iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.48041

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

Sketch-to-Face Image Translation and enhancement using a Multi-GAN approach

Sparsh Nagpal
Thadomal Shahani Engineering College

Abstract: Image to Image Translation using GANs has been introduced for a few years now, and newer methodologies keep originating daily. Our method combined two approaches to generate accurate and more precise results. First, we use inpainting in the Contextual GANs model followed by Generative Facial Prior GAN (GFP GAN) for image enhancement to give realistic results of a face for the sketch input.

I. INTRODUCTION (ALSO, ADD GFP GANS)

Forensic sketch artists create facial composite sketches to assist law enforcement officers in identifying criminal suspects. Because sketches alone do not always provide sufficient detail, it is necessary to match these photos for accuracy. Criminal investigation's technical capabilities and methodologies have recently advanced. However, criminal identification through face sketches still requires good human visual and memory skills. Unless we can efficiently generate images from thoughts, we must rely on the memory and description provided to the person creating the sketch. Sketch images include basic face profile information but lack the detailing found in photogenic photos. Recognizing actual human faces from those sketches becomes difficult as a result. Incorporating computer vision into this will reduce the human effort required to relate a black-and-white drawing to real faces via image translation. Previous works focused on forensic sketches and handled the task of transforming viewed sketches into mugshot images. This research aims to turn viewed sketches into realistic facial photos. In recent years, deep convolutional neural networks have been used to solve many image transformation tasks (CNNs). The networks receive an input image and transform it into a corresponding output image, requiring detailed information and complex image textures. Several researchers have recently used specific mechanisms to tackle these individual tasks.

Many studies have recently been conducted in an attempt to solve this problem. Deep convolutional generative adversarial networks (DCGAN)-based text-to-image task demonstrated that GAN could generate images effectively conditioned on text descriptions. Zhang, likewise, proposed stacked generative adversarial networks (StackGAN) for developing photorealistic images from text. There is an exciting demo using edge to create shoes for the sketch-to-image task, which proposed a pixel-to-pixel image translation network, which sparked the image-to-image research boom (cat). The image generation problem was posed by Luet as an image completion problem, with sketch serving as a weak contextual constraint. We can train a network to generate impressively.

Common approaches in cGANs now include challenging conditions such as pixel-wise correspondence alongside the translation process, ensuring that the output edges strictly align with the input edges. The input is a free-hand sketch, which can be highly problematic in sketch-to-image generation. Sketch-to-image generation aims to automatically generate a photographic image of the hand-sketched object. Even a crudely drawn illustration allows non-artists to easily specify an object's attributes in many situations where verbose text description would be cumbersome. On the other hand, the translation should respect the sparse input content but may require some shape deviation to generate a realistic image. To address these issues, we propose a new contextual generative adversarial network for a sketch-to-image generation.

We use the sketch as a weak constraint to find the closest mapping and define our objective function, consisting of a contextual loss and a traditional GAN loss. We also propose a simple scheme for improving sketch initialization.

This innovative method has the following benefits:

1) A joint sketch-image pair, which is a single image, is understood by a single network; there are no distinct domains for image and sketch learning. The picture translation process, in contrast, only accepts sketches as input.

2) The resulting image may display different poses and shapes beyond the input sketches that may not strictly correspond to photographic objects by utilizing a weak sketch constraint while relating to its input edges.

3) Since there is no difference between the image and the sketch from the perspective of the combined image, they can be switched to provide the other the necessary context to be completed. Consequently, the reverse or image-to-sketch generation can be done using the exact same sketch-to-image generating approach or network.

Our system is universal and may be used with any cutting-edge generative model. We have a two-phase recipe to learn the sketch-image connection inherent in a joint image and impose the weak sketch constraint, capitalizing on the GAN for image completion. The network is trained on the connection between sketches and photographs using uncropped joint images.
To "complete" the image based on a modified objective, we look for an encoding of the provided corrupted image using only the sketch as the weak context. By feeding it to the generator, which creates the photographic object from the drawing, this encoding is then utilized to reconstruct the image.

II. RELATED WORK

A. GANs

Ian Goodfellow et al. [1] introduced the Generative adversarial networks framework, constituting two convolutional neural network models, a generator and a discriminator. A generator generates an image and learns to come close to the target image, while the discriminator acts as a classifier indicating how well the generator performs and gives feedback. Every feedback provided helps the model improve incrementally. DCGANs were proposed by Radford and Metz et al. [2], which allowed training GAN stably over more settings. Conditional Conditioning variables were then introduced to GAN [3,4].

B. Image Completion

Deep image completion with perceptual and contextual losses is related to our contextual GAN for joint images [5]. Pretrained with uncorrupted data, the G and D networks are trained to reconstruct a complete picture. Their impressive examples show that even though a sizeable facial image region is cropped from the input, the entire facial image looks realistic. Another remarkable work on image completion [6] is based on autoencoders with a standard reconstruction loss and an adversarial loss. Autoencoders have also been successfully applied to generate images from visual attributes [7]. The whole input sketch is considered the “uncropped context” for completing the entire raw image portion of the joint image, a bit like how image completion uses the uncropped part of the image to complete the facial image.

C. Image Enhancement

Contextual GANs based on the given dataset provide results with certain image quality and realness limitations. The model's training could be modified to attain better quality. In this, we used an additional layer of GFP GANs [8] to achieve more realistic face images after the image-to-image translation. Our method balances fidelity and realness by incorporating this Generative Facial Prior (GFP) into the face restoration process through channel-split spatial feature transform layers. GFP-GAN can jointly restore facial details and enhance colors with just a single forward pass, while GAN inversion methods require expensive image-specific optimization at inference. Experiments have shown that this method achieves superior performance to the prior art on both synthetic and real-world datasets.

III. PROPOSED METHODOLOGY

A. Inpainting

‘Inpainting’ generates an image from itself. One small portion of the input image is corrupted with a mask. This image, which is partially covered with a mask, is called the context. The model tries to generate a whole image concerning this context. We have employed this concept, but our input image consists of the sketch and the face image side by side. Say we have our sketch image to the left of our face image. We corrupt the right side of our input image, which consists of the face image with a mask. Now, our model uses this partially masked image as the context and generates a completed image concerning this context.

B. Network Architecture

We use seven up-convolutional layers with kernel size 4 and stride 2. Each up-convolutional layer is followed by a batch normalization layer to accelerate training and stabilize learning. All the layers utilize rectified linear unit (ReLU) activation. Finally, we apply tanh to the output layer. This series of up-convolutions and non-linearities conduct a nonlinear weighted upsampling of the latent space and generates a higher resolution image of 256 × 256 x 3. For the discriminator, the dimensions of the input image are 256 x 256 x 3. It is followed by 4 convolutional layers. The dimension of the feature map is halved, while the number of channels is doubled compared to the previous layer. Specifically, we add six convolutional layers with kernel size 4 and stride 1 to produce a 64×64×1 output. To reshape the output to one dimension, we add a fully connected layer followed by a softmax layer.

C. Implementation Details

The image-to-image translation model, especially when the input image is a sketch, is a non-trivial operation. Moreover, the non-uniform sketches are tedious. Therefore, we propose using a model with a joint input space. The input corpus would contain joint images of the sketch and real images capturing contextual information in the region having the sketch.

Specifically, we tend to train a GAN model with exploited joint pictures. The generator then mechanically predicts the corrupted image half supported by the context of the corresponding sketch half. Furthermore, the generator embeds the joint pictures onto a nonlinear joint area z, i.e., z could be a joint embedding of sketch and image. In contrast, in previous work (e.g., [9]), z is just a picture embedding. As a result, rather than restricting the generated image directly with the complete z (hard constraint), we will constrain the generated image indirectly via the sketch of a part of the joint embedding z of the input, therefore remaining devoted whereas exhibiting a point of freedom within the look of the output image. Figure four illustrates this pipeline which can be careful in resulting sections.

The following output images were passed through a layer of pre-trained GFP GANs trained over low quality portrait enhancer to increase the image quality and generate more human like output even on a lower trained model.

IV. EXPERIMENTS

A. Dataset

1) Data Augmentation: We used the CUHK Face Sketch database (CUFS) [10] . It includes 188 faces from the Chinese University of Hong Kong (CUHK) student database, 295 faces from the XM2VTS database [11], and 123 faces from the AR database [12]. There are 606 faces in total. For each face, there is a sketch drawn by an artist based on a photo taken in a frontal pose, under normal lighting conditions, and with a neutral expression. Since the generative model requires a much larger dataset, we used rotation, translation, shearing, and flipping and built a dataset with 12k images.

Fig. 1. CUFS Dataset

2) Implementation: The network is pretrained by the use of contextual GAN. We make use of the Adam optimizer [13] with a learning rate of 0.0002 as well as a beta of 0.5 for both the generator and discriminator network. We train the network with a batch size of 16 and epochs of 100, which takes 6 to 48 hours for training, depending on the size of the training set. During back-propagation, Stochastic clipping is applied. A relatively small λ is set so that the contextual loss is more critical in test-time optimization and that the sketch portion in generated image best resembles the input sketch. During the back-propagation, the generator and the discriminator are fixed. For the purpose of experimental results, this update can be done in 500 iterations (the loss converges very fast with our refined initialization and typically becomes stable after 100 iterations, which takes <1s). We use the same network architecture for all three categories. 10 Y. Lu, S. Wu, Y.W. Tai and C.K. Tang. The following output images were passed through a layer of pre-trained GFP GANs trained over low quality portrait enhancer to increase the image quality and generate more human like output even on a lower trained model.

Fig .2. Sketch (A), Image on Contextual GANs (B), Final Image after GFP GANs (C), Actual Image (D)

3) *Evaluation:* Although we clearly outperformed CE [14] and pix2pix [9] on poorly drawn sketches, for the sake of fairness, we also performed two quantitative experiments on good sketches whose edges correspond to the photographic objects they correspond to: (a) SSIM with ground truth; (b) face verification accuracy. We use CUFS test images for both (a) and (b) evaluation.
4) *SSIM:* Structural similarity metric (SSIM) [16] is used to compare generated images with ground truth. On typical sketches, we achieved comparable results to pix2pix, but much better than CE. It should be noted that pix2pix and CE strictly follow the input sketches, and SSIM may not incorporate measures of human perception if the input sketch is poorly drawn.
5) *Verification Accuracy:* The purpose of this study is to verify whether the generated faces have the same identity label as ground truth if they are plausible. We extracted identity-preserving features using the pretrained Light CNN [17] and compared them with the L2 norm. Our model outperformed pix2pix in Table 1, showing that it not only learns to capture essential details, but also is more resilient to different sketches.

Structural Similarity Index Measure : 0.78
L2-Normalization Score : 93.75

V. CONCLUSION AND FUTURE WORK

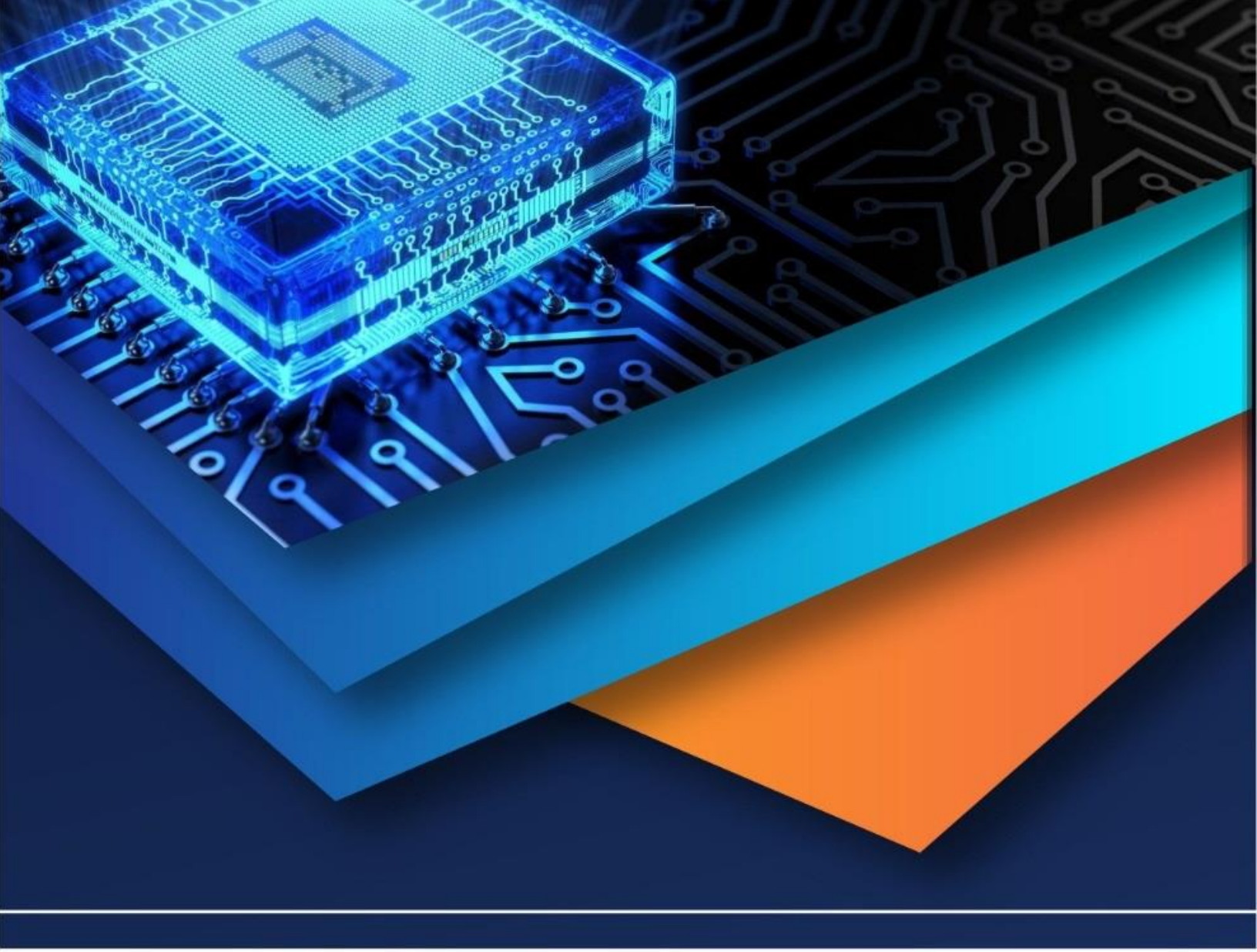
We have showcased that the problem of sketch-to-image generation can be taken up as the joint image completion problem, where the sketch provides the context for completion. Based on this novel idea, we propose the contextual GAN framework approach with the integration of GFP GANs for image enhancement. Using a generative adversarial network, the joint distribution and corresponding image are captured, avoiding the cross-domain problems associated with cross-domain learning. By encoding the "corrupted" joint image into the closest "uncorrupted" joint image in the latent space, the output image part of the joint image can be predicted. This is followed by the GFP GAN model, which recreates a realistic version of human faces from the generated face model of the sketch to face.

Our three-stage method requires longer inferring time during testing than the end-to-end methods. However, the three-stage approach allows us to separate the training, testing and enhancement. In training, our generator learns the natural appearance of faces so that any noise vector in the latent space could generate a plausible visual image. On testing, although we have augmented the sketch drawing with three different sketch styles, we do not strictly restrict the human free-hand drawing following the three augmented styles. The GFP GANs towards the end converts the low pixelated and unclear image and provides a more human touch. We conduct thorough experiments to demonstrate the advantages of the proposed framework. In the future, we plan to investigate more approached including generative models and current state-of-art diffusion models and explore more application scenarios. The diffusion models [18] including the stable-diffusion architecture recently release allows text-to-image translation with a base image. This use case and a study in this approach can further generate efficient solutions. While our output is faithful to the input sketch, the new quantitative measurement may be needed to measure the “perceptual” correspondence between a (badly drawn) input sketch and our generated image (e.g., Figure 1), a subject and complex problem in its own right.

REFERENCES

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014), http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
[2] Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR abs/1511.06434 (2015)
[3] Wang, X., Gupta, A.: Generative image modeling using style and structure adversarial networks. In: ECCV (2016)
[4] Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: IEEE Int. Conf. Comput. Vision (ICCV). pp. 5907–5915 (2017)
[5] Yeh, R.A., Chen, C., Lim, T.Y., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: CVPR (2017)
[6] Pathak, D., Kr¨ahenb¨uhl, P., Donahue, J., Darrell, T., Efros, A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)
[7] Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: European Conference on Computer Vision. pp.776–791. Springer (2016)
[8] Xintao Wang, Yu Li, Honglun Zhang, Ying Shan: Towards Real-World Blind Face Restoration with Generative Facial Prior. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
[9] Zhu, J.Y., Kr¨ahenb¨uhl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: European Conference on ComputerVision. pp. 597–613. Springer (2016)
[10] X. Wang and X. Tang, “Face Photo-Sketch Synthesis and Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 31, 2009.

[11] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: the Extended of M2VTS Database," in Proceedings of International Conference on Audio- and Video-Based Person Authentication, pp. 72-77, 1999.
[12] A. M. Martinez, and R. Benavente, "The AR Face Database," CVC Technical Report #24, June 1998.
[13] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRRabs/1412.6980 (2014), http://arxiv.org/abs/1412.6980
[14] Pathak, D., Kr¨ahenb¨uhl, P., Donahue, J., Darrell, T., Efros, A.: Context encoders: Feature learning by inpainting. In: CVPR (2016) Y. Lu, S. Wu, Y.W. Tai and C.K. Tang
[15] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
[16] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13(4), 600–612 (2004)
[17] Wu, X., He, R., Sun, Z., Tan, T.: A light cnn for deep face representation with noisy labels. arXiv preprint arXiv:1511.02683 (2015)
[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer.: High-Resolution Image Synthesis with Latent Diffusion Models. In: (2021) 2112.10752 arXiv cs.CV



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY