# Compositing Foreground and Background Using Variational Autoencoders

Zezhen Zeng, Jonathon Hare, and Adam Prügel-Bennett

University of Southampton, Southampton, United Kingdom
{zz8n17, jsh2, apb}@ecs.soton.ac.uk

**Abstract.** We consider the problem of composing images by combining an arbitrary foreground object to some background. To achieve this we use a factorized latent space. Thus we introduce a model called the "Background and Foreground VAE" (BFVAE) that can combine arbitrary foreground and background from an image dataset to generate unseen images. To enhance the quality of the generated images we also propose a VAE-GAN mixed model called "Latent Space Renderer-GAN" (LSR-GAN). This substantially reduces the blurriness of BF-VAE images.

**Keywords:** Representation learning · VAE · Disentanglement

## 1 Introduction

Learning factorized representations of visual scenes is a challenging problem in computer vision. Human brains can process the realistic scene as a whole and decompose it into different parts using visual clues and prior knowledge. This cognitive ability also enables humans to imagine different scenes. Objects form the basis of humans' high-level cognition [36]. Thus, learning good object representations could be an important step towards making artificial intelligence closer to human intelligence. In visually inspecting a scene, one object is often attended to as the foreground and the rest of the scene is the background. There exists a considerable body of work learning representation for each object in a scene and achieve objects segmentation [3, 14, 28, 32]. We argue that a good object representation should not only benefit the downstream tasks such as classification, or segmentation, but also enable generative models to create images conditioned on the object representations.

Our aim is to build a generative model for classes of images that allows us to alter the foreground objects independently of the background. This requires building a model that factorizes and composites these two part representations of the image. Existing works that can factorize the foreground and background of images are all based on hierarchical Generative Adversarial Networks [26, 35, 41]. Here we introduce a new VAE-based model that can be used to factorize the background and foreground objects in a continuous latent space and composite factors of those training images to generate new images in one shot. Compared to GAN-based models, our VAE-based models can infer the latent representation of existing images in addition to performing generation.

We consider the decomposition of an image $\boldsymbol{x}$ into a set of foreground pixels $\boldsymbol{f}$ and background pixels $\boldsymbol{b}$ such that

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

where $\boldsymbol{m}$ is the binary mask of the foreground and $\odot$ denotes elementwise multiplication. Thus we require a mask or bounding box to crop the foreground object $\boldsymbol{f}$ out, as shown in Figure 1. Row C is the generated images of our model by combining different

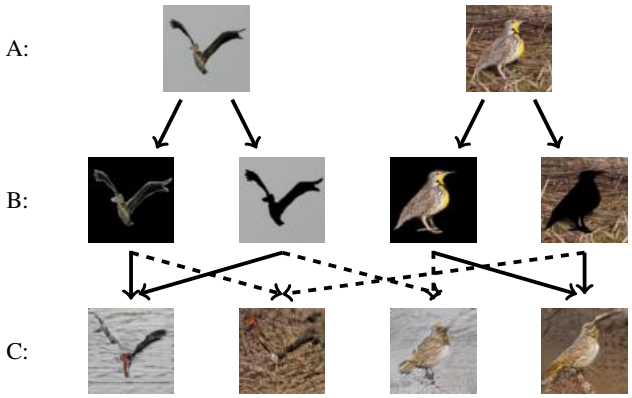


Fig. 1: Our model can disentangle the foreground and background of images and combine factors from different images to generate new images. Row A is the original images. Row B is the decomposition of Row A. Row C is the generations of combining different factors. Between row B and row C, the solid arrow means reconstruction and dashed arrow means generation.

factors in row B. The background and foreground of images in row C are not totally the same as images in row A and row B, this is due to a trade-off between the similarity and reality in our model which will be addressed in the following section.

We name this model the *Background and Foreground VAE* (BFVAE). BFVAE consists of two different VAEs: VAE-B and VAE-F. We train VAE-F on the foreground object, $\boldsymbol{f}$, while the VAE-B encoder is given the full image, $\boldsymbol{x}$, as input. We concatenate the latent representation for VAE-B with that of VAE-F before feeding it into the decoder of VAE-B. This operation encourages the encoder in VAE-B to ignore information about the foreground object as this is already encoded. It is crucial to use the full image, $\boldsymbol{x}$, and not just the background image, $\boldsymbol{b}$, as the input of VAE-B since the original information in $\boldsymbol{x}$ helps the model to generate pure background images without a hole when we use a pure black image as the input of VAE-F. It also helps to fix the hole when we exchange the foreground objects among images.

The drawback of VAE-based model is the generated images tend to be blurry, thus we propose another VAE-GAN mixed model which can generate high-quality images but also obtain an approximate latent space of a pre-trained VAE model (BFVAE in this case). We name it Latent Space Renderer-GAN (or LSR-GAN). By feeding the output of the generator to the encoder of a pre-trained VAE, LSR-GAN is more stable and can avoid mode drops. In addition, we note that a different pre-processing operation of

the images make a significant difference to the FID scores. Thus, we argue that it is necessary to clarify how the FID score is computed.

The main contributions of our work are fourfold:

- We propose a new VAE-based model called BFVAE that can composite the foreground and background of images and generate images by combining different factors.
- We introduce a VAE-GAN mixed model called LSR-GAN which enables us to generate high-quality images with the approximated disentangled latent space.
- We demonstrate that BFVAE can factorize the foreground and background representation and generate high-quality images when combined with LSR-GAN. Moreover, We show our model can obtain FID scores that are comparable to the state-of-the-art model.
- We demonstrate that BFVAE is able to factorize other factors when we have additional information available for training (like class labels).

## 2 Related Work

There are several papers on composition and decomposition of images [3, 9, 11, 12, 14, 27, 28, 34]. Genesis [12] tries to decompose image into object representations using a recurrent neural network which builds a strong relationship among each objects. All the models generate or reconstruct images part by part and stitch all the parts together.

Not many works focus on compositing images in a background-foreground manner. The existing models are all GAN-based model that generate foreground and background separately and recursively, the generated images are stitched at the final stages [26, 35, 41]. ReDO [5] can segment foreground objects from images by using a GAN but it can not change the shape of the original foreground. Only MixNMatch [26] can encode real data into discrete codes or feature maps (in the Feature-mode). Although MixNMatch and FineGAN [35] do not require masks of images nevertheless they both need bounding boxes and the number of categories for training. Our model generates the whole image in one shot while it can still learn a continuous factorized latent space. Moreover, the model can be trained either with or without supervision. And FBC-GAN [10] is another GAN-based model that generates the foreground and the background concurrently and independently.

There is a vast literature about learning disentangled representations based on deep generative models. Many unsupervised generative models have developed disentanglement in a latent space. GAN-type models are usually based on InfoGAN [7] while most VAE-type models are based on $\beta$-VAE [17]. There is a trade-off in $\beta$-VAE between disentanglement and reconstruction quality. There are many attempts to solve this trade-off issue [6, 21, 42]. Both approaches modify the ELBO to avoid getting worse reconstructions while keeping the disentanglement. Kumar *et al.* [25] propose a VAE-based model that penalizing the covariance between latent dimensions without modifying the $\beta$ value.

In contrast to unsupervised models, it is easier for supervised models to learn a factorized representation. VAEs have been used in a semi-supervised manner to factorize the class information and other information [8], or combinations of VAE and GAN aim to learn disentangled representations [29, 30, 38, 40]. ML-VAE [1] is another

VAE-based model that requires weak-supervision. In their work, they propose group supervision that uses a group of images with the same label instead of a single image to learn a separable latent space. Esser *et al.* [13] combine the U-Net [33] with a VAE and one extra encoder to learn a disentangled representation of human pose. Harsh *et al.* [15] is the closest work to our model, we both have two VAEs and a pair of inputs, while they use a different image with the same label and they also swap the factors during training (something we do not require in our model).

The VAE and GAN mixed models has been explored for a long time, and almost all the works train the VAE and GAN simultaneously [2, 4, 29, 37], while our LSR-GAN requires a pre-trained VAE. Some methods feed the output of the decoder into the encoder [2, 19, 37] which is similar to our model. VEEGAN [37] introduces an encoder to the GAN and tries to train all the network simultaneously. However, only LSR-GAN tries to map the sample space into a latent space of a pre-trained VAE by maximum likelihood.

## 3 Model

In this section we describe the components of our model starting from the classic model of VAEs.

### 3.1 VAE

The Variational Autoencoder (VAE) [23] is a deep generative model that learns a distribution over observed data, $\boldsymbol{x}$, in terms of latent variables, $\boldsymbol{z}$. The original VAE approximates the intractable posterior by using a variational approximation to provide a tractable bound on the marginal log-likelihood called the evidence lower bound (ELBO)

$$\log p_\theta(\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}\big(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\big)\,. \tag{2}$$

Commonly, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is the output of an inference network with parameters $\phi$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is generated by a decoder network with parameters $\theta$.

### 3.2 BFVAE

Starting from a mask $\boldsymbol{m}$ of the foreground object we can extract the foreground $\boldsymbol{f}$ from the image $\boldsymbol{x}$ using $\boldsymbol{f} = \boldsymbol{x} \odot \boldsymbol{m}$. We use $\boldsymbol{f}$ and $\boldsymbol{x}$ as inputs to our two VAEs. The architecture of our model is shown in Figure 2. For simplicity, we omit symbols of the parameters. The top network is VAE-F and it acts like a vanilla VAE. The encoder $E_f$ generates a probability distribution, $q(\boldsymbol{z}_f|\boldsymbol{f})$ that acts as a latent space representation of $\boldsymbol{f}$. A sample from this distribution, $\boldsymbol{z}_f$, is used by the decoder $D_f$ to generate a reconstruction, $\hat{\boldsymbol{f}}$, of the input $\boldsymbol{f}$. The top VAE ensures the $\boldsymbol{z}_f$ contain representations of foreground objects. The bottom network is VAE-B. The original image, $\boldsymbol{x}$, is given to the encoder, $E_b$ that generates a probability distribution, $q(\boldsymbol{z}_b|\boldsymbol{x})$. A latent variables $\boldsymbol{z}_b$, is sampled from this distribution and concatenated with $\boldsymbol{z}_f$. This concatenated vector is sent to decoder $D_b$ that must reconstruct the original image. Thus, we modify the ELBO for VAE-B to be

$$\mathscr{L}_b = \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|(\boldsymbol{z}_b, \boldsymbol{z}_f))\right] - D_{KL}(q(\boldsymbol{z}_b|\boldsymbol{x})||p(\boldsymbol{z}_b))\,. \tag{3}$$

Since the input $\boldsymbol{f}$ does not contain any information about the background, it is assured that the latent variables $\boldsymbol{z}_f$ only contain information about the foreground. For

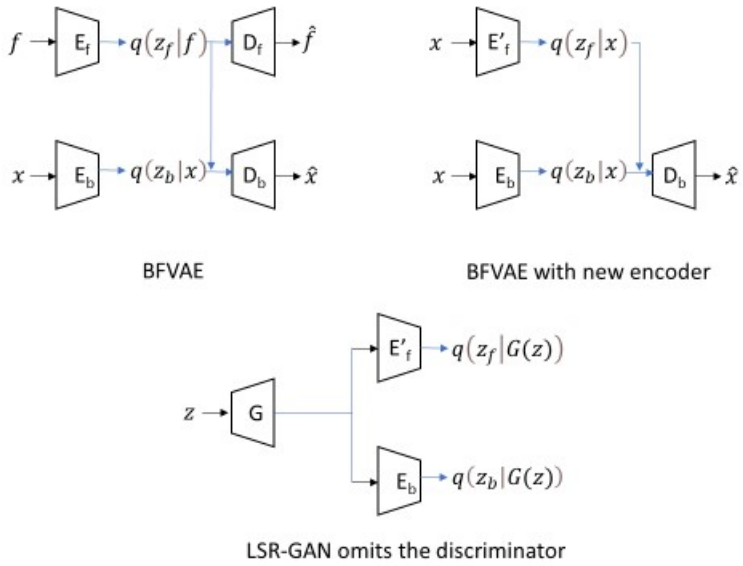


Fig. 2: The diagram of BFVAE, BFVAE with new encoder and the LSR-GAN which we omits the discriminator.

VAE-B, the encoder can extract information about both foreground and background from $\boldsymbol{x}$. When we train the decoder with both $\boldsymbol{z}_f$ and $\boldsymbol{z}_b$, it can force $\boldsymbol{z}_b$ to discard the information about the foreground and only leave information about background. This also enable us to extract the pure background from images by using $\boldsymbol{z}_f$ obtained from a pure black image. In the initial stages of training, $\boldsymbol{z}_b$ contain all the information about the image. This makes the decoder of VAE-B prone to ignore $\boldsymbol{z}_f$ especially when the dataset is complicated. There are two methods to alleviate this issue. The first method is to set the size of $\boldsymbol{z}_b$ to be reasonably small, it forces the $\boldsymbol{z}_b$ to discard information, but this design makes it hard to find an accurate size for $\boldsymbol{z}_b$. Thus, we recommend the second method which is turning the model into $\beta$-VAE,

$$\mathscr{L}_b = \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|(\boldsymbol{z}_b,\boldsymbol{z}_f))\right] - \beta D_{KL}(q(\boldsymbol{z}_b|\boldsymbol{x})||p(\boldsymbol{z}_b)) \tag{4}$$

It is well known (see, for example, Hoffman *et al.* [18] and Kim *et al.* [21]) that the expected KL term in Equation (4) can be rewritten as

$$\mathbb{E}_{p_{data(x)}}[D_{KL}(q(\boldsymbol{z}_b|\boldsymbol{x})||p(\boldsymbol{z}_b))] = D_{KL}(q(\boldsymbol{z}_b)||p(\boldsymbol{z}_b)) + I(\boldsymbol{x},\boldsymbol{z}_b) \tag{5}$$

By setting $\beta > 1$, we penalize both terms on the right side of Equation (5). Penalizing $D_{KL}(q(\boldsymbol{z}_b)||p(\boldsymbol{z}_b))$ encourages factorization of the latent space, while at the same time

it pushes $q(\boldsymbol{z}_b|\boldsymbol{x})$ towards a standard Gaussian distribution. But the most important part in BFVAE is that we penalize the mutual information term $I(\boldsymbol{x}, \boldsymbol{z}_b)$ which helps $\boldsymbol{z}_b$ to discard information about the foreground.

### 3.3 LSR-GAN

A prominent problem of vanilla VAEs is the blurriness of the output. Thus we introduce a VAE-GAN mixed model that can learn a latent space from BFVAE and generate high-quality images. The idea is that we pass the output $G(\boldsymbol{z})$ of the generator $G$ into the two encoders of BFVAE, and ask the two encoders to map $G(\boldsymbol{z})$ to the latent vectors $\boldsymbol{z}$ that we used to generate $G(\boldsymbol{z})$. By doing this, the generator will generate an image with a latent space encoding, $\boldsymbol{z}$, of the pre-trained BFVAE. It can be seen as a simple regularization term of the normal GAN loss function for the generator

$$\mathscr{L}_G = \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(D(G(\boldsymbol{z})))] + \lambda \log(q(\boldsymbol{z}|(E_b(G(\boldsymbol{z})), E'_f(G(\boldsymbol{z})))) \tag{6}$$

where $E_b$ means the encoder of VAE-B, and the $(E_b(G(\boldsymbol{z})), E'_f(G(\boldsymbol{z}))$ in the second term represents the concatenation of $E_b(G(\boldsymbol{z}))$ and $E'_f(G(\boldsymbol{z}))$. We train a new encoder $E'_f$ that can extract the $\boldsymbol{z}_f$ from $G(\boldsymbol{z})$, we freeze all the other parts of BFVAE and replace $E_f$ with $E'_f$, then train the encoder with the original loss function. Note that when training the LSR-GAN we freeze the weights of the $E_b$ and $E'_f$. The constant $\lambda$ is an adjustable hyper-parameter providing a trade-off between how realistic the image looks and the similarity between reconstructions and real images. Although the idea is simple, it provides a powerful method to improve the image quality of BFVAE. The generator $G$ can either be a new network or a pre-trained decoder of BFVAE. The pre-trained decoder can be a strong initialization for the generator when the generator meets with model collapse.

## 4 Experiments

In this section, we show that BFVAE can factorize the foreground and background efficiently and can composite factors from different images to generate high-quality images when we combine LSR-GAN with BFVAE. Our model achieves state-of-the-art FID scores on images of size $64 \times 64$ compared to baseline models. We show the results quantitatively and qualitatively on natural images (CUB [39], Stanforddogs [20] and Stanfordcars [24]). Moreover, we notice that our model can factorize other kinds of attributes in the image as long as we change the input of VAE-F (See details in the following context). We evaluate this on MNIST. We set $\beta = 8$ for Dogs dataset and $\beta = 5$ otherwise; $\lambda$ is 1 for all datasets.

The architecture of the encoder is 4 layers CNN network with Batch Normalization and 2 layers fully connected network, and decoder consists of 5 layers CNN network. When we combine the BFVAE with LSR-GAN, we use 4 residual blocks in both encoder and decoder of BFVAE with downsampling and upsampling operation respectively. And we add one extra linear layer at the end of the encoder and the beginning of the decoder. The generator of LSR-GAN is similar to the decoder of BFVAE and the discriminator consists of 4 residual blocks with a spectral norm [31]. We apply orthogonal initialization to both LSR-GAN and the new Encoder $E'_f$ and optimize the model using ADAM [22] optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We train the BFVAE and the new encoder $E'_f$ for only 100 epochs and train the LSR-GAN for 600 epochs.

### 4.1 BFVAE with LSR-GAN

Due to the blurriness of images created by the VAE decoders, BFVAE cannot perform well on natural images. Thus, we use the LSR-GAN described above to generate high-quality images. We first train a new encoder $E'_f$ before training the new generator. We find it is better to choose the same $\beta$ value for VAE-B and VAE-F when we combine BFVAE with LSR-GAN. Our experiments use three datasets, the whole dataset of CUB, the training set of *Stanford cars* and 12000 images from *Stanford dogs* (120 classes $\times$100 images).

Table 1: The FID scores (lower is better) of saving resized images (left part) and feeding resized images directly (right part) at $64 \times 64$ scale.

| | FID (saved) | | | FID (directly) | | |
|---|---|---|---|---|---|---|
| | CUB | Cars | Dogs | CUB | Cars | Dogs |
| SNGAN | 41.63 | 42.67 | 54.85 | 53.45 | 43.83 | 69.54 |
| LR-GAN | 35.22 | 30.30 | 86.67 | 51.85 | 38.80 | 104.45 |
| FineGAN | 24.51 | 31.32 | **33.66** | **16.79** | 23.61 | **39.43** |
| MixNMatch | 28.25 | 37.42 | 36.62 | 20.63 | 25.53 | 44.42 |
| LSR-GAN | **19.12** | **18.01** | 44.22 | 28.15 | **18.99** | 61.54 |

BFVAE+LSR-GAN obtains superior FID scores [16] compared to previous models for images of size $64 \times 64$. It is well-known that the number of images and the implementation (Pytorch or Tensorflow) we use to calculate FID can strongly affect the results. Moreover, we notice an additional pre-processing operation that can make differences to the results. Commonly, when we calculate the FID scores of the dataset we need to resize the original images to the same size of our generated images, whether we save the resized images and reload them or feed the resized images to the inception model directly makes a significant difference to the FID scores we obtain. Given the extreme sensitivity of FID scores to these details, it is necessary for the process of computing FID scores to be fully documented to make meaningful comparisons. Thus, we report two FID results of feeding resized images directly and saving resized images. The outputs of our model are saved as PNG image. (Given these changes are not noticeable to humans it raises some concerns about how seriously we should take FID scores. However, given these are the standard metric in this field we present the results as honestly as we can.)

**Quantitative Results** We evaluate FID on 10K randomly generated images of size $64 \times 64$ for three different datasets. For LR-GAN, FineGAN and MixNMatch, we use the authors' publicly-available code. For a fair comparison, we use the same architecture of our LSR-GAN to train a SNGAN. We also tried replacing the original discriminator of other models with the same discriminator we use in the LSR-GAN, but it does not improve the results for either FineGAN or MixNMatch. Thus, we present results with those models' original architectures. As shown in Table 1, the results in the two halves of the table are different even though the only difference is whether we saved

the images or kept them in main memory (although saving images will introduce small errors due to truncation, these are not observable to a human viewer). This shows that the small difference in FID scores is not that meaningful. Comparing to previous models, our model is the best overall when we save the resized images while FineGAN is the best one otherwise. But our model has, by a considerable margin, the smallest number of parameters and training time. The size of LRGAN, FineGAN and MixN-Match's saved models are 65.6MB, 158.7MB and 336.7MB respectively. The size of our BFVAE+LSR-GAN is only 22.1MB.

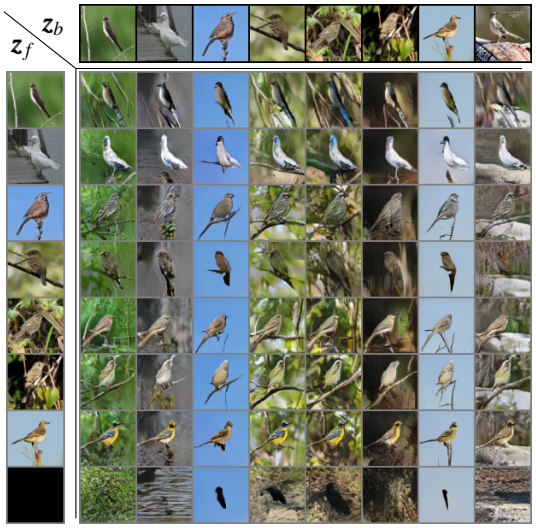


Fig. 3: Generation by swapping $z_f$ and $z_b$. The top row and the first column are both the input images of two encoders.

**Conditional Generation** In Figure 3 we show images generated by our method on three datasets. The top row and the first column are both the input images of two encoders. The other images are generated by combining factors from the two different images. As mentioned before, there is a trade-off between the realism and the similarity of generated images when we train the LSR-GAN, so there is a slight difference between reconstructions and input images. In the last row, some images are not pure background images. Because the discriminator has never seen images without foreground and can easily classify a pure background image as fake, this prevents the generator from generating a pure background image for some backgrounds especially pure colour background (e.g. sky). For the same reason, the background changes a little bit in the same column when the foreground is not harmonious with the background. This demonstrates how the LSR-GAN reduces its similarity when it tries to learn an approximate latent space of the pre-trained BFVAE. The nice part of this phenomena is the generator can adjust details in the background, such as the orientation of branches, to fit the foreground.

**Continuous Latent Space** We demonstrate the continuity of the latent space learnt by BFVAE in Figure 4 where we show the interpolation between two images. The top-

Fig. 4: Interpolation in the latent space, the left-top image and the right bottom image are the original images, others are the interpolations between the two images.

left image and the bottom-right image are the original images. Other images are the interpolations between the two images. As we move along the axes, we change $\boldsymbol{z}_f$ or $\boldsymbol{z}_b$. Both transitions between real images and fake images are smooth in the latent space, but it is obvious that even if we do not change $\boldsymbol{z}_b$ for each column, the birds (foreground) are slightly changing, this also happens for background in each row. The two reasons for this change are the same as above: firstly, the approximate latent space loses some similarity; and secondly, the discriminator can classify the unreal images like waterbirds on the branch or non-aquatic bird on the water as fake images, then the discriminator forces the generator to generate non-aquatic bird on a branch or waterbirds with water, which results in the slight change of both foregrounds and background even we do not change one of $\boldsymbol{z}_f$ and $\boldsymbol{z}_b$. This change is also a trade-off between reality and similarity.

### 4.2 Substitute Attributes for Foreground

Apart from factorizing the background and foreground in the latent space, our model is also capable to factorize attributes such as style and content. The only thing we need to do is to substitute foreground images $\boldsymbol{f}$ with images that represent each different class, and $\boldsymbol{x}$ should be images from the same class as $\boldsymbol{f}$. For example, if we want VAE-F to learn the information about digit label of MNIST, we choose 10 images from 10 classes randomly and use these 10 images as fixed input of VAE-F. This differs from previous work on conditional image generation as we use images instead of one-hot vectors as labels. Then the VAE-B will learn a latent space about the style of images. And Figure 5 shows images generated using MNIST analogous to Figure 4, where the top row is $\boldsymbol{x}$ and the first column is $\boldsymbol{f}$. It can be observed that the generated images obtain class information from the first column and the style information from the top row.

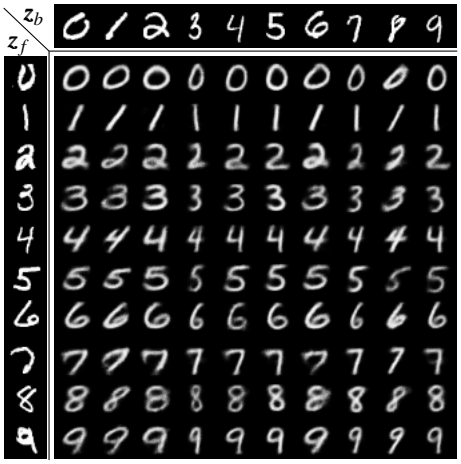


Fig. 5: Generation by swapping $\boldsymbol{z}_f$ and $\boldsymbol{z}_b$ on MNIST. The top row is $\boldsymbol{x}$ and the first column is $\boldsymbol{f}$.

## 5 Discussion

Although several works have shown great success by representing scenes using their components [3, 14, 28], what defines a good object representation is still in discussion. We argue that a good object representation should also benefit the image generation task. We believe that enabling generative models to generate certain objects with random backgrounds should also be a property of good object representations.

Moreover, conditional image generation tasks such as the one discussed here are useful in clarifying what we require of a good image representation. After all the ability of dreaming and imagining scenes seem to be an intrinsic human ability. In the mammalian visual system there is plenty of evidence that scenes are disentangled in different areas of the visual cortex and later re-integrated to obtain a complete understanding of a scene. Although very much simplified the BFVAE makes a step towards learning such a disentangled representation of the foreground and background.

## References

1. Bouchacourt, D., Tomioka, R., Nowozin, S.: Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In: Proceedings of the AAAI Conference on Artificial Intelligence (2018)
2. Brock, A., Lim, T., Ritchie, J.M., Weston, N.: Neural photo editing with introspective adversarial networks. In: Proceedings of International Conference on Learning Representations (2017)
3. Burgess, C.P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., Lerchner, A.: Monet: Unsupervised scene decomposition and representation. arXiv preprint arXiv:1901.11390 (2019)
4. Che, T., Li, Y., Jacob, A.P., Bengio, Y., Li, W.: Mode regularized generative adversarial networks. In: Proceedings of International Conference on Learning Representations (2017)

5. Chen, M., Artières, T., Denoyer, L.: Unsupervised object segmentation by redrawing. In: Advances in Neural Information Processing Systems (2019)
6. Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems. pp. 2610–2620 (2018)
7. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2172–2180 (2016)
8. Cheung, B., Livezey, J.A., Bansal, A.K., Olshausen, B.A.: Discovering hidden factors of variation in deep networks. arXiv preprint arXiv:1412.6583 (2014)
9. Crawford, E., Pineau, J.: Spatially invariant unsupervised object detection with convolutional neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 3412–3420 (2019)
10. Cui, K., Zhang, G., Zhan, F., Huang, J., Lu, S.: Fbc-gan: Diverse and flexible image synthesis via foreground-background composition. arXiv preprint arXiv:2107.03166 (2021)
11. Dubrovina, A., Xia, F., Achlioptas, P., Shalah, M., Groscot, R., Guibas, L.J.: Composite shape modeling via latent space factorization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8140–8149 (2019)
12. Engelcke, M., Kosiorek, A.R., Jones, O.P., Posner, I.: Genesis: Generative scene inference and sampling with object-centric latent representations. In: International Conference on Learning Representations. (2020)
13. Esser, P., Sutter, E., Ommer, B.: A variational u-net for conditional appearance and shape generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8857–8866 (2018)
14. Greff, K., Kaufman, R.L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., Lerchner, A.: Multi-object representation learning with iterative variational inference. In: International Conference on Machine Learning. pp. 2424–2433. PMLR (2019)
15. Harsh Jha, A., Anand, S., Singh, M., Veeravasarapu, V.: Disentangling factors of variation with cycle-consistent variational auto-encoders. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 805–820 (2018)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. pp. 6626–6637 (2017)
17. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: Beta-VAE: Learning basic visual concepts with a constrained variational framework. In: Proceedings of International Conference on Learning Representations (2017)
18. Hoffman, M.D., Johnson, M.J.: Elbo surgery: yet another way to carve up the variational evidence lower bound. In: Workshop in Advances in Approximate Bayesian Inference (2016)
19. Huang, H., He, R., Sun, Z., Tan, T., et al.: IntroVAE: Introspective variational autoencoders for photographic image synthesis. In: Advances in Neural Information Processing Systems. pp. 52–63 (2018)
20. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization. In: First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO (June 2011)
21. Kim, H., Mnih, A.: Disentangling by factorising. In: International Conference on Machine Learning. pp. 2649–2658. PMLR (2018)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of International Conference on Learning Representations (2015)
23. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Proceedings of International Conference on Learning Representations. (2013)

24. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013)
25. Kumar, A., Sattigeri, P., Balakrishnan, A.: Variational inference of disentangled latent concepts from unlabeled observations. arXiv preprint arXiv:1711.00848 (2017)
26. Li, Y., Singh, K.K., Ojha, U., Lee, Y.J.: Mixnmatch: Multifactor disentanglement and encoding for conditional image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
27. Lin, C.H., Yumer, E., Wang, O., Shechtman, E., Lucey, S.: ST-GAN: Spatial transformer generative adversarial networks for image compositing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9455–9464 (2018)
28. Lin, Z., Wu, Y.F., Peri, S.V., Sun, W., Singh, G., Deng, F., Jiang, J., Ahn, S.: Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. In: Proceedings of International Conference on Learning Representations. (2020)
29. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. In: Proceedings of International Conference on Learning Representations (2016)
30. Mathieu, M.F., Zhao, J.J., Zhao, J., Ramesh, A., Sprechmann, P., LeCun, Y.: Disentangling factors of variation in deep representation using adversarial training. In: Advances in Neural Information Processing Systems. pp. 5040–5048 (2016)
31. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
32. Nash, C., Eslami, S.A., Burgess, C., Higgins, I., Zoran, D., Weber, T., Battaglia, P.: The multi-entity variational autoencoder. In: NeurIPS Workshops (2017)
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
34. Schor, N., Katzir, O., Zhang, H., Cohen-Or, D.: Componet: Learning to generate the unseen by part synthesis and composition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8759–8768 (2019)
35. Singh, K.K., Ojha, U., Lee, Y.J.: Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
36. Spelke, E.S.: Principles of object perception. Cognitive science **14**(1), 29–56 (1990)
37. Srivastava, A., Valkov, L., Russell, C., Gutmann, M.U., Sutton, C.: VEEGAN: Reducing mode collapse in GANs using implicit variational learning. In: Advances in Neural Information Processing Systems. pp. 3308–3318 (2017)
38. Szabó, A., Hu, Q., Portenier, T., Zwicker, M., Favaro, P.: Challenges in disentangling independent factors of variation. arXiv preprint arXiv:1711.02245 (2017)
39. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200 (2010)
40. Xiao, T., Hong, J., Ma, J.: DNA-GAN: Learning disentangled representations from multi-attribute images. arXiv preprint arXiv:1711.05415 (2017)
41. Yang, J., Kannan, A., Batra, D., Parikh, D.: LR-GAN: Layered recursive generative adversarial networks for image generation. arXiv preprint arXiv:1703.01560 (2017)
42. Zhao, S., Song, J., Ermon, S.: InfoVAE: Information maximizing variational autoencoders. arXiv preprint arXiv:1706.02262 (2017)