Enabling Deep Hierarchical Image-to-Image Translation by Transferring from GANs

Jimut Bahan Pal [1]
Team name: **Zero1 (22D1594)**

Under the Guidance Of
P. Balamurugan [2]
[1] Centre for Machine Intelligence and Data Science
[2] Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

IE643 Course Project (2022)

Table of Contents

Outline of the presentation

This is the work done by **Yaxing Wang, Lu Yu and Joost van de Weijer**, presented at NeurIPS 2020 conference. In this presentation, we will be going through their work. The presentation is outlined as follows:

- Firstly, we discuss the main problem along with the background of the work.
- We also look into some of the existing work in the field of Image to Image (*I2I*) translations.
- We discuss the methodology for DeepI2I network, the novelties, the architectural choices and the loss functions used.
- Next, we discuss about the dataset used, the training and evaluation metrics.
- We further discuss about the training and experiments conducted from our side and the status of the work.
- We conclude the discussion by proposing some modifications to the existing work.

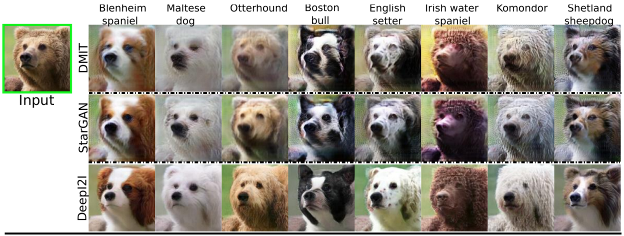
Table of Contents

1. Outline
2. Motivation
3. Current Problem
4. Experiments
5. Status of Work
6. Modifications proposed
7. References
8. Acknowledgements

Background of the problem

- *I2I* translation is an application of Computer Graphics (CG), used in movie industries widely (for e.g.: Morphing).
- This is a labor-intensive process. **The proposed technique can be used to automatically translate faces/objects between images**.
- Previous state-of-the-art method showed inferior performances when **translation between classes required large shape changes**.

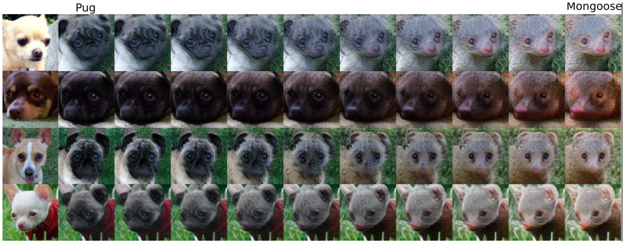
Any object can be **translated** to any other object by passing one image to the model. *This can also be used to create fake content, i.e.,* **deep fakes!!**



Background of the problem

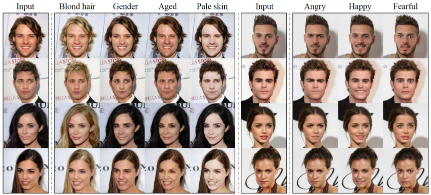
- First to implement **transfer learning framework using GANs** – this improves the performance on small dataset both qualitatively and quantitatively.
- They have done translation over 1000 classes in animal faces and food dataset.
- Proposed **hierarchical translation framework** which extracts **abstract semantic information in the deep low-resolution layers** of the network and **structural information from the shallow layers**.

Objects in images can also be **interpolated from one image to another with varying shape changes**, hence creating a **morphing effect**.

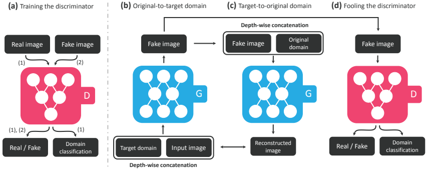


StarGAN: Unified GAN for Multi-Domain I2I translation. [1]

- Scalable approach that performs I2I translations for **multiple domains using a single model**.
- Simultaneous training of multiple domains with a single network.
- High visual quality compared to other methods at that time.
- Generator *G* takes both image and target domain label and generates fake label. It tries to reconstruct the original image from the fake image by using original domain label.



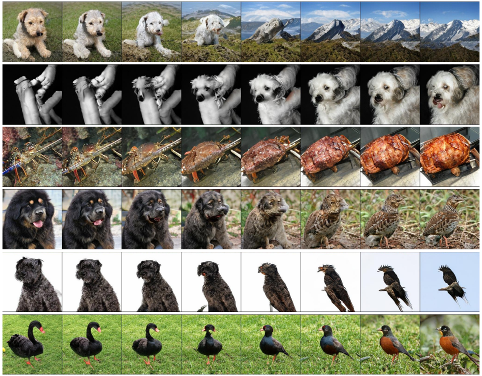
StarGAN Training on CelebA dataset (40 labels facial attribute, hair, eye color etc.) and transferring knowledge from RaFD (8 labels for facial expression, e.g. happy, angry etc.) dataset.



[1] Choi et al. 2017

Large scale GAN training for high fidelity natural image synthesis [2]

- **First to generate high resolution diverse complex samples from ImageNet dataset**.
- Previous architectures were not good at generating samples when the images were scaled up.
- **Class-conditional image synthesis**.
- Good Inception Score and Frechet Inception distance.
- This model can also be used for I2I translation between high resolution images.

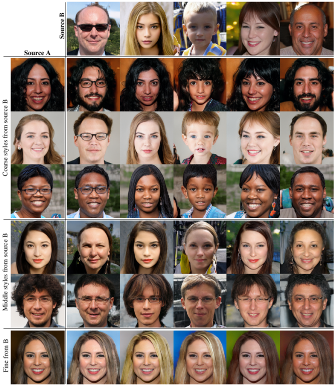


High quality image translations using BigGAN.

[2] Brock et al. 2018

A style based generator architecture for GANs [3]

- Borrowed ideas from style transfer literature to apply in the Generator of GANs.
- The architecture automatically learns unsupervised separation of high level attributes, e.g. pose, identity etc when trained on human faces.
- They build method for **synthesizing stochastic variation in generated images e.g., frekles, hair etc.** that enables scale specific control synthesis.
- Other popular architectures are **pix2pix** (2016), **CycleGAN** (2017), **SDIT** (Scalable and diverse cross domain image translation, 2019) and **DMIT** (Multi-mapping I2I translation via learning disentanglement, 2019).



[3] Karras et al. 2019

Table of Contents

1. Outline
2. Motivation
3. Current Problem
4. Experiments
5. Status of Work
6. Modifications proposed
7. References
8. Acknowledgements

Methodology for DeepI2I

- Current I2I architecture have a limited capacity to translate between classes with **significant shape changes** (e.g., from dog to meerkat face) due to the high resolution bottlenecks, which apply only **two down-sampling blocks** in most architectures.
- These models are successful in style transfer, but it is difficult to extract **abstract semantic information**, since these information are present in the **deep low-resolution layers** of a network.
- The main challenge lies in **inverting deep low-resolution bottleneck (latent) representation into a high-fidelity image**, since the deep **layers contains many attribute level information from which it is difficult to reconstruct realistic images which closely follow the structure of input images**.

Method Overview

- Let $\mathcal{X}, \mathcal{Y} = \mathbb{R}^{H \times W \times 3}$ be the source and target domains.
- The architecture is composed of four neural networks: **encoder Υ, adaptor Λ, generator Φ and discriminator Ψ**.
- They aim to learn a network to map the input source image $\mathbf{x} \in \mathcal{X}$ into a target domain image $\hat{\mathbf{y}} \in \mathcal{Y}$ conditioned on the target domain label $\mathbf{c} \in \{1, \ldots, C\}$ and a random noise vector $\mathbf{z} \in \mathbb{R}^{\mathbf{Z}}$, $\Phi\left(\Lambda\left(\Upsilon\left(\mathbf{x}\right)\right), \mathbf{c}, \mathbf{z}\right) \rightarrow \hat{\mathbf{y}} \in \mathcal{Y}$.
- Latent representation from different layers of encoder Υ is used to extract **structural information (shallow layers)** and **semantic information (deep layers).**
- Let $\Upsilon_l\left(x\right)$ be the l-th $(l = m, ..., n(n > m))$ ResBlock [4] output of the encoder, which is fed into the corresponding adaptor Λ_l, from which it continues as input to the corresponding layer of the generator.

[4] The encoder consists of a series of ResBlock. After each ResBlock the feature resolution is half of the previous one.

Method Overview

- A hierarchical representation $\Upsilon(\mathbf{x}) = \{\Upsilon(\mathbf{x})_l\}$ of input image $\mathbf{x}$ is extracted and fed to the adaptor network as input, that is $\Lambda(\Upsilon(\mathbf{x})) = \{\Lambda_l\}$, where (Λ_l) is the output of each adaptor Λ_l which is further summed to the activations of the corresponding layer of the generator Φ.
- In some cases when we train the DeepI2I from scratch, the adaptor could be the identity function.
- The generator takes as input the output of adaptor $\Lambda(\Upsilon(\mathbf{x}))$, the random noise $\mathbf{z}$ and the target label $\mathbf{c}$. The generator Φ outputs a $\hat{\mathbf{y}} = \Phi(\Lambda(\Upsilon(\mathbf{x})), \mathbf{z}, \mathbf{c})$ which is supposed to mimic the distribution of the target domain images with label $\mathbf{c}$.
- Sampling different $\mathbf{z}$ leads to diverse output results $\hat{\mathbf{y}}$.
- This way the adaptor of the network aligns the representations between the encoder and decoder of the BigGAN network to create meaningful translations, since there is no connection between them in pre-trained network.

Proposed architecture

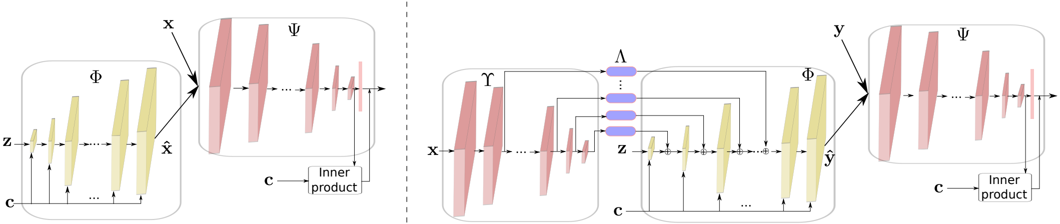


Figure 1: *Left*: the traditional form of conditional GAN (i.e., BigGAN) which contains the generator Φ and the discriminator Ψ. *Right*: the proposed DeepI2I method based on conditional GAN (left). The method consists of four terms: the encoder Υ, the adaptor Λ, the generator Φ and the discriminator Ψ. The encoder Υ is initialized by pre-trained discriminator (left), as well as both the generator Φ and the discriminator Ψ by pre-trained GANs (left). The adaptor Λ aims to align the pre-trained encoder Υ and the pre-trained generator Ψ.

Method Overview - Losses

- The discriminator has three functions:
  - **Distinguish** real target images from generated images.
  - **Guide the generator** Φ to synthesize images which belong to the class $\mathbf{c}$.
  - **Compute the reconstruction loss**, which aims to preserve a similar pose in both input source image $\mathbf{x}$ and the output $\Phi\left(\Lambda\left(\Upsilon\left(\mathbf{x}\right)\right), \mathbf{z}, \mathbf{c}\right)$.
- The reconstruction is computed from the discriminator based on the l-th ResBlock Ψ, and referred to by $\left\{\Psi_l\left(y\right)\right\}$.
- The overall loss is a multi-task objective comprising of:
  - **A conditional adversarial loss** - It optimizes the adverserial game between the generator and the discriminator, i.e., maximize Ψ while minimize $\{\Upsilon, \Lambda, \Phi\}$ to generate class specific images which correspond to label $\mathbf{c}$.
  - **Reconstruction loss** - guarantees that the synthesized image $\hat{\mathbf{y}} = \Phi\left(\Lambda\left(\Upsilon\left(\mathbf{x}\right)\right), \mathbf{z}, \mathbf{c}\right)$ preserve the same pose as the input image $\mathbf{x}$.

Method Overview - Losses

Conditional adversarial loss employing GANs

$$\mathcal{L}_{adv} = \mathbb{E}_{y\sim\mathcal{Y}}\left[\log\Psi\left(\mathbf{y},\mathbf{c}\right)\right] + \mathbb{E}_{\hat{\mathbf{x}}\sim\mathcal{X},\mathbf{z}\sim p(\mathbf{z}),\mathbf{c}\sim p(\mathbf{c})}\left[\log(1-\Psi\left(\Phi\left(\Lambda\left(\Upsilon\left(\mathbf{x}\right)\right),\mathbf{z},\mathbf{c}\right),\mathbf{c}\right)\right]$$

Here $\mathbf{p}\left(\mathbf{z}\right)$ follows the normal distribution , and $\mathbf{p}\left(\mathbf{c}\right)$ is the domain label distribution.

Final loss is optimized by mini-max game

$$\{\Upsilon,\Lambda,\Phi,\Psi\} = \arg\min_{\Upsilon,\Lambda,\Phi}\max_{\Psi}\mathcal{L}_{adv}.$$

Method Overview - Losses

Reconstruction Loss- based on set of activations extracted from multiple layers of discriminator Ψ

$$\mathcal{L}_{rec} = \sum_{l} \alpha_l \left\| \Psi\left(\mathbf{x}\right) - \Psi\left(\hat{\mathbf{y}}\right) \right\|_1$$

Here parameters α_l are scalars which balance the terms, are 0.1 except for $\alpha_3 = 0.01$. Note that this loss is only used to update the encoder Υ, adaptor Λ, and generator Φ.

Full objective function of the model

$$\min_{\Upsilon,\Lambda,\Phi} \max_{\Psi} \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec}$$

Here both λ_{adv} and λ_{rec} are hyper-parameters that balance the importance of each terms.

Table of Contents

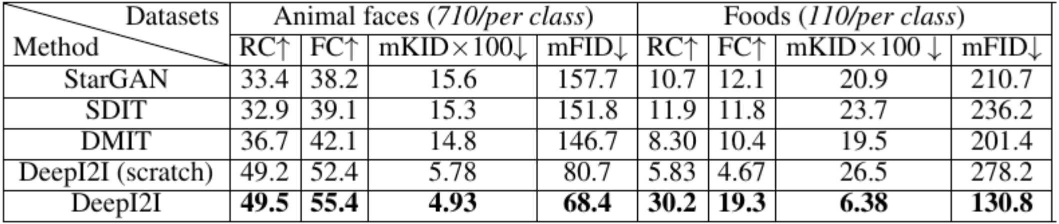
Datasets and Frameworks

- Three datasets were used for this study:
  - **Animal faces** - contains 117,574 images and 149 classes in total.
  - **Foods** - consists of 31,395 images and 256 classes in total
  - **cat2dog** - composes of 2235 images and 2 classes in total.
- The images were resized to 128×128, and divided by **255.0**.
- The dataset is split into training set **(90 %) and test set (10 %)**.
- The datasets were pre-processed to **.HDF5 files for faster I/O** and batch-sized pre-processing, this was borrowed from BigGAN code.
- The code is done in **Python3 and Pytorch1.12.1 framework** was used in this study.

Evaluation metrics

- Four evaluation metrics were considered for this training:
  - **Fréchet Inception Distance (FID)** - measures the similarity between two sets in the embedding space given by the features of a convolutional neural network.
  - **Kernel Inception Distance (KID)** - calculates the squared maximum mean discrepancy to indicate the visual similarity between real and synthesized images.
  - We further evaluate on two metrics of translation accuracy to show the ability of the learned model to synthesizing the correct class-specific images, called **real classifier (RC) (trained on real data and evaluated on the generated data)** and **fake classifier (FC) (trained on the generated samples and evaluated on the real data)**.
- The mean values of all categories in terms of **FID** and **KID** is denoted as **mFID** and **mKID**.

Training and Evaluation metrics



Datasets	Animal faces (*710/per class*)				Foods (*110/per class*)			
Method	RC↑	FC↑	mKID×100↓	mFID↓	RC↑	FC↑	mKID×100 ↓	mFID↓
StarGAN	33.4	38.2	15.6	157.7	10.7	12.1	20.9	210.7
SDIT	32.9	39.1	15.3	151.8	11.9	11.8	23.7	236.2
DMIT	36.7	42.1	14.8	146.7	8.30	10.4	19.5	201.4
DeepI2I (scratch)	49.2	52.4	5.78	80.7	5.83	4.67	26.5	278.2
DeepI2I	**49.5**	**55.4**	**4.93**	**68.4**	**30.2**	**19.3**	**6.38**	**130.8**

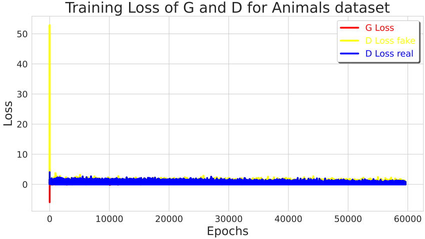
- These are the main results which shows that DeepI2I scratch performs best in case of Animal faces dataset.
- When transfer learning is applied, the metrics improves significantly for the Foods dataset.

Table of Contents

1. Outline
2. Motivation
3. Current Problem
4. Experiments
5. **Status of Work**
6. Modifications proposed
7. References
8. Acknowledgements

Results of experiments conducted

- We were sucessfully able to reproduce the results as shown in the paper.
- We were also able to generate some samples between the training and generate the videos in the transitions.
- We are planning to add one more dataset for the final review.
- The more we train the more the images looks real, but there might be a chance of mode collapse.
- **Show the videos generated**.



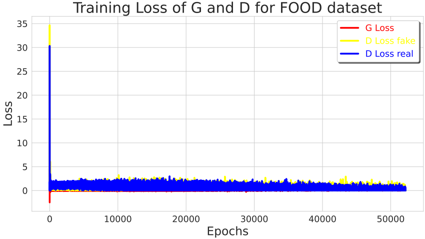


Table of Contents

1. Outline
2. Motivation
3. Current Problem
4. Experiments
5. Status of Work
6. Modifications proposed
7. References
8. Acknowledgements

Modifications proposed

- We are planning to run the model on **NABirds datasets** and also check different types of **Loss functions**. (`https://dl.allaboutbirds.org/nabirds`)
- The dataset will need to be converted to .HDF5 format for faster pre-processing which is required by this architecture.
- This dataset is a collection of 48,000 annotated photographs of the 400 species of birds that are commonly observed in North America.
- Over 100 photographs are available for each species, including separate annotations for males, females and juveniles that comprise 700 visual categories.

Table of Contents

1. Outline
2. Motivation
3. Current Problem
4. Experiments
5. Status of Work
6. Modifications proposed
7. References
8. Acknowledgements

References I

Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 8789–8797. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00916. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Choi_StarGAN_Unified_Generative_CVPR_2018_paper.html.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, Computer Vision – ECCV 2018, pages 179–196, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01219-9.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.OpenReview.net, 2019. URL: https://openreview.net/forum?id=B1xsqj09Fm

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4396–4405, 2019.doi: 10.1109/CVPR.2019.00453.

Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. DRIT++: diverse image-to-image translation via disentangled representations. Int. J. Comput. Vis., 128(10): 2402–2417, 2020. doi: 10.1007/s11263-019-01284-z. URL: https://doi.org/10.1007/s11263-019-01284-z.

Justin N. M. Pinkney and Doron Adler. Resolution dependent GAN interpolation for controllable image synthesis between domains. CoRR, abs/2010.05334, 2020. URL: https://arxiv.org/abs/2010.05334.

References II

Yaxing Wang, Abel Gonzalez-Garcia, Joost van de Weijer, and Luis Herranz. SDIT: scalable and diverse cross-domain image translation. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi, editors, Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019, pages 1267–1276. ACM, 2019. doi: 10.1145/3343031.3351004. URL: https://doi.org/10.1145/3343031.3351004.

Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: Effective knowledge transfer from gans to target domains with few images. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 9329–9338. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00935. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_MineGAN_Effective_Knowledge_Transfer_From_GANs_to_Target_Domains_With_CVPR_2020_paper.html.

Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas H. Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché- Buc, Emily B. Fox, and Roman Garnett, editors, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 2990–2999, 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/5a142a55461d5fef016acfb927fee0bd-Abstract.html.

Table of Contents

Acknowledgements

It is ritual that scholars express their gratitude to their instructors. This acknowledgement is very special to me to express my deepest sense of gratitude and pay respect to my instructor, **P. Balamurugan**, Department of Industrial Engineering and Operations Research, for his constant encouragement, guidance, supervision, and support throughout the completion of my project. His close scrutiny, constructive criticism, and intellectual insight have immensely helped me in every stage of my work. I would like to thank him for patiently answering my often-naive questions related to Computer Vision.

I would like to thank my thesis supervisor **Suyash P. Awate** for arranging free compute resources and help me select such a wonderful course. I would also like to thank the **awesome T.A.'s** for the stimulating discussions that they had shared with me during the Deep Learning Theory and Practices course. Finally I would like to thank my father **Dr. Jadab Kumar Pal**, Indian Statistical Institute, for supporting me.

Any Questions . . . ?

Thank You

22d1594@iitb.ac.in
jimutbahanpal@yahoo.com