LR-GAN: Layered Recursive Generative Adversarial Networks for Image Generation

Jianwei Yang[*]
Virginia Tech
Blacksburg, VA
jw2yang@vt.edu

Anitha Kannan
Facebook AI Research
Menlo Park, CA
akannan@fb.com

Dhruv Batra[*] **and Devi Parikh**[*]
Georgia Institute of Technology
Atlanta, GA
{dbatra, parikh}@gatech.edu

Abstract

We present LR-GAN: an adversarial image generation model which takes scene structure and context into account. Unlike previous generative adversarial networks (GANs), the proposed GAN learns to generate image background and foregrounds separately and recursively, and stitch the foregrounds on the background in a contextually relevant manner to produce a complete natural image. For each foreground, the model learns to generate its appearance, shape and pose. The whole model is unsupervised, and is trained in an end-to-end manner with gradient descent methods. The experiments demonstrate that LR-GAN can generate more natural images with objects that are more human recognizable than DCGAN.

1 Introduction

Generative adversarial networks (GANs) (Goodfellow et al., 2014) have shown significant promise as generative models for natural images. A flurry of recent work has proposed improvements over the original GAN work for image generation (Radford et al., 2015; Denton et al., 2015; Salimans et al., 2016; Chen et al., 2016; Zhu et al., 2016; Zhao et al., 2016), multi-stage image generation including part-based models (Im et al., 2016; Kwak & Zhang, 2016), image generation conditioned on input text or attributes (Mansimov et al., 2015; Reed et al., 2016b;a), image generation based on 3D structure (Wang & Gupta, 2016), and even video generation (Vondrick et al., 2016).

While the holistic 'gist' of images generated by these approaches is beginning to look natural, there is clearly a long way to go. For instance, the foreground objects in these images tend to be deformed, blended into the background, and not look realistic or recognizable.

One fundamental limitation of these methods is that they attempt to generate images without taking into account that images are 2D projections of a 3D visual world, which has a lot of structures in it. This manifests as structure in the 2D images that capture this world. One example of this structure is that images tend to have a background, and foreground objects are placed in this background in contextually relevant ways.

We develop a GAN model that explicitly encodes this structure. Our proposed model generates images in a recursive fashion: it first generates a background, and then conditioned on the background generates a foreground along with a shape (mask) and a pose (affine transformation) that together define how the background and foreground should be composed to obtain a complete image. Conditioned on this composite image, a second foreground and an associated shape and pose are generated, and so on. As a byproduct in the course of recursive image generation, our approach generates some object-shape foreground-background masks in a completely unsupervised way, without access to *any* object masks for training. Note that decomposing a scene into foreground-background layers is a classical ill-posed problem in computer vision. By explicitly factorizing appearance and transformation, LR-GAN encodes natural priors about the images that the same foreground can be 'pasted' to the different backgrounds, under different affine transformations. According to the experiments, the absence of these priors result in degenerate foreground-background decompositions, and thus also degenerate final composite images.

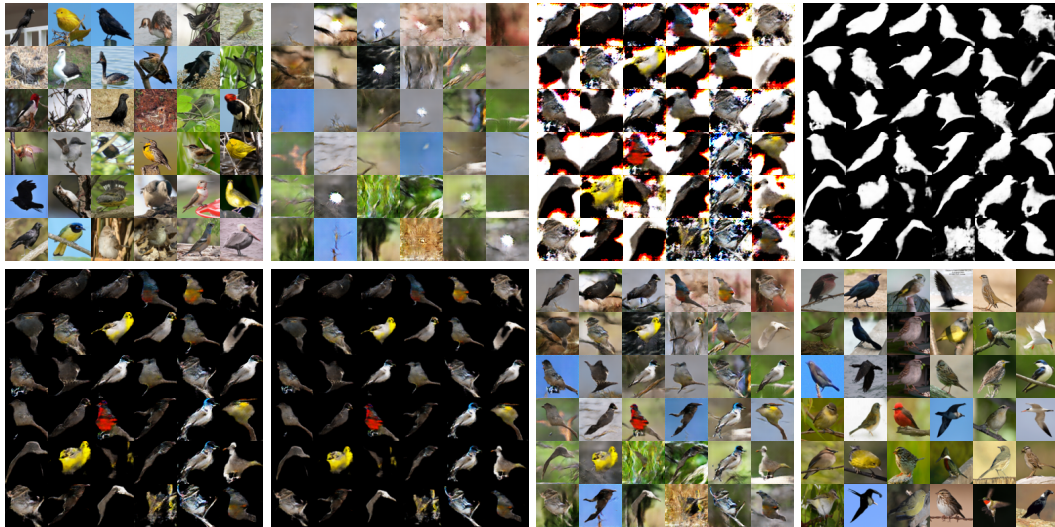[*]Work was done while visiting Facebook AI Research.

Figure 1: Generation results of our model on CUB-200 (Welinder et al., 2010). It generates images in two timesteps. At the first timestep, it generates background images, while generates foreground images, masks and transformations at the second timestep. Then, they are composed to obtain the final images. From top left to bottom right (row major), the blocks are real images, generated background images, foreground images, foreground masks, carved foreground images, carved and transformed foreground images, final composite images, and their nearest neighbor real images in the training set. Note that the model is trained in a completely unsupervised manner.

We mainly evaluate our approach on four datasets: MNIST-ONE (one digit) and MNIST-TWO (two digits) synthesized from MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky & Hinton, 2009) and CUB-200 (Welinder et al., 2010). We show qualitatively (via samples) and quantitatively (via evaluation metrics and human studies on Amazon Mechanical Turk) that LR-GAN generates images that globally look natural *and* contain clear background and object structures in them that are realistic and recognizable by humans as semantic entities. An experimental snapshot on CUB-200 is shown in Fig. 1. We also find that LR-GAN generates foreground objects that are contextually relevant to the backgrounds (e.g., horses on grass, airplanes in skies, ships in water, cars on streets, etc.). For quantitative comparison, besides existing metrics in the literature, we propose two new quantitative metrics to evaluate the quality of generated images. The proposed metrics are derived from the sufficient conditions for the closeness between generated image distribution and real image distribution, and thus supplement existing metrics.

2 RELATED WORK

Early work in parametric texture synthesis was based on a set of hand-crafted features (Portilla & Simoncelli, 2000). Recent improvements in image generation using deep neural networks mainly fall into one of the two stochastic models: variational autoencoders (VAEs) (Kingma et al., 2016) and generative adversarial networks (GANs) (Goodfellow et al., 2014). VAEs pair a top-down probabilistic generative network with a bottom up recognition network for amortized probabilistic inference. Two networks are jointly trained to maximize a variational lower bound on the data likelihood. GANs consist of a generator and a discriminator in a minmax game with the generator aiming to fool the discriminator with its samples with the latter aiming to not get fooled.

Sequential models have been pivotal for improved image generation using variational autoencoders: DRAW (Gregor et al., 2015) uses attention based recurrence conditioning on the canvas drawn so far. In Eslami et al. (2016), a recurrent generative model that draws one object at a time to the canvas was used as the decoder in VAE. These methods are yet to show scalability to natural images. Early compelling results using GANs used sequential coarse-to-fine multiscale generation and class-conditioning (Denton et al., 2015). Since then, improved training schemes (Salimans et al., 2016) and better convolutional structure (Radford et al., 2015) have improved the generation results using

GANs. PixelRNN (van den Oord et al., 2016) is also recently proposed to sequentially generates a pixel at a time, along the two spatial dimensions.

In this paper, we combine the merits of sequential generation with the flexibility of GANs. Our model for sequential generation imbibes a recursive structure that more naturally mimics image composition by inferring three components: appearance, shape, and pose. One closely related work combining recursive structure with GAN is that of Im et al. (2016) but it does not explicitly model object composition and follows a similar paradigm as by Gregor et al. (2015). Another closely related work is that of Kwak & Zhang (2016). It combines recursive structure and alpha blending. However, our work differs in three main ways: (1) We explicitly use a generator for modeling the foreground poses. That provides significant advantage for natural images, as shown by our ablation studies; (2) Our shape generator is separate from the appearance generator. This factored representation allows more flexibility in the generated scenes; (3) Our recursive framework generates subsequent objects conditioned on the current and previous hidden vectors, *and* previously generated object. This allows for explicit contextual modeling among generated elements in the scene. See Fig. 17 for contextually relevant foregrounds generated for the same background, or Fig. 6 for meaningful placement of two MNIST digits relative to each.

Models that provide supervision to image generation using conditioning variables have also been proposed: Style/Structure GANs (Wang & Gupta, 2016) learns separate generative models for style and structure that are then composed to obtain final images. In Reed et al. (2016a), GAN based image generation is conditioned on text and the region in the image where the text manifests, specified during training via keypoints or bounding boxes. While not the focus of our work, the model proposed in this paper can be easily extended to take into account these forms of supervision.

3 PRELIMINARIES

3.1 GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs) consist of a generator G and a discriminator D that are simultaneously trained with competing goals: The generator G is trained to generate samples that can 'fool' a discriminator D, while the discriminator is trained to classify its inputs as either real (coming from the training dataset) or fake (coming from the samples of G). This competition leads to a minmax formulation with a value function:

$$\min_{\theta_G} \max_{\theta_D} \Big(\mathrm{E}_{\boldsymbol{x} \sim p_{data}(x)}[\log(D(\boldsymbol{x}; \theta_D))] + \mathrm{E}_{\boldsymbol{z} \sim p_z(z)}[\log(1 - D(G(\boldsymbol{z}; \theta_G); \theta_D))] \Big), \qquad (1)$$

where $\boldsymbol{z}$ is a random vector from a standard multivariate Gaussian or a uniform distribution $p_z(z)$, $G(\boldsymbol{z}; \theta_G)$ maps $\boldsymbol{z}$ to the data space, $D(\boldsymbol{x})$ is the probability that $\boldsymbol{x}$ is real estimated by D. The advantage of the GANs formulation is that it lacks an explicit loss function and instead uses the discriminator to optimize the generative model. The discriminator, in turn, only cares whether the sample it receives is on the data manifold, and not whether it exactly matches a particular training example (as opposed to losses such as MSE). Hence, the discriminator provides a gradient signal *only* when the generated samples do not lie on the data manifold so that the generator can readjust its parameters accordingly. This form of training enables learning the data manifold of the training set and not just optimizing to reconstruct the dataset, as in autoencoder and its variants.

While the GANs framework is largely agnostic to the choice of G and D, it is clear that generative models with the 'right' inductive biases will be more effective in learning from the gradient information (Denton et al., 2015; Im et al., 2016; Gregor et al., 2015; Reed et al., 2016a; Yan et al., 2015). With this motivation, we propose a generator that models image generation via a recurrent process – in each time step of the recurrence, an object with its own appearance and shape is generated and warped according to a generated pose to compose an image in layers.

3.2 LAYERED STRUCTURE OF IMAGE

An image taken of our 3D world typically contains a layered structure. One way of representing an image layer is by its appearance and shape. As an example, an image $\boldsymbol{x}$ with two layers, foreground $\boldsymbol{f}$ and background $\boldsymbol{b}$ may be factorized as:

$$\boldsymbol{x} = \boldsymbol{f} \odot \boldsymbol{m} + \boldsymbol{b} \odot (1 - \boldsymbol{m}), \qquad (2)$$

where $\boldsymbol{m}$ is the mask depicting the shapes of image layers, and $\odot$ the element wise multiplication operator. Some existing methods assume the access to the shape of the object either during training (Isola & Liu, 2013) or *both* at train and test time (Reed et al., 2016a; Yan et al., 2015). Representing images in layered structure is even straightforward for video with moving objects (Darrell & Pentland, 1991; Wang & Adelson, 1994; Kannan et al., 2005). Vondrick et al. (2016) generates videos by separately generating a fixed background and moving foregrounds. A similar way of generating single image can be found in Kwak & Zhang (2016).

Another way is modeling the layered structure with object appearance and pose as:

$$\boldsymbol{x} = ST(\boldsymbol{f}, \boldsymbol{a}) + \boldsymbol{b}, \tag{3}$$

where $\boldsymbol{f}$ and $\boldsymbol{b}$ are foreground and background, respectively; $\boldsymbol{a}$ is the affine transformation; ST is the spatial transformation operator. Several works fall into this group (Roux et al., 2011; Huang & Murphy, 2015; Eslami et al., 2016). In Huang & Murphy (2015), images are decomposed into layers of objects with specific poses in a variational autoencoder framework, while the number of objects (i.e., layers) is adaptively estimated in Eslami et al. (2016).

To contrast with these works, LR-GAN uses a layered composition, and the foreground layers simultaneously model all three dominant factors of variation: appearance $\boldsymbol{f}$, shape $\boldsymbol{m}$ and pose $\boldsymbol{a}$. We will elaborate it in the following section.

4 LAYERED RECURSIVE GAN (LR-GAN)

The basic structure of LR-GAN is similar to GAN: it consists of a discriminator and a generator that are simultaneously trained using the minmax formulation of GAN, as described in §.3.1. The key innovation of our work is the layered recursive generator, which is what we describe in this section.

The generator in LR-GAN is recursive in that the image is constructed recursively using a recurrent network. Layered in that each recursive step composes an object layer that is 'pasted' on the image generated so far. Object layer at timestep t is parameterized by the following three constituents – 'canonical' appearance $\boldsymbol{f}_t$, shape (or mask) $\boldsymbol{m}_t$, and pose (or affine transformation) $\boldsymbol{a}_t$ for warping the object before pasting in the image composition.

Fig. 2 shows the architecture of the LR-GAN with the generator architecture unrolled for generating background $\boldsymbol{x}_0$ ($\doteq \boldsymbol{x}_b$) and foreground $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. At each time step t, the generator composes the next image $\boldsymbol{x}_t$ via the following recursive computation:

$$\boldsymbol{x}_t = \underbrace{ST(\boldsymbol{m}_t, \boldsymbol{a}_t)}_{\text{affine transformed mask}} \odot \underbrace{ST(\boldsymbol{f}_t, \boldsymbol{a}_t)}_{\text{affine transformed appearance}} + \underbrace{(1 - ST(\boldsymbol{m}_t, \boldsymbol{a}_t)) \odot \boldsymbol{x}_{t-1}}_{\text{pasting on image composed so far}}, \quad \forall t \in [1, T] \tag{4}$$

where $ST(\diamond, \boldsymbol{a}_t)$ is a spatial transformation operator that outputs the affine transformed version of $\diamond$ with $\boldsymbol{a}_t$ indicating parameters of the affine transformation.

Since our proposed model has an explicit transformation variable $\boldsymbol{a}_t$ that is used to warp the object, it can learn a canonical object representation that can be re-used to generate scenes where the object occurs as mere transformations of it, such as different scales or rotations. By factorizing the appearance, shape and pose, the object generator can focus on separately capturing regularities in these three factors that constitute an object. We will demonstrate in our experiments that removing these factorizations from the model leads to its spending capacity in variability that may not solely be about the object in Section 5.5 and 5.6.

4.1 DETAILS OF GENERATOR ARCHITECTURE

Fig. 2 shows our LR-GAN architecture in detail – we use different shapes to indicate different kinds of layers (convolutional, fractional convolutional, (non)linear, etc), as indicated by the legend. Our model consists of two main pieces – a background generator G_b and a foreground generator G_f. G_b and G_f do not share parameters with each other. G_b computation happens only once, while G_f is recurrent over time, i.e., all object generators share the same parameters. In the following, we will introduce each module and connections between them.

Temporal Connections. LR-GAN has two kinds of temporal connections – informally speaking, one on 'top' and one on 'bottom'. The 'top' connections perform the act of sequentially 'pasting'
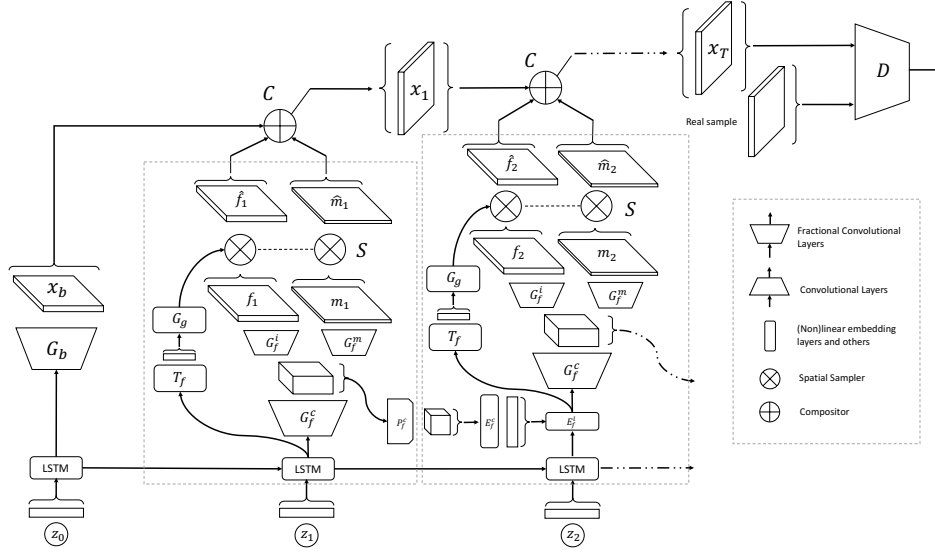
Figure 2: LR-GAN architecture unfolded to three timesteps. It mainly consists of one background generator, one foreground generator, temporal connections and one discriminator. The meaning of each component is explained in the legend.

object layers (Eqn. 4). The 'bottom' connections are constructed by a LSTM on the noise vectors z_0, z_1, z_2. Intuitively, this noise-vector-LSTM provides information to the foreground generator about what else has been generated in past. Besides, when generating multiple objects, we use a pooling layer P_f^c and a fully-connected layer E_f^c to extract the information from previous generated object response map. By this way, the model is able to 'see' previously generated objects.

Background Generator. The background generator G_b is purposely kept simple. It takes the hidden state of noise-vector-LSTM h_l^0 as the input and passes it to a number of fractional convolutional layers (also called 'deconvolution' layer in some papers) to generate images at its end. The output of background generator x_b will be used as the canvas for the following generated foregrounds.

Foreground Generator. The foreground generator G_f is used to generate an object with appearance and shape. Correspondingly, G_f consists of three sub-modules, G_f^c, which is a common 'trunk' whose outputs are shared by G_f^i and G_f^m. G_f^i is used to generate the foreground appearance f_t, while G_f^m generates the mask m_t for the foreground. All three sub-modules consists of one or more fractional convolutional layers combined with batch-normalization and nonlinear layers. The generated foreground appearance and mask have the same spatial size as the background. The top of G_f^m is a sigmoid layer in order to generate one channel mask whose values range in $(0, 1)$.

Spatial Transformer. To spatially transform foreground objects, we need to estimate the transformation matrix. As in Jaderberg et al. (2015), we predict the affine transformation matrix with a linear layer T_f that has six-dimensional outputs. Then based on the predicted transformation matrix, we use a grid generator G_g to generate the corresponding sampling coordinates in the input for each location at the output. The generated foreground appearance and mask share the same transformation matrix, and thus the same sampling grid. Given the grid, the sampler S will simultaneously sample the f_t and m_t to obtain $\hat{f}_t$ and $\hat{m}_t$, respectively. Different from Jaderberg et al. (2015), our sampler here normally performs downsampling, since the the foreground typically has smaller size than the background. Pixels in $\hat{f}_t$ and $\hat{m}_t$ that are from outside the extent of f_t and m_t are set to zero. Finally, $\hat{f}_t$ and $\hat{m}_t$ are sent to the compositor C which combines the canvas x_{t-1} and $\hat{f}_t$ through layered composition with blending weights given by $\hat{m}_t$ (Eqn. 4).

Pseudo-code for our approach and detailed model configuration are provided in the Appendix.

4.2 New Evaluation Metrics

Several metrics have been proposed to evaluate GANs, such as Gaussian parzen window (Good-fellow et al., 2014), Generative Adversarial Metric (GAM) (Im et al., 2016) and Inception Score (Salimans et al., 2016). The common goal is to measure the similarity between the generated data distribution $P_g(\boldsymbol{x}) = G(\boldsymbol{z}; \theta_z)$ and the real data distribution $P(\boldsymbol{x})$. Most recently, Inception Score has been used in several works (Salimans et al., 2016; Zhao et al., 2016). However, it is an assymetric metric and could be easily fooled by generating centers of data modes.

In addition to these metrics, we present two new metrics based on the following intuition – a sufficient (but not necessary) condition for closeness of $P_g(\boldsymbol{x})$ and $P(\boldsymbol{x})$ is closeness of $P_g(\boldsymbol{x}|y)$ and $P(\boldsymbol{x}|y)$, i.e., distributions of generated data and real data conditioned on all possible variables of interest y, e.g., category label. One way to obtain this variable of interest y is via human annotation. Specifically, given the data sampled from $P_g(\boldsymbol{x})$ and $P(\boldsymbol{x})$, we ask people to label the category of the samples according to some rules. Note that such human annotation is often easier than comparing samples from the two distributions (e.g., because there is no 1:1 correspondence between samples to conduct forced-choice tests).

After the annotations, we need to verify whether the two distributions are similar in each category. Clearly, directly comparing the distributions $P_g(\boldsymbol{x}|y)$ and $P(\boldsymbol{x}|y)$ may be as difficult as comparing $P_g(\boldsymbol{x})$ and $P(\boldsymbol{x})$. Fortunately, we can use Bayes rule and alternatively compare $P_g(y|\boldsymbol{x})$ and $P(y|\boldsymbol{x})$, which is a much easier task. In this case, we can simply train a discriminative model on the samples from $P_g(\boldsymbol{x})$ and $P(\boldsymbol{x})$ together with the human annotations about categories of these samples. With a slight abuse of notation, we use $P_g(y|\boldsymbol{x})$ and $P(y|\boldsymbol{x})$ to denote probability outputs from these two classifiers (trained on generated samples vs trained on real samples). We can then use these two classifiers to compute the following two evaluation metrics:

Adversarial Accuracy: Computes the classification accuracies achieved by these two classifiers on a validation set, which can be the training set or another set of real images sampled from $P(\boldsymbol{x})$. If $P_g(\boldsymbol{x})$ is close to $P(\boldsymbol{x})$, we expect to see similar accuracies.

Adversarial Divergence: Computes the KL divergence between $P_g(y|\boldsymbol{x})$ and $P(y|\boldsymbol{x})$. The lower the adversarial divergence, the closer two distributions are. The low bound for this metric is exactly zero, which means $P_g(y|\boldsymbol{x}) = P(y|\boldsymbol{x})$ for all samples in the validation set.

As discussed above, we need human efforts to label the real and generated samples. Fortunately, we can further simplify this. Based on the labels given on training data, we split the training data into categories, and train one generator for each category. With all these generators, we generate samples of all categories. This strategy will be used in our experiments on the datasets with labels given.

5 Experiment

We conduct qualitative and quantitative evaluations on three datasets: 1) MNIST (LeCun et al., 1998); 2) CIFAR-10 (Krizhevsky & Hinton, 2009); 3) CUB-200 (Welinder et al., 2010). To add variability to the MNIST images, we randomly scale (factor of 0.8 to 1.2) and rotate ($-\frac{\pi}{4}$ to $\frac{\pi}{4}$) the digits and then stitch them to 48×48 uniform backgrounds with random grayscale value between [0, 200]. Images are then rescaled back to 32×32. Each image thus has a different background grayscale value and a different transformed digit as foreground. We rename this sythensized dataset as **MNIST-ONE** (single digit on a gray background). We also synthesize a dataset **MNIST-TWO** containing two digits on a grayscale background. We randomly select two images of digits and perform similar transformations as described above, and put one on the left and the other on the right side of a 78×78 gray background. We resize the whole image to 64×64.

We develop LR-GAN based on open source code[1]. We assume the number of objects is known. Therefore, for MNIST-ONE, MNIST-TWO, CIFAR-10, and CUB-200, our model has two, three, two, and two timesteps, respectively. Since the size of foreground object should be smaller than that of canvas, we set the minimal allowed scale [2] in affine transforamtion to be 1.2 for all datasets except for MNIST-TWO, which is set to 2 (objects are smaller in MNIST-TWO). In LR-GAN, the

---

[1] https://github.com/soumith/dcgan.torch

[2] Scale corresponds to the size of the target canvas with respect to the object – the larger the scale, the larger the canvas, and the smaller the relative size of the object in the canvas. 1 means the same size as the canvas.
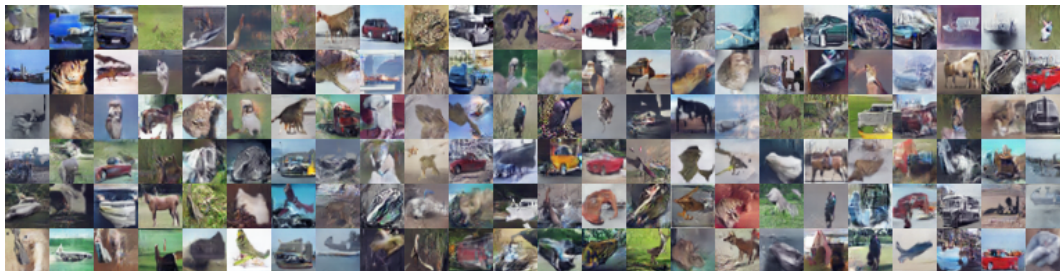
Figure 3: Generated images on CIFAR-10 based on our model.

Figure 4: Generated images on CUB-200 based on our model.

background generator and foreground generator have similar architectures. One difference is that the number of channels in the background generator is half of the one in the foreground generator. We compare our results to that of DCGAN (Radford et al., 2015). Note that LR-GAN without LSTM at the first timestep corresponds exactly to the DCGAN. This allows us to run controlled experiments. In both generator and discriminator, all the (fractional) convolutional layers have 4×4 filter size with stride 2. As a result, the number of layers in the generator and discriminator automatically adapt to the size of training images. Please see the Appendix (Section 6.2) for details about the configurations. We use three metrics for quantitative evaluation, including Inception Score (Salimans et al., 2016) and the proposed Adversarial Accuracy, Adversarial Divergence. Note that we report two versions of Inception Score. One is based on the pre-trained Inception net, and the other one is based on the pre-trained classifier on the target datasets.

5.1 QUALITATIVE RESULTS

In Fig. 3 and 4, we show the generated samples for CIFAR-10 and CUB-200, respectively. MNIST results are shown in the next subsection. As we can see from the images, the compositional nature of our model results in the images being free of blending artifacts between backgrounds and foregrounds. For CIFAR-10, we can see the horses and cars with clear shapes. For CUB-200, the bird shapes tend to be even sharper.

5.2 MNIST-ONE AND MNIST-TWO

We now report the results on MNIST-ONE and MNIST-TWO. Fig. 5 shows the generation results of our model on MNIST-ONE. As we can see, our model generates the background and the foreground in separate timestep, and can disentangle the foreground digits from background nearly perfectly. Though initial values of the mask randomly distribute in the range of (0, 1), after training, the masks are nearly binary and accurately carve out the digits from the generated foreground. More results on MNIST-ONE (including human studies) can be found in the Appendix (Section 6.3).

Fig. 6 shows the generation results for MNIST-TWO. Similarly, the model is also able to generate background and the two foreground objects separately. The foreground generator tends to generate a single digit at each timestep. Meanwhile, it captures the context information from the previous time steps. When the first digit is placed to the left side, the second one tends to be placed on the right side, and vice versa.
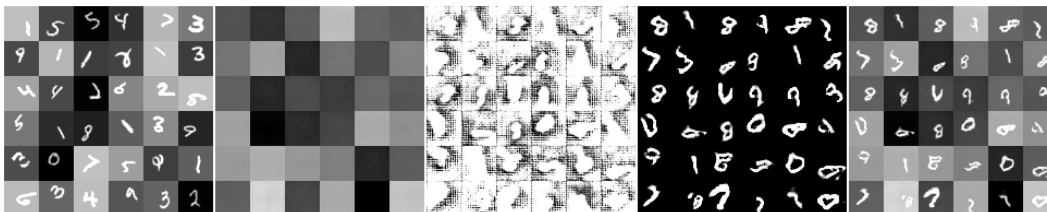
Figure 5: Generation results of our model on MNIST-ONE. From left to right, the image blocks are real images, generated background images, generated foreground images, generated masks and final composite images, respectively.
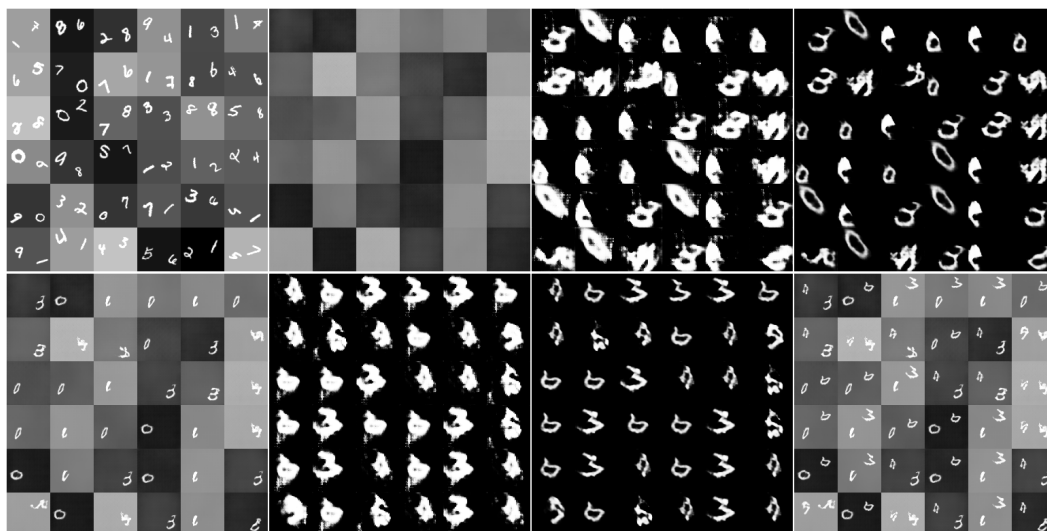

Figure 6: Generation results of our model on MNIST-TWO. From top left to bottom right (row major), the image blocks are real images, generated background images, foreground images and masks at the second timestep, composite images at the second time step, generated foreground images and masks at the third timestep and the final composite images, respectively.

5.3 CUB-200

We study the effectiveness of our model trained on the CUB-200 bird dataset. In Fig. 1, we have shown a random set of generated images, along with the intermediate generation results of the model. While being *completely unsupervised*, the model, for a large fraction of the samples, is able to

Figure 7: Matched pairs of generated images based on DCGAN and LR-GAN, respectively. The odd columns are generated by DCGAN, and the even columns are generated by LR-GAN. These are paired according to the perfect matching based on Hungarian algorithm.
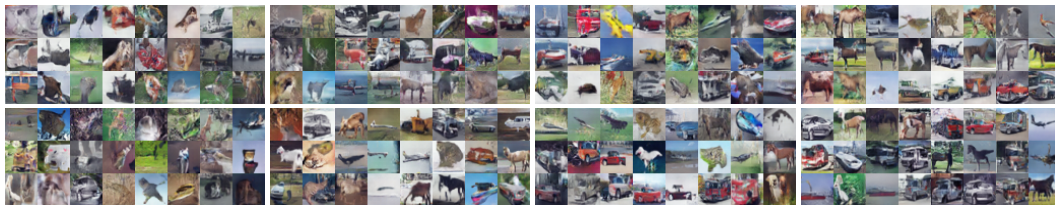
Figure 8: Qualitative comparison on CIFAR-10. Top three rows are images generated by DCGAN; Bottom three rows are by LR-GAN. From left to right, the blocks display generated images with increasing quality level as determined by human studies.

successfully disentangle the foreground and the background. This is evident from the generated bird-like masks.

We do a comparative study based on Amazon Mechanical Turk (AMT) between DCGAN and LR-GAN to quantify relative visual quality of the generated images. We first generated 1000 samples from both the models. Then, we performed perfect matching between the two image sets using the Hungarian algorithm on $L2$ norm distance in the pixel space. This resulted in 1000 image pairs. Some examplar pairs are shown in Fig. 7. For each image pair, 9 judges are asked to choose the one that is more realistic. Based on majority voting, we find that our generated images are selected 68.4% times, compared with 31.6% times for DCGAN. This demonstrates that our model has generated more realistic images than DCGAN. We can attribute this difference to our model's ability to generate foreground separately from the background, enabling stronger edge cues.

5.4 CIFAR-10

We now qualitatively and quantitatively evaluate our model on CIFAR-10, which contains multiple object categories and also various backgrounds.

Comparison of image generation quality: We conduct AMT studies to compare the fidelity of image generation. Towards this goal, we generate 1000 images from DCGAN and LR-GAN, respectively. We ask 5 judges to label each image to either belong to one of the 10 categories or as 'non recognizable' or 'recognizable but not belonging to the listed categories'. We then assign each image a quality level between [0,5] that captures the number of judges that agree with the majority choice. Fig. 8 shows the images generated by both approaches, ordered by increasing quality level. We merge images at quality level 0 (all judges said non-recognizable) and 1 together, and similarly images at level 4 and 5. Visually, the generated samples by our model have clearer boundaries and object structures. We also computed the fraction of non-recognizable images: Our model had a 10% absolute drop in non-recognizability rate (67.3% for ours vs. 77.7% for DCGAN). For reference, 11.4% of real CIFAR images were categorized as non-recognizable. Fig. 9 shows more generated (intermediate) results of our model.

Quantitative evaluation on generators: We evaluate the generators based on three metrics: 1) Inception Score; 2) Adversarial Accuracy; 3) Adversarial Divergence. To obtain a classifier model for evaluation, we remove the top layer in the discriminator used in our model, and then append two fully connected layers on the top of it. We train this classifier using the training samples of CIFAR-10 based on the annotations. Following Salimans et al. (2016), we generated 50,000 images

Table 1: Quantitative comparison between DCGAN and LR-GAN on CIFAR-10.

Training Data	Real Images	DCGAN	Ours
Inception Score[†]	11.18±0.18	6.64±0.14	7.17±0.07
Inception Score[††]	7.23±0.09	5.69±0.07	6.11±0.06
Adversarial Accuracy	83.33±0.08	37.81±0.02	44.22 ±0.08
Adversarial Divergence	0	7.58±0.04	5.57±0.06

[†]Evaluate using the pre-trained Inception net as Salimans et al. (2016)
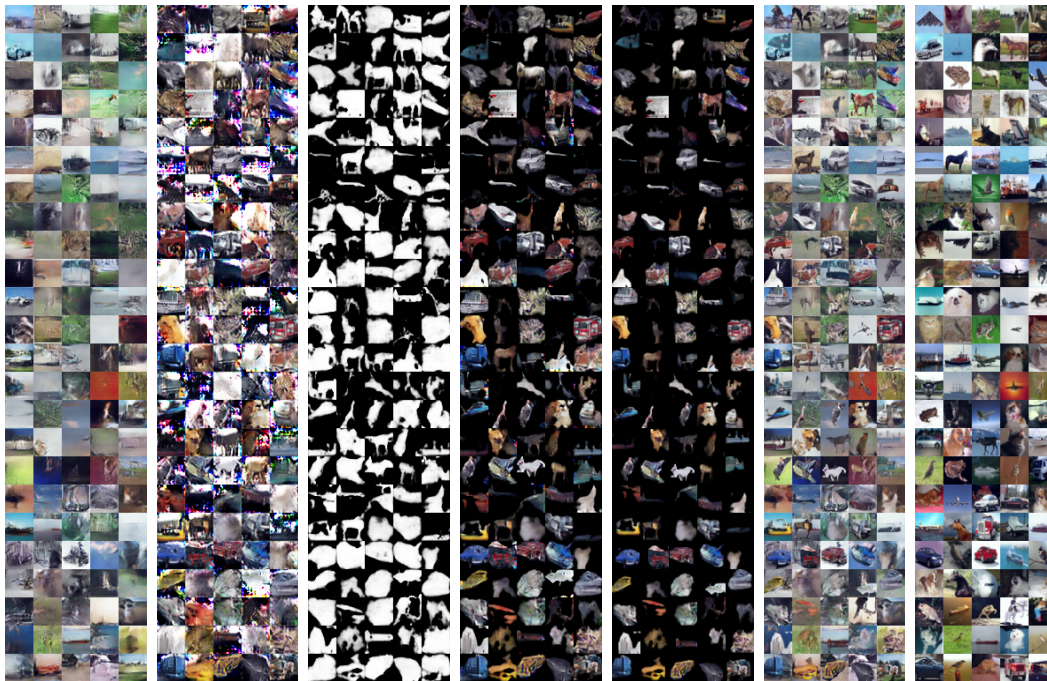[††]Evaluate using the supervisedly trained classifier based on the discriminator in LR-GAN.

Figure 9: Generation results of our model on CIFAR-10. From left to right, the blocks are: generated background images, foreground images, foreground masks, foreground images carved out by masks, carved foregrounds after spatial transformation, final composite images and nearest neighbor training images to the generated images.

Figure 10: Category specific generation results of our model on CIFAR-10 categories of horse, frog, and cat (top to bottom). The blocks from left to right are: generated background images, foreground images, foreground masks, foreground images carved out by masks, carved foregrounds after spatial transformation and final composite images.

based on DCGAN and LR-GAN, repsectively. We compute two types of Inception Scores. The standard Inception Score is based on the Inception net as in Salimans et al. (2016), and the contextual Inception Score is based on our trained classifier model. To distinguish, we denote the standard one as 'Inception Score[†]', and the contextual one as 'Inception Score[††]'. To obtain the Adversarial Accuracy and Adversarial Divergence scores, we train one generator on each of 10 categories for DCGAN and LR-GAN, respectively. Then, we use these generators to generate samples of different categories. Given these generated samples, we train the classifiers for DCGAN and LR-GAN separately. Along with the classifier trained on the real samples, we compute the Adversarial Accuracy

and Adversarial Divergence on the real training samples. In Table 1, we report the Inception Scores, Adversarial Accuracy and Adversarial Divergence for comparison. We can see that our model outperforms DCGAN across the board. To point out, we obtan different Inception Scores based on different classifier models, which indicates that the Inception Score varies with different models.

Quantitative evaluation on discriminators: We evaluate the discriminator as an extractor for deep representations. Specifically, we use the output of the last convolutional layer in the discriminator as features. We perform a 1-NN classification on the test set given the full training set. Cosine similarity is used as the metric. On the test set, our model achieves 62.09%±0.01% compared to DCGAN's 56.05%±0.02%.

Contextual generation: We also show the efficacy of our approach to generate diverse foregrounds conditioned on fixed background. The results in Fig. 17 in Appendix showcase that the foreground generator generates objects that are compatible with the background. This indicates that the model has captured contextual dependencies between the image layers.

Category specific models: The objects in CIFAR-10 exhibit huge variability in shapes. That can partly explain why some of the generated shapes are not as compelling in Fig. 9. To test this hypothesis, we reuse the generators trained for each of 10 categories used in our metrics to obtain the generation results. Fig. 10 shows results for categories 'horse', 'frog' and 'cat'. We can see that the model is now able to generate object-specific appearances and shapes, similar in vein to our results on the CUB-200 dataset.

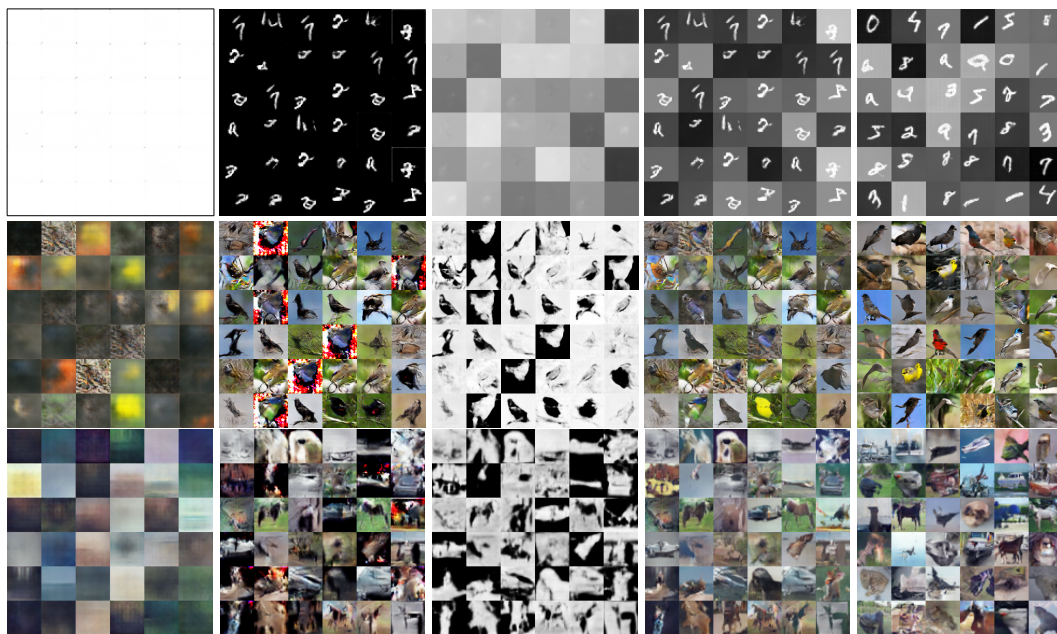## 5.5 IMPORTANCE OF TRANSFORMATIONS

Figure 11: Generation results from an ablated LR-GAN model without affine transformations. From top to bottom, the block rows correspond to different datasets: MNIST-ONE, CUB-200, CIFAR-10. From left to right, the blocks show generated background images, foreground images, foreground masks, and final composite images. For comparison, the rightmost column block shows final generated images from a non-ablated model with affine transformations.

Fig. 11 shows results from an ablated model without affine transformations in the foreground layers, and compares the results with the full model that does include these transformations. We note that one significant problem emerges that the decompositions are degenerate, in the sense that the model is unable to break the symmetry between foreground and background layers, often generating object appearances in the model's background layer and vice versa. For CUB-200, the final generated images have some blendings between foregrounds and backgrounds. This is particularly the case for
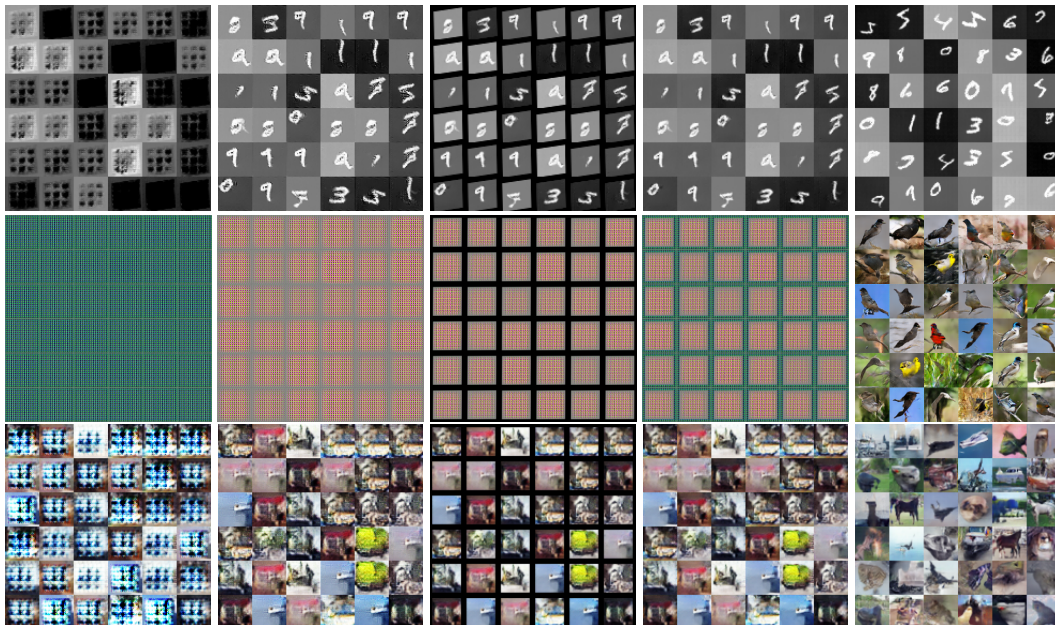
Figure 12: Generation results from an ablated LR-GAN model without mask generator. The block rows correspond to different datasets (from top to bottom: MNIST-ONE, CUB-200, CIFAR-10). From left to right, the blocks show generated background images, foreground images, transformed foreground images, and final composite images. For comparison, the rightmost column block shows final generated images from a non-ablated model with mask generator.

those images without bird-shape masks. For CIFAR-10, a number of generated masks are inverted. In this case, the background images are carved out as the foreground objects. The foreground generator takes almost all the duty to generate the final images, which make it harder to generate images as clear as the model with transformation. From these comparisons, we qualitatively demonstrate the importance of modeling transformations in the foreground generation process. Another merit of using transformation is that the intermediate outputs of the model are more interpretable and faciliate to the downstreaming tasks, such as scene paring, which is demonstrated in Section 6.8.

5.6   IMPORTANCE OF SHAPES

We perform another ablation study by removing the mask generator to understand the importance of modeling object shapes. In this case, the generated foreground is simply pasted on top of the generated background after being transformed. There is no alpha blending between the foregrounds and backgrounds. The generation results for three datasets, MNIST-ONE, CUB-200, CIFAR-10 are shown in Fig. 12. As we can see, though the model works well for the generation of MNIST-ONE, it fails to generate reasonable images across the other datasets. Particularly, the training does not even converge for CUB-200. Based on these results, we qualitatively demonstrate that mask generator in our model is fairly important to obtain plausible results, especially for realistic images.

REFERENCES

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Info-gan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016.

Trevor Darrell and Alex Pentland. Robust estimation of a multi-layered motion representation. *IEEE Workshop on Visual Motion*, 1991.

Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pp. 1486–1494, 2015.

S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, infer, repeat: Fast scene understanding with generative models. *CoRR*, abs/1603.08575, 2016.

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*, pp. 2672–2680, 2014.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.046239*, 2015.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

Jonathan Huang and Kevin Murphy. Efficient inference in occlusion-aware generative models of images. *CoRR*, abs/1511.06362, 2015.

Daniel Jiwoong Im, Chris Dongjoo Kim, Hui Jiang, and Roland Memisevic. Generating images with recurrent adversarial networks. *arXiv preprint arXiv:1602.05110*, 2016.

Phillip Isola and Ce Liu. Scene collaging: Analysis and synthesis of natural images with semantic layers. In *IEEE International Conference on Computer Vision*, pp. 3048–3055, 2013.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28*, pp. 2017–2025, 2015.

Anitha Kannan, Nebojsa Jojic, and Brendan Frey. Generative model for layers of appearance and deformation. *AISTATS*, 2005.

Diederik P Kingma, Tim Salimans, and Max Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

Hanock Kwak and Byoung-Tak Zhang. Generating images part by part with composite generative adversarial networks. *arXiv preprint arXiv:1607.05387*, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015.

Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.

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

Scott Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. *arXiv preprint arXiv:1610.02454*, 2016a.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016b.

Nicolas Le Roux, Nicolas Heess, Jamie Shotton, and John Winn. Learning a generative model of images by factoring appearance and shape. *Neural Computation*, 23:593–650, 2011.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016.

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *CoRR*, abs/1601.06759, 2016.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *arXiv preprint arXiv:1609.02612*, 2016.

John Wang and Edward Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 1994.

Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. *arXiv preprint arXiv:1603.05631*, 2016.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. *CoRR*, abs/1512.00570, 2015.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pp. 597–613. Springer, 2016.

6 APPENDIX

6.1 ALGORITHM

Algo. 1 illustrates the generative process in our model. $g(\star)$ evaluates the function g at $\star$. $\circ$ is a composition operator that composes its operands so that $f \circ g(\star) = f(g(\star))$.

Algorithm 1 Stochastic Layered Recursive Image Generation

1: $\boldsymbol{z}_0 \sim \mathcal{N}(0, I)$
2: $\boldsymbol{x}_0 = G_b(\boldsymbol{z}_0)$ ▷ background generator
3: $\boldsymbol{h}_l^0 \leftarrow \mathbf{0}$
4: $\boldsymbol{c}_l^0 \leftarrow \mathbf{0}$
5: **for** $t \in [1 \cdots T]$ **do**
6:     $\boldsymbol{z}_t \sim \mathcal{N}(0, I)$
7:     $\boldsymbol{h}_l^t, \boldsymbol{c}_l^t \leftarrow \text{LSTM}([\boldsymbol{z}_t, \boldsymbol{h}_l^{t-1}, \boldsymbol{c}_l^{t-1}])$ ▷ pass through LSTM
8:     **if** t = 1 **then**
9:         $\boldsymbol{y}_t \leftarrow \boldsymbol{h}_l^t$
10:    **else**
11:        $\boldsymbol{y}_t \leftarrow E_f^l([\boldsymbol{h}_l^t \ \boldsymbol{h}_f^{t-1}])$ ▷ pass through non-linear embedding layers E_f^l
12:    **end if**
13:    $\boldsymbol{s}_t \leftarrow G_f^c(\boldsymbol{y}_t)$ ▷ predict shared cube for G_f^i and G_f^m
14:    $a_t \leftarrow T_f(\boldsymbol{y}_t)$ ▷ object transformation
15:    $\boldsymbol{f}_t \leftarrow G_f^i(\boldsymbol{s}_t)$ ▷ generate object appearance
16:    $\boldsymbol{m}_t \leftarrow G_f^m(\boldsymbol{s}_t)$ ▷ generate object shape
17:    $\boldsymbol{h}_f^t \leftarrow E_f^c \circ P_f^c(\boldsymbol{s}_t)$ ▷ predict shared represenation embedding
18:    $\boldsymbol{x}_t \leftarrow ST(\boldsymbol{m}_t, \boldsymbol{a}_t) \odot ST(\boldsymbol{f}_t, \boldsymbol{a}_t) + (1 - ST(\boldsymbol{m}_t, \boldsymbol{a}_t)) \odot \boldsymbol{x}_{t-1}$
19: **end for**

6.2 MODEL CONFIGURATIONS

Table 2 lists the information and model configuration for different datasets. The dimensions of random vectors and hidden vectors are all set to 100. We also compare the number of parameters in DCGAN and LR-GAN. The numbers before '/' are our model, after '/' are DCGAN. Based on the same notation used in (Zhao et al., 2016), the architectures for the different datasets are:

Table 2: Information and model configurations on different datasets.

Dataset	MNIST-ONE	MNIST-TWO	CIFAR-10	CUB-200
Image Size	32	64	32	64
#Images	60,000	60,000	50,000	5,994
#Timesteps	2	3	2	2
#Parameters	5.25M/4.11M	7.53M/6.33M	5.26M/4.11M	27.3M/6.34M

- MNIST-ONE: G_b: (256)4c-(128)4c2s-(64)4c2s-(3)4c2s; G_f^c: (512)4c-(256)4c2s-(128)4c2s; G_f^i: (3)4c2s; G_f^m: (1)4c2s; D: (64)4c2s-(128)4c2s-(256)4c2s-(256)4p4s-1

- MNIST-TWO: G_b: (256)4c-(128)4c2s-(64)4c2s-(32)4c2s-(3)4c2s; G_f^c: (512)4c-(256)4c2s-(128)4c2s-(64)4c2s; G_f^i: (3)4c2s; G_f^m: (1)4c2s;D: (64)4c2s-(128)4c2s-(256)4c2s-(512)4c2s-(512)4p4s-1

- CUB-200: G_b: (512)4c-(256)4c2s-(128)4c2s-(64)4c2s-(3)4c2s; G_f^c: (1024)4c-(512)4c2s-(256)4c2s-(128)4c2s; G_f^i: (3)4c2s; G_f^m: (1)4c2s;D: (128)4c2s-(256)4c2s-(512)4c2s-(1024)4c2s-(1024)4p4s-1

- CIFAR-10: G_b: (256)4c-(128)4c2s-(64)4c2s-(3)4c2s; G_f^c: (512)4c-(256)4c2s-(128)4c2s; G_f^i: (3)4c2s; G_f^m: (1)4c2s D: (64)4c2s-(128)4c2s-(256)4c2s-(256)4p4s-1

6.3 RESULTS ON MNIST-ONE

We conduct human studies on generation results on MNIST-ONE. Specifically, we generate 1,000 images using both LR-GAN and DCGAN. As references, we also include 1000 real images. Then we ask the users on AMT to label each image to be one of the digits (0-9). We also provide them an option 'non recognizable' in case the generated image does not seem to contain a digit. Each image was judged by 5 unique workers. Similar to CIFAR-10, if an image is recognized to be the same digit by all 5 users, it is assigned to quality level 5. If it is not recognizable according to all users, it is assigned to quality level 0. Fig. 13 (left) shows the number of images assigned to all six quality levels. Compared to DCGAN, our model generated more samples with high quality levels. As expected, the real images have many samples with high quality levels. In Fig. 13 (right), we show the number of images that are recognized to each digit category (0-9). For qualitative comparison, we show examplar images at each quality level in Fig. 14. From left to right, the quality level increases from 0 to 5. As expected, the images with higher quality level are more clear.

For quantitative evaluation, we use the same way as for CIFAR-10. The classifier model used for contextual Inception Score is trained based on the training set. We generate 60,000 samples based on DCGAN and LR-GAN for evaluation, respectively. To obtain the Adversarial Accuracy and Adversarial Divergence, we first train 10 generators for 10 digit categories separately, and then use the generated samples to train the classifier. As shown in Table 3, our model has higher scores than DCGAN on both standard and contextual Inception Score. Also, our model has a slightly higher
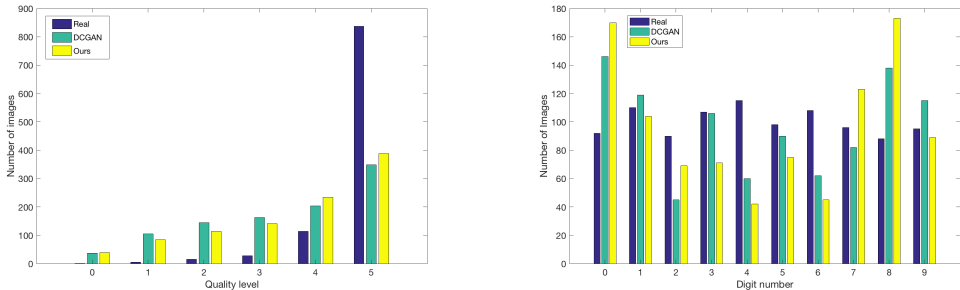

Figure 13: Statistics of annotations in human studies on MNIST-ONE. Left: distribution of quality level; Right: distribution of recognized digit categories.
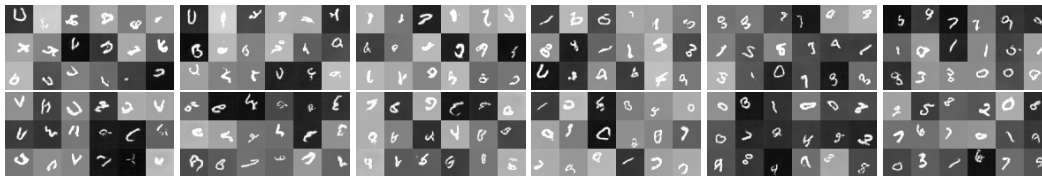
Figure 14: Qualitative comparison on MNIST-ONE. Top three rows are samples generated by DC-GAN. Bottom three rows are samples generated by LR-GAN. The quality level increases from left to right as determined via human studies.

Table 3: Quantitative comparison on MNIST-ONE.

Training Data	Real Images	DCGAN	Ours
Inception Score[†]	1.83±0.01	2.03±0.01	2.06±0.01
Inception Score[††]	9.15±0.04	6.42±0.03	7.15±0.04
Adversarial Accuracy	95.22 ± 0.25	26.12 ± 0.07	26.61 ± 0.06
Adversarial Divergence Score	0	8.47 ± 0.03	8.39 ± 0.04

[†]Evaluate using the pre-trained Inception net as Salimans et al. (2016)
[††]Evaluate using the supervisedly trained classifier based on the discriminator in LR-GAN.

adversarial accuracy, and lower adversarial divergence than DCGAN. We find that the all three image sets have low standard Inception Scores. This is mainly because the Inception net is trained on ImageNet, which has a very different data distribution from the MNIST dataset. Based on this, we argue that the standard Inception Score is not suitable for some image datasets.

6.4 MORE RESULTS ON CUB-200

In this experiment, we reduce the minimal allowed object scale to 1.1, which allows the model to generate larger foreground objects. The results are shown in Fig. 15. Similar to the results when the constraint is 1.2, the crisp bird-like masks are generated automatically by our model.
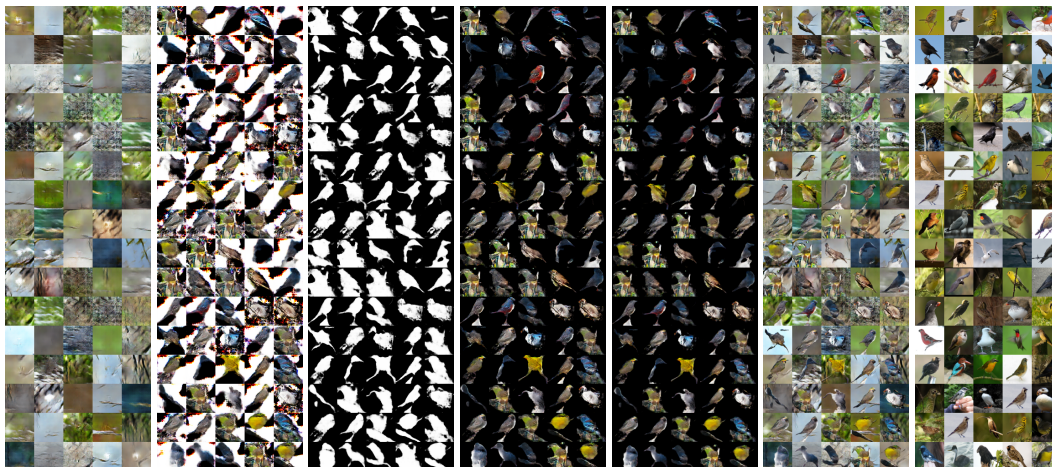

Figure 15: Generation results of our model on CUB-200 when setting minimal allowed scale to 1.1. From left to right, the blocks show the generated background images, foreground images, foreground masks, foreground images carved out by masks, carved foreground images after spatial transformation. The sixth and seventh blocks are final composite images and the nearest neighbor real images.

16

6.5 MORE RESULTS ON CIFAR-10

6.5.1 QUALITATIVE RESULTS

In Fig. 16, we show more results on CIFAR-10 when setting minimal allowed object scale to 1.1. The rightmost column block also shows the training images that are closest to the generated images (cosine similarity in pixel space). We can see our model does not memorize the training data.

Figure 16: Generation results of our model on CIFAR-10 with minimal allowed scale be 1.1, From left to right, the layout is same to Fig. 15.

6.5.2 WALKING IN THE LATENT SPACE

Similar to DCGAN, we also show results by walking in the latent space. Note that our model has two or more inputs. So we can walk along any of them or their combination. In Fig. 17, we generate multiple foregrounds for the same fixed generated background. We find that our model consistently generates contextually compatible foregrounds. For example, for the grass-like backgrounds, the foreground generator generates horses and deer, and airplane-like objects for the blue sky.

6.5.3 WORD CLOUD BASED ON HUMAN STUDY

As we mentioned above, we conducted human studies on CIFAR-10. Besides asking people to select a name from a list for an image, we also conducted another human study where we ask people to use one word (free-form) to describe the main object in the image. Each image was 'named' by 5 unique people. We generate word clouds for real images, images generated by DCGAN and LR-GAN, as shown in Fig. 18.

6.6 RESULTS ON LFW FACE DATASET

We conduct experiment on face images in LFW dataset (Huang et al., 2007). Different from previous works which work on cropped and aligned faces, we directly generate the original images which contains a large portion of backgrounds. This configuration helps to verify the efficiency of LR-GAN to model the object appearance, shape and pose. In Fig. 19, we show the (intermediate) generation results of LR-GAN. Surprisingly, without any supervisions, the model generated background and faces in separate steps, and the generated masks accurately depict face shapes. Moreover, the model
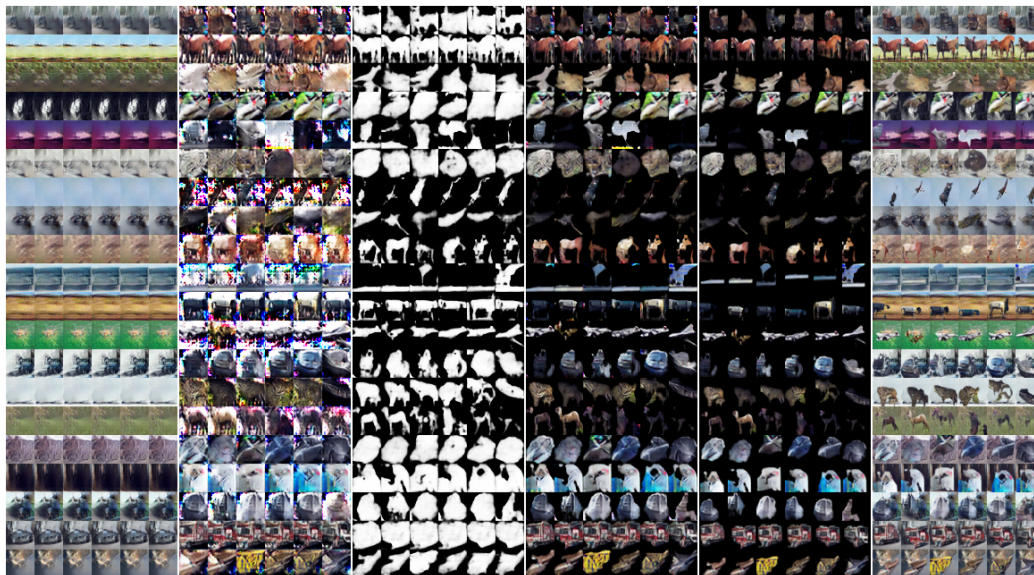
Figure 17: Walking in the latent foreground space by fixing backgrounds in our model on CIFAR-10. From left to right, the blocks are: generated background images, foreground images, foreground masks, foreground images carved out by masks, carved out foreground images after spatial transformation, and final composite images. Each row has the same background, but different foregrounds.

Figure 18: Statistics of annotations in human studies on CIFAR-10. Left to right: word cloud for real images, images generated by DCGAN, images generated by LR-GAN.
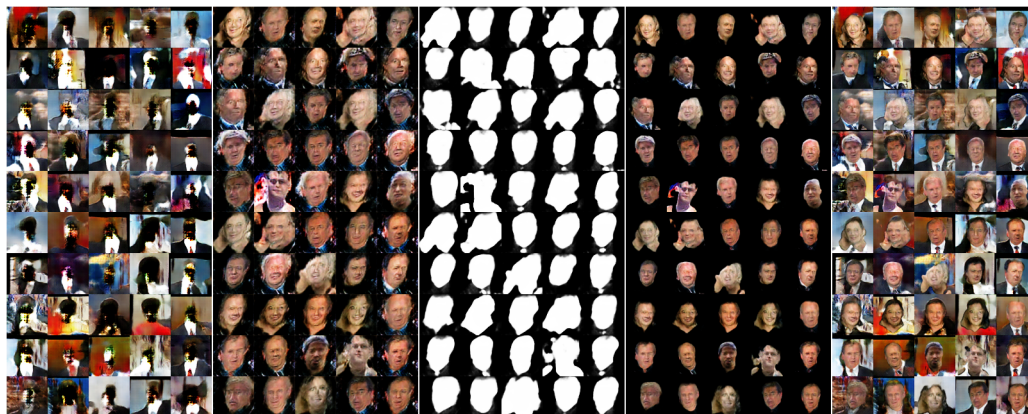

Figure 19: Generation results of our model on LFW. From left to right, the blocks are: generated background images, foreground images, foreground masks, carved out foreground images after spatial transformation, and final composite images.

learns where to place the generated faces so that the whole image looks natural. For comparison, please refer to (Kwak & Zhang, 2016) which does not model the transformation. We can find the generation results degrade much.

6.7 Statistics on Transformation Matrices

In this part, we analyze the statistics on the transformation matrices generated by our model for different datasets, including MNIST-ONE, CUB-200, CIFAR-10 and LFW. We used affine transformation in our model. So there are 6 parameters, scaling in the x coordinate (s_x), scaling in the y coordinate (s_y), translation in the x coordinate (t_x), translation in the y coordinate (t_y), rotation in the x coordinate (r_x) and rotation in the y coordinate (r_y). In Fig. 20, we show the histograms on different parameters for different datasets.These histograms show that the model produces non-trivial varied scaling, translation and rotation on all datasets. For different datasets, the learned transformation have different patterns. We hypothesize that this is mainly determined by the configurations of objects in the images. For example, on MNIST-ONE, all six parameters have some fluctuations since the synthetic dataset contains digits randomly placed at different locations. For the other three datasets, the scalings converge to single value since the object sizes do not vary much, and the variations on rotation and translation suffice to generate realistic images. Specifically, we can find the generator largely relies on the translation on x coordinate for generating CUB-200. This makes sense since birds in the images have similar scales, orientations but various horizontal locations. For CIFAR-10, since there are 10 different object categories, the configurations are more diverse, hence the generator uses all parameters for generation except for the scaling. For LFW, since faces have similar configurations, the learned transformations have less fluctuation as well. As a result, we can see that LR-GAN indeed models the transformations on the foreground to generate images.

6.8 Conditional Image Generation

Considering our model can generate object-like masks (shapes) for images, we conducted an experiment to evaluate whether our model can be potentially used for image segmentation and object detection. We make some changes to the model. For the background generator, the input is a real image instead of a random vector. Then the image is passed through an encoder to extract the hidden features, which replaces the random vector z_0 and are fed to the background generator. For the foreground generator, we subtract the image generated by the background generator from the input image to obtain a residual image. Then this residual image is fed to the same encoder to get the hidden features, which are used as the input for foreground generator. In our conditional model, we want to reconstruct the image, so we add a reconstruction loss along with the adversarial loss. We train this conditional model on CIFAR-10. The (intermediate) outputs of the model is shown in Fig. 21. Interestingly, the model successfully learned to decompose the input images into background and foreground. The background generator tends to do an image inpainting by generating a complete background without object, while the foreground generator works as a segmentation model to get object mask from the input image.

Similarly, we also run the conditional LR-GAN on LFW dataset. As we can see in Fig. 22, the foreground generator automatically and consistently learned to generate the face regions, even though there are large portion of background in the input images. In other words, the conditional LR-GAN successfully learned to detection faces in images. We suspect this success is due to that it has low cost for the generator to generate similar images, and thus converge to the case that the first generator generate background, and the second generator generate face images.

Based on these experiments, we argue that our model can be possibly used for image segmentation and object detection in a generative and unsupervised manner. One future work would be verifying this by applying it to high-resolution and more complicate datasets.
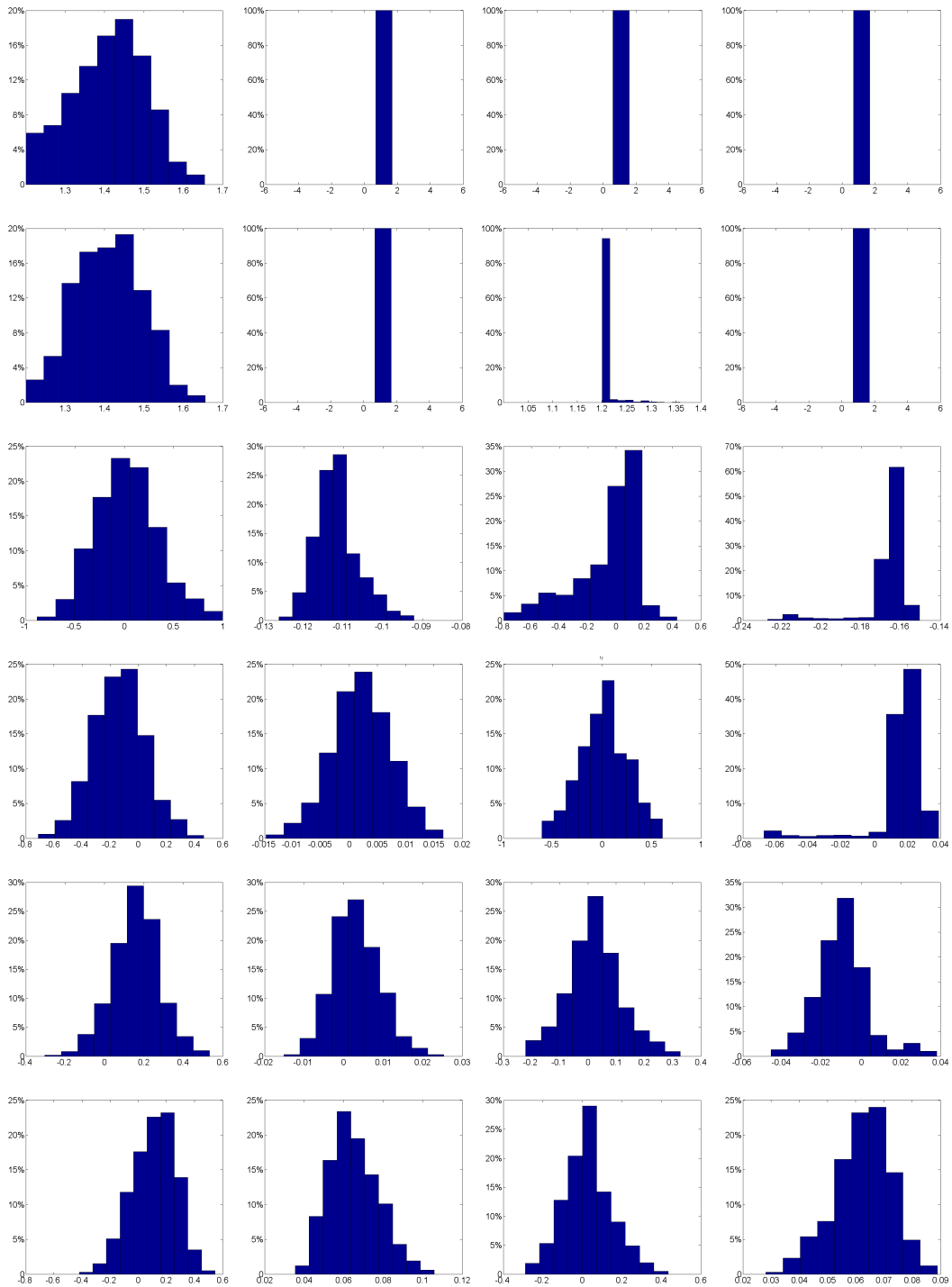
Figure 20: Histograms of transformation parameters learnt in our model for different datasets. From left to right, the datasets are: MNIST-ONE, CUB-200, CIFAR-10 and LFW. From top to bottom, they are scaling s_x, s_y, translation t_x, t_y, and rotation r_x, r_y in x and y coordinate, respectively.
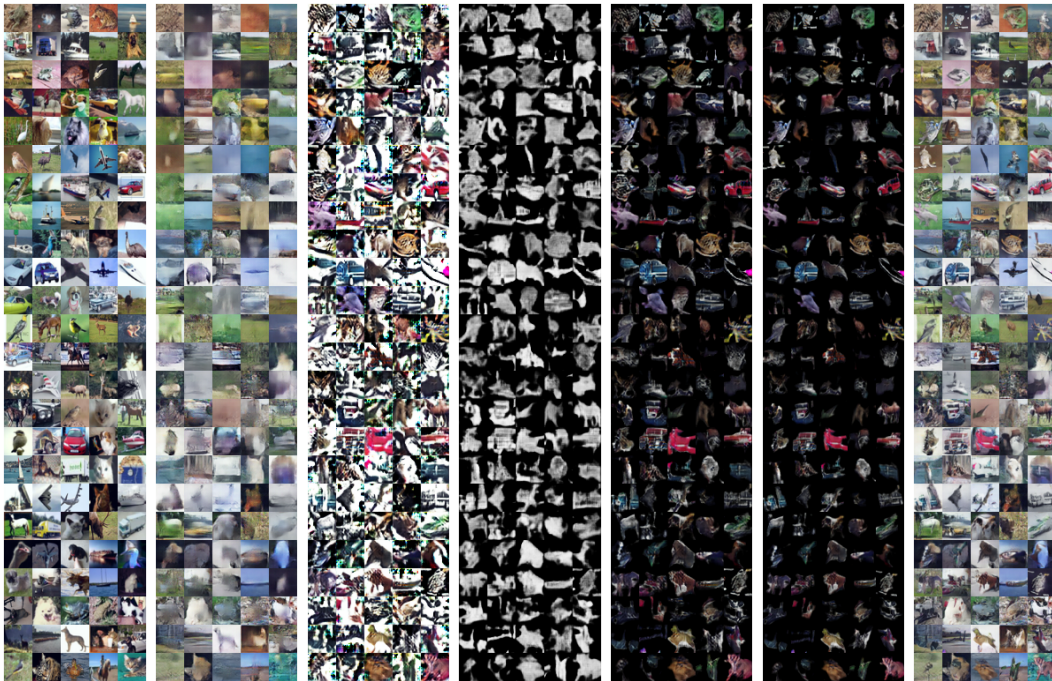
Figure 21: Conditional generation results of our model on CIFAR-10. From left to right, the blocks are: real images, generated background images, foreground images, foreground masks, foreground images carved out by masks, carved foreground images after spatial transformation, and final composite (reconstructed) images.
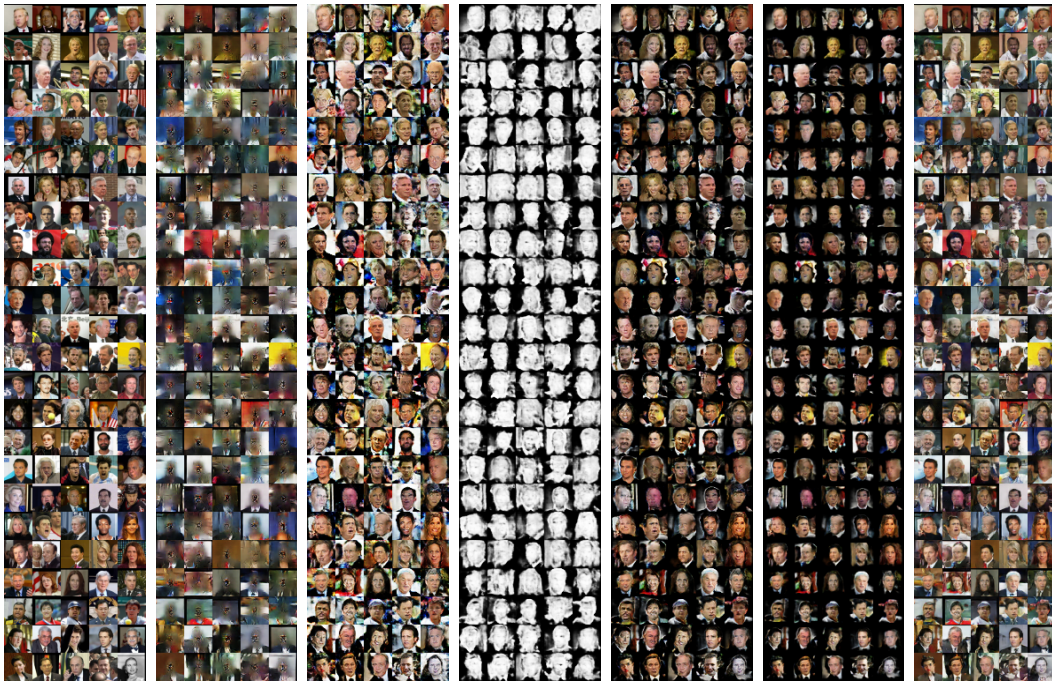

Figure 22: Conditional generation results of our model on LFW, displayed with the same layout to Fig. 21.