

A U-Net Based Discriminator for Generative Adversarial Networks

Edgar Schönfeld

Bosch Center for Artificial Intelligence
edgar.schoenfeld@bosch.com

Bernt Schiele

Max Planck Institute for Informatics
schiele@mpi-inf.mpg.com

Anna Khoreva

Bosch Center for Artificial Intelligence
anna.khoreva@bosch.com

Abstract

Among the major remaining challenges for generative adversarial networks (GANs) is the capacity to synthesize globally and locally coherent images with object shapes and textures indistinguishable from real images. To target this issue we propose an alternative U-Net based discriminator architecture, borrowing the insights from the segmentation literature. The proposed U-Net based architecture allows to provide detailed per-pixel feedback to the generator while maintaining the global coherence of synthesized images, by providing the global image feedback as well. Empowered by the per-pixel response of the discriminator, we further propose a per-pixel consistency regularization technique based on the CutMix data augmentation, encouraging the U-Net discriminator to focus more on semantic and structural changes between real and fake images. This improves the U-Net discriminator training, further enhancing the quality of generated samples. The novel discriminator improves over the state of the art in terms of the standard distribution and image quality metrics, enabling the generator to synthesize images with varying structure, appearance and levels of detail, maintaining global and local realism. Compared to the BigGAN baseline, we achieve an average improvement of 2.7 FID points across FFHQ, CelebA, and the newly introduced COCO-Animals dataset. The code is available at <https://github.com/boschresearch/unetgan>.

1. Introduction

The quality of synthetic images produced by generative adversarial networks (GANs) has seen tremendous improvement recently [5, 20]. The progress is attributed to large-scale training [32, 5], architectural modifications [50, 19, 20, 27], and improved training stability via the use of different regularization techniques [34, 51]. However, despite the recent advances, learning to synthesize images with global semantic coherence, long-range structure and the exactness of detail remains challenging.

One source of the problem lies potentially in the discrim-

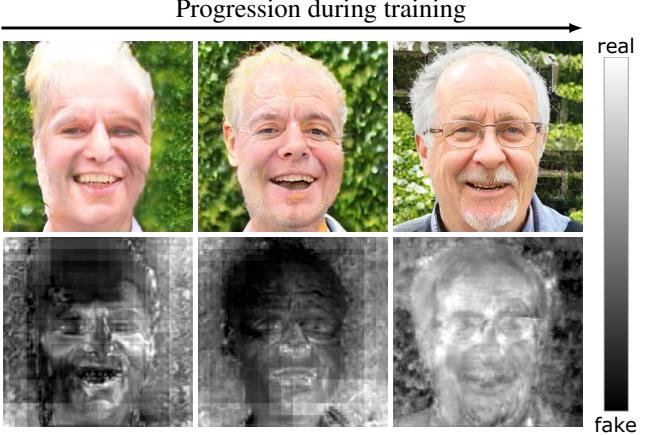


Figure 1: Images produced throughout the training by our U-Net GAN model (top row) and their corresponding per-pixel feedback of the *U-Net discriminator* (bottom row). The synthetic image samples are obtained from a fixed noise vector at different training iterations. Brighter colors correspond to the discriminator confidence of pixel being real (and darker of being fake). Note that the U-Net discriminator provides very detailed and spatially coherent response to the generator, enabling it to further improve the image quality, e.g. the unnaturally large man’s forehead is recognized as fake by the discriminator and is corrected by the generator throughout the training.

inator network. The discriminator aims to model the data distribution, acting as a loss function to provide the generator a learning signal to synthesize realistic image samples. The stronger the discriminator is, the better the generator has to become. In the current state-of-the-art GAN models, the discriminator being a classification network learns only a representation that allows to efficiently penalize the generator based on the most discriminative difference between real and synthetic images. Thus, it often focuses either on the global structure or local details. The problem amplifies as the discriminator has to learn in a non-stationary envi-

ronment: the distribution of synthetic samples shifts as the generator constantly changes through training, and is prone to forgetting previous tasks [7] (in the context of the discriminator training, learning semantics, structures, and textures can be considered different tasks). This discriminator is not incentivized to maintain a more powerful data representation, learning both global and local image differences. This often results in the generated images with discontinued and mottled local structures [27] or images with incoherent geometric and structural patterns (e.g. asymmetric faces or animals with missing legs) [50].

To mitigate this problem, we propose an alternative discriminator architecture, which outputs simultaneously both global (over the whole image) and local (per-pixel) decision of the image belonging to either the real or fake class, see Figure 1. Motivated by the ideas from the segmentation literature, we re-design the discriminator to take a role of both a classifier and segmenter. We change the architecture of the discriminator network to a U-Net [39], where the encoder module performs per-image classification, as in the standard GAN setting, and the decoder module outputs per-pixel class decision, providing spatially coherent feedback to the generator, see Figure 2. This architectural change leads to a stronger discriminator, which is encouraged to maintain a more powerful data representation, making the generator task of fooling the discriminator more challenging and thus improving the quality of generated samples (as also reflected in the generator and discriminator loss behavior in Figure 8). Note that we do not modify the generator in any way, and our work is orthogonal to the ongoing research on architectural changes of the generator [20, 27], divergence measures [25, 1, 37], and regularizations [40, 15, 34].

The proposed U-Net based discriminator allows to employ the recently introduced CutMix [47] augmentation, which is shown to be effective for classification networks, for consistency regularization in the two-dimensional output space of the decoder. Inspired by [47], we cut and mix the patches from real and synthetic images together, where the ground truth label maps are spatially combined with respect to the real and fake patch class for the segmenter (U-Net decoder) and the class labels are set to fake for the classifier (U-Net encoder), as globally the CutMix image should be recognized as fake, see Figure 3. Empowered by per-pixel feedback of the U-Net discriminator, we further employ these CutMix images for consistency regularization, penalizing per-pixel inconsistent predictions of the discriminator under the CutMix transformations. This fosters the discriminator to focus more on semantic and structural changes between real and fake images and to attend less to domain-preserving perturbations. Moreover, it also helps to improve the localization ability of the decoder. Employing the proposed consistency regularization leads to a stronger generator, which pays more attention to local and

global image realism. We call our model U-Net GAN.

We evaluate the proposed U-Net GAN model across several datasets using the state-of-the-art BigGAN model [5] as a baseline and observe an improved quality of the generated samples in terms of the FID and IS metrics. For unconditional image synthesis on FFHQ [20] at resolution 256×256 , our U-Net GAN model improves 4 FID points over the BigGAN model, synthesizing high quality human faces (see Figure 4). On CelebA [29] at resolution 128×128 we achieve 1.6 point FID gain, yielding to the best of our knowledge the lowest known FID score of 2.95. For class-conditional image synthesis on the introduced COCO-Animals dataset [28, 24] at resolution 128×128 we observe an improvement in FID from 16.37 to 13.73, synthesizing diverse images of different animal classes (see Figure 5).

2. Related work

Generative adversarial networks. GAN [14] and its conditional variant [33] have recently demonstrated impressive results on different computer vision tasks, including image synthesis [38, 50, 19, 5, 20, 27, 10]. Plenty of efforts have been made to improve the training and performance of GANs, from reformulation of the objective function [31, 1, 26, 37], integration of different regularization techniques [51, 34, 40, 48] and architectural changes [38, 19, 13, 27]. To enhance the quality of generated samples, [38] introduced the DCGAN architecture that employs strided and transposed convolutions. In SAGAN [50] the self-attention block was added to improve the network ability to model global structure. PG-GAN [19] proposed to grow both the generator and discriminator networks to increase the resolution of generated images. Other lines of work focused mainly on improving the discriminator by exploiting multiple [36, 13, 11] and multi-resolution [45, 42] discriminators, using spatial feedback of the discriminator [17], an auto-encoder architecture with the reconstruction-based feedback to the generator [52] or self-supervision to avoid catastrophic forgetting [7]. Most recently, the attention has been switched back to the generator network. StyleGAN [20] proposed to alter the generator architecture by injecting latent codes to each convolution layer, thus allowing more control over the image synthesis process. COCO-GAN [27] integrated the conditional coordination mechanism into the generator, making image synthesis highly parallelizable. In this paper, we propose to alter the discriminator network to a U-Net based architecture, empowering the discriminator to capture better both global and local structures, enabled by per-pixel discriminator feedback. Local discriminator feedback is also commonly applied through PatchGAN discriminators [18]. Our U-Net GAN extends this idea to dense prediction over the whole image plane, with visual information being integrated over up- and down-sampling pathways and through the encoder-decoder skip

connections, without trading off local over global realism.

Mix&Cut regularizations. Recently, a few simple yet effective regularization techniques have been proposed, which are based on augmenting the training data by creating synthetic images via mixing or/and cutting samples from different classes. In MixUp [49] the input images and their target labels are interpolated using the same randomly chosen factor. [43] extends [49] by performing interpolation not only in the input layer but also in the intermediate layers. CutOut [9] augments an image by masking a rectangular region to zero. Differently, CutMix [47] augments training data by creating synthetic images via cutting and pasting patches from image samples of different classes, marrying the best aspects of MixUp and CutOut. Other works employ the Mix&Cut approaches for consistency regularization [44, 4, 51], i.e. penalizing the classification network sensitivity to samples generated via MixUp or CutOut [49, 9]. In our work, we propose the consistency regularization under the CutMix transformation in the pixel output space of our U-Net discriminator. This helps to improve its localization quality and induce it to attend to non-discriminative differences between real and fake regions.

3. U-Net GAN Model

A “vanilla” GAN consists of two networks: a generator G and a discriminator D , trained by minimizing the following competing objectives in an alternating manner:

$$\begin{aligned}\mathcal{L}_D &= -\mathbb{E}_x[\log D(x)] - \mathbb{E}_z[\log(1 - D(G(z)))] , \\ \mathcal{L}_G &= -\mathbb{E}_z[\log D(G(z))]^1.\end{aligned}\quad (1)$$

G aims to map a latent variable $z \sim p(z)$ sampled from a prior distribution to a realistic-looking image, while D aims to distinguish between real x and generated $G(z)$ images. Ordinarily, G and D are modeled as a decoder and an encoder convolutional network, respectively.

While there are many variations of the GAN objective function and its network architectures [23, 30], in this paper we focus on improving the discriminator network. In Section 3.1, we propose to alter the D architecture from a standard classification network to an encoder-decoder network – U-Net [39], leaving the underlying basic architecture of D – the encoder part – untouched. The proposed discriminator allows to maintain both global and local data representation, providing more informative feedback to the generator. Empowered by local per-pixel feedback of the U-Net decoder module, in Section 3.2 we further propose a consistency regularization technique, penalizing per-pixel inconsistent predictions of the discriminator under the CutMix transformations [47] of real and fake images. This helps to improve

¹This formulation is originally proposed as non-saturating (NS) GAN in [14].

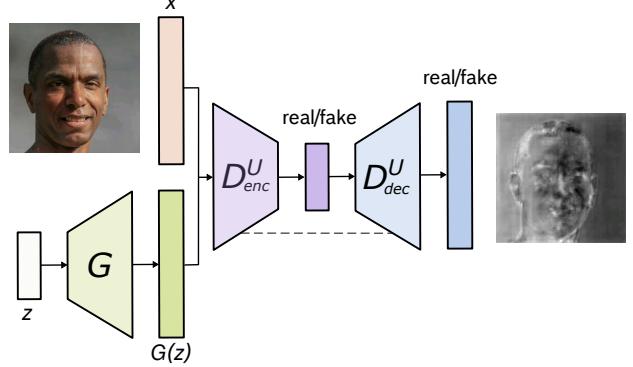


Figure 2: U-Net GAN. The proposed U-Net discriminator classifies the input images on a global and local *per-pixel* level. Due to the skip-connections between the encoder and the decoder (dashed line), the channels in the output layer contain both high- and low-level information. Brighter colors in the decoder output correspond to the discriminator confidence of pixel being real (and darker of being fake).

the localization quality of the U-Net discriminator and induce it to attend more to semantic and structural changes between real and fake samples. We call our model *U-Net GAN*. Note that our method is compatible with most GAN models as it does not modify the generator in any way and leaves the original GAN objective intact.

3.1. U-Net Based Discriminator

Encoder-decoder networks [2, 39] constitute a powerful method for dense prediction. U-Nets [39] in particular have demonstrated state-of-art performance in many complex image segmentation tasks. In these methods, similarly to image classification networks, the encoder progressively downsamples the input, capturing the global image context. The decoder performs progressive upsampling, matching the output resolution to the input one and thus enabling precise localization. Skip connections route data between the matching resolutions of the two modules, improving further the ability of the network to accurately segment fine details.

Analogously, in this work, we propose to extend a discriminator to form a U-Net, by reusing building blocks of the original discriminator classification network as an encoder part and building blocks of the generator network as the decoder part. In other words, the discriminator now consists of the original downsampling network and a new upsampling network. The two modules are connected via a bottleneck, as well as skip-connections that copy and concatenate feature maps from the encoder and the decoder modules, following [39]. We will refer to this discriminator as D^U . While the original $D(x)$ classifies the input image x into being real and fake, the U-Net discriminator $D^U(x)$ additionally performs this classification on a *per-pixel* basis, segmenting image x into real and fake regions, along with the original image classification of x from the encoder,

see Figure 2. This enables the discriminator to learn both global and local differences between real and fake images.

Hereafter, we refer to the original encoder module of the discriminator as D_{enc}^U and to the introduced decoder module as D_{dec}^U . The new discriminator loss is now can be computed by taking the decisions from both D_{enc}^U and D_{dec}^U :

$$\mathcal{L}_{DU} = \mathcal{L}_{D_{enc}^U} + \mathcal{L}_{D_{dec}^U}, \quad (2)$$

where similarly to Eq. 1 the loss for the encoder $L_{D_{enc}^U}$ is computed from the scalar output of D_{enc}^U :

$$\mathcal{L}_{D_{enc}^U} = -\mathbb{E}_x[\log D_{enc}^U(x)] - \mathbb{E}_z[\log(1 - D_{enc}^U(G(z)))] , \quad (3)$$

and the loss for the decoder $L_{D_{dec}^U}$ is computed as the mean decision over all pixels:

$$\begin{aligned} \mathcal{L}_{D_{dec}^U} &= -\mathbb{E}_x \left[\sum_{i,j} \log[D_{dec}^U(x)]_{i,j} \right] \\ &\quad - \mathbb{E}_z \left[\sum_{i,j} \log(1 - [D_{dec}^U(G(z))]_{i,j}) \right]. \end{aligned} \quad (4)$$

Here, $[D_{dec}^U(x)]_{i,j}$ and $[D_{dec}^U(G(z))]_{i,j}$ refer to the discriminator decision at pixel (i, j) . These per-pixel outputs of D_{dec}^U are derived based on global information from high-level features, enabled through the process of upsampling from the bottleneck, as well as more local information from low-level features, mediated by the skip connections from the intermediate layers of the encoder network.

Correspondingly, the generator objective becomes:

$$\begin{aligned} \mathcal{L}_G &= -\mathbb{E}_z \left[\log D_{enc}^U(G(z)) \right. \\ &\quad \left. + \sum_{i,j} \log[D_{dec}^U(G(z))]_{i,j} \right], \end{aligned} \quad (5)$$

encouraging the generator to focus on both global structures and local details while synthesizing images in order to fool the more powerful discriminator D^U .

3.2. Consistency Regularization

Here we present the consistency regularization technique for the U-Net based discriminator introduced in the previous section. The per-pixel decision of the well-trained D^U discriminator should be equivariant under any class-domain-altering transformations of images. However, this property is not explicitly guaranteed. To enable it, the discriminator should be regularized to focus more on semantic and structural changes between real and fake samples and to pay less attention to arbitrary class-domain-preserving perturbations. Therefore, we propose the consistency regularization of the D^U discriminator, explicitly encouraging the decoder module D_{dec}^U to output equivariant predictions under the CutMix transformations [47] of real and fake samples. The

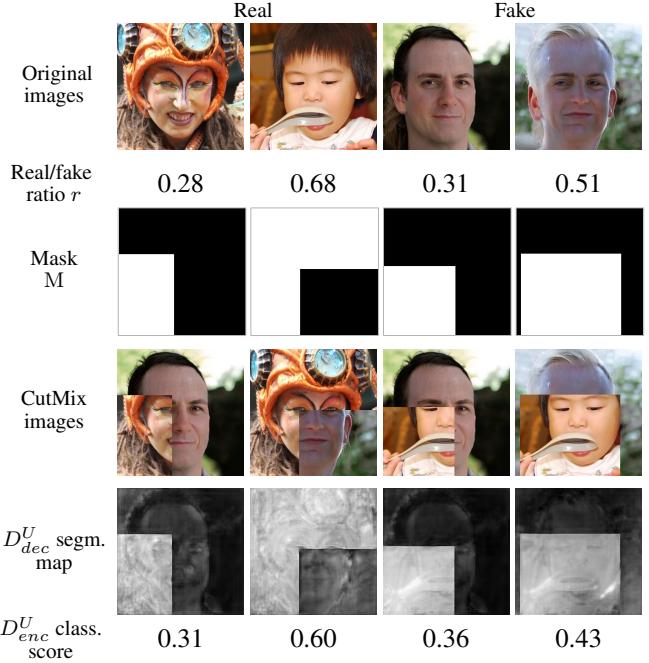


Figure 3: Visualization of the CutMix augmentation and the predictions of the U-Net discriminator on CutMix images. 1st row: real and fake samples. 2nd&3rd rows: sampled real/fake CutMix ratio r and corresponding binary masks M (color code: white for real, black for fake). 4th row: generated CutMix images from real and fake samples. 5th&6th row: the corresponding real/fake segmentation maps of D^U with its predicted classification scores.

CutMix augmentation creates synthetic images via cutting and pasting patches from images of different classes. We choose CutMix among other Mix&Cut strategies (cf. Section 2) as it does not alter the real and fake image patches used for mixing, in contrast to [49], preserving their original class domain, and provides a large variety of possible outputs. We visualize the CutMix augmentation strategy and the D^U predictions in Figure 3.

Following [47], we synthesize a new training sample \tilde{x} for the discriminator D^U by mixing x and $G(z) \in \mathbb{R}^{W \times H \times C}$ with the mask M :

$$\begin{aligned} \tilde{x} &= \text{mix}(x, G(z), M), \\ \text{mix}(x, G(z), M) &= M \odot x + (1 - M) \odot G(z), \end{aligned} \quad (6)$$

where $M \in \{0, 1\}^{W \times H}$ is the binary mask indicating if the pixel (i, j) comes from the real ($M_{i,j} = 1$) or fake ($M_{i,j} = 0$) image, 1 is a binary mask filled with ones, and \odot is an element-wise multiplication. In contrast to [47], the class label $c \in \{0, 1\}$ for the new CutMix image \tilde{x} is set to be fake, i.e. $c = 0$. Globally the mixed synthetic image should be recognized as fake by the encoder D_{enc}^U , otherwise the generator can learn to introduce the CutMix augmentation into generated samples, causing undesirable artifacts. Note that for the synthetic sample \tilde{x} , $c = 0$ and M

are the ground truth for the encoder and decoder modules of the discriminator D^U , respectively.

Given the CutMix operation in Eq. 6, we train the discriminator to provide consistent per-pixel predictions, i.e. $D_{dec}^U(\text{mix}(x, G(z), M)) \approx \text{mix}(D_{dec}^U(x), D_{dec}^U(G(z)), M)$, by introducing the consistency regularization loss term in the discriminator objective:

$$\mathcal{L}_{D_{dec}^U}^{cons} = \left\| D_{dec}^U \left(\text{mix}(x, G(z), M) \right) - \text{mix} \left(D_{dec}^U(x), D_{dec}^U(G(z)), M \right) \right\|^2, \quad (7)$$

where denotes $\|\cdot\|$ the L^2 norm. This consistency loss is then taken between the per-pixel output of D_{dec}^U on the CutMix image and the CutMix between outputs of the D_{dec}^U on real and fake images, penalizing the discriminator for inconsistent predictions.

We add the loss term in Eq. 7 to the discriminator objective in Eq. 2 with a weighting hyper-parameter λ :

$$\mathcal{L}_{D^U} = \mathcal{L}_{D_{enc}^U} + \mathcal{L}_{D_{dec}^U} + \lambda \mathcal{L}_{D_{dec}^U}^{cons} \dots \quad (8)$$

The generator objective \mathcal{L}_G remains unchanged, see Eq. 5.

In addition to the proposed consistency regularization, we also use CutMix samples for training both the encoder and decoder modules of D^U . Note that for the U-Net GAN we use the non-saturating GAN objective formulation [14]. However, the introduced consistency regularization as well as the U-Net architecture of the discriminator can be combined with any other adversarial losses of the generator and discriminator [1, 26, 37].

3.3. Implementation

Here we discuss implementation details of the U-Net GAN model proposed in Section 3.1 and 3.2.

U-Net based discriminator. We build upon the recent state-of-the-art BigGAN model [5], and extend its discriminator with our proposed changes. We adopt the BigGAN generator and discriminator architectures for the 256×256 (and 128×128) resolution with a channel multiplier $ch = 64$, as described in detail in [5]. The original BigGAN discriminator downsamples the input image to a feature map of dimensions $16ch \times 4 \times 4$, on which global sum pooling is applied to derive a $16ch$ dimensional feature vector that is classified into real or fake. In order to turn the discriminator into a U-Net, we copy the generator architecture and append it to the 4×4 output of the discriminator. In effect, the features are successively upsampled via ResNet blocks until the original image resolution ($H \times W$) is reached. To make the U-Net complete, the input to every decoder ResNet block is concatenated with the output features of the encoder blocks that share the same intermediate resolution. In this way, high-level and low-level information are

effectively integrated on the way to the output feature map. Hereby, the decoder architecture is almost identical to the generator, with the exception of that we change the number of channels of the final output from 3 to ch , append a final block of 1×1 convolutions to produce the $1 \times H \times W$ output map, and do not use class-conditional BatchNorm [8, 12] in the decoder, nor the encoder. Similarly to [5], we provide class information to D^U with projection [35] to the ch -dimensional channel features of the U-Net encoder and decoder output. In contrast to [5] and in alignment with [6], we find it beneficial not to use a hierarchical latent space, but to directly feed the same input vector z to BatchNorm at every layer in the generator. Lastly, we also remove the self-attention layer in both encoder and decoder, as in our experiments they did not contribute to the performance but led to memory overhead. While the original BigGAN is a class-conditional model, we additionally devise an unconditional version for our experiments. For the unconditional model, we replace class-conditional BatchNorm with self-modulation [6], where the BatchNorm parameters are conditioned only on the latent vector z , and do not use the class projection of [35] in the discriminator.

All these modifications leave us with a two-headed discriminator. While the decoder head is already sufficient to train the network, we find it beneficial to compute the GAN loss at both heads with equal weight. Analogously to BigGAN, we keep the hinge loss [50] in all basic U-Net models, while the models that also employ the consistency regularization in the decoder output space benefit from using the non-saturating loss [14]. Our implementation builds on top of the original BigGAN PyTorch implementation².

Consistency regularization. For each training iteration a mini-batch of CutMix images $(\tilde{x}, c = 0, M)$ is created with probability p_{mix} . This probability is increased linearly from 0 to 0.5 between the first n epochs in order to give the generator time to learn how to synthesize more real looking samples and not to give the discriminator too much power from the start. CutMix images are created from the existing real and fake images in the mini-batch using binary masks M . For sampling M , we use the original CutMix implementation³: first sampling the combination ratio r between the real and generated images from the uniform distribution $(0, 1)$ and then uniformly sample the bounding box coordinates for the cropping regions of x and $G(z)$ to preserve the r ratio, i.e. $r = \frac{|M|}{W \times H}$ (see Figure 3). Binary masks M also denote the target for the decoder D_{dec}^U , while we use *fake*, i.e $c = 0$, as the target for the encoder D_{enc}^U . We set $\lambda = 1.0$ as it showed empirically to be a good choice. Note that the consistency regularization does not impose much overhead during training. Extra computational cost comes only from

²<https://github.com/ajbrock/BigGAN-PyTorch>

³<https://github.com/clovaai/CutMix-PyTorch>

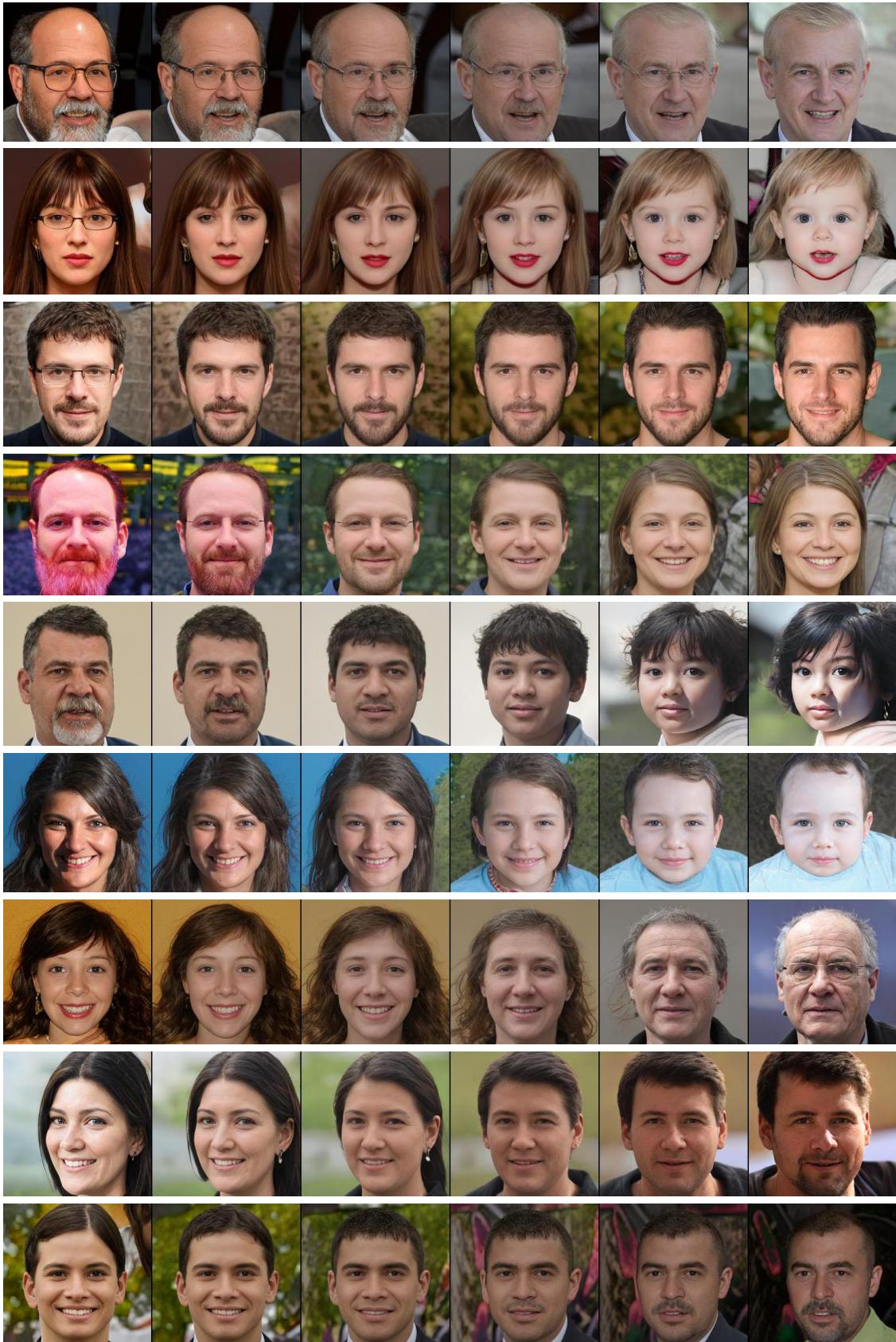


Figure 4: Images generated with U-Net GAN trained on FFHQ with resolution 256×256 when interpolating in the latent space between two synthetic samples (left to right). Note the high quality of synthetic samples and very smooth interpolations, maintaining *global* and *local* realism.



Figure 5: Images generated with U-Net GAN trained on COCO-Animals with resolution 128×128 .

feeding additional CutMix images through the discriminator while updating its parameters.

4. Experiments

4.1. Experimental Setup

Datasets. We consider three datasets: FFHQ [20], CelebA [29] and the subset of the COCO [28] and Open-Images [24] images containing animal classes, which we will further refer to as COCO-Animals. We use FFHQ and CelebA for unconditional image synthesis and COCO-Animals for class-conditional image synthesis, where the class label is used. We experiment with 256×256 resolution for FFHQ and 128×128 for CelebA and COCO-Animals.

CelebA is a human face dataset of 200k images, featuring ~ 10 k different celebrities with a variety of facial poses and expressions. Similarly, FFHQ is a more recent dataset of human faces, consisting of 70k high-quality images with higher variation in terms of age, ethnicity, accessories, and viewpoints. The proposed COCO-Animals dataset consists of ~ 38 k training images belonging to 10 animal classes, where we choose COCO and OpenImages (using the human verified subset with mask annotations) samples in the categories *bird*, *cat*, *dog*, *horse*, *cow*, *sheep*, *giraffe*, *zebra*, *elephant*, and *monkey*. With its relatively small size and imbalanced number of images per class as well as due to its variation in poses, shapes, number of objects, and backgrounds, COCO-Animals presents a challenging task for class-conditional image synthesis. We choose to create this dataset in order to perform conditional image generation in the mid- to high-resolution regime, with a reasonable computational budget and feasible training time. Other datasets in this order of size either have too few examples per class (e.g. AWA [46]) or too little inter- and intra-class variability. In contrast, the intra-class variability of COCO-Animals is very high for certain classes, e.g. bird and monkey, which span many subspecies. For more details, we refer to Section C in the supplementary material.

Evaluation metrics. For quantitative evaluation we use the Fréchet Inception distance (FID) [16] as the main metric,

Method	FFHQ				COCO-Animals			
	Best		Median		Best		Median	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑
BigGAN [5]	11.48	3.97	12.42	4.02	16.37	11.77	16.55	11.78
U-Net GAN	7.48	4.46	7.63	4.47	13.73	12.29	13.87	12.31

Table 1: Evaluation results on FFHQ and COCO-Animals. We report the best and median FID score across 5 runs and its corresponding IS, see Section 4.2 for discussion.

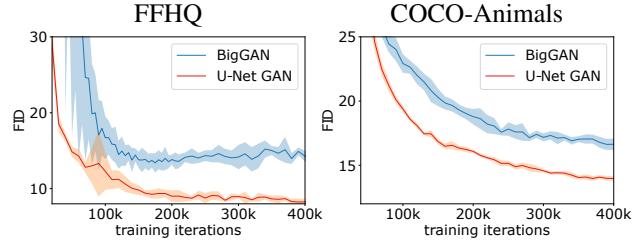


Figure 6: FID curves over iterations of the BigGAN model (blue) and the proposed U-Net GAN (red). Depicted are the FID mean and standard deviation across 5 runs per setting.

and additionally consider the Inception score (IS) [41]. Between the two, FID is a more comprehensive metric, which has been shown to be more consistent with human evaluation in assessing the realism and variation of the generated images [16], while IS is limited by what the Inception classifier can recognise, which is directly linked to its training data [3]. If one learns to generate something not present in the classifier’s training data (e.g. human faces) then IS can still be low despite generating high quality images since that image does not get classified as a distinct class.

In all our experiments, FID and IS are computed using 50k synthetic images, following [19]. By default all reported numbers correspond to the best or median FID of five independent runs achieved with 400k training iterations for FFHQ and COCO-Animals, and 800k training iterations for CelebA. For evaluation, we employ moving averages of the generator weights following [5, 19], with a decay of 0.9999. Note that we do not use any truncation tricks or rejection sampling for image generation.

Method	COCO-Animals	FFHQ
BigGAN [5]	16.55	12.42
U-Net based discriminator	15.86	10.86
+ CutMix augmentation	14.95	10.30
+ Consistency regularization	13.87	7.63

Table 2: Ablation study of the U-Net GAN model on FFHQ and COCO-Animals. Shown are the median FID scores. The proposed components lead to better performance, on average improving the median FID by 3.7 points over BigGAN [5]. See Section 4.2 for discussion.

Training details. We adopt the original training parameters of [5]. In particular, we use a uniformly distributed noise vector $z \in [-1, 1]^{140}$ as input to the generator, and the Adam optimizer [22] with learning rates of $1e-4$ and $5e-4$ for G and D^U . The number of warmup epochs n for consistency regularization is chosen to be 200 for COCO-Animals and 20 for FFHQ and CelebA. In contrast to [5], we operate with considerably smaller mini-batch sizes: 20 for FFHQ, 50 for CelebA and 80 for COCO-Animals. See Section E in the supplementary material for more details.

4.2. Results

We first test our proposed U-Net discriminator in two settings: unconditional image synthesis on FFHQ and class-conditional image synthesis on COCO-Animals, using the BigGAN model [5] as a baseline for comparison. We report our key results in Table 1 and Figure 6.

In the unconditional case, our model achieves the FID score of 7.48, which is an improvement of 4.0 FID points over the canonical BigGAN discriminator (see Table 1). In addition, the new U-Net discriminator also improves over the baseline in terms of the IS metric (3.97 vs. 4.46). The same effect is observed for the conditional image generation setting. Here, our U-Net GAN achieves an FID of 13.73, improving 2.64 points over BigGAN, as well as increases the IS score from 11.77 to 12.29. Figure 6 visualizes the mean FID behaviour over the training across 5 independent runs. From Figure 6 it is evident that the FID score drops for both models at the similar rate, with a constant offset for the U-Net GAN model, as well as the smaller standard deviation of FID. These results showcase the high potential of the new U-Net based discriminator. For a detailed comparison of the FID mean, median and standard deviation across 5 runs we refer to Table S2 in the supplementary material.

Qualitative results on FFHQ and COCO-Animals are shown in Figure 4 and Figure 5. Figure 4 displays human faces generated by U-Net GAN through linear interpolation in the latent space between two synthetic samples. We observe that the interpolations are semantically smooth between faces, i.e. an open mouth gradually becomes a closed mouth, hair progressively grows in length, beards or glasses smoothly fade or appear, and hair color changes seamlessly.

Method	FID ↓	IS ↑
PG-GAN [19]	7.30	–
COCO-GAN [27]	5.74	–
BigGAN [5]	4.54	3.23
U-Net GAN	2.95	3.43

Table 3: Comparison with the state-of-the-art models on CelebA (128×128). See Section 4.2 for discussion.

Furthermore, we notice that on several occasions men appear with pink beards. As FFHQ contains a fair share of people with pink hair, we suspect that our generator extrapolates hair color to beards, enabled by the global and local D^U feedback during the training. Figure 5 shows generated samples on COCO-Animals. We observe diverse images of high quality. We further notice that employing the class-conditional projection (as used in BigGAN) in the pixel output space of the decoder does not introduce class leakage or influence the class separation in any other way. These observations confirm that our U-Net GAN is effective in both unconditional and class-conditional image generation.

Ablation Study. In Table 2 we next analyze the individual effect of each of the proposed components of the U-Net GAN model (see Section 3 for details) to the baseline architecture of BigGAN on the FFHQ and COCO-Animals datasets, comparing the median FID scores. Note that each of these individual components builds on each other. As shown in Table 2, employing the U-Net architecture for the discriminator alone improves the median FID score from 12.42 to 10.86 for FFHQ and 16.55 to 15.86 for COCO-Animals. Adding the CutMix augmentation improves upon these scores even further, achieving FID of 10.30 for FFHQ and 14.95 for COCO-Animals. Note that we observe a similar improvement if we employ the CutMix augmentation during the BigGAN training as well. Employing the proposed consistency regularization in the segmenter D_{dec}^U output space on the CutMix images enables us to get the most out of the CutMix augmentation as well as allows to leverage better the per-pixel feedback of the U-Net discriminator, without imposing much computational or memory costs. In effect, the median FID score drops to 7.63 for FFHQ and to 13.87 for COCO-Animals. Overall, we observe that each proposed component of the U-Net GAN model leads to improved performance in terms of FID.

Comparison with state of the art. Table 3 shows that U-Net GAN compares favourably with the state of the art on the CelebA dataset. The BigGAN baseline computed already outperforms COCO-GAN, the best result reported in the literature to the best of our knowledge, lowering FID from 5.74 to 4.54, whereas U-Net GAN further improves FID to 2.95⁴. It is worth noting that BigGAN is the rep-

⁴FID scores for CelebA were computed with the standard TensorFlow Inception network for comparability. The PyTorch and TensorFlow FIDs for all datasets are presented in Table S1.

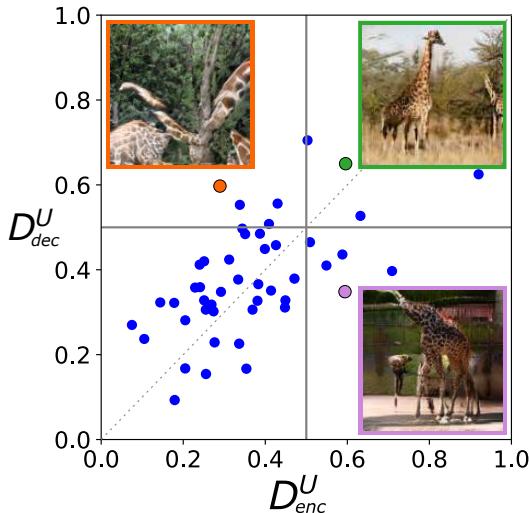


Figure 7: Visualization of the predictions of the encoder D_{enc}^U and decoder D_{dec}^U modules during training, within a batch of 50 generated samples. For visualization purposes, the D_{dec}^U score is averaged over all pixels in the output. Note that quite often decisions of D_{enc}^U and D_{dec}^U are not coherent with each other. As judged by the U-Net discriminator, samples in the upper left consist of locally plausible patterns, while not being globally coherent (example in orange), whereas samples in the lower right look globally coherent but have local inconsistencies (example in purple: giraffe with too many legs and vague background).

resentative of just one of the two well known state-of-the art GAN families, led by BigGAN and StyleGAN, and their respective further improvements [51, 53, 21]. While in this paper we base our framework on BigGAN, it would be interesting to also explore the application of the U-Net based discriminator for the StyleGAN family.

Discriminator response visualization. Experimentally we observe that D_{enc}^U and D_{dec}^U often assign different real-/fake scores per sample. Figure 7 visualizes the per-sample predictions for a complete training batch. Here, the decoder score is computed as the average per-pixel prediction. The scores correlate with each other but have a high variance. Points in the upper left quadrant correspond to samples that are assigned a high probability of being real by the decoder, but a low probability by the encoder. This implies realism on a local level, but not necessarily on a global one. Similarly, the lower right quadrant represents samples that are identified as realistic by the encoder, but contain unrealistic patches which cause a low decoder score. The fact that the encoder and decoder predictions are not tightly coupled further implies that these two components are complementary. In other words, the generator receives more pronounced feedback by the proposed U-Net discriminator than it would get from a standard GAN discriminator.

Characterizing the training dynamics. Both BigGAN and U-Net GAN experience similar stability issues, with

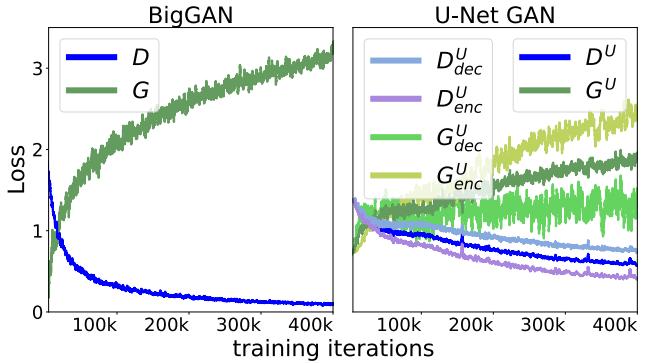


Figure 8: Comparison of the generator and discriminator loss behavior over training for U-Net GAN and BigGAN. The generator and discriminator loss of U-Net GAN is additionally split up into its encoder- and decoder components.

$\sim 60\%$ of all runs being successful. For U-Net GAN, training collapse occurs generally much earlier (~ 30 k iterations) than for BigGAN (> 200 k iterations, as also reported in [5]), allowing to discard failed runs earlier. Among successful runs for both models, we observe a lower standard deviation in the achieved FID scores, compared to the BigGAN baseline (see Table S2 in the supplementary material). Figure 8 depicts the evolution of the generator and discriminator losses (green and blue, respectively) for U-Net GAN and BigGAN over training. For U-Net GAN, the generator and discriminator losses are additionally split into the loss components of the U-Net encoder D_{enc}^U and decoder D_{dec}^U . The U-Net GAN discriminator loss decays slowly, while the BigGAN discriminator loss approaches zero rather quickly, which prevents further learning from the generator. This explains the FID gains of U-Net GAN and shows its potential to improve with longer training. The generator and discriminator loss parts from encoder (image-level) and decoder (pixel-level) show similar trends, i.e. we observe the same decay for D_{enc}^U and D_{dec}^U losses but with different scales. This is expected as D_{enc}^U can easily classify image as belonging to the real or fake class just by looking at one distinctive trait, while to achieve the same scale D_{dec}^U needs to make a uniform real or fake decision on all image pixels.

5. Conclusion

In this paper, we propose an alternative U-Net based architecture for the discriminator, which allows to provide both global and local feedback to the generator. In addition, we introduce a consistency regularization technique for the U-Net discriminator based on the CutMix data augmentation. We show that all the proposed changes result in a stronger discriminator, enabling the generator to synthesize images with varying levels of detail, maintaining global and local realism. We demonstrate the improvement over the state-of-the-art BigGAN model [5] in terms of the FID score on three different datasets.

References

- [1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 5
- [2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3
- [3] Shane T. Barratt and Rishi Sharma. A note on the inception score. *arXiv:1801.01973*, 2018. 7
- [4] David Berthelot, Nicholas Carlini, Ian G Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3
- [5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 5, 7, 8, 9, 12, 16, 19, 23
- [6] Ting Chen, Mario Lucic, Neil Houlsby, and Sylvain Gelly. On self modulation for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018. 5
- [7] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [8] Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C. Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5
- [9] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv:1708.04552*, 2017. 3
- [10] Rahul Dey, Felix Juefei-Xu, Vishnu Naresh Boddeti, and Marios Savvides. Rankgan: A maximum margin ranking gan for generating faces. In *Asian Conference on Computer Vision (ACCV)*, 2018. 2
- [11] Thang Doan, João Monteiro, Isabela Albuquerque, Bogdan Mazoure, Audrey Durand, Joelle Pineau, and R. Devon Hjelm. Online adaptative curriculum learning for gans. In *AAAI*, 2018. 2
- [12] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *International Conference on Learning Representations (ICLR)*, 2017. 5
- [13] Ishan P. Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2017. 2
- [14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2, 3, 5
- [15] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2
- [16] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 7, 12
- [17] Minyoung Huh, Shao-Hua Sun, and Ning Zhang. Feedback adversarial learning: Spatial feedback for improving generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [18] 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, 2017. 2
- [19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 7, 8, 12
- [20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 7, 12
- [21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019. 9
- [22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 8
- [23] Karol Kurach, Mario Lučić, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. The GAN landscape: Losses, architectures, regularization, and normalization. *arXiv:1807.04720*, 2018. 3
- [24] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallochi, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018. 2, 7, 12, 19
- [25] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2
- [26] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv:1705.02894*, 2017. 2, 5
- [27] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: Generation by parts via conditional coordinating. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 8, 12
- [28] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 2, 7, 12, 19

- [29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015. 2, 7, 12
- [30] Mario Lučić, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3
- [31] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, and Zhen Wang. Multi-class generative adversarial networks with the L2 loss function. *arXiv:1611.04076*, 2016. 2
- [32] Jacob Menick and Nal Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. In *International Conference on Learning Representations (ICLR)*, 2019. 1
- [33] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014. 2
- [34] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2
- [35] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations (ICLR)*, 2018. 5
- [36] Gonçalo Mordido, Haojin Yang, and Christoph Meinel. Dropout-gan: Learning from a dynamic ensemble of discriminators. *arXiv:1807.11346*, 2018. 2
- [37] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2, 5
- [38] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016. 2
- [39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2, 3
- [40] Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2
- [41] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 7, 12
- [42] Rishi Sharma, Shane T. Barratt, Stefano Ermon, and Vijay S. Pande. Improved training with curriculum gans. *arXiv:1807.09295*, 2018. 2
- [43] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning (ICML)*, 2019. 3
- [44] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *IJCAI*, 2019. 3
- [45] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2
- [46] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *Transactions on Pattern Analysis and Machine Intelligence*, 2018. 7
- [47] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyun Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 4
- [48] Dan Zhang and Anna Khoreva. PA-GAN: Improving GAN training by progressive augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2
- [49] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 3, 4
- [50] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine learning (ICML)*, 2019. 1, 2, 5
- [51] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 3, 9
- [52] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016. 2
- [53] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans. *arXiv preprint arXiv:2002.04724*, 2020. 9

Supplementary Material

This supplementary material complements the presentation of U-Net GAN in the main paper with the following:

- Additional quantitative results in Section A;
- Exemplar synthetic images on FFHQ in Section B and on COCO-Animals in Section C.
- Network architectures and hyperparameter settings in Section E.

A. Additional Evaluations

Here we provide more detailed evaluation of the results presented in the main paper. In Table S1 we report the inception metrics for images generated on FFHQ [20], COCO-Animals [28, 24] and CelebA [29] at resolution 256×256 , 128×128 , and 128×128 , respectively. In particular, we report the Fréchet Inception distance (FID) [16] and the Inception score (IS) [41] computed by both the PyTorch⁵ and TensorFlow⁶ implementations. Note that the difference between two implementations lies in using either the TensorFlow or the PyTorch in-built inception network to calculate IS and FID, resulting in slightly different scores. In all experiments, FID and IS are computed using 50k synthetic images, following [19]. By default all reported numbers correspond to the best FID achieved with 400k training iterations for FFHQ and COCO-Animals, and 800k iterations for CelebA, using the PyTorch implementation.

In the unconditional case, on FFHQ, our model achieves FID of 7.48 (8.88 in TensorFlow), which is an improvement of 4.0 (6.04 in TensorFlow) FID points over the BigGAN discriminator [5]. The same effect is observed for the conditional image generation setting on COCO-Animals. Here, our U-Net GAN achieves FID of 13.73 (13.96 in TensorFlow), improving 2.64 (2.46 in TensorFlow) points over BigGAN. To compare with other state-of-the-art models we additionally evaluate U-Net GAN on CelebA for unconditional image synthesis. Our U-Net GAN achieves 2.95 FID (in TensorFlow), outperforming COCO-GAN [27], PG-GAN [19], and the BigGAN baseline [5].

Table S2 shows that U-Net GAN does not only outperform the BigGAN baseline in terms of the best recorded FID, but also with respect to the mean, median and standard deviation computed over 5 independent runs. Note the strong drop in standard deviation from 0.24 to 0.11 on COCO-Animals and from 0.16 to 0.04 on CelebA.

B. Qualitative Results on FFHQ

Here we present more qualitative results of U-Net GAN on FFHQ [20]. We use FFHQ for unconditional image synthesis and generate images with a resolution of 256×256 .

⁵<https://github.com/ajbrock/BigGAN-PyTorch>

⁶<https://github.com/bioinf-jku/TTUR>

Dataset	Method	PyTorch		TensorFlow	
		FID ↓	IS ↑	FID ↓	IS ↑
FFHQ (256×256)	BigGAN [5]	11.48	3.97	14.92	3.96
	U-Net GAN	7.48	4.46	8.88	4.50
COCO-Animals (128×128)	BigGAN [5]	16.37	11.77	16.42	11.34
	U-Net GAN	13.73	12.29	13.96	11.77
CelebA (128×128)	PG-GAN [19]	–	–	7.30	–
	COCO-GAN [27]	–	–	5.74	–
	BigGAN [5]	3.70	3.08	4.54	3.23
U-Net GAN		2.03	3.33	2.95	3.43

Table S1: Evaluation results on FFHQ, COCO-Animals and CelebA with PyTorch and TensorFlow FID/IS scores. The difference lies in the choice of framework in which the inception network is implemented, which is used to extract the inception metrics. See Section A for discussion.

Method	Dataset	FID			
		Best	Median	Mean	Std
BigGAN U-Net GAN	COCO-Animals	16.37	16.55	16.62	0.24
		13.73	13.87	13.88	0.11
BigGAN U-Net GAN	FFHQ	11.48	12.42	12.35	0.67
		7.48	7.63	7.73	0.56
BigGAN U-Net GAN	CelebA	3.70	3.89	3.94	0.16
		2.03	2.07	2.08	0.04

Table S2: Best, median, mean and std of FID values across 5 runs.

Generated FFHQ samples

Figure S1 shows samples of human faces generated by U-Net GAN on FFHQ. We observe diverse images of high quality, maintaining local and global realism.

Per-pixel U-Net discriminator feedback

In Figure S2 we visualize synthetic images and their corresponding per-pixel feedback of the U-Net discriminator. Note that the U-Net discriminator provides a very detailed and spatially coherent response, which enables the generator to further improve the image quality.

Interpolations in the latent space

Figure S3 displays human faces generated by U-Net GAN through linear interpolation in the latent space between two synthetic samples. We observe that the interpolations are semantically smooth between faces, e.g. an open mouth gradually becomes a closed mouth, hair progressively grows or gets shorter in length, beards or glasses smoothly fade or appear, and hair color changes seamlessly.



Figure S1: Images generated by U-Net GAN trained on FFHQ with resolution 256×256 .



Figure S2: Samples generated by U-Net GAN and the corresponding real-fake predictions of the U-Net decoder. Brighter colors correspond to the discriminator confidence of pixel being real (and darker of being fake).

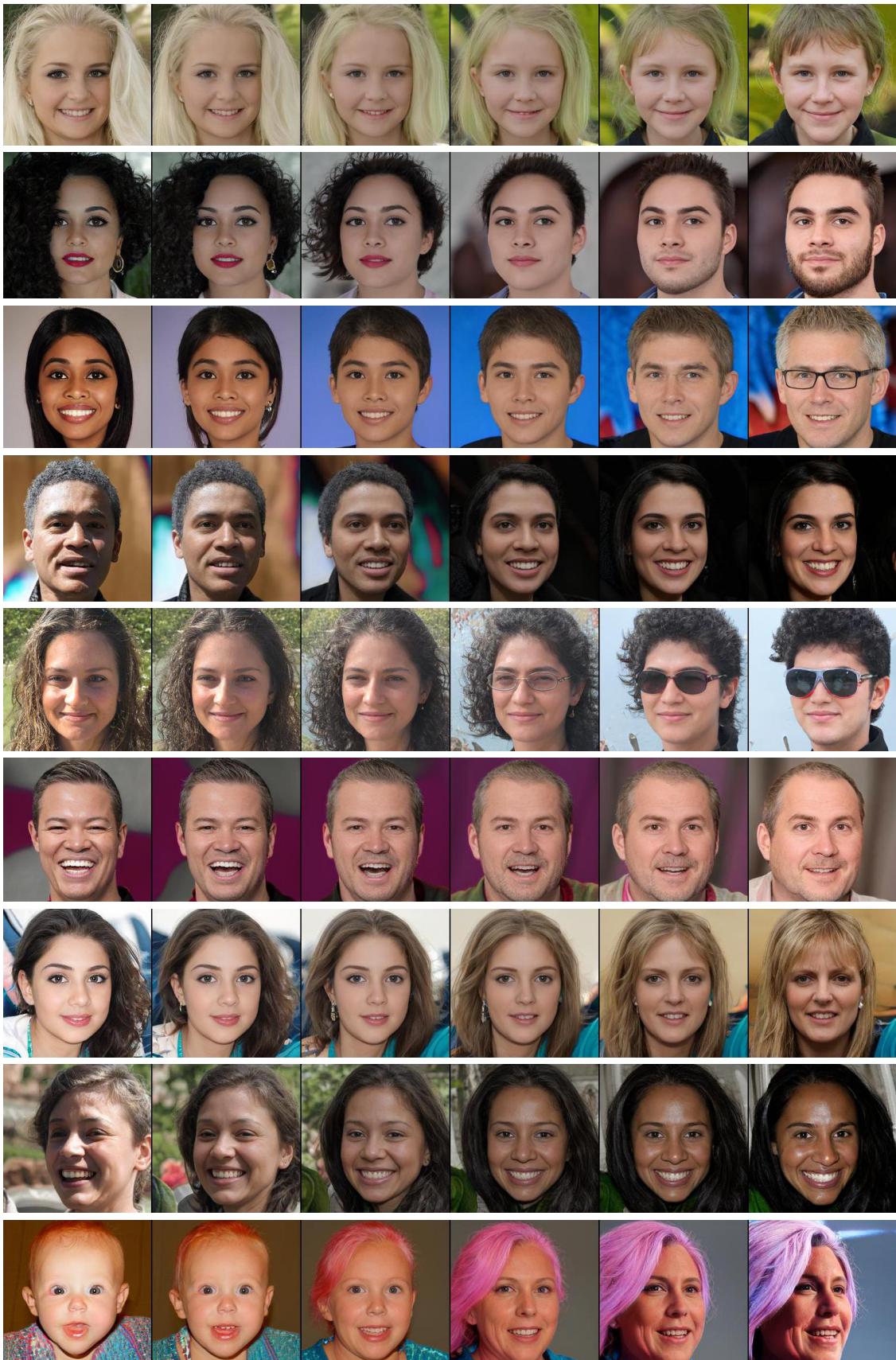


Figure S3: Images generated with U-Net GAN on FFHQ with resolution 256×256 when interpolating in the latent space.

Comparison between BigGAN and U-Net GAN

In Figure S4 we present a qualitative comparison of uncurated images generated with the unconditional BigGAN model [5] and our U-Net GAN. Note that the images generated by U-Net GAN exhibit finer details and maintain better local realism.

CutMix images and U-Net discriminator predictions

In Figure S5 we show more examples of the CutMix images and the corresponding U-Net based discriminator D^U predictions. Note that in many cases, the decoder output for fake image patches is darker than for real image ones. However, the predicted intensity for an identical local patch can change for different mixing scenarios. This indicates that the U-Net discriminator takes contextual information into account for local decisions.

BigGAN



U-Net GAN

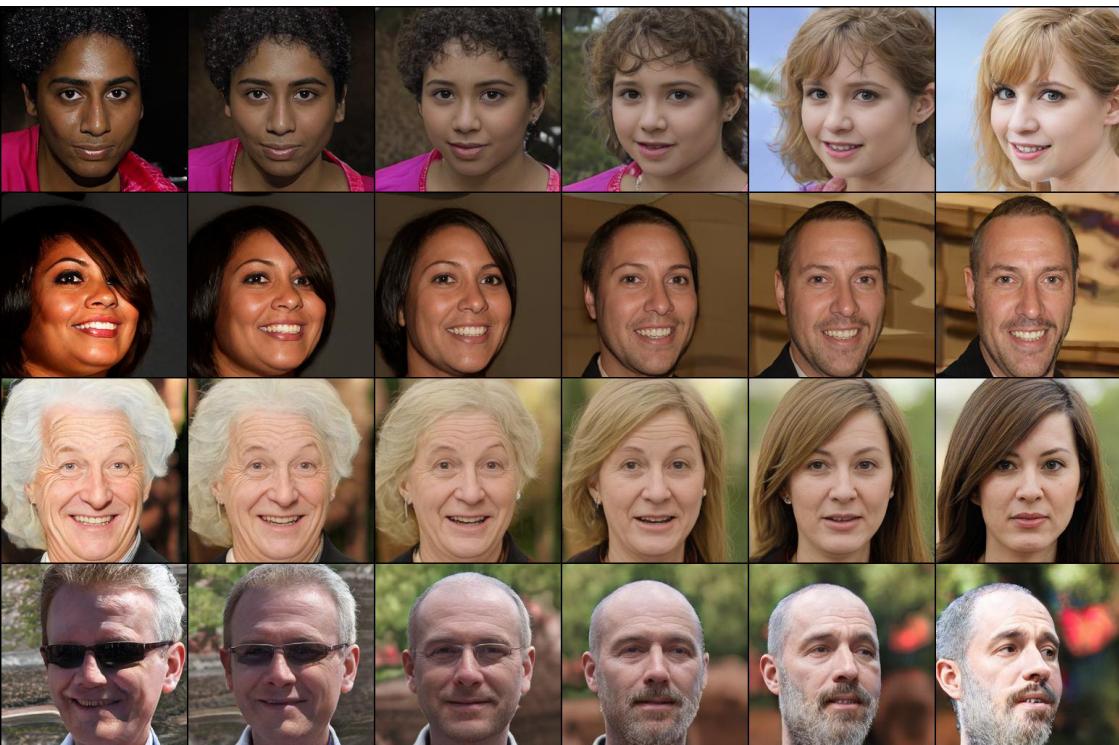


Figure S4: Qualitative comparison of uncurated images generated with the unconditional BigGAN model (top) and our U-Net GAN (bottom) on FFHQ with resolution 256×256 . Note that the images generated by U-Net GAN exhibit finer details and maintain better local realism.

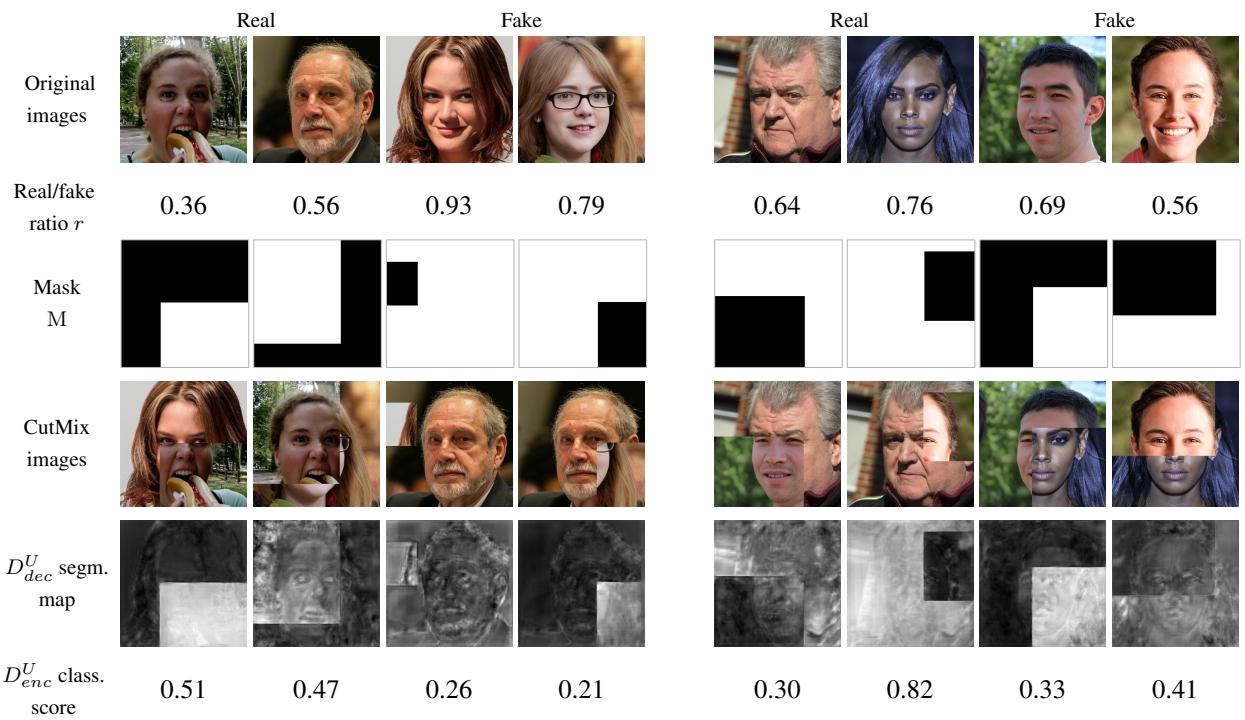
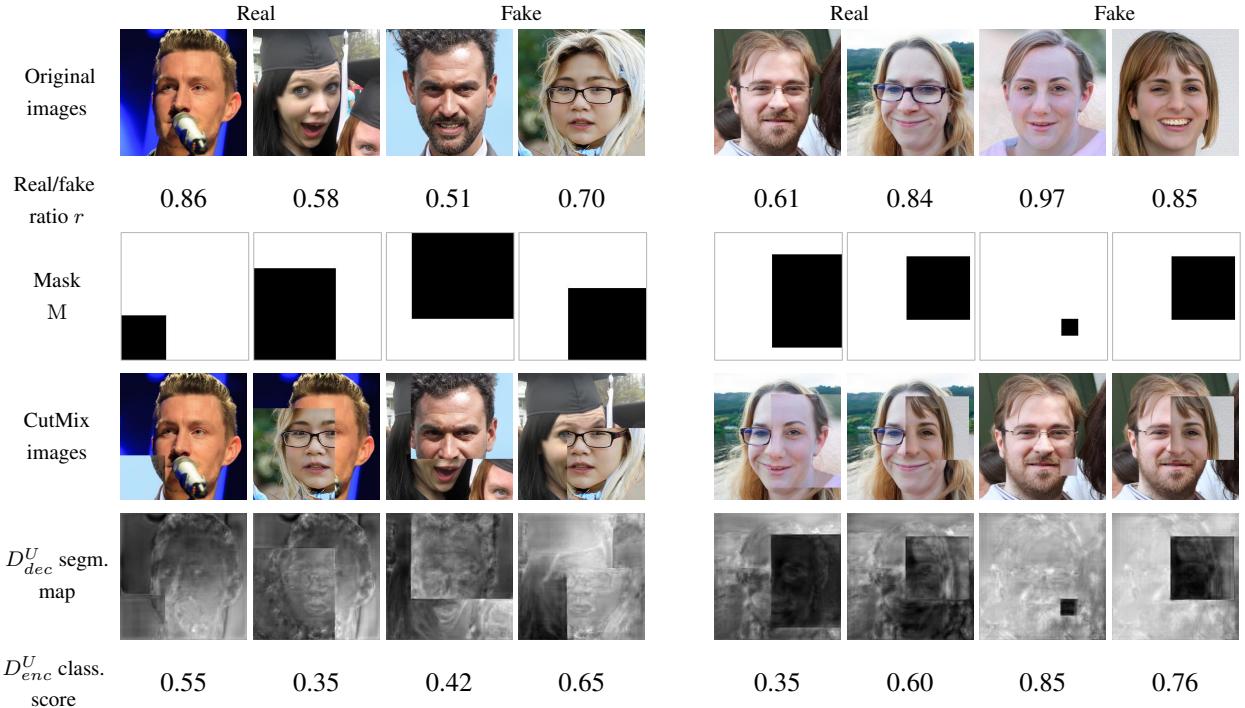


Figure S5: Visualization of the CutMix augmentation and the predictions of the U-Net discriminator on CutMix images. 1st row: real and fake samples. 2nd&3rd rows: sampled real/fake CutMix ratio r and corresponding binary masks M (color code: white for real, black for fake). 4th row: generated CutMix images from real and fake samples. 5th&6th row: the corresponding real/fake segmentation maps of the U-Net GAN decoder D_{dec}^U with the corresponding predicted classification scores by the encoder D_{enc}^U below.

C. Qualitative Results on COCO-Animals

Here we present more qualitative results of U-Net GAN on COCO-Animals [28, 24]. We use COCO-Animals for class conditional image synthesis and generate images with the resolution of 128×128 .

Generated COCO-Animals samples

Figure S6 shows generated samples of different classes on COCO-Animals. We observe images of good quality and high intra-class variation. We further notice that employing the class-conditional projection (as used in BigGAN) in the pixel output space of the decoder does not introduce class leakage or influence the class separation in any other way. These observations further confirm that our U-Net GAN is effective in class-conditional image generation as well.

Per-pixel U-Net discriminator feedback

Figure S7 shows generated examples and the corresponding per-pixel predictions of the U-Net discriminator. We observe that the resulting maps often tend to exhibit a bias towards objects.

Interpolations in the latent space

Figure S8 displays images generated on COCO-Animals by U-Net GAN through linear interpolation in the latent space between two synthetic samples. We observe that the interpolations are semantically smooth between different classes of animals, e.g. background seamlessly changes between two scenes, number of instances gradually increases or decreases, shape and color of objects smoothly changes from left to right.

D. Details on the COCO-Animals Dataset

COCO-Animals is a medium-sized ($\sim 38k$) dataset composed of 10 animal classes, and is intended for experiments that demand a high-resolution equivalent for CIFAR10. The categories are *bird*, *cat*, *dog*, *horse*, *cow*, *sheep*, *giraffe*, *zebra*, *elephant*, and *monkey*. The images are taken from COCO [28] and the OpenImages [24] subset that provides semantic label maps and binary mask and is also human-verified. The two datasets have a great overlap in animal classes. We take *all* images from COCO and the aforementioned OpenImages split in the categories *horse*, *cow*, *sheep*, *giraffe*, *zebra* and *elephant*. The *monkey* images are taken over directly from OpenImages, since this category contained more training samples than the next biggest COCO animal class *bear*. The class *bear* and *monkey* are not shared between COCO and OpenImages. Lastly, the categories *bird*, *cat* and *dog* contained vastly more samples than all other categories. For this reason, we took

over only a subset of the total of all images in these categories. These samples were picked from OpenImages only, for their better visual quality. To ensure good quality of the picked examples, we used the provided bounding boxes to filter out images in which the animal of interest is either too small or too big ($> 80\%$, $< 30\%$ of the image area for cats, $> 70\%$, $< 50\%$ for birds and dogs). The thresholds were chose such that the number of appropriate images is approximately equal.

E. Architectures and Training Details

Architecture details of the BigGAN model [5] and our U-Net discriminator are summarized in Table S3 and Table S4. From these tables it is easy to see that the encoder and decoder of the U-Net discriminator follow the original BigGAN discriminator and generator setups, respectively. One difference is that the number of input channels in the U-Net decoder is doubled, since encoder features are concatenated to the input features.

Table S4 presents two U-Net discriminator networks: a class-conditional discriminator for image resolution 128×128 , and an unconditional discriminator for resolution 256×256 . The decoder does not have 3 output channels (like the BigGAN generator that it is copied from), but $ch = 64$ channels, resulting in a feature map h of size $64 \times 128 \times 128$, to which a 1×1 convolution is applied to reduce the number of channels to 1. In the class-conditional architecture, a learned class-embedding is multiplied with the aforementioned 64-dimensional output h at every spatial position, and summed along the channel dimension (corresponding to the inner product). The resulting map of size $1 \times 128 \times 128$ is added to the output, leaving us with 128×128 logits.

We follow [5] for setting up the hyperparameters for training U-Net GAN, which are summarized in Table S5.

Hyperparameter	Value
Optimizer	Adam ($\beta_1 = 0, \beta_2 = 0.999$)
G's learning rate	1e-4 (256), 5e-5 (128)
D's learning rate	5e-4 (256), 2e-4 (128)
Batch size	20 (256), 80 (128)
Weight Initialization	Orthogonal

Table S5: Hyperparameters of U-Net GAN

Regarding the difference between class-conditional and unconditional image generation, it is worth noting that the CutMix regularization is applied only to samples within the same class. In other words, real and generated samples are mixed only within the class (e.g. real and fake zebras, but not real zebras with fake elephants).

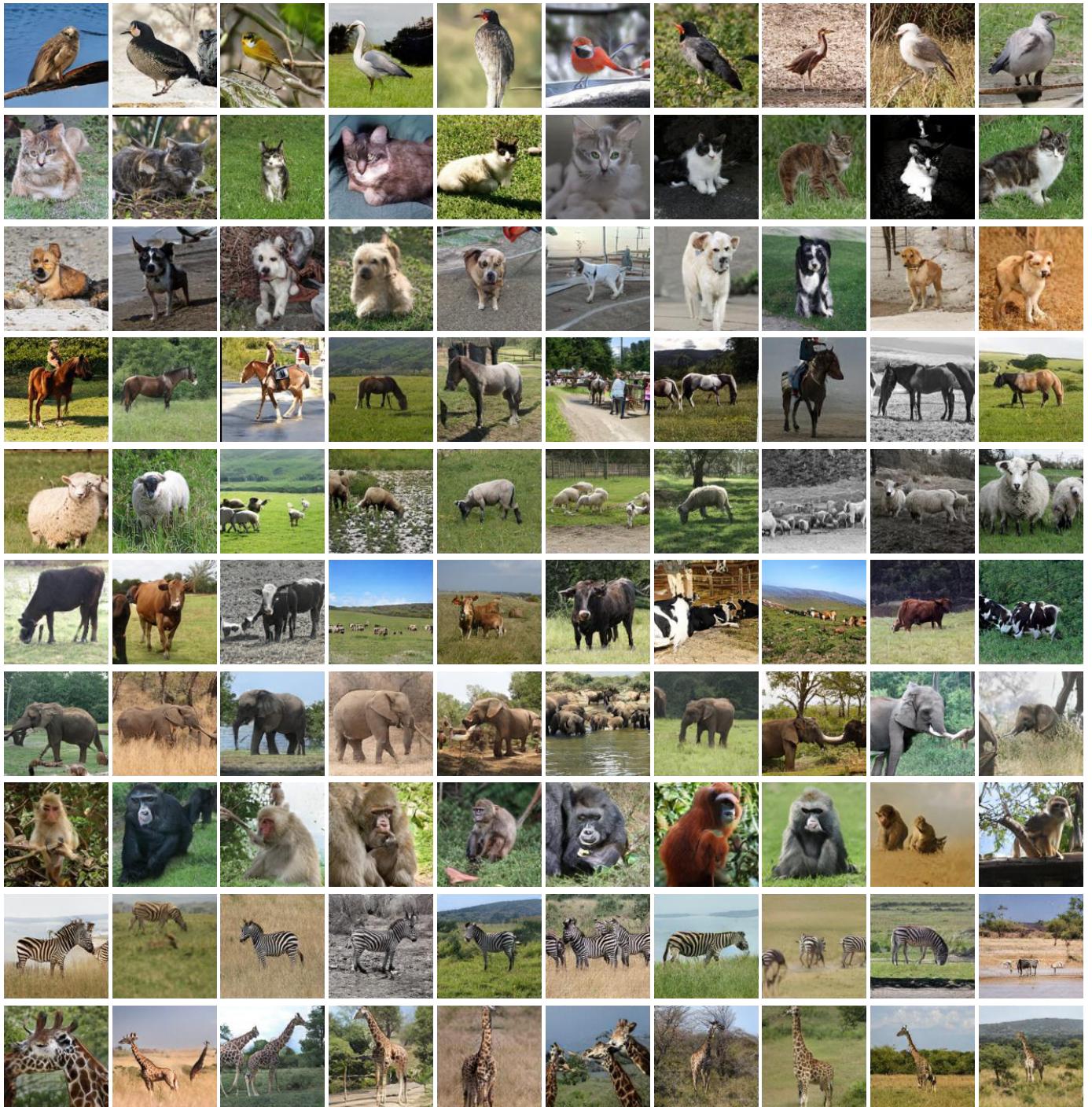


Figure S6: Images generated with U-Net GAN trained on COCO-Animals with resolution 128×128 .

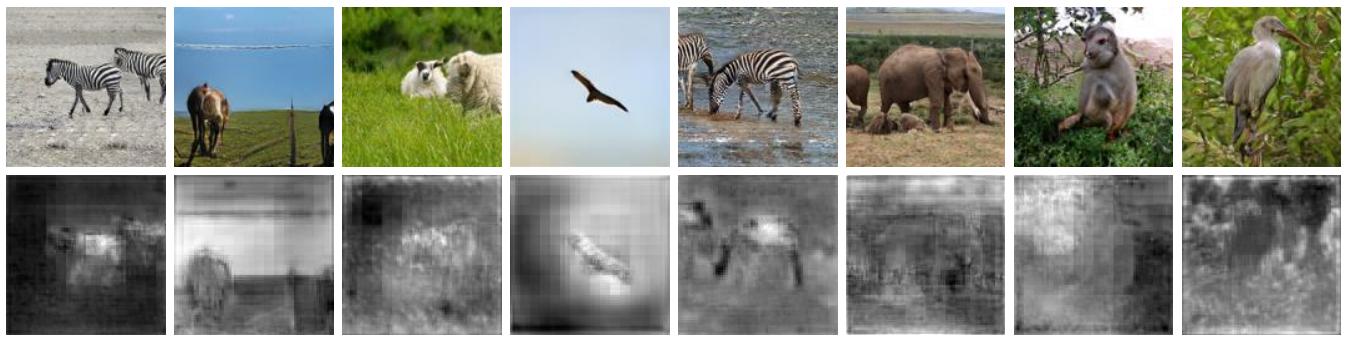


Figure S7: Generated samples on COCO-Animals and the corresponding U-Net decoder predictions. Brighter colors correspond to the discriminator confidence of pixel being real (and darker of being fake).

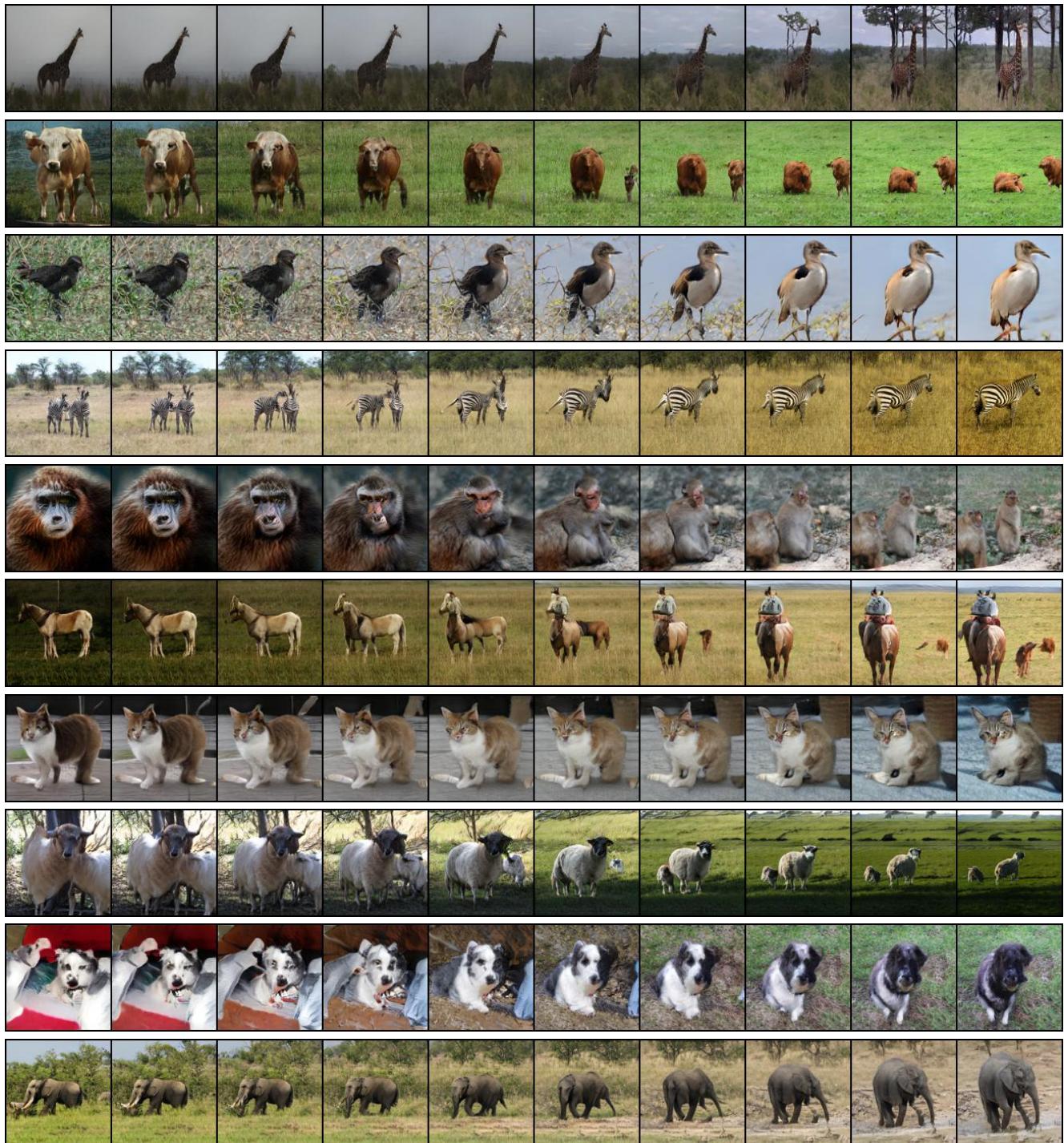


Figure S8: Images generated with U-Net GAN on COCO-Animals with resolution 128×128 when interpolating in the latent space between two synthetic samples (left to right).

(a) BigGAN Generator (128×128 , class-conditional)

$z \in \mathbb{R}^{120} \sim \mathcal{N}(0, I)$
Embed(y) $\in \mathbb{R}^{128}$
Linear ($20 + 128$) $\rightarrow 4 \times 4 \times 16ch$
ResBlock up $16ch \rightarrow 16ch$
ResBlock up $16ch \rightarrow 8ch$
ResBlock up $8ch \rightarrow 4ch$
ResBlock up $4ch \rightarrow 2ch$
Non-Local Block (64×64)
ResBlock up $2ch \rightarrow ch$
BN, ReLU, 3×3 Conv $ch \rightarrow 3$
Tanh

(b) BigGAN Discriminator (128×128 , class-conditional)

RGB image $x \in \mathbb{R}^{128 \times 128 \times 3}$
ResBlock down $ch \rightarrow 2ch$
Non-Local Block (64×64)
ResBlock down $2ch \rightarrow 4ch$
ResBlock down $4ch \rightarrow 8ch$
ResBlock down $8ch \rightarrow 16ch$
ResBlock down $16ch \rightarrow 16ch$
ReLU, Global sum pooling
Embed(y) $\cdot h + (\text{linear} \rightarrow 1)$

(c) BigGAN Generator (256×256 , unconditional)

$z \in \mathbb{R}^{140} \sim \mathcal{N}(0, I)$
Linear ($20 + 128$) $\rightarrow 4 \times 4 \times 16ch$
ResBlock up $16ch \rightarrow 16ch$
ResBlock up $16ch \rightarrow 8ch$
ResBlock up $8ch \rightarrow 8ch$
ResBlock up $8ch \rightarrow 4ch$
ResBlock up $4ch \rightarrow 2ch$
Non-Local Block (128×128)
ResBlock up $2ch \rightarrow ch$
BN, ReLU, 3×3 Conv $ch \rightarrow 3$
Tanh

(d) BigGAN Discriminator (256×256 , unconditional)

RGB image $x \in \mathbb{R}^{256 \times 256 \times 3}$
ResBlock down $ch \rightarrow 2ch$
ResBlock down $2ch \rightarrow 4ch$
Non-Local Block (64×64)
ResBlock down $4ch \rightarrow 8ch$
ResBlock down $8ch \rightarrow 8ch$
ResBlock down $8ch \rightarrow 16ch$
ResBlock down $16ch \rightarrow 16ch$
ReLU, Global sum pooling
linear $\rightarrow 1$

Table S3: The BigGAN [5] generator and discriminator architectures for class-conditional and unconditional tasks of generating images at different resolutions. Top (a and b): The class-conditional BigGAN model for resolution 128×128 . Bottom (c and d): The BigGAN model for resolution 256×256 , modified to be *unconditional*.

(a) U-Net GAN Discriminator (256×256 , unconditional)

RGB image $x \in \mathbb{R}^{256 \times 256 \times 3}$
ResBlock down $ch \rightarrow 2ch$
ResBlock down $2ch \rightarrow 4ch$
Optional Non-Local Block (64×64)
ResBlock down $4ch \rightarrow 8ch$
ResBlock down $8ch \rightarrow 8ch$
ResBlock down $8ch \rightarrow 16ch$ *(see below)
ResBlock up $16ch \rightarrow 8ch$
ResBlock up $(8 + 8)ch \rightarrow 8ch$
ResBlock up $(8 + 8)ch \rightarrow 4ch$
ResBlock up $(4 + 4)ch \rightarrow 2ch$
ResBlock up $(2 + 2)ch \rightarrow ch$
ResBlock up $(ch + ch) \rightarrow ch$
ResBlock $ch \rightarrow 1$
Sigmoid
* ReLU, Global sum pooling, linear $\rightarrow 1$

(b) U-Net GAN Discriminator (128×128 , class-conditional)

RGB image $x \in \mathbb{R}^{128 \times 128 \times 3}$
ResBlock down $ch \rightarrow 2ch$
Optional Non-Local Block (64×64)
ResBlock down $2ch \rightarrow 4ch$
ResBlock down $8ch \rightarrow 8ch$
ResBlock down $8ch \rightarrow 16ch$ *(see below)
ResBlock up $16ch \rightarrow 8ch$
ResBlock up $(8 + 8)ch \rightarrow 4ch$
ResBlock up $(4 + 4)ch \rightarrow 2ch$
ResBlock up $(2 + 2)ch \rightarrow ch$
ResBlock up $(ch + ch) \rightarrow ch$
Embed(y) $\cdot h + (\text{Conv } ch \rightarrow 1)$
Sigmoid
* ReLU, Global sum pooling
Embed(y) $\cdot h + (\text{linear} \rightarrow 1)$

Table S4: The U-Net GAN discriminator architectures for class-conditional (a) and unconditional (b) tasks of generating images at resolution 128×128 and 256×256 , respectively.