Aerial to Street View Image Translation using Cascaded Conditional GANs

Kshitij Singh, Alexia Briassouli[a] and Mirela Popa
Department of Data Science and Knowledge Engineering, Maastricht University, The Netherlands

Keywords: Cross View Image Translation, Conditional GANs, Semantic Segmentation, U-net.

Abstract: Cross view image translation is a challenging case of viewpoint translation which involves generating the street view image when the aerial view image is given and vice versa. As there is no overlap in the two views, a single stage generation network fails to capture the complex scene structure of objects in these two views. Our work aims to tackle the task of generating street level view images from aerial view images on the benchmarking CVUSA dataset by a cascade pipeline consisting of three smaller stages: street view image generation, semantic segmentation map generation, and image refinement, trained together in a constrained manner in a Conditional GAN (CGAN) framework. Our contributions are twofold: (1) The first stage of our pipeline examines the use of alternate architectures ResNet, ResUnet++ in a framework similar to the current State-of-the-Art (SoA), leading to useful insights and comparable or improved results in some cases. (2) In the 3rd stage, ResUNet++ is used for the first time for image refinement. U-net performs the best for street view image generation and semantic map generation as a result of the skip connections between encoders and decoders, while ResU-Net++ performs the best for image refinement because of the presence of the attention module in the decoders. Qualitative and quantitative comparisons with existing methods show that our model outperforms all others on the KL Divergence metric and ranks amongst the best for other metrics.

1 INTRODUCTION

The task of generating outdoor scenes from a variety of viewpoints is a challenging one that is gaining a lot of attention recently with applications in domains like autonomous driving, virtual reality, geo-tagging etc. Generation of a novel viewpoint involves transforming objects in a scene from a given view to the desired view in a natural setting, while maintaining the photo-realism of the transformation.

Cross view image translation is a special case of viewpoint translation, where the desired view has no overlap with the given view (aerial to street or vice versa). This is much more challenging due to occlusion and the large degree of deformation while transforming from one view to another. Moreover, when transforming from aerial view to street view, there is uncertainty in the orientation in which the street view will be synthesized. Existing methods (Zhai et al., 2017) (Regmi and Borji, 2018) (Tang et al., 2019) show that a single stage image translation model fails to transfer fine details of the objects. Thus, a multi step process is needed, with image refinement after street view image generation (Tang et al., 2019). A semantic segmentation map generator, for comparison with the ground truth semantic map, is added to the multi-step process, to guide image generation. The final pipeline consists of 3 steps, where a street view image is first generated, and is then provided to image refinement and semantic map generation networks.

Our work builds upon (Tang et al., 2019), investigating which architectures are best suited for each of the 3 steps/subtasks (street view image generation, semantic map generation, and image refinement). Our contributions are: (1) Stage 1 of our pipeline examines alternate SoA architectures ResNet, ResUnet++ in a framework similar to (Tang et al., 2019), leading to useful insights and comparable or improved results. (2) In stage 3, ResUNet++ is used for the first time for image refinement.

Conditional GANs (CGAN), proven to be very effective in image translation (Isola et al., 2017), are used as the framework for each step. In addition to U-Net, which is the standard for image translation, CGAN, ResNet and ResU-Net++ (Jha et al., 2019) are

[a] https://orcid.org/0000-0002-0545-3215

Singh, K., Briassouli, A. and Popa, M.
Aerial to Street View Image Translation using Cascaded Conditional GANs.
DOI: 10.5220/0010814000003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 372-379
ISBN: 978-989-758-555-5; ISSN: 2184-4321

studied for their suitability in each subtask: they are trained together in a constrained manner in a cascade network framework, to study the effect of each one of them in the final street view image generation process. The investigation of these generators provides important insights into the suitability of network features like skip connections and attention for each subtask.

This paper is structured as follows. Sec. 2 presents the background of this work, Sec. 3 provides details on the architectures and methods used. Experiments and results are presented in Sec. 4, with qualitative and quantitative comparisons of the proposed pipeline against existing methods. Finally, conclusions and plans for future work are presented in Sec. 5.

2 BACKGROUND, RELATED WORK

Conditional GANs (CGANs) (Mirza and Osindero, 2014) generate domain-specific images in a controlled setting, so they have been used for image to image translation (Isola et al., 2017), (Odena et al., 2017), (Choi et al., 2018). In order to make the task feasible, semantic segmentation maps of the opposite view are provided to the CGAN generator network, allowing the network to focus on reproducing color, texture, and structure. (Regmi and Borji, 2018) proposed X-Fork and X-Seq for cross view image translation, using CGANs. The SoA in this task is SelectionGAN (Tang et al., 2019) which used CGAN, similarly to (Regmi and Borji, 2018), but with multi channel attention. This multi channel attention is handcrafted to a certain extent, which potentially constrains its generalizability. Recently, (Toker et al., 2021) introduced a novel multi-task architecture with joint image synthesis and retrieval, achieving SoA cross-view image based geo-localization.

We examine the following research aspects: (1) how newer architectures perform in cascaded CGANs, (2) how our system compares to the SoA on quantitative metrics used in (Tang et al., 2019), (3) the effect of adding tasks (like segmentation map generation, image refinement), while training concurrently with cross view image synthesis, (4) the role of attention in image generation. We focus on street level view generation when the corresponding top level view is given as an input, which is more difficult than the inverse (generating aerial level images from street view), as it requires generating a view which has more details using an input view which has fewer details.

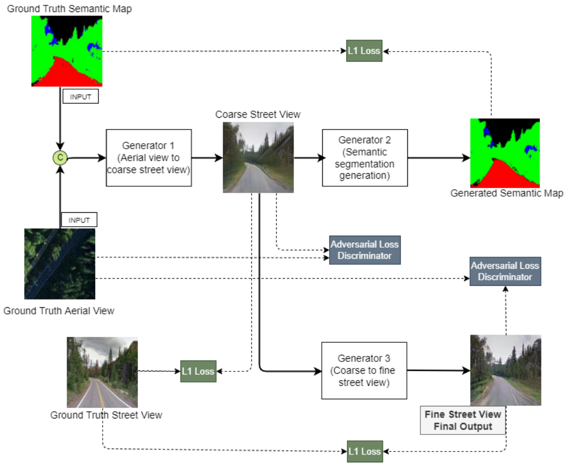
Ground Truth Semantic Map
INPUT
L1 Loss
Coarse Street View
C
Generator 1 (Aerial view to coarse street view)
Generator 2 (Semantic segmentation generation)
Generated Semantic Map
INPUT
Adversarial Loss Discriminator
Adversarial Loss Discriminator
Ground Truth Aerial View
L1 Loss
Generator 3 (Coarse to fine street view)
Ground Truth Street View
Fine Street View Final Output
L1 Loss

Figure 1: Diagram of the pipeline. Dashed lines indicate the different adversarial and L1 losses used to train the network.

3 METHODOLOGY

Motivation for the Chosen Pipeline: Objects in aerial and street views have different dimensions and orientation, leading to ambiguity when transferring from one view to another. A single stage image to image translation model cannot capture details, especially when the degree of overlap between the views is small (as in our case). This issue is addressed by using a cascade CGANs pipeline (Tang et al., 2019), with 3 generators trained together in an adversarial manner guided by pixel losses (Figure 1). The three generators have different tasks, described below.

1. Generator 1 (coarse street view): This generator deals with **coarse street view image generation**. The aerial view and semantic segmentation map of the street view are input and the network generates a coarse reconstruction of the street view.
2. Generator 2 (street view semantic map): This generator deals with **street view semantic map generation**. It takes the coarse street view (i.e output of Generator 1) as input and generates the semantic segmentation map of the street view.
3. Generator 3 (coarse to fine street view): This generator deals with the task of **image refinement**. Similar to Generator 2, it takes the coarse street view (i.e output of Generator 1) as input and generates a fine version of the street view. This refined street level view is the final output.

Generator Architecture Choices:
U-Net: CGANs for image to image translation (Isola et al., 2017) used U-Nets, as they share low level features between the input and output views by skip connections between layer i and layer $n-i$ of the stan-

dard encoder-decoder module. This is appropriate for image to image translation, where input and output images share low level features such as color, shape, dimensions, orientation, etc. We use the implementation of U-Net by (Isola et al., 2017), which can also be used with fewer filters for coarse image generation, greatly reducing the number of parameters of the network and saving on training time.

ResNet: Residual networks for image classification (He et al., 2016) replaced the layers in the middle of the network by blocks with skip connections, connecting alternate blocks. In our two image translation tasks (street view image generation and coarse to fine image refinement) the input and output images have high level of structural similarities, so we use a ResNet style architecture. There are 3 blocks, each downsampling by a factor of 2, so a 256x256 image is downsampled to 64x64 with 256 channels. It is then passed through 6 ResNet blocks, containing a convolutional block that does not change the image dimensions. Output from block *i* is added to output of block $i+1$, which is then given as input to block $i+2$. After the ResNet block, the image is passed to 3 upsampling blocks to bring the output back to a 256x256 image.

ResU-Net++: ResU-Net++ (Jha et al., 2019) combines key features of U-Net and ResNet. It has skip connections between block i and block $n-i$ similar to U-Net, and skip connections between consecutive blocks similar to ResNet. Attention has recently been proven useful for image translation (Zhang et al., 2019), (Kim et al., 2019), and is explored here in the context of cross view image translation and coarse to fine image refinement. ResU-Net++ has never been used previously for image translation related tasks. We examine it for our three subtasks, since U-Net and ResNet have already proven to be effective for image related tasks using CGAN, and its inclusion of attention is expected to show promising results.

PatchGAN Discriminator: (Pathak et al., 2016) introduced pixel loss (L1 or L2) in addition to adversarial GAN loss. We use L1 loss, as it leads to less blurring than L1/L2 (Zhao et al., 2016), (Larsen et al., 2016). (Isola et al., 2017) argued that L1 loss captures low, but fails to capture high frequencies, so we use a 70x70 PatchGAN discriminator, which has gives the sharpest images both in spatial and spectral domains.

Loss Function Formulation: In CGANs, both the discriminator and generator receive the conditioning variable (aerial view image in this case). For the discriminator, we first construct fake sample pairs using the aerial view and the street view from the generator, pass this through the discriminator, and calculate the loss based on the prediction by the discriminator (2nd component in Eqs 1, 2). Next,we do the same with real sample pairs (aerial view and ground truth street view) (1st component in Eqs 1, 2). The discriminator's combined adversarial loss L_{adv1} for identifying generated coarse street view ($I_{g'}$) from ground truth street view(I_g) is the sum of these two:

$$\begin{aligned} L_{adv1} = \min_G \max_D L_{cGAN}(G, D_1) = \\ \mathbb{E}_{I_a,I_g}[\log D(I_a, I_g)] + \mathbb{E}_{I_a,I_{g'}}[\log(1 - D(I_a, I_{g'}))] \end{aligned} \tag{1}$$

The discriminator adversarial loss L_{adv2} for the refined generated street view image ($I_{g''}$) is:

$$\begin{aligned} L_{adv2} = \min_G \max_D L_{cGAN}(G, D_2) = \\ \mathbb{E}_{I_a,I_g}[\log D(I_a, I_g)] + \mathbb{E}_{I_a,I_{g''}}[\log(1 - D(I_a, I_{g''}))] \end{aligned} \tag{2}$$

L1 Loss: Each of the 3 generators produces an image which is compared to the ground truth with L1 loss.

Between the coarse ($I_{g'}$) and ground truth street view image (I_g):

$$L1_{gen1} = \min_G L1(G) = \mathbb{E}_{I_g,I_{g'}}[||I_g - I_{g'}||_1] \tag{3}$$

Between the generated semantic map ($I_{s'}$) and ground truth semantic map of street view (I_s):

$$L1_{gen2} = \min_G L1(G) = \mathbb{E}_{I_s,I_{s'}}[||I_s - I_{s'}||_1] \tag{4}$$

Between the refined ($I_{g''}$) and ground truth street view image(I_g):

$$L1_{gen3} = \min_G L1(G) = \mathbb{E}_{I_g,I_{g''}}[||I_g - I_{g''}||_1] \tag{5}$$

The total L1 loss is then defined as:

$$L_{L1} = \min_G (L1_{gen1} + L1_{gen2} + L1_{gen3}) \tag{6}$$

Overall Loss: The Overall Loss is the weighted sum of the adversarial and L1 loss. The weighting factors λ_1 and λ_2 are hyper-parameters which can be tuned.

$$\min_G \max_D L = L_{adv1} + \lambda_1 L_{adv2} + \lambda_2 L_{L1} \tag{7}$$

Selecting λ_1 and λ_2: In Eq. 7, regularization term λ_1 denotes the importance of reducing the adversarial loss in coarse to fine image refinement, in comparison to coarse street view generation. Higher values indicate that discriminating the refined street view from the ground truth is more difficult than discriminating the coarse view, so $\lambda_1 > 1$ seems logical. Experiments for $\lambda_1 = 0.5, 1, 5, 10$ showed λ_1 = 5 to be optimum.

λ_2 is the regularization term for reducing L1 loss compared to the other terms in the overall loss. (Isola et al., 2017) observed that, for image to image translation, the discriminator trains faster than the generator (i.e it can identify fake images before the generator can keep up with generating realistic images and thus fails to train efficiently). Also, color is a very important in image translatio,n especially during viewpoint translation where objects in the scene (mountains, sky etc) have a narrow range of realistic spectral values for humans. This motivates using to a high value of λ_2 (λ_2 = 10). For cross view image translation, (Regmi and Borji, 2018) found it should be even higher and used $\lambda_2 = 100$ which is also used here. Tests using higher or lower values of λ_2 validated this choice.

Network Training: In order to optimize the network, we perform gradient descent on the discriminator and the generator alternatively. We first train the discriminator and update weights via backpropagation, keeping the 3 generators fixed. Then we fix the discriminator weights, train the 3 generators and update their weights via backpropagation. The process is end-to-end and continues for 30 epochs. Similarly to (Regmi and Borji, 2018) and (Tang et al., 2019), an Adam optimizer is used for both discriminator and generator training with initial learning rate = 0.0002 and moment parameters β_1 = 0.5 and β_2 = 0.999. For each epoch, the intermediate model is saved so results can be compared and progress can be monitored.

4 EXPERIMENTS, RESULTS

Dataset: The dataset used is Crossview USA (CVUSA) (Workman et al., 2015), with millions of pairs of aerial images obtained from Bing Maps and ground-level images from Google Street view. For our task of cross view image translation, we use a subset of CVUSA created by (Zhai et al., 2017), with 35,552 training and 8,884 testing tuples. This exact split and distribution of images is used to compare our results with the SoA. The semantic segmentation map of the ground-level view is included with the aerial and ground-level images, to guide the image generation process. As shown in Figure 2, a sample image consists of a tuple (aerial view, street view, semantic segmentation of street view). The resolution is 768 X 256 (i.e each image in the tuple has a resolution of 256 X 256). In order to train the model, they have to be cropped and converted into separate paired images.

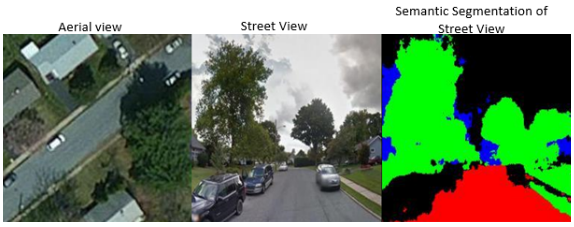

Figure 2: The entire image is divided into 3 windows of equal width and height (256x256) showing different views of the same place.

In this dataset, the street level views are consistent in attributes like height from the ground and direction, so the network can learn to generate street level images for this consistent set of attribute values. Therefore data augmentation techniques like flipping, rotation etc are not applied, as they would change the direction attribute, breaking the consistency. However, all existing methods for this task using this dataset perform random cropping from 256x256 to 224x224 and then resizing back to 256x256. Since we want to compare the performance of our proposed model with the SoA, we applied this cropping and resizing.

Qualitative Evaluation of Generated Images: In order to visually assess the performance of the generator in GANs for photo-realistic images of a viewpoint we visually assess structure, color, and textures.

Quantitative Evaluation of Generated Images: Qualitative assessment is subjective, time consuming and labor intensive, so the quantitative metrics are also measured. They are also used in the SoA, allowing us to quantitatively compare our results to it.

Pixel Level Similarity Measures: Peak Signal to Noise Ratio (PSNR) between a reconstructed and input view: high PSNR indicates better reconstruction. Sharpness Difference (SD), for loss in sharpness during viewpoint reconstruction. Structural-Similarity (SSIM), between the ground truth and generated image based on structure, luminance, and contrast. A higher value of SSIM indicates higher similarity.

High Level Feature Similarity Measures: Inception Score (IS), to evaluate the diversity of a generated image within a class while simultaneously being representative of that class. We use AlexNet(Krizhevsky et al., 2012) trained on Places (Zhou et al., 2017) with 365 categories, as in (Regmi and Borji, 2018). Then we calculate the probability of the generated street view image belonging to each class, for Top-1 and Top-5 Inception Scores. Kullback-Leibler divergence (KL) is also used. The probability distribution is calculated by passing the image through the pre-trained model used for the IS calculation.

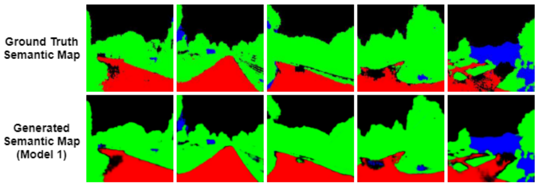

Figure 3: Generated semantic segmentation vs ground truth.

Architecture Selection: U-Net performs extremely well in the semantic segmentation of Generator 2 (Figure 3). Generator 2 generates a semantic segmentation map of the street view, trained using a L1 loss and added to the overall loss of the generator to help the other two generators train faster.

Semantic segmentation is only used for loss guidance, while the other two generators actually produce street level views, so we fix the architecture of Generator 2 as U-Net, reducing the number of configurations to be tried from 27 to 9. The 9 configurations, with U-Net as Generator 2, contain the following combinations for Generators 1 and 3 (G1, G3): (a) U-Net, U-Net, (b) U-Net, ResNet, (c) U-Net, ResU-Net++, (d) ResNet, U-Net, (e) ResNet, ResNet, (f) ResNet, ResU-Net++, (g) ResU-Net++, U-Net, (h) ResU-Net++, ResNet, (i) ResU-Net++, ResU-Net++.

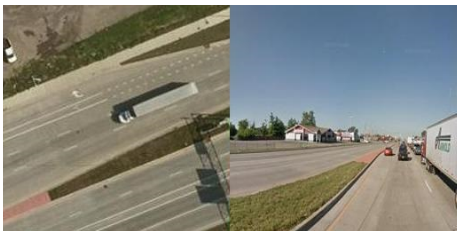

Figure 4: Ground truth image; Left window is the aerial view and the right window is the street view.

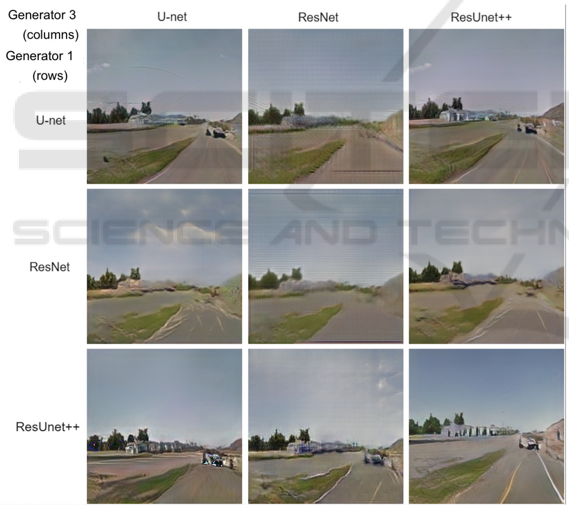

Figure 5: Generated refined street view for models (a)-(i).

Figure 5 shows the generated street view of these configurations for the ground truth image in Figure 4. Some observations on our results are listed below.

1. The color reproduction of all the tried configuration seems to be accurate compared to the ground truth. This shows that the L1 loss works well.
2. ResNet performs poorly as G1, which deals with coarse street views, i.e. structure. ResNet lacks skip connections between encoder and decoder (unlike U-Net, ResU-Net++), which are important for structure in viewpoint translation (Isola et al., 2017). The quality of images generated in G3 is also lower, hence, ResNet is not ideal for viewpoint translation or image refinement.
3. **Best Configuration, Role of Attention:** The best results are in (c) (**Generator 1: U-Net Generator 3: ResU-Net++**) and (i) (**Generator 1: ResU-Net++ Generator 3: ResU-Net++**), so they are evaluated quantitatively. ResU-Net++ has skip connections between encoders and decoders with an attention module in the decoders. Attention improves the ability of the network to convert coarse generated images into finer grained by focusing on relevant parts of the image.
4. The output using configuration (g) has structure and details comparable to the other 2 selected configurations. However, on closer inspection on multiple test images, it seems it introduces random brightly colored artifacts (Figure 6).

Table 1: High level feature qualitative metrics for the 9 configurations (except for KL Div., higher is better).

	Accuracy(%)		Inception Score			
	Top 1	Top 5	All	Top 1	Top 5	KL Div.
(a)	50.45	79.84	3.33	2.38	3.58	4.98 ± 1.25
(b)	47.29	75.92	3.23	2.30	3.47	9.87 ± 1.60
(c)	**58.23**	87.91	**3.76**	**2.67**	**3.89**	**2.84** ± **0.93**
(d)	31.82	57.65	3.16	2.28	3.36	14.09 ± 1.65
(e)	25.65	54.32	3.09	2.12	3.23	17.61 ± 1.63
(f)	33.21	62.39	3.13	2.22	3.26	13.29 ± 1.66
(g)	57.49	**88.51**	3.32	2.39	3.449	6.27 ± 1.52
(h)	49.45	77.49	3.29	2.29	3.32	8.27 ± 1.56
(i)	55.06	85.23	3.58	2.54	3.66	3.55 ± 1.05

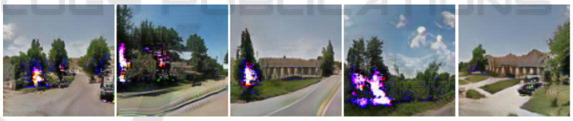

Figure 6: Artifacts with G1: ResU-Net++, G3: U-Net.

Quantitative Assessment of Generated Images: Tables 1 and 2 show quantitative results for all 9 configurations. The two configurations with the best qualitative results also show the best quantitative results. Configuration (g) had some comparable/better metrics (Top-5 accuracy and SD), but introduces artifacts (Figure 6), so we do not examine it further.

Table 2: Pixel level qualitative metrics of the configurations.

	PSNR	SSIM	SD
(a)	21.549	0.479	18.707
(b)	21.267	0.469	18.732
(c)	**22.355**	**0.509**	19.459
(d)	19.728	0.456	17.170
(e)	18.521	0.428	16.091
(f)	19.853	0.449	17.670
(g)	20.473	0.483	**19.893**
(h)	21.161	0.474	18.876
(i)	21.769	0.498	19.482

The Two Best Configurations: As seen from the qualitative and quantitative assessment, configuration (c), which we will call Model 1, (U-Net, ResU-Net++) and configuration (i), which we will call Model 2 (ResU-Net++, ResU-Net++) give the best street view transformation image. Figure 7 shows the generated refined street views for both models are similar to the ground truth for different test samples, encompassing various viewpoints. The semantic map provided is not detailed, so some inaccuracies in it carry over to the generated street view images.

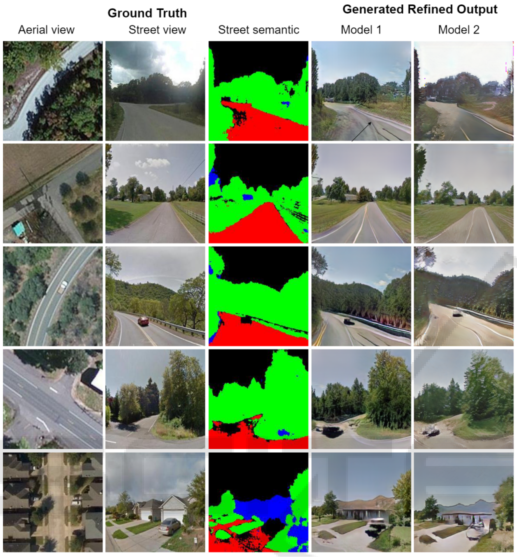

Figure 7: Model 1, 2 refined generated street view.

Coarse vs Refined: We first generate a coarse intermediate image with Generator 1 and refine it with Generator 3. Figure 8 shows the coarse output for Model 1 lacks detail compared to Model 2. This can be attributed to the fact that for Generator 1, Model 1 uses U-Net and Model 2 uses ResU-Net++, the latter being more complex, despite having fewer parameters. For Model 2, the coarse image displays the same artifacts that appeared in the refined image of configuration (g), but are mostly resolved in the refined image when using Model 2, while adding a few details. Model 1 on the other hand recovers more details in the image refinement process resulting in a starker difference between the coarse and refined images.

In conclusion, the refined output of Model 1 is slightly better than Model 2, so **Model 1 (G1: U-Net, G2: U-Net, G3: ResU-Net++) is our final chosen model**. However, it should be noted that the coarse intermediate output of Model 2 (G1: ResU-Net++, G2: U-Net, G3: ResU-Net++) shows promising results, so there is a room for further improvement if a suitable image refinement process can be designed for it.

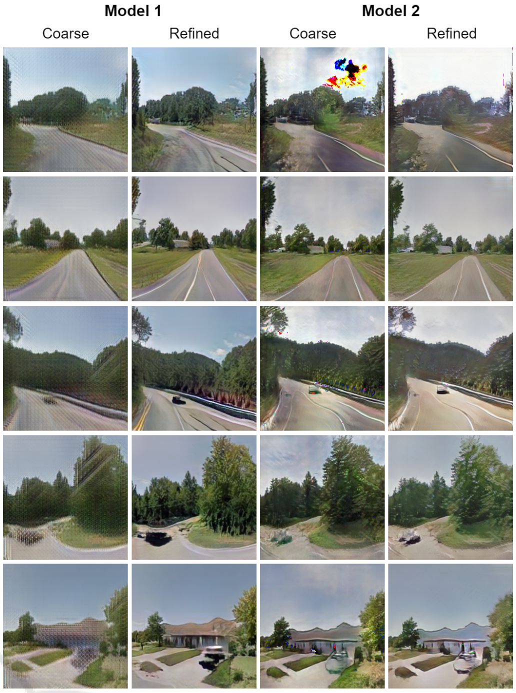

Figure 8: Model 1, 2 coarse vs refined street view.

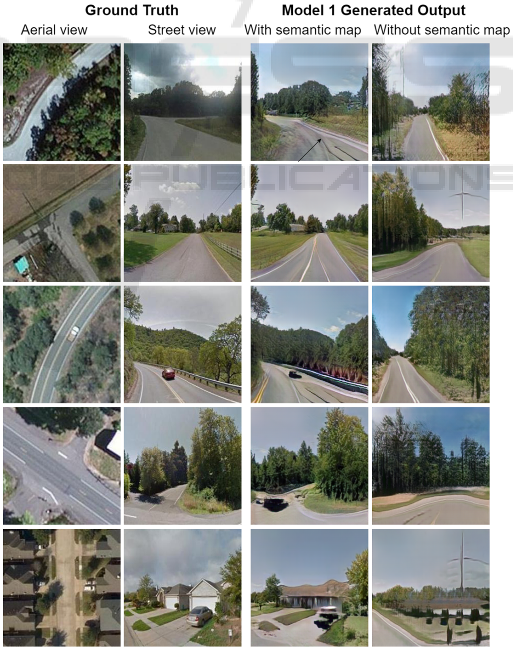

Figure 9: Comparison of refined street view generated for Model 1 with and without semantic segmentation map.

Removing Semantic Guidance: We investigate the effects of removing semantic segmentation from the input, to obtain insights into the limitations of this task. In Figure 9, we observe that we usually obtain photo-realistic images (or parts that are photo-realistic) of the street view even when semantic maps are not used. This shows the model learns to generate the structure of objects in the street, although there remains a significant difference when compared to the ground truth, as there is no way of knowing the exact dimensions of the objects (e.g. the height of trees or the design of the windows/doors of houses). However, there are still some improvements that can be made with regards to the general layout of the objects.

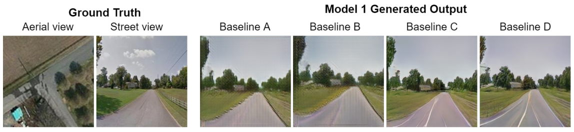

Figure 10: Ablation study: Comparison of 4 baselines.

Table 3: High level feature qualitative metrics for ablation study (except for KL Div., higher is better).

	Accuracy(%)		Inception Score			
	Top 1	Top 5	All	Top 1	Top 5	KL Div.
A	44.18	81.50	3.12	2.28	3.25	9.47 ± 1.83
B	44.91	82.98	3.17	2.33	3.29	9.11 ± 1.85
C	57.44	87.17	3.70	2.64	3.79	3.59 ± 1.13
D	**58.23**	**87.91**	**3.76**	**2.67**	**3.89**	**2.84 ± 0.93**

Ablation Study: We perform an ablation study to see how each of the networks in the pipeline affects the quality of images generated, and their qualitative and quantitative relative importance. We compare 4 baselines: Baseline A uses G1. Baseline B uses G1 and G2. Baseline C uses G1 and G3. Baseline D uses all 3 generators. From the results of the ablation study (Figure 10, Tables 3, 4), it can be said that Baselines A and B are similar to each other, B being slightly higher. When compared to Baseline C, it can be said that even though the addition of semantic segmentation generator in the pipeline (Baseline B) improves performance, the impact is not as significant as Baseline C, where the coarse to fine image refinement is added to coarse street view generation. Baseline C and Baseline D metrics are quite similar and much better than Baseline A and B. This again emphasizes the importance of image refinement in this pipeline.

Comparison with SoA: In this section, we compare the results from Model 1 and Model 2 with existing methods: SelectionGAN (Tang et al., 2019), X-Fork and X-Seq (Regmi and Borji, 2018), Zhai et al(Zhai et al., 2017), Pix2Pix (Isola et al., 2017).

Table 4: Pixel level qualitative metrics for ablation study.

Baseline (Generators used)	PSNR	SSIM	SD
A (1)	21.098	0.456	18.809
B (1 and 2)	21.032	0.461	19.016
C (1 and 3)	22.090	0.498	19.312
D (All 3)	**22.355**	**0.509**	**19.459**

Table 5: High level feature metrics for existing methods (except for KL Div., higher is better).

	Accuracy(%)		Inception Score			
	Top 1	Top 5	All	Top 1	Top 5	KL Div.
Pix2Pix	41.87	72.87	3.26	2.42	3.51	9.47 ± 1.69
Zhai et al.	14.03	52.29	1.84	1.52	1.87	27.43 ± 1.63
X-Fork	49.65	81.16	3.38	2.54	3.57	7.18 ± 1.56
X-Seq	54.61	83.46	**3.82**	2.67	**4.01**	5.19 ± 1.31
SelGAN	**65.51**	**89.66**	3.81	**2.72**	3.92	2.96 ± 0.97
Our Model 1	58.23	87.91	3.76	2.67	3.89	**2.84** ± **0.93**

Table 6: Pixel level metrics for existing methods.

Method	PSNR	SSIM	SD
Pix2Pix	21.57	0.46	18.90
Zhai et al.	17.49	0.42	16.62
X-Fork	21.65	0.48	18.99
X-Seq	21.67	0.47	18.99
SelectionGAN	**23.15**	**0.53**	**19.61**
Proposed Model 1	22.36	0.51	19.46

Quantitative Comparison: When compared to existing methods on quantitative metrics, in Tables 5 and 6, it can be seen that our Model 1 outperforms all models except SelectionGAN on all metrics (the only exception is Inception Score for which it is worse than X-Seq). When compared to SelectionGAN, our proposed Model 1 has comparable performance on all metrics, and beats SelectionGAN in KL Divergence.

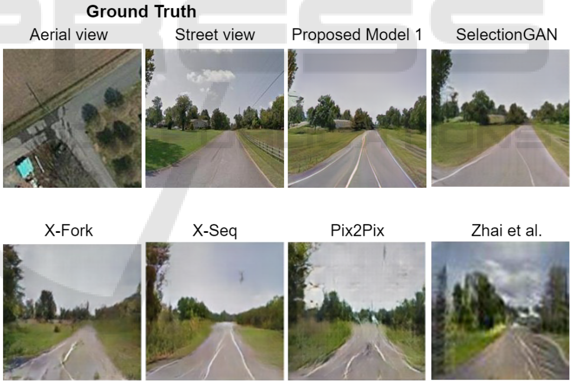

Figure 11: Comparison of our Model 1 with SoA methods.

Qualitative Comparison: In Figure 11 we qualitatively compare the results of Model 1 with existing methods for a test image. The output images of the other models are taken from the (Tang et al., 2019) paper, so it is not fair to compare image resolution, but only structure. We see that our model performs comparably well to the SoA SelectionGAN, maintaining the structure of objects. The other models do not maintain structure that well, and produce significant distortion, or miss some objects in the scene completely (for example the house in the background).

5 CONCLUSIONS, FUTURE WORK

In this paper we examine cross view image translation, generating a street view from the corresponding aerial view using a cascade pipeline, where coarse street view image generation, semantic segmentation, and image refinement, are combined and trained together . We tested SoA generator models U-Net, ResNet, and ResU-Net++ and found best results were obtained for the configuration (Generator 1: U-Net, Generator 2: ResNet, Generator 3: ResU-Net++). This demonstrates the importance of sjkip connections for street view generation and of attention for image refinement. The role of each of the 3 subtasks in the pipeline was studied and it was concluded that each subtask improved overall performance qualitatively and quantitatively. Future work includes investigating appropriate networks for further refinement of the output images to address artifacts related to perspective projection, and how to incorporate varying sources of input data (such as aerial input data from drones at varying heights, or video input).

REFERENCES

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Jha, D., Smedsrud, P. H., Riegler, M. A., Johansen, D., De Lange, T., Halvorsen, P., and Johansen, H. D. (2019). Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255. IEEE.

Kim, J., Kim, M., Kang, H., and Lee, K. (2019). U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Odena, A., Olah, C., and Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544.

Regmi, K. and Borji, A. (2018). Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3501–3510.

Tang, H., Xu, D., Sebe, N., Wang, Y., Corso, J. J., and Yan, Y. (2019). Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2417–2426.

Toker, A., Zhou, Q., Maximov, M., and Leal-Taixe, L. (2021). Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6488–6497.

Workman, S., Souvenir, R., and Jacobs, N. (2015). Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969.

Zhai, M., Bessinger, Z., Workman, S., and Jacobs, N. (2017). Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR.

Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.