Group A
Motion/Style Transfer

Ajinkya R, Branden Lopez, Indranil Patil, Rushikesh P, Warada K

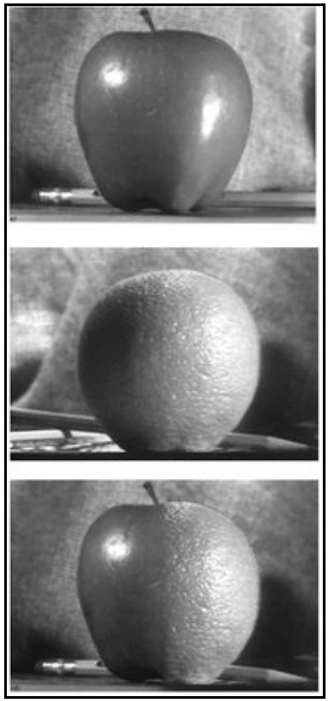
Background Problem

- Computer vision (CV) aims to give computers “sight”.
- With sight we can recognize motion, art styles, and create new images with these properties.
- CV attempts to see these properties and “Transfer” them into new images.
- Leading to Motion Transfer, Styles Transfer and Image Generation.

- State-of-the art cannot motion transfer two objects at once.
- We will explore extending motion transfer to two objects.

Classical Approach

- One of the applications for domain adaptation or image generation can be to split two component images and merge them.
- [5] has described a multiresolution spline technique that first decomposes the component images into a set of band-pass filtered images and then the images in each frequency are assembled into the corresponding band-pass mosaic.
- Lastly, these band-pass mosaic images are then added to get the desired image.
- The working of this technique is explained in Figure (2).



Source: P. Burt and E. Adelson, "A multiresolution spline with application to image mosaics," *ACM Transactions on Graphics,* vol. 2, *(4),* pp. 217-236, 1983. . DOI: 10.1145/245.247.

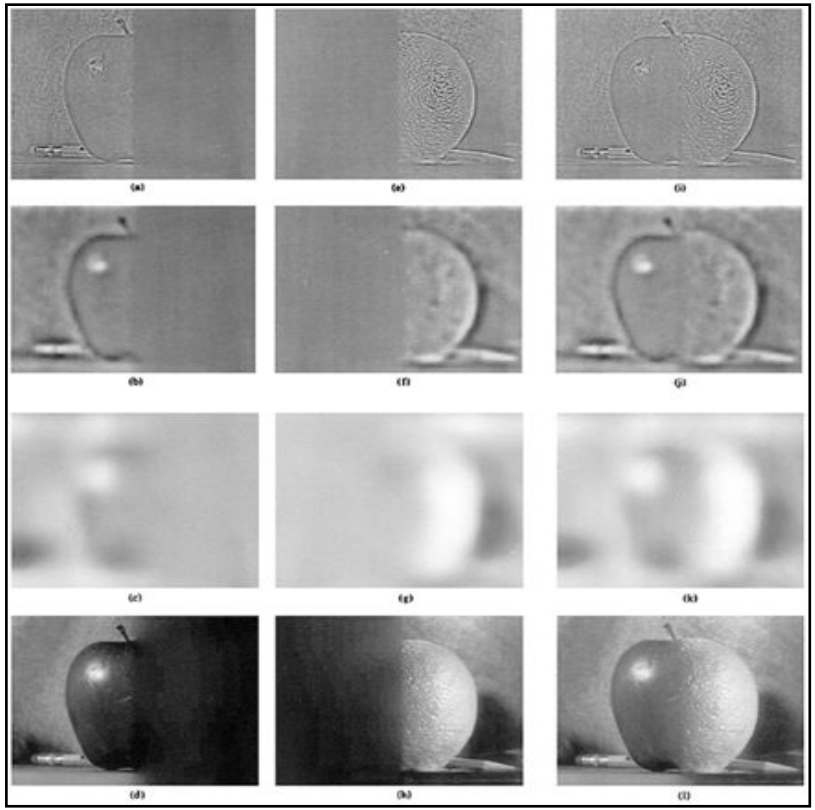


Figure 2. Contributions of the various band-pass filtered images to the apple-orange mosaic. Figure 10a-10c shows low, medium, and high frequency components of apple. Figure 10d is obtained by summing the figures 10a-10c. Same frequency components are obtained for orange in the Figures 10e-10h. Images in figure 10i-10l are obtained by summing the frequency components of half apple and half orange.

Source: P. Burt and E. Adelson, "A multiresolution spline with application to image mosaics," *ACM Transactions on Graphics,* vol. 2, *(4),* pp. 217-236, 1983. . DOI: 10.1145/245.247.

CNN

- *A Neural Algorithm of Artistic Style* was introduced by [6] that uses the feature space derived by the VGG network.
- The proposed algorithm enables the generation of new images which combines the content of a photograph with the style of artworks as depicted in figure (3).
- The functioning of the algorithm is as illustrated in figure (4).
- However, a good tradeoff between the style and content that needs to be obtained is difficult to synthesize.

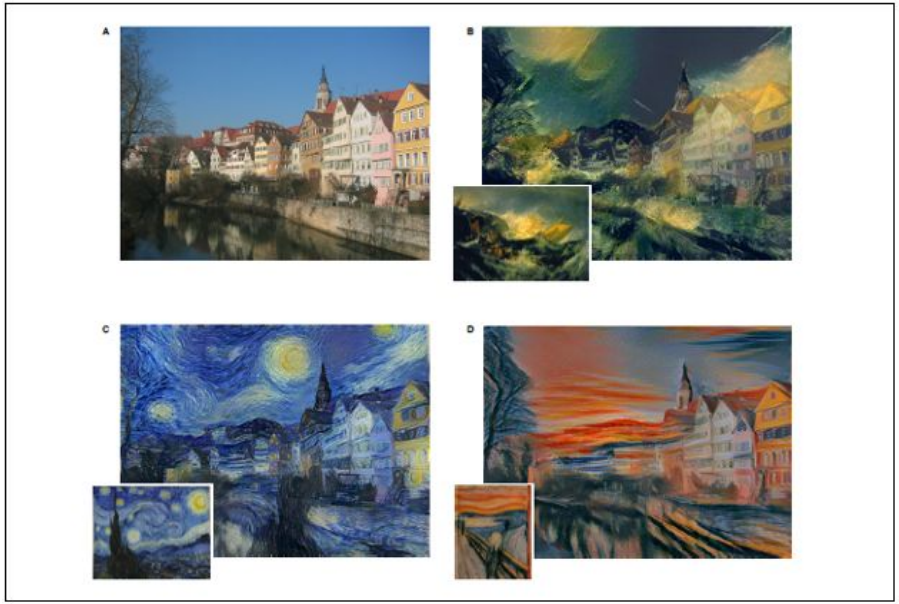


Figure 3. Images combining the content of A with artworks B, C, and D

Source: L. A. Gatys, A. S. Ecker and M. Bethge, "Image style transfer using convolutional neural networks," in 2016, . DOI: 10.1109/CVPR.2016.265.

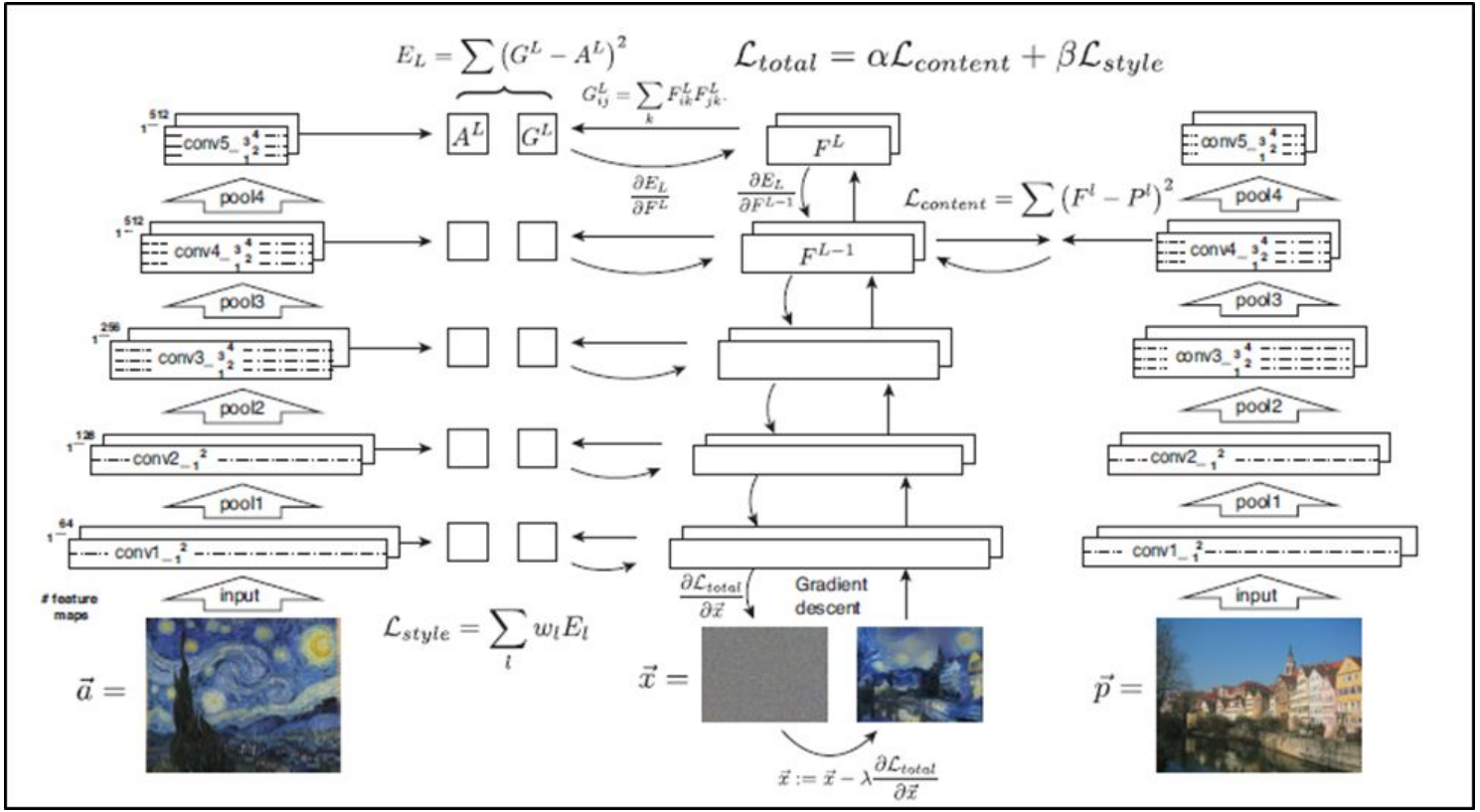


Figure 4. Style Transfer Algorithm

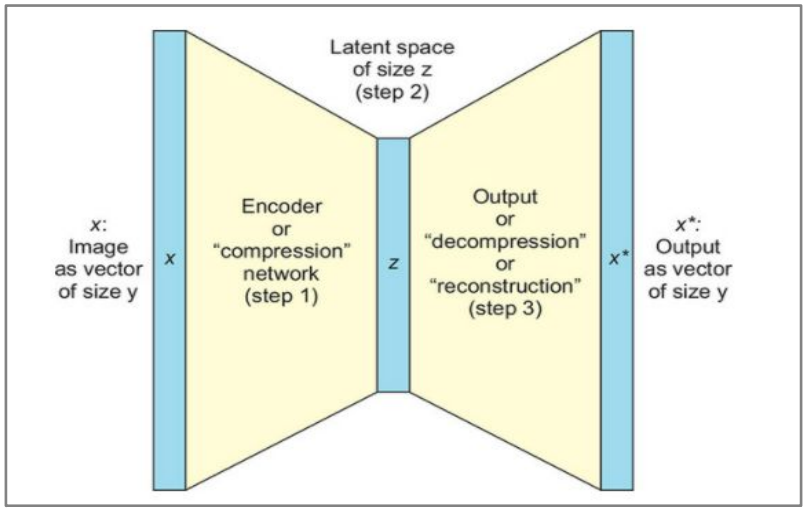
Source: L. A. Gatys, A. S. Ecker and M. Bethge, "Image style transfer using convolutional neural networks," in 2016, . DOI: 10.1109/CVPR.2016.265.

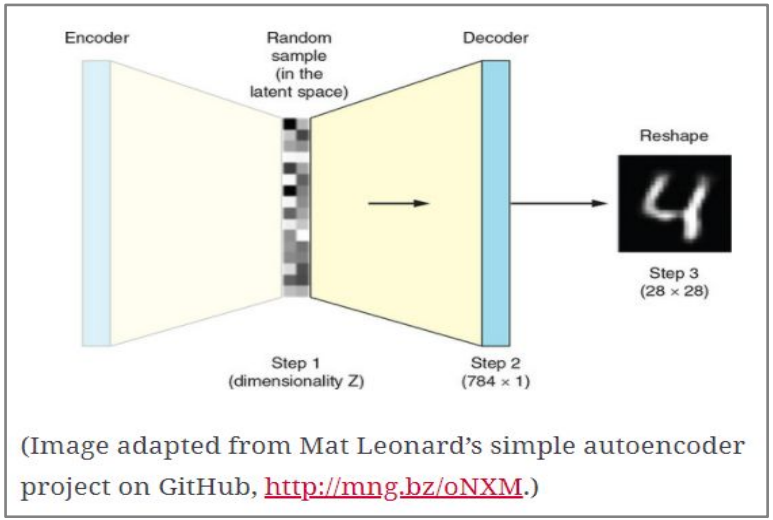
Generative Models

- Typical machine learning problems are classification and regression.
- These type of models do not generate new data.
- Generative models - a new type of task in machine learning.
- These models create new data which looks realistic based on some input noise.
- Two broad categories of generative models are-
  - Autoencoders
  - Generative Adversarial Networks (GANs)

Autoencoders

- A single deep neural network with one loss function.
- Architecture contains compression network and reconstruction network.

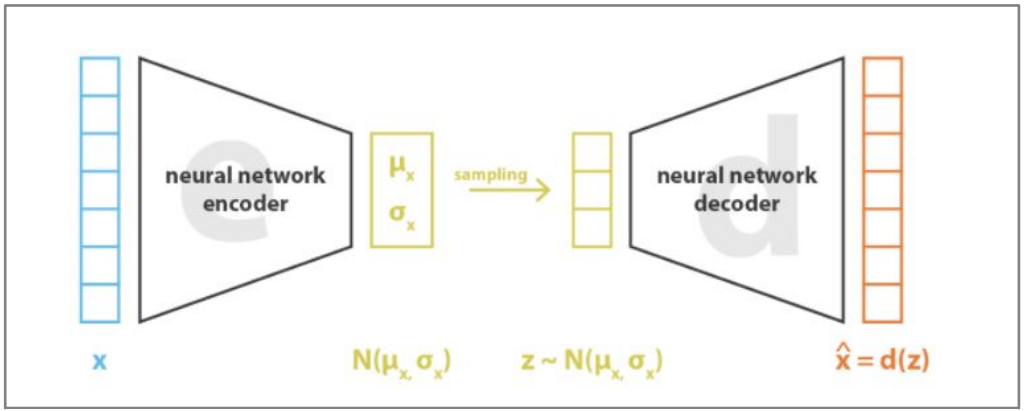


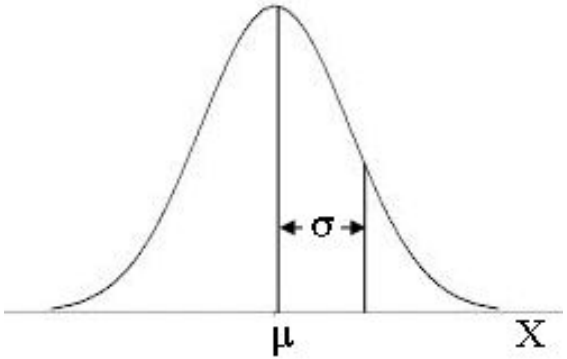


Source: Langr, J., & Bok, V. (2019). "GANs in Action: deep learning with generative adversarial networks"

Variational Autoencoders (VAE)

- Finding random vector from latent space is non-trivial.
- VAE creates distribution of vector from latent, so that samples can be drawn from latent space with the choice of variance.

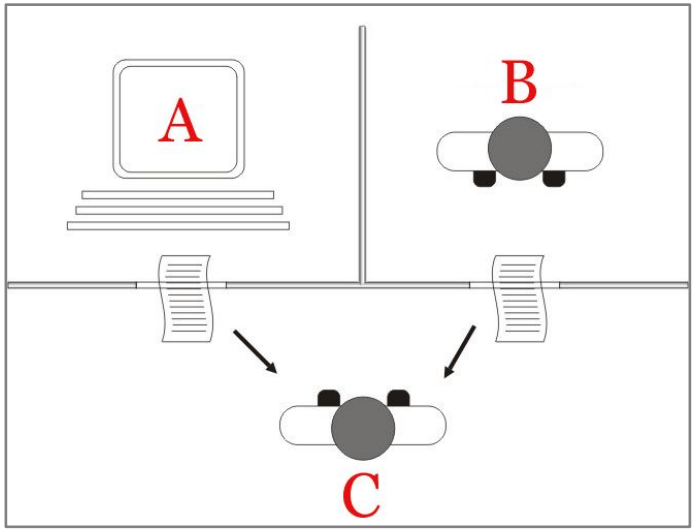




Source:https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Generative Adversarial Networks (GANs)

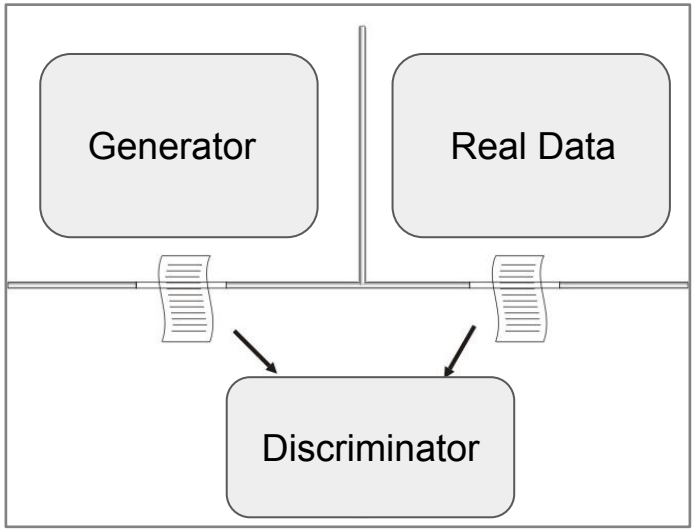
- Alan Turing, in his famous paper - ‘Imitation Game’ proposed turing test for artificial intelligence.
- An unknowing observer talks with two counterparts - a human and a machine behind closed doors.
- If observer cannot distinguish between human and machine, then turing test suggests machine passed the test and can be deemed as intelligent.



Juan Alberto Sánchez Margallo - https://commons.wikimedia.org/wiki/File:Test_de_Turing.jpg

Generative Adversarial Networks (GANs)

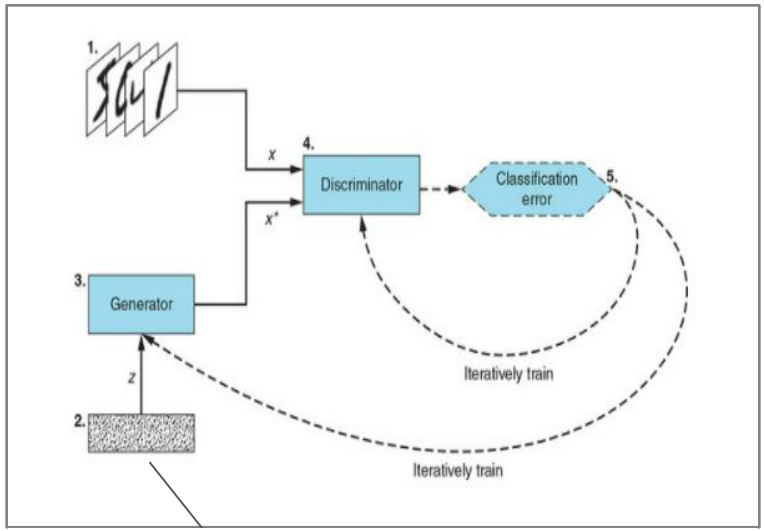
- GAN contains two models - Generator and Discriminator.
- Both of the networks are trained simultaneously.
- Generator generates fake data that looks like a sample in training data.
- Discriminator tries to distinguish between real data and fake data generated by the generator.



Adapted from: Juan Alberto Sánchez Margallo - https://commons.wikimedia.org/wiki/File:Test_de_Turing.jpg

Generative Adversarial Networks (GANs)

- Objective of discriminator is to differentiate between real and fake data.
- Objective of generator is to fool discriminator and increase misclassification of discriminator.



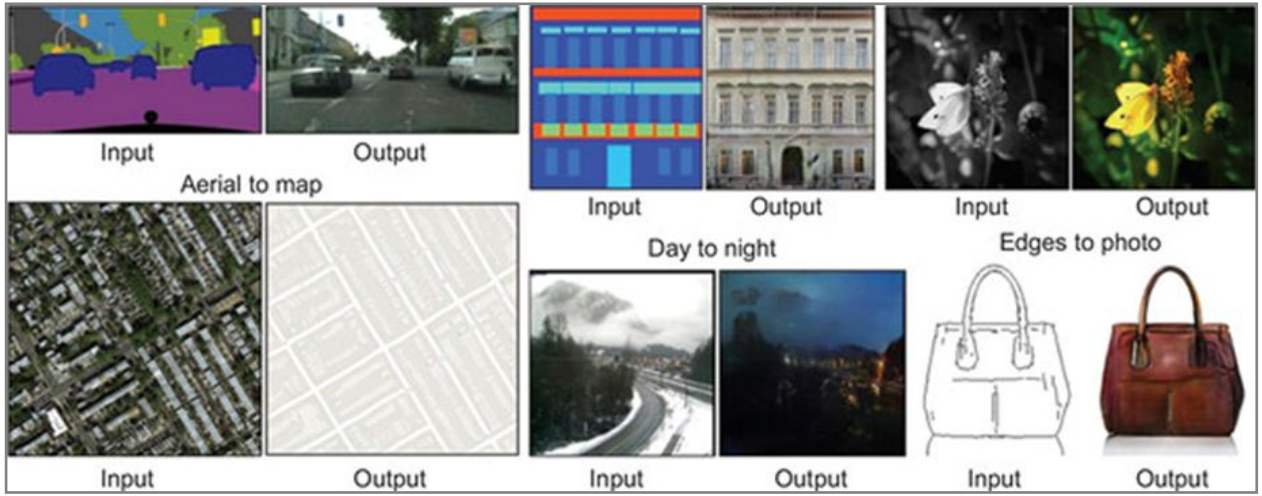
Source: Langr, J., & Bok, V. (2019). "GANs in Action: deep learning with generative adversarial networks"

Domain Adaptation & Style Transfer

Conditional GANs

Image-to-Image Translation with Conditional Adversarial Networks

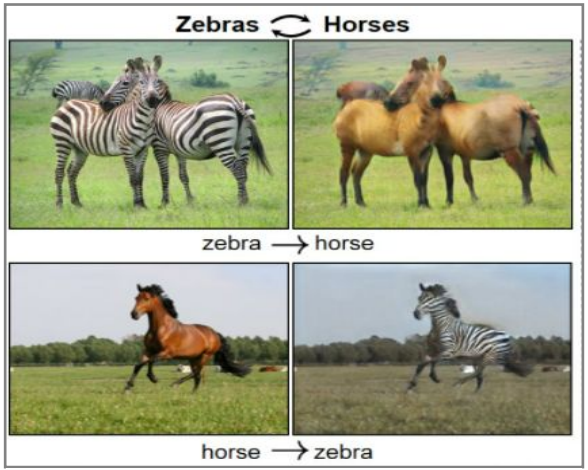
- Additional label is given as an input to the generator along with input image.
- The issue with the approach is that it requires ‘perfect pairs’ in both domains.



Source: “Image-to-Image Translation with Conditional Adversarial Networks,” by Phillip Isola [8]

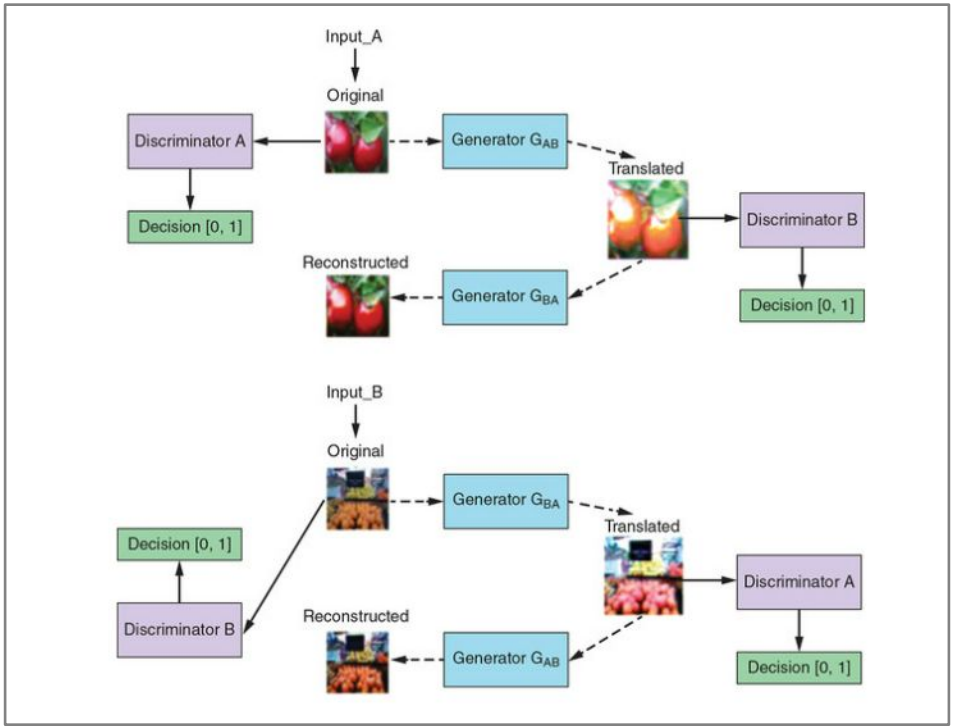
CycleGANs

- CycleGAN introduced an additional loss function called cycle-consistency loss along with adversarial loss.
- Image from one domain is translated into the other domain and then back in original domain.
- Ideally, reconstructed image should have been the same as original image.
- Difference between original and reconstructed image is calculated to compute the loss.



Source: Jun-Yan Zhu et al., 2017

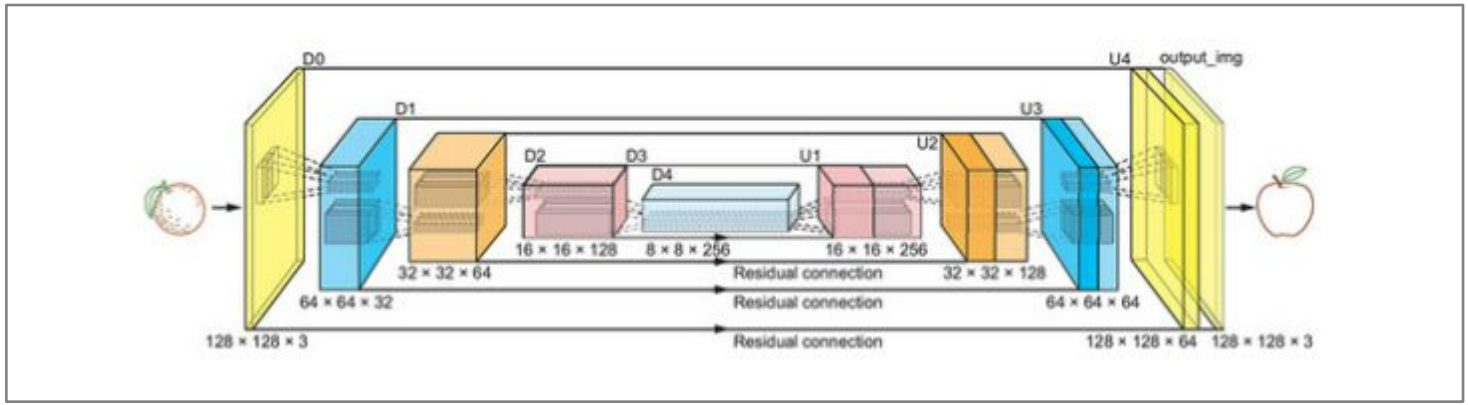
CycleGANs



Source: Langr, J., & Bok, V. (2019). "GANs in Action: deep learning with generative adversarial networks"

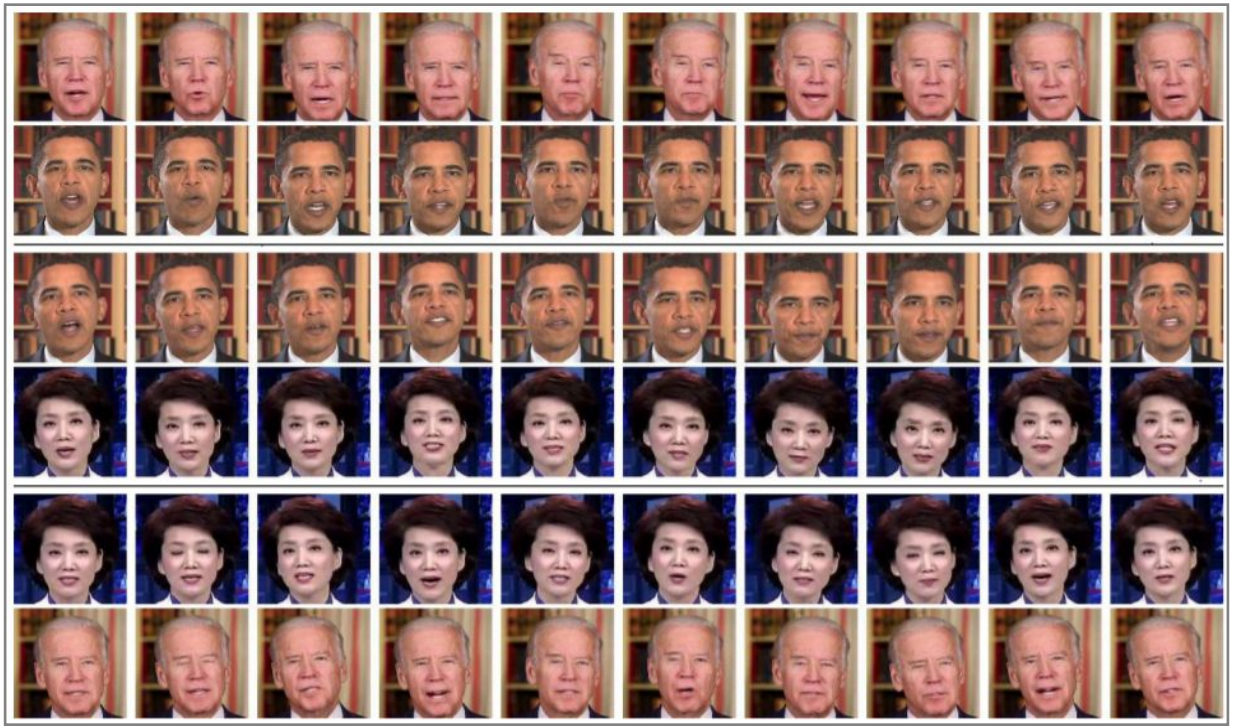
CycleGANs Architecture

- Generator in CycleGAN uses U-Net architecture.
- Characteristics -
  - Encoder-decoder model
  - Skip connections
- Discriminator uses PatchGAN architecture.



Source: Langr, J., & Bok, V. (2019). "GANs in Action: deep learning with generative adversarial networks"

CycleGANs Results

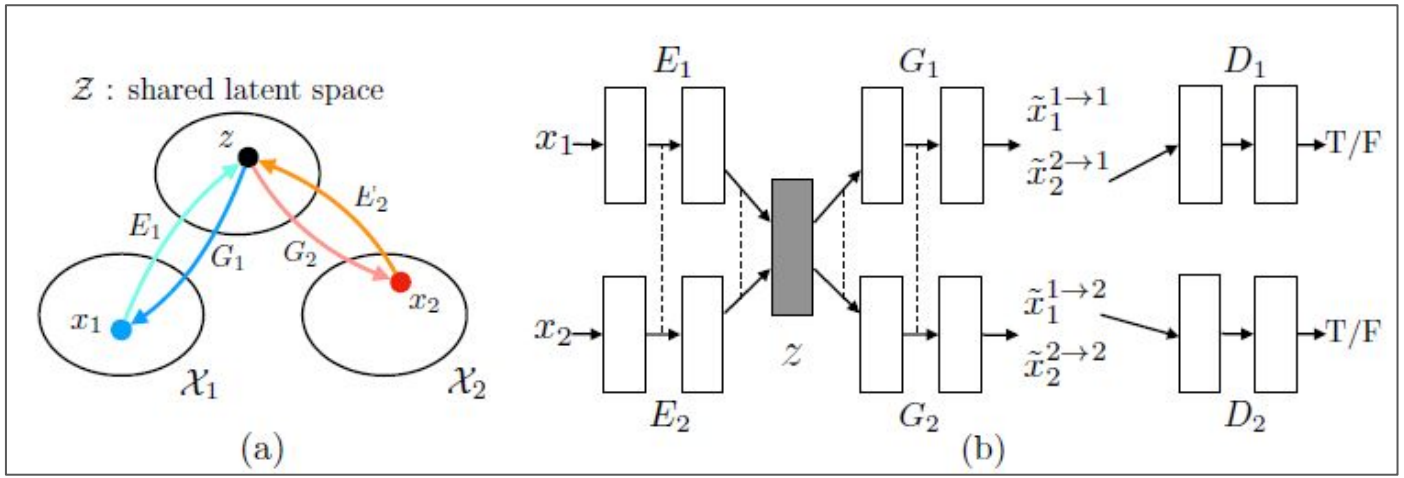


Source: Face Transfer with Generative Adversarial Network [11]

VAE and Coupled GANs

- An algorithm of unsupervised image-to-image translation that aims at learning the joint distribution of images in separate domains by using images from marginal distributions in individual domains [12].
- Learning of translation takes place in both directions in one shot.
- The proposed framework was applied on the CelebA dataset with images having facial attributes.
- The attributes were then translated from domain 1 to the images of domain 2 as visualized in figure (9), without having any corresponding images in both the domains in the training dataset.

Framework



(a) The shared latent space assumption.
(b) The proposed framework. E1, E2, G1, and G2 are represented using CNNs and the latent space is implemented by using the weights of E1 and E2 tied to the weights of G1 and G2.

Source: M. Liu, T. Breuel and J. Kautz, "Unsupervised Image-to-Image Translation Networks," 2017.

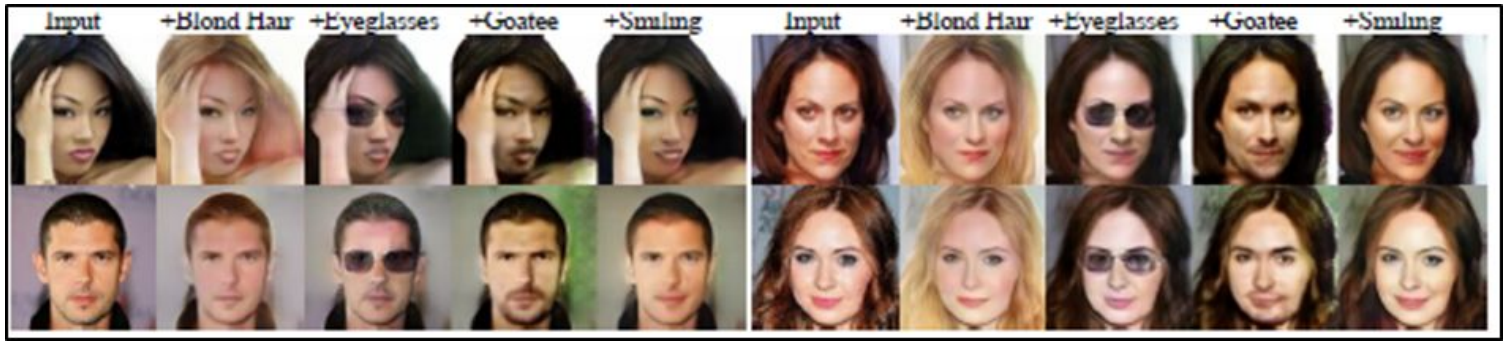
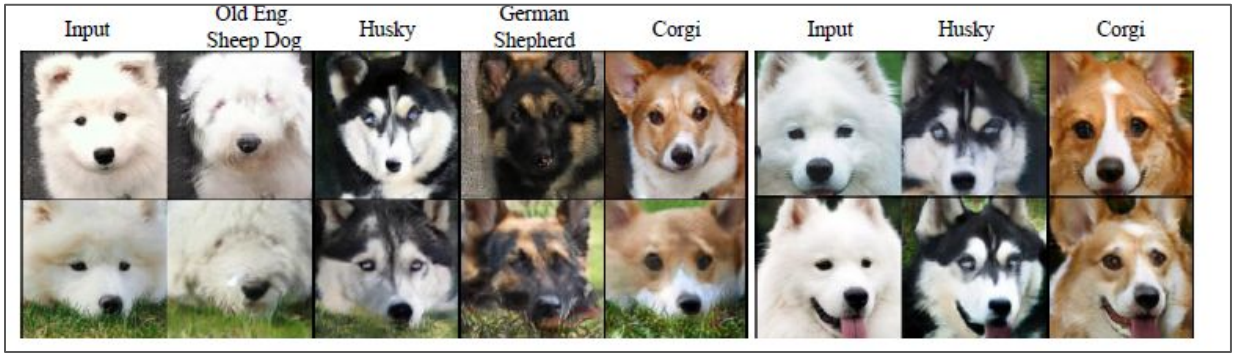


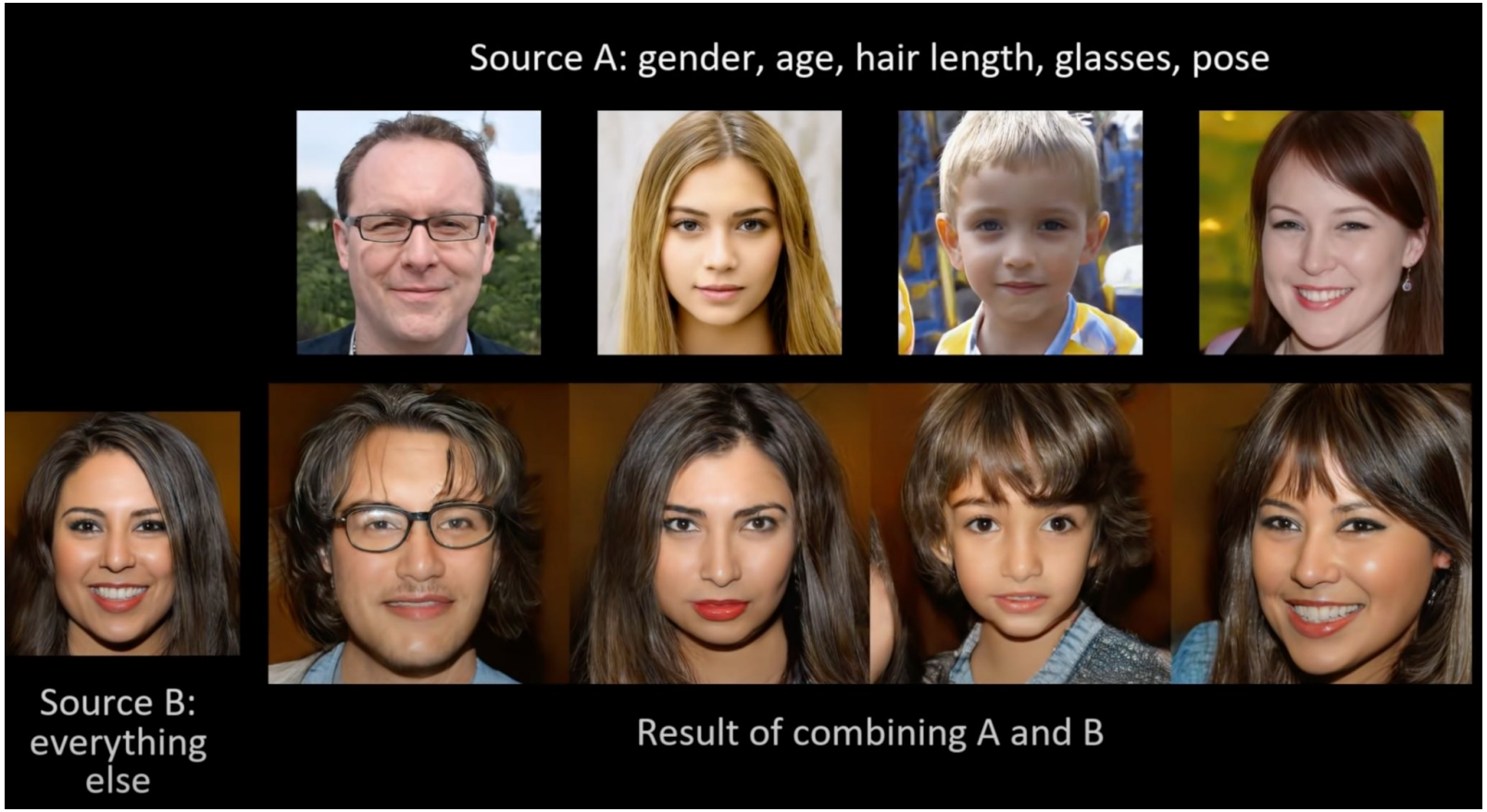
Figure 9. Attribute-based face translation results

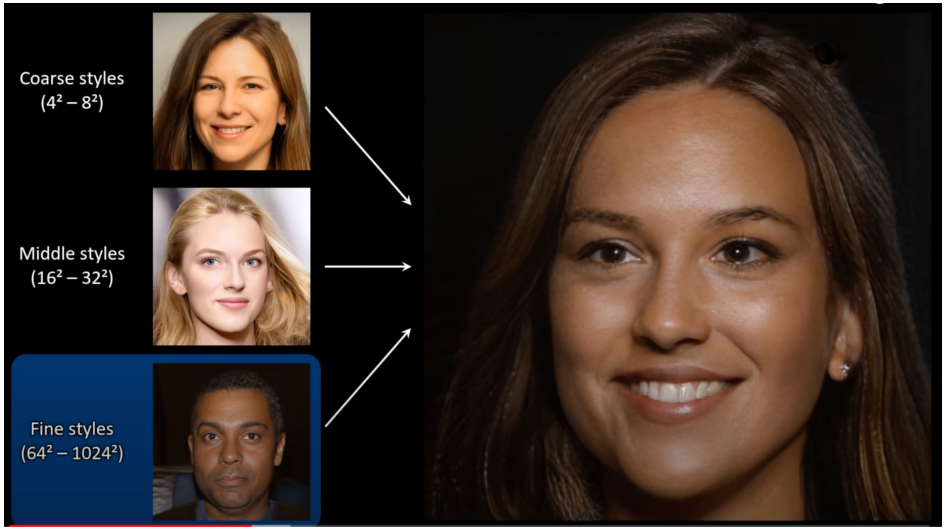


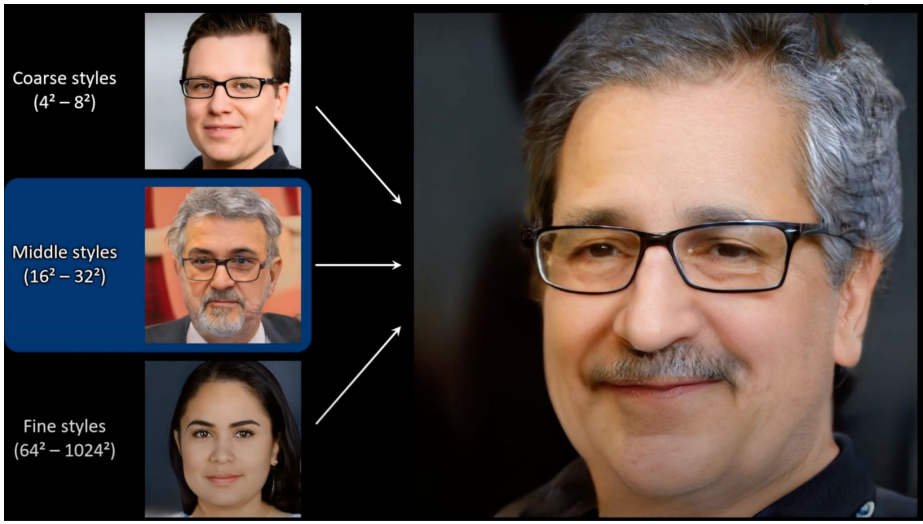
Source: M. Liu, T. Breuel and J. Kautz, "Unsupervised Image-to-Image Translation Networks," 2017.

StyleGAN

- A new architecture for Generative adversarial networks adapted from the style transfer literature.
- Depicts an image as a collection of styles which control effects at particular scales.
- Allows for unsupervised separation of high level attributes in an image -
    - Coarse styles - Pose, hair, face shape
    - Middle styles - facial features, eyes
    - Fine styles - color scheme

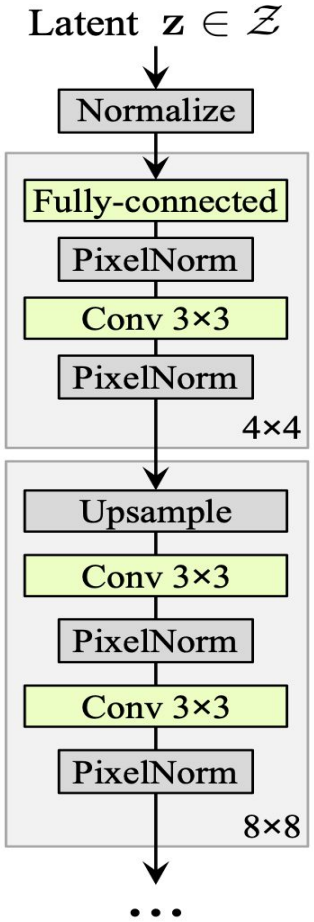

Source A: gender, age, hair length, glasses, pose
Source B: everything else
Result of combining A and B



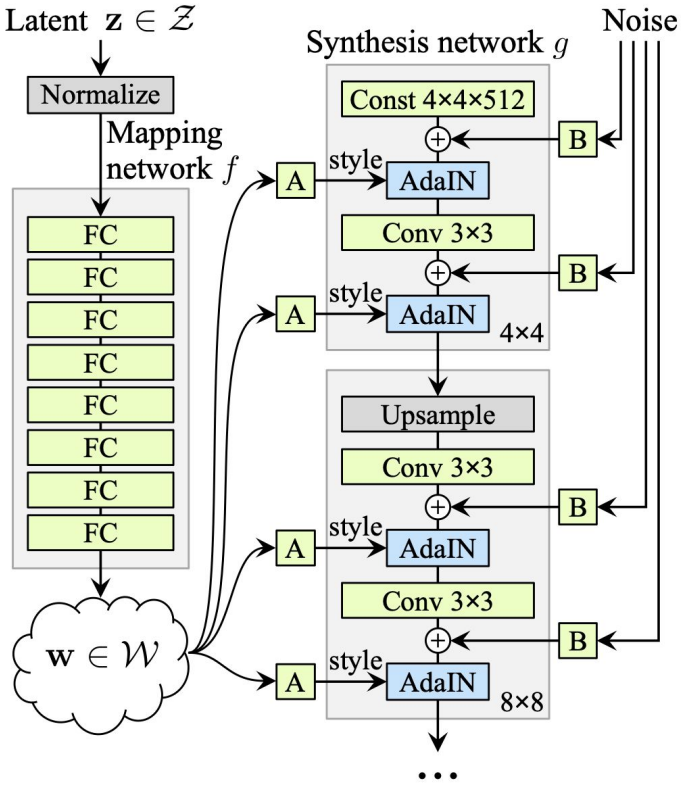


- Coarse styles - Pose, hair, face shape $(4_2 – 8_2)$
- Middle styles - facial features, eyes $(16_2 – 32_2)$
- Fine styles - color scheme $(64_2 – 1024_2)$

Traditional Generator Architecture

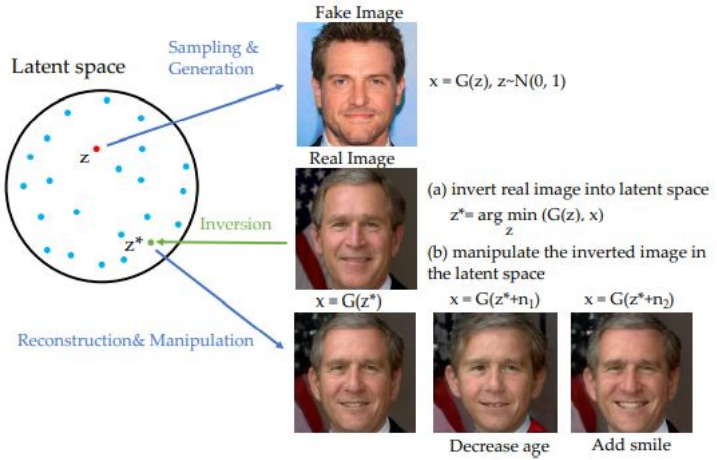


Style-based Generator



GAN Inversion

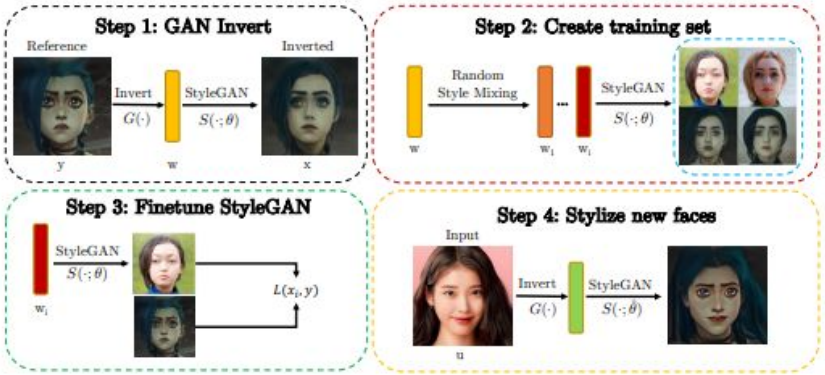
- Latent Space is a representation of data
- GAN inversion takes a generated image and attempts to construct the original image from the latent space.
- Used in JoJoGAN to create a dataset from few art samples



[14] Illustration of GAN Inversion.

JoJo GAN

- Built on StyleGan for Style Transfer (ST).
- ST takes the style of a reference image and applies it to a input image.
- A One-Shot Method.
- Finetunes StyleGan and performs layer swapping to achieve ST



[15] Steps of JoJo Gan

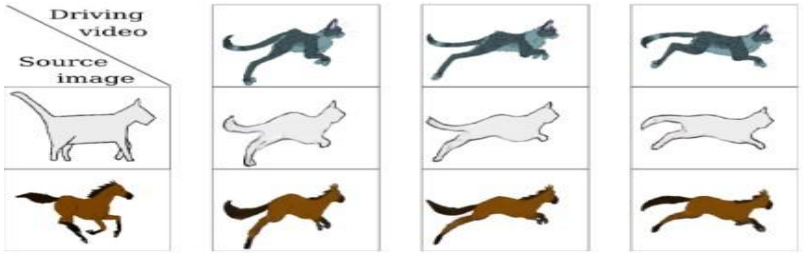
We looked at architectures and models which did image-image translations or modifications and generation.

Now let’s get things moving.

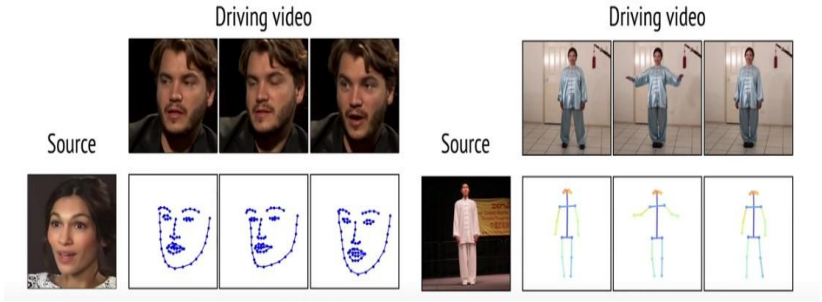
Image animation consists of generating a video sequence so that an object in a source image is animated according to the motion of a driving video.

First Order Motion Model

- It extracts and retargets motion from the driving video onto a source image.



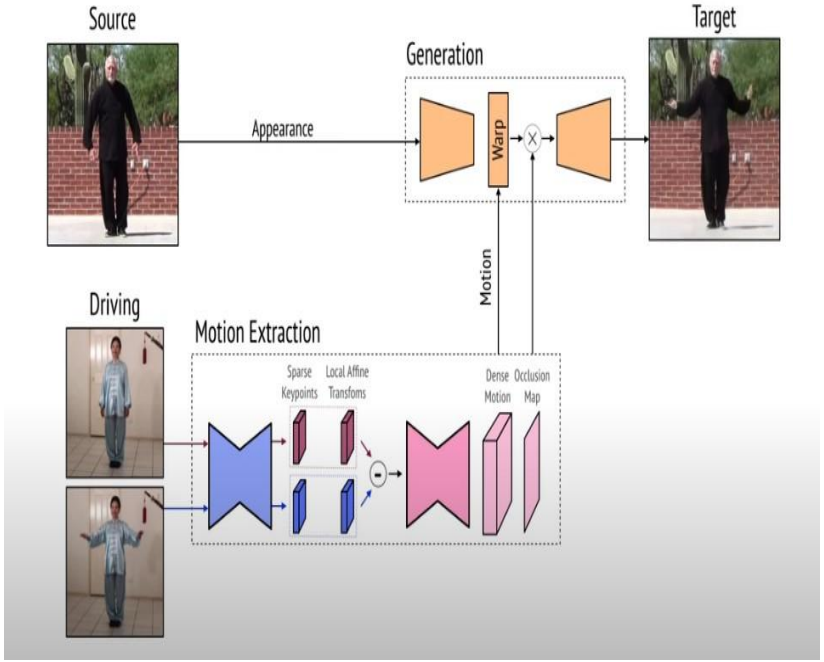
- The framework addresses this problem without using any annotation or prior information about the specific object to animate.

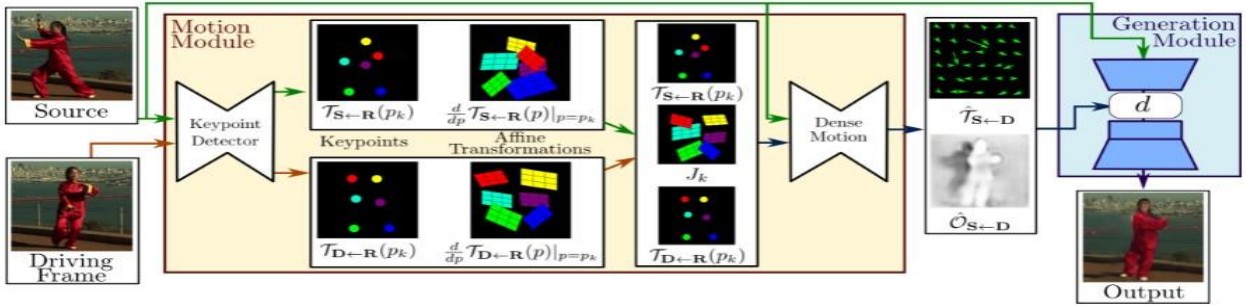


Working

- It splits the process into two:

  Motion Extraction and Generation.
- Consecutive frames from driving video are passed to unsupervised keypoint detector to get sparse keypoint for each frame.
- For each keypoint, local affine transformation in the neighborhood of these keypoints are approximated. The two outputs are subtracted to get sparse movement.
- To predict dense motion another network is used to find dense motion and occlusion map.
- The appearance is extracted from the source image by passing it through the encoder and then the features are warped using the dense motion and multiplied with the occlusion map this way the decoder knows which areas of the image require inpainting.





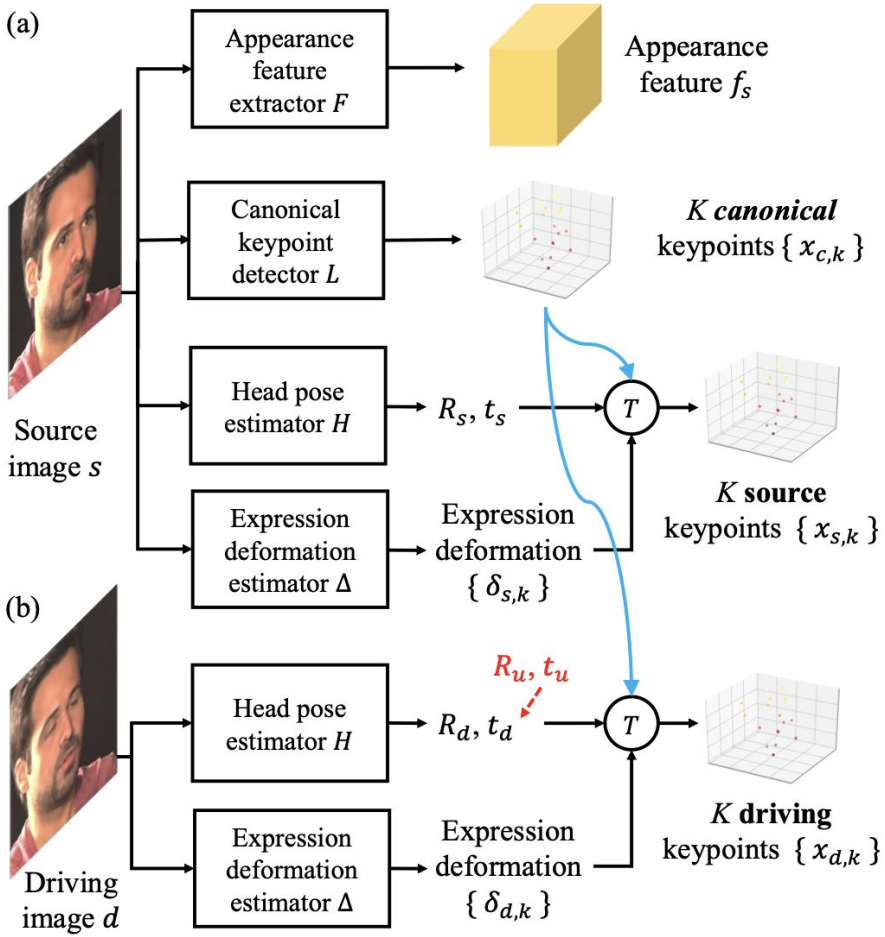
Learned Keypoints

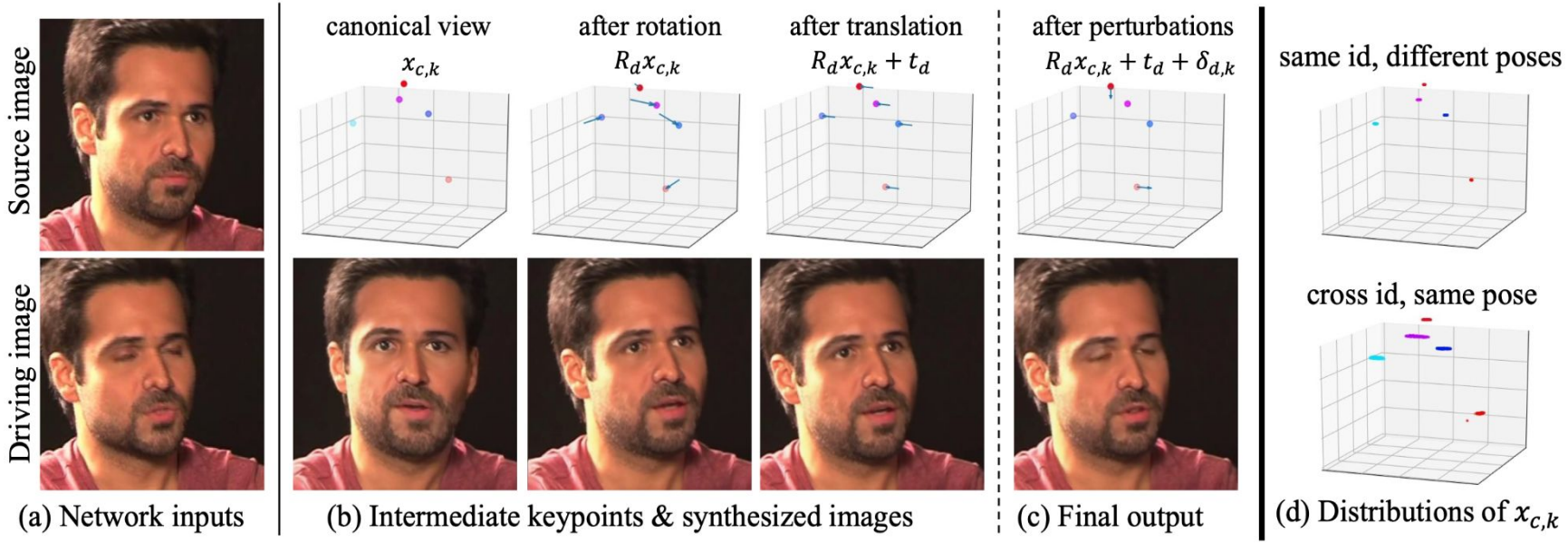
Free-View Neural Talking-Head Synthesis

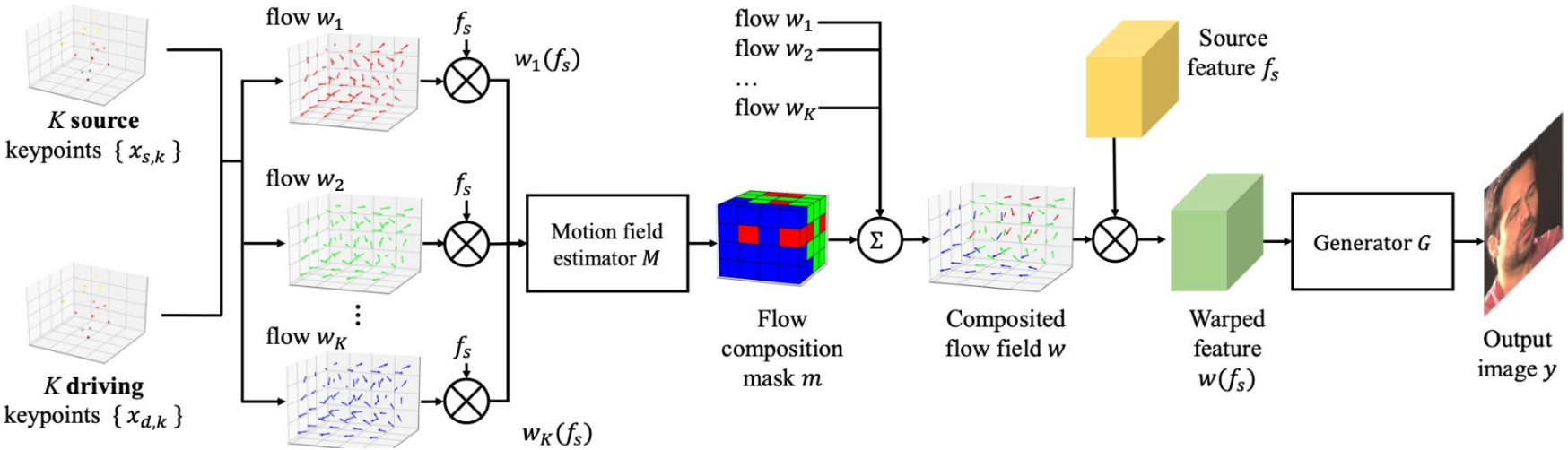
- A new model to generate a video of a talking head derived from a source image (appearance) and a driving video.
- The paper proposes a technique where fine-grained features derived from the source and target image allow for a much finer level of control over generated images.
- As the size of extracted features from the driving image is much lesser than the image, it allows for more efficient compression with video quality on par with commercial standards while using only one-tenth of the bandwidth.

Approach contains three major steps:

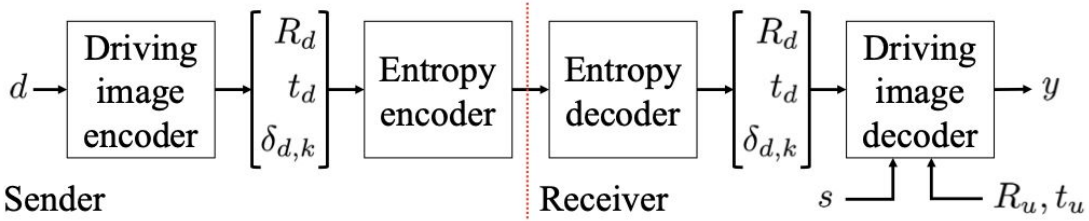
- Source image feature extraction,
- Driving video feature extraction,
- Video generation

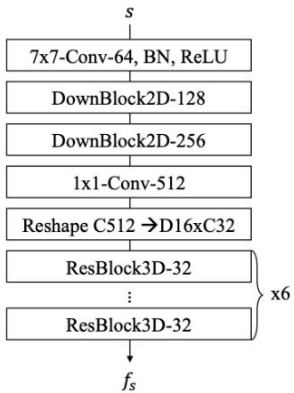

(a)
Appearance feature extractor F
Appearance feature f_s
Canonical keypoint detector L
K *canonical* keypoints $\{ x_{c,k} \}$
Source image s
Head pose estimator H
R_s, t_s
T
K **source** keypoints $\{ x_{s,k} \}$
Expression deformation estimator Δ
Expression deformation $\{ \delta_{s,k} \}$
(b)
R_u, t_u
Head pose estimator H
R_d, t_d
T
K **driving** keypoints $\{ x_{d,k} \}$
Expression deformation estimator Δ
Expression deformation $\{ \delta_{d,k} \}$
Driving image d


Source image
Driving image
canonical view
$x_{c,k}$
after rotation
$R_d x_{c,k}$
after translation
$R_d x_{c,k} + t_d$
after perturbations
$R_d x_{c,k} + t_d + \delta_{d,k}$
same id, different poses
cross id, same pose
(a) Network inputs
(b) Intermediate keypoints & synthesized images
(c) Final output
(d) Distributions of $x_{c,k}$

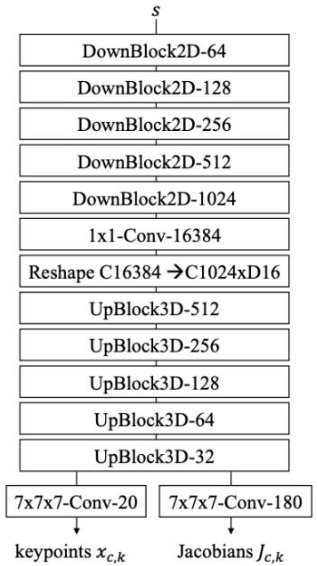

K source
keypoints { $x_{s,k}$ }
K driving
keypoints { $x_{d,k}$ }
flow w_1
flow w_2
flow w_K
f_s
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
flow w_1
flow w_2
…
flow w_K
Σ
Flow composition mask m
Composited flow field w
Source feature f_s
Warped feature $w(f_s)$
Generator G
Output image y

Video Compression framework

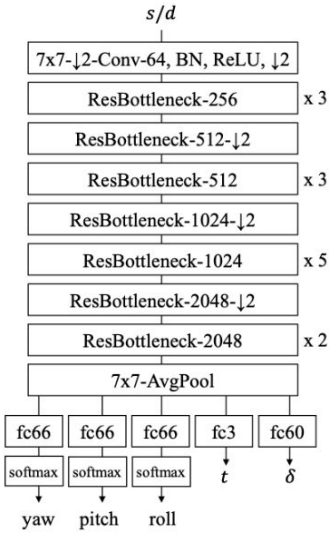




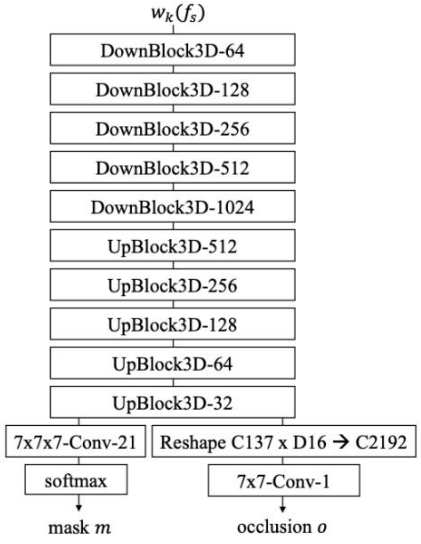
(a) Appearance feature extractor F



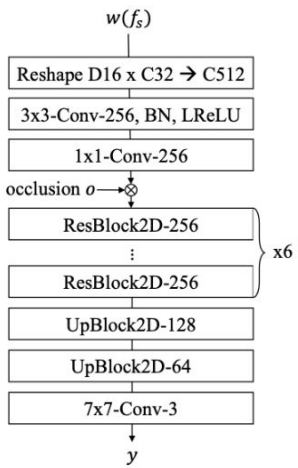
(b) Canonical keypoint detector L



(c) Head pose estimator H & expression deformation estimator Δ

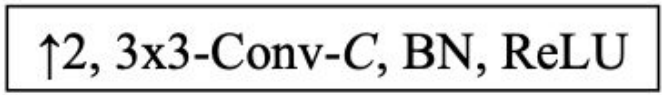


(d) Motion field estimator M

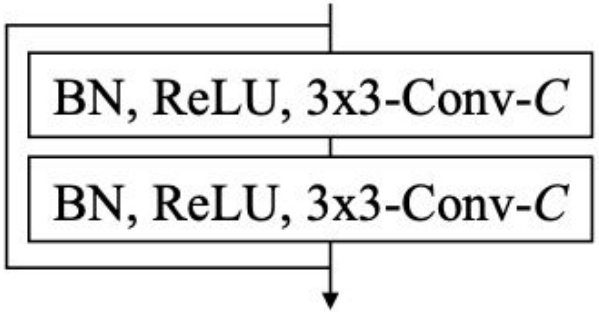


(e) Generator G

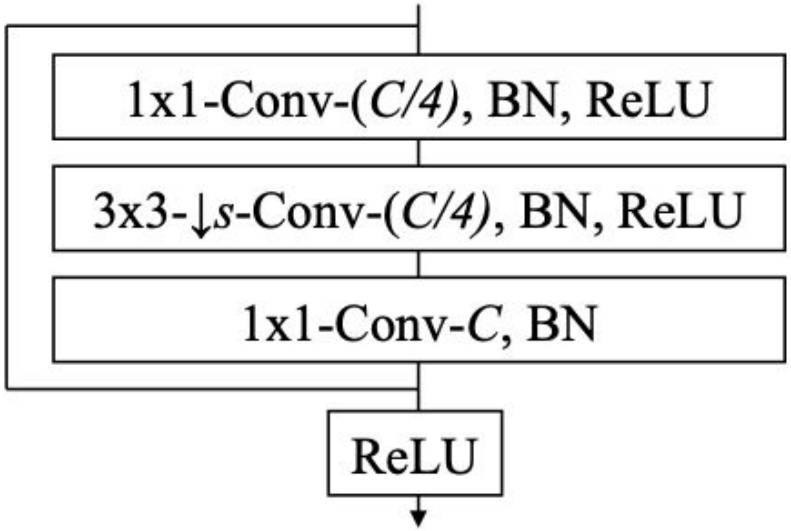
(a) DownBlock2D-*C*



(b) UpBlock2D-*C*



(c) ResBlock2D-*C*



(d) ResBottleneck-*C*-↓*s*

References

1. Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. “VoxCeleb2: deep speaker recognition”. In INTERSPEECH, 2018.
2. Wang, T.-C., Mallya, A., & Liu, M.-Y. (2021). “One-Shot Free-View neural talking-head synthesis for video conferencing”. CVPR.
3. Martin Heusel, et al, “GANs trained by a two time-scale update rule converge to a local nash equilibrium”. In NeurIPS, 2017 References
4. Davis E. King. “Dlib-ml: A machine learning toolkit”. JMLR, 2009
5. P. Burt and E. Adelson, "A multiresolution spline with application to image mosaics," ACM Transactions on Graphics, vol. 2, (4), pp. 217-236, 1983. . DOI: 10.1145/245.247.
6. L. A. Gatys, A. S. Ecker and M. Bethge, "Image style transfer using convolutional neural networks," in 2016. doi: 10.1109/CVPR.2016.265.
7. M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.
8. P. Isola, et al, “Image-to-image translation with conditional adversarial networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.
9. Langr, J., & Bok, V. (2019). “GANs in Action: deep learning with generative adversarial networks”. Manning. https://books.google.com/books?id=HojvugEACAAJ
10. Zhu, J.-Y., et al, “Unpaired image-to-image translation using cycle-consistent adversarial networks”. Computer Vision (ICCV), 2017 IEEE International Conference On.
11. R. Xu et al, "face transfer with generative adversarial network," 2017.
12. M. Liu, T. Breuel and J. Kautz, "Unsupervised image-to-image translation networks," 2017.
13. T. Karras, S. Laine, and T. Aila, “A Style-Based generator architecture for generative adversarial networks,” presented at the - 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4396–4405, doi: 10.1109/CVPR.2019.00453.
14. Xia. Weihao, et al, "GAN inversion: a survey," 2021, doi: arXiv:2101.05278
15. Chong. Min, and Forsyth. D.A, “JoJoGAN: one shot face stylization”, 2021 doi: arXiv:2112.11641
16. Pinkney, J.N., Adler, D.: “Resolution dependent gan interpolation for controllable image synthesis between domains”. arXiv preprint arXiv:2010.05334 (2020)
17. Siarohin, Aliaksandr, et al. "First order motion model for image animation." Advances in Neural Information Processing Systems 32, 2019. https://github.com/AliaksandrSiarohin/first-order-model
18. T.-C. Wang, A. Mallya, and M.-Y. Liu, “One-Shot Free-View neural talking-head synthesis for video conferencing,” presented at the - 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10034–10044, doi: 10.1109/CVPR46437.2021.00991.