HYPE-C: EVALUATING IMAGE COMPLETION MOD-ELS THROUGH STANDARDIZED CROWDSOURCING

Anonymous authors

Paper under double-blind review

Abstract

A significant obstacle to the development of new image completion models is the lack of a standardized evaluation metric that reflects human judgement. Recent work has proposed the use of human evaluation for image synthesis models, allowing for a reliable method to evaluate the visual quality of generated images. However, there does not yet exist a standardized human evaluation protocol for image completion. In this work, we propose such a protocol. We also provide experimental results of our evaluation method applied to many of the current state-of-the-art generative image models and compare these results to various automated metrics. Our evaluation yields a number of interesting findings. Notably, GAN-based image completion models are outperformed by autoregressive approaches.

1 INTRODUCTION

Image completion is a form of image generation in which a set of missing pixels within an image are filled in by a generative model conditioned on the visible pixels. It is is an important problem in machine learning and has a wide range of real-world applications, including photo repair, content-aware filling, and denoising. Image completion also acts as a useful proxy task to measure the progress of generative image models in general, as it is much easier to detect problems such as mode collapse or memorization in image completion models than in image synthesis models.

Over the past few years, an increasing number of image completion models based on GAN and autoregressive architectures have been released. But how do we best evaluate and compare these models? This question is of critical importance for measuring current progress and informing future research. Despite the development of a plethora of new models, there does not yet exist a standard-ized evaluation method appropriate for the task of image completion.

Reconstruction error metrics-such as l_1 error, l_2 error and peak signal-to-noise ratio (PSNR)-are commonly used, but are a poor method to evaluate such models. There may be many different valid ways to complete an image other than the original content, and these metrics only judge how well a model recreates the missing portion of an image in its original state.

Image synthesis models are often compared using automatic metrics based on deep networks features, such as the Frechét Inception Distance (Heusel et al., 2017) and the Inception Score (Salimans et al., 2016). Such techniques have been adapted to provide alternative metrics applicable to image completion (Zhang et al., 2018), but the use of deep networks to evaluate models is flawed in a number of ways. Not only can deep networks diverge considerably from human perception, as evidenced by their failures on adversarial examples (Nguyen et al., 2015), but the evaluation results can vary substantially depending on the architecture of the evaluation model and the dataset on which it was trained (Barratt & Sharma, 2018).

For these reasons it has been well recognized that there does not yet exist a satisfactory automatic procedure for generative model evaluation–particularly in the case of image completion–and that the gold standard remains human assessment. While it is common to use crowdsourcing platforms such as Amazon Mechanical Turk to evaluate models, these evaluations are usually ad-hoc and do not allow for a direct comparison of different models. Recent work has attempted to standardize human evaluation for image synthesis models and provide an easy to use benchmark. In particular, the HYPE protocol (Zhou et al., 2019) accepts a set of fully synthetic images, randomly mixes them with real images taken from the models training set, and asks humans on crowdsouring platforms

to rate each image as real or fake. The HYPE score of a model measures how often the generated images are rated as real by humans.

However, HYPE is not well suited to evaluating image completion models. As it does not provide standardized image masks to define the regions to complete, it is not possible to guarantee a fair and consistent score when applied to image completion. Additionally, due to the fact that HYPE does not define a standardized test set on which to perform completions and selects comparison images randomly from the training set, it is not possible to directly compare models on a per-image basis.

Our two main contributions are as follows. First, we develop a modified version of HYPE, called "HYPE-Completion (HYPE-C)", that resolves these issues and makes it possible to apply HYPE to image completion models. ¹ Second, we use HYPE-C to benchmark a number image completion models, well-known GAN architectures which we modify to perform image completion, and autoregressive models which are able to perform image completion out-of-the-box, in order to create a baseline for future work. We evaluate the models at different image resolutions on commonly used datasets, including FFHQ (Karras et al., 2019), CUB (Wah et al., 2011), LSUN-Bedrooms (Yu et al., 2015), and Stanford Cars (Krause et al., 2013). On each dataset, all models are evaluated on the same test split that is disjoint from the training images.

Additionally, our HYPE-C evaluation reveals a number of interesting findings. First, state-of-theart GAN-based image synthesis models modified to perform image completion often outperform models designed for image completion from the ground up. Second, GAN-based image completion methods are substantially outperformed by autoregressive image completion. Third, high quality image completion remains difficult for all of our evaluated methods across datasets, with the only exception of faces at lower resolutions. Finally, that a model trained to predict human evaluations directly can outperform previous automatic evaluation methods, but is still not a reasonable approximation of human judgement.

2 RELATED WORK

2.1 PER-PIXEL PERCEPTUAL METRICS

Image completion models are often evaluated using simple perceptual or reconstruction error metrics. These include metrics such as l_1 or l_2 error, as well as peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM). However, while these metrics are well suited to evaluating the quality of lossy compression algorithms or similar tasks, they are unable to provide a reasonable assessment of image completion quality as they simply measure the difference between the original and partially synthetic images at the scale of individual pixels. Under these metrics, an image completion model may receive a poor score despite being able to consistently create realistic and believable images, if the completed regions diverge significantly from the original image in pixel space.

2.2 AUTOMATIC EVALUATION METRICS FOR GENERATIVE MODELS

Many methods for evaluating generative image models have been proposed, particularly for unconditional GANs (Borji, 2019). As GANs are prone to overfitting and mode collapse (Arora et al., 2017; Arora & Zhang, 2017), it is important to not only evaluate the visual fidelity of generated images, but also their diversity. The most commonly used automatic evaluation metrics are the Inception Score (Salimans et al., 2016) and the Frechét Inception Distance (FID) (Heusel et al., 2017).

The Inception Score uses an Inception Net model pre-trained on ImageNet to label a large number of generated samples, and measures the KL divergence between the conditional label distribution $p(y \mid x)$ and the marginal distribution $p(y) = p(y \mid x = G(z))dz$. Generated images with a high visual fidelity should be more easily labeled and therefore have a conditional label distribution with low entropy, while the marginal distribution should have a high entropy when the images are diverse. Therefore, the Inception Score $\exp(\mathbb{E}_x[\mathbb{KL}(p(y \mid x) || p(y))])$ simultaneously evaluates both image

¹The supporting code used to launch HYPE-C evaluations through Amazon Mechanical Turk will be released publicly upon publication.

quality and diversity. However, the Inception Score may produce deceptive results when applied to models trained on a dataset other than ImageNet (Salimans et al., 2016; Barratt & Sharma, 2018).

The FID uses the same pre-trained Inception Net model, but does not utilize the label distributions. Instead, it embeds a set of real images and a set of generated images into the latent feature space of the Inception Net. Then, modelling the distributions of the embeddings of the real and generated images as multivariate Gaussians $\mathcal{N}(\mu_R, \Sigma_R)$ and $\mathcal{N}(\mu_G, \Sigma_R)$, the FID is the Frechét distance between them. FID is sensitive to both the visual quality of generated images as well as their diversity. However, it makes the strong assumption that the feature vectors follow a Gaussian distribution. Additionally, as in the case of the Inception Score, the FID may not provide accurate results when applied to models trained on datasets other than ImageNet.

Techniques similar to FID have been used for purposes other than GAN evaluation as well. Loss functions computed from the features of a pre-trained network are a key component of neural style transfer (Johnson et al., 2016; Jing et al., 2019) and are frequently used for super-resolution (Johnson et al., 2016; Ledig et al., 2017; Wang et al., 2018). Zhang et al. (2018) recently proposed adapting such techniques to create a perceptual distance metric, which outperforms classical metrics such as PSNR and SSIM in correlation with human evaluations. Although these techniques have produced impressive results in their respective domains, they are still reliant on pre-trained networks, meaning that they may produce vastly different results depending upon the specific network architecture used and the dataset on which the network was trained, and can only act as a rough approximation of how human beings perceive images.

2.3 HUMAN EVALUATION METRICS

Direct human evaluation through crowdsourcing platforms such as Amazon Mechanical Turk provides an attractive alternative, as it avoids many of the pitfalls of automatic metrics. The recently produced HYPE (HUMAN EYE PERCEPTUAL EVALUATION) benchmark (Zhou et al., 2019) standardized the process of performing human evaluations for generative image models, allowing for direct comparison. The first HYPE evaluation method, HYPE_{time}, tasks human evaluators from Amazon Mechanical Turk with classifying a random sequence of real and fake images. Images are shown for a varying length of time, and the generative model is scored based on the average minimum length of time required for the human evaluators to reliably discern real from fake images. The second evaluation method, HYPE_{∞}, again shows evaluators a random sequence of real and fake images does not limit the amount of time that each image is displayed, and scores models based on the percentage of images incorrectly classified by the evaluators (both real and fake).

3 EVALUATION METHOD

Although HYPE provides an effective standardized method to perform human evaluations of image synthesis models, it is not directly applicable to image completion. In order to adapt HYPE to the task of image completion, we must define a standardized set of test images, as well as standardized mask or set of masks to apply to the test images denoting which regions to complete.

For our modified HYPE evaluation protocol, we first select 100 random test images and 100 random comparison images from the dataset. We then mask the bottom half of the test images and use the respective image completion model to replace the masked portion, while leaving the comparison images unchanged. All 200 images are then shown in random order to each of the human evaluators, who classify each image as real or synthetic, and are given an unlimited amount of time to view each image. The model's HYPE-C score is simply the percentage of images (both real and completed) that are misclassified by the human evaluators. In the scenario that the model produces perfect completed images, the human evaluators are no longer able to distinguish the difference between real and partially synthetic images, reaching a HYPE-C score of 50% or more. In our experiments we use 15 human evaluators for each model evaluation. The primary difference between HYPE-C and the original HYPE is the use of a standardized test set and image mask.

Note that while masking the bottom half of each image does not necessarily replicate a real-world use case, it allows us to evaluate the widest range of model architectures, and provides a useful baseline.

Quality Control: To ensure high-quality results, we follow the approach used in HYPE of requiring evaluators to pass a qualification exam. Potential evaluators are shown a random sequence of 50 real and 50 synthetic images, and must correctly classify at least 65 of the given images. Additionally, to encourage workers to remain fully focused on their given tasks, we pay evaluators a bonus of \$0.02 for every correctly classified image.

4 EXPERIMENTS

4.1 DATASETS

We trained models on the Flickr-Faces-HQ (FFHQ) dataset (Karras et al., 2019), the Caltech-UCSD Birds-200-2011 (CUB) dataset (Wah et al., 2011), the Stanford Cars dataset (Krause et al., 2013), the LSUN-Bedroom dataset (Yu et al., 2015), and the LSUN-Cat dataset (Yu et al., 2015) at 32x32, 64x4, and 128x128 resolution.

Train and Test Split: From each dataset, we randomly selected 100 test images and 100 comparison images from the dataset's predefined test-set if available. For datasets that did not have a predefined test-set, we generated a new train-validation-test split, with 100 test images and 100 comparison images set aside for HYPE-C evaluation. The same train-validation-test splits were used for the evaluation of all models at all resolutions.

Inpainting Method	Model	Feasible Resolutions		
Latent Space Search	StyleGAN (Karras et al., 2019) ProGAN (Karras et al., 2018) WGAN-GP (Gulrajani et al., 2017)	32x32, 64x64, 128x128 32x32, 64x64, 128x128 32x32, 64x64, 128x128 32x32, 64x64, 128x128		
Conditional GAN	Conditional StyleGAN (Karras et al., 2019) Conditional ProGAN (Karras et al., 2018) Conditional WGAN-GP (Gulrajani et al., 2017) DeepFill (Yu et al., 2020)	32x32, 64x64, 128x128 32x32, 64x64, 128x128 32x32, 64x64, 128x128 32x32, 64x64, 128x128 32x32, 64x64, 128x128		
Autoregressive Generation	PixelCNN++ (Salimans et al., 2019) PixelSNAIL (Chen et al., 2018) Pixel Constrained CNN (Dupont & Suresha, 2018)	32x32, 64x64 32x32, 64x64 32x32, 64x64		

4.2 EVALUATED MODELS

Table 1: Model Architectures

In Table 1 we list all of the tested model architectures. We evaluated a combination of both GANbased image completion methods and autoregressive methods. The evaluated methods can be divided into three categories:

Latent Space Search: We adapt several unconditional GAN architectures to image completion using the latent space search technique proposed by Yeh et al. (2018). Given any trained GAN model, we search the latent space for an optimal latent vector, one that maximizes the discriminator score of the generated image, while minimizing the difference between the generated and the input image.

Conditional GAN Generation: We also adapt unconditional GAN architecture to perform image completion by modifying the architecture directly. We swap the latent inputs to the generator with features conditioned on the upper-half of the image, while making minimal changes to other parts of the model. More specifically, we adopt an encoder-decoder architecture as the generator, where the encoder consists of multiple convolutional layers, and the decoder is the one used in the original generator. Skip connections are made to propagate the high-resolution details from the encoder to the decoder. In terms of training objectives, in addition to the adversarial term, the generator is also trained with an image reconstruction term that minimizes the difference between the inpainted and the original image (Wang et al., 2020). Additionally, we test the DeepFill architecture (Yu et al., 2020), which is designed for image completion.

Autoregressive Generation: Our evaluation method can in general be directly applied to autoregressive generative models. As PixelCNN++ (Salimans et al., 2019) and PixelSNAIL (Chen et al., 2018) generate images pixel-by-pixel in raster scan order, it is not necessary to modify the model architecture to generate the bottom half of test images. Additionally, Pixel Constrained CNN (Dupont & Suresha, 2018) is an autoregressive model specifically designed for image completion and requires no modification for our evaluation.

4.3 EVALUATION RESULTS

Model	FFHQ	Stanford Cars
StyleGAN (Karras et al., 2019)	0.119	0.126
ProGAN (Karras et al., 2018)	0.303	0.159
WGAN-GP (Gulrajani et al., 2017)	0.095	0.164
Conditional StyleGAN (Karras et al., 2019)	0.453	0.160
Conditional ProGAN (Karras et al., 2018)	0.440	0.111
Conditional WGAN-GP (Gulrajani et al., 2017)	0.156	0.135
DeepFill (Yu et al., 2020)	0.259	0.080
PixelCNN++ (Salimans et al., 2019)	0.486	0.306
PixelSNAIL (Chen et al., 2018)	0.458	0.293
Pixel Constrained CNN (Dupont & Suresha, 2018)	0.200	0.166

Table 2: HYPE-C scores of different models on FFHQ and Stanford Cars at 32x32 resolution.

In order to identify and filter out models that are not capable of higher resolution image generation, we first performed an initial set of evaluations of all model architectures on the FFHQ and Stanford Cars datasets at 32x32 resolution, the results of which can be seen in Table 2. Autoregressive methods in particular are often unable to cope with high resolution image generation. On the other hand, autoregression and conditional GAN models achieve the best results. For a qualitative comparison of images completed by the selected models, see Figure 1.

Based on these initial results, we selected four top-performing model architectures (ProGAN, Con-StyleGAN, Con-ProGAN, and PixelCNN++), including at least one model from each category of image inpainting methods. We then evaluated these models on the remaining datasets and resolutions (Table 3). Figure 2, 3 show qualitative examples. More are provided in appendix.

Resolution	Dataset	ProGAN	Con-StyleGAN	Con-ProGAN	PixelCNN++
32x32	FFHQ	0.303	0.453	0.4403	0.486
	Stanford Cars	0.159	0.160	0.111	0.306
	CUB	0.331	0.399	0.377	0.376
	LSUN-Bedroom	0.387	0.310	0.415	0.465
	LSUN-Cat	0.353	0.403	0.407	0.461
64x64	FFHQ	0.123	0.307	0.374	0.369
	Stanford Cars	0.088	0.086	0.138	0.239
	CUB	0.122	0.156	0.303	0.337
	LSUN-Bedroom	0.245	0.303	0.352	0.283
	LSUN-Cat	0.129	0.368	0.317	0.374
128x128	FFHQ	0.077	0.218	0.298	X
	Stanford Cars	0.047	0.084	0.050	Х
	CUB	0.106	0.146	0.167	Х
	LSUN-Bedroom	0.087	0.214	0.269	Х
	LSUN-Cat	0.109	0.252	0.227	Х

Table 3: Full results for ProGAN, Con-StyleGAN, Con-ProGAN, and PixelCNN++.

There are a few notable aspects of these results. First, our modified GAN architectures are able to outperform models designed for image completion. Second, PixelCNN++ outperforms all of the other models on every dataset at both 32x32 and 64x64 resolution. Third, even at these low resolutions, the evaluated models perform poorly on all datasets, with the exception of FFHQ at 32x32 resolution. The current state-of-the-art unconditional GAN models are capable of generating realistic images at a resolution of 1024x1024 or higher (Karras et al., 2019; 2018), yet perform poorly when they are tasked with completing half of a low resolution image.



Figure 1: Qualitative comparison between different methods at 32x32 resolution. More results can be found in the appendix.



Figure 2: Qualitative results at 64x64 resolution.

Figure 3: Qualitative results at 128x128 resolution.

4.4 COMPARISON WITH AUTOMATIC METRICS

We compare FID scores with HYPE-C for each of our evaluated models in Figure 4 in order to determine how well corresponds to human perception. To evaluate an unconditional GAN model, one would usually generate a large set of synthetic images and then compute the FID between that set and the model's training set. Since we are evaluating image completion models, we complete a subset of the training set and compute the FID between it and the full unmodified training set. We complete the same subset of images when evaluating all models at all resolutions.

Since we are performing image completions, we are limited to generating a set of samples of equal size to the training set with which to compute activation statistic. This is particularly problematic with smaller datasets such as Stanford Cars and CUB, where the training set may have only a few thousand images. Due to the increased computational cost of image completion compared to image generation, we also limit the size of our samples to at most 10,000 images per dataset.



Figure 4: FID vs HYPE-C Scores

It is clear from the figure that the evaluations of model performance using FID and HYPE-C do not coincide. This is not entirely unexpected, as $HYPE_{time}$ and $HYPE_{\infty}$ scores were shown to be uncorrelated with FID by Zhour et al. (Zhou et al., 2019). However, FID measures both image quality and diversity, while $HYPE_{time}$ and $HYPE_{\infty}$ only measure image quality, making it somewhat of an unfair comparison. Since HYPE-C evaluates models based on their ability to complete a set of images representative of the target distribution, it is sensitive to image diversity, and we can much more directly compare our results with FID.

We use Spearman rank correlation coefficients to determine how strongly FID correlates with HYPE-C scores across all models at each resolution. We also measure the correlation between the human evaluations on individual images and four different perceptual metrics – perceptual distance (Zhang et al., 2018), peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM), and mean squared error (MSE). The results can be found in Table 4. A value of $\rho = \pm 1$ indicates a perfect positive or negative correlation, while a value of $\rho = 0$ indicates no correlation. We see that FID is nearly completely uncorrelated with HYPE-C, while the perceptual metrics achieve no more than a weak correlation with human evaluations.

Metric	32x32		64x64		128x128	
	ρ	p	ρ	p	ρ	p
FID	-0.003	0.991	-0.052	0.850	-0.028	0.931
Perceptual Distance	-0.342	0.00	-0.322	0.00	-0.306	0.00
PSNR	0.133	0.00	0.196	0.00	0.240	0.00
SSIM	0.278	0.00	0.373	0.00	0.326	0.00
MSE	-0.133	0.00	-0.196	0.00	-0.240	0.00

Table 4: Spearman Rank Correlation Coefficients of human evaluations and automatic metrics

4.5 HUMAN SCORE PREDICTOR

We attempt to create an automatic evaluation metric in a way similar to the perceptual distance metric (Zhang et al., 2018). We created training, validation, and test sets using the full set of completed image from our earlier experiments labeled by their individual average human HYPE-C score. We then trained multiple model architectures to predict the human ratings of the completed images.

We tested a variety of core CNN architectures, including Inception Net, AlexNet, ResNet, and VGG. For each architecture, we replaced the final classification layer with a small fully connected network. To prevent over-training, we used early stopping triggered by a stalled validation score. We tested one version of each model where the core CNN weights are held constant, and one where the core CNN is fine-tuned. The results can be found in Table 5 and Table 6.

Model	l_1 Error	ρ	p	_	Model	l_1 Error	ρ	p
Inception Net V3	0.145	0.764	0.00		Inception Net V3	0.142	0.773	0.00
AlexNet	0.186	0.637	0.00		AlexNet	0.162	0.704	0.00
ResNet-18	0.158	0.728	0.00		ResNet-18	0.139	0.783	0.00
ResNet-34	0.143	0.766	0.00		ResNet-34	0.139	0.772	0.00
ResNet-50	0.141	0.778	0.00		ResNet-50	0.140	0.771	0.00
ResNet-101	0.136	0.784	0.00		ResNet-101	0.141	0.774	0.00
ResNet-152	0.132	0.795	0.00		ResNet-152	0.138	0.777	0.00
VGG-11	0.183	0.642	0.00		VGG-11	0.143	0.764	0.00
VGG-13	0.183	0.631	0.00		VGG-13	0.160	0.710	0.00
VGG-16	0.179	0.648	0.00		VGG-16	0.207	0.523	0.00
VGG-19	0.182	0.640	0.00		VGG-19	0.215	0.517	0.00

networks as feature extractors.

Table 5: Score prediction using pre-trained Table 6: Score prediction with fine-tuning of feature extraction networks.

While these results are a substantial improvement over the automatic metrics we have previously discussed, our score predictor models are not necessarily sufficient to act as a replacement for evaluation by real humans. They will also be vulnerable to the same issues as other metrics based on neural network features, namely that they may produce deceiving results when applied to other datasets, and may have certain "blindspots" with respect to specific images i.e. adversarial examples.

5 CONCLUSION

In this work we introduced HYPE-C, a modified form of HYPE capable of evaluating image completion models. We provided both qualitative and quantitative experimental results of HYPE-C applied to a variety of image completion models, forming a baseline for comparison. We showed that wellknown GAN-based image synthesis models modified to perform image completion can outperform more complex methods in our setting and that autoregressive models can often outperform GANs in terms of human evaluation at low resolutions. Using our evaluation method, we were able to perform an analysis of the efficacy of automatic evaluation metrics, and show that they only weakly correlate with human evaluations. Finally, we evaluated the use of features extracted from a variety of pre-trained networks as a means to create a proxy for human evaluation.

REFERENCES

- Sanjeev Arora and Yi Zhang. Do GANs actually learn the distribution? An empirical study. pp. 1–11, 2017. URL http://arxiv.org/abs/1706.08224.
- Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). 34th International Conference on Machine Learning, ICML 2017, 1:322–349, 2017.
- Shane Barratt and Rishi Sharma. A Note on the Inception Score. 2018. URL http://arxiv.org/abs/1801.01973.
- Ali Borji. Pros and cons of GAN evaluation measures. Computer Vision and Image Understanding, 179:41–65, 2019. ISSN 1090235X. doi: 10.1016/j.cviu.2018.10.009.
- Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. PixelSNAIL: An improved autoregressive generative model. 35th International Conference on Machine Learning, ICML 2018, 2:1364–1372, 2018. URL https://github.com/neocxi/pixelsnail-public.http://arxiv.org/abs/1712.09763.
- Emilien Dupont and Suhas Suresha. Probabilistic Semantic Inpainting with Pixel Constrained CNNs. 89, 2018. URL http://arxiv.org/abs/1810.03728.
- Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (eds.), Advances in Neural Information Processing Systems, volume 2017-Decem, pp. 5768–5778. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.pdf.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):6627–6638, 2017. ISSN 10495258.
- Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 2019.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.
- Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, pp. 1–26, 2018.
- Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision* and Pattern Recognition, 2019-June:4396–4405, 2019. ISSN 10636919. doi: 10.1109/CVPR. 2019.00453.
- Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia, 2013.
- 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.
- Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015:427–436, 2015. ISSN 10636919. doi: 10.1109/CVPR.2015.7298640.
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pp. 2234–2242, 2016.

- Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. PixelCNN++: Improving the PixelCnn with discretized logistic mixture likelihood and other modifications. 5th International Conference on Learning Representations, ICLR 2017 Conference Track Proceedings, 2019. URL http://arxiv.org/abs/1701.05517.
- C Wah, S Branson, P Welinder, P Perona, and S Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.
- Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings* of the European Conference on Computer Vision (ECCV), pp. 0–0, 2018.
- Yi Wang, Ying-Cong Chen, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Attentive normalization for conditional image generation. *arXiv preprint arXiv:2004.03828*, 2020.
- Raymond A. Yeh, Teck Yian Lim, Chen Chen, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minhn Do. Image Restoration with Deep Generative Models. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018-April:6772–6776, 2018. ISSN 15206149. doi: 10.1109/ICASSP.2018.8462317.
- Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, 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.
- Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-Form Image Inpainting With Gated Convolution. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4470–4479, 2020. doi: 10.1109/iccv.2019.00457.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.
- Sharon Zhou, Mitchell L. Gordon, Ranjay Krishna, Austin Narcomey, Durim Morina, and Michael S. Bernstein. Hype: Human eye perceptual evaluation of generative models. *Deep Generative Models for Highly Structured Data*, DGS@ICLR 2019 Workshop, (NeurIPS), 2019.

A QUALITATIVE AND AUTOMATIC METRIC RESULTS

- A.1 FFHQ
- A.1.1 32x32



Figure 5: Ground truth for FFHQ at 32x32 resolution.





Human: 0.000 Perceptual: 3.431 MSE: 525.703 PSNR: 20.923 SSIM: 0.576



Human: 0.000 Perceptual: 3.239 MSE: 2988.611 PSNR: 13.376 SSIM: 0.621



Human: 0.000 Perceptual: 2. MSE: 1393.11 PSNR: 16.691 SSIM: 0.732





Human: 0.067 Perceptual: 2.5 MSE: 4378.838 PSNR: 11.717 SSIM: 0.614



Human: 0.600 Perceptual: 1.968 MSE: 1236.185 PSNR: 17.210 SSIM: 0.680



Human: 0.067 Perceptual: 2.304 MSE: 1188.358 PSNR: 17.381 SSIM: 0.622



Human: 0.133 Perceptual: 2.219 MSE: 3140.401 PSNR: 13.161 SSIM: 0.533





Human: 0.067 Perceptual: 3.604 MSE: 3631.109 PSNR: 12.530 SSIM: 0.609



Human: 0.400 Perceptual: 1.313 MSE: 792.463 PSNR: 19.141 SSIM: 0.733





Figure 7: FFHQ 32x32 results for ProGAN.



Figure 8: FFHQ 32x32 results for WGAN-GP.



Figure 9: FFHQ 32x32 results for Conditional StyleGAN.



Figure 10: FFHQ 32x32 results for Conditional ProGAN.



Figure 11: FFHQ 32x32 results for Conditional WGAN-GP.



Figure 12: FFHQ 32x32 results for DeepFill.



Perceptual: 0. MSE: 487.326 PSNR: 21.253 SSIM: 0.867







Human: 0.800 Perceptual: 1.992 MSE: 3662.557 PSNR: 12.493 SSIM: 0.743



Human: 0.733 Perceptual: 2.0 MSE: 1338.182 PSNR: 16.866 SSIM: 0.809





Human: 0.667 Perceptual: 1.380 MSE: 870.726 PSNR: 18.732 SSIM: 0.780



Human: 0.667 Perceptual: 1.534 MSE: 1090.963 PSNR: 17.753 SSIM: 0.735



Human: 0.733 Perceptual: 1.573 MSE: 1805.711 PSNR: 15.564 SSIM: 0.727



Human: 0.933 Perceptual: 1.185 MSE: 1797.344 PSNR: 15.584 SSIM: 0.792





Perceptual: 0.933 Perceptual: 0.469 MSE: 444.485 PSNR: 21.652 SSIM: 0.939



Perceptual: 1.939 MSE: 1316.162 PSNR: 16.938 SSIM: 0.725



Figure 13: FFHQ 32x32 results for PixelCNN++.

19



Human: 0.800 Perceptual: 1.752 MSE: 1863.583 PSNR: 15.427 SSIM: 0.783







Figure 14: FFHQ 32x32 results for PixelSNAIL.



Figure 15: FFHQ 32x32 results for Pixel Constrained CNN.

A.1.2 64x64



Figure 16: Ground truth for FFHQ at 64x64 resolution.























Human: 0.000 Perceptual: 2.780 MSE: 1479.806 PSNR: 16.429 SSIM: 0.668











Human: 0.200 Perceptual: 2.3 MSE: 1398.785 PSNR: 16.673 SSIM: 0.657









Figure 17: FFHQ 64x64 results for ProGAN.



Figure 18: FFHQ 64x64 results for Conditional StyleGAN.



Figure 19: FFHQ 64x64 results for Conditional ProGAN.



Figure 20: FFHQ 64x64 results for PixelCNN++.

```
A.1.3 128x128
```



Figure 21: Ground truth for FFHQ at 128x128 resolution.



Human: 0.000 Perceptual: 3.331 MSE: 438.432 PSNR: 21.712 SSIM: 0.724





Human: 0.000 Perceptual: 3.36 MSE: 2828.716 PSNR: 13.615 SSIM: 0.580



Human: 0.067 Perceptual: 3.480 MSE: 2844.778 PSNR: 13.590 SSIM: 0.575



Human: 0.000 Perceptual: 3.611 MSE: 1789.822 PSNR: 15.603 SSIM: 0.608



Human: 0.533 Perceptual: 2.9 MSE: 2004.922 PSNR: 15.110 SSIM: 0.637







Human: 0.000 Perceptual: 3.145 MSE: 1645.202 PSNR: 15.969 SSIM: 0.638

Perceptual: 2.637 MSE: 839.892 PSNR: 18.889 SSIM: 0.701

Figure 22: FFHQ 128x128 results for ProGAN.

Human: 0.067 Perceptual: 2.709 MSE: 2742.653 PSNR: 13.749 SSIM: 0.660

.091

Perceptual: 2. MSE: 957.700 PSNR: 18.319 SSIM: 0.735

Human: 0.067 Perceptual: 2.697 MSE: 2154.122 PSNR: 14.798 SSIM: 0.640

Perceptual: 3.7 MSE: 2329.184 PSNR: 14.459 SSIM: 0.560

Perceptual: 3.125 MSE: 5378.356 PSNR: 10.824 SSIM: 0.583

Perceptual: 3.519 MSE: 1774.299 PSNR: 15.641 SSIM: 0.649

Human: 0.000 Perceptual: 3.682 MSE: 2580.470 PSNR: 14.014 SSIM: 0.617



Figure 23: FFHQ 128x128 results for Conditional StyleGAN.



Figure 24: FFHQ 128x128 results for Conditional ProGAN.

A.2 STANFORD CARS

A.2.1 32x32



Figure 25: Ground truth for Stanford Cars at 32x32 resolution.



Figure 26: Stanford Cars 32x32 results for StyleGAN.



Figure 27: Stanford Cars 32x32 results for ProGAN.



Figure 28: Stanford Cars 32x32 results for WGAN-GP.



Figure 29: Stanford Cars 32x32 results for Conditional StyleGAN.



Figure 30: Stanford Cars 32x32 results for Conditional ProGAN.


Figure 31: Stanford Cars 32x32 results for Conditional WGAN-GP.



Figure 32: Stanford Cars 32x32 results for DeepFill.



Figure 33: Stanford Cars 32x32 results for PixelCNN++.



Figure 34: Stanford Cars 32x32 results for PixelSNAIL.



Figure 35: Stanford Cars 32x32 results for Pixel Constrained CNN.

A.2.2 64x64



Figure 36: Ground truth for Stanford Cars at 64x64 resolution.



Figure 37: Stanford Cars 64x64 results for ProGAN.



Figure 38: Stanford Cars 64x64 results for Conditional StyleGAN.



Figure 39: Stanford Cars 64x64 results for Conditional ProGAN.



Figure 40: Stanford Cars 64x64 results for PixelCNN++.

```
A.2.3 128x128
```



Figure 41: Ground truth for Stanford Cars at 128x128 resolution.



Figure 42: Stanford Cars 128x128 results for ProGAN.



Figure 43: Stanford Cars 128x128 results for Conditional StyleGAN.



Figure 44: Stanford Cars 128x128 results for Conditional ProGAN.

A.3 CUB

A.3.1 32x32



Figure 45: Ground truth for CUB at 32x32 resolution.



Figure 46: CUB 32x32 results for ProGAN.



Figure 47: CUB 32x32 results for Conditional StyleGAN.

Human: 0.200	Human: 0.133	Human: 0.667	Human: 0.133	Human: 0.733	Human: 0.600	Human: 0.067	Human: 0.733	Human: 0.733	Human: 0.400
Perceptual: 3.290	Perceptual: 4.014	Perceptual: 2.888	Perceptual: 2.571	Perceptual: 3.754	Perceptual: 3.406	Perceptual: 3.810	Perceptual: 1.831	Perceptual: 4.000	Perceptual: 2.326
MSE: 1301.679	MSE: 985.025	M5E: 320.294	M5E: 415:935	MSE: 2694.605	M5E: 987.857	MSE: 3125.396	M5: 1.937.601	MSE: 898.065	MSE: 740.546
PSNR: 16.986	PSNR: 18.196	PSINI: 23.075	P5NR: 21.941	PSNR: 13.826	PSNR: 18.184	PSNR: 13.182	PSNR: 15.349	PSNR: 18.598	PSNR: 19.435
SSIM: 0.549	SSIM: 0.535	SSIMI: 0.601	SSIM: 0.637	SSIM: 0.551	SSIM: 0.513	SSIM: 0.597	SSIM: 0.537	SSIM: 0.538	SSIM: 0.570
Human: 0.667	Human: 0.133	Human: 0.467	Human: 0.533	Human: 0.067	Human: 0.867	Human: 0.667	Human: 0.667	Human: 0.533	Human: 0.533
Perceptual: 2.761	Perceptual: 2.541	Perceptual: 3.558	Perceptual: 3.292	Perceptual: 3.975	Perceptuai: 3.649	Perceptual: 3.352	Perceptual: 4.086	Perceptuai: 3.838	Perceptual: 2.861
MSE: 1339.244	MSE: 920.854	MSE: 448.43	MSE: 1413.619	MSE: 349.194	MSE: 495.082	M55: 1110.643	MSE: 1397.209	MSE: 2230.210	MSE: 729.569
PSRN: 10.862	PSNR: 18.489	PSNR: 21.610	PSNR: 16.627	PSNR: 22.700	PSNR: 21.184	PSNR: 17.675	PSNR: 16.678	PSNR: 14.647	PSNR: 19.500
SSIM: 0.529	SSIM: 0.580	SSIM: 0.520	SSIM: 0.551	SSIM: 0.668	SSIM: 0.619	SSIM: 0.624	SSIM: 0.573	SSIM: 0.542	SSIM: 0.570
Human: 0.733	Human: 0.800	Human: 0.333	Human: 0.133	Human: 0.267	Human: 0.667	Human: 0.800	Human: 0.600	Human: 0.933	Human: 0.600
Perceptual: 3.973	Perceptual: 2.083	Perceptual: 4.003	Perceptual: 3.071	Perceptual: 3.844	Perceptual: 2.893	Perceptual: 3.587	Perceptual: 2.702	Perceptual: 4.002	Perceptual: 2.997
M5E: 2652.396	MSE: 771.501	MSE: 942.738	MSE: 2757.755	MSF: 2120.033	MSE: 351.644	MSE: 1267.974	MSE: 850.855	MSE: 712.262	MSE: 634.333
PSNR: 13.894	PSNR: 19.257	PSNR: 18.887	PSNR: 13.725	PSNR: 14.867	PSNR: 22.670	PSNR: 17.100	PSNR: 18.832	PSNR: 13.604	PSNR: 20.108
SSIM: 0.615	SSIM: 0.483	SSIM: 0.461	SSIM: 0.581	SSIM: 0.600	SSIM: 0.700	SSIM: 0.563	SSIM: 0.652	SSIM: 0.621	SSIM: 0.609
Human: 0.800	Human: 0.733	Human: 0.400	Human: 0.733	Human: 0.067	Human: 0.467	Human: 0.467	Human: 0.667	Human: 0.067	Human: 0.600
Perceptual: 3.590	Perceptual: 3.536	Perceptual: 3.006	Perceptual: 2.486	Perceptual: 2.572	Perceptual: 3.162	Perceptual: 3.135	Perceptual: 4.640	Perceptual: 2.523	Perceptual: 4.277
MSE: 1673.106	MSE: 1636.627	MSE: 940.016	MSE: 1887.801	MSE: 522.765	MSE: 313.701	MSE: 2066.755	MSF: 2151.743	MSE: 75.632	MSE: 1977.041
PSNR: 15.896	PSNR: 15.991	PSNR: 18.399	PSNR: 15.371	PSNR: 20.948	PSNR: 23.166	PSNR: 14.999	PSNR: 14.803	PSNR: 19.342	PSNR: 15.396
SSIM: 0.557	SSIM: 0.594	SSIM: 0.687	SSIM: 0.603	SSIM: 0.541	SSIM: 0.758	SSIM: 0.558	SSIM: 0.620	SSIM: 0.486	SSIM: 0.571
Human: 0.400	Human: 0.933	Human: 0.733	Human: 0.667	Human: 0.733	Human: 0.600	Human: 0.333	Human: 0.600	Human: 0.733	Human: 0.800
Perceptual: 2,781	Perceptual: 2.724	Perceptual: 2.658	Perceptual: 4.328	Perceptual: 3.571	Perceptual: 4.450	Perceptual: 3.353	Perceptual: 3.204	Perceptual: 4.315	Perceptual: 2.804
MSE: 560.967	MSE: 1013.587	MSE: 392.337	MSE: 366.341	MSE: 1296.415	MSE: 1265.938	M5E: 369.777	MSE: 711.830	MSE: 1066.097	MSE: 1776.638
PSNR: 20.641	PSNR: 18.072	PSNR: 22.194	PSNR: 22.492	PSNR: 17.003	PSNR: 17.107	PSNR: 22.451	PSNR: 19.607	PSNR: 17.732	PSNR: 15.635
SSIM: 0.545	SSIM: 0.640	SSIM: 0.641	SSIM: 0.601	SSIM: 0.592	SSIM: 0.588	SSIM: 0.669	SSIM: 0.620	SSIM: 0.621	SSIM: 0.589
Human: 0.333 Perceptual: 4.543 MSE: 1628.263 PSNR: 15.510 SSIM: 0.596	Human: 0.333 Perceptual: 3.526 MSE: 626.075 PSNR: 20.165 SSIM: 0.529	Human: 0.000 Perceptual: 3.382 MSE: 975.977 PSNR: 18.236 SSIM: 0.545	Human: 0.133 Perceptual: 4.642 MSE: 1606.739 PSNR: 16.071 SSIM: 0.497	Human: 0.067 Perceptual: 2.872 MSE: 378.668 PSNR: 22.348 SSIM: 0.622	Human: 0.667 Perceptual: 3.155 PSN: 22.350 SSIM: 0.635	Human: 0.467 Perceptual: 3.086 MSE: 965.269 PSNR: 18.284 SSIM: 0.554	Human: 0.667 Perceptual: 2.910 MSE: 319.761 PSNR: 23.083 SSIM: 0.655	Human: 0.333 Perceptual: 3.258 MSE: 877.453 PSNR: 18.699 SSIM: 0.604	Human: 0.333 Perceptual: 3.140 MSE: 510.856 PSNR: 21.048 SSIM: 0.583
Human: 0.200	Human: 0.800	Human: 0.200	Human: 0.333	Human: 0.867	Human: 0.867	Human: 0.067	Human: 0.333	Human: 0.200	Human: 0.867
Perceptual: 2.677	Perceptual: 3.687	Perceptual: 2.410	Perceptual: 2.357	Perceptual: 3.935	Perceptual: 3.058	Perceptual: 3.043	Perceptual: 3.467	Perceptual: 2.005	Perceptual: 3.297
MSE: 559.432	MSE: 1841.096	MSE: 1127.463	MSE: 696.860	MSE: 674.975	MSE: 1290.275	MSE: 1101.580	MSE: 511.964	MSE: 461.770	MSE: 303.960
PSNR: 20.576	PSNR: 15.480	PSNR: 17.610	PSNR: 19.699	PSNR: 19.838	PSNR: 17.024	PSNR: 17.711	PSNR: 21.038	PSNR: 21.487	PSNR: 23.303
SSIM: 0.607	SSIM: 0.509	SSIM: 0.568	SSIM: 0.579	SSIM: 0.597	SSIM: 0.629	SSIM: 0.567	SSIM: 0.533	SSIM: 0.587	SSIM: 0.625
Human: 0.667	Human: 0.267	Human: 0.867	Human: 0.800	Human: 0.400	Human: 0.000	Human: 0.867	Human: 0.733	Human: 0.800	Human: 0.467
Perceptual: 3.546	Perceptual: 3.634	Perceptual: 3.300	Perceptual: 3.764	Perceptual: 2.378	Perceptual: 4.453	Perceptual: 4.151	Perceptual: 3.509	Perceptual: 3.696	Perceptual: 2.814
MSE: 1134.085	MSE: 2943.497	MSE: 697.217	MSE: 1073.511	MSE: 951.902	MSE: 2680.813	MSE: 1099.790	MSE: 3384.819	MSE: 1308.356	MSE: 1042.061
PSNR: 17.584	PSNR: 13.442	PSNR: 19.697	PSNR: 17.823	PSNR: 18.345	PSNR: 13.848	PSNR: 17.718	PSNR: 12.835	PSNR: 16.964	PSNR: 17.952
SSIM: 0.593	SSIM: 0.459	SSIM: 0.654	SSIM: 0.534	SSIM: 0.619	SSIM: 0.552	SSIM: 0.570	SSIM: 0.567	SSIM: 0.558	SSIM: 0.607
Human: 0.000	Human: 0.267	Human: 0.533	Human: 0.667	Human: 0.800	Human: 0.400	Human: 0.933	Human: 0.867	Human: 0.867	Human: 0.600
Perceptual: 3.220	Perceptual: 3.271	Perceptual: 2.414	Perceptual: 3.959	Perceptual: 4.871	Perceptual: 2.363	Perceptuai: 2.804	Perceptual: 2.863	Perceptual: 2.477	Perceptual: 3.344
MSE: 1555:241	MSE: 903.598	MSE: 256.194	MSE: 976.151	MSE: 2463,398	MSE: 218.487	MSE: 195.021	MSE: 977.392	MSE: 359.090	MSE: 1819.131
PSNR: 16.213	PSNR: 18.571	PSNR: 24.045	PSNR: 18.236	PSNR: 14.215	PSNR: 24.737	PSNR: 25.230	PSNR: 18.230	PSNR: 22.579	PSNR: 15.532
SSIM: 0.528	SSIM: 0.591	SSIM: 0.657	SSIM: 0.637	SSIM: 0.548	SSIM: 0.573	SSIM: 0.665	SSIM: 0.638	SSIM: 0.660	SSIM: 0.587
Human: 0.200	Human: 0.200	Human: 0.533	Human: 0.667	Human: 0.000	Human: 0.533	Human: 0.200	Human: 0.600	Human: 0.800	Human: 0.733
Perceptual: 3.930	Perceptual: 2.821	Perceptual: 4.126	Perceptual: 3.139	Perceptual: 3.784	Perceptual: 4.312	Perceptual: 3.118	Perceptual: 3.762	Perceptual: 4.322	Perceptual: 4.020
MSE: 1446.275	MSE: 2217.483	MSE: 1008.765	MSE: 1466.006	MSE: 1802.259	MSE: 1913.836	MSE: 1165.424	MSE: 2100.191	MSE: 1659.373	M5E: 2012.086
PSNR: 16.528	PSNR: 14.672	PSNR: 18.093	PSNR: 16.469	PSNR: 115.573	PSNR: 15.312	PSNR: 17.466	PSNR: 14.908	PSNR: 15.931	PSNR: 15.094
SSIM: 0.647	SSIM: 0.495	SSIM: 0.567	SSIM: 0.509	SSIM: 0.525	SSIM: 0.544	SSIM: 0.582	SSIM: 0.622	SSIM: 0.573	SSIM: 0.569

Figure 48: CUB 32x32 results for Conditional ProGAN.



Figure 49: CUB 32x32 results for PixelCNN++.





Figure 50: Ground truth for CUB at 64x64 resolution.

the second	P		- demo	-		~	15-1	1	al start
Human: 0.333	Human: 0.533	Human: 0.067	Human: 0.000	Human: 0.267	Human: 0.200	Human: 0.200	Human: 0.333	Human: 0.133	Human: 0.000
Perceptual: 3.958	Perceptual: 2.654	Perceptual: 4.023	Perceptual: 4.006	Perceptual: 4.238	Perceptual: 4.553	Perceptual: 4.680	Perceptual: 3.389	Perceptual: 4.150	Perceptual: 4.194
MSE: 2193.261	MSE: 1108.099	MSE: 408.763	MSE: 783.493	MSE: 4776.855	MSE: 1462.088	MSE: 4291.429	MSE: 2279.982	MSE: 2575.894	MSE: 2107.474
PSNR: 14.720	PSNR: 17.685	PSNR: 22.016	PSNR: 19.190	PSNR: 11.339	PSNR: 16.481	PSNR: 11.805	PSNR: 14.551	PSNR: 14.022	PSNR: 14.893
SSIM: 0.511	SSIM: 0.677	SSIM: 0.651	SSIM: 0.563	SSIM: 0.542	SSIM: 0.513	SSIM: 0.551	SSIM: 0.531	SSIM: 0.560	SSIM: 0.529
1 and	New -				A				
Human: 0.200	Human: 0.133	Human: 0.067	Human: 0.400	Human: 0.333	Human: 0.133	Human: 0.000	Human: 0.067	Human: 0.000	Human: 0.067
Perceptual: 4.013	Perceptual: 3.767	Perceptual: 4.767	Perceptual: 3.468	Perceptual: 4.799	Perceptual: 4.623	Perceptual: 4.032	Perceptual: 4.340	Perceptual: 4.813	Perceptual: 3.359
MSE: 2188.096	MSE: 1312.883	MSE: 524.875	MSE: 2031.532	MSE: 3473.973	MSE: 4931.812	MSE: 2518.601	MSE: 1220.650	MSE: 2452.510	MSE: 965.292
PSNR: 14.730	PSNR: 16.949	PSNR: 20.930	PSNR: 15.053	PSNR: 12.723	PSNR: 11.201	PSNR: 14.119	PSNR: 17.265	PSNR: 14.235	PSNR: 18.284
SSIM: 0.506	SSIM: 0.520	SSIM: 0.565	SSIM: 0.525	SSIM: 0.614	SSIM: 0.600	SSIM: 0.568	SSIM: 0.599	SSIM: 0.573	SSIM: 0.551
X			15	K			4		Y
Human: 0.400	Human: 0.133	Human: 0.000	Human: 0.133	Human: 0.067	Human: 0.467	Human: 0.533	Human: 0.400	Human: 0.400	Human: 0.133
Perceptual: 3.442	Perceptual: 3.728	Perceptual: 4.332	Perceptual: 4.417	Perceptual: 4.437	Perceptual: 4.110	Perceptual: 3.496	Perceptual: 3.379	Perceptual: 4.525	Perceptual: 3.893
MSE: 2330.680	MSE: 1100.143	MSE: 1604.806	MSE: 4210.345	MSE: 6012.778	MSE: 6010.405	MSE: 1746.084	MSE: 2377,452	MSE: 2577.028	MSE: 662.601
PSNR: 14.456	PSNR: 17.716	PSNR: 16.077	PSNR: 11.888	PSNR: 10.340	PSNR: 10.342	PSNR: 15.710	PSNR: 14.370	PSNR: 14.020	PSNR: 19.918
SSIM: 0.602	SSIM: 0.551	SSIM: 0.510	SSIM: 0.499	SSIM: 0.532	SSIM: 0.591	SSIM: 0.590	SSIM: 0.682	SSIM: 0.581	SSIM: 0.712
3		+	N	and the second sec	ale.		ate.		4
Human: 0.133	Human: 0.200	Human: 0.467	Human: 0.267	Human: 0.000	Human: 0.000	Human: 0.000	Human: 0.667	Human: 0.000	Human: 0.133
Perceptual: 3.737	Perceptual: 3.668	Perceptual: 3.877	Perceptual: 3.861	Perceptual: 4.358	Perceptual: 4.706	Perceptual: 4.483	Perceptual: 4.489	Perceptual: 4.385	Perceptual: 3.664
MSE: 2256.436	MSE: 3201.350	MSE: 2523,776	MSE: 2733.164	MSE: 1179.434	MSE: 2043.086	MSE: 5739.060	MSE: 4944,792	MSE: 821.047	MSE: 2555.913
PSNR: 14.597	PSNR: 13.077	PSNR: 14.110	PSNR: 13.764	PSNR: 17.414	PSNR: 15.028	PSNR: 10.542	PSNR: 11.189	PSNR: 18.987	PSNR: 14.055
SSIM: 0.528	SSIM: 0.568	SSIM: 0.583	SSIM: 0.530	SSIM: 0.537	SSIM: 0.628	SSIM: 0.492	SSIM: 0.551	SSIM: 0.553	SSIM: 0.688
M		AL	- Com	<u> </u>	2.	-	7-	r	-F
Human: 0.200	Human: 0.533	Human: 0.067	Human: 0.133	Human: 0.133	Human: 0.333	Human: 0.067	Human: 0.200	Human: 0.133	Human: 0.067
Perceptual: 4.720	Perceptual: 4.027	Perceptual: 3.311	Perceptual: 3.761	Perceptual: 3.427	Perceptual: 4.009	Perceptual: 3.574	Perceptual: 3.778	Perceptual: 3.854	Perceptual: 4.128
MSE: 775.241	MSE: 2216.442	MSE: 898.257	MSE: 523.065	MSE: 1611.103	MSE: 1839.010	MSE: 1210.770	MSE: 752.559	MSE: 2049.001	MSE: 1427.536
PSNR: 19.236	PSNR: 14.674	PSNR: 18.597	PSNR: 20.945	PSNR: 16.060	PSNR: 15.485	PSNR: 17.300	PSNR: 19.365	PSNR: 15.015	PSNR: 16.585
SSIM: 0.582	SSIM: 0.561	SSIM: 0.565	SSIM: 0.591	SSIM: 0.576	SSIM: 0.537	SSIM: 0.556	SSIM: 0.718	SSIM: 0.555	SSIM: 0.544
	E.		An		*		P	5	
Human: 0.200	Human: 0.600	Human: 0.067	Human: 0.000	Human: 0.133	Human: 0.000	Human: 0.267	Human: 0.133	Human: 0.067	Human: 0.067
Perceptual: 4.585	Perceptual: 3.585	Perceptual: 4.089	Perceptual: 4.975	Perceptual: 4.125	Perceptual: 4.086	Perceptual: 4.906	Perceptual: 4.123	Perceptual: 3.760	Perceptual: 4.558
MSE: 2241.812	MSE: 651.985	MSE: 773.066	MSE: 1634.921	MSE: 737.485	MSE: 657.405	MSE: 1938.907	MSE: 963.782	MSE: 1763.389	MSE: 1045.049
PSNR: 14.625	PSNR: 19.988	PSNR: 19.249	PSNR: 15.996	PSNR: 19.453	PSNR: 19.952	PSNR: 15.255	PSNR: 18.291	PSNR: 15.667	PSNR: 17.939
SSIM: 0.530	SSIM: 0.592	SSIM: 0.569	SSIM: 0.550	SSIM: 0.578	SSIM: 0.572	SSIM: 0.512	SSIM: 0.576	SSIM: 0.537	SSIM: 0.541
	S-			No.	1				
Human: 0.067	Human: 0.133	Human: 0.133	Human: 0.267	Human: 0.400	Human: 0.200	Human: 0.000	Human: 0.133	Human: 0.000	Human: 0.133
Perceptual: 3.984	Perceptual: 4.157	Perceptual: 4.246	Perceptual: 3.474	Perceptual: 3.518	Perceptual: 3.721	Perceptual: 3.900	Perceptual: 4.308	Perceptual: 4.033	Perceptual: 4.189
MSE: 971.256	MSE: 2310.207	MSE: 1246.688	MSE: 1218.680	MSE: 1185.674	MSE: 1992.886	MSE: 1981.406	MSE: 2137.631	MSE: 646.898	MSE: 1070.431
PSNR: 18.257	PSNR: 14.494	PSNR: 17.173	PSNR: 17.272	PSNR: 17.391	PSNR: 15.136	PSNR: 15.161	PSNR: 14.831	PSNR: 20.022	PSNR: 17.835
SSIM: 0.553	SSIM: 0.567	SSIM: 0.566	SSIM: 0.546	SSIM: 0.587	SSIM: 0.634	SSIM: 0.517	SSIM: 0.543	SSIM: 0.553	SSIM: 0.551
		-	É	2				A	-
Human: 0.133	Human: 0.067	Human: 0.200	Human: 0.067	Human: 0.467	Human: 0.133	Human: 0.400	Human: 0.133	Human: 0.333	Human: 0.000
Perceptual: 4.355	Perceptual: 4.298	Perceptual: 4.064	Perceptual: 3.574	Perceptual: 2.746	Perceptual: 4.648	Perceptual: 3.849	Perceptual: 4.085	Perceptual: 3.727	Perceptual: 4.621
MSE: 2550.201	MSE: 3203.249	MSE: 2118.915	MSE: 2147.205	MSE: 1284.603	MSE: 1477.494	MSE: 3647.860	MSE: 4511.529	MSE: 3009.686	MSE: 3119.876
PSNR: 14.065	PSNR: 13.075	PSNR: 14.870	PSNR: 14.812	PSNR: 17.043	PSNR: 16.436	PSNR: 12.510	PSNR: 11.588	PSNR: 13.346	PSNR: 13.189
SSIM: 0.480	SSIM: 0.494	SSIM: 0.549	SSIM: 0.555	SSIM: 0.641	SSIM: 0.515	SSIM: 0.536	SSIM: 0.551	SSIM: 0.564	SSIM: 0.512
		A.	\$		2	1)	1	
Human: 0.000	Human: 0.067	Human: 0.267	Human: 0.133	Human: 0.067	Human: 0.133	Human: 0.667	Human: 0.267	Human: 0.333	Human: 0.467
Perceptual: 4.385	Perceptual: 4.323	Perceptual: 4.842	Perceptual: 3.350	Perceptual: 4.420	Perceptual: 4.779	Perceptual: 4.135	Perceptual: 4.282	Perceptual: 3.909	Perceptual: 3.513
MSE: 1738.191	MSE: 1276.339	MSE: 1611.334	MSE: 2178.938	MSE: 3421.896	MSE: 1222.715	MSE: 522.253	MSE: 2317.713	MSE: 1221.416	MSE: 2367.704
PSNR: 15.730	PSNR: 17.071	PSNR: 16.059	PSNR: 14.748	PSNR: 12.788	PSNR: 17.258	PSNR: 20.952	PSNR: 14.480	PSNR: 17.262	PSNR: 14.388
SSIM: 0.571	SSIM: 0.577	SSIM: 0.575	SSIM: 0.558	SSIM: 0.503	SSIM: 0.572	SSIM: 0.646	SSIM: 0.548	SSIM: 0.577	SSIM: 0.581
te		- 8			1	12			-
Human: 0.000	Human: 0.067	Human: 0.267	Human: 0.200	Human: 0.200	Human: 0.000	Human: 0.000	Human: 0.200	Human: 0.467	Human: 0.267
Perceptual: 5.066	Perceptual: 4.472	Perceptual: 4.211	Perceptual: 4.107	Perceptual: 4.128	Perceptual: 3.887	Perceptual: 4.913	Perceptual: 4.442	Perceptual: 4.309	Perceptual: 4.263
MSE: 2531.948	MSE: 3405.464	MSE: 1537.860	MSE: 2259.324	MSE: 3593.818	MSE: 2440.742	MSE: 2625.052	MSE: 3666.098	MSE: 2676.839	MSE: 1983.832
PSNR: 14.096	PSNR: 12.809	PSNR: 16.262	PSNR: 14.591	PSNR: 12.575	PSNR: 14.256	PSNR: 13.939	PSNR: 12.489	PSNR: 13.855	PSNR: 15.156
SSIM: 0.515	SSIM: 0.543	SSIM: 0.560	SSIM: 0.552	SSIM: 0.537	SSIM: 0.556	SSIM: 0.494	SSIM: 0.556	SSIM: 0.556	SSIM: 0.562

Figure 51: CUB 64x64 results for ProGAN.



Figure 52: CUB 64x64 results for Conditional StyleGAN.



Figure 53: CUB 64x64 results for Conditional ProGAN.

man: 0.200

Perceptual: 3. MSE: 677.000 PSNR: 19.825 SSIM: 0.650

MSE: 3 PSNR: SSIM: 3504.222 12.685 0.562

Q

Perceptual: 2.781 MSE: 3633.937 PSNR: 12.527 SSIM: 0.701

Perceptual: 4.081 MSE: 4915.741 PSNR: 11.215 SSIM: 0.540

Human: 0.400 Perceptual: 3.476 MSE: 1832.323 PSNR: 15.501 SSIM: 0.558

Human: 0.400 Perceptual: 2. MSE: 566.229 PSNR: 20.601 SSIM: 0.585

MSE: 3012.645 PSNR: 13.341 SSIM: 0.640



Perceptual: 4.161 MSE: 4639.706 PSNR: 11.466 SSIM: 0.559

Perceptual: 3.6 MSE: 2918.179 PSNR: 13.480 SSIM: 0.628







Human: 0.533 Perceptual: 3.429 MSE: 2535.653 PSNR: 14.090 SSIM: 0.609

Figure 54: CUB 64x64 results for PixelCNN++.

60

Perceptual: 3.792 MSE: 3570.136 PSNR: 12.604 SSIM: 0.520

Human: 0.467 Perceptual: 3.791 MSE: 1831.564 PSNR: 15.503 SSIM: 0.555



Human: 0.400 Perceptual: 2.668 MSE: 2219.021 PSNR: 14.669 SSIM: 0.582

Human: 0.667 Perceptual: 2.9 MSE: 1166.638 PSNR: 17.461 SSIM: 0.628

Human: 0.467 Perceptual: 3.1 MSE: 1110.495 PSNR: 17.676 SSIM: 0.569











Human: 0.533 Perceptual: 2.978 MSE: 792.531 PSNR: 19.141



































Human: 0.533 Perceptual: 3.451 MSE: 1632.512 PSNR: 16.002 SSIM: 0.557

Human: 0.533 Perceptual: 3.255 MSE: 1042.928 PSNR: 17.948 SSIM: 0.518









A.3.3 128x128



Figure 55: Ground truth for CUB at 128x128 resolution.



Figure 56: CUB 128x128 results for ProGAN.



Figure 57: CUB 128x128 results for Conditional StyleGAN.



Figure 58: CUB 128x128 results for Conditional ProGAN.

A.4 LSUN-BEDROOM

A.4.1 32x32



Figure 59: Ground truth for LSUN-Bedroom at 32x32 resolution.



Figure 60: LSUN-Bedroom 32x32 results for ProGAN.



Figure 61: LSUN-Bedroom 32x32 results for Conditional StyleGAN.



Figure 62: LSUN-Bedroom 32x32 results for Conditional ProGAN.



Figure 63: LSUN-Bedroom 32x32 results for PixelCNN++.





Figure 64: Ground truth for LSUN-Bedroom at 64x64 resolution.



Figure 65: LSUN-Bedroom 64x64 results for ProGAN.



Figure 66: LSUN-Bedroom 64x64 results for Conditional StyleGAN.


Figure 67: LSUN-Bedroom 64x64 results for Conditional ProGAN.



Figure 68: LSUN-Bedroom 64x64 results for PixelCNN++.

```
A.4.3 128x128
```



Figure 69: Ground truth for LSUN-Bedroom at 128x128 resolution.



Figure 70: LSUN-Bedroom 128x128 results for ProGAN.



Figure 71: LSUN-Bedroom 128x128 results for Conditional StyleGAN.



Figure 72: LSUN-Bedroom 128x128 results for Conditional ProGAN.

A.5 LSUN-CAT

A.5.1 32x32



Figure 73: Ground truth for LSUN-Cat at 32x32 resolution.



Figure 74: LSUN-Cat 32x32 results for ProGAN.



Figure 75: LSUN-Cat 32x32 results for Conditional StyleGAN.



Figure 76: LSUN-Cat 32x32 results for Conditional ProGAN.

man: 0.667



Human: 0.733 Perceptual: 3.564 MSE: 3367.831 PSNR: 12.857 SSIM: 0.540



Human: 0.333 Perceptual: 2.845 MSE: 1041.152 PSNR: 17.956 SSIM: 0.762

Perceptual: 3.684 MSE: 3099.789 PSNR: 13.217





Human: 0.200 Perceptual: 2.764 MSE: 2557.924 PSNR: 14.052



Human: 0.467 Perceptual: 2.028 MSE: 1337.378 PSNR: 16.868 SSIM: 0.751



Human: 0.933 Perceptual: 1.854 MSE: 766.889 PSNR: 19.283 SSIM: 0.794





Perceptual: 2.882 MSE: 1467.240 PSNR: 16.466 SSIM: 0.730

Per MSI PSN

Perceptual: 2.546 MSE: 3188.136 PSNR: 13.095 SSIM: 0.591

Perceptual: 2.8 MSE: 2222.461 PSNR: 14.662 SSIM: 0.623

48

Perceptual: 2.727 MSE: 1741.525 PSNR: 15.722 SSIM: 0.616





Perceptual: 3.533 MSE: 4490.450 PSNR: 11.608 SSIM: 0.576

Human: 0.333 Perceptual: 2.735 MSE: 1408.275 PSNR: 16.644 SSIM: 0.685

Numan: 0.467 Perceptual: 4.044 MSE: 2605.378 PSNR: 13.972 SSIM: 0.559





Figure 77: LSUN-Cat 32x32 results for PixelCNN++.

```
A.5.2 64x64
```



Figure 78: Ground truth for LSUN-Cat at 64x64 resolution.



Figure 79: LSUN-Cat 64x64 results for ProGAN.



Figure 80: LSUN-Cat 64x64 results for Conditional StyleGAN.



Figure 81: LSUN-Cat 64x64 results for Conditional ProGAN.

Perce MSE: PSNR SSIM:

Perc MSE PSN

Perceptual: 3.381 MSE: 5817.350 PSNR: 10.484 SSIM: 0.583

Perceptual: 3.3 MSE: 1437.930 PSNR: 16.553 SSIM: 0.623

luman: 0.867



Perceptual: 2.4 MSE: 1325.557 PSNR: 16.907 SSIM: 0.546











Human: 0.600 Perceptual: 3.793 MSE: 2678.404 PSNR: 13.852 SSIM: 0.600



Human: 0.000 Perceptual: 3.731 MSE: 3191.954 PSNR: 13.090



Human: 0.267 Perceptual: 3.986 MSE: 8980.595 PSNR: 8.598 SSIM: 0.555



Human: 0.800 Perceptual: 2.7 MSE: 1274.897 PSNR: 17.076 SSIM: 0.673





numan: 0.400 Perceptual: 3.582 MSE: 4428.549 PSNR: 11.668 SSIM: 0.580 Perceptual: 4.1 MSE: 2190.245 PSNR: 14.726 SSIM: 0.585 Perceptual: 3.6 MSE: 4721.623 PSNR: 11.390 SSIM: 0.645





Perceptual: 3.260 MSE: 1465.315 PSNR: 16.471 SSIM: 0.578

luman: 0.733

Perceptual: 3.331 MSE: 4494.124 PSNR: 11.604 SSIM: 0.560











Figure 82: LSUN-Cat 64x64 results for PixelCNN++.

```
A.5.3 128x128
```



Figure 83: Ground truth for LSUN-Cat at 128x128 resolution.

Per MSI PSN





Human: 0.000 Perceptual: 4.263 MSE: 2366.316 PSNR: 14.390 SSIM: 0.528









Perceptual: 4.497 MSE: 2851.889 PSNR: 13.579 SSIM: 0.545

90

Figure 84: LSUN-Cat 128x128 results for ProGAN.



Figure 85: LSUN-Cat 128x128 results for Conditional StyleGAN.



Figure 86: LSUN-Cat 128x128 results for Conditional ProGAN.