GeneGAN: Learning Object Transfiguration and Attribute Subspace from Unpaired Data

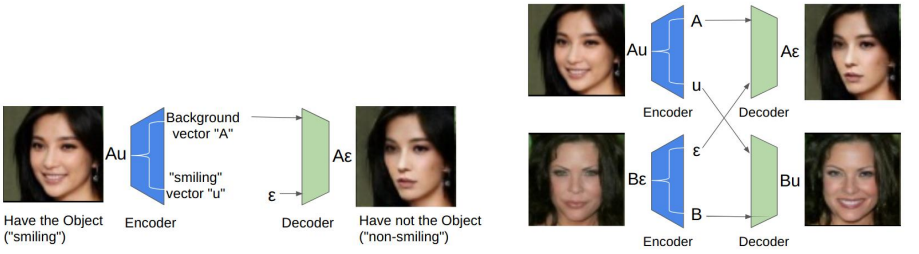
Shuchang Zhou[1]
zsc@megvii.com
Taihong Xiao[12]
xiaotaihong@pku.edu.cn
Yi Yang[1]
yangyi@megvii.com
Dieqiao Feng[1]
fdq@megvii.com
Qinyao He[1]
hqy@megvii.com
Weiran He[1]
hwr@megvii.com

[1] Megvii Inc.
Beijing, China

[2] Department of Information Science,
School of Mathematical Sciences,
Peking University
Beijing, China

Abstract

Object Transfiguration generates diverse novel images by replacing an object in the given image with particular objects from exemplar images. It offers fine-grained controls of image generation, and can perform tasks like "put exactly those eyeglasses from image A onto the nose of the person in image B". However, object transfiguration often requires disentanglement of objects from backgrounds in feature space, which is challenging and previously requires learning from paired training data: two images sharing the same background but with different objects. In this work, we propose a deterministic generative model that learns disentangled feature subspaces by adversarial training. The training data are two unpaired sets of images: a positive set containing images that have some kind of object, and a negative set being the opposite. The model encodes an image into two complement features: one for the object, and the other for the background. The object and background features from a "positive" parent and a "negative" parent, can be recombined to produce four children, of which two are exact reproductions, and the other two are crossbreeds. Minimizing the adversarial loss between crossbreeds and parents will ensure the crossbreeds inherit the specific objects of parents. On the other hand, minimizing the reconstruction loss between reproductions and parents can ensure the completeness of the features. Overall, the object and background features are complete and disentangled representations of images. Moreover, the object features are found to constitute a multidimensional attribute subspace. Experiments on CelebA and Multi-PIE datasets validate the effectiveness of the proposed model on real world data, for generating images with specified eyeglasses, smiling, hair styles, and lighting conditions.



(a) Object Removal (b) Object Transplanting

Figure 1: Components and working mechanism of GeneGAN. (a) Encoder of GeneGAN decomposes an image to the background feature A and the object feature u. The decoder can reconstruct an image without the object (a non-smiling face), from background feature A and the zero object feature ε. (b) Decomposed object feature can be used to transplant the object to another image. When "smiling" feature u and background feature B are fed to a decoder, the generated image Bu would have the same level and style of smiling as Au. The outputs of Encoder and inputs of Decoder in the second row are flipped respectively for clarity of presentation.

1 Introduction

Object transfiguration is a type of conditional image generation. It first decomposes an image into an object part and a background part. The object can be modified before recombination with the unchanged background to generate a novel image. Object transfiguration can produce images of desired attributes and has found applications in image editing [1, 2, 3, 4, 5, 6], and image synthesizing [7, 8, 9]. Depending on the task, the object can be physical objects like eyeglasses, or more abstract concepts like facial expressions. The generated images would then be answers to questions like "what if her eyeglasses is on my nose?" or "what if I smile like her?"

Under the Linear Feature Space conjecture [10] for features extracted by Deep Neural Networks, we may achieve complex object transfiguration tasks like removal and transplanting of objects, by linear operations on the feature vectors. The images generated from the modified feature vectors can still be natural-looking and have negligible artifacts. In fact, previous works [6, 11] have shown that making a face in an image smile, is as simple as addition with an vector in feature space:

$$\text{smiling face} = \phi^{-1}(\phi(\text{non-smiling face}) + \mathbf{v}_{\text{smiling}}), \tag{1}$$

where ϕ is a mapping from images to features, and the transform vector $\mathbf{v}_{\text{smiling}}$ can be computed as the difference between cluster centers of features of smiling faces and non-smiling faces.

However, there are many styles and levels of smiling. For example, some kinds of smiling do not expose teeth, and some are more manifested in eyes than in mouths. Hence representing smiling by a single transform vector will severely limit the diversity of smiling in generated images. To address this diversity issue, the Visual Analogy-Making [12] method proposes to employ Object Transfiguration for image generation. A pair of reference images, for example two images where the same person smiles in one and not in the other, is used to to specify the object and consequently the transform vector. Though the method significantly

increases the diversity of generated images, such paired data are hard to acquire except in controlled environments.

Yet another approach to Object Transfiguration is the recently proposed GAN with cyclic loss approach, which exploits Dual Learning [13, 14, 15, 16] to map between the source images (non-smiling faces) and the target images (smiling faces). Nevertheless, Dual Learning relies on the invertibility of the mapping for the cyclic loss to work. When the intrinsic dimensions of the source and target domains are not the same, for example when the source domain does not have the object and the target domain has it, the cyclic loss cannot be applied.

In this work, we propose a model that can generate an object feature vector from a single image. The object feature can then be transplanted to other images to generate novel images with that particular object. Our model is made up of two parts: an Encoder that decomposes an image to a background feature part and an object feature part, and a Decoder that can combine a background feature and an object feature to produce an image. Figure 1 illustrates typical usage patterns of the proposed model: object removal and object transplanting, which are both done by some wiring of Decoders and Encoders. All instances of Decoders and Encoders share parameters respectively. Direct chaining of the Encoder and Decoder is effectively an autoencoder [17]. Unlike previous works that require class labels to ensure disentanglement of the two spaces [12, 18, 19, 20], we do not assume that any of the two spaces have additional fine-grained labels, but relies on preservation of information and adversarial training to achieve the disentanglement.

Moreover, object features produced by the Decoder are found to constitute an attribute vector space: the non-presence of objects is mapped to the zero vector, the norm is proportional to the strength of the object (like level of smiling), and linear combinations produce other feasible object features. Though previous works [6, 11] observe that the one-dimensional attribute vector also demonstrates these properties, we are able to extract higher dimensional attribute subspace, due to the increase of diversity of object features. Experiments on transfigurations of hair styles and eyeglasses demonstrate the effectiveness of GeneGAN in disentangling features, and the richness of resulting attribute subspaces. A Tensorflow implementation of our method is available at https://github.com/Prinsphield/GeneGAN.

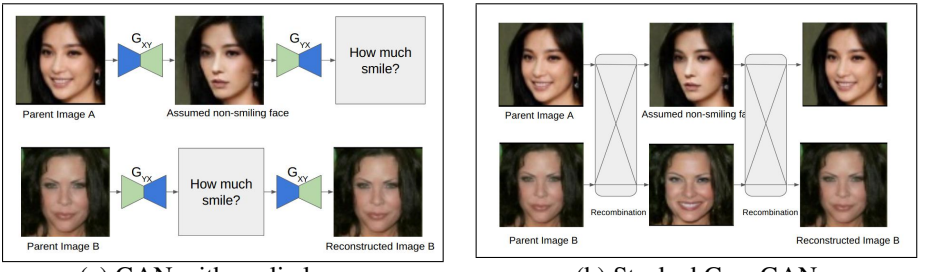
2 Method

In this section, we formally outline our method and present the problem of learning the disentangled features for backgrounds and objects.

The training data is made of two sets: the positive set containing images having the objects is $\{x^i_{Au}\}_{i=1}^N$, and the negative set is $\{x^i_{B\varepsilon}\}_{i=1}^M$. u stands for the feature of an object (also an instance of an attribute). While ε stands for the feature of the "null" object, indicating the non-presence of the attribute. x_{Au} stands for an image that will be encoded to background feature A and object feature u. We will sometimes refer to the image x_{Au} simply as Au when it is clear from context. The two sets of images need not be paired.

2.1 Model

Our model is made up of an Encoder that maps an image to two complement feature parts, and a Decoder that is the inverse of Encoder. Division between object and background is



(a) GAN with cyclic losses (b) Stacked GeneGAN

Figure 2: (a) The method of GAN with cyclic losses will suffer loss of information when removing an object: there will be incomplete information when going from non-smiling faces to smiling faces to determine the level and style of smiling in generated images. (b) The Stacked GeneGAN model. The smiling of two images are swapped twice, through recombinations, to reconstruct the original images.

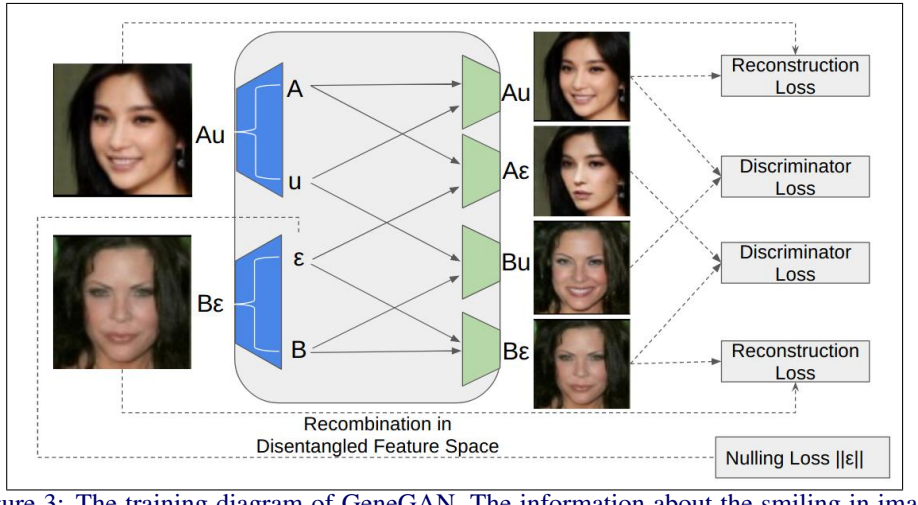


Figure 3: The training diagram of GeneGAN. The information about the smiling in image *Au* is flowed to its reproduction *Au* and crossbreed *Bu* through object feature *u*.

unknown and can be learned from the following constraints:

1. An image with $\boldsymbol{\varepsilon}$ object does not have an object, so that it is indistinguishable from the negative set $\{x^i_{B\boldsymbol{\varepsilon}}\}_{i=1}^N$ in training data.

2. An image with $\neq \boldsymbol{\varepsilon}$ object does have an object, so that it is indeed indistinguishable from the positive set $\{x^i_{Au}\}_{i=1}^N$.

Such "indistinguishable" constraint can be enforced by introduction of adversarial discriminators [21, 22], which interprets indistinguishable as "there does not exist discriminator that can assigns different score to two sets".

This inspires us to introduce the training diagram as illustrated in Figure 3 for our model, namely GeneGAN. During training, four children Au, $A\boldsymbol{\varepsilon}$, Bu, $B\boldsymbol{\varepsilon}$ are created out of object and background features of two parent images Au and $B\boldsymbol{\varepsilon}$ as follows: first, the Encoder will create four pieces of codes for the two images, namely A, u, B and $\boldsymbol{\varepsilon}$; then Decoders will create four legal recombinations as children: Au, $A\boldsymbol{\varepsilon}$, Bu, $B\boldsymbol{\varepsilon}$.

Out of four children, two recombinations Au and $B\varepsilon$ are exact reconstructions, while $A\varepsilon$ and Bu are novel crossbreeds. By using an Adversarial Discriminator to require that Au being indistinguishable from Bu, and that Au can be decoded from A and u, we can enforce that all information about the object to be carried by u. Similarly, if $A\varepsilon$ is not distinguishable from $B\varepsilon$, we can ensure that A does not contain any information about the objects. Overall, we can achieve the disentanglement of the object information from the background factors. Moreover, inclusion of reconstruction loss also stabilizes the training of Adversarial Discriminator [14, 15, 16].

2.1.1 Loss Function of Training

Given two images x_{Au} and $x_{B\varepsilon}$, the data flow of training of GeneGAN can be summarized in following equations:

$$\begin{array}{ll} (A,u) = \text{Encoder}(x_{Au}) & (B,\varepsilon) = \text{Encoder}(x_{B\varepsilon}) \\ x_{A\varepsilon} = \text{Decoder}(A,0) & x_{Bu} = \text{Decoder}(B,u) \\ x'_{Au} = \text{Decoder}(A,u) & x'_{B\varepsilon} = \text{Decoder}(B,0) \end{array} . \tag{2}$$

We use 0 instead of ε when decoding to $x_{A\varepsilon}$ and $x_{B\varepsilon}$, to ensure the hard constraint that $A\varepsilon$ should not contain any information about $B\varepsilon$, and that no information should be contained in ε.

The generator receives four types of losses: (1) the standard GAN losses, which measures how realistic the generated images are; (2) the reconstruction losses, which measures how well the original input is reconstructed after a sequence of encoding and decoding. (3) the nulling losses, which reflects how well the object features are disentangled from background features and (4) the optional parallelogram losses, which enforces a constraint between the children and the parents in image pixel values, confer Section 2.1.2 for more details. We omit the weights of the losses and left the details to online implementation. P_ε and $P_{\neq\varepsilon}$ stand for the distributions of images without and with the objects, respectively.

$$\begin{array}{cc} L^{Au}_{reconstruct} = \|x_{Au} - x'_{Au}\|_1 & L^{B\varepsilon}_{reconstruct} = \|x_{B\varepsilon} - x'_{B\varepsilon}\|_1 \\ L^{\varepsilon}_{GAN} = -\mathbb{E}_{z\sim P_\varepsilon}[\log D(x_{A\varepsilon}, z)] & L^{\neq\varepsilon}_{GAN} = -\mathbb{E}_{z\sim P_{\neq\varepsilon}}[\log D(x_{Bu}, z)] \\ L_\varepsilon = \|\varepsilon\|_1 & L_{parallelogram} = \|x_{Au} + x_{B\varepsilon} - x_{A\varepsilon} - x_{Bu}\|_1 \\ L_G = L^{Au}_{reconstruct} + L^{B\varepsilon}_{reconstruct} + L^{\varepsilon}_{GAN} + L^{\neq\varepsilon}_{GAN} + L_\varepsilon + L_{parallelogram} & \end{array} \tag{3}$$

The nulling loss will force the background information to be encoded in the background feature B. In fact, the object feature ε will contain neither object nor background information, as ε is forced to be zero.

The reconstruction loss will serve multiple purposes. First, the losses associated with Au and $B\varepsilon$ will ensure that Decoder and Encoder are inverse to each other. In addition, background feature A is forced to contain background information, to allow Decoder to reconstruct Au from A and u.

The discriminator receives the standard GAN discriminator losses, which will ensure the crossbreeds to inherit particular objects of parents.

$$\begin{aligned} L^{\varepsilon}_D &= -\mathbb{E}_{z\sim P_\varepsilon}[\log D(x_{Au}, z)] - \mathbb{E}_{z\sim P_\varepsilon}[\log(1 - D(x_{Bu}, z))] \\ L^{\neq\varepsilon}_D &= -\mathbb{E}_{z\sim P_{\neq\varepsilon}}[\log D(x_{A\varepsilon}, z)] - \mathbb{E}_{z\sim P_{\neq\varepsilon}}[\log(1 - D(x_{B\varepsilon}, z))] \\ L_D &= L^{\varepsilon}_D + L^{\neq\varepsilon}_D \end{aligned} \tag{4}$$

As the constraints are only approximately enforced by losses, there will be potential leakage of information between the object and feature part. We leave it as future work to explore even stronger enforcement of constraints.

2.1.2 Attribute Drift Problem and Parallelogram Constraint

In experiments, we observe that occasionally the encoding and decoding processes may happen to "drift" an object, i.e. have an endomorphism over the objects, while still achieving low reconstruction and discriminator losses. This happens when the same u exhibits different visual appearance when combining with different background. For example, when Au is a face with eyeglasses and the crossbreed Bu is a face with sun glasses, both the reconstruction loss and discriminator losses can still be minimal.

To avoid this drift, for spatially aligned objects, we propose the following parallelogram constraint on the image domain:

$$L_{parallelogram} = \|x_{Au} + x_{B\varepsilon} - x_{A\varepsilon} - x_{Bu}\|_1 \tag{5}$$

The loss will encourage a sun glass, when transplanted, to stay as a sun glass, and not mutating into an eyeglass. Note it will not make sense to include the parallelogram loss for GAN with cyclic losses, as the transformation of two original images are completely independent of each other. Adding a parallelogram loss will only increase the overfitting level of the model.

2.2 Comparison against GAN with cyclic loss

Method of GAN with cyclic loss suffers from under-determination problem when performing object removal and reconstruction. For example, in Figure 2(a), though the non-smiling version of a smiling face is well defined, there is no well-grounded information to determine how much smiling should be present when mapping from the non-smiling faces to smiling faces. In contrast, GeneGAN ensures preservation of information, as all information of parents Au and $B\varepsilon$ are still present in the crossbreeds $A\varepsilon$ and Bu.

However, it is also possible to add cross-links to allow of exchange information. In fact, we have also experimented with Stacked GeneGAN, a training diagram that bears more resemblance to the method of GAN with cyclic losses, as illustrated in Figure 2. With Stacked GeneGAN, only crossbreeds are created as children of the parent images. The children will be recombined again to produce grand-children, which ideally have the same features as their grand-parents. Reconstruction losses can then be used to enforce this invariance, exactly as is done in the method of GAN with cyclic losses.

However, we find this "double-swap" training diagram to be inferior to GeneGAN in experiments. We observe that the reconstruction losses will be significantly higher as the grand-children went through more nonlinear transformations than children. We conjecture that the higher reconstruction losses will compete more with the adversarial losses in the optimization, and may cause instability of training and degrades the quality of generated images.

3 Experiments

In this section we perform experiments on real-world datasets to validate the effectiveness of our method. For training, we use learning rate of 5e-5 and momentum of 0, under RMSProp [23] learning rule. The Neural Network models used in this section are all equipped with Batch Normalization [24] to speedup convergence. The encoder has three 4×4 stride-2 convolution layers with Leaky ReLU activation functions [25] and zero-padded by radius of 1. The number of output channels are 128, 256 and 512, respectively. The decoder has three 4×4 Deconvolution layers with fractional stride of $\frac{1}{2}$ [26] and zero padded by radius of 1. The number of output channels are 512, 256 and 3, respectively. The output of the last layer is clipped by tanh and then affine transformed to be in value range $[0, 1]$. Experiments are performed on Linux machines with Intel Xeon CPUs and NVidia TitanX Graphic Processing Units.

3.1 Dataset

The CelebA [27] dataset is a large-scale face attributes dataset including 202599 face images of 10177 identities, each with 40 attributes annotations and 5 landmark locations. The landmark locations can be used to spatially align the faces.

The Multi-PIE database [28] contains over 754,200 images from 337 subjects, captured under 15 viewpoints and 19 illumination conditions.

3.2 Swapping of Objects

As GeneGAN only exploits the weak 0/1 labels of the images, and that there will be no training data about the recombined versions, we will not distinguish between the training data and test data in most experiments carried out in this subsection.

An object in an image can be replaced by passing another object to the decoder. Similarly, an object can be removed by passing ε to the decoder. As our model can disentangle the object feature from the background feature, the crossing usage pattern as illustrated in Figure 1(b) can also be used to swap the objects.

In each row of Figure 4, the object parts of the images are overridden by the objects of the first image in the row. Of particular interest is Figure 4(a). It can be observed that the hair styles follow closely the source images on the top row. In fact, the directions of hairs are in good agreement with the source images. We note that some examples contain serious artifacts, for example the fourth column of Figure 4(c). We conjecture that as the helmet is a quite uncommon background, the model has difficulty in disentangling it from the object. An obvious solution is to increase the amount of training data covering such situations. There are also some difficult cases when transplanting eyeglasses between faces of different poses, as it would be necessary to rotate the eyeglasses. The parallelogram loss is no longer applicable and need be replaced with a more robust constraint.

3.3 Generalization to Unseen Images and Comparisons with GANs with cyclic losses

Figure 6 compares GeneGAN with DiscoGAN. Though DiscoGAN may also reconstruct images that are of good quality, the objects in the reconstructed images are not quite related

(a) hair

(b) smiling

(c) glasses

Figure 4: Replacing the object for images in each row with those of the column heads.

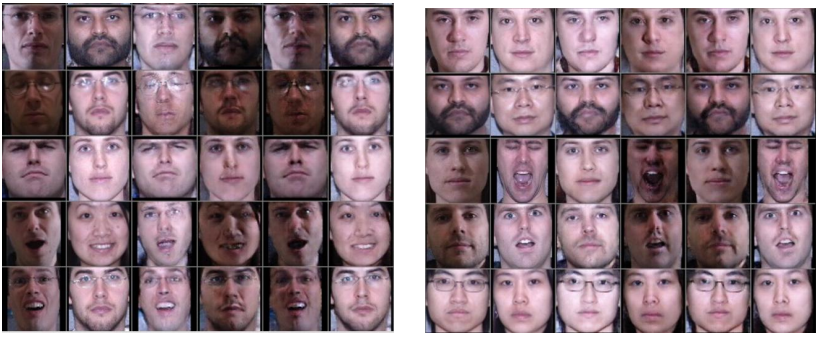
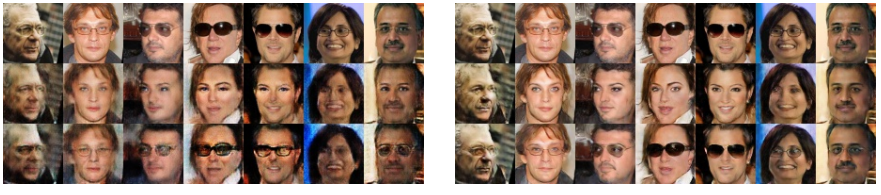


Figure 5: Swapping the lighting conditions of two faces on Multi-PIE dataset. From left to right, the six images in a row are: original Au and $B\varepsilon$, recombined $A\varepsilon$ and Bu, and reconstructed Au and $B\varepsilon$ respectively.



(a) GAN with cyclic losses (b) GeneGAN

Figure 6: Comparison between GAN with cyclic losses (the best model before divergence), and GeneGAN. The top row, middle row, and the bottom row are images with original, removed and reconstructed objects respectively.

to the original images, and the novel instances have random objects. In contrast, GeneGAN produces consistent reconstructions and the crossbreeds are more predictable.

Figure 7 gives results of testing a GeneGAN model trained on CelebA dataset on the Wider Face dataset, which contains face images in even less constrained environments. It can be seen that the GeneGAN model generalizes well to unseen data. Despite the large variations of faces of the Wider Face dataset, the crossbreeds are still natural-looking and carries the desired backgrounds and objects.

3.4 Interpolation in Attribute Subspace

Figure 8 gives interpolation of object features in attribute subspace. Note the backgrounds (human identities) are approximately the same, while objects (hair styles) are interpolated. Let the features of the four corner images be $A_{11}u_{11}$, $A_{12}u_{12}$, $A_{21}u_{21}$ and $A_{22}u_{22}$, novel images are generated from the combination of the unchanged background B and bi-linearly interpolated object features:

$$\frac{u_{11}(x_2-x)(y_2-y)+u_{21}(x-x_1)(y_2-y)+u_{12}(x_2-x)(y-y_1)+u_{22}(x-x_1)(y-y_1)}{(x_2-x_1)(y_2-y_1)} \tag{6}$$

where (x,y) is the coordinate in the diagram.

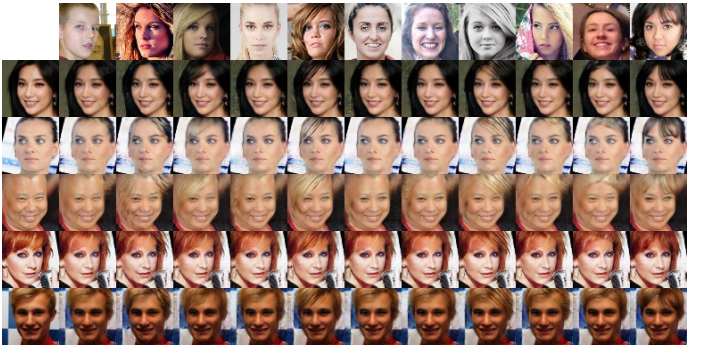


Figure 7: Performance of GeneGAN on unseen data from Wider Face [29].

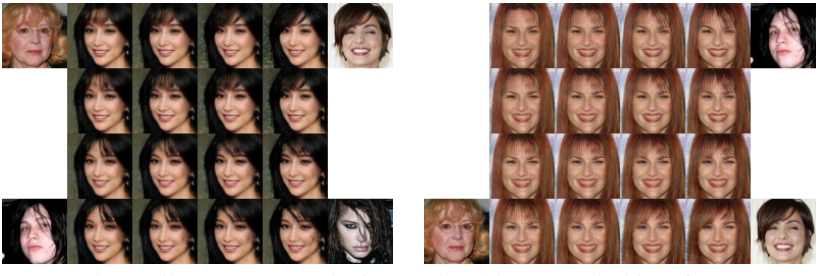


Figure 8: The attribute space can be interpolated by several object feature vectors.

4 Conclusion

In this paper, we propose GeneGAN, a deterministic conditional generative model that can perform object transfiguration task. The proposed model learns to disentangle the object from background in feature space, by learning from the labels of presence/non-presence of objects in unpaired training data. Consequently, our model can extract an object feature vector from a single image and transplant it to another image, hence allows fine-grained control of generated images. The objects can be abstract and difficult to characterize, like hair styles and lighting conditions. The training method for our model is symmetric and allows exploiting the reconstruction loss, which improves stability of training. The setup of our model also gives rise to an attribute subspace, which contains multiple vectors that are representatives of different objects. The vectors can be scaled, inverted and interpolated to manipulate the objects in generated images.

As future work, it would be interesting to investigate whether more complex crossbreeding patterns between more parents would allow further improvements of stability of training, quality and diversity of generated images.

References

[1] Jacob R. Gardner, Matt J. Kusner, Yixuan Li, Paul Upchurch, Kilian Q. Weinberger, and John E. Hopcroft. Deep manifold traversal: Changing labels with convolutional features. *CoRR*, abs/1511.06421, 2015.

[2] Andrew Brock, Theodore Lim, James M. Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. *CoRR*, abs/1609.07093, 2016.

[3] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Álvarez. Invertible conditional gans for image editing. *CoRR*, abs/1611.06355, 2016.

[4] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 597–613, 2016.

[5] Paul Upchurch, Jacob R. Gardner, Kavita Bala, Robert Pless, Noah Snavely, and Kilian Q. Weinberger. Deep feature interpolation for image content changes. *CoRR*, abs/1611.05507, 2016.

[6] Weidong Yin, Yanwei Fu, Leonid Sigal, and Xiangyang Xue. Semi-latent gan: Learning to generate and modify facial images from attributes. *CoRR*, abs/1704.02166, 2017.

[7] Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. *arXiv preprint arXiv:1411.5928*, 2014.

[8] Tejas D. Kulkarni, William F. Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2539–2547, 2015.

[9] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 776–791, 2016.

[10] Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 552–560, 2013.

[11] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[12] Scott E. Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1252–1260, 2015.

[13] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 820–828, 2016.

[14] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *CoRR*, abs/1703.05192, 2017.

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.

[16] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *CoRR*, abs/1704.02510, 2017.

[17] Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. Transforming auto-encoders. In *Artificial Neural Networks and Machine Learning - ICANN 2011 - 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I*, pages 44–51, 2011.

[18] Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3581–3589, 2014.

[19] Brian Cheung, Jesse A. Livezey, Arjun K. Bansal, and Bruno A. Olshausen. Discovering hidden factors of variation in deep networks. *CoRR*, abs/1412.6583, 2014.

[20] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. *CoRR*, abs/1511.00830, 2015.

[21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[22] Ian J. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *CoRR*, abs/1701.00160, 2017.

[23] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1139–1147, 2013.

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[25] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013.

[26] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014.

[27] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1096–1104, 2016.

[28] Stephen Moore and Richard Bowden. Multi-view pose and facial expression recognition. In *Proc. BMVC*, volume 2, 2010.

[29] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A face detection benchmark. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5525–5533, 2016.