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

Andrei Halitchi - 1836170

Sapienza University of Rome – Computer Vision Course

23/07/2025

Outline

Problem Statement

- **Task**: Match a panorama ground-level photo to its corresponding satellite image within a large gallery.
- **Input**: Ground image, Satellite view image and Segmented satellite image.
- **Output**: Rank list of satellite candidates evaluated using Recall@K.

Related Work

- **DSM** — Dynamic Similarity Matching netowrk (Shi et al., 2020).
- **SAN** — Semantic Align Net (siamese-like) (Pro et al., 2024).
- **SAN-QUAD**, Quadruple Semantic Align Net (four-stream siamese-like), (Pro et al., 2025).

Overview of the implemented Pipeline

Implemented pipeline

1. VGG-16 backbone (pre-trained on ImageNet).
2. Orientation correlation + circular padding.
3. Two enhancements:
   - **Attention maps**.
   - **Sky removal** using pre-trained model.
4. Triplet loss on correlation distance.

Tested configurations

1. **BASE**: identical as in Pro et al. 2020.
2. **ATTENTION**: BASE + attention map.
3. **SKYREMOVAL**: BASE + sky removal.
4. **FULL**: BASE + attention map and sky removal.

Overview of Our Pipeline

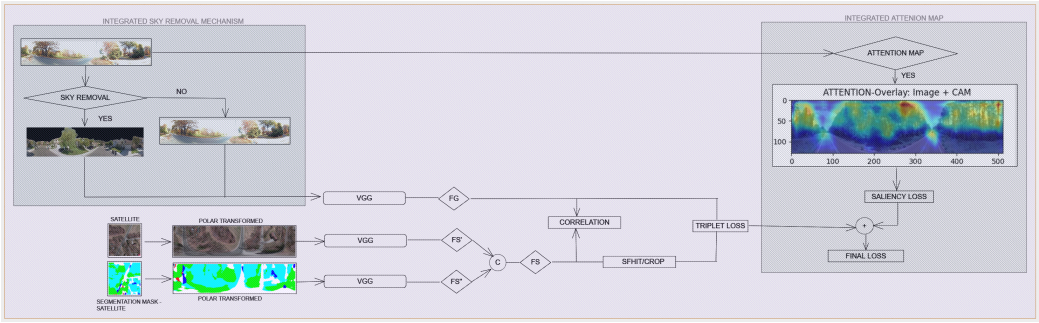


Figure: Used Architecture

Loss definition

Let N be the mini-batch size. For each ground panorama $\mathbf{g}_i$ and its matched satellite patch $\mathbf{s}_i$ we obtain a cosine-derived distance

$$d_{ij} \;=\; 2 \;-\; 2\,\langle \mathbf{g}_i, \mathbf{s}_j \rangle, \qquad i, j = 1, \ldots, N,$$

so that d_{ii} is the *positive* distance and d_{ij} with $i \neq j$ are *negative* distances. With a scale factor $\gamma > 0$ (named `loss_weight` in the code, default $\gamma = 10$) the loss is

$$\mathcal{L}_{\text{triplet}} = \frac{1}{2N\,(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \Big[\underbrace{\log\!\big(1 + e^{\gamma\,(d_{ii} - d_{ij})}\big)}_{\text{ground}\to\text{satellite}} + \underbrace{\log\!\big(1 + e^{\gamma\,(d_{ii} - d_{ji})}\big)}_{\text{satellite}\to\text{ground}} \Big]. \tag{1}$$

Saliency Loss

For the ATTENTION and FULL variants a Grad-CAM map $\mathrm{CAM}_i \in [0,1]^{H \times W}$ is extracted from the ground branch for each $\mathbf{g}_i$. Following the implementation,

$$\mathcal{L}_{\text{saliency}} = -\frac{1}{N\,H\,W} \sum_{i=1}^{N} \sum_{p=1}^{H \times W} \mathrm{CAM}_i(p), \tag{2}$$

so that higher average activation (larger CAM values) *reduces* the loss and is therefore encouraged.
The total objective becomes

$$\boxed{\mathcal{L} = \mathcal{L}_{\text{triplet}} + \lambda_{\text{sal}}\,\mathcal{L}_{\text{saliency}}} \tag{3}$$

with the weighting coefficient λ_{sal} given in the configuration (`gradcam_config['lambda_saliency']`).

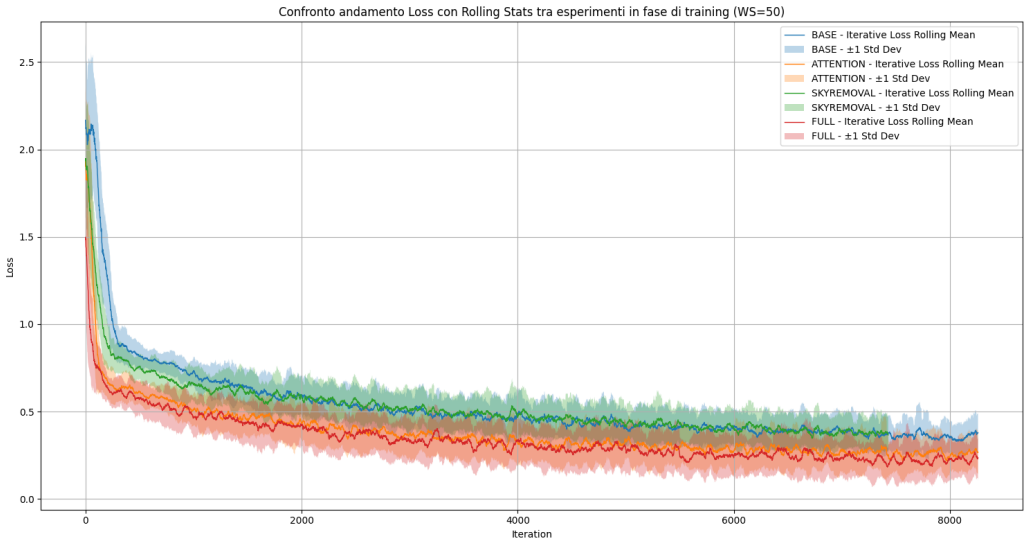
Dataset – CVUSA Subset

- **CVUSA** originally 35k train, 8k test pairs
- Train set composed by 6647 samples
- Test set composed by 2215 samples
- Each sample stores satellite image, segmented satellite image and ground view.
- Pre-processing:
  - Resize satellite and segmented satellite images to 512×128 using polar transformed operation.
  - Apply sky-mask when experiment `SKYREMOVAL` or `FULL`.

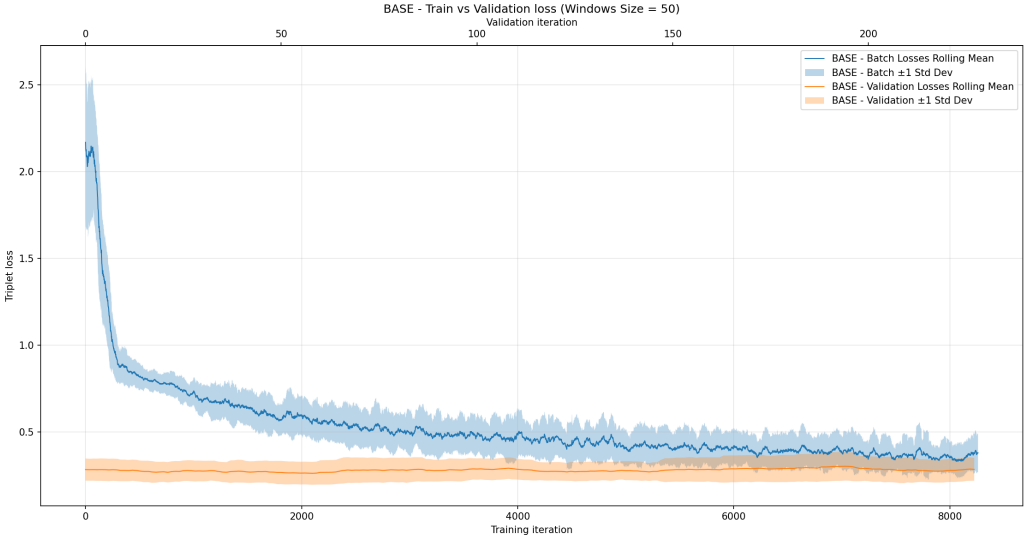
Settings

- Framework: PyTorch 2.3 (rewrite of legacy TensorFlow code).
- Optimiser: Adam ($\eta = 1 \times 10^{-5}$), batch = 4*8=32, epochs = 10.
- Train iterations = 6647 / 8 * 10 = 8308
- Test iterations = 2215 / 8 * 1 = 277
- Triplet loss weight = 10.0
- Saliency loss weight = 5.0

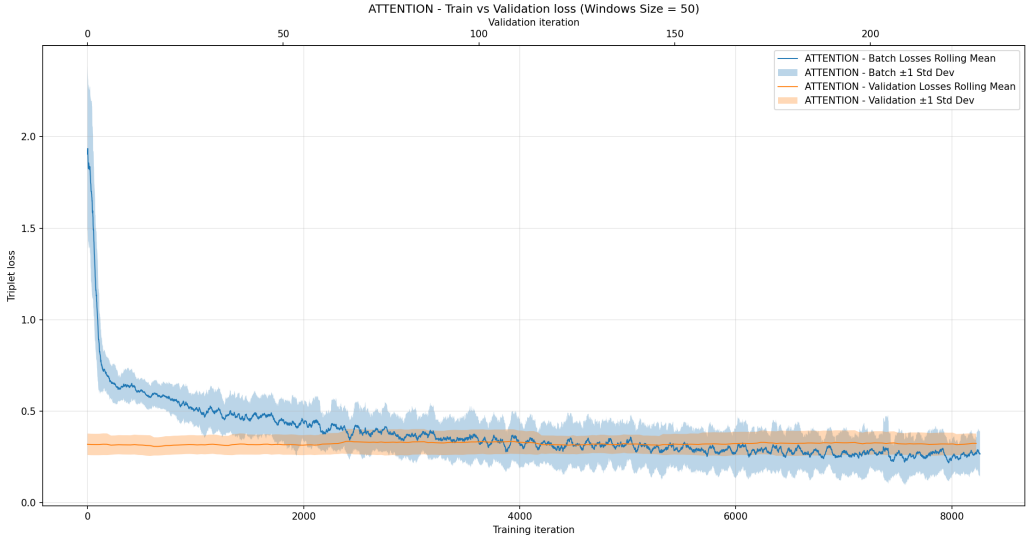
Train Loss Curves



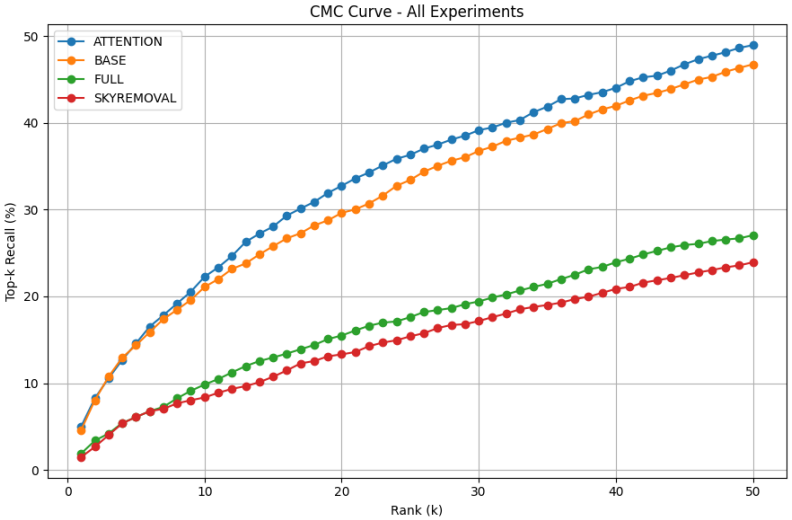
Validation Loss Curves



Validation Loss Curves



Cumulative Match Characteristic Curve



ATTENTION - Some Attention Maps

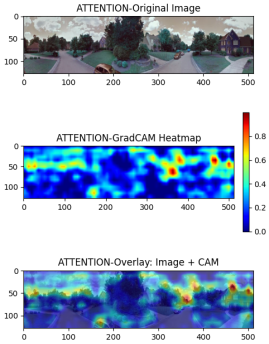


Figure: epoch 3 - iter 30

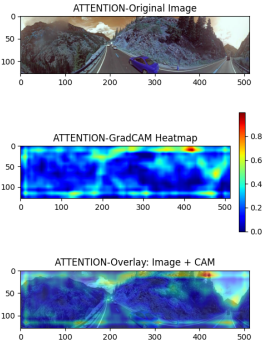


Figure: epoch 5 - iter 150

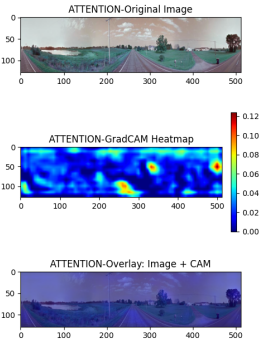


Figure: epoch 10 - iter 40

FULL - Some Attention Maps

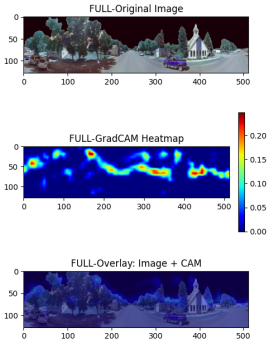


Figure: epoch 3 - iter 120

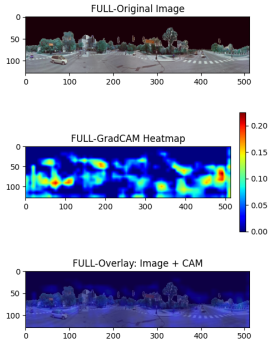


Figure: epoch 5 - iter 170

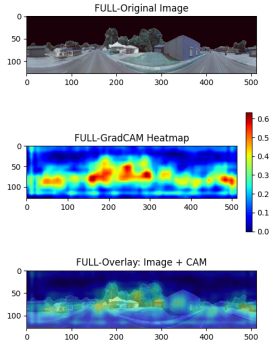


Figure: epoch 10 - iter 60

BASE - Image retrieval examples

ATTENTION - Image retrieval examples

Recall@K on CVUSA Subset

Model	R@1	R@5	R@10	R@1 %
BASE	4.6	14.4	21.1	30.7
ATTENTION	5.0	14.5	22.3	34.3
SKYREMOVAL	1.5	6.1	8.4	14.3
FULL	1.9	6.1	9.8	16.6

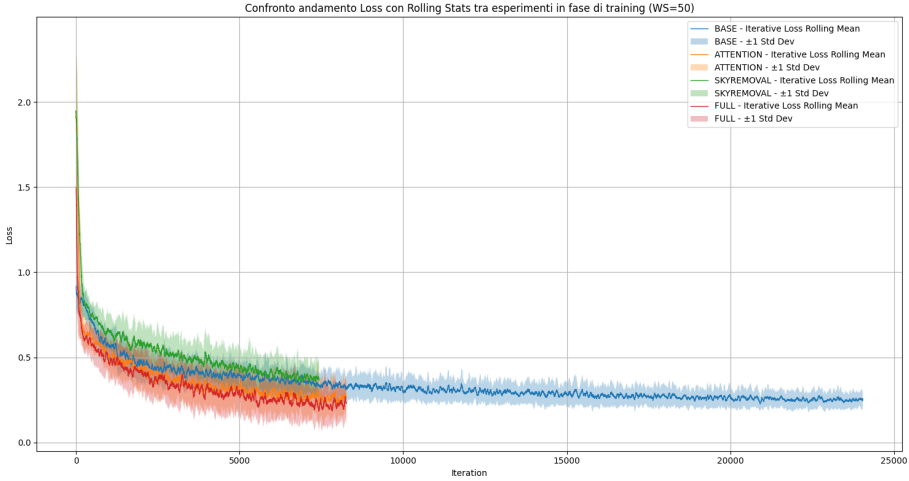
- Attention improves retrieval modestly (+0.4 pp R@1).
- Sky removal alone degrades, likely due to over-masking.

Corrections

- Normalization of ground features in forward pass
- Wrong division in training loss
- Evaluation.py to be reviewed

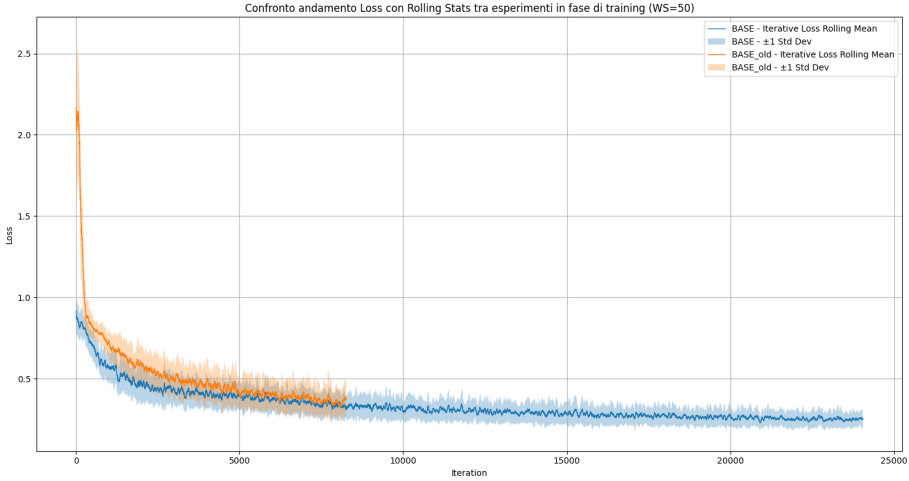
BASE - Updates

Updated Train Loss Curves

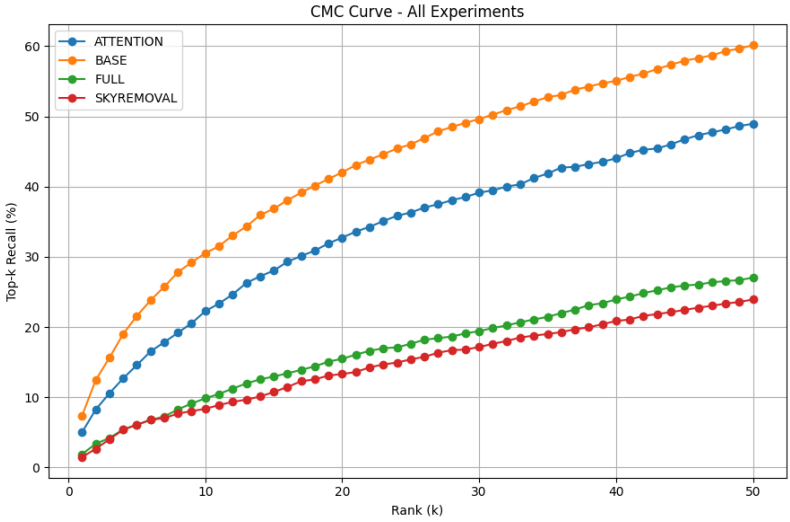


BASE - Updates

Comparison between old and new BASE Train Loss



BASE - Updates



Updated CMC Curve

References I

- [1] Shi, Yu, et al. "Where am I looing at? Joint Location and Orientatio Estimation by Cross-View Matching" 2020
- [2] Pro, Dionelis, et al. "A Semantic Segmentation-Guied Approach For Gournd-to-Aerial Image Matching" 2024
- [3] Mule, Pannacci, et al. "Enhancing Ground-toAerial Image Matching for VIsual Misinformation Detection Using Semantic Segmentation" 2025