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

Computer Vision A.A. 2024-2025

Project presentation by:

Filippo Tatafiore 1934988
Lorenzo Ciappetta 2011296

DIPARTIMENTO DI INGEGNERIA INFORMATICA AUTOMATICA E GESTIONALE ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

Table of Contents
Presentation outline

Objective

Area of interest of the project

- Ground-to-Aerial image matching is the problem of associating a query ground-view with the corresponding satellite image, despite of the extreme view-point difference.

- Objective:
    - Learn to synthesize aerial images through a transformer model for image generation
    - Learn to extract the characteristics and features of the images.
    - Correlate the features from the different images, and evaluate it with a score.
    - Discriminate non-matching views from matching ones

State of the art

Current research approaches to this challenge

- Popular Approaches based on:
  - Feature extraction through VGG models and correlation between separate branches
  - Triplet loss

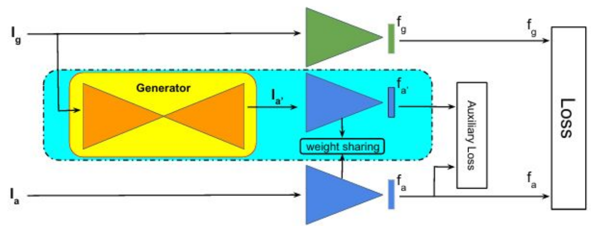


figure 1: JointFeatureLearningNet architecture from [1]

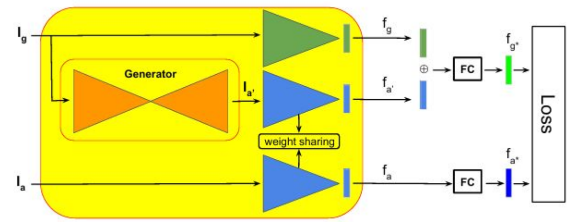


figure 2: FeatureFusionNett architecture from [1]

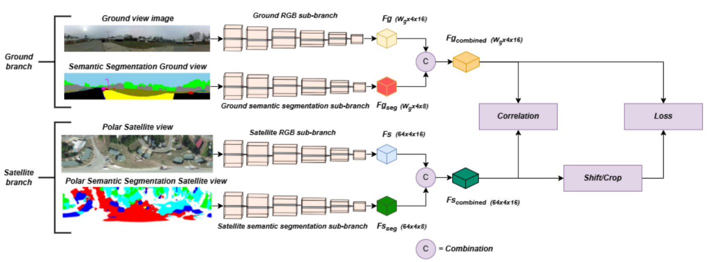


figure 3: SAN-QUAD Architecture from [2]

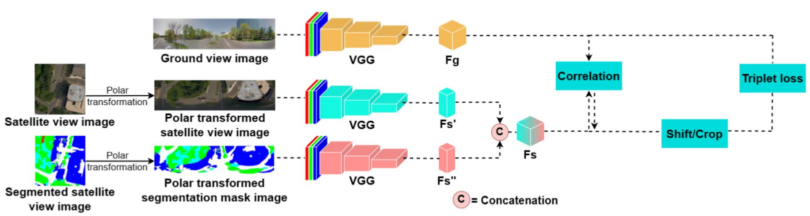


figure 4: SAN Architecture from [3]

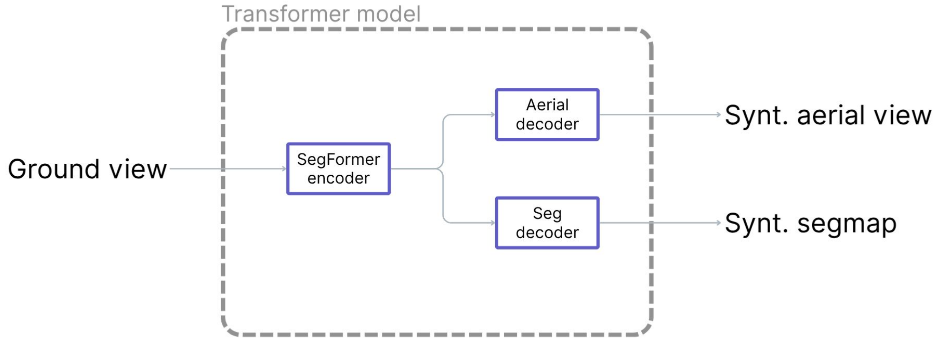
Proposed method

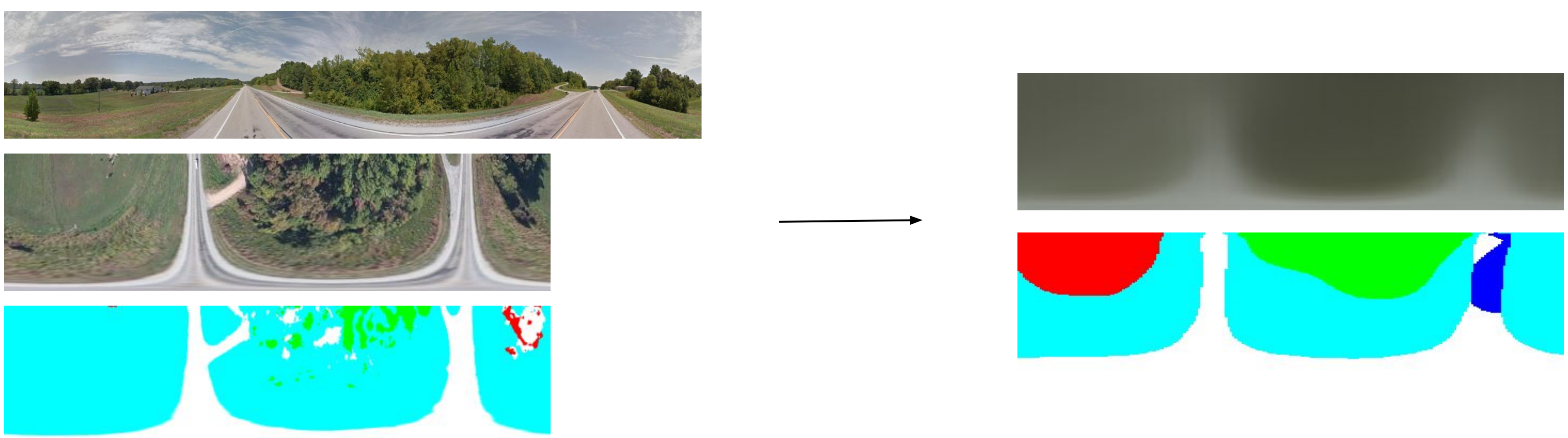
Project pipeline

- Image Generation, Feature Extraction and Fusion:
  - Fine-tune Image Generation model to produce expected aerial views and segmentation maps from ground images
  - Use generated images along with the others to aid in feature extraction
  - Fuse salient features to elaborate the compatiblity between query ground view and aerial candidate image
- Image Discrimination
  - Train the model for matching to true aerial image representation from wrong ones via triplet loss

Proposed method
Models

- Image Generation: Vision Transformer model

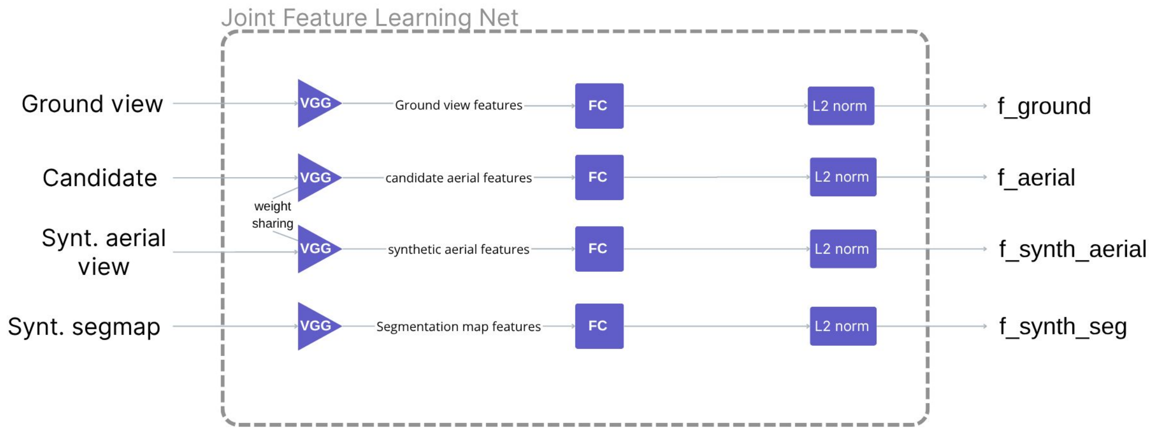


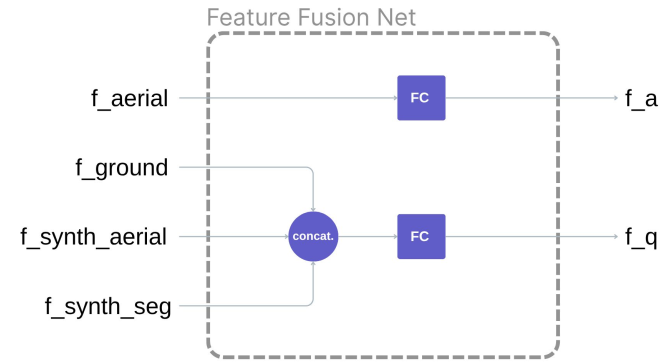


Proposed method

Models

- Feature Extraction and Fusion: Joint Feature Learning Net and Feature Fusion Net

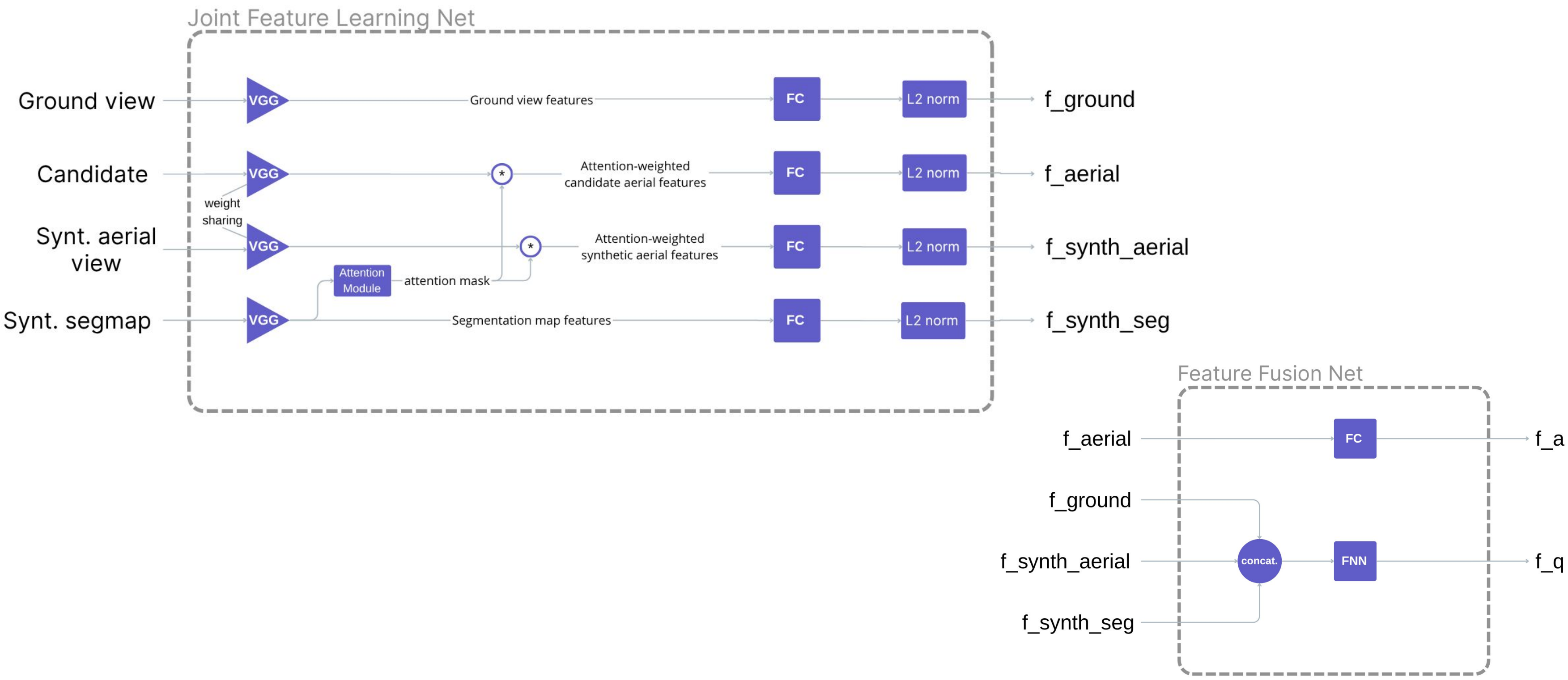




Proposed method

Another approach

- Feature Extraction and Fusion: Joint Feature Learning Net and Feature Fusion Net



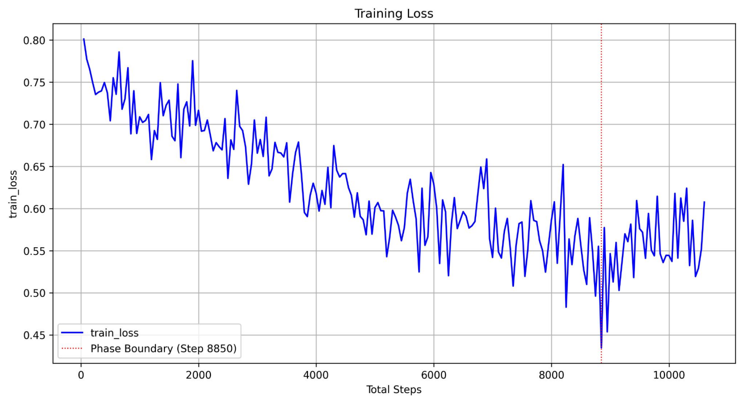
Dataset

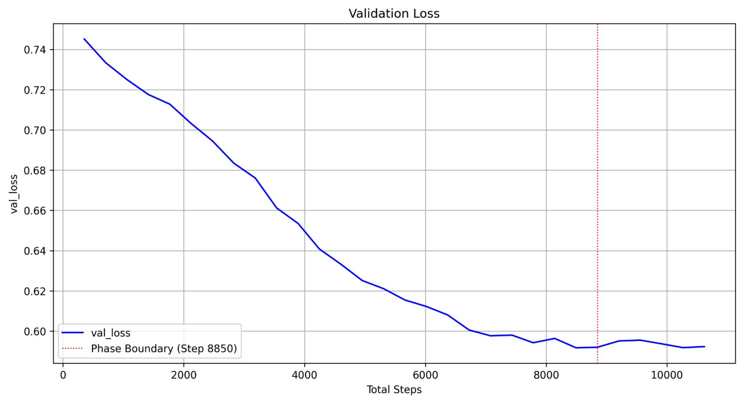
- CVUSA
  - polar-transformed ground view images
  - aerial satellite images
  - aerial satellite segmentation maps
  - polar-transformed aerial satellite images
  - polar-transformed aerial satellite segmentation maps
- Preprocessing
  - Resize (adapt to pretrained models)
  - Data augmentation
  - Normalization

Setup and Training

Configuration of the elements of the project

- Generative model
  - encoder: version of nvidia/segformer-b0-finetuned-ade-512-512, fine-tuned on satellite semantic dataset [4]
  - training strategy: 2 phases

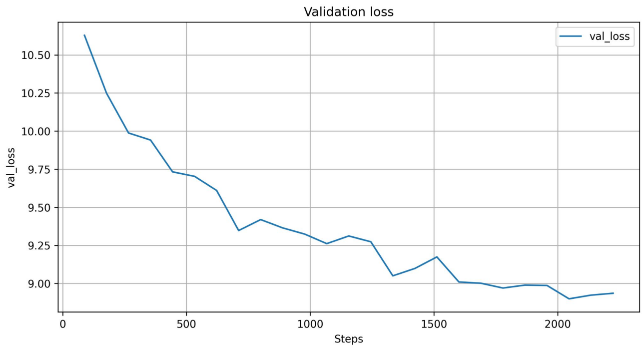




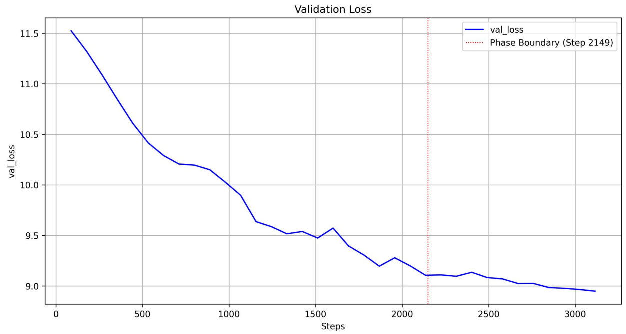
Setup and Training

Configuration of the elements of the project

- Joint Feature Learning Net
  - Logs for first approach:



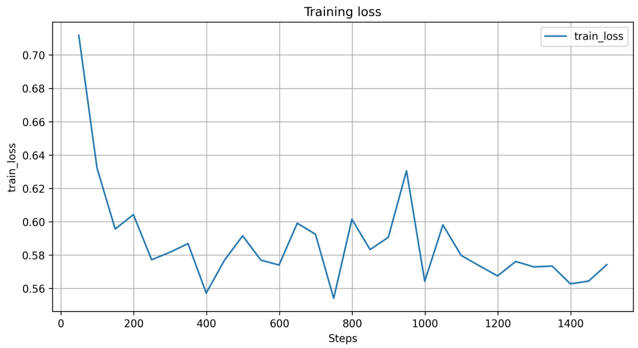
  - Logs for second approach (with attention module):

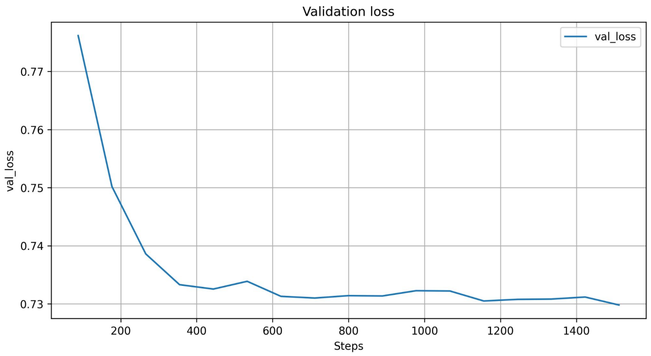


Setup and Training

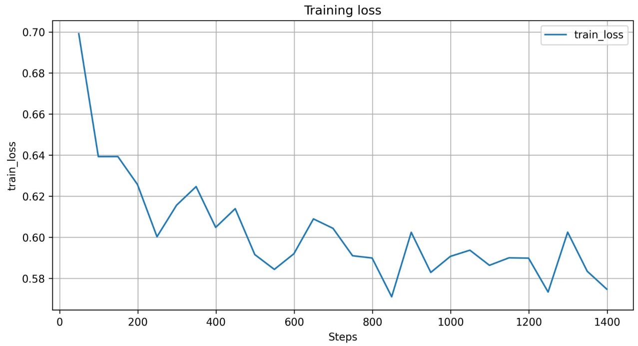
Configuration of the elements of the project

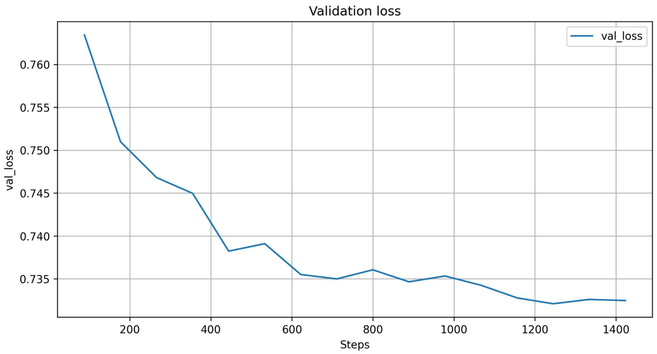
- Feature Fusion Net
  - Logs for first approach:





  - Logs for second approach (shallow NN instead of FC):





Evaluation

- Generative model
  - Aerial Loss: weighted sum of Perceptual Loss and L1 loss
  - Segmentation loss: Weighted Focal Loss + Dice Loss
  - PSNR and IoU Monitoring
- Joint Feature Learning Net model
  - Weighted sum of three triplet losses
- Feature Fusion Net model
  - Triplet loss
  - Recall@k

```
============ Test set results ============
Avg Segmentation Loss: 0.8044
Avg Aerial Reconstruction Loss: 0.5933
Avg Total Loss: 0.7247
```

Test metric	DataLoader 0
loss 1	0.7349947690963745
loss 2	0.8178892135620117
loss 3	0.8212516903877258
test_loss	8.98908805847168

Test metric	DataLoader 0
test_loss	0.7335683703422546

```
Recall@1: 0.5621
Recall@5: 0.8239
Recall@10: 0.9016
Recall@20: 0.9431
```

Conclusion
Final considerations

- Generative model:
  - different encoders: nvidia/mit-b1, sawthiha/segformer-b0-finetuned-deprem-satellite
  - different losses:
    - Aerial loss: L1 loss, Perceptual Loss
    - Segmentation loss: Weighted Cross Entropy, Weighted Focal Loss, Dice loss

- Joint Feature Learning model:
  - different architectures: with or without attention

- Feature Fusion model:
  - different approaches for extracting the query embedding

Conclusion

Future work

- Generative model:
  - larger encoder, pre-trained on satellite data (Prithvi, Swin Transformer, …)
  - deeper decoders
- Joint Feature Learning and Feature Fusion:
  - use also ground view segmentation maps
  - calculate negative not only within the current batch
- Other possibilities:
  - Larger dataset
  - Different training strategies

References

[1] Regmi, K., & Shah, M. (2019). Bridging the Domain Gap for Ground-to-Aerial Image Matching. arXiv.

[2] 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).

[3] 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.

[4] https://huggingface.co/sawthiha/segformer-b0-finetuned-deprem-satellite

Thanks for your attention