

# RPGAN: GANs Interpretability via Random Routing

Andrey Voynov<sup>1</sup> Artem Babenko<sup>1,2</sup>

## Abstract

Generative adversarial networks (GANs) are currently the top choice for applications involving image generation. However, in practice, GANs are mostly used as black-box instruments, and we still lack a complete understanding of an underlying generation process. While several recent works address the interpretability of GANs, the proposed techniques require some form of supervision and cannot be applied for general data.

In this paper, we introduce Random Path Generative Adversarial Network (RPGAN) — an alternative design of GANs that can serve as a data-agnostic tool for generative model analysis. While the latent space of a typical GAN consists of input vectors, randomly sampled from the standard Gaussian distribution, the latent space of RPGAN consists of random paths in a generator network. As we show, this design allows to understand factors of variation, captured by different generator layers, providing their natural interpretability. With experiments on standard benchmarks, we demonstrate that RPGAN reveals several insights about the roles that different layers play in the image generation process. Aside from interpretability, the RPGAN model also provides competitive generation quality and enables efficient incremental learning on new data. The PyTorch implementation of RPGAN is available online<sup>1</sup>.

## 1. Introduction

Nowadays, deep generative models are an active research direction in the machine learning community. The dominant methods for generative modeling, such as Generative Adversarial Networks (GANs), are currently able to produce

<sup>1</sup>Yandex, Russia <sup>2</sup>National Research University Higher School of Economics, Moscow, Russia. Correspondence to: Andrey Voynov <an.voynov@yandex.ru>.

Preprint. Under review.

<sup>1</sup><https://github.com/anvoynov/RandomPathGAN>

diverse photorealistic images (Brock et al., 2019; Karras et al., 2019). These methods are not only popular among academicians but are also a crucial component in a wide range of applications, including image editing (Isola et al., 2017; Zhu et al., 2017), super-resolution (Ledig et al., 2017), video generation (Wang et al., 2018) and many others.

Along with practical importance, a key benefit of accurate generative models is a more complete understanding of the internal structure of the data. Insights about the data generation process can result both in the development of new machine learning techniques as well as advances in industrial applications. However, most state-of-the-art generative models employ deep multi-layer architectures, which are difficult to interpret or explain. While many works investigate the interpretability of discriminative models (Zeiler & Fergus, 2014; Simonyan et al., 2014; Mahendran & Vedaldi, 2015), only a few (Chen et al., 2016; Bau et al., 2019; Yang et al., 2019) address the understanding of generative ones. Moreover, the techniques proposed in (Bau et al., 2019; Yang et al., 2019) require labeled datasets or pretrained models, which can be expensive to obtain, hence, have limited applicability. Overall, at the moment, there is no tool that provides an understanding of how GANs work on general data, and we aim to reduce this gap by our paper.

In this work, we propose the Random Path GAN (RPGAN) — an alternative design of GANs that allows the natural interpretability of the generator network. In traditional GAN generators, the stochastic component that influences individual samples is a noisy input vector, typically sampled from the standard Gaussian distribution. In contrast, RPGAN generators instead use stochastic routing during the forward pass as their source of stochasticity. In a nutshell, the RPGAN generator contains several instances of each generator layer, and only one random instance is activated during generation. The training of the RPGAN can be performed in the same adversarial manner as in traditional GANs. In the sections below, we show how RPGAN allows to understand the factors of variation captured by the particular layer and reveals several interesting findings about the image generation process, e.g., that different layers are “responsible for” coloring or objection location. As a practical advantage, RPGANs can be efficiently updated to new data via the simple addition of new instances of a particular layer, avoiding re-training the full model from scratch. Finally, we

observe that RPGANs allow to construct generative models without nonlinearities, which can significantly speed up the generation process for fully-connected layers. In summary, the main contributions of our paper are:

- We introduce RPGAN — GAN with an alternative source of stochasticity, based on random routing, enabling to analyse the roles of different layers in the generation process. To the best of our knowledge, RPGAN is the first **completely unsupervised** technique for GANs interpretability.
- With experiments on standard benchmarks, we reveal several insights about the generation process. Many insights confirm and extend the findings from prior works (Bau et al., 2019; Yang et al., 2019) that exploit some form of supervision. Note RPGAN is more general compared to the techniques from (Bau et al., 2019; Yang et al., 2019) as RPGAN does not require labeled data or pretrained models.

## 2. Related work

In this section we briefly describe connections of RPGAN to existing ideas from prior works

**Generative adversarial networks.** GANs are currently one of the main paradigms in generative modelling. Since the seminal paper on GANs by (Goodfellow et al., 2014), a plethora of alternative loss functions, architectures, normalizations, and regularization techniques were developed (Kurach et al., 2019). Today, state-of-the-art GANs are able to produce high-fidelity images, often indistinguishable from real ones (Brock et al., 2019; Karras et al., 2019). In essence, GANs consist of two networks – a generator and a discriminator, which are trained jointly in an adversarial manner. In standard GANs, the generation stochasticity is provided by the input noise vector. In RPGANs, we propose an alternative source of stochasticity by using a fixed input but random routes during forward pass in the generator.

**Specific GAN architectures.** Many prior works investigated different design choices for GANs, but to the best of our knowledge, none of them explicitly aimed to propose an interpretable GAN model. (Hoang et al., 2018) proposed the use of several independent generators to address the mode collapse problem. (Chavdarova & Fleuret, 2018) employ several auxiliary local generators and discriminators to improve mode coverage as well. (Huang et al., 2017) use layer-wise generators and discriminators to enforce hidden representations produced by layers of the generator to be similar to the corresponding representations produced by a reversed classification network. Important differences of RPGAN compared to the works described above is that it uses random routes as its latent space and does not enforce to mimic the latent representations of pretrained classifiers.

**Interpretability.** While the interpretability of models based on deep neural networks is an important research direction, most existing work addresses the interpretability of discriminative models. These works typically aim to understand the internal representations of networks (Zeiler & Fergus, 2014; Simonyan et al., 2014; Mahendran & Vedaldi, 2015; Dosovitskiy & Brox, 2016) or explain decisions produced by the network for particular samples (Sundararajan et al., 2017; Bach et al., 2015; Simonyan et al., 2014). However, only a few works address the interpretability of generative models. Related work by (Bau et al., 2019) develops a technique that allows to identify which parts of the generator are responsible for the generation of different objects. Note, that the technique from (Bau et al., 2019) requires a pre-trained segmentation network and cannot be directly applied to several benchmarks, e.g., CIFAR-10 or MNIST. A recent work (Yang et al., 2019) also aims to understand the roles of different generator layers, but their approach relies on pretrained classifiers; hence, it has limited applicability.

In contrast, RPGAN does not require any auxiliary models or other forms of supervision and can be applied to general data. Some of our findings confirm the results from (Bau et al., 2019; Yang et al., 2019), providing stronger evidence about the roles of different layers in the generation process.

## 3. Random Path GAN

### 3.1. Motivation

Before a formal description, we provide intuition behind the RPGAN model. Several prior works have demonstrated that in discriminative convolutional neural networks, different layers are “responsible” for different levels of abstraction (Zeiler & Fergus, 2014; Babenko et al., 2014). For instance, earlier layers detect small texture patterns, while activations in deeper layers correspond to semantically meaningful concepts. Similarly, in our paper we aim to understand the roles that different GAN layers play in image generation. Thus, we propose an architecture that provides a direct way to interpret the impact of individual layers. For a given architecture, RPGAN has several instances of each layer in its architecture. During the forward pass, a random instance of each layer is used to generate a particular image. Therefore, one can analyze the role of RPGAN layers by visualizing how different instances of each layer affect the image.

### 3.2. Model

Here we formally describe a structure of the RPGAN model. Similarly to the standard GAN architectures, RPGAN consists of two networks – a generator and a discriminator. The RPGAN discriminator operates exactly like discriminators in standard GANs, so we focus on the generator description.

The RPGAN generator consists of several consequent *buck-*

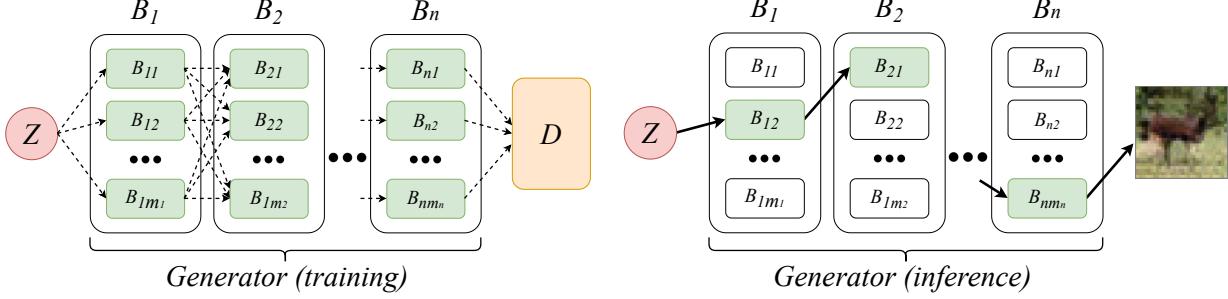


Figure 1. Design of the proposed RPGAN model. The RPGAN generator consists of several buckets  $B_1, \dots, B_n$ , corresponding to layers of the generator network. Each bucket contains several instances of the corresponding layer. During forward pass, a random instance from each bucket is “activated” to generate a particular image.

ets  $B_1, \dots, B_n$ , where each bucket corresponds to a particular computational unit of a generator, e.g., a ResNet block (He et al., 2015), a convolutional layer with a nonlinearity or any other. We associate each bucket with a layer (or several layers) in the generator architecture, which we aim to interpret or analyze. In each bucket, RPGAN maintains several independent instances of the corresponding computational unit  $B_i = \{B_{i1}, \dots, B_{im_i}\}$ , see Figure 1.

RPGAN generator performs forward pass as follows. For each  $i=1, \dots, n-1$  a random instance from the bucket  $B_i$  produces an intermediate output tensor that is passed to a random instance from the next bucket  $B_{i+1}$ . An instance from the first bucket  $B_1$  always receives a fixed input vector  $Z$ , which is the same for different forward passes. Therefore, the generator stochasticity arises only from a random path that goes from  $Z$  to an output image, using only a single instance from each bucket. Formally, during each forward pass, we randomly choose indices  $s_1, \dots, s_n$  with  $1 \leq s_i \leq m_i$ . The generator output is then computed as  $B_{ns_n} \circ \dots \circ B_{2s_2} \circ B_{1s_1}(Z)$ , see Figure 1. In other words, the generator defines a map from the Cartesian product  $\langle m_1 \rangle \times \langle m_2 \rangle \times \dots \times \langle m_n \rangle$  to the image space. Note that we can take an arbitrary existing GAN model, group its generator layers into buckets and replicate them into multiple instances. In these terms, the original model can be treated as RPGAN with a single instance in each bucket and random input noise. Note that during generation, RPGAN performs the same number of operations as standard GANs.

By its design, RPGAN with buckets  $B_1, \dots, B_n$  and a constant input  $Z$  is able to generate at most  $|B_1| \times \dots \times |B_n|$  different samples were  $|B_k|$  is the number of instances in the bucket  $B_k$ . Nevertheless, this number is typically much larger compared to the training set size. We argue that the probability space of random paths can serve as a latent space to generate high-quality images, as confirmed by the experiments below. The model is highly flexible to the choice of generator and discriminator architectures as well as to the loss function and learning strategy.

**Instance diversity loss.** To guarantee that the instances in

a particular bucket are different, we also add a specific diversity term in the generator loss function. The motivation for this term is to prevent instances  $B_{ki}, B_{kj}$  from learning the same weights. Let  $W$  be the set of all parameters of the generator. For each parameter  $w \in W$  there is a set of its instances  $\{w^{(1)}, \dots, w^{(m_w)}\}$  in the RPGAN model. Then we enforce the instances to be different by the additional loss term  $-\sum_{w \in W, i \neq j} \text{MSE}\left(\frac{w^{(i)}}{s_w}, \frac{w^{(j)}}{s_w}\right)$ . Here we also normalize by the standard deviation of  $s_w$  of all parameters from different instances that correspond to the same layer. This normalization effectively guarantees that all buckets contribute to the diversity term.

## 4. Experiments

**Architecture.** In all the experiments, we use ResNet-like generators with spectral normalization and the hinge loss (SN-ResNet) as in (Miyato et al., 2018). The instances in the first bucket are fully-connected layers, the instances in the last bucket are convolutional layers and instances in all other buckets are residual blocks with two convolutions and a skip connection. If not stated otherwise, all the buckets have the same number of instances. Additional experiments with other architectures are provided in supplementary.

**Datasets.** We performed experiments on CIFAR-10 (Krizhevsky et al., 2009), LSUN-bedroom (Yu et al., 2015) and Anime Faces (Jin et al., 2017) datasets. For different datasets, we use different numbers of discriminator steps per one generator step  $d_{steps}$  and different numbers of instances in each bucket  $n_{in}$ . The parameters used for three datasets are summarized in Table 1. In the last column, we also report *Coverage*, which is the ratio of the latent space cardinality (which equals the number of buckets to

Dataset	Size	Buckets	$n_{in}$	$d_{steps}$	Batch	Coverage
CIFAR-10	$32 \times 32$	5	40	5	64	2048
AnimeFaces	$64 \times 64$	6	20	1	32	$\approx 2970$
LSUN	$128 \times 128$	7	20	1	16	$\approx 420$

Table 1. The details of architectures and training protocols.

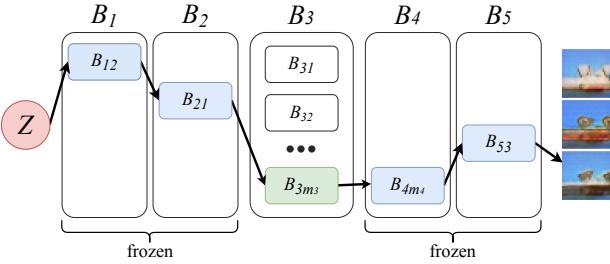


Figure 2. The five-bucket RPGAN model with “frozen” instances in buckets 1, 2, 4 and 5. In this case the stochasticity comes only from the third bucket. Analyzing the generated images allows to identify factors of variation, affected by the third bucket.

the power  $n_{in}$ ) to the dataset size. Intuitively, large coverage guarantees that RPGAN has a sufficiently rich latent space of generator routes to capture the reference dataset. In the experiments below, we demonstrate that even moderate coverage is sufficient to generate high-fidelity images (see the LSUN-bedroom dataset with coverage  $\approx 420$ ).

**Training details.** We use the Adam optimizer with learning rate equal to  $0.25 \times 10^{-3}$ ,  $\beta_1, \beta_2$  equal to 0.5, 0.999 and train the model for  $45 \times 10^4$  generator steps for CIFAR-10 and  $25 \times 10^4$  generator steps for Anime Faces and LSUN-bedroom datasets. During training, we also learn the unique input vector  $Z$ . We observed that a learnable  $Z$  slightly improves the final generation quality and stabilizes the learning process. Training is performed by passing  $Z$  through  $N$  independent random paths. Formally, let  $\{x_1, \dots, x_N\}$  be a batch of samples received from a bucket  $B_k$ . To pass this batch through the bucket  $B_{k+1}$  we take random instances  $B_{ki_1}, \dots, B_{ki_N}$  and form a new batch  $\{B_{ki_1}(x_1), \dots, B_{ki_N}(x_N)\}$ . In all the experiments, we use the same training protocols for both RPGAN and the standard GAN of the same generator architecture. Note that despite a larger number of learnable parameters, RPGAN does not require more data or training time to achieve the same quality, compared to standard GANs.

#### 4.1. Do different generator layers affect different factors of variations?

In the first series of experiments, we investigate the “responsibility” of different generator layers. With RPGAN, this can be performed with a technique schematically presented on Figure 2. In this example, the goal is to understand the role of the third bucket  $B_3$  in a five-bucket generator. To this end, we fix the instances from all other buckets  $B_1, B_2, B_4, B_5$ , shown in blue on Figure 2. Then we generate images corresponding to routes that contain all the fixed instances, with the stochasticity coming only from varying instances from the target bucket  $B_3$ . By inspecting the distribution of the obtained images, one can understand what factors of variation are affected by  $B_3$ .

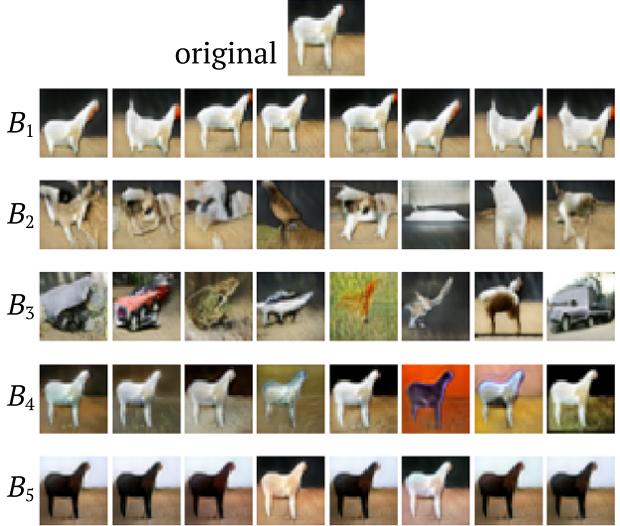


Figure 3. Top image: an image generated with a fixed sequence of instances. Horizontal lines: images generated by the same sequence of instances in all buckets and one *unfrozen* bucket. In the unfrozen bucket we choose eight arbitrary instances to avoid excessively large figures. The generated images allow to interpret factors of variation, captured by different buckets.

Figure 3 shows an example of image distributions obtained by varying instances in different buckets of five-bucket RPGAN for CIFAR-10. Each row shows how the original generated image can change if different instances from the corresponding bucket are used during generation. Other qualitative examples for different datasets are provided in the supplementary material. Several observations from these figures are the following. The first bucket typically does not influence coloring and mostly affects small objects’ deformations. The intermediate buckets have the largest influence on semantics. The last two buckets are mostly responsible for coloring and do not influence the content shape. In particular, on Figure 3, the fourth layer widely varies color, while the fifth acts as a general tone corrector. Note that these observations are consistent with the insights revealed by (Bau et al., 2019; Yang et al., 2019) but are obtained without any supervision. In contrast, techniques from (Bau et al., 2019; Yang et al., 2019) often cannot be applied since they require pretrained segmentation or classification models (which can be unavailable, e.g., for CIFAR and AnimeFaces).

To verify the observations above, we perform more rigorous experiments that evaluate the roles of different layers quantitatively. Let us define a metric  $d_{\text{img}}$  that evaluates the similarity between two generated images. Note that different metrics are able to capture different factors of variations (e.g., in terms of semantic, color histogram, etc.), and we describe two particular choices of  $d_{\text{img}}$  below. Then we choose a random route in the RPGAN generator and for each bucket  $B_l$  generate four images  $Im_1^{(l)}, \dots, Im_4^{(l)}$ , varying instances in  $B_l$ . In other words, we take four ran-

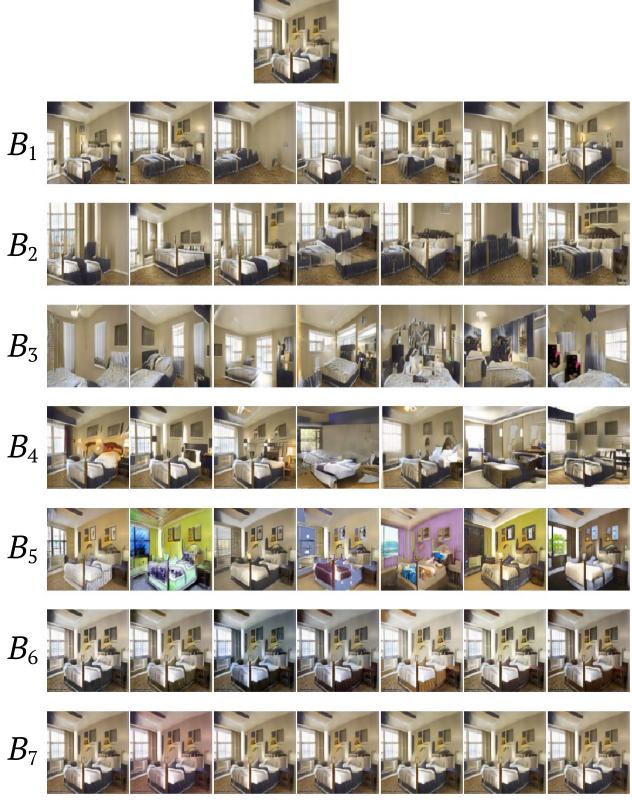


Figure 4. Image distributions for 7-bucket RPGAN and LSUN. dom images from each line of the table in the Figure 3. Then we measure diversity w.r.t.  $d_{\text{img}}$  captured by  $B_l$  as a ratio

$$D_{l \rightarrow 1, d_{\text{img}}} = \frac{\sum_{i \neq j} d_{\text{img}}(Im_i^{(l)}, Im_j^{(l)})}{\sum_{i \neq j} d_{\text{img}}(Im_i^{(1)}, Im_j^{(1)})} \quad (1)$$

Intuitively, the formula above computes the relative diversity with respect to the first bucket, which typically captures the smallest amount of variations in our experiments. We then average these ratios over 100 independent routes. In essence, higher values of averaged ratio  $D_{l \rightarrow 1, d_{\text{img}}}$  imply higher diversity of  $B_l$  compared to the first bucket in terms of the metric  $d_{\text{img}}$ , which implies that  $B_l$  strongly affects the factor of variation captured by particular metric  $d_{\text{img}}$ .

In experiments, we use two following metrics:

- $d_{\text{semantic}}(img_1, img_2)$ , capturing semantic, is based on the recent LPIPS (Zhang et al., 2018), which was shown to indicate perceptual similarity.
- $d_{\text{color}}(img_1, img_2)$ , measuring difference in colorings. Namely, we take the Hellinger distance between color histograms of generated samples. For each color channel, we split the range  $[0, \dots, 255]$  into 25 equal buckets and evaluate the discrete distribution defined

by the frequencies the image pixel intensities appear in a given bucket. Then the Hellinger distance between two quantified color distributions is defined as

$$d_{\text{color}}(img_1, img_2) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{25} (\sqrt{p_i} - \sqrt{q_i})^2} . \text{ We compute it for each RGB channel independently.}$$

The plots of averaged values of both metrics on three datasets are presented on Figure 6, Figure 5 Figure 4. They reaffirm that the semantic diversity is the largest for the intermediate layers. On the contrary, the last buckets, which are closer to the output, do not influence semantics but have a higher impact in terms of color. Note that the first layer always shows the smallest variability in terms of both factors of variation. The last bucket affects color correction and color inversion and has a lower pallet variability impact. Overall, we summarize the findings that are common for CIFAR-10, LSUN, and Anime Faces datasets as:

- The earlier layers have a smaller variability and seem to be responsible for the viewpoint and the position of the object on the image.
- The semantic details of the image content are mostly influenced by the intermediate layers.

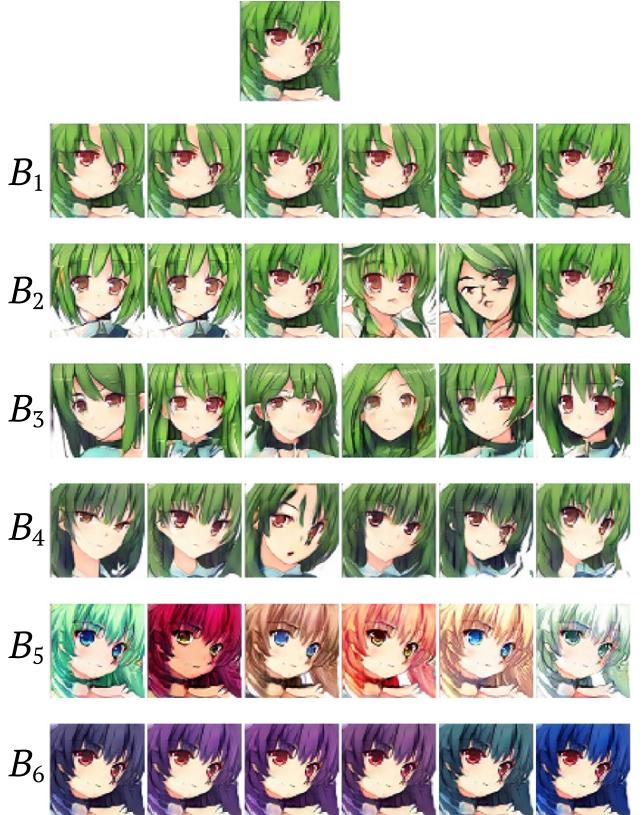


Figure 5. Image distributions for 6-bucket RPGAN and Anime-Faces.

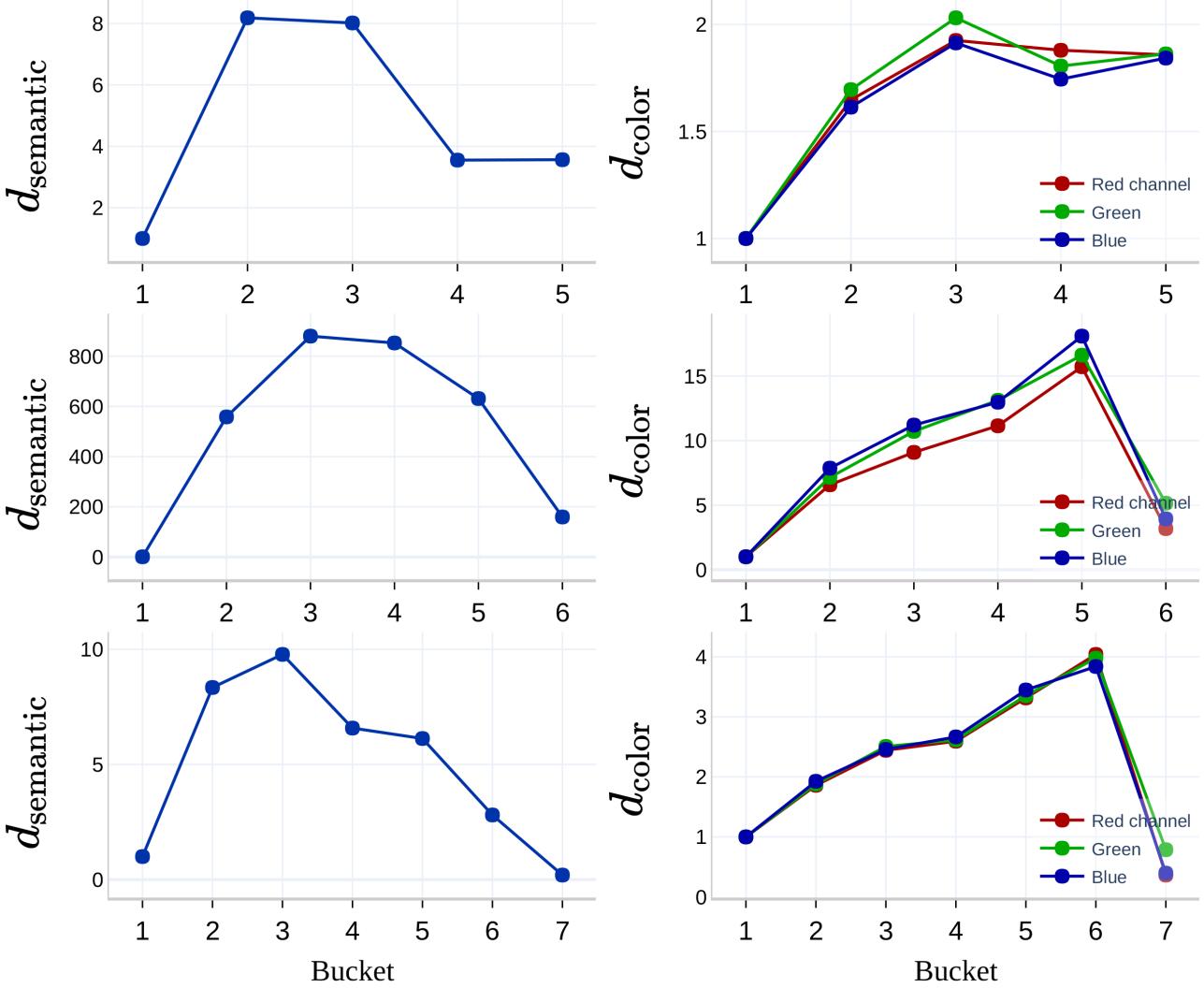


Figure 6. The quantitative evaluation of the extent  $D_{l \rightarrow 1, d}$  to which different generator parts affect different factors of variation for (Top) the five-bucket RPGAN and CIFAR-10; (Middle) the six-bucket RPGAN and AnimeFaces; (Bottom) the seven-bucket RPGAN and LSUN;

- The last layers typically affect only coloring scheme and do not affect content semantics or image geometry.

Note, that these conclusions can differ for other datasets or other generator architectures. For instance, for the four-bucket generator and MNIST (Figure 9, left) or randomly colored MNIST (Figure 8, left) the semantics are mostly determined by the first two buckets. Overall, the layers’ “responsibilities” depend on both an architecture and a particular dataset. Unlike prior techniques, RPGAN can be applied to any generator model and data domain, hence, provides a universal instrument for GAN interpretability.

#### 4.2. Are RPGAN interpretations valid for standard GANs?

In this subsection, we argue that the interpretations of different layers, obtained with RPGAN, are also valid for a

standard GAN generator of the same architecture. First, we demonstrate that both standard GAN and RPGAN trained under the same training protocol provide almost the same generation quality. As a standard evaluation measure, we use the Fréchet Inception Distance (FID) introduced in (Heusel et al., 2017). For evaluation on CIFAR-10, we use 50000 generated samples and the whole train dataset. For both standard GAN and RPGAN, we use ten independently trained generators and report minimal and average FID values in Table 2. In terms of FID RPGAN and the standard GAN perform with comparable generation quality.

model	min FID	average FID
Five-bucket RPGAN	16.9	20.8
SN-ResNet	16.75	18.7

Table 2. FID values for CIFAR-10.

To confirm that the layers of the standard GAN generator can

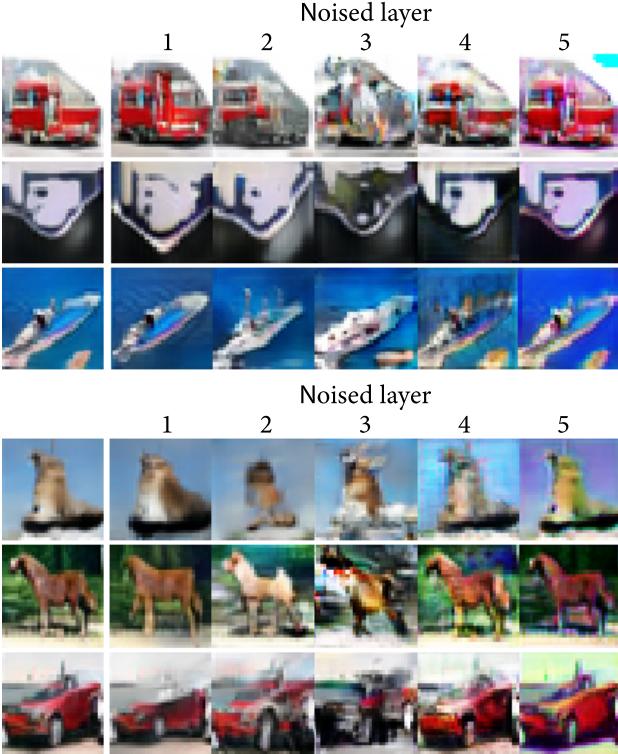


Figure 7. Images, produced by the standard SN-ResNet GAN with noise injection in the parameters of different generator layers. First column: original images, produced without generator perturbation. Other columns: images produced by perturbed generator.

be interpreted in the same way as the corresponding layers of its RPGAN counterpart, we perform the following experiment. We take a standard SN-ResNet GAN, consisting of five layers associated with the correspondent buckets in RPGAN, and train it on CIFAR-10. Then for each layer, we add normal noise to its weights. Intuitively, we expect that the noise injection in the particular layer will change generated images in terms of factors of variation, influenced by this layer. For instance, noise in the last two layers is expected to harm the coloring scheme, while noise in the intermediate layers is expected to bring maximal semantic damage. Several images, produced by perturbed layers, are presented on Figure 7. The images support the intuition described above and confirm that RPGAN may serve as an analysis tool for the underlying generator model. Note, however, that injecting noise per se cannot be used as a stand-alone interpretability method. The perturbed generators produce poor images, which are difficult to analyze. Meanwhile, RPGAN always generates good-looking images, which allows to identify the factors of variation, corresponding to the particular layer. For instance, see Figure 8 for the colored MNIST dataset. Figure 8 (left) shows plausible images, generated by varying RPGAN instances. In contrast, Figure 8 (right) demonstrates images from generators perturbed with

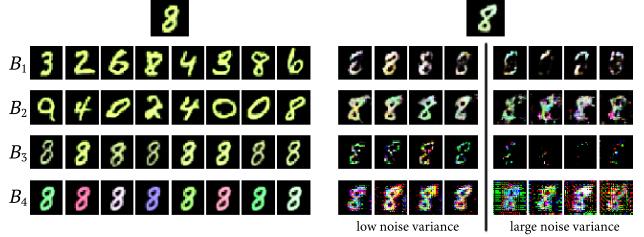


Figure 8. Left: images produced by varying instances in a particular bucket of RPGAN. Right: images produced by the standard GAN after parameters perturbation in a particular generator layer, with low and high normal noise variance.

small and large noise. For both noise magnitudes, these images are difficult to interpret. Of course, given the interpretations obtained via RPGAN, one can perceive similar patterns in the noisy generations, but noise injection alone is not sufficient for interpretability.

#### 4.3. Ablation on number of instances

Here we investigate the impact of  $n_{in}$  on the generation quality. We train RPGAN with the SN-ResNet generator on CIFAR-10 with a different values of  $n_{in}$ . The resulting FID values are presented on Table 3.

$n_{in}$	5	10	15	20	25	30	35	40	45	50
FID	61.0	25.6	23.4	20.0	19.7	18.0	17.4	17.1	18.6	22.9

Table 3. FID values for RPGAN with different  $n_{in}$ .

Overall, if  $n_{in}$  is too low, the latent space has insufficient cardinality to model real data. On the other hand, large  $n_{in}$  results in difficult training and can fail.

#### 4.4. Incremental learning with RPGAN

In the next experiment, we demonstrate that the RPGAN model is also a natural fit for the generative incremental learning task (see, e.g., (Wu et al., 2018)). Let us assume that the whole train dataset  $D$  is split into two disjoint subsets  $D = D_1 \cup D_2$ . Suppose that originally we have no samples from  $D_2$  and train a generative model to approximate a distribution defined by the subset  $D_1$ . Then, given additional samples from  $D_2$ , we aim to solve an incremental learning task — to update the model with new data without re-training it from scratch. The RPGAN model naturally allows to solve this task efficiently. First, we train an RPGAN generator with buckets  $B_1, \dots, B_n$  to approximate the distribution  $D_1$ . Once one aims to extend the generator with samples from  $D_2$ , one can add several new instances to the buckets that are responsible for the features that capture the difference between  $D_1$  and  $D_2$ . Then we optimize the generator to reproduce both  $D_1$  and  $D_2$  by training only the new instances. Thus, instead of training a new generator

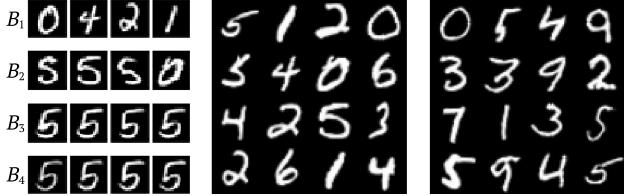


Figure 9. Incremental learning with RPGAN. *Left*: variations captured by different buckets with generator trained on  $\text{MNIST}_{0-6}$ . *Center*: images produced by a generator trained on  $\text{MNIST}_{0-6}$ . *Right*: images produced by a generator tuned on MNIST with only ten new instances training (see the details in the main text).

from scratch, we exploit the pretrained instances that are responsible for features, which are common for  $D_1$  and  $D_2$ . To illustrate this scenario, we take a partition of the MNIST hand-written digits dataset (LeCun, 1989) into two subsets  $\text{MNIST}_{0-6}$  and  $\text{MNIST}_{7-9}$  of digits from 0 to 6 and from 7 to 9 correspondingly. As for generator for  $\text{MNIST}_{0-6}$  we take a four-bucket RPGAN model with  $n_{in}$  equals to 20, 20, 20, 8. Note that the last bucket is much thinner than the others, as it turns out to be responsible for variations in writing style, which does not change much across the dataset. Then we train the generator on the subset  $\text{MNIST}_{0-6}$  of first 7 digits (see Figure 9, left and center). After that, we add five additional instances to each of the first two layers, obtaining a generator with  $n_{in}$  equals to 25, 25, 20, 8, and pretrained weights in all instances except for five in the first and in the second buckets. Then we train the extended model to fit the whole MNIST by optimizing only the ten new instances (see Figure 9, right).

#### 4.5. Linear map generator

As a surprising side effect of the RPGAN model, we discovered that decent generation quality can be achieved by the RPGAN generator **with no nonlinearities**, i.e., one can train the RPGAN generator with all instances consisting of linear transformations only. To demonstrate that, we take an RPGAN with the same ResNet-like generator architecture as in the experiments above. Then we replace all nonlin-

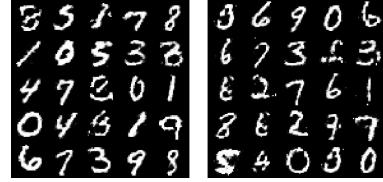


Figure 10. Digits generated by RPGAN without nonlinearities (*left*) and by its  $\times 2.2$  faster compression (*right*).

earities in the generator model by identity operations and train it on the CIFAR-10 dataset. The model demonstrates FID equal to 22.79 that is competitive to the state-of-the-art generative models of comparable sizes. Note that this approach fails for a standard GAN generator that maps a Gaussian distribution to an images distribution. Indeed, that generator would be a linear operator from a latent space with a Gaussian distribution in the images domain.

This purely linear generator architecture allows us to speed up the image generation process for fully-connected layers significantly. We group consequent buckets of fully-connected layers to form a new bucket. The instances in the new bucket are linear transformations that are products of the instances from the original buckets. To demonstrate this, we train a fully-connected generator network on the MNIST dataset, see Table 4. Then we join the last three buckets into a single one. Namely, we form a new bucket by instances defined as the linear operators  $B_{5k} \circ B_{4j} \circ B_{3i}$  where  $i, j, k$  are random indices of instances from the buckets  $B_3, B_4, B_5$  of the original generator. Thus instead of performing three multiplications of features vector from the second layer by matrices of the shapes  $256 \times 512, 512 \times 1024, 1024 \times 784$ , we perform a single multiplication by a  $256 \times 784$  matrix. In our experiments, we achieved  $\times 2.2$  speed up. Note, however, that after the compression, the latent space cardinality can decrease if a small subset of tuples  $(i, j, k)$  is used to populate the new bucket. Nevertheless, as random products of joining buckets are used, we expect that the generated images would be uniformly distributed in the space of images, produced by the uncompressed generator (see Figure 10 for visual comparison).

## 5. Conclusion

In this paper, we have proposed RPGAN — the unsupervised technique to interpret and analyze GAN models. RPGAN is based on an alternative generator design that allows natural interpretation of different layers via using random routing as a source of stochasticity. With experiments on several datasets, we provide evidence that different layers are responsible for the different factors of variation in generated images, which is consistent with findings from previous work. As a possible direction of future research, one can use the RPGAN analysis to construct efficient models, e.g., via identification of redundant parts of the generator for pruning or inference speedup.

Table 4. Fully connected RPGAN without nonlinearities and its compressed modification.

Original	Compressed
$z \in \mathbb{R}^{128}$	
fc, 32 blocks, 128	
fc, 32 blocks, 256	
fc, 32 blocks, 512	
fc, 16 blocks, 1024	fc, 128 blocks, 784
fc, 16 blocks, 784	
Tanh, reshape to $28 \times 28$	

## References

- Babenko, A., Slesarev, A., Chigorin, A., and Lempitsky, V. Neural codes for image retrieval. In *European conference on computer vision*, 2014.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10, 2015.
- Bau, D., Zhu, J.-Y., Strobelt, H., Bolei, Z., Tenenbaum, J. B., Freeman, W. T., and Torralba, A. Gan dissection: Visualizing and understanding generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.
- Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.
- Chavdarova, T. and Fleuret, F. Sgan: An alternative training of generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, 2016.
- Dosovitskiy, A. and Brox, T. Generating images with perceptual similarity metrics based on deep networks. In *Advances in neural information processing systems*, 2016.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.
- Hoang, Q., Nguyen, T. D., Le, T., and Phung, D. MGAN: Training generative adversarial nets with multiple generators. In *International Conference on Learning Representations*, 2018.
- Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., and Belongie, S. Stacked generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.
- Jin, Y., Zhang, J., Li, M., Tian, Y., and Zhu, H. Towards the high-quality anime characters generation with generative adversarial networks. In *Proceedings of the Machine Learning for Creativity and Design Workshop at NIPS*, 2017.
- Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- Krizhevsky, A. et al. Learning multiple layers of features from tiny images. Technical report, 2009.
- Kurach, K., Lučić, M., Zhai, X., Michalski, M., and Gelly, S. A large-scale study on regularization and normalization in gans. In *International Conference on Machine Learning*, 2019.
- LeCun, Y. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1989.
- Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., 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*, 2017.
- Lin, M., Chen, Q., and Yan, S. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.
- Mahendran, A. and Vedaldi, A. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.
- Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Sajjadi, M. S. M., Bachem, O., Lučić, M., Bousquet, O., and Gelly, S. Assessing Generative Models via Precision and Recall. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR)*, 2014.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. Video-to-video synthesis. In *Advances in Neural Information Processing Systems*, 2018.

Wu, C., Herranz, L., Liu, X., van de Weijer, J., Raducanu, B., et al. Memory replay gans: Learning to generate new categories without forgetting. In *Advances In Neural Information Processing Systems*, 2018.

Yang, C., Shen, Y., and Zhou, B. Semantic hierarchy emerges in deep generative representations for scene synthesis. *arXiv preprint arXiv:1911.09267*, 2019.

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 2014.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2017.

## A. Additional quantitative and qualitative results

**Generation quality.** To confirm that RPGAN provides the same generation quality as the standard GAN of the same backbone, we also compare them in terms of the recent precision-recall metrics (Sajjadi et al., 2018), see Figure 11. It shows precision-recall curves for two models trained on the CIFAR-10 dataset and demonstrates almost equal generation quality.

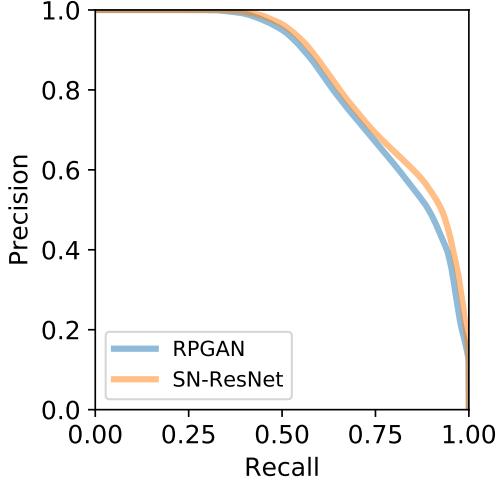


Figure 11. Precision-Recall curves for SN-ResNet GAN and RPGAN trained on CIFAR-10.

We also present several qualitative results of RPGAN generation on CIFAR-10 (Figure 12) Anime Faces (Figure 13) and LSUN (Figure 14). All the figures demonstrate that RPGAN does provide an understanding of responsibilities of different layers in generation.

**Wasserstein GAN.** Here we show that RPGAN can be used for the analysis of different generator architectures and learning strategies. Namely, we present plots for DCGAN-like generators consisting of consequent convolutional layers without skip connections. All the models were trained with the same hyperparameters as described in Section 4. Here we do not use spectral normalization and train generators as WGANs with weight penalty (Gulrajani et al., 2017). On the Figure 15 we show plots for a four-bucket generator trained on CIFAR-10. Additionally, we show plots for the five-bucket generator and CelebA-64x64 dataset on Figure 16. See Figure 17 for the quantitative evaluation for these two GANs and the RPGAN trained on colored MNIST dataset introduced in Section 4.2.

## B. RPGAN inversion for image editing

The discrete nature of the RPGAN latent space implies a simple procedure for RPGAN inversion, which can be useful for image editing. Let us have an RPGAN with buckets  $B_1, \dots, B_n$  with a number of instances  $m_1, \dots, m_n$  respectively. Our goal then is to obtain an *encoder*  $E : I \rightarrow \langle m_1 \rangle \times \dots \times \langle m_n \rangle$  from the image space to a cartesian product of indices. For  $E$  we take  $n$  independent Network in Network (Lin et al., 2013) classification models, where each model is trained to predict an index of the instance that was used for a particular image generation. We train these models for six-bucket RPGAN and the Anime Faces dataset on generated images. The accuracy values of classification models are provided in Table 5. As one can see, for the first two buckets, it is difficult to predict the corresponding instance indices. Meanwhile, instance indices from the latter buckets can be predicted almost perfectly. Given encoder, image editing can be performed as follows. For a real image, we obtain its *reconstruction* in  $\langle m_1 \rangle \times \dots \times \langle m_n \rangle$  with  $E$  and then “tweak” the instances from different buckets to perform semantic manipulations. Examples of reconstructions and editing are presented on Figure 18.

Bucket	1	2	3	4	5	6
Accuracy	0.05	0.21	1.0	0.99	1.0	0.69

Table 5. Instances inversion accuracies for different buckets. All buckets are consisted of 20 instances. Evaluated on 1280 generated images.



Figure 12. Images distributions for 5-bucket RPGAN and CIFAR-10 at resolution  $32 \times 32$  (Top). Random samples of the correspondent RPGAN (Bottom).



Figure 13. Images distributions for 6-bucket RPGAN and Anime Faces at resolution  $64 \times 64$  (Top). Random samples of the correspondent RPGAN (Bottom).

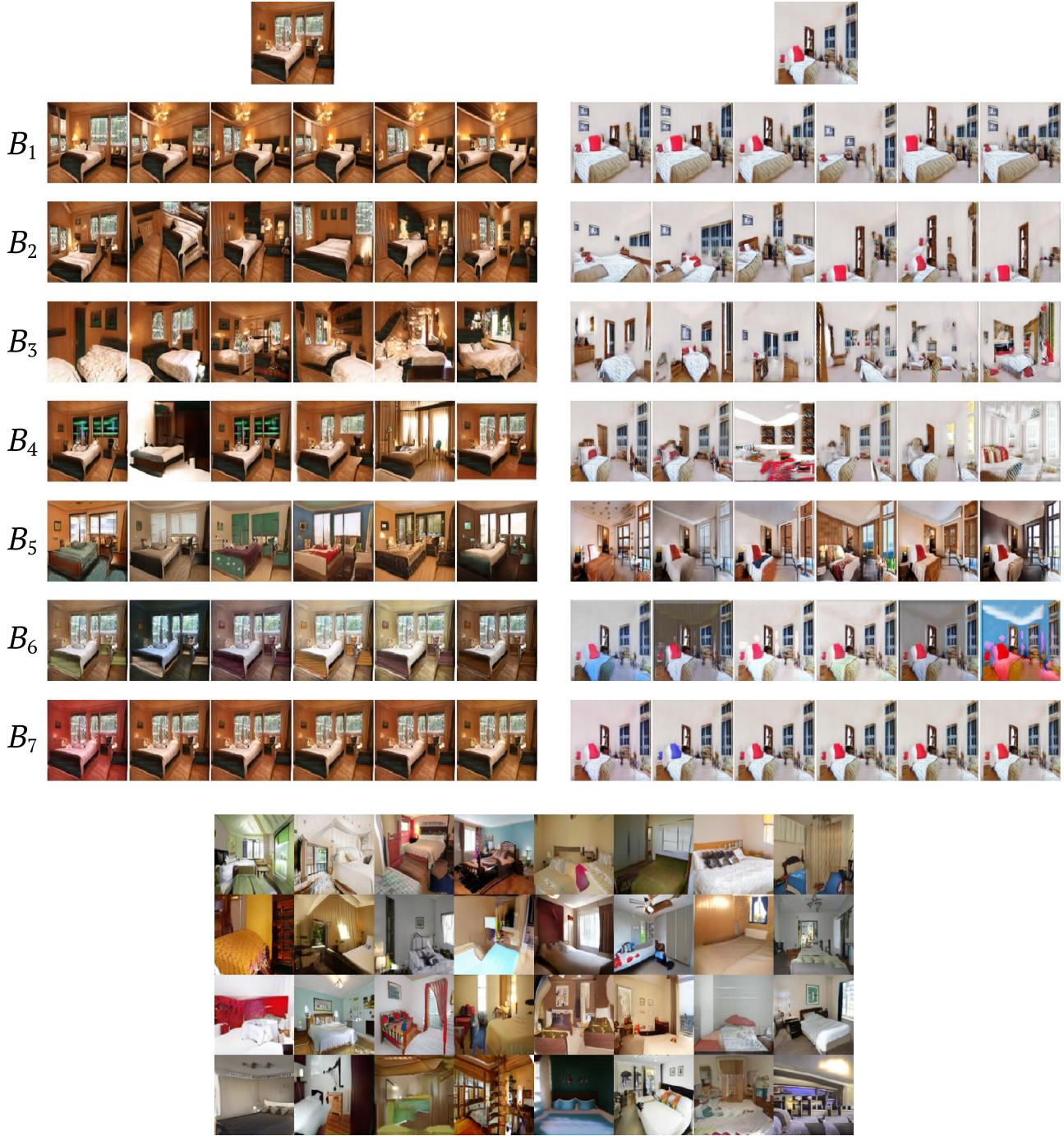


Figure 14. Images distributions for 7-bucket RPGAN and LSUN Bedroom at resolution  $128 \times 128$  (Top). Random samples of the correspondent RPGAN (Bottom).

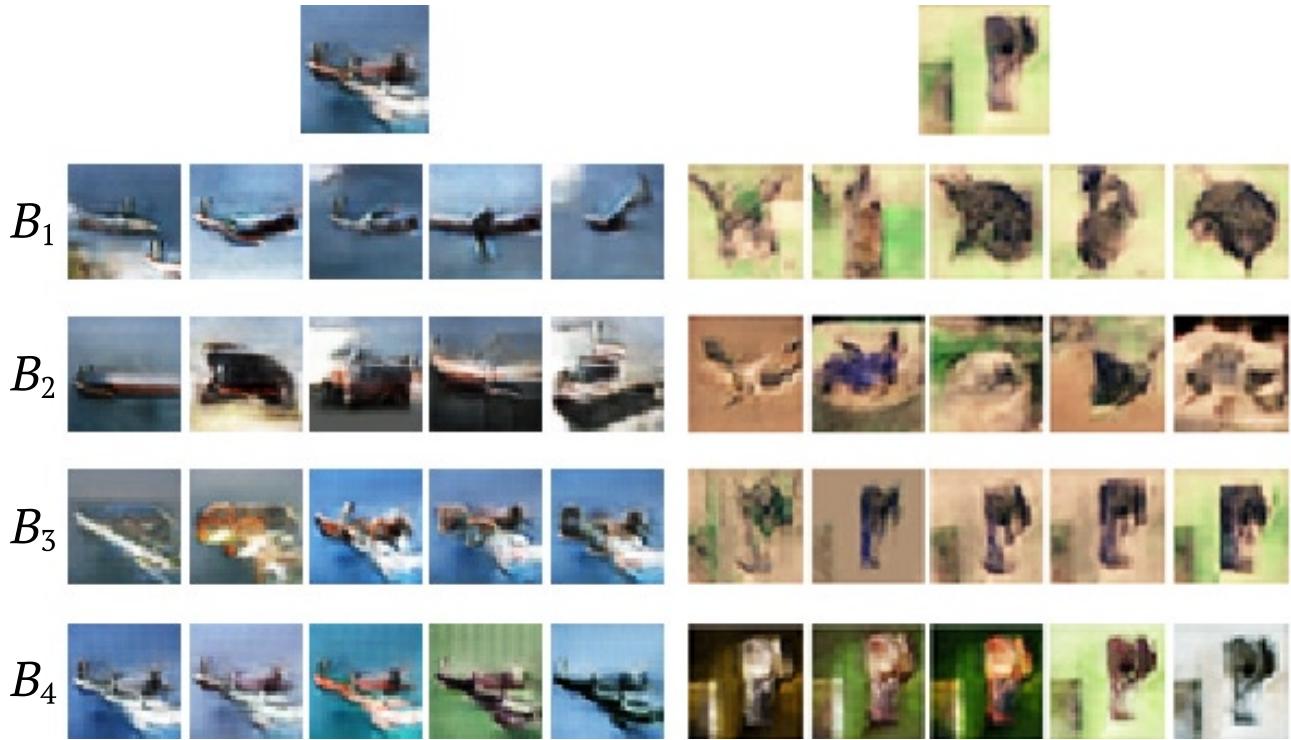


Figure 15. Images distributions for convolutional 4-bucket RPGAN and CIFAR-10 at resolution  $32 \times 32$ .

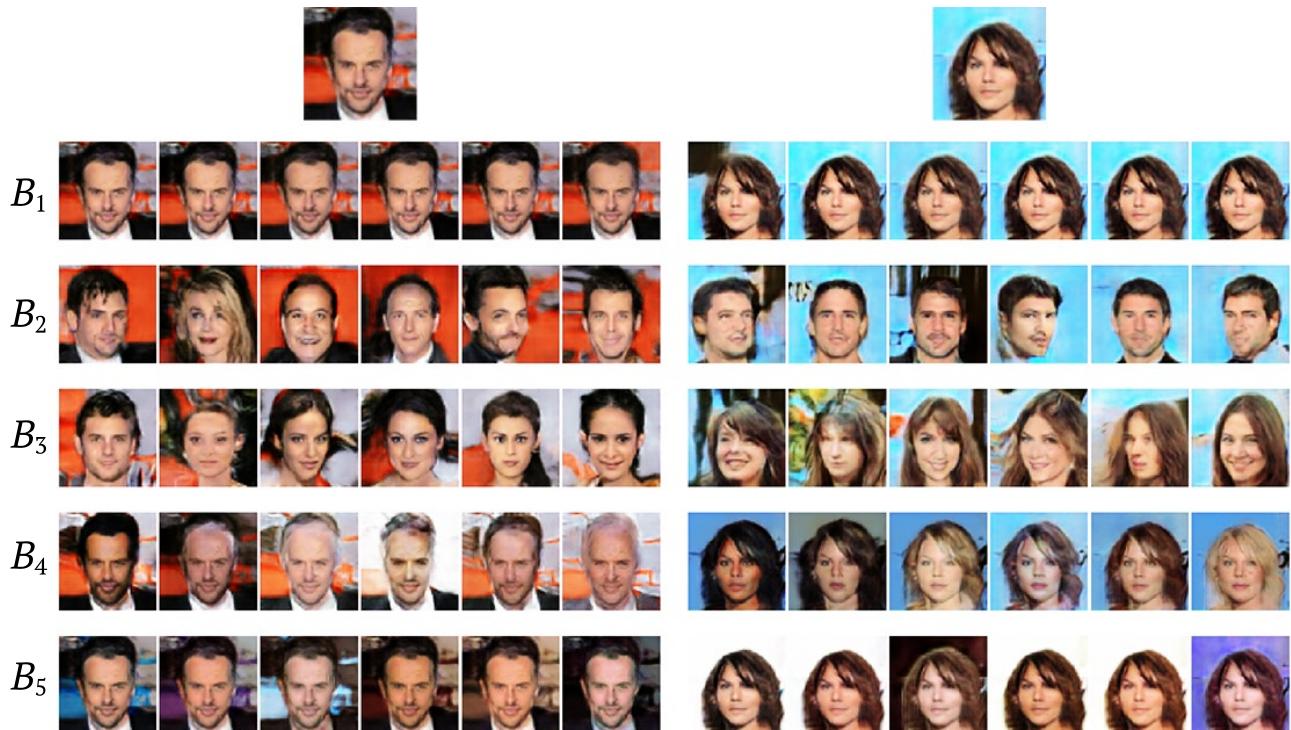


Figure 16. Images distributions for convolutional 5-bucket RPGAN and CelebA at resolution  $64 \times 64$ .

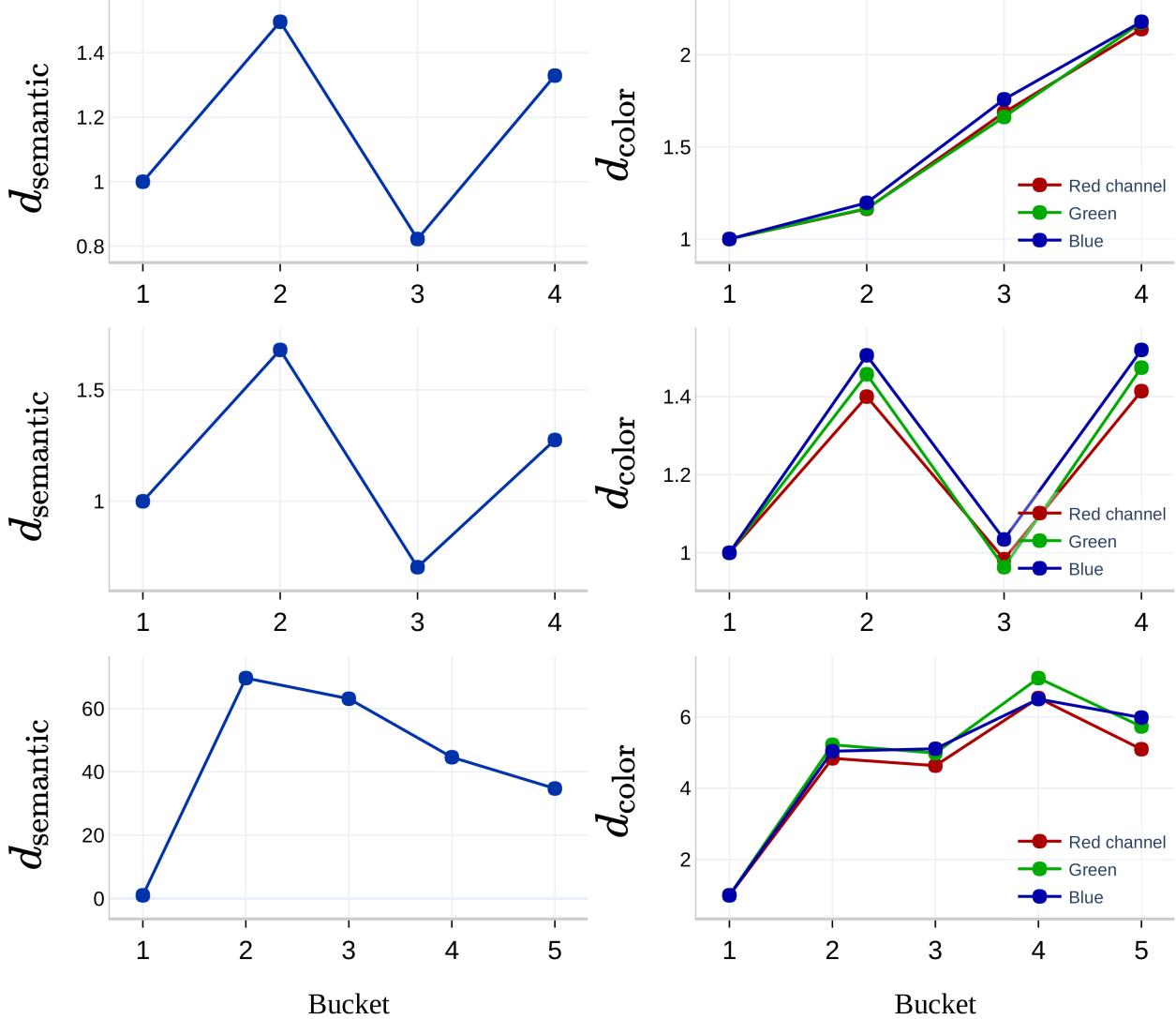


Figure 17. The quantitative evaluation of the extent  $D_{l \rightarrow 1, d}$  to which different generator parts affect different factors of variation for (Top) the four-bucket RPGAN and colored MNIST; (Middle) the four-bucket RPGAN and CIFAR-10; (Bottom) the five-bucket RPGAN and CelebA. In all cases, DCGAN-like generator architectures were used and each RPGAN bucket is associated with a convolutional layer.



Figure 18. Reconstructions of the real data images samples. *First column:* real data sample; *Second column:* its reconstruction with the invertor  $E$ ; *Other images in lines:* the images generated by the RPGAN by replacing a fixed bucket reconstructed index. For the first two images, we modify the bucket responsible for semantics, while for the last two we modify the bucket responsible for coloring.