Transformer-based Satellite Image and Segmentation Generation for Ground-to-Aerial Image Matching

Course of Computer Vision
Alessandro Carotenuto, 1847282

The Problem

The task consists in matching two images from two sets, one consists of ground-level photo images and the other of overhead aerial photos of the same portion of space. The query for this is the **ground image** and the methods in questions aim to associate a similarity score between said ground image and candidate matches.
The highest score is the actual match.

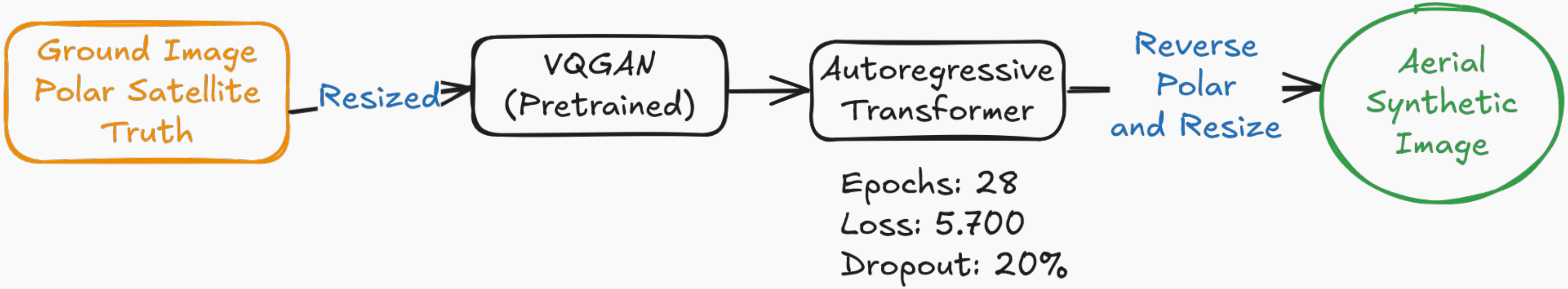
01

Proposed Method: Phase 1

Generate Synthetic Aerial Image (Polar)
Reverse Polar Transformation
6-Class Segmentation

Synthetic Aerial Images

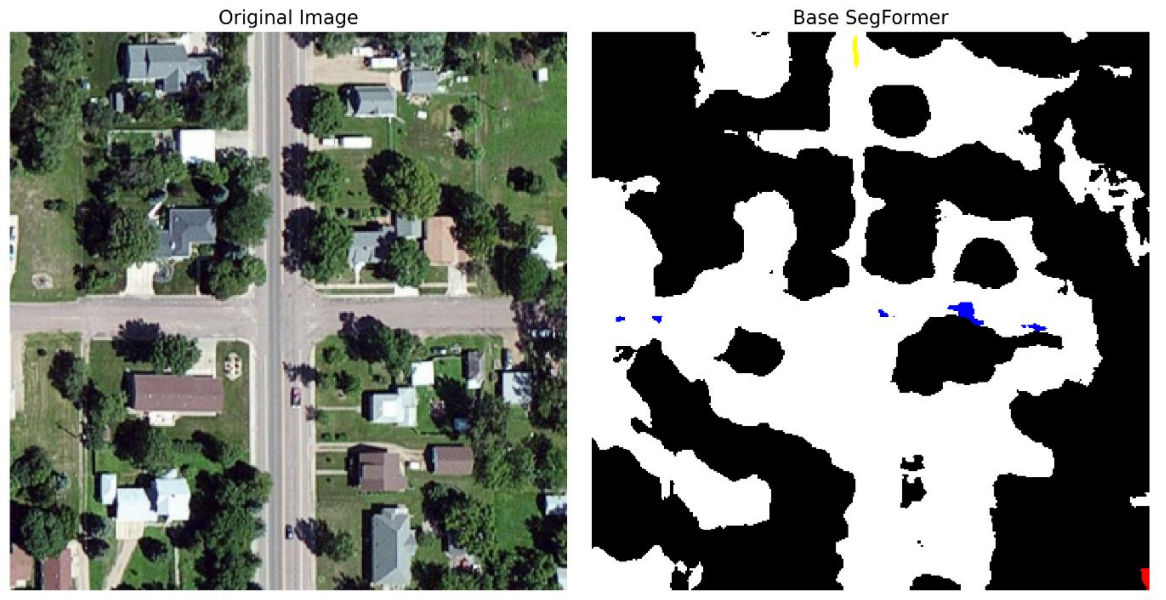
Inspired by the “Taming Transformers for igh-Resolution Image Synthesis paper” i used a VQGAN (which learns a codebook of discrete image tokens and a GAN-based decoder for high-quality detail) i used a VQGAN pretrained on **ImageNet** and then added, as in the paper, an **autoregressive Transformer** for the image generation



Synthetic Aerial Images

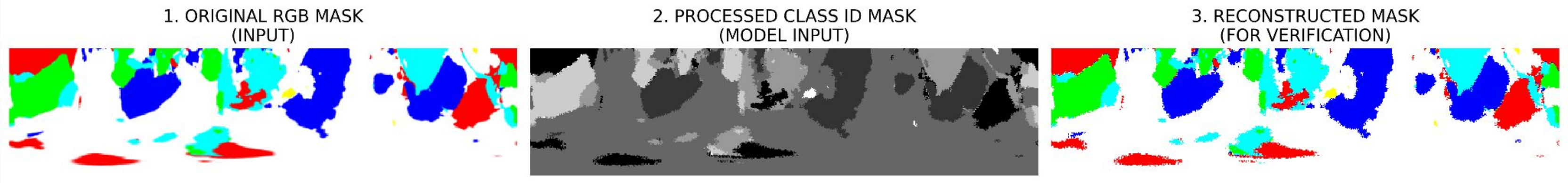
Semantic Segmentation

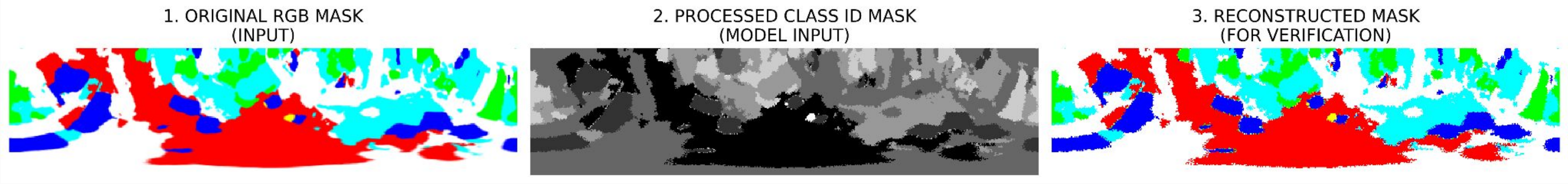
For the Semantic segmentation i started with the **pretrained SegFormer B5 finetuned** on the dataset **Cityscapes** obtaining, of course underwhelming results 'cause of the domain gap between the two datasets.



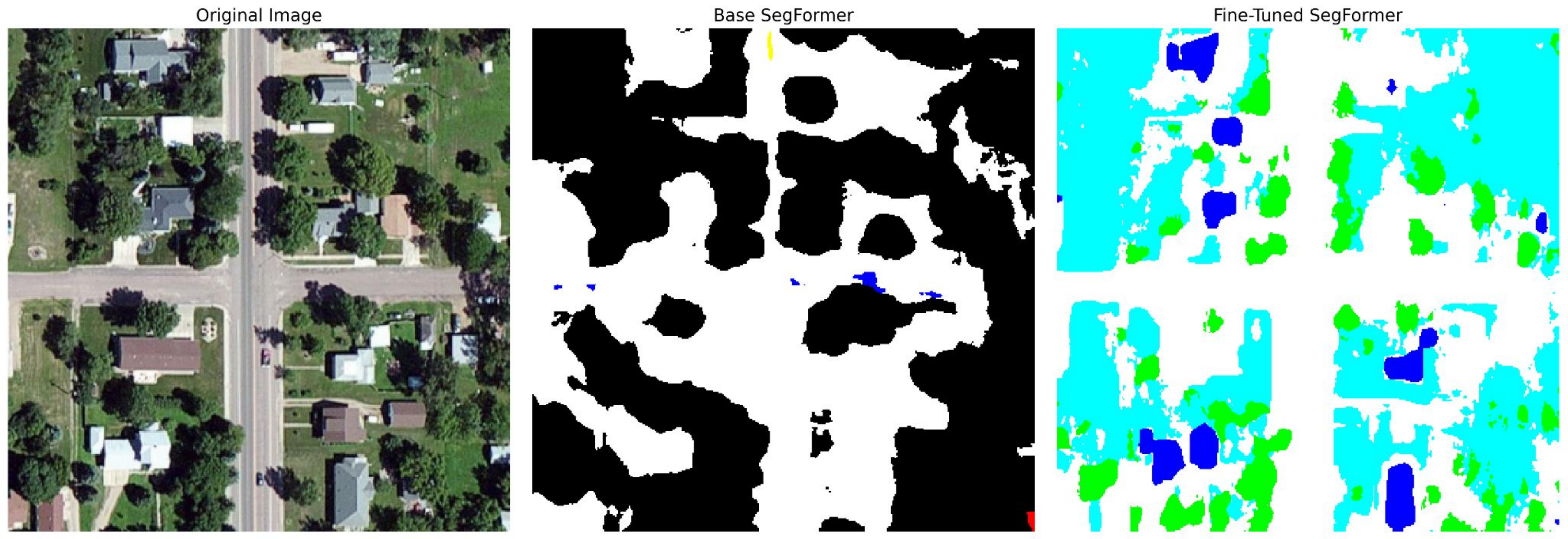
Semantic Segmentation

The Semantic masks had a hard to handle anti-aliasing, that i preprocessed out quantizing everything in the original 6-classes, so i **fine-tuned the model**

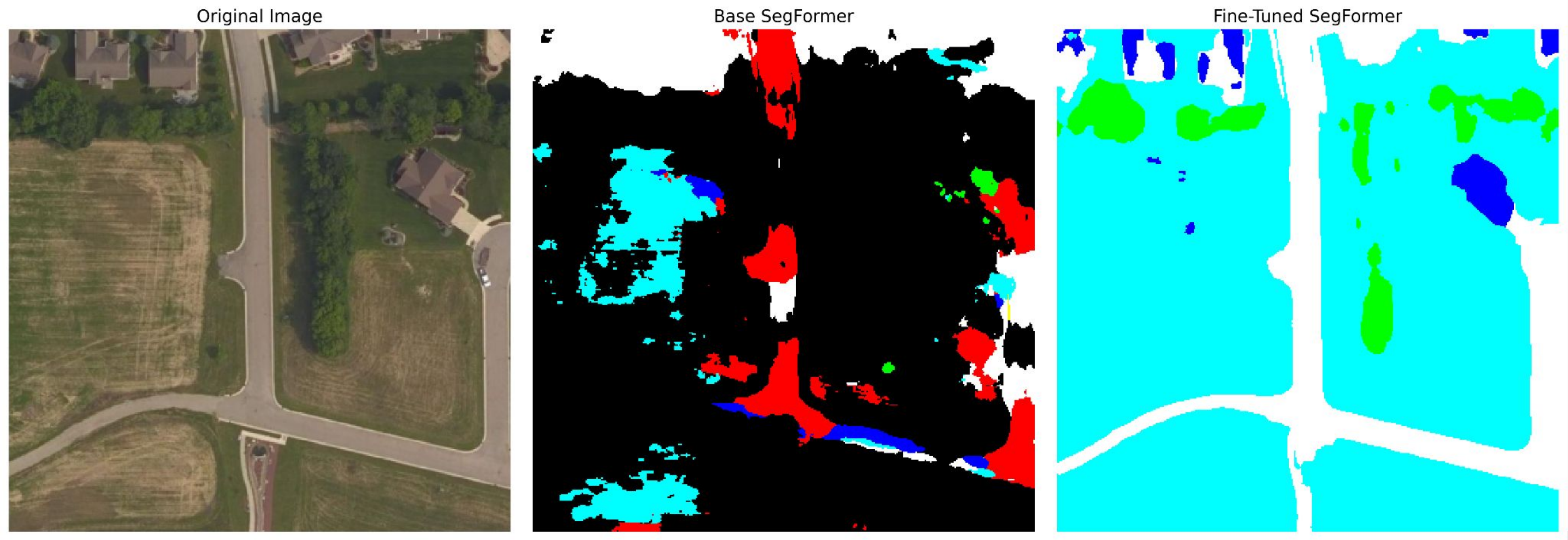




Semantic Segmentation

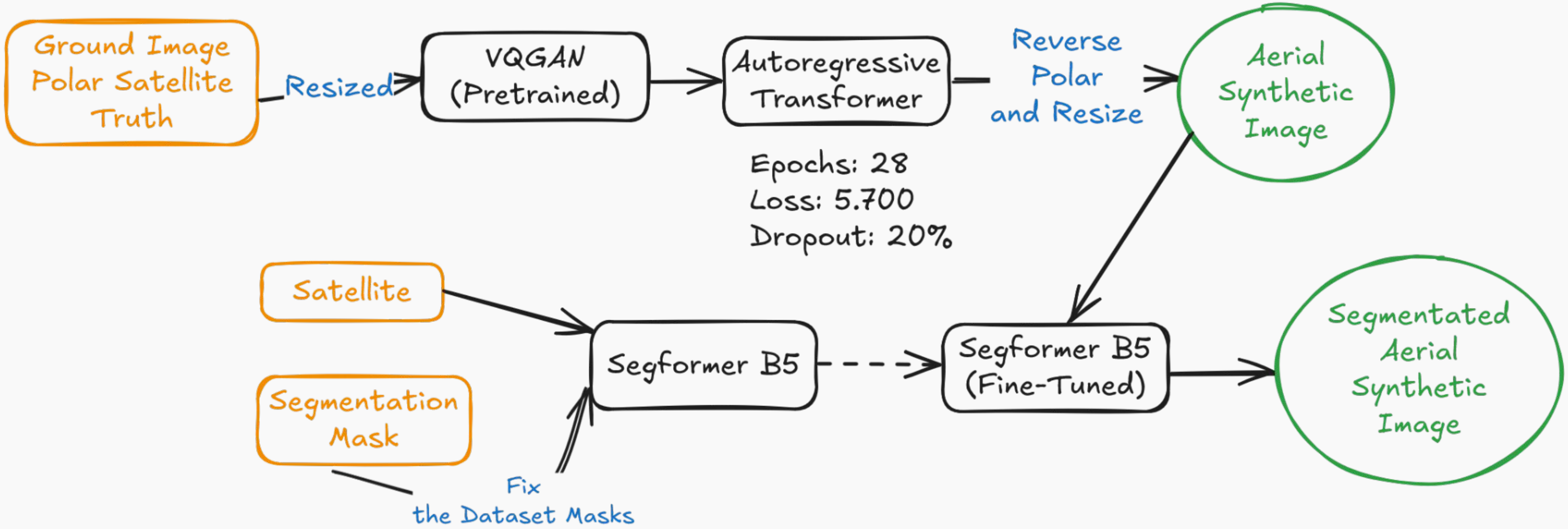


Semantic Segmentation



Phase 1 Example

Phase 1 Inference pipeline so far

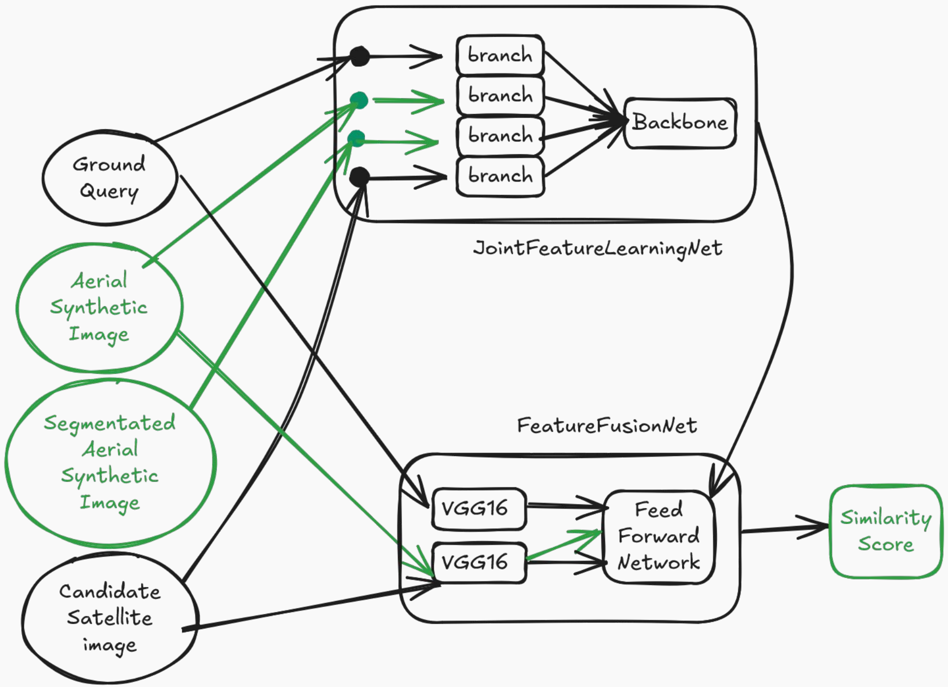


02

Phase 2

Parallel Networks & Similarity Score

Structure



Joint Feature Learning Net:
LR: 1e-3 (trained from scratch)

Each branch contains:

- Conv2D(3→128) → BatchNorm → ReLU → MaxPool
- Conv2D(128→256) → BatchNorm → ReLU → MaxPool
- Global Average Pooling → **256-dim feature vector**

After the common backbone we obtain a 2048-dim joint embedding of the features

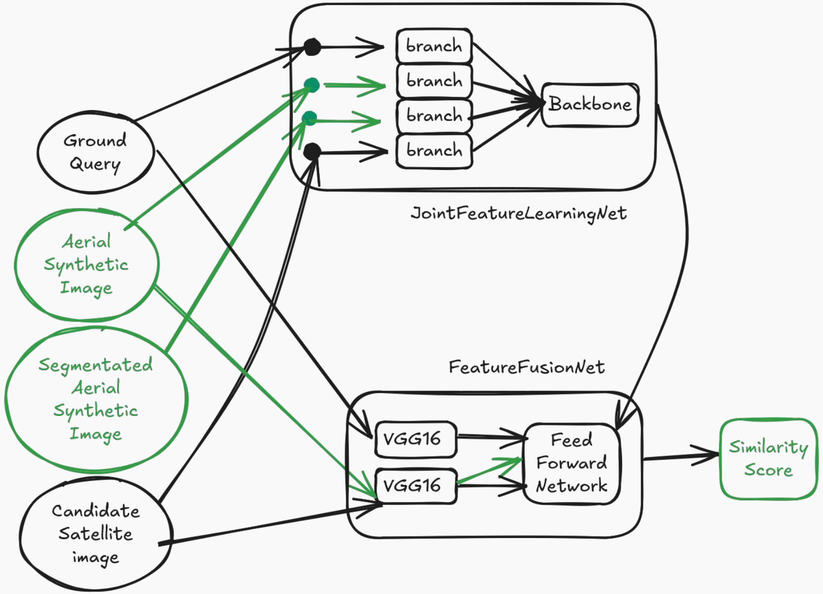
Structure

Feature Fusion Net

- Three VGG16 networks to fine tune (weight shared between two of them)
- **LR: 1e-5 (fine tuned)**

Fully Connected Layers

- Dense(14,336 → 7168) → ReLU → BatchNorm → Dropout(0.2)
- Dense(7168 → 3584) → ReLU → BatchNorm → Dropout(0.2)
- Dense(3584 → 1) → **Sigmoid output (similarity score)**
- **LR: 1e-3 (trained from scratch)**



Triplet loss is the choice in this system, it compares a ground image with a **positive satellite** match and a **negative** one. The loss encourages the model to assign a higher similarity score to the positive pair than to the negative.

References

- Regmi, K., & Shah, M. (2019). Bridging the Domain Gap for Ground-to-Aerial Image Matching. arXiv.
- F. Pro, N. Dionelis, L. Maiano, B. L. Saux and I. Amerini, ”A Semantic Segmentation-Guided Approach for Ground-to-Aerial Image Matching,” IGARSS 2024 - Athens, Greece, 2024, pp. 2630-2635
- Mule, E., Pannacci, M., Goudarzi, A., Pro, F., Papa, L., Maiano, L., and Amerini, I. (2025). Enhancing Ground-to-Aerial Image Matching for Visual Misinformation Detection Using Semantic Segmentation. In Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops (pp. 795-803).
- Esser, P., Rombach, R., & Ommer, B. (2021). Taming Transformers for High-Resolution Image Synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12873–12883.