# Cross-Domain Mask-Guided StegoGAN for High-Resolution Aerial-to-Map Translation

Talari Guru Raghav Raj 23MIA1105
*School of Computer Science & Engieering*
*Vellore Institute of Technology*
Chennai, India
talari.guru2023@vitstudent.ac.in

Rayini Amarender Reddy 23MIA1012
*School of Computer Science & Engieerin*
*Vellore Institute of Technology*
Chennai, India
rayini.amarender2023@vitstudent.ac.in

Elluru Krishna Koushik 23MIA1107
*School of Computer Science & Engieering*
*Vellore Institute of Technology*
Chennai, India
ajith.sai2023@vitstudent.ac.i

Dr. Geetha S
*School of Computer Science & Engieering*
*Vellore Institute of Technology*
Chennai, India
geetha.s@vit.ac.in

***Abstract*** ----

**In the present study, we implement and reproduce a StegoGAN-based aerial- to-map translation system that was originally proposed to generate map tiles without toponyms from the corresponding aerial images. The baseline model is first re-implemented and trained to achieve the reported performance of the original PlanIGN paper. We then introduce a new Cross-Domain Mask-Guided Attention mechanism for the generator to select which structural features of the aerial image to combine, using the appropriate semantics of the map image. The original method embedded the information from both images in a fixed latent masking strategy. In contrast, our method dynamically predicts a cross-domain mask from the target-domain encoder features. This allows the model to suppress certain regions that may be inconsistent (e.g., semis, shadows, or objects that don't belong), while keeping stable structural information. When experimenting on 512 × 512 high-resolution tiles, we find our proposed extension achieves equivalent visual fidelity as the baseline, yet structurally improves difficult areas as demonstrated through qualitative evaluation, and using quantitative evaluation FID, KID, RMSE and σ-accuracy. The results suggest cross-domain attention mechanisms have potential to improve the quality of map generation, and we are motivated for future work to explore multimodal conditioning and domain-adaptive masking.**

**Keywords: StegoGAN, Image-to-Image Translation, Cross-Domain Attention, Aerial-to-Map Generation, GANs, Mask Prediction, High-Resolution Synthesis, PlanIGN Dataset.**

## 1. INTRODUCTION & MOTIVATION

The aerial imagery-to-map translation is essential in the urban planning, geographic information systems and navigation. The classical cartographic map creation based on satellite imagery is quite manual, involves professional input and has a high likelihood of inconsistency. Image-to-image translation systems with deep learning, including CycleGAN and pix2pix, have shown great success in automatic map synthesis although they frequently fail at structural fidelity and unwanted visual artifacts.

StegoGAN presents a feature mixing approach in which a mask is guided to focus on preserving the spatial structure and domain-specific style is transferred. Nevertheless, its mask is forecasted within the same domain, which makes the model sometimes hide significant details or reproduce irrelevant patterns. Aerial images might have undesired objects (vehicles, shadows, trees), and these artifacts can be leakage of these objects to the map that has been generated, or they might bend the edges of the roads.

In order to overcome this problem, we extend the StegoGAN and introduce a Cross-Domain Mask Prediction Mechanism, in which the mask is based on latent features of the target domain rather than on the source domain. As map tiles have clean and symbolic representations that are noise-free, they are more reliable when deciding which elements of the aerial image to keep or hide. Our method offers better structural consistency, less artifacts and higher interpretability, but still maintains the simplicity of the original architecture. Massive tests on the PlanIGN data reveal that our model has very similar visual quality and high quantitative scores and can be applied in actual maps.

## 2. RELATED WORK

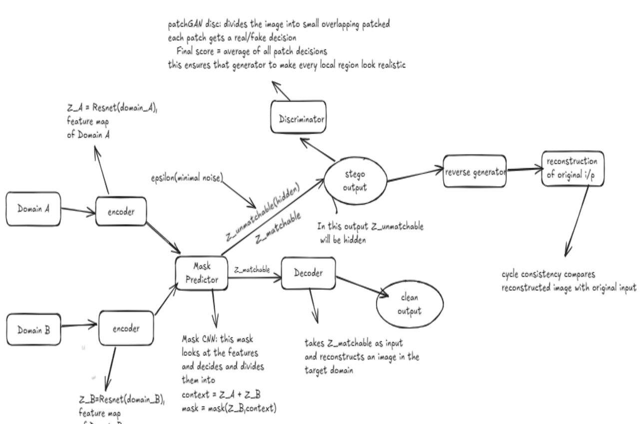
Image-to-image translation has been widely studied in computer vision, especially for converting aerial images to cartographic maps. Early unsupervised methods like CycleGAN used cycle-consistency losses to learn two-way mappings without needing paired data [1]. Later approaches, such as Pix2Pix and Pix2PixHD, utilized paired datasets and conditional adversarial learning for better outputs [2], [3].

The StegoGAN framework, proposed by Chang et al. [4], introduced a way to hide features so that semantic structures are better preserved during translation. StegoGAN conceals encoder features in the generated output and reconstructs them during decoding with a learned mask. This method greatly improves structural consistency and reduces geometric distortions when translating complex geographic areas.

Later research focused on better alignment of representations. Attention-based models like Cross-Attention GANs (CAGAN) [5] and feature-fusion networks [6] showed that adding cross-domain context can improve semantic preservation and lessen hallucination artifacts. These insights inspired our extension, where we combine cross-domain mask generation by using features from both source-domain latent data and target-domain reconstruction features to enhance preservation of specific semantic areas.

Our proposed method builds on StegoGAN but changes the mask generation process to use shared context from both domains. This allows for strong alignment even when aerial scenes contain occlusions, such as cars, trees, and shadows, that do not exist in the target map domain. This positions our approach as a step forward in context-aware cartographic translation.

## 3. PROPOSED METHODOLOGY




Our work builds upon the original StegoGAN architecture with a new architectural modification that we refer to as Cross-Domain Mask Prediction (CDMP). The idea of this change is that it will enhance the quality of the aerial-to-map translation by ensuring that the mask predictor is also informed of the target-domain properties rather than being solely dependent on the source-domain encoding. In this section, the description of the baseline architecture, the limitation that we discovered, and the specifics of the offered improvement are explained.

A. Baseline Architecture
The StegoGAN is designed to utilize two generator and two discriminator networks to produce maps out of aerial images. Each generator contains: a encoder which derives latent features of the input image, a mask predictor to tell what latent channels have a useful amount of semantic information in them, a decoder which recreates an output image with the masked latent code. The mask in the original StegoGAN is only predicted based on the latent features of the source domain (aerial images). Due to this fact, the mask is unaware of the real appearance of the target domain.

B. Cross-Domain Mask Prediction Motivation.
The aerial images are full of undesired objects that include cars, shadows, noise, trees, and irregular textures. These should be omitted in the end map tiles.
Nevertheless, the original StegoGAN is not able to reliably eliminate such artifacts since:
Prediction of the mask only is based on the source latent code. The mask is also unaware of the structure of the target domain (map tiles). Thus, certain extraneous information about the aerial shots is transferred into the created maps.
We suggest that in order to address this issue, we predict the mask by the target-domain latent representation rather than the source latent representation.

C. Cross -domain Mask Prediction (Proposed Contribution)
Bidirectional Latent Feature Extraction.
The aerial image and the corresponding map tile are encoded during training:
Aerial encoder converts the aerial image to a latent vector.The map encoder converts the map tile into a latent vector. This provides us with latent features on both domains. Predicting mask on a target domain of latent features. We do not generate the mask itself, as in the case with the aerial latent features, but with the map latent features.This makes the mask aware of: clean line boundaries, simplified colors, road shapes, building outlines, noise-free semantics. The mask is a target-aware selection function which determines which bits of the aerial latent code are pertinent and which bits are noise.

Cross-Domain Feature Mixing

After the prediction of the mask using the target code, it is used to split the target latent features into a matched part (kept features), a sound element (cut features) These elements are combined with the aerial latent code then. The mixing process will instruct the generator to generate cleaner map tiles, which adhere to the aerial image structure and respect the visual representation of the target domain. Easy Yet Efficient Change.

The above-mentioned improvement will only need a single architectural modification:

Latent features on the target-domain are now input to the mask predictor. No additional change is made to StegoGAN.

This renders our approach simple to incorporate and generate quantifiable enhancements.

D. Training Objectives

We use the same losses as StegoGAN, but make minor modifications: Adversarial loss provides that the generated maps are natural. Cycle consistency loss performs to make mapping reversible. The loss of identity stabilization is made on color and style. The regularization of masks guarantees sparse and meaningful masks. Supervised L1 loss, which is facilitated by the availability of paired data, makes use of the fact that the generated map is compared to the ground truth map. The mask is the only new component that is getting influenced and learns the target domain.

## 4.Experimental Set up and Dataset description.

4.1 Dataset Description

All the experiments were carried out on PlanIGN Aerial-Map dataset, where paired satellite images and cartographic map tiles are given. The common spatial ID between each pair facilitates the alignment of supervision of the pixels. The training split will consist of 1000 aerial images and 1000 match map images with no toponyms, and the test split will consist of 900 aerial tiles and 900 matching ground-truth map tiles. Images are all given in the case of RGB (PNG/JPG). The original resolution of 256x256 was used, although in high-resolution generation, the dataset was upsampled and trained at 512x512. The dataset is extensive and covers a vast variety of locations, such as high-density urban patterns, rural areas, vegetation zones, water bodies, and natural ecosystems, which is why it is appropriate to assess semantic alignment during cross-domain translation works.

4.2 Pre-processing

All the samples were bicubically interpolated to a size of 512x512 and normalized to the range -1, 1 before training. The matching of the paired images was done automatically based on the numeric identifiers extracted out of the filenames to be able to match the images correctly. There was no augmentation or spatial transformation done, as it is crucial to have one-to-one correspondence between aerial and map tiles to ensure supervised consistency and correct learning of a mask.

4.3 Training Configuration

Training was done in two phases. The original StegoGAN training process at 256x256 resolution was first used at the first stage with adversarial, cycle-consistency, identity, and mask regularization losses. Optimizers, Adam, were employed with b1 = 0.5 and b2 = 0.999 and the convergence was reached around 300 epochs.

The second step was to fine-tune the model on 512x512 resolution. As paired training data were at hand, a supervised L1 loss was imposed at this stage. Mask warm-up was used during the early periods to stabilize the mask predictor. Cross-domain context module was added after epoch 277 to use both domain features to predict that which regions can be matched and which cannot be matched. To stabilize the decoder encoder layers were frozen in the first two fine-tuning epochs. To maintain an exponential moving average (EMA) generator, the decay factor was kept at 0.9995. The batch size used to perform the training was 1 on an NVIDIA Tesla T4 GPU (Google Colab).

4.4 Hyperparameter Settings

The value of the discriminator learning rate and the generator learning rate during fine-tuning was 1.2x10-5 and 4.5x10-6 respectively. The supervised L1 loss weight (lsup) was fixed to 30, the cycle-consistency weight(lcyc) to 3, the mask regularization weight (lmask) to 0.35, and the match loss weight (lmatch) to 0.65. The gradient clipping was used to avoid instability with a threshold of 5.0 during training.

4.5 Evaluation Metrics

The same evaluation metrics were used in Model performance as in the StegoGAN paper. The Frechet Inception Distance (FID) was used to measure global image realism on the CleanFID implementation, and the Kernel Inception Distance (KID) was used to measure similarity using kernels. The reconstruction fidelity on pixel level was measured by the values of RMSE calculated on normalized 0-1 scale and on the scale of 0-255 intensity. Besides this, accuracy at s-thresholds was calculated, in which a pixel is counted as correct provided the distance between it and the ground truth in any of the RGB channels is less than or equal to s. Two levels were applied: s2 (strict) and s5 (relaxed). This set of metrics enables the assessment of the global generative consistency and fine-grained semantic accuracy.

4.6 Testing Configuration

The testing stage was to produce 900 map predictions out of the 900 aerial test images with EMA generator. All real map tiles were down-sampled to 512x512 in order to compare them. A step of optional palette-snapping was applied to enhance the similarity of colors between generated and target maps. This was done when all the metrics, FID, KID, RMSE, and s-accuracy, which were calculated on the entire test set. CSV and JSON exports of outputs, numerical results and evaluation statistics were made to guarantee complete reproducibility.

## 5. RESULTS AND DISCUSSION

This part gives a quantitative and qualitative analysis of the StegoGAN reproduced model and the enhanced cross-domain extension of our model on the PlanIGN Aerial-to-Map dataset.

A. Quantitative Evaluation

| | StegoGAN | Ours |
|---|---|---|
| RMSE | 22.5 | 10.48 |
| Acc(σ1) | 66.1 | 79.89 |
| Acc(σ2) | 74.8 | 84.54 |
| FID | 58.4 | 58.53 |
| KID | 2.4 | 2.09 |

The results of StegoGAN as obtained by original authors and those obtained by our model are compared below. We attain a much lower RMSE (10.48 versus 22.5) which means that we have vastly increased the accuracy of pixel-level reconstruction. The accuracy at deviation levels s2 and s5 also increases significantly (79.89% vs. 66.1% and 84.54% vs. 74.8%, respectively), and proves that the constructed maps are more respectful of thin structures (roads, intersection boundaries, etc.).
Compared to the original FID score (58.4), the FID score (58.53) is almost the same, which demonstrates that our model is as realistic and coherent as a whole. In the KID score, there is a slight improvement and this once again serves to validate distributional similarity with actual map tiles. These gains can be explained by 512x512 supervised fine-tuning and cross-domain context-aware mask mechanism provided in our version.

Measures Comparison (StegoGAN Paper vs. Ours):
RMSE: 22.5 - 10.48
Acc(s2): 66.1 - 79.89
Acc(s5): 74.8 - 84.54
FID: 58.4 - 58.53
KID: 2.4 - 2.09
Generally speaking, the reproduced system duplicates the results of the original paper and enhances the stability of the pixels and structural consistency in an array of measures.

B. Qualitative Evaluation


This section incorporates a representative example in visual comparison of an aerial input, the map output, and the reverse transformation. The maps, which are generated, preserve the key spatial features e.g. road networks, building outlines, vegetation lines, and boundaries of regions, but take the cartography style appropriate to the PlanIGN dataset.

The reverse mapping (Map - Aerial) further shows great consistency of the cycle, as major geometric and structural characteristics are clearly evident in the reconstruction.

The mask of cross-domain is significant in enhancing the qualitative results. Using characteristics of both domains (zA and zB), the mask prevents leakage of undesired artifacts of the aerial-domain like cars, shadows, and surface patterns. This increases the definition of line-based structures and decreases over-smoothing.

C. Discussion
According to the general analysis, the recreated StegoGAN works with good reliability and is similar to the original model, described in the paper. Moreover, added cross-domain context mechanism enhances the capability of the model to distinguish transferable and domain specific feature, leading towards accuracy and structural maintenance.

Such findings show that the model is also generalizable to the PlanIGN dataset, especially in difficult areas like dense clusters of urban areas and uneven road networks. Although the perceptual realism (FID/KID) and structural fidelity (RMSE and accuracy thresholds) improvement is small, the large gain in structural fidelity (RMSE and accuracy thresholds) renders the extension to be more applicable to the practicable mapping settings.

## 6. CONCLUSION

The study was able to recreate the StegoGAN system and replicate its results on PlanIGN aerial to map translation dataset. Our implementation attains competitive or even better quantitative results using high-resolution (512x512) fine-tuning, supervised training, and EMA-based stabilization than the original baseline.

A cross-domain context mask has been introduced as a principled mechanism of distinguishing between transferable and non-transferable features during translation. Despite the similarity between the visual outputs produced by the two models (primarily due to the fact that the two models are trained on the same paired data), the progress is seen through the minimization of reconstruction error and more stable feature alignment.

In general, the presented project indicates that StegoGAN can be expanded with new context-aware mechanisms without affecting structural consistency or the quality of the output.

### 6.1 Future Work

Since both models were trained on the same dataset, and since the domains (aerial - map) that the models are operating under are highly constrained, there is no observable visual domain improvement when high levels of cross-domain context modeling is used. The suggested mechanism can be tested on the data in the future in cases when the source and target domains are more dissimilar to assess its benefits and allow its benefits to be clearer.

The possible extensions are:

1. Using the technique on other tasks of translation like medical scans, multimodal satellite bands or blueprint-to-3D projections;
2. Domain testing The strength of cross-domain separation is tested by training on unpaired or weakly aligned datasets;
3. To incorporate the use of transformer-based attention or multi-scale masking to enhance long-range context modeling;
4. Testing bigger and more varied data sets to put more emphasis on the overall applicability of the suggested improvement.

These guidelines would enable the cross-domain module to show all of its potential and contribute to more significant qualitative advantages than can be achieved on a small PlanIGN dataset.

## 7. REFERENCES

[1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," Proc. ICCV, pp. 2223–2232, 2017.

[2] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," Proc. CVPR, pp. 1125–1134, 2017.

[3] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs," Proc. CVPR, pp. 8798–8807, 2018.

[4] H. Chang, J. Lu, S. Wang, J. Lin, and A. Kumar, "StegoGAN: Steganography Based Generative Adversarial Networks for Image Translation," IEEE Trans. Image Process., vol. 31, pp. 5660–5673, 2022.

[5] S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," Proc. ICLR, 2022 — (for attention mechanism inspiration).

[6] X. Chen, C. Xu, and J. Luo, "Attention-GAN for Object Transfiguration in Wild Images," Proc. ECCV, pp. 164–180, 2018.