Face Generation for Low-shot Learning using Generative Adversarial Networks

Junsuk Choe* Song Park* Kyungmin Kim* Joo Hyun Park* Dongseob Kim* Hyunjung Shim
Yonsei University
{skykite, song.park, kyungmin.kim, pwangjoo, kou.k, kateshim} @yonsei.ac.kr

Abstract

Recently, low-shot learning has been proposed for handling the lack of training data in machine learning. Despite of the importance of this issue, relatively less efforts have been made to study this problem. In this paper, we aim to increase the size of training dataset in various ways to improve the accuracy and robustness of face recognition. In detail, we adapt a generator from the Generative Adversarial Network (GAN) to increase the size of training dataset, which includes a base set, a widely available dataset, and a novel set, a given limited dataset, while adopting transfer learning as a backend. Based on extensive experimental study, we conduct the analysis on various data augmentation methods, observing how each affects the identification accuracy. Finally, we conclude that the proposed algorithm for generating faces is effective in improving the identification accuracy and coverage at the precision of 99% using both the base and novel set.

1. Introduction

In recent years, face recognition performance has been greatly improved thanks to powerful machine learning techniques such as Convolutional Neural Networks (CNN). The machine's capability for face recognition reaches even to the point where they perform better than humans under several circumstances. From various studies, there has been a wide agreement that a large amount of data brings out the success of CNN based learning algorithms. However, such massive dataset is not always available in practice. To handle the lack of data, researchers have attempted to either generate more data to increase the size of dataset, or develop an discriminative feature extraction algorithm that understands the target even with extremely limited dataset.

To this end, previous studies using CNN architectures have tried to augment data through modifying images by rotating, flipping, adding noise, or jittering color [14, 21, 11]. However, how the performance changes upon different methods was not thoroughly analyzed. Under the same philosophy, with more variations, generative approaches were popularly employed for various recognition tasks [18, 25, 6, 5, 10]. Although these methods succeeded in generating larger version of the original dataset, since their achievements were in stabilizing the training process, the main issue of data shortage remained unsolved.

In order to solve such limitation, low-shot learning, which suggests training with few images, have been issued. Low-shot learning is currently being actively studied to tackle the limited dataset problem in recognition by mimicking the process of a human brain, and hence getting even closer to the scenarios of real world applications; humans can recognize after learning only a few images of a target, and sometimes can even perceptually understand without learning. In this paper, we focus on face recognition, one of the most popular and interesting objects.

In our paper, we use transfer learning suggested by Thrun [22] to train a model with a large set (i.e. a base set) to build a feature extractor. Then, we fine tune to learn actual label of the target with a new, small image dataset (i.e. a novel set) that does not overlap with the base set. Using transfer learning as a backend, we implement various type of data generations to increase both base and novel set to achieve higher accuracy in recognition. For image generation, we present a training model that combines transfer learning and data generation to a face recognition task. Figure 1 shows the overview of our algorithm. In Section 3, we show how we employed the variants of GAN [7, 2] to generate faces under various poses and attributes. Utilizing the capability of GAN in generalization aspects, we use a base set to train the GAN in order to build a face generator. In Section 4, we analyze face identification results using every different combination for increasing dataset. For the dataset, our base set relies on MS-Celeb-1M Challenge-2 [9, 8] and CelebA [16].

[*] These authors contributed equally to this work

2473-9944/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCVW.2017.229

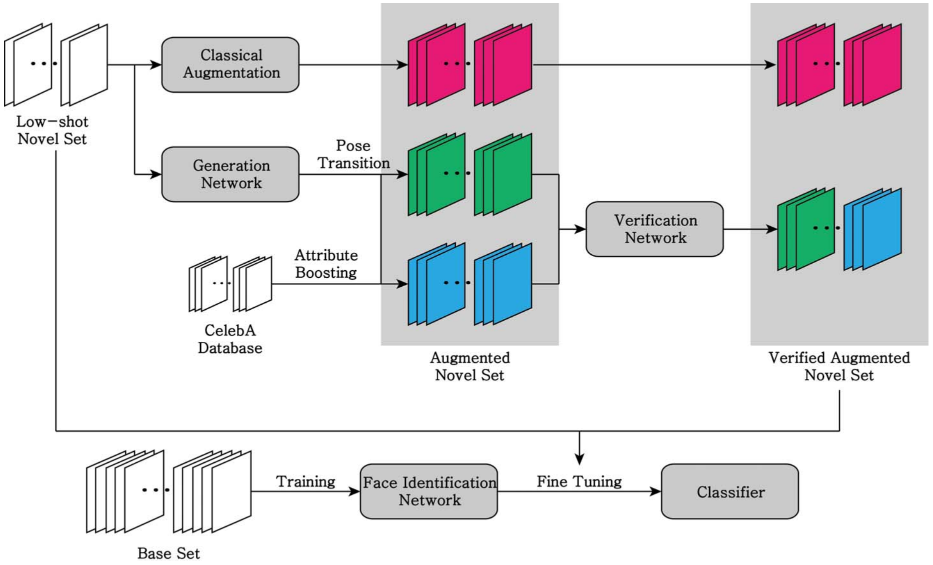


Figure 1: Overview of proposed algorithm.

2. Related Work

Low-shot learning process can be categorized into three groups: 1) discriminative approach which learns powerful features to improve the recognition accuracy, 2) generative approach which augments or generates using given data to enlarge the dataset, and 3) a network structural approach which builds a novel classifier. In this paper, we focus on a generative approach.

Data augmentation and generation The most naive method in data generation is the augmentation via simple image transformation. This method uses image translation, rotation, noise addition, and color jittering to transform original images for creating additional data, and provide prevention to overfitting that can occur within an interclass. These methods are jointly used in the literatures [14, 21, 11].Our model also includes these conventional schemes to increase robustness, and to achieve higher scores.

However, these methods generate duplicated version of the original data; hence the dataset still lacks intra-class variations. To handle this issue, there have been attempts to refine data to transform the original data into features, so that the variation of the dataset is increased. For example, in order to train a network which labels buildings from 1-shot image, Movshotvitz-Attias et al. [17] extracted text information from the fascia and used it as a textual guide when buildings in different categories have similar appearances. Similarly, Park and Ramanan [18] extracted pairs of body parts from an image and used it as a new representation for learning in order to train a motion recognition network. Because there are usually more than one pair of body parts, it resulted in enlarging dataset proportional to the number of objects in a scene. Unlike those of augmenting the data, Zhu et al. [25] recently addressed that well defined data representation could yield better performance than adding more data into dataset.

Despite their achievement, these methods are not a suitable way of enlarging dataset for other types of task. For example, the idea from [18] is not applicable for our problem, because same faces do not appear multiple times in face recognition. Hence, in our model, we suggest using a generative approach using a Generative Adversarial Network (GAN), for which it can be used specifically for face identification task, by generating new facial images.

Generative adversarial network A Generative Adversarial Network (GAN) proposed by Goodfellow et al. [7] is a powerful model for modeling and generating images in an unsupervised manner. Their network is composed of generator and discriminator that works adversarially to learn the generalization process of the given dataset. Upon their success in image generation, several existing studies use [7] to increase the size of dataset for classifications. One is from Denton et al. [5], who generated enlarged output image from interpolating reconstructed latent variables with GAN generator using a Laplacian pyramid. However, despite such immense outcome of GAN, slight difference in parameters and structure choice were likely to result in irregular, poor outcome, making it hard to train in purpose of real world applications.

Despite such limitation, due to its strength in generation, the variations of GAN have been actively studied. The most popular one is the Deep Convolutional GAN (DCGAN) introduced by Radford et al. [19]. DCGAN improved the overall quality of generated images by adapting Convolutional Neural Network (CNN) into GAN architecture, and

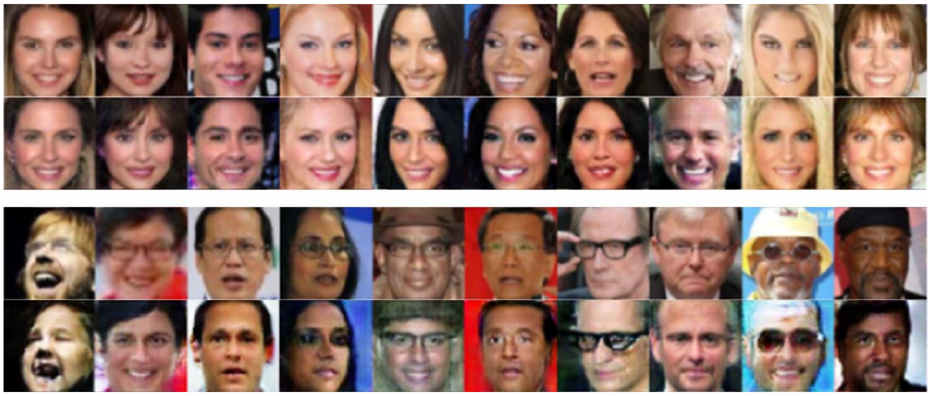


Figure 2: The reconstructed faces using VAE/GAN. Top row: Input faces. Bottom row: Corresponding reconstructed faces. The inputs and outputs of VAE/GAN do not look like same people.

also provided a guideline of building a GAN network, making it easier to choose between parameters. Larsen et al. [15], at the same time, developed VAE/GAN by adapting Variational Autoencoder (VAE) [12] into GAN to maximize the advantages of both models. They used discriminator of the GAN for computing feature-wise loss for training decoder of VAE. This feature-wise loss improved the performance significantly compared to the one with pixel-wise loss.

However, it was not sufficient to facilitate the training of GAN due to the lack of convergence measure. Arjovsky et al. [1] then introduced Wasserstein GAN (WGAN) that suggested a specific measure, which opened the possibility for actual measurement of the convergence. Despite such improvement, due to the instability in balance between generator and discriminator, the generated image quality did not improve as much.

Recently, BEGAN was proposed by Berthelot et al. [2] to tackle the issue of imbalance of generator and discriminator by introducing equilibrium factor to the model. Their efforts significantly improved the performance of image generation. They show their result on faces, successfully presenting its smooth and sharp rebuilt facial images. We also employ BEGAN into our model as a mean to generate based on low-shot dataset.

Transfer learning Transfer learning [22], first introduced by Thrun, utilizes knowledge extracted and saved from a specific topic to help the learning on the other similar field. This concept was carefully scrutinized as a solution to learning with limited dataset. First several studies applied this idea to train the CNN. Bertinetto et al. [3] suggested a second deep network after a network, called a *learnet* to predict parameters in a previous network. They argued that it is applicable in 1-shot learning, however, the result was not satisfactory. Meanwhile, Wang and Herbert [24] trained a paired network, one that learns few samples with annotated, categorized few dataset, and one with larger samples. With a premise that transform is a type of regression, they learned a transform function with a multilayered regression neural network. Similarly, Vinyals et al. [23] trained a matching network, but with small labeled support set and large unlabeled dataset. These ideas were successful in terms of transferring the knowledge to other dataset, but was limited in terms of actual result.

Combining the idea of transfer learning and [25], some researchers added a feature extractor to the commercially used networks such as CNN. Koch et al. [13] connected two convolutional networks, one network trained as a feature extractor, and the other network for 1-shot classification. Similarly, Schroff et al. [20] trained Euclidean embedding CNN with CNN bottleneck. They both argue that efficiency was largely improved with this method.

More closely related work was done by Hariharan and Girshick [10], as they trained a network with parts of ImageNet [4] as a base set to use them as a feature extractor, while training classifier with the rest of ImageNet as a novel set to achieve low-shot learning. They separated the model into procedures of representation learning phase and low-shot learning phase. In addition to this, they trained MLP within the relationship between example images to get hallucinated features, which they use for extra information. With such method, the low-shot learning result has significantly improved. Recently, Guo and Zhang [8] discovered that the magnitude of decision weight vector of novel set is biased and smaller than that of the base set. They introduced the new loss term to compensate the bias of weight vector distribution, and showed that their proposal was effective to improve the accuracy for one-shot face recognition.

Our idea builds upon the same backend, but differs in a way that we create additional images possessing various attributes using BEGAN trained with the base set.

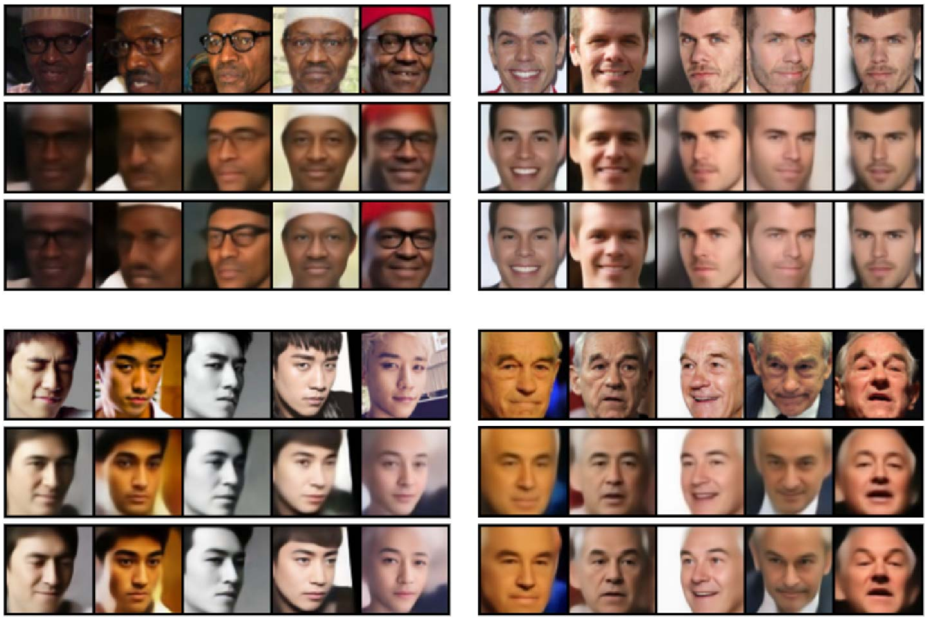


Figure 3: The reconstructed faces using BEGAN. Top row: Input images. Middle row: Outputs from BEGAN without skip connection. Bottom row: Generated images from BEGAN with skip connection. Outputs with Skip connection are more identical with input images.

3. Face generation for face identification

In this section, we describe details of our face identification framework. We first review a generation network and face identification network, then introduce our data augmentation methods and propose our own network training scheme.

3.1. Generation network

We employ VAE/GAN [15] and BEGAN [2] for generation networks. In VAE/GAN, learned similar metric replaces pixel-wise similarity used in original GAN architecture. The learned similar metric measures the distance between features that is extracted from the discriminator. We refer to [15] for further detail and structure of VAE/GAN. In BEGAN, auto-encoder is used as a discriminator and its loss is used as a reconstruction error. Unlike existing variants of GAN, the loss function of BEGAN includes an equilibrium term to balance the loss of discriminator and the generator. To preserve high frequency components in reconstructed image, BEGAN introduce skip connections to facilitate gradient propagation. Since skip connections is effective to produce sharper images, we use skip connection to preserve the facial identity. For details, refer to [2] for the network structure and detail about BEGAN. The generated faces are shown in Figure 2 and 3.

3.2. Face identification network

ResNet [11] is a widely used model in image classification task. The outstanding performance of deep ResNet was first demonstrated in *ILSVRC 2015 competitions*. The success of ResNet architecture was achieved by very deep layers with skip connections and a large amount of data. Under low-shot scenario, however, the amount of data is extremely small. Considering the limited size of training dataset, we chose simple ResNet-10 as our face identification network to avoid overfitting and attain the efficiency of training. ResNet-10 is comprised of one convolutional layer, four layers that consists two residual block and one fully connected layer. Additionally, the architecture of our verification network is identical to the face identification network.

3.3. Data augmentation

Data augmentation is widely used to improve accuracy and stabilize the network training in image classification. We increase low-shot data with classical augmentation methods and the two augmentation methods with generation network, pose transition and attribute boosting, as well. We analyze and demonstrate the effect of several augmentation methods for face identification. Details of the implementation will be explained in section 4.

Classical data augmentation Simple typical augmentation methods can make network more robust.(e.g. flipping, noise addition) The effectiveness of classical data augmentation is empirically shown in several previous studies

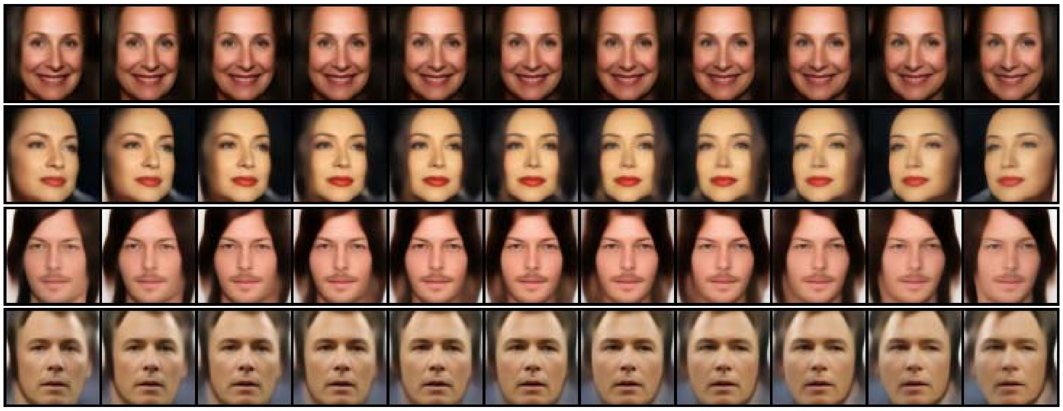


Figure 4: The results of pose transition using BEGAN. First and last columns: Single reconstruction results. Rest columns: Pose transition results. Latent values of these intermediate images are combination of latent value of two end columns. Pose is smoothly changed one from another depending on proportion of latent variable.

[14, 21, 11].

Pose transition using GAN In data augmentation using GAN, we extract latent variables of images from discriminator of GAN and arithmetically modify it with semantic latent variables. For pose transition, we extract latent variables of a image and its horizontal flipped version and linearly interpolate two latent variables. We expect interpolation generates continuous pose transition between two images, while preserving the identity of faces. It is important to note that pose transition does not require extra dataset.

Attribute boosting using GAN In order to achieve attribute boosting, we utilize the extra dataset, CelebA [16], presenting various attributes. Detailed information about CelebA database is described in Table 1. We compute average of latent for attribute set and entire set. We assume difference between two latent values represents corresponding attribute. To avoid gender bias, we split entire set into female and male set and calculate the average of the each gender set. Then we obtain weighted average of two gender average with gender ratio of specific attribute set. We compute 40 latent variables that equivalent to 40 attribute in CelebA database. We anticipate that addition of these latent variables induces manifestation of attribute in image space.

3.4. Network training

Training scheme was inspired by [10], a two phase learning strategy. Our face identification network is pre-trained with base set and then fine-tuned with base set, novel set and our augmented novel set. However, Based on our empirical observation, face identification performance can be degraded with several augmented data. Several images through augmentation rarely retains the identity information,especially those obtained by our generation network. Therefore, we use augmented data if it succeeds in identification using verification network. Verification network is pre-trained with base set and fine-tuned with novel set to recognize face in novel set. We pick out a subset of the augmented data with verification network. Finally, face identification network fine-tuned with original base set, novel set and a qualified augmented novel set as we explained above.

(a) MSCeleb Database [9, 8]

	Subjects	Training set	Test set
Base set	20,000	50 - 100	5
Novel set	1,000	5	20

(b) CelebA Database [16]

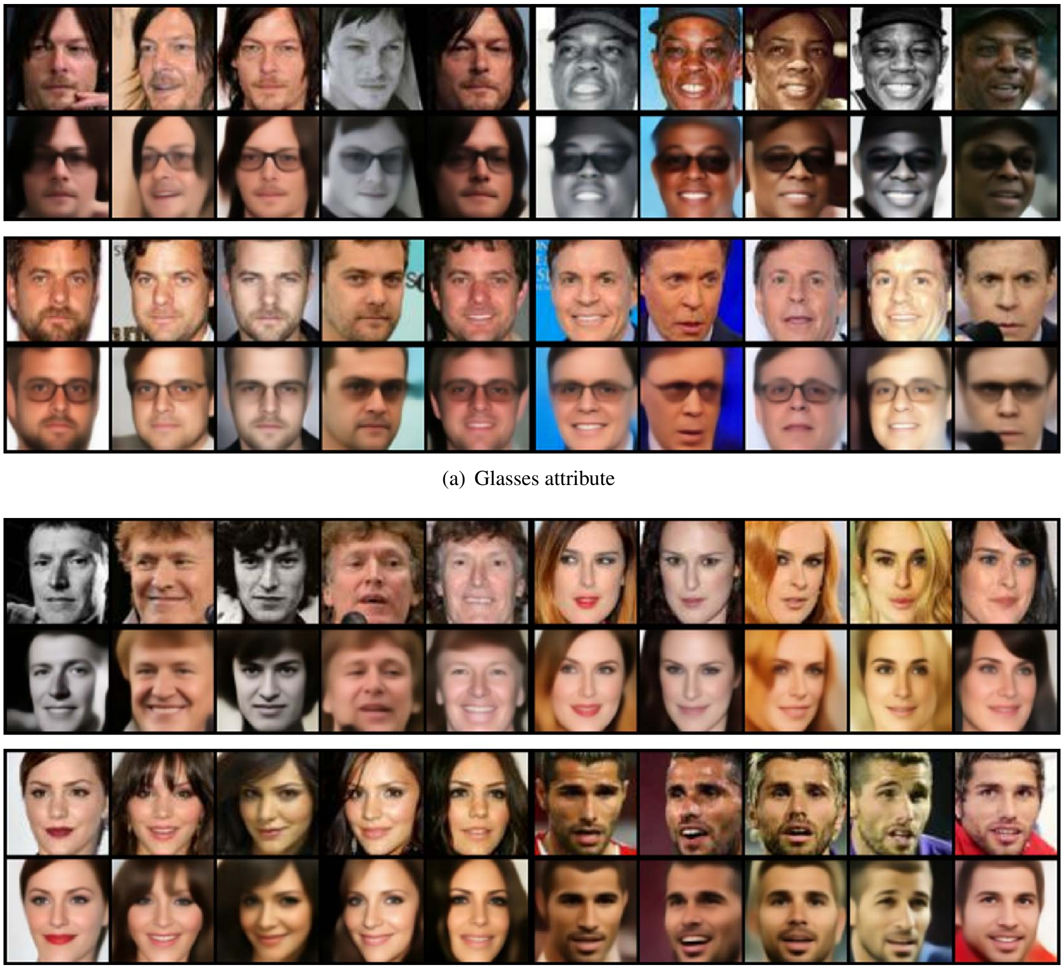
	Subjects	Attributes	Total
CelebA	10,177	40	202,599

Table 1: Dataset information.

4. Experimental results

4.1. Datasets

We use the dataset of *MS-Celeb-1M Challenge-2: Low-shot learning* [9, 8] for training face identification network. This dataset is divided into two sets (i.e., base set and novel set). Detailed description of dataset is shown in Table 1. It is important to note that there is no overlap between base set and novel set. Our goal is to identify individuals in novel set, while keeping the recognition accuracy using base set. In attribute boosting using GAN, additional dataset, CelebA, is utilized to extract attribute variables. CelebA contains more than 200k images with 40 attribute annotations for about 10k identities. Entire database is used to extract attributes values. Note that, this extra dataset is used only for attribute boosting.



(a) Glasses attribute

(b) Smile attribute

Figure 5: The sample results of attribute boosting using BEGAN. (a) Glasses attribute boosting results. (b) Smile attribute boosting results. Top row: Input faces. Bottom row: Reconstructed faces with target attributes. Semantic attribute addition in latent space modify image while preserving identity.

4.2. Verification network

Verification network selects proper images among the augmented novel set generated by BEGAN. Verification network uses the architecture of [10], the slightly modified version of [11] for low-shot learning. We set the batch size to be 100, the training iteration to be 180K, and the initial learning rate to be 0.001, which changes by 0.1 times for every 50000 iterates. This model is trained on NVIDIA Geforce GTX 1080 Ti or NVIDIA Tesla K80 with 8GB GPU memory. Training times are about 8 hours for pre-training and 2 hours for fine-tuning. As we explained above, we pre-train our verification network with base set and then fine-tune with original given novel set to recognize face in novel set.

4.3. Face generation using GAN

Reconstruction We reconstruct face images using VAE/GAN and BEGAN, respectively. As shown in Figure 2, we can observe the discrepancy between the reconstructed faces from VAE/GAN and original input faces. Especially, some results does not look like human faces. Meanwhile, as shown in Figure 3, BEGAN produces better results, more human-like faces, and also the reconstructed faces more similar to original input faces. As a result, we decided to choose BEGAN for the face generation model. To enhance the detail in reconstructed faces, we add skip connections in decoder and generator of BEGAN as we mentioned in Section 3. Note that all the generated faces for identification in this paper are synthesized using BEGAN with skip connection.

Pose transition We expect that interpolating a latent pair

		Verification rate (%)
BEGAN without SC	pose transition	10
	attribute boosting	10
BEGAN with SC	pose transition	31
	attribute boosting	33

Table 2: Data verification results. Skip connection is denoted by SC. Verification networks verify augmented data produced by BEGAN. Results validate that generation network with skip connection produce more qualified images, because it preserves identity better.

from a image and its mirrored version produces continuous pose transition between two poses, while preserving identity. Each image pair produce 11 additional images. We conduct pose transition for 1-shot and 3-shot experiments. The total number of generated additional novel set images are 11,000 and 33,000 for each experiments. Results of the pose transition are shown in Figure 4. Resulted images smoothly changes one to another and strikes a new pose.

Attribute boosting We expect addition of attributes latent in latent space turn into appearance of attribute in image space. We randomly choose test images and extract latent variable of them with discriminator. We add specific attribute latent to latent of test images to conduct attribute boosting experiment. Some attribute boosting results are shown in Figure 5. Note that, alignment between image sets significantly influences generation results. Therefore, we align CelebA database with base set and novel set, according to the location of eye.

Data verification Data augmentation, especially in case of using generation networks, occasionally harms identity of images. For instance, some attributes addition in CelebA dataset cause severe damage to identity. (e.g. beard, cosmetics) Instead of hand crafted selection of appropriate attribute or proper generated results, we verify augmented data from generation networks using verification network. We train verification network to recognize original novel set. Inappropriate, failed to recognize, augmented images are filtered out from our additional augmented novel set. Table 2 shows the verification success rate of generated images in 1-shot experiments. The number of recognizable images generated by BEGAN with skip connection are bigger than BEGAN without skip connection. This was expected, because BEGAN with skip connection preserves the face identity better. Furthermore, we effectively verify proper data using verification network without hand labeling.

Aggregating those data accepted by verification network, we build our final low-shot dataset: classical augmented dataset, verified pose transition, attribute boosted dataset, and original novel set are combined to form the final dataset.

4.4. Data pre-processing

Before the training phase, every data is pre-processed to fit into our face identification network. We cropped rectangle images into a square with the sides size of shorter side of the original image. Images were cropped on the center. Then, we roughly align images based on eye location and resize images into 224x224x3.

4.5. Face identification

Our final face identification network identify individuals in novel set, while keeping recognizable subject in base set. Like verification network, face identification network utilizes architecture and training strategy of [10]. It is important to note that, the only difference between two networks is fine-tuning data. Face identification network is pre-trained with based set and then fine-tuned with base set, novel set and our additional augmented novel set. For evaluation, we measure accuracy, precision and coverage. Suppose that N images are available in the test set, M images are evaluated, and C images are recognized correctly. Then, the accuracy, precision and coverage are defined as C/N, C/M and M/N. The coverage is reported when the precision is at 99%.

The face identification results are shown in Table 3. We gradually increase novel set with proposed augmentation methods and investigate the changes of identification accuracy. Because the amount of base set is significantly greater than that of novel set, baseline network trained with original base set and novel set rarely identifies novel set. It is because the importance of individual is proportional to the number of training images. Therefore, the accuracy and coverage of novel set are increased by adding augmented data. From these experiments, we observe that both augmentation methods, classical augmentation and face generation using GAN, improve identification performance. Considering its importance in performance, pose transition and attribute boosting using BEGAN are more effective in improving performance than the classical method. Notice that accuracy improves from 17.40% to 40.40% for 1-shot scenario and from 57.42% to 80.45% for 3-shot scenario. Likewise, coverage at 99% precision is advanced from 0.02% to 1.75% for 1-shot scenario and from 0.22% to 52.07% in 3-shot scenario. This results demonstrate that face generation with GAN certainly complement lack of intra-class variation.

Note that, face identification accuracy for base set is constantly above 95% whereas accuracy and coverage for identifying novel set is increased thanks to our augmented data. As expected, accuracy for 3-shot experiment came out to be better than 1-shot experiment due to larger dataset. However, the amount of additional generated data is considerably bigger than original real novel set. To prevent generated data dominating the training process, we just duplicate

Training method	Test set	1-shot		3-shot	
		Accuracy(%)	Coverage(%)@P99	Accuracy(%)	Coverage(%)@P99
BS+NS	base set	98.70	99.00	98.68	98.95
	novel set	0.00	0.02	0.00	0.02
BS+NS+CS	base set	98.43	98.97	98.08	98.42
	novel set	17.40	0.02	57.42	0.22
BS+NS+CS+GS	base set	97.00	94.47	95.90	89.50
	novel set	**40.40**	**1.75**	**80.45**	**52.07**

Table 3: Face identification results. In this table, base set, novel set, classical augmented set and generated set using GAN are denoted by BS, NS, CS, and GS each. The best results for novel set are in bold.

Training method	1-shot	
	Accuracy(%)	Coverage(%)@P99
NS + pose(1:11)	29.37	0.0095
+ V pose(1:3)	30.20	1.95
+ 1:1	39.80	3.20
+ 10:1	49.53	**10.65**
+ 20:1	**50.72**	10.12
NS + attr(1:40)	24.55	0.40
+ V attr(1:13)	25.45	0.85
+ 1:1	39.35	6.42
+ 10:1	49.57	7.60
+ 20:1	**49.60**	**11.17**

Table 4: Result of manipulating the ratio (real:fake) of augmented data. Novel set, attribute and verified datasets using verification network are denoted by NS, attr and V each. The best results are highlighted in bold.

real image multiple times and manipulate the ratio between real images and fake images. We observed the positive effect of setting the ratio of the data by increasing the amount of real data. For the computational efficiency, we excluded the base set from fine-tuning data. Table 4 shows the verification network successfully selects useful data to improve accuracy. Also, a ratio of real and fake images greatly influences identification performance. Based on Table 4, experiments in Table 3 are conducted with 20:1 ratio.

5. Conclusion

Image generation techniques are attractive in various computer vision applications, especially because of the difficulties of labeled data collection for training. Among those, we attempted to generate face images with several attributes and poses using GAN, enlarging the novel set to achieve increased performance on low-shot face recognition task. With the increased dataset, we verified increased performance on low-shot recognition using ResNet-10. Moreover, we demonstrate that duplicating original data effectively regularized excessive influences from augmented data.

While the latent variable used in this work is effective for data generation, this may not be optimal for discrimination. In the future, we plan to study the latent representation of faces for face identification purpose.

Acknowledgement

This work was supported by ICT R&D program of MSIP/IITP. [R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding] and also supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the University Information Technology Research Center support program (IITP-2016-R2718-16-0014) supervised by the IITP (Institute for Information & communication Technology Promotion).

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3

[2] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 1, 3, 4

[3] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi. Learning feed-forward one-shot learners. In *Advances in Neural Information Processing Systems*, pages 523–531, 2016. 3

[4] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, June 2009. 3

[5] E. L. Denton, S. Chintala, a. szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015. 1, 2

[6] A. Dosovitskiy, J. T. Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1538–1546, June 2015. 1

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2

[8] Y. Guo and L. Zhang. One-shot face recognition by promoting underrepresented classes. *arXiv preprint arXiv:1707.05574*, 2017. 1, 3, 5

[9] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *European Conference on Computer Vision*. Springer, 2016. 1, 5

[10] B. Hariharan and R. Girshick. Low-shot visual recognition by shrinking and hallucinating features. *arXiv preprint arXiv:1606.02819*, 2016. 1, 3, 5, 6, 7

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 1, 2, 4, 5, 6

[12] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[13] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015. 3

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 2, 5

[15] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. 3, 4

[16] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, Dec 2015. 1, 5

[17] Y. Movshovitz-Attias, Q. Yu, M. C. Stumpe, V. Shet, S. Arnoud, and L. Yatziv. Ontological supervision for fine grained classification of street view storefronts. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1693–1702, June 2015. 2

[18] D. Park and D. Ramanan. Articulated pose estimation with tiny synthetic videos. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 58–66, June 2015. 1, 2

[19] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[20] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015. 3

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2, 5

[22] S. Thrun. Is learning the n-th thing any easier than learning the first? In *Advances in neural information processing systems*, pages 640–646, 1996. 1, 3

[23] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016. 3

[24] Y. Wang and M. Hebert. Learning to learn: Model regression networks for easy small sample learning. In *European Conference on Computer Vision (ECCV)*, October 2016. 3

[25] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan. Do we need more training data? *International Journal of Computer Vision*, 119(1):76–92, 2016. 1, 2, 3