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. The presentation is outlined as follows:

- Problem statement.
- Summary of the work done before mid-term review, and the major comments given during the same.
- Issues that occurred while implementing some of the comments, and the major work that was done after the mid-term review.
- Conclusions and possible future directions.

Table of Contents

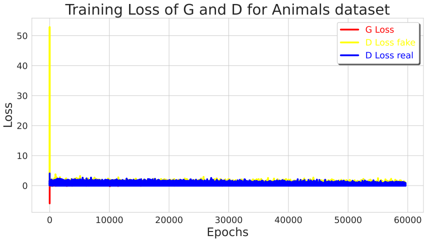
Problem Statement

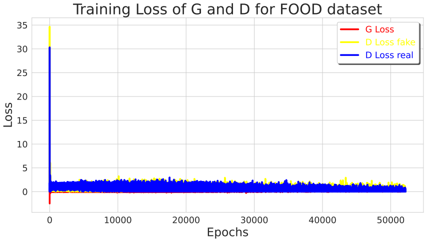
- *I2I* translation is an application of Computer Graphics (CG), used in movie industries widely (for e.g.: Morphing).
- **The proposed technique can be used to automatically translate faces/objects between images**.
- Previous state-of-the-art method showed inferior performances when **translation between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

Table of Contents

Summary of work done before mid term review

- Reproduced the results as shown in the paper.
- Generated some samples between the training and generated the videos in the transitions.
- Planned 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.





Summary of work done before mid term review

- Generated samples were not very good.
- Translated samples were more or less of the same style, hence there were issues of mode collapse.
- Hyper-parameter tuning could be experimented a bit.

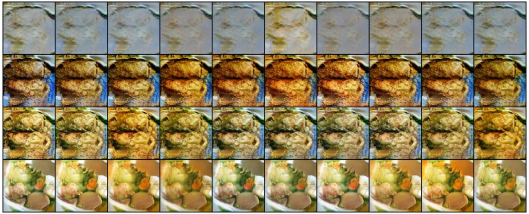


Table of Contents

Major comments given during the mid-term project review

Instructor:

- New loss functions like SI-SDR, SSIM can be tried.
- In the final presentation, the proposed modifications can be demonstrated with new loss functions.

TAs:

- Modification proposed includes working on a different data set and modification in loss function.
- Class names should be included in the images when displayed for translation.

Table of Contents

How the team has addressed the comments given during mid-term project review

- **Instructor** - New loss functions like SI-SDR, SSIM can be tried - experimented with these loss function, but didn't give any significant results (added 1-SSIM in Discriminator).
- **Instructor** - In the final presentation, the proposed modifications can be demonstrated with new loss functions - A different loss function is tried which was integrated with discriminator. - **Gives slightly better results!!**
- **TAs** - Class names should be included in the images when displayed for translation - done.
- **TAs** - New dataset could be tried in the final experimentation - done.

On the new recommended loss functions

- Structural Similarity (SSIM) [1] is a measurement of how degraded an image is, by comparing two images.
- Scale Invariant Signal to Distortion Ratio (SI-SDR) [2] is used in speech enhancement and source separation.
- The issue is, these methods need two images to be present for comparing and evaluating a deterministic output.
- For our output, the latent vector learns a distribution, by using a single activation from the discriminator while computing the loss.

[1] Zhou et al., Image Quality Assessment: From Error Visibility to Structural Similarity (2004).
[2] Roux et al., SDR – HALF-BAKED OR WELL DONE? (2018)

Proposed architecture

The reconstruction loss is computed by taking activations from the different layers of the network.

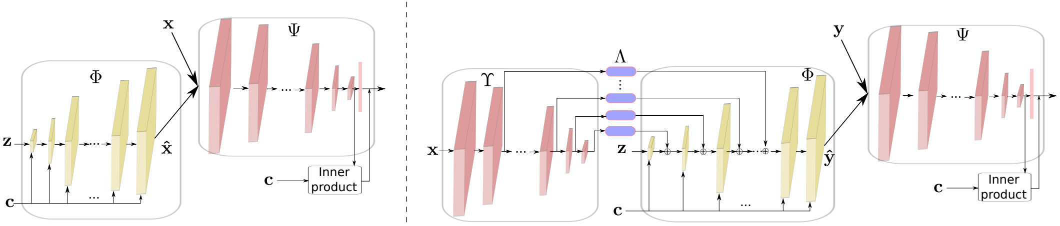


Figure 1: *Left*: the traditional form of conditional GAN (i.e., BigGAN) which contains the generator Φ and the discriminator Ψ. *Right*: the proposed DeepI2I method based on conditional GAN (left). The method consists of four terms: the encoder Υ, the adaptor Λ, the generator Φ and the discriminator 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.

Table of Contents

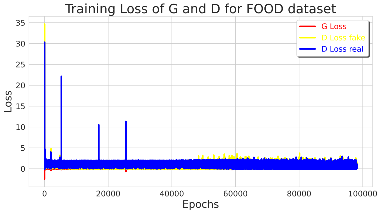
Work done after mid-term project review

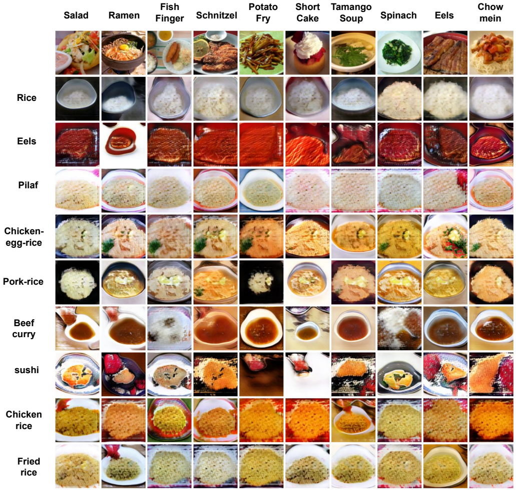
- System: 8 x NVIDIA GeForce RTX 2080 Ti GPUs with Intel Xeon Gold 6130 @ 64x 2.101GHz processor, 5.4 TB space of solid-state drive, Ubuntu 18.04 LTS Operating system and a main memory of 128 GB (RAM).
- Training for Foods dataset took about 5 days for 98000 iterations, NABirds dataset took about 10 days for 151700 iterations and Animals dataset took about 17 days for 367138 iterations.
- The model consists of Generator = 70.43 M, Discriminator D = 87.98 M, Encoder = 87.98 M and Adaptor = 87.36 M parameters.
- Batch size of 4 was used, a learning rate of 1e-04 was used for the Generator and a learning rate of 4e-04 was used for the discriminator.

NABirds Dataset

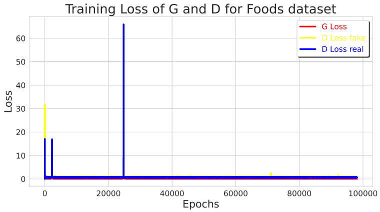
- The dataset is converted to .HDF5 format for faster pre-processing which is required by this architecture, i.e., 128x128 sized images in binary.
- This dataset is a collection of 48,000 annotated photographs of the 400 species of birds that are commonly observed in North America.
- Over 100 photographs are available for each species, including separate annotations for males, females and juveniles that comprise 700 visual categories.
- Different types of eagles are shown below:

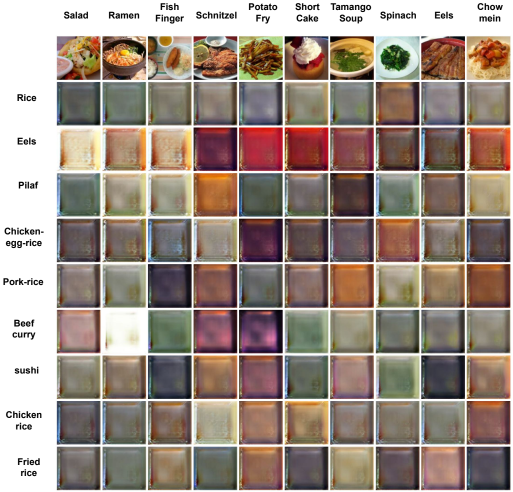
For Foods dataset (Normal)



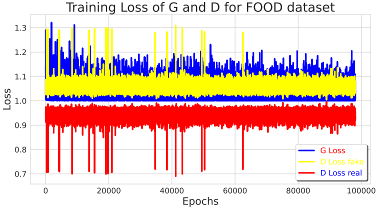


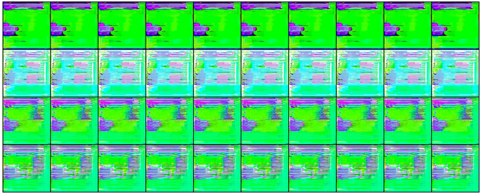
For Foods dataset (Our loss - Unable to generate properly)

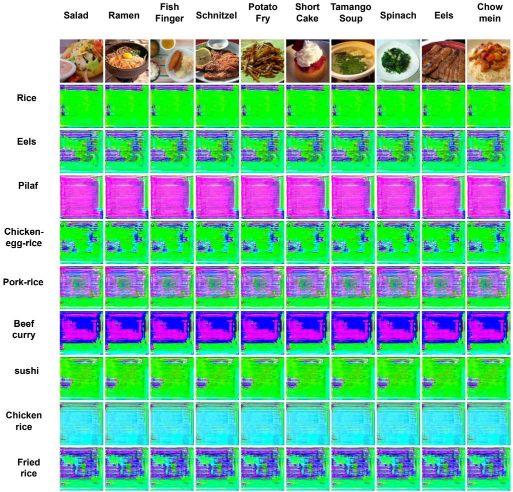




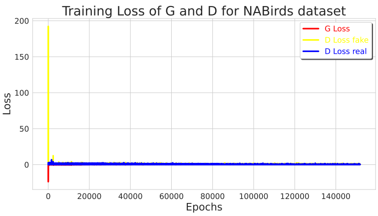
For Foods dataset (SSIM loss - Unable to generate properly)

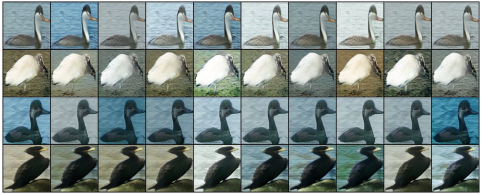


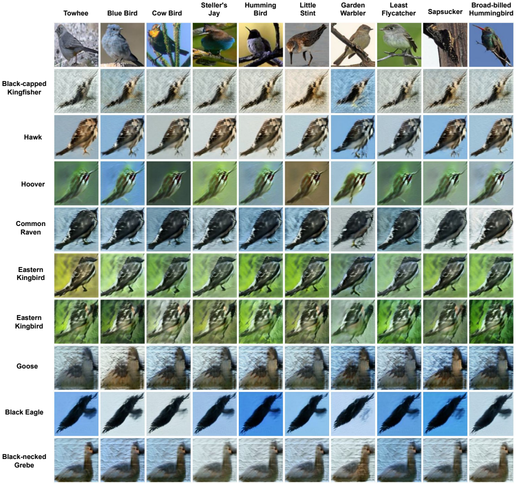




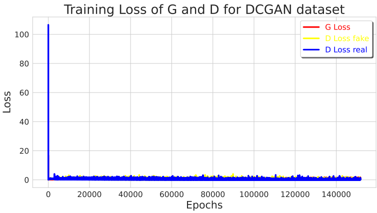
For NABirds dataset (Normal)

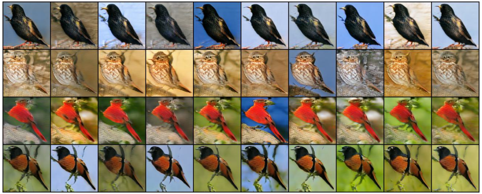


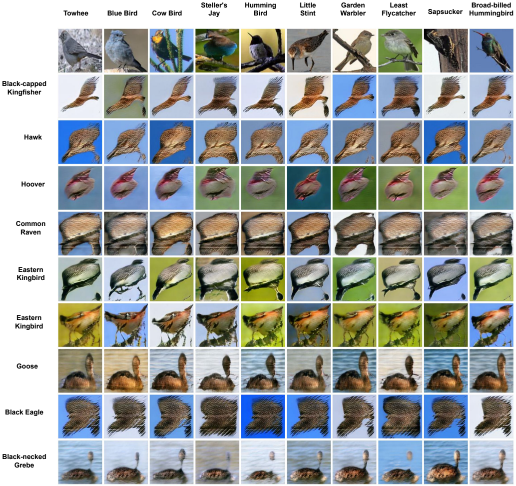




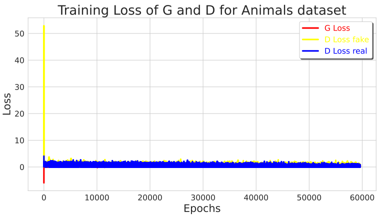
For NABirds dataset (SoftPlus loss - Good Results)

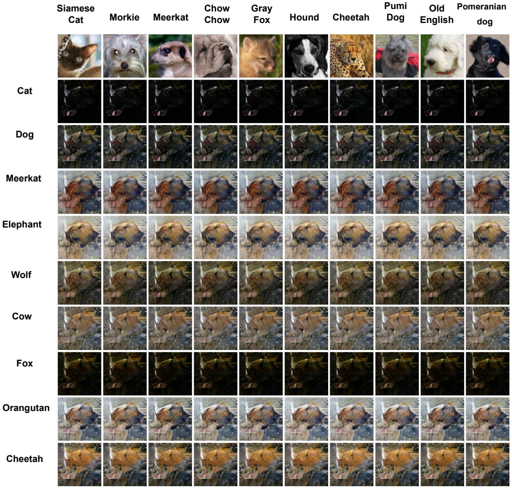




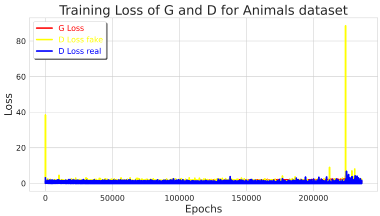


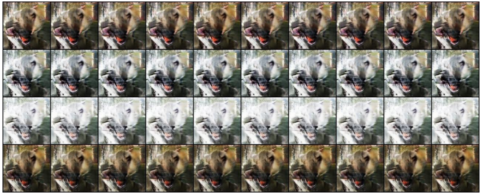
For Animals dataset (Normal - Mode collapse)

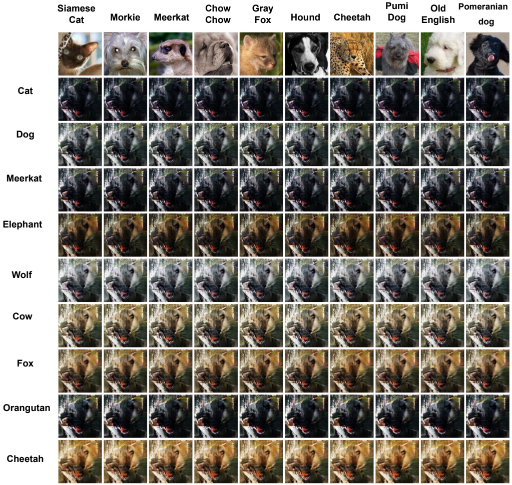




For NABirds dataset (SoftPlus loss - Mode collapse)







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

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

Losses

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

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

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.

Final Results

	RC ↑	FC ↑	mKIDx100 ↓	mFID ↓
Animal (Ori. Loss)	49.2	52.4	5.78	80.7
Food (Ori. Loss)	5.83	4.67	26.5	278.2
Birds (Ori. Loss)	3.24	5.84	30.5	301.7
Birds (Our Loss - SoftPlus)	**3.57**	**5.93**	**30.71**	**301.9**

- **Fréchet Inception Distance (FID)** - similarity between two sets in the embedding space given by the features of a convolutional neural network.
- **Kernel Inception Distance (KID)** - squared maximum mean discrepancy to indicate the visual similarity between real and synthesized images.

Table of Contents

Conclusions

- GAN training can be an extremely difficult process and is prone to mode collapse problem.
- Designing of new loss function needs additional constraints apart from direct theoretical derivations.
- Successfully reproduced the code, tried new dataset and designed a loss function.

Table of Contents

Future directions

- The SSIM and SI-SDR loss functions can be tried with VAE based Generative networks (but getting comparable results might be very difficult).
- The quality of the images could be increased more, like 1080x1080 px, by using different up-sampling architectures.

Table of Contents

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