unsuccessful in holistically translating images with consistent structural and textural regularity. Diffusion models [Sohl-Dickstein et al., 2015] recently emerged with impressive results on image generation [Ho et al., 2020, 2021, Dhariwal and Nichol, 2021], audio synthesis [Chen et al., 2021a, Kong et al., 2020], and image super-resolution [Saharia et al., 2021a, Kadkhodaie and Simoncelli, 2021], as well as unpaired image-to-image translation [Sasaki et al., 2021] and image editing [Meng et al., 2021, Sinha et al., 2021]. Our conditional diffusion models build on these recent advances, showing versatility on a suite of image-to-image translation tasks.

Early **inpainting** approaches Bertalmio et al. [2000], Barnes et al. [2009], He and Sun [2012], Hays and Efros [2007], Roth and Black [2005] work well on textured regions but often fall short in generating semantically consistent structure. GANs are now widely used, but often require auxiliary objectives on structures, context, edges, contours and hand-engineered features Iizuka et al. [2017], Yu et al. [2018a, 2019], Nazeri et al. [2019], Yi et al. [2020], Liu et al. [2020], Kim et al. [2021b], and a lack of stochasticity and diversity in their outputs has been observed [Zheng et al., 2019, Zhao et al., 2021]. Generic diffusion models for image generation can be used for linear inverse tasks like inpainting [Sohl-Dickstein et al., 2015, Song et al., 2020, Kadkhodaie and Simoncelli, 2021], but we posit that conditional models trained for inpainting will be superior. **Image uncropping** (a.k.a. outpainting) is considered more challenging than inpainting as it entails generating open-ended content with less context. Early methods relied on retrieval [Kopf et al., 2012, Wang et al., 2014, Shan et al., 2014]. GAN-based methods are now predominant Teterwak et al. [2019], but are often domain-specific Yang et al. [2019b], Bowen et al. [2021], Wang et al. [2019a], Cheng et al. [2021], Lin et al. [2021]. We show that conditional diffusion models trained on large datasets reliably address both inpainting and uncropping on a wide range of image domains.

Colorization is a well-studied image-to-image task Kumar et al. [2021], Guadarrama et al. [2017], Royer et al. [2017], Ardizzone et al. [2019], requiring a degree of scene understanding, which makes it a natural choice for self-supervised learning [Larsson et al., 2016]. Challenges include diverse colorization [Deshpande et al., 2017], respecting semantic categories [Zhang et al., 2016], and producing high-fidelity color Guadarrama et al. [2017]. While some prior work makes use of specialized auxiliary classification losses, we find that generic image-to-image diffusion models work well without task-specific specialization. **JPEG restoration** (aka. JPEG artifact removal) is a nonlinear inverse problem for restoring realistic textures and removing compression artifacts. Dong et al. [2015] applied deep CNN architectures for JPEG restoration, and Galteri et al. [2017, 2019] successfully applied GANs for artifact removal, but they have been restricted to quality factors of above 10. We show the effectiveness of Palette in removing compression artifacts for quality factors as low as 5.

Multi-task training is a relatively under-explored area in image-to-image translation. A few papers [Qian et al., 2019, Yu et al., 2018b] have focused on simultaneous training over multiple tasks, but they are primarily focused on enhancement tasks like deblurring, denoising, and super-resolution, and they use smaller, modular networks. Several works have also dealt with simultaneous training over multiple degradations on a single task *e.g.,* multi-scale super-resolution [Kim et al., 2016], and JPEG restoration on multiple quality factors [Galteri et al., 2019, Liu et al., 2018b]. With Palette we take a first step toward building multi-task image-to-image diffusion models for a wide variety of tasks.

3 Palette

Diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020] convert samples from a satandard Gaussian distribution into samples from an empirical data distribution through an iterative denoising process. Conditional diffusion models [Chen et al., 2021a, Saharia et al., 2021b] make the denoising process conditional on an input signal. Image-to-image diffusion models are conditional diffusion models of the form $p(\boldsymbol{y} \mid \boldsymbol{x})$, where both $\boldsymbol{x}$ and $\boldsymbol{y}$ are images, *e.g.,* $\boldsymbol{x}$ is a grayscale image and $\boldsymbol{y}$ is a color image. These models have been applied to image super-resolution [Saharia et al., 2021a, Nichol and Dhariwal, 2021]. We study the general applicability of image-to-image diffusion models on a broad set of tasks.

For a detailed treatment of diffusion models, please see Appendix A. Here, we briefly discuss the denoising loss function. Given a training output image $\boldsymbol{y}$, we generate a noisy version $\widetilde{\boldsymbol{y}}$, and train a neural network f_θ to denoise $\widetilde{\boldsymbol{y}}$ given $\boldsymbol{x}$ and a noise level indicator γ, for which the loss is

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\mathbb{E}_{\gamma}\left\| f_\theta(\boldsymbol{x}, \underbrace{\sqrt{\gamma}\,\boldsymbol{y} + \sqrt{1-\gamma}\,\boldsymbol{\epsilon}}_{\widetilde{\boldsymbol{y}}}, \gamma) - \boldsymbol{\epsilon} \right\|_p^p, \tag{1}$$

Chen et al. [2021a] and Saharia et al. [2021a] suggest using the L_1 norm, *i.e.,* $p = 1$, whereas the standard formulation is based on the usual L_2 norm [Ho et al., 2020]. We perform careful ablations below, and analyze the impact of the choice of norm. We find that L_1 yields significanlty lower sample diversity compared to L_2. While L_1 may be useful, to reduce potential hallucinations in some applications, here we adopt L_2 to capture the output distribution more faithfully.

Architecture. Palette uses a standard U-Net architecture [Ho et al., 2020] with several modifications inspired by recent work [Song et al., 2021, Saharia et al., 2021a, Dhariwal and Nichol, 2021]. The network architecture is based on the 256×256 class-conditional U-Net model of Dhariwal and Nichol [2021]. The two main differences between our architecture and theirs are (i) absence of class-conditioning, and (ii) additional conditioning of the source image via concatenation, following Saharia et al. [2021a].

4 Evaluation protocol

Evaluating image-to-image translation models is challenging. Prior work on colorization [Zhang et al., 2016, Guadarrama et al., 2017, Kumar et al., 2021] has relied on FID scores and human evaluation for model comparison. Tasks like inpainting [Yu et al., 2019, 2018a] and uncropping [Teterwak et al., 2019, Wang et al., 2019b] have often heavily relied on qualitative evaluation. For many tasks, such as JPEG restoration [Dong et al., 2015, Liu et al., 2018b, Galteri et al., 2019], it has been common to use reference-based pixel-level similarity scores such as PSNR and SSIM. It is also notable that many tasks lack a standardized dataset for evaluation, *e.g.,* different test sets with method-specific splits are used for evaluation.

We propose a unified evaluation protocol for inpainting, uncropping, and JPEG restoration on ImageNet [Deng et al., 2009], due to its scale, diversity, and public availability. For inpainting and uncropping, existing work has primarily relied on Places2 dataset [Zhou et al., 2017] for evaluation. Hence, we also use a standard evaluation setup on Places2 for these tasks. Specifically, we advocate the use of ImageNet ctest10k split proposed by Larsson et al. [2016] as a standard subset for benchmarking of all image-to-image translation tasks on ImageNet. We also introduce a similar category balanced 10,950 image subset of Places2 validation set called *places10k*. We further advocate the use of automated metrics that capture both image quality and diversity, in addition to controlled human evaluation. We avoid the use of pixel-level metrics such as PSNR and SSIM since such metrics are not reliable measures of sample quality for difficult tasks that require hallucination, like recent super-resolution work, where [Ledig et al., 2017, Dahl et al., 2017, Menon et al., 2020] observe that PSNR and SSIM tend to prefer blurry regression outputs which do not correlate well with human judgement.

We use four automated quantitative measures of sample quality for image-to-image translation: **Inception Score (IS)** [Salimans et al., 2017]; **Fréchet Inception Distance (FID)**; **Classification Accuracy (CA)** (top-1) of a pre-trained ResNet-50 classifier; and a simple measure of **Perceptual**

Distance (PD), *i.e.,* Euclidean distance in Inception-v1 feature space. In order to encourage and facilitate benchmarking on our proposed subsets, we release our model outputs on these benchmarks together with other data such as the inpainting masks used.[2] More details regarding our evaluation setup can be found in Appendix C.5. For some tasks, we also assess **sample diversity** through pairwise SSIM and LPIPS scores between multiple model outputs. Sample diversity is challenging and has been a key limitation of many existing GAN-based methods [Zhu et al., 2017b, Yang et al., 2019a].

The ultimate evaluation of image-to-image translation models is **human evaluation**; *i.e.,* whether or not humans can discriminate model outputs from natural images. To this end we use 2-alternative forced choice (2AFC) trials to evaluate the perceptual quality of model outputs against the original images from which we obtained the test inputs (*c.f.,* the Colorization Turing Test [Zhang et al., 2016]). We summarize the results in terms of the **fool rate**, the percentage of human raters who select model outputs over the original natural images when they were asked "Which image would you guess is from a camera?". (See Appendix C for details.)

5 Experiments

We study the application of *Palette* to a suite of challenging image-to-image translation tasks:

1. **Colorization** transforms an input grayscale image to a plausible color image.
2. **Inpainting** fills in user-specified masked regions of an image with realistic content.
3. **Uncropping** extends an input image along one or more directions to enlarge the image.
4. **JPEG restoration** corrects for JPEG compression artifacts, restoring plausible image detail.

We evaluate Palette on these tasks without task-specific hyper-parameter tuning, architecture customization, or any auxiliary loss function. Inputs and outputs for all tasks are represented as RGB 256×256 images. Each of these tasks presents its own unique challenges. Colorization entails some representation of objects, segmentation and layout, often with long-range image dependencies. Inpainting is challenging with large masks and diverse image datasets comprising cluttered scenes. Outpainting is widely considered even more challenging than inpainting as there is less surrounding context to constrain semantically meaningful generation. While the other tasks are linear inverse problems, JPEG restoration is a non-linear inverse problem; it requires a good local model of natural image statistics to detect and correct compression artifacts. While previous work has studied these problems extensively, it is rarely the case that a model with no task-specific engineering achieves strong performance in all tasks, beating strong task-specific GAN and regression baselines. Palette uses an L_2 loss for the denoising objective, unless otherwise specified. More implementation details can be found in Appendix B.

5.1 Colorization

While prior works [Zhang et al., 2016, Kumar et al., 2021] have adopted LAB or YCbCr color spaces to represent output images for colorization, we use the RGB color space to maintain generality across tasks. Preliminary experiments indicated that Palette is equally effective in YCbCr and RGB spaces. We compare Palette with pix2pix [Isola et al., 2017b], PixColor [Guadarrama et al., 2017], and ColTran [Kumar et al., 2021]. Qualitative results are shown in Fig. 3, with quantitative scores in Table 1. Palette establishes a new SoTA, outperforming existing works by a large margin. Further, the performance measures (FID, IS, and CA) indicate that Palette outputs are close to being indistinguishable from the original images that were used to create the test greyscale inputs. Surprisingly, our L_2 Regression baseline also outperforms prior task-specific techniques, highlighting the importance of modern architectures and large-scale training, even for a basic Regression model. On human evaluation, Palette improves upon human raters' fool rate of ColTran by more than 10%, approaching an ideal fool rate of 50%.

5.2 Inpainting

We follow Yu et al. [2019] and train inpainting models on free-form generated masks, augmented with simple rectangular masks. To maintain generality of Palette across tasks, in contrast to prior work, we do not pass a binary inpainting mask to the models. Instead, we fill the masked region with standard Gaussian noise, which is compatible with denoising diffusion models. The training loss only considers the masked out pixels, rather than the entire image, to speed up training. We compare Palette with DeepFillv2 [Yu et al., 2019], HiFill [Yi et al., 2020], Photoshop's *Content-aware Fill*,

[2] https://bit.ly/eval-pix2pix

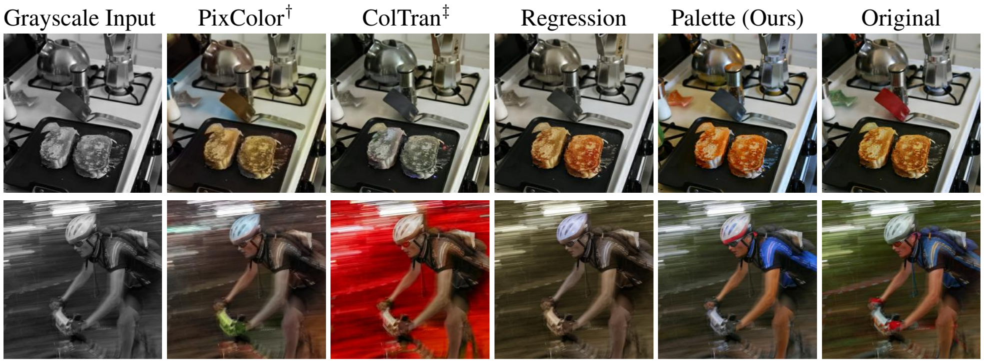

Figure 3: Illustration of colorization methods on ImageNet validation images. Baselines: †[Guadarrama et al., 2017], ‡[Kumar et al., 2021], and our own strong regression baseline. Figure C.3 shows more samples.

Model	FID-5K ↓	IS ↑	CA ↑	PD ↓	Fool rate ↑
Prior Work					
pix2pix [Isola et al., 2017b]	24.41	-	-	-	-
PixColor [Guadarrama et al., 2017]	24.32	-	-	-	29.90%
Coltran [Kumar et al., 2021]	19.37	-	-	-	36.55%
This paper					
Regression	17.89	169.8	68.2%	60.0	39.45%
Palette	**15.78**	**200.8**	**72.5%**	**46.2**	**47.80%**
Original images	14.68	229.6	75.6%	0.0	-

Table 1: Colorization quantitative scores and fool rates on ImageNet val set indicate that Palette outputs are bridging the gap to being indistinguishable from the original images from which the greyscale inputs were created. Appendix C.1 has more results.

and Co-ModGAN [Zhao et al., 2021]. While there are other important prior works on inpainting such as [Liu et al., 2018a, 2020, Zheng et al., 2019], we were not able to compare Palette with all of them.

Qualitative and quantiative results are given in Fig. 4 and Table 2. Palette exhibits strong performance across inpainting datasets and mask configurations, outperforming DeepFillv2, HiFill and Co-ModGAN by a large margin. Importantly, like the colorization task above, the FID scores for Palette outputs in the case of 20-30% free-form masks, are extremely close to FID scores on the original images from which we created the masked test inputs. See Appendix C.2 for more results.

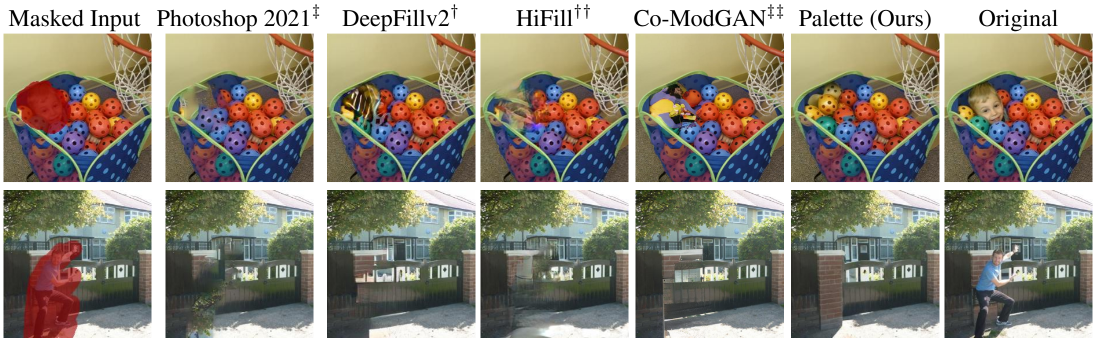

Figure 4: Comparison of inpainting methods on object removal. Baselines: ‡Photoshop's *Content-aware Fill* built on PatchMatch [Barnes et al., 2009], †[Yu et al., 2019], ††[Yi et al., 2020] and ‡‡[Zhao et al., 2021]. See Figure C.5 for more samples.

5.3 Uncropping

Recent works [Teterwak et al., 2019, Lin et al., 2021] have shown impressive visual effects by extending (extrapolating) input images along the right border. We train Palette on uncropping in any one of the four directions, or around the entire image border on all four sides. In all cases, we keep the area of the masked region at 50% of the image. Like inpainting, we fill the masked region

Mask Type	Model	ImageNet				Places2	
		FID ↓	**IS ↑**	**CA ↑**	**PD ↓**	**FID ↓**	**PD ↓**
20-30% *free form*	DeepFillv2 [Yu et al., 2019]	9.4	174.6	68.8%	64.7	13.5	63.0
	HiFill [Yi et al., 2020]	12.4	157.0	65.7%	86.2	15.7	92.8
	Co-ModGAN [Zhao et al., 2021]	-	-	-	-	12.4	51.6
	Palette (Ours)	**5.2**	**205.5**	**72.3%**	**27.6**	**11.7**	**35.0**
128×128 *center mask*	DeepFillv2 [Yu et al., 2019]	18.0	135.3	64.3%	117.2	15.3	96.3
	HiFill [Yi et al., 2020]	20.1	126.8	62.3%	129.7	16.9	115.4
	Co-ModGAN [Zhao et al., 2021]	-	-	-	-	13.7	86.2
	Palette (Ours)	**6.6**	**173.9**	**69.3%**	**59.5**	**11.9**	**57.3**
	Original images	5.1	231.6	74.6%	0.0	11.4	0.0

Table 2: Quantitative evaluation for free-form and center inpainting on ImageNet and Places2 validation images.

with Gaussian noise, and keep the unmasked region fixed during inference. We compare Palette with Boundless [Teterwak et al., 2019] and InfinityGAN [Lin et al., 2021]. While other uncropping methods exist (e.g., [Guo et al., 2020, Wang et al., 2019b]), we only compare with two representative methods. From the results in Fig. 5 and Table 3, one can see that Palette outperforms baselines on ImageNet and Places2 by a large margin. On human evaluation, Palette has a 40% fool rate, compared to 25% and 15% for Boundless and InfinityGAN (see Fig. C.2 for details).

We further assess the robustness of Palette by generating panoramas through repeated application of uncropping (Fig. 2). We observe that Palette is surprisingly robust, generating realistic and coherent outputs even after 8 repeated applications of uncrop. We also generate zoom-out sequences by repeated uncropping around the entire border of the image with similarly appealing results[3].

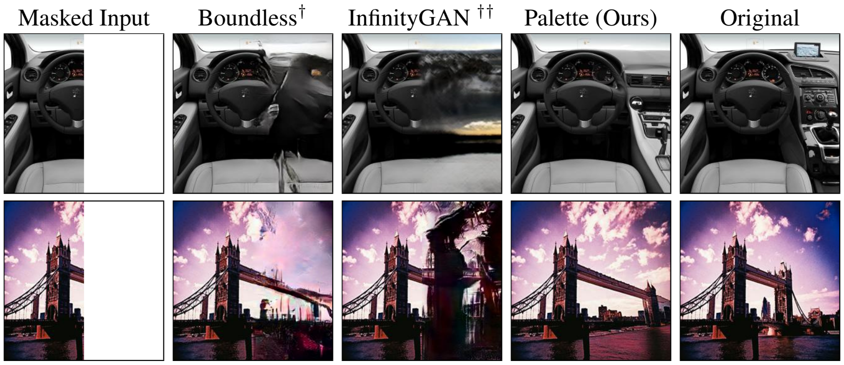

Figure 5: Image uncropping results on Places2 validation images. Baselines: Boundless† [Teterwak et al., 2019] and InfinityGAN†† [Lin et al., 2021] trained on a scenery subset of Places2. Figure C.7 shows more samples.

Model	ImageNet				Places2	
	FID ↓	**IS ↑**	**CA ↑**	**PD ↓**	**FID ↓**	**PD ↓**
Boundless [Teterwak et al., 2019]	18.7	104.1	58.8%	127.9	11.8	129.3
Palette (Ours)	**5.8**	**138.1**	**63.4%**	**85.9**	**3.53**	**103.3**
Original images	2.7	250.1	76.0%	0.0	2.1	0.0

Table 3: Quantitative scores and human raters' fool rates on uncropping. Appendix C.3 has more results.

5.4 JPEG restoration

Finally, we evaluate Palette on the task of removing JPEG compression artifacts, a long standing image restoration problem [Dong et al., 2015, Galteri et al., 2019, Liu et al., 2018b]. Like prior work [Ehrlich et al., 2020, Liu et al., 2018b], we train Palette on inputs compressed with various quality

[3] https://bit.ly/palette-diffusion

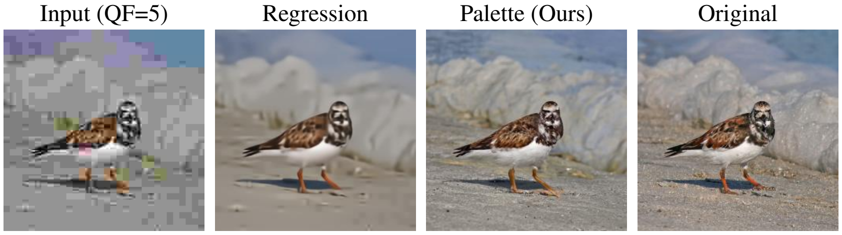

Figure 6: An example of JPEG restoration results. Fig. C.13 shows more samples.

factors (QF). While prior work has typically limited itself to a Quality Factor ≥ 10, we increase the difficulty of the task and train on Quality Factors as low as 5, producing severe compression artifacts. Table 4 summarizes the ImageNet results, with Palette exhibiting strong performance across all quality factors, outperforming the regression baseline. As expected, the performance gap between Palette and the regression baseline widens with decreasing quality factor. Figure 6 shows the qualitative comparison between Palette and our Regression baseline at a quality factor of 5. It is easy to see that the regression model produces blurry outputs, while Palette produces sharper images.

QF	Model	FID-5K ↓	IS ↑	CA ↑	PD ↓
5	Regression	29.0	73.9	52.8%	155.4
	Palette (Ours)	**8.3**	**133.6**	**64.2%**	**95.5**
10	Regression	18.0	117.2	63.5%	102.2
	Palette (Ours)	**5.4**	**180.5**	**70.7%**	**58.3**
20	Regression	11.5	158.7	69.7%	65.4
	Palette (Ours)	**4.3**	**208.7**	**73.5%**	**37.1**
	Original images	2.7	250.1	76.0%	0.0

Table 4: Quantitative evaluation for JPEG restoration for various Quality Factors (QF).

5.5 Self-attention in diffusion model architectures

Self-attention layers [Vaswani et al., 2017] have been an important component in recent U-Net architectures for diffusion models [Ho et al., 2020, Dhariwal and Nichol, 2021]. While self-attention layers build a direct form of global dependency into the model, they prevent generalization to unseen image resolutions. Generalization to new resolutions at test time is convenient for many image-to-image translation tasks, and therefore previous works have relied primarily on fully convolutional architectures [Yu et al., 2019, Galteri et al., 2019].

We analyze the impact of these self-attention layers on sample quality for inpainting, one of the more difficult image-to-image translation tasks. In order to enable input resolution generalization for Palette, we explore replacing global self-attention layers with different alternatives each of which represents a trade-off between large context dependency, and resolution robustness. In particular, we experiment with the following four configurations:

1. **Global Self-Attention**: Baseline configuration with global self-attention layers at 32×32, 16×16 and 8×8 resolutions.
2. **Local Self-Attention**: Local self-attention layers [Vaswani et al., 2021] at 32×32, 16×16 and 8×8 resolutions, at which feature maps are divided into 4 non-overlapping query blocks.
3. **More ResNet Blocks w/o Self-Attention**: 2 × residual blocks at 32×32, 16×16 and 8×8 resolutions allowing deeper convolutions to increase receptive field sizes.
4. **Dilated Convolutions w/o Self-Attention**: Similar to 3. ResNet blocks at 32×32, 16×16 and 8×8 resolutions with increasing dilation rates [Chen et al., 2017] allowing exponentially increasing receptive fields.

We train these models for 500K steps, and a batch size of 512. Table 5 reports the performance of different configurations for inpainting. Global self-attention offers much better performance than the fully-convolutional alternatives (even with 15% more parameters), re-affirming the importance of self-attention layers for such architectures. Surprisingly, local self-attention performs worse than fully-convolutional alternatives. We leave more detailed analysis of local self-attention in image generation architectures to future work.

Architecture	Model	# Params	FID ↓	IS ↑	PD ↓
Fully Convolutional	Dilated Convolutions	624M	8.0	157.5	70.6
	More ResNet Blocks	603M	8.1	157.1	71.9
Self-Attention	Local Self-Attention	552M	9.4	149.8	78.2
	Global Self-Attention	552M	**7.4**	**164.8**	**67.1**

Table 5: Architecture ablation for inpainting.

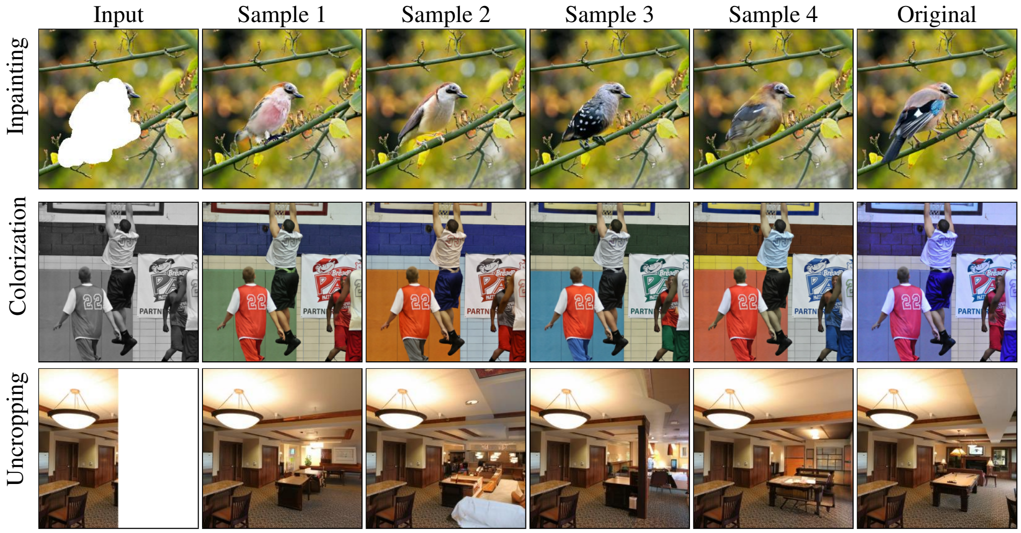

Figure 7: Diversity of Palette outputs on inpainting, colorization, and uncropping. Figures C.4, C.6, C.8 and C.9 have more samples.

5.6 Sample diversity

We next analyze sample diversity of Palette on two tasks, colorization and inpainting. Specifically, we analyze the impact of the changing the diffusion loss function L_{simple} [Ho et al., 2020], and compare L_1 vs. L_2 on sample diversity. While existing conditional diffusion models, SR3 [Saharia et al., 2021a] and WaveGrad [Chen et al., 2021a], have found L_1 norm to perform better than the conventional L_2 loss, there has not been a detailed comparison of the two. To quantitatively compare sample diversity, we use multi-scale SSIM Guadarrama et al. [2017] and the LPIPS diversity score Zhu et al. [2017b]. Given multiple generated outputs for each input image, we compute pairwise multi-scale SSIM between the first output sample and the remaining samples. We do this for multiple input images, and then plot the histogram of SSIM values (see Fig. 8). Following Zhu et al. [2017b], we also compute LPIPS scores between consecutive pairs of model outputs for a given input image, and then average across all outputs and input images. Lower SSIM and higher LPIPS scores imply more sample diversity. The results in Table 6 thus clearly show that models trained with the L_2 loss have greater sample diversity than those trained with the L_1 loss.

Interestingly, Table 6 also indicates that L_1 and L_2 models yield similar FID scores (i.e., comparable perceptual quality), but that L_1 has somewhat lower Perceptual Distance scores than L_2. One can speculate that L_1 models may drop more modes than L_2 models, thereby increasing the likelihood that a single sample from an L_1 model is from the mode containing the corresponding original image, and hence a smaller Perceptual Distance.

Some existing GAN-based models explicitly encourage diversity; [Zhu et al., 2017b, Yang et al., 2019a] propose methods for improving diversity of conditional GANs, and [Zhao et al., 2020, Han et al., 2019] explore diverse sample generation for image inpainting. We leave comparison of sample diversity between Palette and other such GAN based techniques to future work.

5.7 Multi-task learning

Multi-task training for image-to-image translation is a challenging but largely underexplored area of research. In prior work, the training paradigm is often restricted to multiple levels of corruptions

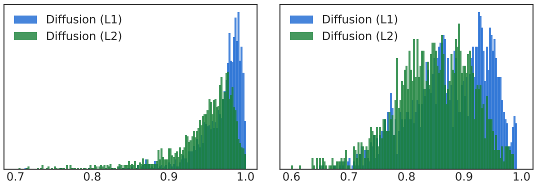

Figure 8: Distribution of pairwise multi-scale SSIM for Colorization (left) and Inpainting (right).

Model	**FID** ↓	**PD** ↓	**LPIPS** ↑
Inpainting			
Diffusion L_1	3.6	**41.9**	0.11
Diffusion L_2	3.6	43.8	**0.13**
Colorization			
Diffusion L_1	3.4	**45.8**	0.09
Diffusion L_2	3.4	48.0	**0.15**

Table 6: Comparison of L_p norm for Palette objective function.

within a single translation task. For example, Yu et al. [2019, 2018a] train inpainting models on different proportions of masked regions, Galteri et al. [2019], Liu et al. [2018b] train JPEG restoration models on multiple quality factors and [Zhang et al., 2017] train denoising models with a large range of noise level degradation. Here we train a single generalist Palette model on multiple translation tasks simultaneously. Specifically, we train Palette on tasks of JPEG restoration, inpainting, and colorization.

Table 7 indicates that multi-task generalist Palette outperforms the task-specific JPEG restoration specialist model, but slightly lags behind task-specific Palette models on inpainting and colorization. Due to compute limitations, we currently train multi-task Palette for the same number of training steps as task-specific models. We expect its performance to improve with more training.

Task	**Model**	**FID** ↓	**IS** ↑	**CA** ↑	**PD** ↓
Inpainting	Palette *(Task-specific)*	**6.6**	**173.9**	**69.3%**	**59.5**
128×128 center	Palette *(Multi-task)*	6.8	165.7	68.9%	65.2
Colorization	Regression *(Task-specific)*	5.5	176.9	68.0%	61.1
	Palette *(Task-specific)*	**3.4**	**212.9**	**72.0%**	**48.0**
	Palette *(Multi-task)*	3.7	187.4	69.4%	57.1
JPEG Restoration	Regression *(Task-specific)*	29.0	73.9	52.8%	155.4
(QF = 5)	Palette *(Task-specific)*	8.3	133.6	64.2%	95.5
	Palette *(Multi-task)*	**7.0**	**137.8**	**64.7%**	**92.4**

Table 7: Performance of multi-task Palette on various tasks.

6 Conclusion

We present Palette, a simple and general framework for image-to-image translation. Palette achieves strong results on four challenging image-to-image translation tasks (colorization, inpainting, uncropping, and JPEG restoration), outperforming strong GAN and regression baselines. Unlike many GAN models, Palette produces diverse outputs consistent with natural images. This is accomplished without task-specific customization or optimization instability. We also present a multi-task Palette model, that performs just as well or better over their task-specific counterparts. Further exploration and investigation of multi-task diffusion models is an exciting avenue for future work. This paper shows some of the potential of image-to-image diffusion models, but we look forward to seeing new applications.

7 Acknowledgement

We thank Jon Shlens, Doug Eck, Rif A. Saurous, Hossein Talebi and the Google Brain team for research discussions and technical assistance.

References

TensorFlow Datasets, a collection of ready-to-use datasets. https://www.tensorflow.org/datasets.

Eirikur Agustsson and Radu Timofte. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *CVPRW*, 2017.

Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Guided Image Generation with Conditional Invertible Neural Networks. In *arXiv:1907.02392*, 2019.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. In *arXiv*, 2017.

Jacob Austin, Daniel Johnson, Jonathan Ho, Danny Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *arXiv preprint arXiv:2107.03006*, 2021.

Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), August 2009.

Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

Richard Strong Bowen, Huiwen Chang, Charles Herrmann, Piotr Teterwak, Ce Liu, and Ramin Zabih. Oconet: Image extrapolation by object completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2307–2317, 2021.

Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. Learning Gradient Fields for Shape Generation. In *ECCV*, 2020.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating Gradients for Waveform Generation. In *ICLR*, 2021a.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. WaveGrad 2: Iterative Refinement for Text-to-Speech Synthesis . In *INTERSPEECH*, 2021b.

Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. In&out: Diverse image outpainting via gan inversion. *arXiv preprint arXiv:2104.00675*, 2021.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. Pixel recursive super resolution. In *ICCV*, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. Learning diverse image colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6837–6845, 2017.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. *arXiv:1605.08803*, 2016.

Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 576–584, 2015.

Max Ehrlich, Larry Davis, Ser-Nam Lim, and Abhinav Shrivastava. Quantization guided jpeg artifact correction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 293–309. Springer, 2020.

Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep generative adversarial compression artifact removal. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4826–4835, 2017.

Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep universal generative adversarial compression artifact removal. *IEEE Transactions on Multimedia*, 21(8):2131–2145, 2019.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *NIPS*, 2014.

Sergio Guadarrama, Ryan Dahl, David Bieber, Mohammad Norouzi, Jonathon Shlens, and Kevin Murphy. Pixcolor: Pixel recursive colorization. *arXiv preprint arXiv:1705.07208*, 2017.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

Dongsheng Guo, Hongzhi Liu, Haoru Zhao, Yunhao Cheng, Qingwei Song, Zhaorui Gu, Haiyong Zheng, and Bing Zheng. Spiral generative network for image extrapolation. In *European Conference on Computer Vision*, pages 701–717. Springer, 2020.

Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. Finet: Compatible and diverse fashion image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4481–4491, 2019.

James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (ToG)*, 26(3):4–es, 2007.

Kaiming He and Jian Sun. Statistics of patch offsets for image completion. In *European conference on computer vision*, pages 16–29. Springer, 2012.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded Diffusion Models for High Fidelity Image Generation. In *arXiv*, 2021.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. *arXiv preprint arXiv:2102.05379*, 2021.

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

Phillip Isola, Jun-Yan Zhu, and Tinghui Zhou ajnd Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Nets. In *CVPR*, 2017a.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017b.

Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021.

Zahra Kadkhodaie and Eero P Simoncelli. Solving linear inverse problems using the prior implicit in a denoiser. *arXiv preprint 2007.13640*, 2021.

Eungyeup Kim, Sanghyeon Lee, Jeonghoon Park, Somi Choi, Choonghyun Seo, and Jaegul Choo. Deep edge-aware interactive colorization against color-bleeding effects. *arXiv preprint arXiv:2107.01619*, 2021a.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645, 2016.

Soo Ye Kim, Kfir Aberman, Nori Kanazawa, Rahul Garg, Neal Wadhwa, Huiwen Chang, Nikhil Karnad, Munchurl Kim, and Orly Liba. Zoom-to-inpaint: Image inpainting with high-frequency details, 2021b.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *NIPS*, 2018.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2013.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *arXiv preprint arXiv:2107.00630*, 2021.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. *ICLR*, 2021.

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

Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. In *ICLR 2021*, 2021. URL https://openreview.net/forum?id=5NA1PinlGFu.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *ICCV*, 2017.

Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. Infinitygan: Towards infinite-resolution image synthesis. *arXiv preprint arXiv:2104.03963*, 2021.

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

Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 725–741. Springer, 2020.

Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10551–10560, 2019.

Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 773–782, 2018b.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.

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.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018.

Guocheng Qian, Jinjin Gu, Jimmy Ren, Chao Dong, Furong Zhao, and Juan Lin. Trinity of Pixel Enhancement: a Joint Solution for Demosaicking, Denoising and Super-Resolution. In *arXiv:1905.02538*, 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *arXiv preprint arXiv:1905.10887*, 2019.

Stefan Roth and Michael J Black. Fields of experts: A framework for learning image priors. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 860–867. IEEE, 2005.

Amelie Royer, Alexander Kolesnikov, and Christoph H. Lampert. Probabilistic Image Colorization. In *arXiv:1705.04258*, 2017.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. *arXiv preprint arXiv:2104.07636*, 2021a.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021b.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. In *ICLR*, 2017.

Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021.

Qi Shan, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M Seitz. Photo uncrop. In *European Conference on Computer Vision*, pages 16–31. Springer, 2014.

Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-denoising models for few-shot conditional generation. *arXiv preprint arXiv:2106.06819*, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015.

Yang Song and Stefano Ermon. Improved Techniques for Training Score-Based Generative Models. *arXiv preprint arXiv:2006.09011*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*, 2021.

Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7968–7977, 2020.

Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10521–10530, 2019.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *NeurIPS*, 2020.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *arXiv preprint arXiv:2106.05931*, 2021.

Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with PixelCNN decoders. In *NIPS*, pages 4790–4798, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NIPS*, 2017.

Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12894–12904, 2021.

Miao Wang, Yu-Kun Lai, Yuan Liang, Ralph R Martin, and Shi-Min Hu. Biggerpicture: data-driven image extrapolation using graph matching. *ACM Transactions on Graphics*, 33(6), 2014.

Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1399–1408, 2019a.

Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2019b.

Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*, 2021.

Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. *arXiv preprint arXiv:1901.09024*, 2019a.

Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10561–10570, 2019b.

Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018a.

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

Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2443–2452, 2018b.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5741–5750, 2020.

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.

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

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*, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017a.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in neural information processing systems*, pages 465–476, 2017b.

A Diffusion Models

Diffusion models comprise a forward diffusion process and a reverse denoising process that is used at generation time. The forward diffusion process is a Markovian process that iteratively adds Gaussian noise to a data point $\boldsymbol{y}_0 \equiv \boldsymbol{y}$ over T iterations:

$$q(\boldsymbol{y}_{t+1}|\boldsymbol{y}_t) \quad = \quad \mathcal{N}(\boldsymbol{y}_{t-1}; \sqrt{\alpha_t}\boldsymbol{y}_{t-1}, (1-\alpha_t)I) \tag{2}$$

$$q(\boldsymbol{y}_{1:T}|\boldsymbol{y}_0) \quad = \quad \prod_{t=1}^{T} q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}) \tag{3}$$

where α_t are hyper-parameters of the noise schedule. The forward process with α_t is constructed in a manner where at $t = T$, $\boldsymbol{y}_T$ is virtually indistinguishable from Gaussian noise. Note, we can also marginalize the forward process at each step:

$$q(\boldsymbol{y}_t|\boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{y}_t; \sqrt{\gamma_t}\boldsymbol{y}_0, (1-\gamma_t)I) \ , \tag{4}$$

where $\gamma_t = \prod_{t'}^{t} \alpha'_t$.

The Gaussian parameterization of the forward process also allows a closed form formulation of the posterior distribution of $\boldsymbol{y}_{t-1}$ given $(\boldsymbol{y}_0, \boldsymbol{y}_t)$ as

$$q(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_0, \boldsymbol{y}_t) \ = \ \mathcal{N}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \tag{5}$$

where $\boldsymbol{\mu} = \frac{\sqrt{\gamma_{t-1}}\,(1-\alpha_t)}{1-\gamma_t}\,\boldsymbol{y}_0 + \frac{\sqrt{\alpha_t}\,(1-\gamma_{t-1})}{1-\gamma_t}\boldsymbol{y}_t$ and $\sigma^2 = \frac{(1-\gamma_{t-1})(1-\alpha_t)}{1-\gamma_t}$. This result proves to be very helpful during inference as shown below.

Learning: Palette learns a reverse process which inverts the forward process. Given a noisy image $\widetilde{\boldsymbol{y}}$,

$$\widetilde{\boldsymbol{y}} = \sqrt{\gamma}\,\boldsymbol{y}_0 + \sqrt{1-\gamma}\,\boldsymbol{\epsilon} \ , \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \ , \tag{6}$$

the goal is to recover the target image $\boldsymbol{y}_0$. We parameterize our neural network model $f_\theta(x, \widetilde{\boldsymbol{y}}, \gamma)$ to condition on the input x, a noisy image $\widetilde{\boldsymbol{y}}$, and the current noise level γ. Learning entails prediction of the noise vector $\boldsymbol{\epsilon}$ by optimizing the objective

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\mathbb{E}_{\boldsymbol{\epsilon},\gamma}\bigg\| f_\theta(\boldsymbol{x}, \underbrace{\sqrt{\gamma}\,\boldsymbol{y}_0 + \sqrt{1-\gamma}\,\boldsymbol{\epsilon}}_{\widetilde{\boldsymbol{y}}}, \gamma) - \boldsymbol{\epsilon} \bigg\|_p^p \ . \tag{7}$$

This objective, also known as L_{simple} in Ho et al. [2020], is equivalent to maximizing a weighted variational lower-bound on the likelihood [Ho et al., 2020].

Inference: Palette performs inference via the learned reverse process. Since the forward process is constructed so the prior distribution p(y_T) approximates a standard normal distribution $\mathcal{N}(\boldsymbol{y}_T|\boldsymbol{0}, \boldsymbol{I})$, the sampling process can start at pure Gaussian noise, followed by T steps of iterative refinement.

Also recall that the neural network model f_θ is trained to estimate $\boldsymbol{\epsilon}$, given any noisy image $\widetilde{\boldsymbol{y}}$, and $\boldsymbol{y}_t$. Thus, given $\boldsymbol{y}_t$, we approximate $\boldsymbol{y}_0$ by rearranging terms in equation 6 as

$$\hat{\boldsymbol{y}}_0 = \frac{1}{\sqrt{\gamma_t}}\left(\boldsymbol{y}_t - \sqrt{1-\gamma_t}\, f_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t)\right) \ . \tag{8}$$

Following Ho et al. [2020], we substitute our estimate $\hat{\boldsymbol{y}}_0$ into the posterior distribution of $q(\boldsymbol{y}_{t-1}|\boldsymbol{y}_0, \boldsymbol{y}_t)$ in equation 5 to parameterize the mean of $p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x})$ as

$$\mu_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t)\right) \ . \tag{9}$$

And we set the variance of $p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{x})$ to $(1-\alpha_t)$, a default given by the variance of the forward process Ho et al. [2020].

With this parameterization, each iteration of the reverse process can be computed as

$$\boldsymbol{y}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t)\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t \ ,$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. This resembles one step of Langevin dynamics for which f_θ provides an estimate of the gradient of the data log-density.

Algorithm 1 Training a denoising model f_θ

```
repeat
    $(\boldsymbol{x}, \boldsymbol{y}_0) \sim p(\boldsymbol{x}, \boldsymbol{y})$
    $\gamma \sim p(\gamma)$
    $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    Take a gradient descent step on
        $\nabla_\theta \left\| f_\theta(\boldsymbol{x}, \sqrt{\gamma}\boldsymbol{y}_0 + \sqrt{1-\gamma}\boldsymbol{\epsilon}, \gamma) - \boldsymbol{\epsilon} \right\|_p^p$
until converged
```

Algorithm 2 Inference in T iterative refinement steps

```
$\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ do
    $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\boldsymbol{z} = \mathbf{0}$
    $\boldsymbol{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t) \right) + \sqrt{1-\alpha_t}\boldsymbol{z}$
end for
return $\boldsymbol{y}_0$
```

B Implementation Details

Training Details : We train all models with a mini batch-size of 1024 for 1M training steps. We do not find over fitting to be an issue, and hence use the model checkpoint at 1M steps for reporting the final results. Consistent with previous works [Ho et al., 2020, Saharia et al., 2021a], we use standard Adam optimizer with a fixed 1e-4 learning rate and 10k linear learning rate warmup schedule. We use 0.9999 EMA for all our experiments. We do not perform any task-specific hyper-parameter tuning, or architectural modifications.

Diffusion Hyper-parameters : Following [Saharia et al., 2021a, Chen et al., 2021a] we use α conditioning for training Palette. This allows us to perform hyper-parameter tuning over noise schedules and refinement steps for Palette during inference. During training, we use a linear noise schedule of ($1e^{-6}$, 0.01) with 2000 time-steps, and use 1000 refinement steps with a linear schedule of ($1e^{-4}$, 0.09) during inference.

Task Specific Details: We specify specific training details for each of the tasks below:

- **Colorization** : We use RGB parameterization for colorization. We use the grayscale image as the source image and train Palette to predict the full RGB image. During training, following [Kumar et al., 2021], we randomly select the largest square crop from the image and resize it to 256×256.
- **Inpainting** : We train Palette on a combination of free-form and rectangular masks. For free-form masks, we use Algorithm 1 in [Yu et al., 2019]. For rectangular masks, we uniformly sample between 1 and 5 masks. The total area covered by the rectangular masks is kept between 10% to 40% of the image. We randomly sample a free-form mask with 60% probability, and rectangular masks with 40% probability. Note that this is an arbitrary training choice. We do not provide any additional mask channel, and simply fill the masked region with random Gaussian noise. During training, we restrict the L_{simple} loss function to the spatial region corresponding to masked regions, and use the model's prediction for only the masked region during inference. We train Palette on two types of 256×256 crops. Consistent with previous inpainting works [Yu et al., 2019, 2018a, Yi et al., 2020], we use random 256×256 crops, and we combine these with the resized random largest square crops used in colorization literature [Kumar et al., 2021].
- **Uncropping** : We train the model for image extension along all four directions, or just one direction. In both cases, we set the masked region to 50% of the image. During training, we uniformly choose masking along one side, or masking along all 4 sides. When masking along one side, we further make a uniform random choice over the side. Rest of the training details are identical to inpainting.
- **JPEG Restoration** : We train Palette for JPEG restoration on quality factors in (5, 30). Since decompression for lower quality factors is a significantly more difficult task, we use an exponential distribution to sample the quality factor during training. Specifically, the sampling probability of a quality range Q is set to $\propto e^{-\frac{Q}{10}}$.

C Additional Experimental Results

C.1 Colorization

Following prior work [Zhang et al., 2016, Guadarrama et al., 2017, Kumar et al., 2021], we train and evaluate models on ImageNet [Deng et al., 2009]. In order to compare our models with existing works in Table 1, we follow ColTran [Kumar et al., 2021] and use the first 5000 images from ImageNet validation set to report performance on standard metrics. We use the next 5000 images as the reference distribution for FID to mirror ColTran's implementation (as returned by TFDS [TFD] data loader). For benchmarking purposes, we also report the performance of Palette on ImageNet ctest10k [Larsson et al., 2016] dataset in Table C.1.

Human Evaluation: The ultimate evaluation of image-to-image translation models is human evaluation; *i.e.,* whether or not humans can discriminate model outputs from reference images. To this end we use controlled human experiments. In a series of two alternative forced choice trials, we ask subjects which of two side-by-side images is the real photo and which has been generated by the model. In particular, subjects are asked *"Which*

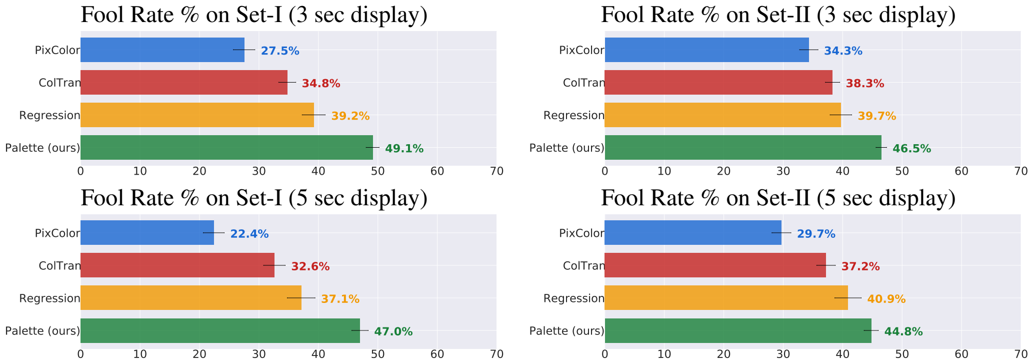

Figure C.1: Human evaluation results on ImageNet colorization.

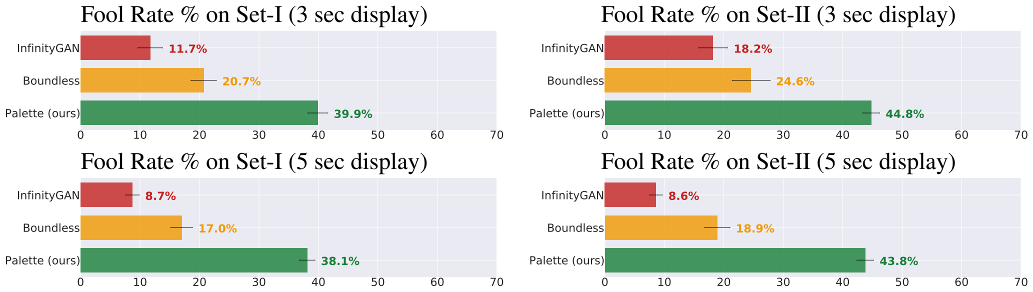

Figure C.2: Human evaluation results on Places2 uncropping.

image would you guess is from a camera?" Subjects viewed images for either 3 or 5 seconds before having to respond. For the experiments we compare outputs from four models against reference images, namely, PixColor [Guadarrama et al., 2017], Coltran [Kumar et al., 2021], our Regression baseline, and Palette. To summarize the result we compute the subject *fool rate*, i.e., the fraction of human raters who select the model outputs over the reference image. We use a total of 100 images for human evaluation, and divide these into two independent subsets - Set-I and Set-II, each of which is seen by 50 subjects.

As shown in Figure C.1, the fool rate for Palette is close to 50% and higher than baselines in all cases. We note that when subjects are given less time to inspect the images the fool rates are somewhat higher, as expected. We also note the strength of our regression baseline, which also performs better than PixColor and Coltran. Finally, to provide insight into the human evaluation results we also show several more examples of Palette output, with comparisons to benchmarks, in Figure C.3. One can see that in several cases, Palette has learned colors that are more meaningful and consistent with the reference images and the semantic content of the images. Figure C.4 also shows the natural diversity of Palette outputs for colorization model.

C.2 Inpainting

Comparison on 256×256 images: We report all inpainting results on 256×256 center cropped images. Since the prior works we use for comparison are all trained on random 256×256 crops, evaluation on 256×256 center crops ensures fair comparison. Furthermore, we use a fixed set of image-mask pair for each configuration for all models during evaluation. Since HiFill [Yi et al., 2020] and Co-ModGAN [Yi et al., 2020] are primarily trained on 512×512 images, we use 512×512 center crops with exact same mask within the central 256×256 region. This provides these two models with 4× bigger inpainting context compared to DeepFillv2 and Palette.

We train two Palette models for Inpainting - i) Palette (I) trained on ImageNet dataset, and ii) Palette (I+P) trained on mixture of ImageNet and Places2 dataset. For Palette (I+P), we use a random sampler policy to sample from ImageNet and Places2 dataset with a uniform probability. Table C.2 shows full comparison of Palette with existing methods on all inpainting configurations. Based on the type of mask, and the area covered, we report results for the following categories - i) 10-20% free-form region, ii) 20-30% free-form region, iii) 30-40% free-form region and iv) 128×128 center rectangle region. Palette consistently outperforms existing works by a significant margin on all configurations. Interestingly Palette (I) performs slightly better than Palette (I+P) on ImageNet indicating that augmentation with Places2 images during training doesn't boost to ImageNet performance. Furthermore, Palette (I) is only slightly worse compared to Palette (I+P) on Places2 even though it is not trained on Places2 images. We observe a significant drop in the performance of HiFill [Yi et al., 2020]

Model	**FID-10K** ↓	**IS** ↑	**CA** ↑	**PD** ↓
Palette (L_2)	3.4	212.9	72.0%	48.0
Palette (L_1)	3.4	215.8	71.9%	45.8
Ground Truth	2.7	250.1	76.0%	0.0

Table C.1: Benchmark numbers on ctest10k ImageNet subset for Image Colorization.

Mask Type	**Model**	**ImageNet**				**Places2**	
		FID ↓	**IS** ↑	**CA** ↑	**PD** ↓	**FID** ↓	**PD** ↓
10-20% Free-Form Mask	DeepFillv2 [Yu et al., 2019]	6.7	198.2	71.6%	38.6	12.2	38.1
	HiFill [Yi et al., 2020]	7.5	192.0	70.1%	46.9	13.0	55.1
	Palette (I) (Ours)	**5.1**	**221.0**	**73.8%**	15.6	11.6	22.1
	Palette (I+P) (Ours)	5.2	219.2	73.7%	**15.5**	**11.6**	**20.3**
20-30% Free-Form Mask	DeepFillv2 [Yu et al., 2019]	9.4	174.6	68.8%	64.7	13.5	63.0
	HiFill [Yi et al., 2020]	12.4	157.0	65.7%	86.2	15.7	92.8
	Co-ModGAN [Zhao et al., 2021]	-	-	-	-	12.4	51.6
	Palette (I) (Ours)	**5.2**	**208.6**	**72.6%**	**27.4**	11.8	37.7
	Palette (I+P) (Ours)	**5.2**	205.5	72.3%	27.6	**11.7**	**35.0**
30-40% Free-Form Mask	DeepFillv2 [Yu et al., 2019]	14.2	144.7	64.9%	95.5	15.8	90.1
	HiFill [Yi et al., 2020]	20.9	115.6	59.4%	131.0	20.1	132.0
	Palette (I)	**5.5**	**195.2**	**71.4%**	**39.9**	12.1	53.5
	Palette (I+P)	5.6	192.8	71.3%	40.2	**11.6**	**49.2**
128×128 Center Mask	DeepFillv2 [Yu et al., 2019]	18.0	135.3	64.3%	117.2	15.3	96.3
	HiFill [Yi et al., 2020]	20.1	126.8	62.3%	129.7	16.9	115.4
	Palette (I)	**6.4**	173.3	**69.7%**	**58.8**	12.2	62.8
	Co-ModGAN [Zhao et al., 2021]	-	-	-	-	13.7	86.2
	Palette (I+P)	6.6	**173.9**	69.3%	59.5	**11.9**	**57.3**
	Ground Truth	5.1	231.6	74.6%	0.0	11.4	0.0

Table C.2: Quantitative evaluation for inpainting on ImageNet and Places2 validation images.

with larger masks. It is important to note that DeepFillv2 and HiFill are not trained on ImageNet, but we report their performance on ImageNet ctest10k primarily for benchmarking purposes.

C.3 Uncropping

Many existing uncropping methods [Cheng et al., 2021, Teterwak et al., 2019] have been trained on different subsets of Places2 [Zhou et al., 2017] dataset. In order to maintain uniformity, we follow a similar setup as inpainting and train Palette on a combined dataset of Places2 and ImageNet. While we train Palette to extend the image in all directions or just one direction, to compare fairly against existing methods we evaluate Palette on extending only the right half of the image. For Table 3, we use ctest10k and places10k to report results on ImageNet and Places2 validation sets respectively.

We also perform category specific evaluation of Palette with existing techniques - Boundless [Teterwak et al., 2019] and InfinityGAN [Lin et al., 2021]. Since Boundless is only trained on top-50 categories from Places2 dataset, we compare Palette with Boundless specifically on these categories from Places2 validation set in Table C.3. Palette achieves significantly better performance compared to Boundless re-affirming the strength of our model. Furthermore, we compare Palette with a more recent GAN based uncropping technique - InfinityGAN [Lin et al., 2021]. In order to fairly compare Palette with InfinityGAN, we specifically evaluate on the scenery categories from Places2 validation and test set. We use the samples generously provided by Lin et al. [2021], and generate outputs for Boundless, and Palette. Table C.4 shows that Palette is significantly better than domain specific model InfinityGAN on scenery images in terms of automated metrics.

Human Evaluation: Like colorization, we also report results from human evaluation experiments. Obtaining high fool rates for uncropping is a significantly more challenging task than colorization, because one half of the image area is fully generated by the model. As a consequence there are more opportunities for synthetic artifacts. Because the baselines available for uncropping are trained and tested on Places2, we run human evaluation experiments only on Places2. Beyond the choice of dataset, all other aspects of experimental design are identical to that used above for colorization, with two disjoint sets of test images, namely, Set-I and Set-II.

Model	FID ↓	PD ↓
Boundless [Teterwak et al., 2019]	28.3	115.0
Palette	**22.9**	**93.4**
Ground Truth	23.6	0.0

Table C.3: Comparison with uncropping method Boundless [Teterwak et al., 2019] on top-50 Places2 categories.

Model	FID ↓
Boundless [Teterwak et al., 2019]	12.7
InfinityGAN [Lin et al., 2021]	15.7
Palette	**5.6**

Table C.4: Comparison with uncropping method InfinityGAN [Lin et al., 2021] and Boundless [Teterwak et al., 2019] on scenery categories.

The results are characterized in terms of the fool rate, and are shown in Figure C.2. Palette obtains significantly higher fool rates on all human evaluation runs compared to existing methods, i.e., Boundless [Teterwak et al., 2019] and InfinityGAN [Lin et al., 2021]. Interestingly, when raters are given more time to inspect each pair of images, the fool rates for InfinityGAN and Boundless worsen considerably. Palette, on the other hand, observes approximately similar fool rates.

C.4 JPEG Restoration

In order to be consistent with other tasks, we perform training and evaluation on ImageNet dataset. Note that this is unlike most prior work [Dong et al., 2015, Liu et al., 2018b], which mainly use small datasets such as DIV2K [Agustsson and Timofte, 2017] and BSD500 [Martin et al., 2001] for training and evaluation. Recent works such as [Galteri et al., 2019] use a relatively larger MS-COCO dataset for training, however, to the best of our knowledge, we are the first to train and evaluate JPEG restoration on ImageNet. We compare Palette with a strong Regression baseline which uses an identical architecture. We report results on JPEG quality factor settings of 5, 10 and 20 in Table 4.

C.5 Evaluation and Benchmarking Details

Several existing works report automated metrics such as FID, Inception Score, etc. [Kumar et al., 2021, Lin et al., 2021, Yi et al., 2020] but often lack key details such as the subset of images used for computing these metrics, or the reference distribution used for calculating FID scores. This makes direct comparison with such reported metrics difficult. Together with advocating for our proposed benchmark validation sets, we also provide all the necessary details to exactly replicate our reported results. We encourage future works to adopt a similar practice of reporting all the necessary evaluation details in order to facilitate direct comparison with their methods.

Benchmark datasets: For ImageNet evaluation, we use the 10,000 image subset from ImageNet validation set - **ctest10k** introduced by [Larsson et al., 2016]. While this subset has been primarliy used for evaluation in the colorization literature [Guadarrama et al., 2017, Su et al., 2020, Kim et al., 2021a], we extend its use for other image-to-image translation tasks. Many image-to-image translation tasks such as inpainting, uncropping are evaluated on Places2 dataset [Zhou et al., 2017]. However, to the best of our knowledge, there is no such standardized subset for Places2 validation set used for benchmarking. To this end, we introduce **places10k**, a 10,950 image subset of Places2 validation set. Similar to ctest10k, we make places10k class balanced with 30 images per class (Places2 dataset has 365 classes/categories in total.).

Metrics: We report several automated metrics for benchmarking and comparison with existing methods. Specifically, we report **Fréchet Inception Distance (FID)**, **Inception Score**, **Perceptual Distance** and **Classification Accuracy** for qualitative comparison. When computing FID scores, the choice of the reference distribution is important, but is often not clarified in existing works. In our work, we use the full validation set as the reference distribution, i.e. 50k images from ImageNet validation set for computing scores on ImageNet subset ctest10k, and 36.5k images from Places2 validation set for computing scores on Places2 subset places10k. For Perceptual Distance, we use the Euclidean distance in the *pool*_3 feature space of the pre-trained InceptionV1 network (same as the features used for calculating FID scores). We use EfficientNet-B0 [4] top-1 accuracy for reporting Classification Accuracy scores.

[4] https://tfhub.dev/google/efficientnet/b0/classification/1

C.6 Limitations and Social Impact

While Palette achieves strong results on several image-to-image translation tasks demonstrating the generality and versatility of the emerging diffusion models, there are many important limitations to address. Diffusion models generally require large number of refinement steps during sample generation (e.g. we use 1k refinement steps for Palette throughout the paper) resulting in significantly slower inference compared to GAN based models. This is an active area of research, and several new techniques [Nichol and Dhariwal, 2021, Watson et al., 2021, Jolicoeur-Martineau et al., 2021] have been proposed to reduce the number of refinement steps significantly. We leave application of these techniques on Palette to future work. Palette's use of group-normalization and self-attention layers prevents its generalizability to arbitrary input image resolutions, limiting its practical usability. Techniques to adapt such models to arbitrary resolutions such as fine-tuning, or patch based inference can be an interesting direction of research.

Bias is a key issue in all generative models. This includes conditional diffusion models like Palette, which generate samples from a posterior distribution over images conditioned on a task and an input image. In some tasks, like the JPEG artifact removal, this entails the generation of local image structure to repair compression artifacts. For other tasks, like inpainting and uncropping, relatively large regions of images are hallucinated, which are meant to be plausible and consistent with the surrounding image context. Their ability to solve this task, in part, depende on the data they are trained with. Our models have been primarily trained on ImageNet and Places2 data. Salient objects that do not occur with sufficient frequency in these training data will not be rendered well by Palette. Examples may include images containing text, or images of people, especially faces, to which human vision is quite sensitive. While our log-likelihood based objective is mode covering (*e.g.,* unlike some GAN-based objectives), we believe it is likely our diffusion-based models may drop modes and may produce somewhat biased results. Hence, we recommend training the models on large datasets that are well suited to the test domain of interest before practical deployment.

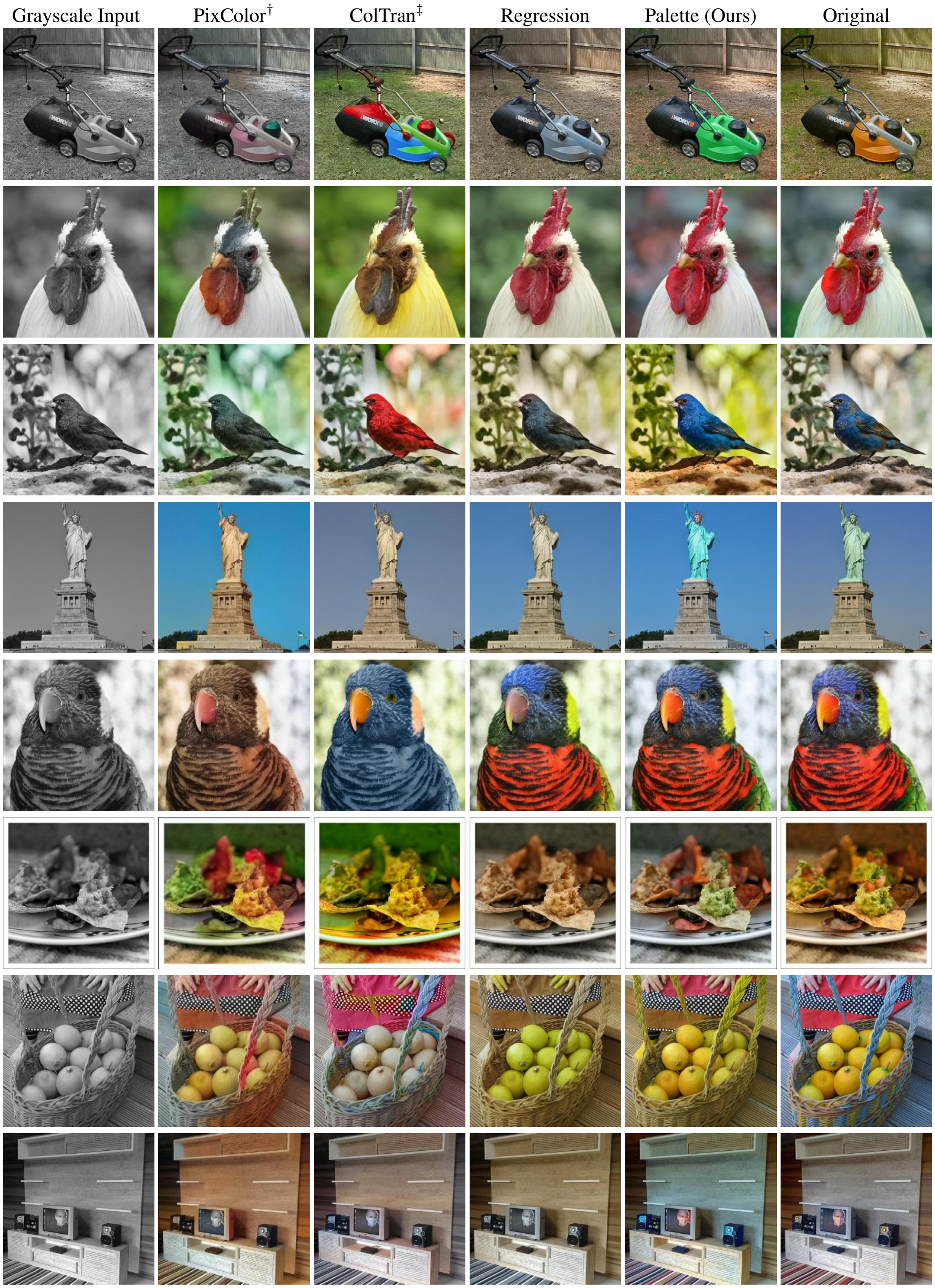

Figure C.3: Comparison of different methods for colorization on ImageNet validation images. Baselines: †[Guadarrama et al., 2017] and ‡[Kumar et al., 2021].

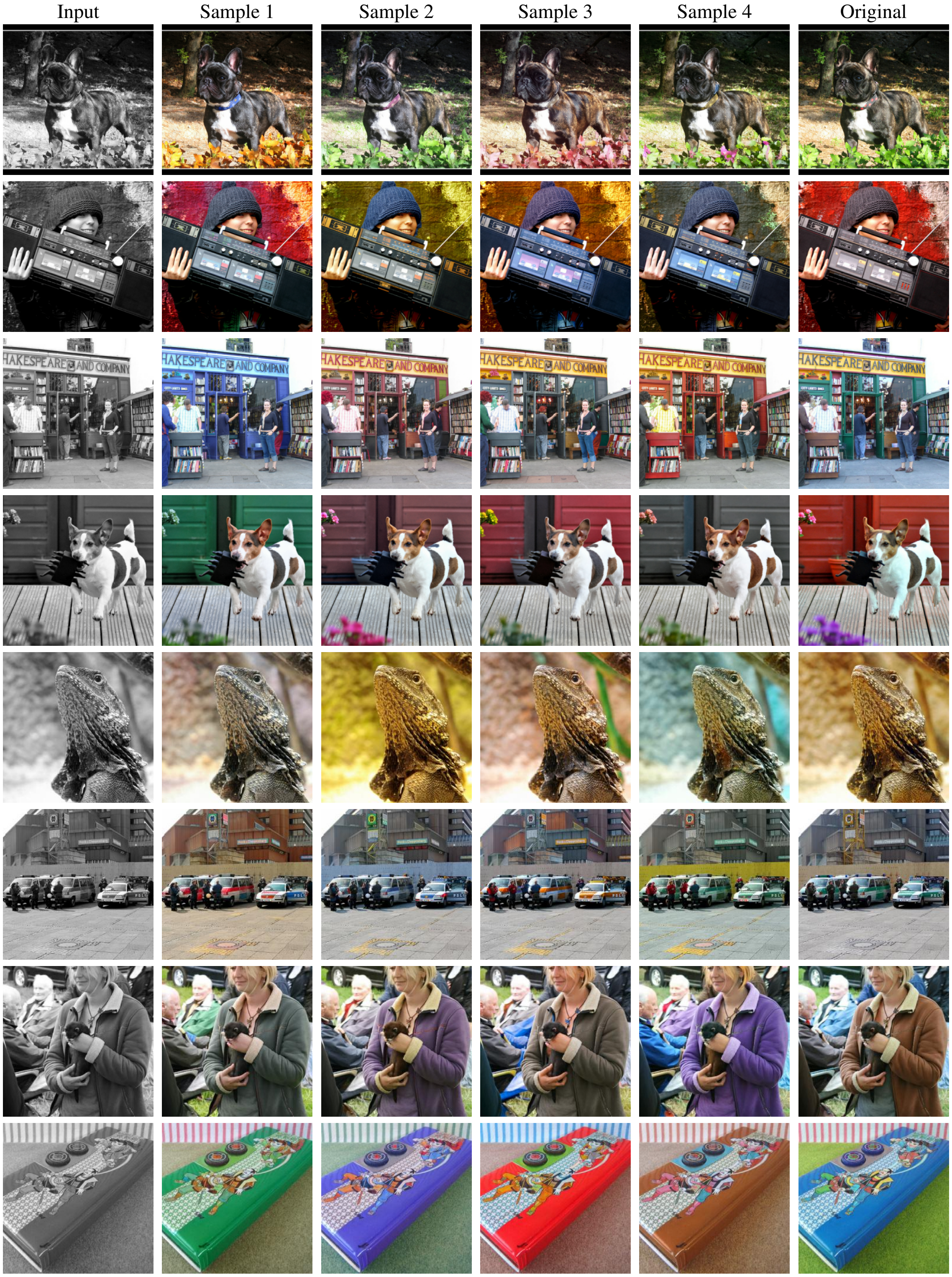

Figure C.4: Diversity of Palette outputs on ImageNet colorization validation images.

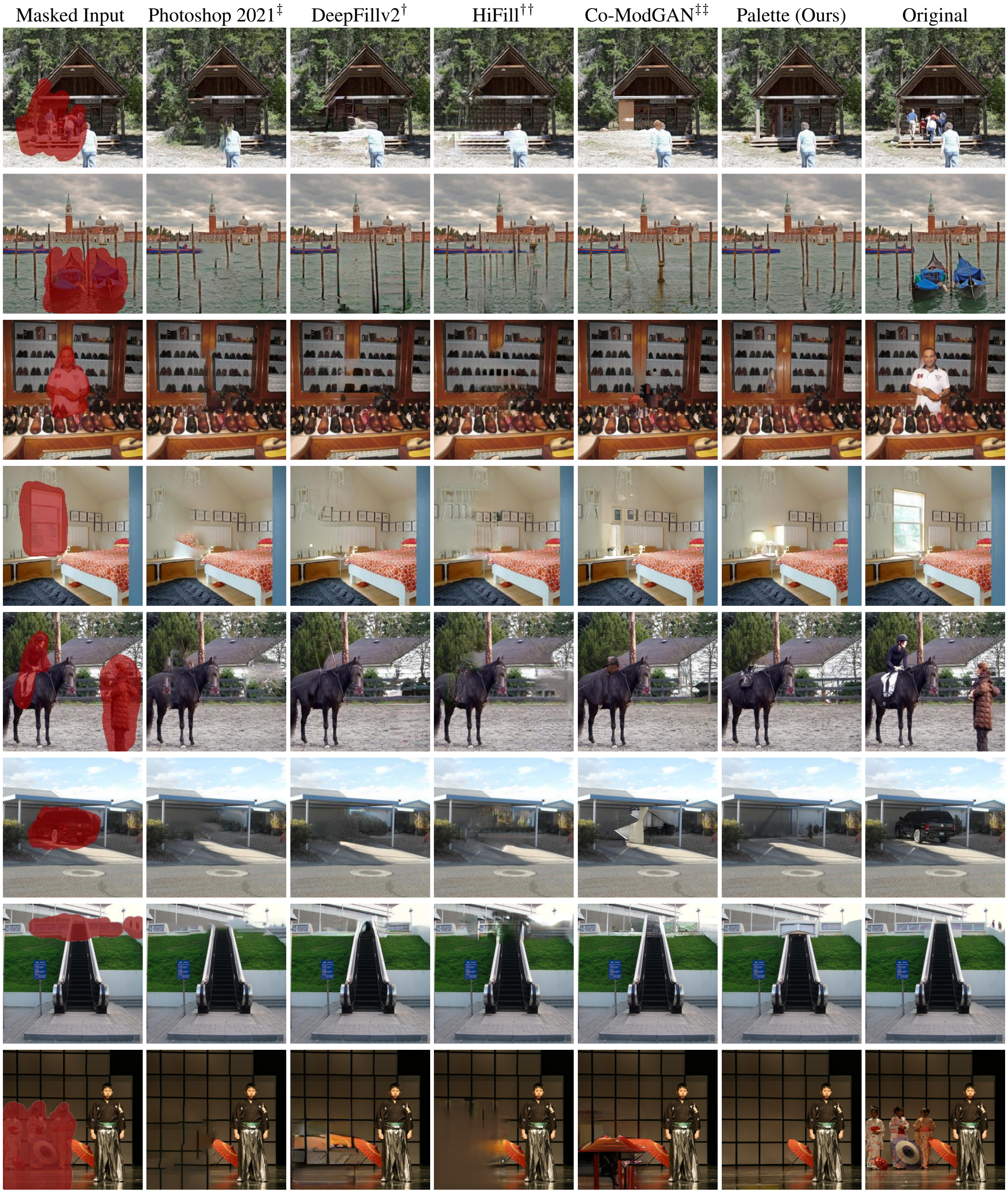

Figure C.5: Comparison of inpainting methods on object removal. Baselines: ‡Photoshop's *Content-aware Fill*, based on PatchMatch [Barnes et al., 2009], †[Yu et al., 2019], ††[Yi et al., 2020] and ‡‡[Zhao et al., 2021].

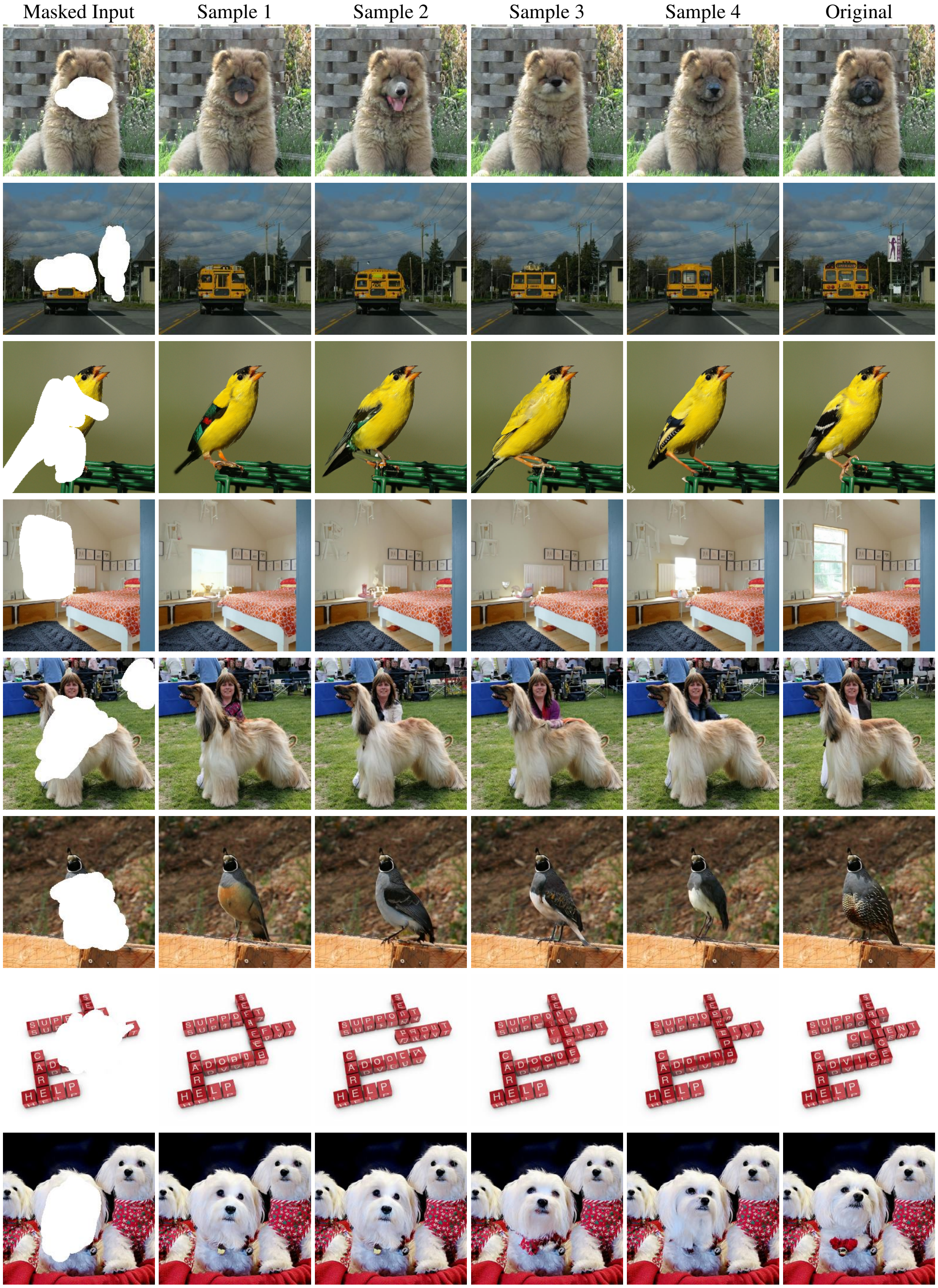

Figure C.6: Diversity of Palette outputs on image inpainting.

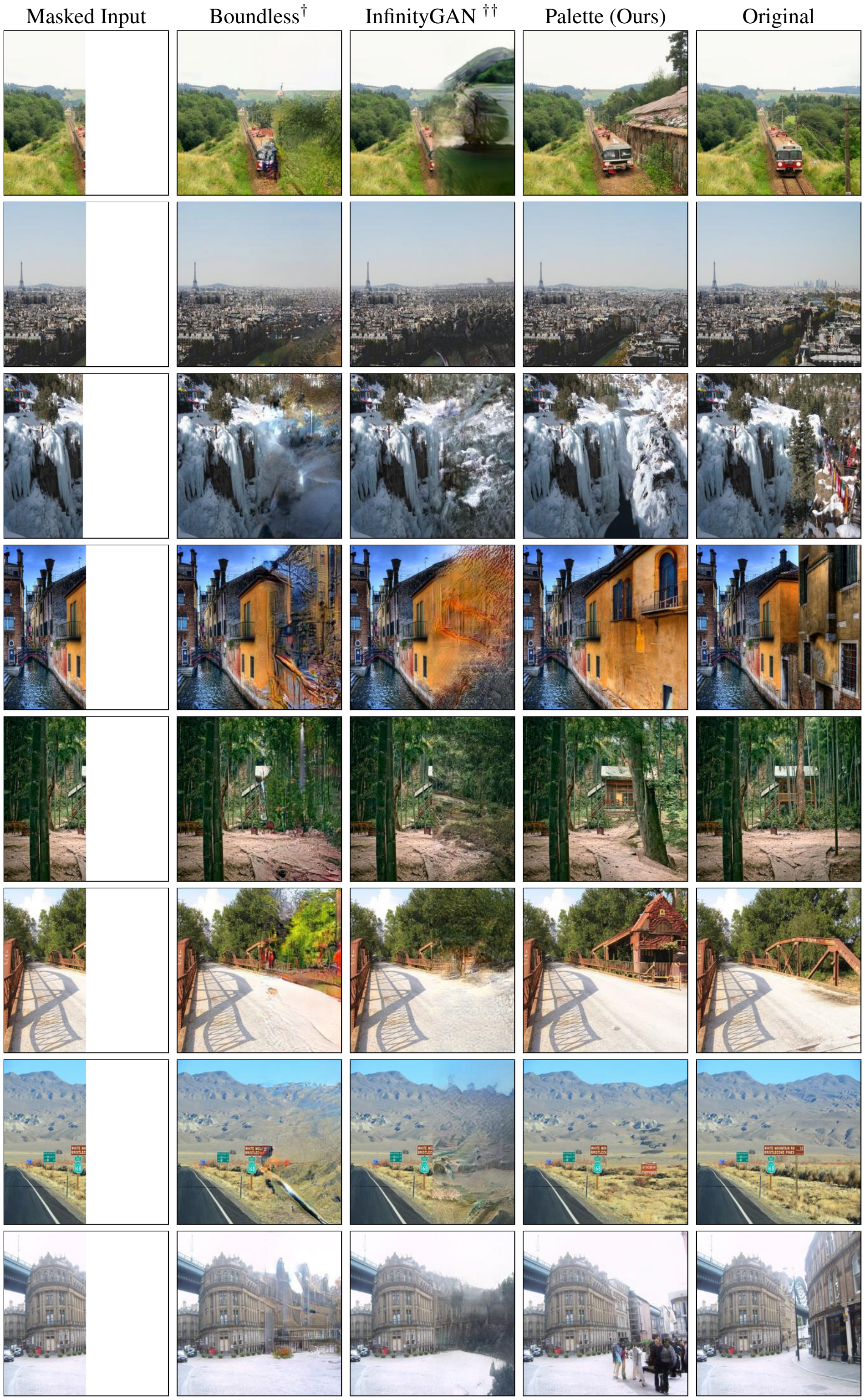

Figure C.7: Image uncropping results on Places2 validation images. Baselines: Boundless† [Teterwak et al., 2019] and InfinityGAN†† [Lin et al., 2021] trained on a scenery subset of Places2. Samples for both baselines are generously provided by their respective authors.

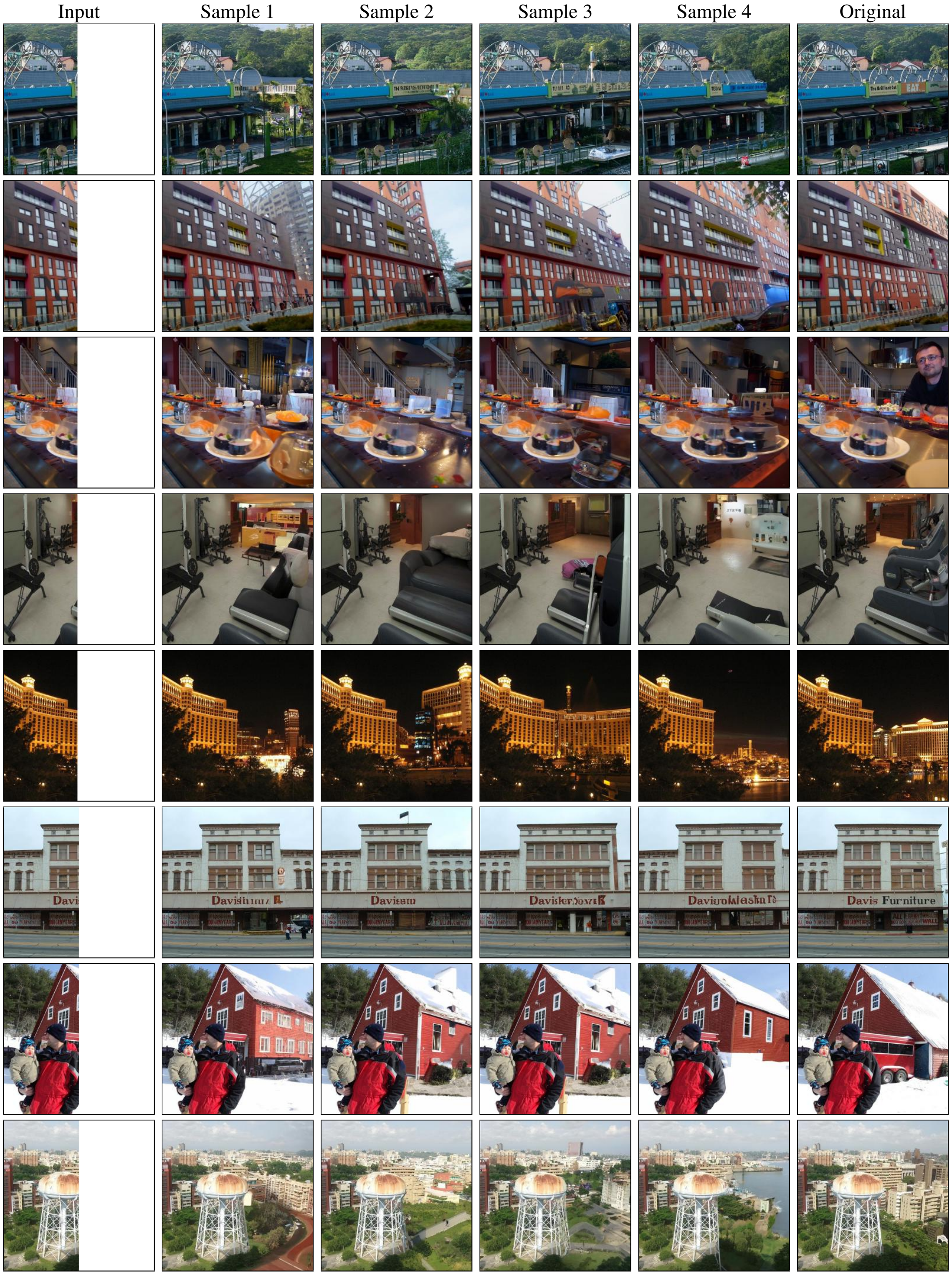

Figure C.8: Diversity of Palette outputs on Right Uncropping on Places2 dataset.

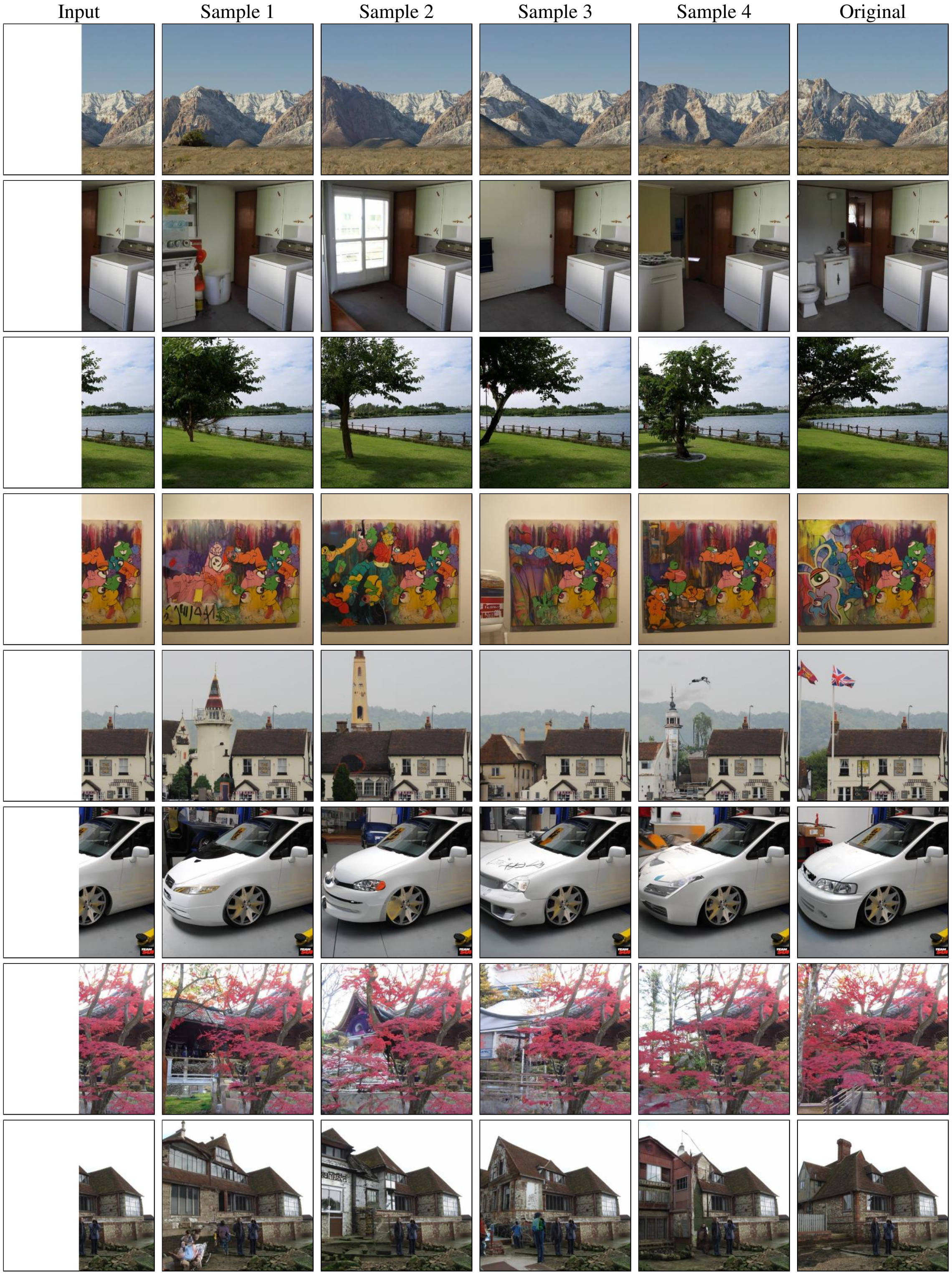

Figure C.9: Diversity of Palette outputs on Left uncropping on Places2 dataset.

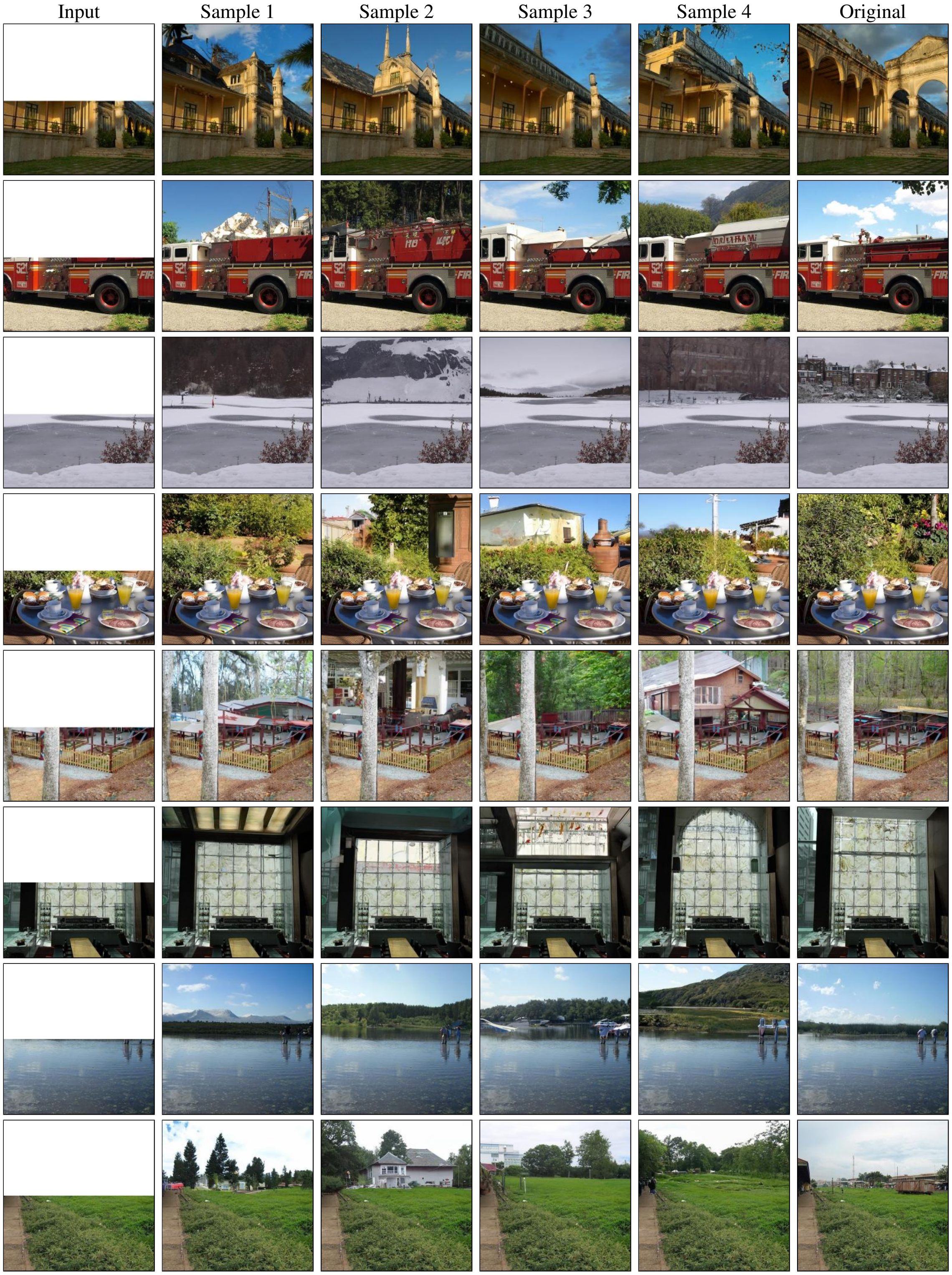

Figure C.10: Diversity of Palette outputs on Top uncropping on Places2 dataset.

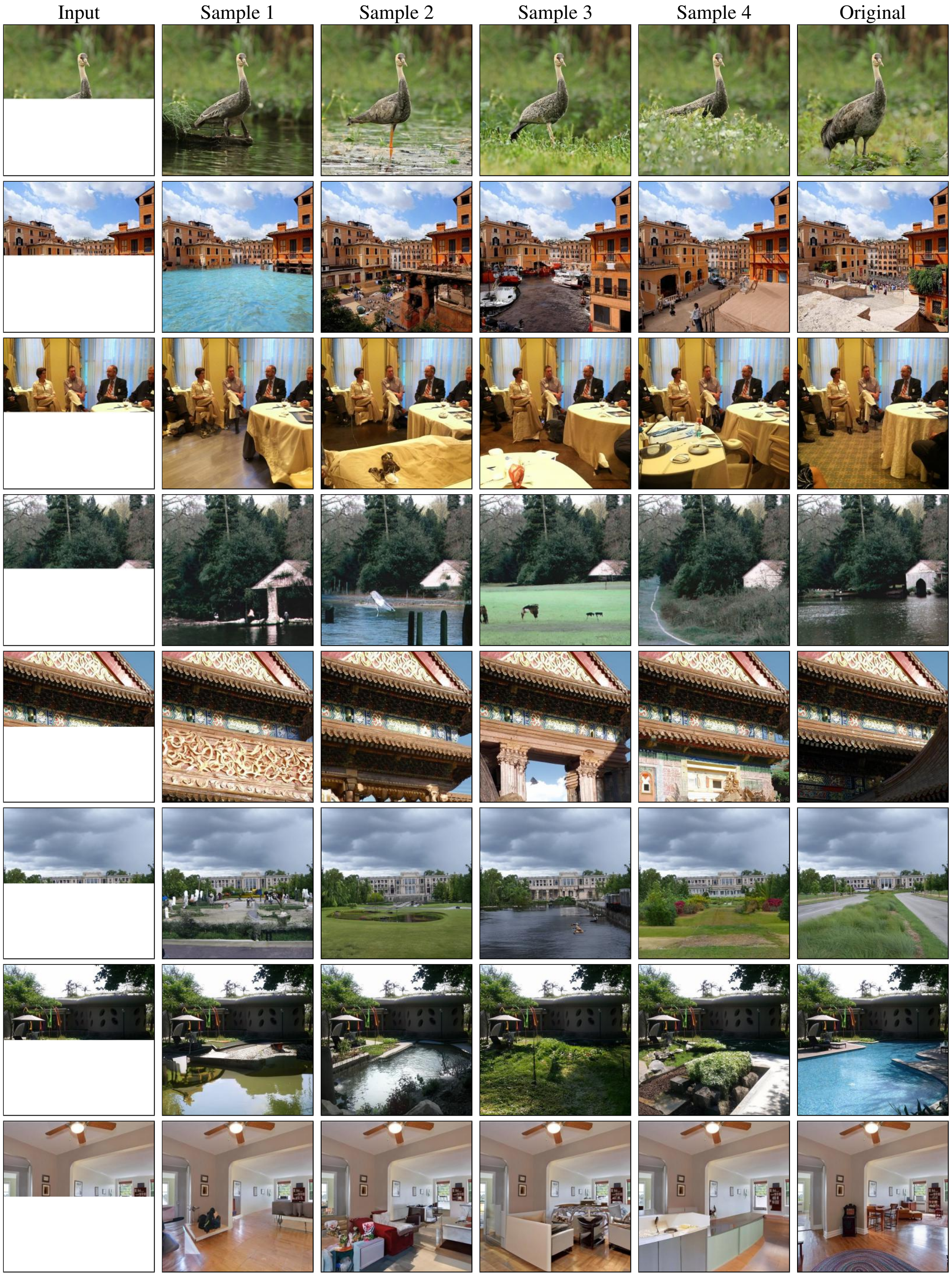

Figure C.11: Diversity of Palette outputs on Bottom uncropping on Places2 dataset.

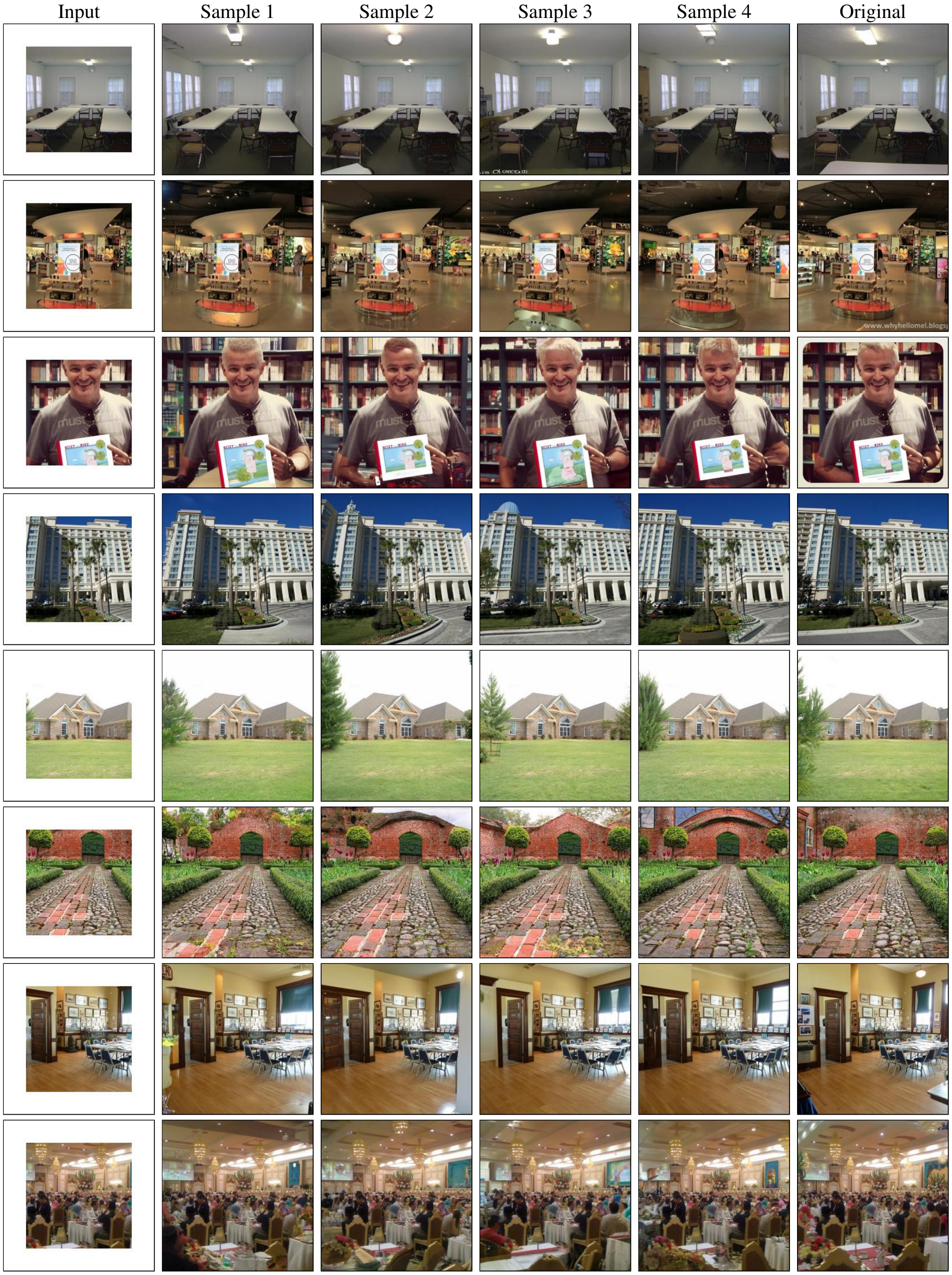

Figure C.12: Diversity of Palette outputs on Four Sided uncropping on Places2 dataset.

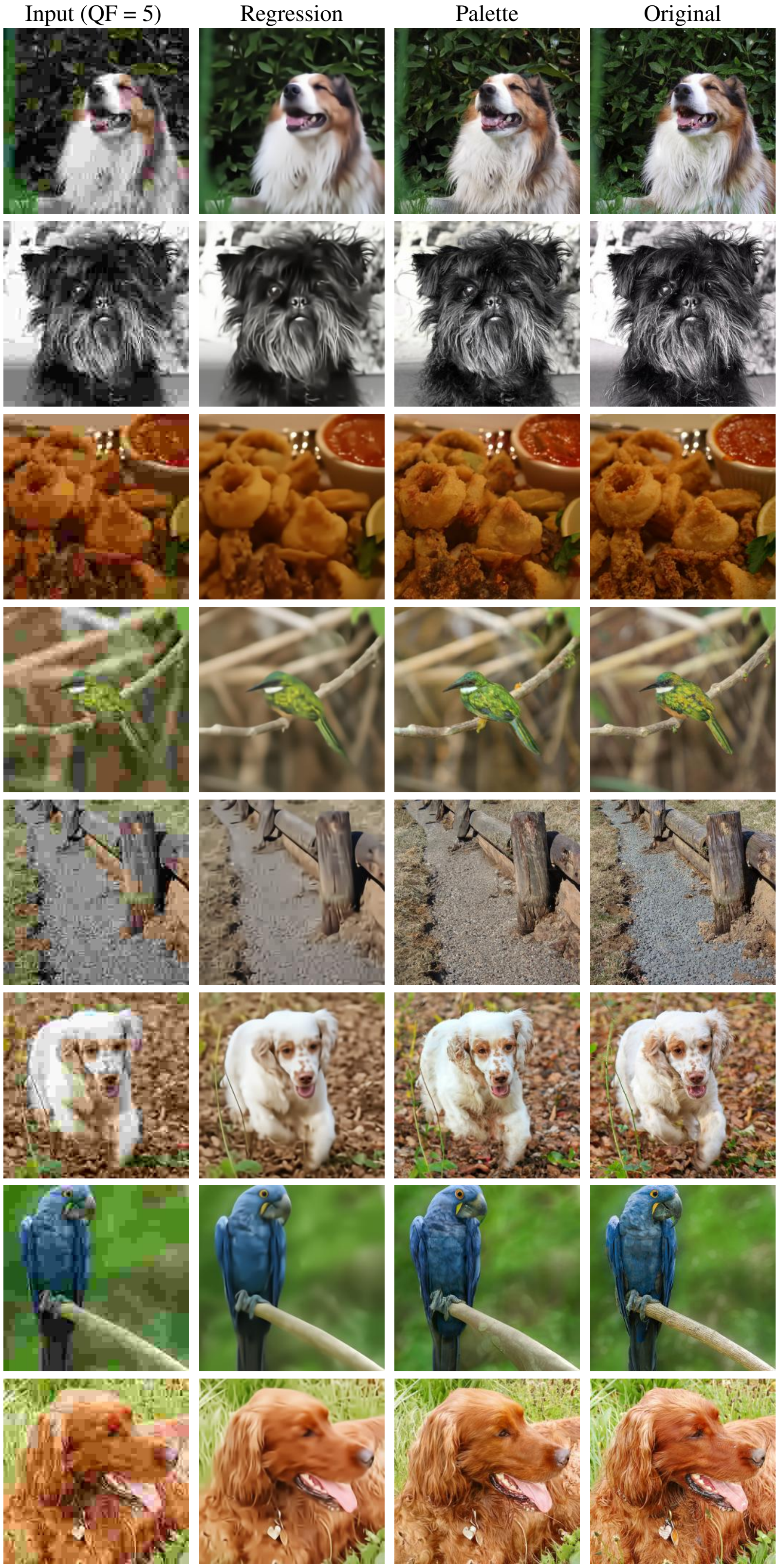

Figure C.13: JPEG Restoration results on ImageNet images.

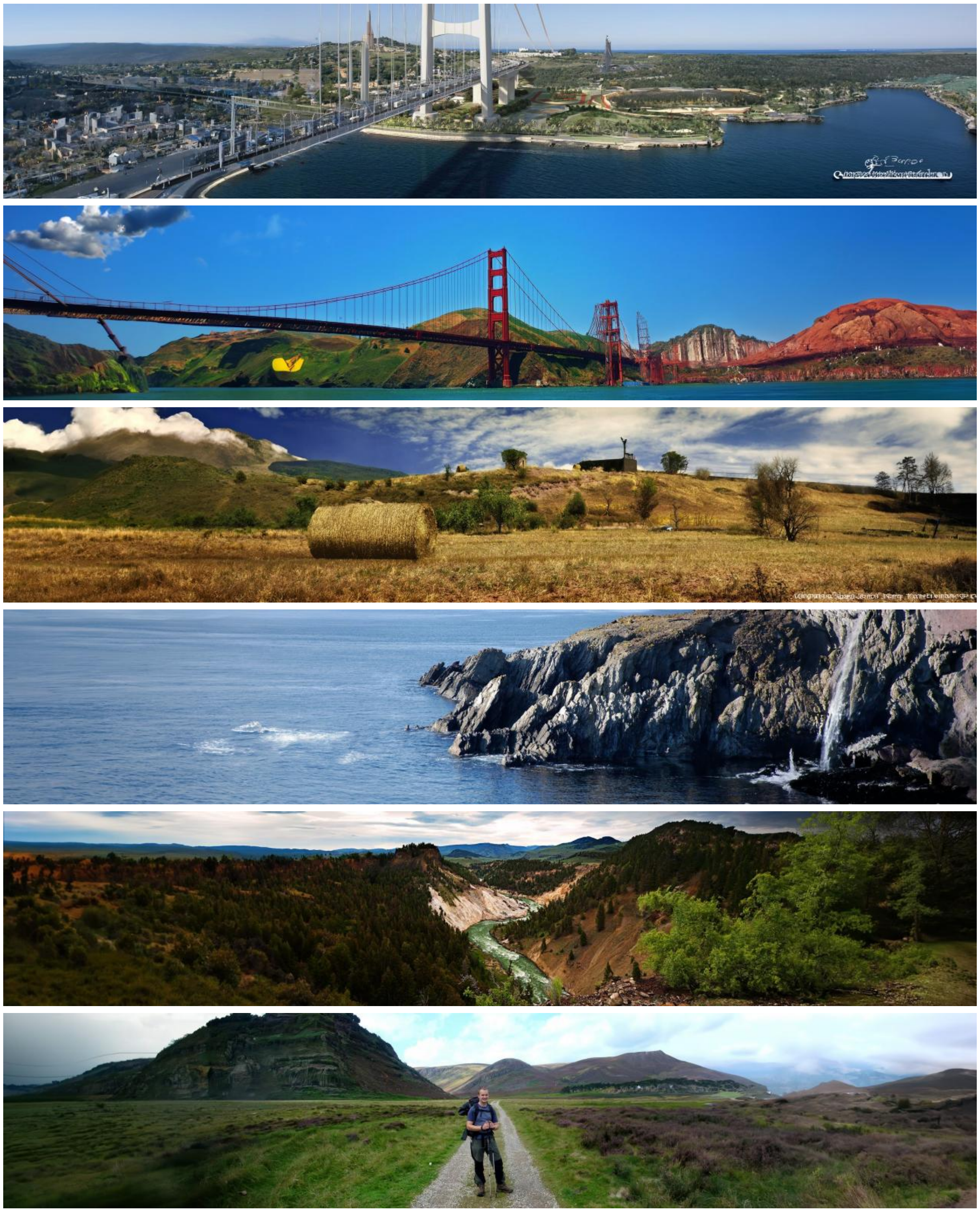

Figure C.14: Palette panorama uncropping. Given the center 256×256 pixels, we extrapolate 512 pixels to the right and to the left, in steps of 128 (via 50% uncropping tasks), yielding a 256×1280 panorama.

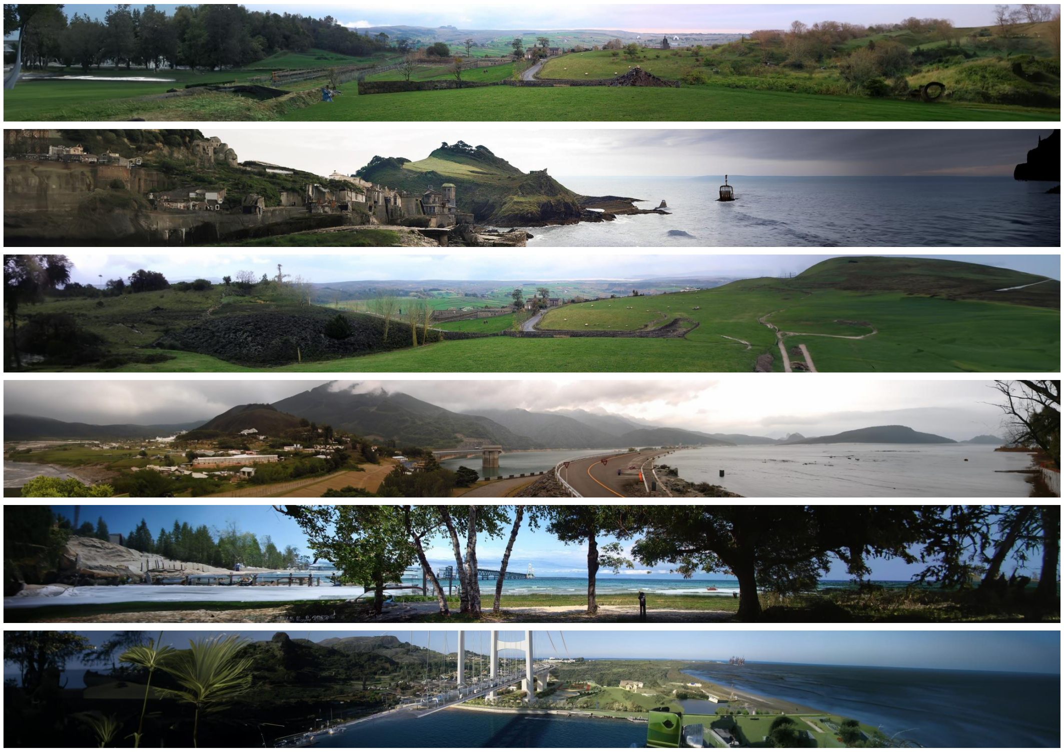

Figure C.15: Palette panorama uncropping. Given the center 256×256 pixels, we extrapolate 1024 pixels to the right and to the left, in steps of 128 (via 50% uncropping tasks), yielding a 256×2304 panorama.