Ensembled Approach To Cross View Image Translation Using Combination Of SelectionGAN And CycleGAN

Abhishek Sundar Raman - Gayatri Ganapathy - Ria Gupta - Varnnitha Venugopal
Department of Computer Science
Simon Fraser University
8888 University Dr, Burnaby, BC V5A 1S6
asundarr@sfu.ca, gganapat@sfu.ca, riag@sfu.ca, vvenugop@sfu.ca

Abstract

Enhancing the image quality and improving the evaluation metrics has always been the main subject for images generated through GANs. Our project proposes a novel approach, EnsembledGAN, for improving evaluation metrics and the quality of images generated from SelectionGAN. A SelectionGAN produces a street-view image as output, given an aerial view image and semantic map of the side view as inputs. EnsembledGAN consists of 2 stages. In stage 1, SelectionGAN is used which produces the street-view as the output. In stage 2, we attempt to improve the image quality by providing the output of SelectionGAN as input to the CycleGAN. The CycleGAN is a two-way GAN which is trained on the output of SelectionGAN and ground truth. It learns how to convert a SelectionGAN generated output closer to a ground truth image and vice versa. This enables the CycleGAN to generate better quality images. These experiments were performed on CVUSA dataset and the results show that our EnsembledGAN model is able to improve the evaluation metrics in terms of Top-k classes accuracy, Inception score and KL score when compared with the SelectionGAN model.

1 Introduction and Motivation

Improving the evaluation metrics and the quality of the image generated from CNNs and GANs has always been the subject of active research. In the past, we have seen [4] [1] that attempts have been made to obtain better quality images through the usage of CNNs and GANs. Our EnsembledGAN approach is an attempt to help improve the quality of the final cross-view image generated from SelctionGAN [7] along with the improvements in the evaluation metrics such as Top-k classes accuracy, Inception score and KL score.

A CycleGAN [13] is an image-to-image translation system with a boundless number of applications. One of the applications of this paradigm includes enhancing an image with shallower depth of field at the pixel level. Earlier works also show that CNNs could be employed to generate high-quality photos with shallower depth of field [4]. In the CycleGAN [13] the model was trained on flower photos downloaded from Flickr dataset. The source domain consisted of flower photos taken by smartphones, which usually have deep depth of field due to a small aperture. The target contains photos captured by DSLRs with a larger aperture. CycleGAN [13] was able to successfully generate photos with shallower depth of field from the photos taken by smartphones.

CycleGAN [13] works with images that mostly involve colour and texture changes, and is tailored for good performance on appearance changes. Moreover, both our input and output domains during training do not have much varied and extreme transformations for which CycleGAN generally achieves good results.

We propose a novel approach of ensembling two existing GAN models SelectionGAN [7] and CycleGAN [13] to obtain enhanced images closer to the ground truth. In EnsembledGAN model, we are experimenting on the above discussed technique for improving the evaluation metrics of the image generated by SelectionGAN [7] by capturing special characteristics of one image collection and figuring out how these characteristics could be translated into the other image collection. This approach tries to resolve unsatisfactory aspects in the generated scene structure and details of the image generated by SelectionGAN [7]. Using EnsembledGAN approach we are able to improve some of the evaluation metrics as well as the quality of the final image generated from the model by appending the trained CycleGAN [13] model to the existing SelectionGAN [7].

2 Related Work

2.1 GANs

Generative Adversarial Networks (GANs) [3] have proved to produce better quality images for the image to image translation tasks [6]. GANs consists of a generator used to produce images closer to the real images using a noise vector and discriminator tries to distinguish between the produced image and the real image. By varying the structure of GANs and the inputs provided to the GAN, the model performance can be improved. We have also seen improvements in the quality of images produced by GANs when provided with low-quality images [1].

2.2 View-Point Translations

Early works in view-point translation applications involve generating one view of an object given another view of the object [2] [8] [10]. Other applications of view-point translation involve generation of street-view of images for the provided aerial view [6] [12]. This is a challenging task as the aerial image covers a larger area whereas a street-view consists of more details of a specific area. Secondly, similar appearing images in aerial view may not be the same in the street-view. One of the earlier works in this field is by Predicting Ground-Level Scene Layout from Aerial Imagery by Zhai et al [12]. This approach uses a Convolutional Neural Network to extract features from aerial imagery to generate ground-level panoramic images. Another approach by Krishna Regmi et al [6] using ConditionalGANs proposes two architectures X-Fork and X-Seq. Both these architectures are built on the underlying baseline that generating semantic maps along with the street-view images would improve the quality of generated output [6]. Hao Tang et al [7] worked on the approach that providing semantic maps as input to the GAN instead of generating them, would lead to better results.

2.3 Image Enhancement

Prior work in image enhancement involving Deep Convolutional Networks [4] uses a direct mapping between the poor and better quality images to improve the colour rendition and image sharpness. The approach by Nelson Chong et al [1] uses a CycleGAN. In this method, two generator-discriminator pairs are used to generate target images from input images and to generate back the input images using the target images. The loss measure between the original input image and the generated input image is used to improve the model. Since CycleGANs do not need pair information of input and target, it can perform better on a variety of input sets. Jun-Yan et al [13] used CycleGAN to generate flower photographs of DSLR camera quality using a dataset of photographs taken from smartphones.

3 Approach

3.1 Datasets

CVUSA dataset [9] consists of matching ground panoramas with their corresponding annotations and aerial view satellite images. It contains 35532 image pairs for training and 8884 image pairs for testing. In the data pre-processing step, we follow [11] [6] where the aerial images are center-cropped to 224×224 and resized to 256×256. The coloured semantic annotations maps are obtained by providing annotation images as input to the MATLAB script.

The input image to the EnsembledGAN model is generated by concatenating four images as follows: an aerial-view satellite image, the ground truth or the street-view image, a black image which acts as a placeholder for the semantic map of the aerial-view image and the coloured annotated semantic map of the street-view. We consider a quarter of an image from the panoramic street-view image and the coloured annotation maps for the above dataset preparation. All images are scaled to 256×256 image size in order to enable image flipping and random crops for data augmentation while the model is being trained.

3.2 EnsembledGAN Model

The proposed EnsembledGAN model is formed by a combination of two existing GANs - CycleGAN [13] and SelectionGAN [7] for the suggested approach to improve the evaluation metrics as well as the quality of the image generated by this method when compared with the existing SelectionGAN [7].

3.2.1 Architecture

The architecture for our EnsembledGAN after combining SelectionGAN [7] and CycleGAN [13] is as follows:

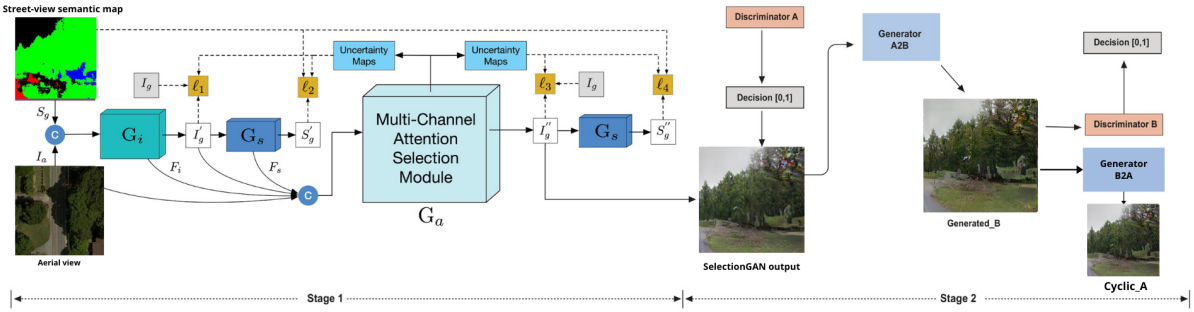

Figure 1: EnsembledGAN architecture involving SelectionGAN [7] and CycleGAN [13]

It is composed of two stages.

Stage 1:

The SelectionGAN [7] module has the following network. First, the input aerial image and the annotated semantic map are concatenated and provided into Generator G_i to produce target image I'_g. This I'_g is fed into the semantic generator G_s to get new semantic map S'_g. The G_i and G_s generators use U-Net [5] neural architecture which is a network with skip connections between a downsampling encoder and an up-sampling decoder. This architecture helps to retain detailed contextual and textural information. The model makes use of a deeper G_i network compared to G_s as cross-view image generation is a more important task as it is related to the final output to be generated by the network. Thus, the number of filters in first convolution layers of G_i and G_s are 64 and 4 respectively. PatchGAN [5] is adopted for discriminator D.

The I'_g and S'_g generated are passed into multi-channel attention selection module to generate ten intermediate images and attention maps. The uncertainty maps are generated by concatenation of these attention maps which help to guide pixel loss for effective optimization. The neural network here attentively selects parts which are more important for generating a scene image with a new viewpoint. Further, in the network, total variational regularization (spatial smoothness to denoise the output image) is applied to the final image and multi-scale spatial pooling is employed to refine the features.

Stage 2:

We append an additional CycleGAN [13] model to stage 1 model. The model takes an input image from Domain A (output from SelectionGAN) to generate an image in the target Domain B(ground truth street-view image) via GeneratorA2B. This newly generated image is then fed to another GeneratorB2A which converts it back into the corresponding image of Domain A (Cyclic A). Two inputs are provided into DiscriminatorA - one is the ground truth of street-view and the other is the generated image from GeneratorA2B. The DiscriminatorA tries to distinguish between them and takes a decision either to accept or reject the images generated by GeneratorA2B. The GeneratorA2B keeps on generating images close to images in Domain B until the DiscriminatorA accepts them. This happens vice versa for DiscriminatorB.

3.2.2 Working

Training:

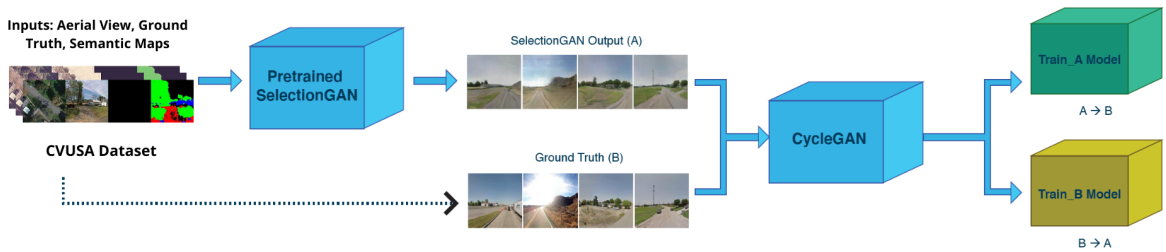

Figure 2: EnsembledGAN Training Phase

CycleGAN is trained with the output image from the SelectionGAN and the ground truth street-view image. Here the output image from the SelectionGAN belongs to domain A and the ground truth image belongs to domain B. While training we also compute mapping losses, for mapping function G from A to B as:

$$L_{GAN}(G, |D_B, A, B) = \mathbb{E}_{b \sim p_{data}(b)}[logD_B(b)] + \mathbb{E}_{a \sim p_{data}(a)}[log(1 - D_B(G(a)))] \quad (1)$$

And F from B to A as:

$$L_{GAN}(F, |D_A, B, A) = \mathbb{E}_{a \sim p_{data}(a)}[logD_A(a)] + \mathbb{E}_{b \sim p_{data}(b)}[log(1 - D_A(F(b)))] \quad (2)$$

where:
$a \, \epsilon$ domain A (output from SelectionGAN) and $b \, \epsilon$ domain B (ground truth street-view image)

These are adversarial loss and is computed as least squares loss. Least squared loss is more stable during training and generates higher quality results [13]. However, to guarantee that each individual image can map to the desired output (ground truth), we learn from CycleGAN [13] that mapping should be cycle-consistent for each image a from domain A, the image translation cycle should be able to bring image a back to the original image. This is forward cycle consistency. Similarly, for each image b from domain B, mapping functions G and F should also satisfy backward cycle consistency. The cycle loss thus is calculated as:

$$L_{cyc}(G, F) = \mathbb{E}_{a \sim p_{data}(a)}[||F(G(a)) - a||] + \mathbb{E}_{b \sim p_{data}(b)}[||G(F(b)) - b||] \quad (3)$$

So, total loss is calculated as:

$$L(G, F, D_A, D_B) = L_{GAN}(G, |D_B, A, B) + L_{GAN}(F, |D_A, B, A) + \lambda L_{cyc}(G, F) \quad (4)$$

We observe the below cycle loss over 200 epochs during our training.

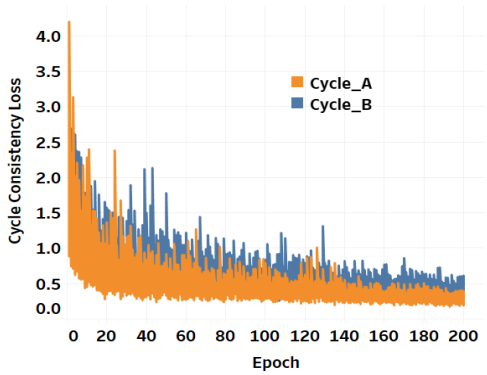

Figure 3: Cycle loss over 200 epochs during our training

The cycle loss is reducing with epochs, which infers that our model is doing a better mapping for an unpaired image to image translation. The training approach for our image-to-image translation task is as follows: The Adam optimizer, a common variant of gradient descent, is used to make training more stable and efficient. The learning rate was set to 0.0002 with moment parameter of 0.5. λ in Equation [4] is set to 10. We perform batch normalization with a batch size of 2. We get two trained models namely latest_net_G_A.pth and latest_net_G_B.pth at the end of this stage.

Testing:

Figure 4: EnsembledGAN Testing Phase

In this phase, we use n output images obtained from SelectionGAN [7] as input to our trained CycleGAN [13] model. We use latest_net_G_A.pth model from the training phase as this model is trained to bring images closer to the ground truth. The corresponding n output images generated as output from this model are further evaluated based on metrics mentioned below. Batch size is set to 2 during this phase.

3.3 Evaluation Metrics

Similar to [7], our evaluation protocol employs Inception Score(IS), Top-k Prediction Accuracy, Kullback-Leibler Divergence score (KL Score), Structural Similarity Index Measure (SSIM) and Peak Signal to Noise Ratio (PSNR).

4 Experiments and Results

4.1 Experimental Setting

We consider a sample from CVUSA [9] dataset with 500 train and 500 test images for the first round of experiments and $2,500$ train and $2,500$ test images for the second round of experiments. Small sample space is taken for training/analysis due to time constraint given it takes weeks of time to train the model with such large datasets.

All the input images are scaled to 256×256 for fair comparison and we train the EnsembledGAN network for around 200 epochs with a batch size of 2 for both rounds of experiments. It took around 32 hours and 78 hours for training 500 and 2500 images respectively. The model was trained using Nvidia GeForce RTX 2080 GPU with 8 GB Memory to accelerate our training and inference.

4.2 Experimental Results

We compare our EnsembledGAN model with the SelectionGAN [7] with all the above mentioned evaluation metrics. For the first round of experiments involving 500 images, our model performs better than SelectionGAN [7] for **inception score** and **top-k prediction accuracy** (refer Table [1] and Table [2]). For the second round of experiment involving $2,500$ images, our model is performing better than SelectionGAN [7] for the **inception score**, **top-k prediction accuracy** and the **KL score** (refer Table [1], Table [2] and Table [3]). We achieve comparable results for SSIM and PSNR for both rounds of experiments which is evident from the Table [3].

Method	All classes	Top-1 Classes	Top-5 Classes
SelectionGAN (for N* = 500)	3.7589	2.7395	3.8158
EnsembledGAN (for N* = 500)	**4.2267**	**2.9125**	**4.3052**
Real Data (for N* = 500)	5.164	3.307	5.2779
SelectionGAN (for N* = 2500)	3.5378	2.6544	3.6123
EnsembledGAN (for N* = 2500)	**3.9263**	**2.8786**	**4.0393**
Real Data (for N* = 2500)	5.0803	3.3585	5.1833

Table 1: Inception Score of different methods. For this metric, higher is better. These results are calculated as per [6].*here N = no. of images considered for the analysis

Method	Top-1 Accuracy		Top-5 Accuracy	
Number of Images	500	135 (with prob >0.5)	500	135 (with prob >0.5)
SelectionGAN	41.6000	69.6296	**77.0000**	91.8519
EnsembledGAN	**44.0000**	**71.1111**	74.6000	**92.5926**
Number of Images	2500	682 (with prob >0.5)	2500	682 (with prob >0.5)
SelectionGAN	40.2400	63.0499	74.1200	90.4692
EnsembledGAN	**42.9200**	**68.9150**	**75.5600**	**91.2023**

Table 2: Top-k accuracies of different methods. For this metric, higher is better.These results are calculated as per [6]

Method	SSIM	PSNR	KL Score
SelectionGAN (for N* = 500)	**0.4669**	**17.4515**	$\mathbf{3.07 \pm 0.95}$
EnsembledGAN (for N* = 500)	0.4462	17.0272	3.17 ± 1.06
SelectionGAN (for N* = 2500)	**0.4631**	**17.3583**	3.17 ± 1.07
EnsembledGAN (for N* = 2500)	0.4470	16.9583	$\mathbf{3.10 \pm 1.10}$

Table 3: SSIM, PSNR and KL score of different methods. For these metrics except KL score, higher is better. These results are calculated as per [6]. *here N = no. of images considered for the analysis.

4.3 Qualitative Evaluation

We can see that the EnsembledGAN modelled approach generates more clearer and sharper images with less smudging, along with better contrast on the objects such as roads, grass, trees, clouds, etc. The results generated render clear distinction of these objects and also is leading closer to the ground truth for the provided aerial image input generating the street view as can be seen from the figure below (Figure: 5).

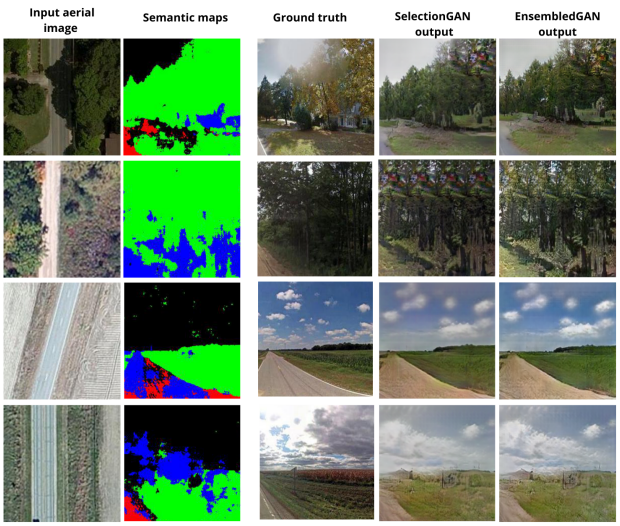

Figure 5: Qualitative Results of SelectionGAN and our EnsembledGAN on CVUSA dataset

5 Conclusion

We make use of many state-of-the-art existing deep learning techniques to ensure that our final outcome improves some of the evaluation metrics used to quantitatively analyse a Generative Adversarial Network (GAN).

We have proposed a novel approach EnsembledGAN of combining the two existing GAN architectures to ultimately generate images which are more closer to the ground truth. We leverage the abilities of SelectionGAN [7] to perform cross-view image translation and the ability of CycleGAN [13] to generate enhanced images when trained with output of SelectionGAN and ground truth.

The final obtained results on a small subset of the CVUSA [9] dataset is a proof of this experimental approach of EnsembledGAN's capability to produce better quantitative and qualitative results of the generated images.

6 Contributions

Everyone in the project has contributed equally for the success of the project and everyone has been involved equally during the brainstorming for the project proposal stage to the stage involving generation of the final outcome achieved in the project .

Here are some of the highlights involved in each one's contribution:

Abhishek Sundar Raman :

Instrumental in proposing the Novel approach to use the Ensembling technique of two GANS, Training of the EnsembledGAN Model and Generation of the evaluation results improved in the proposed model.

Gayatri Ganapathy :

Did the initial Exploratory analysis on the Dayton dataset for the proposed technique. Involved in the dataset preparation and analysis of the results.

Ria Gupta :

Instrumental towards generation of the evaluation metrics and did extensive networking with Hao Tang(https://www.linkedin.com/in/hao-tang-887475138/) to gain information about the subject, built graphs for the report.

Varnnitha Venugopal :

Contributed towards designing of the architecture diagram and involved during the testing phase of the EnsembledGAN approach. Instrumental in designing the poster for the presentation.

Project report involved contribution from every team member along with peer reviews done by each member.

References

[1] Nelson Chong, Lai-Kuan Wong, and John See. Ganmera: Reproducing aesthetically pleasing photographs using deep adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[2] Alexey Dosovitskiy, Jost Tobias Springenberg, Maxim Tatarchenko, and Thomas Brox. Learning to generate chairs, tables and cars with convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):692–705, 2016.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[4] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3277–3285, 2017.

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2016.

[6] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3501–3510, 2018.

[7] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2417–2426, 2019.

[8] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network: Supplementary material.

[9] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.

[10] Christopher B Choy Danfei Xu, JunYoung Gwak, and Kevin Chen Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. *arXiv preprint arXiv:1604.00449*, 1, 2016.

[11] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. *CoRR*, abs/1612.02709, 2016.

[12] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875, 2017.

[13] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.