Generation of Aerial Images from Ground-Level Views for Cross-View Matching

Alessandro De Luca

Academic Year 2024/2025

SAPIENZA
UNIVERSITÀ DI ROMA

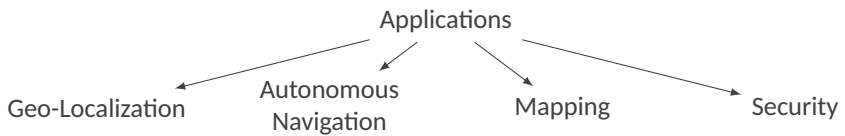
Table of Contents

1] Introduction

Background and Motivations

1] Introduction

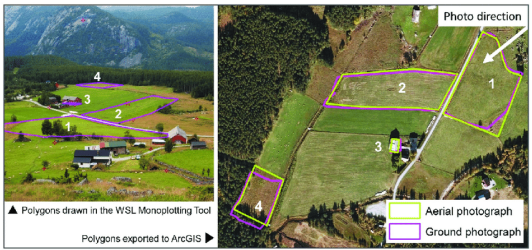
The goal of this research is to match aerial images with ground-level surveys. Identifying an effective feature extractor is crucial for accurately pairing images. This has several applications:



General Problem

1] Introduction

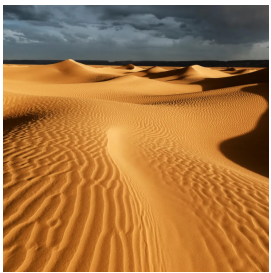
Aerial imagery provides an overhead view of large areas, capturing structural details such as buildings, roads, and green spaces but without perspective. In contrast, street-level imagery offers a detailed, three-dimensional view, revealing building facades, traffic signs, and other elements only visible from the ground. Identifying common features between these two perspectives is a significant challenge.



Dataset

1] Introduction

This research relies on images of urban environments or, more broadly, images that contain distinct features making them recognizable.



No Good

Good

Dataset

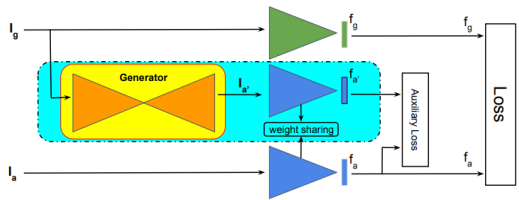
1] Introduction

The training process is conducted using the CVUSA dataset. In particular, the dataset has been modified by removing polar images and retaining only the original ones.

Paper Approach

1] Introduction

Instead of traditional SIFT or SURF methods, which are highly dependent on lighting and scale, an AI-based approach offers greater robustness. My contribution involves reproducing the pipeline of the Feature Learning Extractor:



Paper Approach

1] Introduction

The work is divided into two phases:

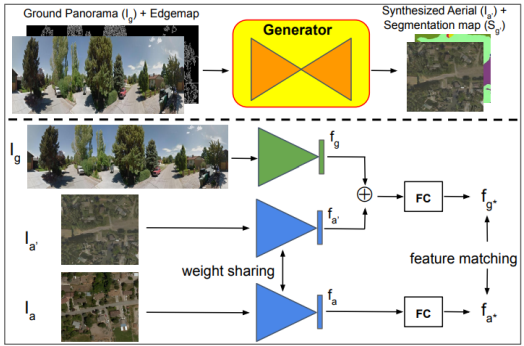


Table of Contents

2] GAN

GAN

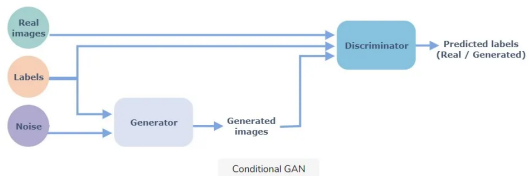
2] GAN

A generative adversarial network (GAN) is a class of machine learning models comprising two neural networks that compete against each other to minimize their own error while maximizing the opponent's error.

Case Study

2] GAN

In this specific case, I used a conditional GAN (cGAN), which generates an image based on a conditional input—the real image we aim to ”reproduce.” Image generation is a challenging problem, as it must capture fine details and establish a coherent structure in the aerial image. Moreover, for the same input image, multiple artificial results can be generated.

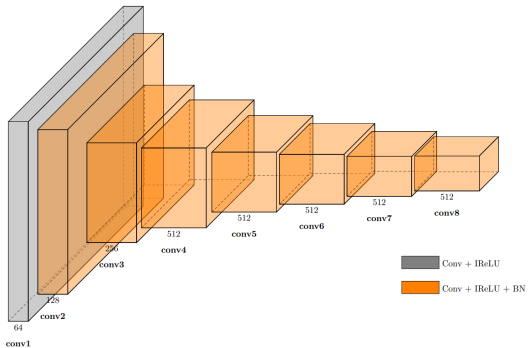


Conditional GAN

Generator - Encoder

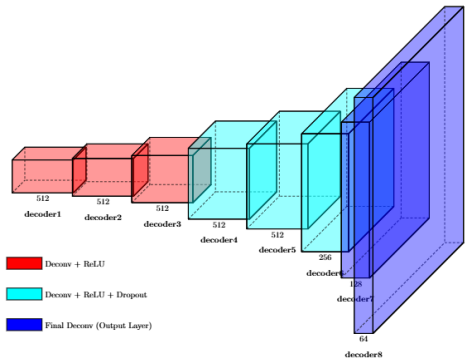
2] GAN

In game theory terms, the generator acts as the first player, aiming to make the artificial image indistinguishable from the real one.



Generator - Decoder

2] GAN



Discriminator

2] GAN

The discriminator serves as the second player, attempting to determine whether the aerial image is artificial or real.

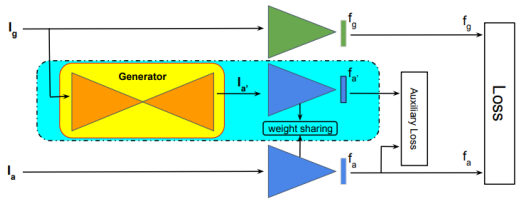
$$\begin{aligned}\min_G \max_D L_{\text{cGAN}}(G, D) &= \mathbb{E}_{I_g, I_a \sim p_{\text{data}}(I_g, I_a)} \left[\log D(I_g, I_a)\right] \\ &\quad + \mathbb{E}_{I_a, I'_g \sim p_{\text{data}}(I'_a, I_g)} \left[\log(1 - D(I_g, I'_a))\right] \\ \min_G L_{L1}(G) &= \mathbb{E}_{I_a, I'_a \sim p_{\text{data}}(I_a, I'_a)} \left[\|I_a - I'_a\|_1\right]\end{aligned}$$

Table of Contents

3] Joint Feature Extractor

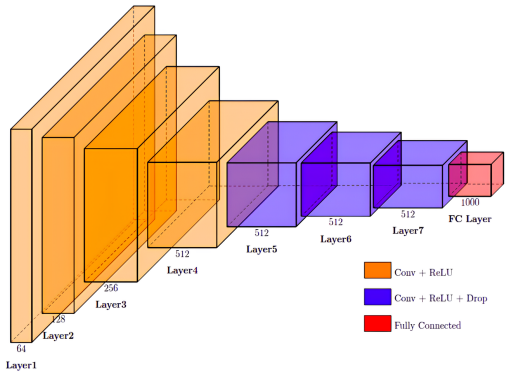
Recall

3] Joint Feature Extractor



VGG

3] Joint Feature Extractor



Triplet Loss + Joint Loss

3] Joint Feature Extractor

The *triplet loss* ensures that the feature representations of matching image pairs (ground and aerial) are drawn closer together in the embedding space, while non-matching pairs are pushed further apart.

$$L_{\text{TripletLoss}} = \max(0, d_p - d_n + m) \tag{1}$$

$$L_{\text{joint}} = \lambda_1 \cdot L(I_g, I_a) + \lambda_2 \cdot L(I'_a, I_a) \tag{2}$$

where:

- A = Anchor (reference sample)
- P = Positive sample (same class as Anchor)
- N = Negative sample (different class)
- d = Distance function
- m = Margin to enforce separation

Table of Contents

4] Conclusion

Simulations

4] Conclusion

As suggested in the paper, the dataset I considered is CVUSA. To verify some results, I reduced the dataset to 1,200 pairs and trained models on Kaggle.

- GAN: link1
- Feature Extractor: link2

Results

4] Conclusion

The use of synthesized aerial images as a bridge between the two views helped reduce the domain gap, leading to better performance in matching.

Method	Top-1	Top-10	Top-1%
Two-stream baseline ($I_{a'}$, I_a)	10.23%	35.10%	72.58%
Two-stream baseline (I_g, I_a)	18.45%	48.98%	82.94%
Joint Feat. Learning ($I_{a'}$, I_a)	14.31%	48.75%	86.47%
Joint Feat. Learning (I_g, I_a)	29.75%	66.34%	92.09%
Feature Fusion	**48.75%**	**81.27%**	**95.98%**

My Results

4] Conclusion

Satellite Image original

Satellite Image artificial

Thank You!