# EDGE-GAN: EDGE CONDITIONED MULTI-VIEW FACE IMAGE GENERATION

*Heqing Zou[1], Kenan E. Ak[2], Ashraf A. Kassim[1]*

[1]National University of Singapore, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore

## ABSTRACT

Reconstructing photorealistic multi-view images from an image with an arbitrary view has a wide range of applications in the field of face generation. However, most current pixel-based generation models cannot generate sufficiently realistic enough images. To address this problem, we propose an edge-conditioned multi-view image generation model called Edge-GAN. Edge-GAN utilizes edge information to guide the image generation based on the perspective of the target view while the details of the input image are used to influence the target image. Edge-GAN combines the input image with the target pose information to generate a coarse image with an approximate target outline which is then refined to a better quality using adversarial training. Experiments conducted show that our Edge-GAN is able to generate high-quality images of people with convincing details.

***Index Terms***— Image generation, edge-conditioned, GAN, multi-view, face synthesis

## 1. INTRODUCTION

Human-based multi-view image generation is a hot topic in computer vision and has good potential for commercial uses. Recently, several style/pose relevant models have been proposed for pose generation [23, 14, 15], which generally synthesize new images based on input poses followed up with a refinement step on the generated images. In the field of face generation, related works include attribute-related face generation [2] and spatially varying face synthesis [22].

In Generative Adversarial Networks (GANs) -based models, introduced by Goodfellow et al. [6], images are generated in a coarse-to-fine fashion [3]. The generator and the discriminator are made to work against each other, with the generator being trained to generate images that try to confuse the discriminator, while the discriminator is used to distinguish fake samples from real ones. Conditional GANs [16] achieve comparative results in image translation, such as pix2pix [8] and cycleGAN [24].

There are two major GAN-based approaches for the spatial transformation of faces. One relies on spatial rotation to perform new image synthesis by learning the similarities and differences of face characteristics between input and target images [8, 21]. TP-GAN [8], for example, utilizes two pathways to model out-of-plane rotation and non-linear transformation of the local texture. However, TP-GAN is not able to generate images with side views.


**Figure 1**: Photorealistic side view (45°, -45°) and frontal view (0°) images generated by EdgeGAN conditioned on input RaFD frontal and side view images, respectively.

Another approach is based on conditioned inputs. For example, [20] introduces a framework for meta-learning of adversarial generative models which is based on learning the characteristics of faces with landmarks. However, this approach cannot be directly applied to different people, due to the mismatch of detailed landmarks. Variations of GANs have also been successfully applied for image synthesis [10,11].

Our proposed Edge-GAN also takes the input image as the conditioned information to generate new images and uses an inverse mapping like cycleGAN and BiGANs [4]. The Edge-GAN is designed to synthesize new frontal face images using random side-view images as well as generate different side-view images from other views, as shown in Figure 1. The input information used in Edge-GAN comprises two parts; the content image and the edge image which are obtained from the image set with the target pose. The detailed information on the target image is determined by the details in the input content image. The extracted edge image of the input is used to shape the generated face image with the target view.

The unique approach in using edge information to shape the generated images with the target pose enables Edge-GAN to synthesize images with better-generated details even for "unseen" regions. Compared with other studies, Edge-GAN can obtain better synthesis results when using high-angle images. And images from the Radboud Faces Database (RaFD) [13] and the CMU Multi-PIE Face Database (Multi-PIE) [7] are used in this work.

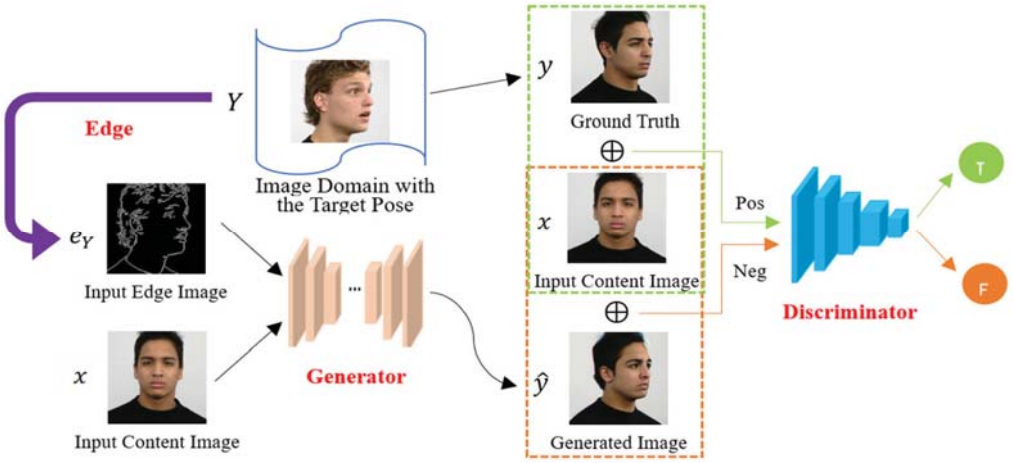



**Figure 2**: General framework of Edge-GAN. The input information for Edge-GAN is the combination of the edge image randomly obtained from the target image set and the content image. The network works in an adversarial way to refine the output image.

## 2. RELATED WORK

### 2.1 Pose Representation Learning

Ranzato et al. [18] is the first to use the encoder-decoder structure for representation learning. Recently, different deep learning models have been developed to determine the pose and identity of objects. The methods presented in [26, 17] use reconstruction loss to synthesize images and disentangle pose and identity factors whereas in [26] information is transferred from pose variant inputs to a fractalized appearance. VariGANs [23] and PoseGAN [14] use images conditioned with the target style and the target pose to generate the target image, both methods divide the generating process into two steps: coarse image generation and image refinement.

### 2.2 Multi-view Synthesis

StarGAN [2] is the first to use a single generator network to synthesize images from multi-domains where the generated results need to keep the inherent feature information unchanged. DR-GAN [20] follows a single-pathway framework to learn identity features that are invariant to a viewpoint. Both TP-GAN and CR-GAN [19] utilize a framework that comprises two pathways. TP-GAN uses two distinct encoder-decoder networks to model the out-of-plane rotation of the global structure and the non-linear transformation of the local texture. However, CR-GAN shares all modules in the two pathways and concentrates on learning complete representations in multi-view generation, guaranteeing high-quality generations for "unseen" inputs.

## 3. EDGE-GAN

Canny [1] introduced a universal multi-stage algorithm to detect varieties of edges from images. Our proposed Edge-GAN first uses edge information as a condition to control the target image generation where the model is fed with a combination of the edge image $e_Y$ and the content image $x$. Using these inputs, the blurry image with uncertain details is preliminarily produced by Edge-GAN as shown in Figure 2.

**Adversarial loss.** We divide the adversarial loss [3] into two parts: generator loss and discriminator loss which are defined as follows:

$$\mathcal{L}_{adv} = \mathcal{L}_{adv_{gen}} + \mathcal{L}_{adv_{dis}} \quad (1)$$

where $\mathcal{L}_{adv_{gen}}$ and $\mathcal{L}_{adv_{dis}}$ represents the generator loss and discriminator loss, respectively.

Generator loss is used when we generate a new sample and want to minimize the gap between the generated and the target images. Similar to the generating process in [2], we build a two-way generative network and express the generator loss as:

$$\mathcal{L}_{adv_{gen}}(D_X, D_Y, x, y, e_X, e_Y) = \mathbb{E}_y\left[\log(1 - D_X(F(y, e_X))\right] + \mathbb{E}_x\left[\log(1 - D_Y(G(x, e_Y))\right] \quad (2)$$

where $x$, $y$ and $e_X$, $e_Y$ are the input images of content information and random edge images from the X domain and Y domain respectively. $G(x, e_Y)$, $F(y, e_X)$ are the generators which try to generate samples to confuse the discriminators $D_X$ and $D_Y$, using different weights in the training networks.

Discriminator loss is used for distinguishing generated results with the ground truth.

$$\mathcal{L}_{adv_{dis}}(D_X, D_Y, x, y, e_X, e_Y) = \mathbb{E}_y[\log D_Y(y, e_X)] + \mathbb{E}_x[\log D_X(x, e_Y)] \quad (3)$$

**Reconstruction Loss.** Like the cycle consistency loss proposed in [26], we introduce a reconstruction loss to determine the generation direction, guaranteeing that the model can map an individual input to the desired output.

$$\mathcal{L}_{rec}(G, F, x, y, e_X, e_Y) = \mathbb{E}_x[\|F(G(x, e_Y), e_X) - x\|_1] + \mathbb{E}_y[\|G(F(y, e_X), e_Y) - y\|_1] \quad (4)$$

(a) (b)

**Figure 3**: **(a)** Multi-view results generated by Edge-GAN using some specific poses. The images in the left column are the input conditioned content images. And the target generated images in the 1st, 3rd and 5th row are a series images with different poses (-90°, -60°, -30°, -15°, 15°, 30°, 60° and 90°). The images in the 2nd, 4th and 6th rows are the corresponding ground truth images. **(b)** General frontal results generated by Edge-GAN using different poses. The images in the 1st, 3rd and 5th rows are the input conditioned content images of faces with different views, -90°, -60°, -30°, -15°, 15°, 30°, 60° and 90°. The images in the 2nd, 4th and 6th row are the corresponding generated results. The corresponding frontal ground truth images are shown in the last column.

**L1 distance.** L1 distance explored in [9] is also preferred here as it encourages less blurring. The following loss function is used for comparing the generated samples with the target images.

$$\mathcal{L}_{L1} = \mathbb{E}_y[\|y - G(x, e_Y)\|_1] + \mathbb{E}_x[\|x - F(y, e_X)\|_1] \quad (5)$$

**Full Objective.** Finally, the objective functions to optimize G and D are written respectively as follows:

$$\mathcal{L}_G = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{gen}\mathcal{L}_{adv_{gen}} + \lambda_{rec}\mathcal{L}_{rec} \quad (6)$$

$$\mathcal{L}_D = \lambda_{adv}\mathcal{L}_{adv} \quad (7)$$

where $\lambda_{L1}$, $\lambda_{gen}$, and $\lambda_{rec}$ are the required hyper-parameters that manage the relative importance of L1-distance, generating loss and reconstruction loss, and $\lambda_{adv}$ is the hyper-parameter for the discriminator. The set values of the above parameters in all experiments are 50, 1, 10 and 0.5 respectively.

## 4. EXPERIMENTS

Our proposed Edge-GAN is able to achieve the goal of reconstructing multi-view face images in high quality using any inputs with different views. In this section, we present the experimental results of applying our model on the Multi-PIE dataset. The steps involved in pre-processing and applying the dataset is elaborated in Sec. 4.1, while the results and a comparison with other works are presented in Sec. 4.2 and Sec. 4.3, respectively.

### 4.1 Experimental Settings

**Databases.** RAFD is a face image dataset containing 67 persons of different races, ages and genders; with each person presenting 8 expressions, (namely: anger, disgust, fear, joy, sadness, surprise, contempt and neutrality) with 3 three different gaze directions for each facial expression that were captured simultaneously from 5 different angles using 5 cameras. For this work, we mainly focus on face images in different views. Multi-PIE [7] contains 337 people, and all subjects were shot from 15 viewpoints and 19 illumination conditions with a range of facial expressions. For comparison, this work uses the Multi-PIE dataset presented in [19], which contains 250 subjects (image size of 128*128) of which the first 200 subjects are used for training and the rest for testing.

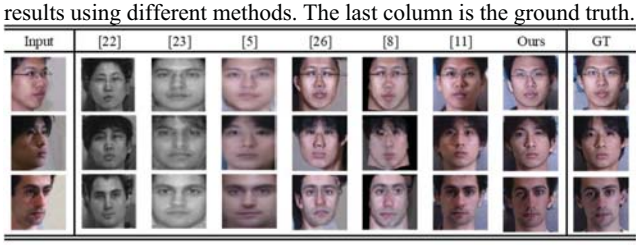
**Table 1**: Generator Network Architecture.

| Part | Layer Information |
|---|---|
| Down-sampling | Conv-(N64, K7*7, S1, P1), ReLU<br>Conv-(N128, K3*3, S2, P1), ReLU<br>Conv-(N256, K3*3, S2, P1), ReLU |
| Residual Block (6) | Conv-(N256, K3*3, S1, P1), ReLU<br>…….<br>Conv-(N256, K3*3, S1, P1), ReLU |
| Up-sampling | DeConv-(N256, K3*3, S2, P1), ReLU<br>DeConv-(N128, K3*3, S2, P1), ReLU<br>DeConv-(N64, K7*7, S1, P1), ReLU |

**Table 2**: Discriminator Network Architecture.

| Part | Layer Information |
|---|---|
| Input Layer | Conv-(N64, K4*4, S2, P1), Leaky ReLU |
| Hidden Layer | Conv-(N128, K4*4, S1, P1), Leaky ReLU<br>Conv-(N256, K4*4, S1, P1), Leaky ReLU<br>Conv-(N512, K4*4, S1, P1), Leaky ReLU |
| Output Layer | Conv-(N512, K4*4, S2, P1), Leaky ReLU<br>Conv-(N1, K4*4, S1, P1), Leaky ReLU |

**Table 3**: Comparison with state-of-art frontal-view generating results on Multi-PIE. The images in the left column are the input images with poses of 45° (1st and 2nd row) and 30° (3rd row). And the images in the later columns are the corresponding generating results using different methods. The last column is the ground truth.




The results presented in Sec 4.2 and Sec 4.3 are synthesized from images with views ranging from -90° to 90°.

**Implementation Details.** The architecture for our generator and discriminator is similar to [12]. The network contains stride-2 convolutions, several residual blocks, and 1/2-strided convolutions. Instance normalization is used here, similar to [24]. Also, the image data are normalized before the model training or testing. The learning rate used is 0.0002 and we use minibatch SGD and the Adam solver, where the batch size is set to 1.

### 4.2 Qualitative Results on Multi-PIE

In this section, we study the generation results with different baselines. Edge-GAN can easily recover photorealistic frontal views from any other face with different poses, as shown in Figure 3. The edge information fed to the model can shape the generated image with the desired pose and the input content image can easily keep the main content of the target image.

Figure 3 (a) presents the multi-view results generated by our Edge-GAN from some specifically conditioned images, while Figure 3(b) presents the general frontal generated results by Edge-GAN using different images with different views. The views involved in both cases range from -90°, -60°, -30°, -15°, 15°, 30°, 60° and 90°, correspond to almost all practical situation.

These results above show that our model can guarantee relatively stable and less distorted generated results from various angles. Additionally, Edge-GAN perfectly preserves observed face attributes in the original profile image; e.g. the eyeglasses for the second person and the hair color for the third person in (a) and the eyeglasses for the second person and the hairstyle for the third person in (b) that are all generated with a good identity preserving quality.

### 4.3 Comparison Study

**Comparison of generated images.** Table 3 which presents some visual results from the same face images, with poses of 30°and 45°, Enables a basis of comparison of the efficacy of this model. Like TP-GAN [8], our Edge-GAN can also generate good results in handling poses larger than 45°.

**Table 4**: Rank-1 Recognition Accuracy (%) on Multi-PIE under setting as Sec 4.1.

| Method | ±90° | ±75° | ±60° | ±45° | ±30° | ±15° |
|---|---|---|---|---|---|---|
| Light-CNN | 9.00 | 32.35 | 73.30 | 97.45 | 99.80 | 99.78 |
| TP-GAN | 64.03 | 84.10 | **92.93** | **98.58** | **99.85** | 99.78 |
| Edge-GAN (Without edge) | 39.98 | 63.88 | 83.16 | 97.24 | 99.49 | 99.80 |
| Edge-GAN | **68.47** | **85.31** | 91.42 | 97.65 | 99.59 | **99.80** |

**Table 5**: Identity similarities between real and generated images on Multi-PIE.

| Model | Edge-GAN | Edge-GAN (Without edge) | CR-GAN | DR-GAN |
|---|---|---|---|---|
| Identity Similarities | **0.997±0.016** | 1.157±0.017 | 1.018±0.019 | 1.073±0.013 |

**Rank-1 recognition.** After extracting deep features using Light-CNN [29], we obtain a 256-dimensional vector representation of all test images. A cosine-distance metric is used to characterize the similarity of images and the accuracy of Rank-1 recognition is evaluated. Table 4 provides a comparison with TP-GAN and Edge-GAN without edge information as input.

**Identify similarities.** In addition to the experiments discussed above, we generated 9 views of -60°, -45°, -30°, -15°, 0°, 15°, 30°, 45° and 60° for images in Multi-PIE and evaluate the identity similarities between the real and generated images using squared L2 distance. Table 5 shows the average L2 distance of generated results using Edge-GAN, CR-GAN and DR-GAN, obtained from FaceNet [28]. The average L2 distance of the generated images and the corresponding real images can be seen as a quantitative index to describe the difference between the two domains. With this approach, we can clearly compare different baselines with numerical values.

## 5. CONCLUSION

In this paper, we present an edge-detection based GAN (Edge-GAN) model for synthesizing multi-view images from a single-view face image. The proposed method utilizes the edge information to shape the pose of the generated image and fill the content of the target image by the connection with the conditioned input image. The experimental results on Multi-PIE and RaFD datasets have shown that our method can synthesis multi-view images using an input image the arbitrary views. EdgeGAN is not only able to generate high-quality frontal face images, especially getting better-generated results from high-angle face images, but also highly restore the side-view of faces which would be useful for the realization of synthetic high-quality 3D images.

## 6. REFERENCES


[1] Canny, J., "A computational approach to edge detection." *IEEE Transactions on pattern analysis and machine intelligence*, (6), pp. 679-698, 1986.

[2] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J., "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation." In *CVPR*, pp. 8789-8797, 2018.

[3] Denton, E. L., Chintala, S., & Fergus, R., "Deep generative image models using a laplacian pyramid of adversarial networks." In *NeurIPS*, pp. 1486-1494, 2015.

[4] Donahue, J., Krähenbühl, P., & Darrell, T., "Adversarial feature learning." *arXiv preprint arXiv:1605.09782*, 2016.

[5] Ghodrati, A., Jia, X., Pedersoli, M., & Tuytelaars, T., "Towards automatic image editing: Learning to see another you." *arXiv preprint arXiv:1511.08446*, 2015

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y., "Generative adversarial nets." In *NeurIPS*, pp. 2672-2680, 2014.

[7] Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S., "Multi-pie." *Image and Vision Computing*, *28*(5), 807-813, 2010.

[8] Huang, R., Zhang, S., Li, T., & He, R., "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis." In *ICCV*, pp. 2439-2448, 2017.

[9] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A., "Image-to-image translation with conditional adversarial networks." In *CVPR*, pp. 1125-1134, 2017.

[10] K. E. Ak, J. H. Lim, J. Y. Tham, and A. A. Kassim., "Attribute manipulation generative adversarial networks for fashion images." In *ICCV*, 2019.

[11] K. E. Ak, J. H. Lim, J. Y. Tham, and A. A. Kassim., "Semantically consistent text to fashion image synthesis with an enhanced attentional generative adversarial network." *Pattern Recognition Letters*, 2020.

[12] Karacan, L., Akata, Z., Erdem, A., & Erdem, E., "Learning to generate images of outdoor scenes from attributes and semantic layouts." *arXiv preprint arXiv:1612.00215*, 2016.

[13] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. D., "Presentation and validation of the Radboud Faces Database." *Cognition and emotion*, *24*(8), pp. 1377-1388, 2010.

[14] Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., & Van Gool, L., "Pose guided person image generation." In *NeurIPS*, pp. 406-416, 2017.

[15] Mejjati, Y. A., Richardt, C., Tompkin, J., Cosker, D., & Kim, K. I., "Unsupervised attention-guided image-to-image translation." In *NeurIPS*, pp. 3693-3703, 2018.

[16] Mirza, M., & Osindero, S., "Conditional generative adversarial nets." *arXiv preprint arXiv:1411.1784*, 2014.

[17] Peng, X., Yu, X., Sohn, K., Metaxas, D. N., & Chandraker, M., "Reconstruction-based disentanglement for pose-invariant face recognition." In *ICCV*, pp. 1623-1632, 2017.

[18] Ranzato, M. A., Huang, F. J., Boureau, Y. L., & LeCun, Y., "Unsupervised learning of invariant feature hierarchies with applications to object recognition." In *CVPR*, pp. 1-8, 2007, June.

[19] Tian, Y., Peng, X., Zhao, L., Zhang, S., & Metaxas, D. N., "CR-GAN: learning complete representations for multi-view generation." *arXiv preprint arXiv:1806.11191*, 2018.

[20] Tran, L., Yin, X., & Liu, X., "Disentangled representation learning gan for pose-invariant face recognition." In *CVPR*, pp. 1415-1424, 2017.

[21] Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., & Kim, J., "Rotating your face using multi-task deep neural network." In *CVPR*, pp. 676-684, 2015.

[22] Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V., "Few-Shot Adversarial Learning of Realistic Neural Talking Head Models." *arXiv preprint arXiv:1905.08233*, 2019.

[23] Zhao, B., Wu, X., Cheng, Z. Q., Liu, H., Jie, Z., & Feng, J., "Multi-view image generation from a single-view." *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 383-391. 2018, October.

[24] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A., "Unpaired image-to-image translation using cycle-consistent adversarial networks." In *ICCV*, pp. 2223-2232, 2017.

[25] Zhu, X., Lei, Z., Yan, J., Yi, D., & Li, S. Z., "High-fidelity pose and expression normalization for face recognition in the wild." In *CVPR*, pp. 787-796, 2015.

[26] Zhu, Z., Luo, P., Wang, X., & Tang, X., "Multi-view perceptron: a deep model for learning face identity and view representations." In *NeurIPS*, pp. 217-225, 2014.

[27] Zhu, Z., Luo, P., Wang, X., & Tang, X., "Deep learning identity-preserving face space." In *ICCV*, pp. 113-120, 2013.

[28] Schroff, F., Kalenichenko, D., & Philbin, J., "Facenet: A unified embedding for face recognition and clustering." In *CVPR*, pp. 815-823, 2015.

[29] Wu, X., He, R., Sun, Z., & Tan, T., "A light cnn for deep face representation with noisy labels." IEEE Transactions on Information Forensics and Security, 13(11), pp. 2884-2896, 2018.