# Enabling Rapid Disaster Response: Multimodal Remote Sensing Coregistration

Evan Twarog
Stanford University
etwarog@stanford.edu

## Abstract

*Aligning keypoints between electro-optical (EO) and synthetic aperture radar (SAR) satellite imagery is crucial for rapid disaster response but remains challenging due to fundamentally different imaging modalities. In this work, we use coarsely geo-referenced EO–SAR tile pairs from the SpaceNet 9 Challenge (approx. 400 tie-points) as input and train convolutional models to predict pixel-wise keypoint heatmaps. We first benchmark a Vanilla U-Net baseline, then shift to a residual blocks (ResU-Net) approach which improves stability on our limited dataset. Finally, we introduce cross-modal attention—letting SAR and EO features explicitly attend to one another at each encoder scale—which yields a further reduction in keypoint localization error compared to both Vanilla U-Net and standard ResU-Net. Our best model achieves a 43.0% decrease in End Point Error on held-out tiles. Future work will expand the training set and explore real-time inference, with the goal of accelerating life-saving mapping tasks in disaster scenarios.*

## 1. Introduction

Accurate sensor fusion of Synthetic Aperture Radar (SAR) and electro-optical (EO) satellite imagery is essential for reliable change detection and damage assessment in large-scale disaster response. In the aftermath of earthquakes, floods, or other major events, optical imagery is often are limited due to cloud cover, smoke, or inadequate lighting, whereas SAR imaging can penetrate clouds and collect data both during the day and night. By aligning or ”coregistering” SAR and EO images, emergency responders gain a more complete and complementary view of the affected region, allowing rapid damage mapping and informed allocation of resources [6]. Despite its importance, precise SAR - EO coregistration remains challenging because the two modalities record fundamentally different imaging physics: EO sensors capture reflected visible-spectrum radiance, while SAR sensors measure microwave backscatter intensity. This leads to divergent contrast patterns, speckle noise in SAR, and occlusion artifacts in optical images, causing traditional feature-matching methods to fail in many real-world scenarios [6].

Manual annotation of keypoint correspondences between SAR and EO over hundreds of square kilometers is labor-intensive and prone to error [6]. Our motivation is to develop a deep-learning approach that learns modality-invariant representations to automatically detect and match keypoints, thereby reducing human workload and accelerating end-to-end registration pipelines.

The input to our algorithm is a pair of coarsely geo-referenced GeoTIFF images—one RGB optical and one single-channel SAR—covering the same area of interest (AOI) at roughly 0.3–0.5 m ground sampling distance. Each GeoTIFF includes metadata that guarantees coarse alignment [6]. We preprocess by tiling each image into $512 \times 512$ patches containing candidate keypoint regions. Given a tile pair $(I_{\text{opt}}, I_{\text{SAR}})$, our U-Net-based network with cross-attention modules outputs a per-pixel likelihood heatmap $H$, whose peak corresponds to the predicted keypoint match. In other words, the final output is a set of predicted coordinate pairs $\{(x_i^{\text{opt}}, y_i^{\text{opt}}) \leftrightarrow (x_i^{\text{SAR}}, y_i^{\text{SAR}})\}$ for each tile. Figure 1 illustrates an example of cross-modal keypoint matching between SAR and EO imagery.

In the remainder of this paper, Section 2 reviews related work on multimodal registration and deep keypoint matching. Section 3 details our U-Net with cross-attention architecture, loss formulation, and training procedure. Section 4 describes data preprocessing, tile generation, and feature extraction. Section 5 presents quantitative and qualitative experiments on SpaceNet 9 test sets, including ablation studies on hyperparameters and architectural choices. Finally, Section 6 concludes and outlines future directions, such as extending to other sensor pairs (e.g., SAR - hyperspectral) and incorporating geometric consistency constraints.

## 2. Related Work

Computer vision for remote sensing has advanced quickly thanks to more multimodal imagery and the use of transformer models. We group prior work into three

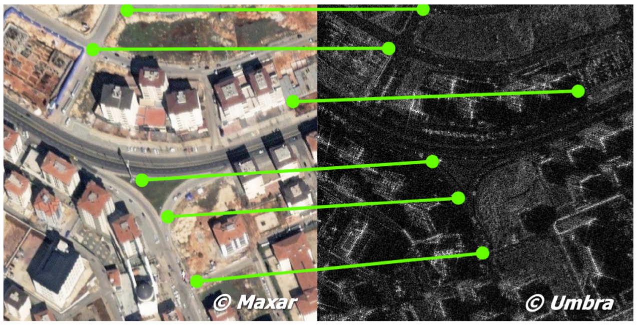

Figure 1. Example of cross-modal keypoint matching between SAR and EO imagery.

categories: (i) transformer-based segmentation and change detection, (ii) attention-augmented and lightweight CNN fusion, and (iii) multimodal registration and feature-based alignment. Below, we describe key papers in each group.

## Transformer-Based Segmentation and Change Detection

Transformers have gained traction in recent years for their performance when deployed at scale, and different approaches tackle remote sensing problems. FTransUNet [10] combines a CNN branch with a transformer branch. It uses Squeeze-and-Excitation and Adaptively Mutually Boosted Attention (Ada-MBA) to fuse features at multiple levels. On the WHU-RS19 dataset, it gets very high accuracy, but it needs about 150 million parameters and around 48 GB of GPU memory, so it is slow to train. BIT [2] cuts down on tokens by turning image patches into smaller semantic tokens before feeding them to the transformer. This saves a lot of computation: on LEVIR-CD, BIT reaches about 85 % mIoU with roughly half the FLOPs of a full transformer. Dahal et al. [4] compare MaskFormer (with a Swin-Large backbone) to a U-Net CNN that uses a special weighted loss. On the iSAID dataset, they show that with the right loss, a 42 million-parameter U-Net can reach within 7 % of MaskFormer's 88 % mIoU while cutting inference cost in half. These papers show that transformers are very good at looking at long-range and cross-modal context but can be too big or slow for real-time use.

## Attention-Augmented and Lightweight CNN Fusion

Innovation continues with CNNs, particularly when focused on incorporating principles such as attention. MAResU-Net [9] adds Linear Attention Mechanisms (LAM) into U-Net skip connections. It uses a ResNet-34 backbone to get global context with less computation. On the ISPRS Vaihingen dataset, MAResU-Net achieves about 83.3% mIoU while using around 60 % fewer parameters than big transformers. Xiao et al. [17] propose Enhanced Interlayer Feature Correlation (EFC), which replaces the standard FPN. They use two modules—Grouped Feature Focus (GFF) and Multilevel Feature Reconstruction (MFR)—to boost small-object detection mAP by 1.7–3.1 % on VisDrone and UAVDT, while reducing GFLOPs by up to 42.7 %. SuperYOLO [20] adds a super-resolution stage (residual dense blocks) before YOLO detection. This improves small-object mAP by 5–8 % on DOTA and HRSC. These models show that you can get many of the benefits of transformers but with much lower cost: MAResU-Net balances accuracy and memory, EFC focuses on interlayer links, and SuperYOLO uses super-resolution to help detection.

## Multimodal Registration and Feature-Based Alignment

Sensor fusion remains an active area of research. PSRNet [18] treats registration as a two-way regression problem. It uses a Twins-SVT backbone with two branches (one for each modality) to get features at three scales. Then a Progressive Cross-Modal Transformer refines the deformation field step by step, using a consistency loss that keeps the two directions aligned. On the HMRSIR dataset, PSRNet cuts endpoint error by 52–69 % compared to older methods, but it uses about 300 GFLOPs and 120 million parameters, which is very heavy. CIRSM-Net [16] adds SAR physics into the feature extraction and uses a cyclic LSTM optimizer guided by RIFT2 supervision. On SEN1-2 and WHU-OPT-SAR, it halves endpoint error. Both PSRNet and CIRSM-Net are top of the line but require a lot of computation. Older feature-based methods are still useful when compute is limited: Yu et al. [19] use "triangular features" from road intersections and get over 80 % correct matches even under strong radiometric changes (Potsdam and Niagara). Chen et al. [3] propose ISIFT, which is an iterative loop of SIFT and RANSAC. On TerraSAR-X vs. Landsat-8, ISIFT reduces RMSE by about 30 %. Roadcross works well in cities but fails outside of road networks; ISIFT is simple but can break when there is little texture.

## Discussion

Transformer methods like FTransUNet, BIT, and MaskFormer are the best when you have enough compute. BIT's token reduction and FTransUNet's Ada-MBA are especially clever. But they need a lot of memory and take a long time to run. Attention-augmented CNNs like MAResU-Net, EFC, and SuperYOLO get much of the same benefit without such high cost—MAResU-Net's linear attention, EFC's GFF/MFR, and SuperYOLO's super-resolution all show good trade-offs. For registration, PSRNet's bidirectional refinement is top-performing but heavy, while roadcross and ISIFT are lighter and still work well when data or compute is limited (often combined with manual checking). Overall, most people use a mix of automated methods plus manual validation; fully unsupervised end-to-end registration is still an open challenge.

## 3. Data

In this project, we leverage the SpaceNet 9 dataset, which consists of high-resolution multi-modal satellite imagery and associated ground-truth tie-points specifically curated for cross-modal registration tasks in earthquake-affected regions. The primary data modalities are:

- **Optical (Electro-Optical) Imagery:** Three-band (RGB) GeoTIFFs at approximately 0.3–0.5 m spatial resolution, provided through the Maxar Open Data Program [14]. These images capture the visual (chemical-reflectance) properties of the Earth's surface shortly before or after seismic events.

- **Synthetic Aperture Radar (SAR) Imagery:** Single-band GeoTIFFs at approximately 0.3–0.5 m resolution, supplied by UMBRA [8]. SAR data measure physical (backscatter) properties of the same areas, collected via side-looking geometry, enabling cloud-penetrating, day-night acquisition.

Each Optical–SAR pair corresponds to one Area of Interest (AOI). For the development (training) set, there are three image pairs, each accompanied by a CSV of manually labeled tie-points manually labeled by SpaceNet 9 organizers.. Each AOI's tie-point CSV contains roughly 100-150 point correspondences—that is, pixel locations in the optical image matched to pixel locations in the SAR image. In total, the training set includes approximately 400 tie-points across the three scene pairs.

**Dataset Size**

- **Training Imagery:** 3 optical + 3 SAR GeoTIFFs, each about 13,000 × 13,000 pixels (approximately $1.69 \times 10^8$ pixels per image).

- **Tie-Points:** The challenge curated a list of approximately 100-150 tie-points per AOI for a total of 400 tie-points.

- **Tile-Based Subsets:** For training keypoint detectors, we extract $512 \times 512$ pixel tiles, with each tile containing a labeled tie-point. Each $512 \times 512$ patch spans approx. 150 m × 150 m on the ground (at 0.3 m per px).

**Pre-Processing and Special Treatment** To train deep networks for cross-modal keypoint detection and to estimate pixel-wise transformation maps, we performed the following preprocessing steps:

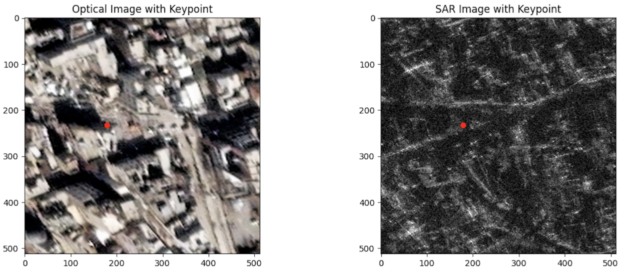

Figure 2. Sample EO and SAR tiles generated during data processing. Notice the keypoint is slightly offset in each image (hence the registration problem).

**Tie-Point Tile Extraction** For each labeled tie-point, we extract a $512 \times 512$ patch from both optical and SAR images, containing the optical pixel coordinate (for training the SAR detector) or containing the SAR pixel coordinate (for training the optical detector). Figure 2 shows a pair of $512 \times 512$ tiles extracted from optical (left) and SAR (right), with a manually labeled keypoint highlighted.

**Heatmap Label Generation** From each tile pair, we create a single-channel 2D Gaussian heatmap centered on the corresponding keypoint location in the target modality. Each heatmap occupies the same $512 \times 512$ grid, producing a floating-point label map where values peak at 1.0 at the ground-truth keypoint and decay toward 0 at distant pixels. These heatmaps serve as regression targets for U-Net–style keypoint detection networks, compelling the model to predict a continuous keypoint likelihood rather than a discrete coordinate.

**Normalization & Augmentation** Optical tiles are normalized per channel to zero mean and unit variance, using training-set statistics computed over all extracted $512 \times 512$ crops. SAR tiles (single-channel) are likewise normalized. Basic data augmentation (horizontal/vertical flips, 90° rotations) is applied on the fly during training to improve robustness to orientation changes.

**Dataset Splitting** The full tile dataset (approximately 400 pairs) is randomly split 80% / 20% into training/validation and test sets, ensuring that no tile from the same AOI appears in both sets.

## 4. Methods

There are multiple challenges associated with satellite coregistration. Cross-modal mismatch motivates a network that can learn modality-invariant features, since SAR and EO imagery exhibit distinct contrast mechanisms, noise characteristics, and spatial distortions. Secondly, our dataset is small (400 tile pairs), suggesting we avoid extremely large models like Vision Transformers. Finally, our methodology must remain scalable, allowing us to incorporate additional SAR–EO pairs without a complete overhaul of the pipeline.

In our approach, we experiment with three model families: 1) Vanilla U-Net, 2) MAResU-Net, 3) MAResU-Net variants with cross-attention.

### 4.1. Input, Output, and Loss

Each training example consists of a pair of coarsely georeferenced tiles,

$$I_{\text{opt}} \in \mathbb{R}^{3\times 512\times 512}, \quad I_{\text{SAR}} \in \mathbb{R}^{1\times 512\times 512}.$$

We concatenate them into a single 4-channel tensor:

$$X = [\, I_{\text{opt}};\, I_{\text{SAR}} \,] \;\in\; \mathbb{R}^{4\times 512\times 512}.$$

The network predicts a heatmap

$$\hat{H} \;=\; f_\theta(X) \;\in\; \mathbb{R}^{512\times 512},$$

whose peak indicates the predicted keypoint location. Each ground-truth keypoint $(x^\star, y^\star)$ is encoded as a Gaussian heatmap

$$H^\star(x,y) \;=\; \exp\!\Bigl(-\tfrac{\|\,(x,y)-(x^\star,y^\star)\|_2^2}{2\sigma^2}\Bigr), \quad \sigma = 1 \text{ px}.$$

Training minimizes the mean squared error between $\hat{H}$ and $H^\star$:

$$\mathcal{L}(\hat{H},\, H^\star) \;=\; \frac{1}{512^2}\sum_{x=1}^{512}\sum_{y=1}^{512}\bigl(\hat{H}(x,y) - H^\star(x,y)\bigr)^2.$$

### 4.2. Vanilla and Residual U-Nets

Because our dataset is small (400 tile pairs), Vision Transformers are unlikely to outperform. We therefore use U-Nets for coregistration. A U-Net's encoder downsamples via convolution and pooling to capture context, while its decoder upsamples and fuses encoder features through skip connections to produce a pixel-wise heatmap [12]. Extending beyond Vanilla implementations, Residual U0Nets (ResU-Nets) improve feature extraction and training stability [7]. These ResU-Nets are implemented in different forms throughout this report.

### 4.3. Baseline Architectures

The SpaceNet 9 Challenge sponsors supplied reference implementations for several baselines: KeypointArchitecture, SiameseU-Net, and MAResU-Net [6]. KeypointArchitecture is a U-Net–based baseline that concatenates SAR and EO channels and regresses a single-channel heatmap [13]. SiameseU-Net uses two parallel encoder–decoder streams—one for SAR, one for EO—merging their feature maps in the decoder, which encourages modality-specific feature learning before fusion [5]. MAResU-Net extends ResU-Net by inserting multi-scale attention modules at each skip connection: spatial attention captures long-range context within a modality, while channel attention highlights complementary features shared between SAR and EO [9]. These baselines provide a spectrum—from simple U-Net to dual-stream fusion and attention-augmented models—against which we benchmark and refine our MAResU-Net modifications.

We used the sponsor's reference code as a foundation, evaluating each baseline on our dataset before extending it. Beyond running their implementations, we integrated cross-validation, richer data augmentation, and an improved tile-extraction pipeline. Although these models gave us a head start, most of the work—particularly the MAResU-Net cross-attention modifications—was developed from scratch.

### 4.4. Multistage Attention ResU-Net (MAResU-Net) [9]

After establishing baseline results with simpler U-Net variants, we adopted MAResU-Net because it explicitly addresses two coregistration challenges: capturing long-range context across high-resolution tiles (512×512) and maintaining efficiency on our limited dataset ( 400 scenes). Vanilla U-Net skip connections relay only local features, while full non-local attention $\mathrm{softmax}(QK^\top)V$ over $N = H \times W$ pixels scales as $\mathcal{O}(N^2)$, which is prohibitive for $N = 512^2$.

MAResU-Net uses a *Linear Attention Mechanism (LAM)* to approximate the softmax kernel under unit-norm queries and keys $\|q_i\| = \|k_j\| = 1$. Instead of computing

$$\bigl[\mathrm{softmax}(QK^\top)V\bigr]_i = \sum_{j=1}^{N} \frac{e^{\,q_i^\top k_j}}{\sum_m e^{\,q_i^\top k_m}}\, v_j,$$

we apply a first-order Taylor expansion $\exp(q_i^\top k_j) \approx 1 + q_i^\top k_j$, yielding

$$D(Q,K,V)_i = \frac{\sum_j v_j \;+\; \bigl(q_i^\top \sum_j k_j\, v_j^\top\bigr)}{N + q_i^\top \sum_j k_j},$$

which can be computed in $\mathcal{O}(N)$ time by precomputing $\sum_j k_j$ and $\sum_j k_j\, v_j^\top$ [9, Eq. (12)]. This captures global feature interactions without quadratic cost [9].

Each attention block also includes: - **Channel Attention:** $\mathrm{softmax}(XX^\top)\, X$ over $C$ channels (cost $\mathcal{O}(C^2 N)$, but $C \ll N$). - **Spatial LAM:** Project $X \in \mathbb{R}^{C\times H\times W}$ into queries $Q$, keys $K$, and values $V \in \mathbb{R}^{N\times d}$ via 1×1 convolutions, then apply $D(Q,K,V)$.

The summed channel- and spatial-attention outputs pass through a final 1×1 convolution, producing a refined feature map $\widetilde{F} \in \mathbb{R}^{C\times H\times W}$.

MAResU-Net's encoder is ResNet-34, which extracts multiscale features $\{F_1, F_2, F_3, F_4\}$ at resolutions $\{1/2, 1/4, 1/8, 1/16\}$. Each $F_k$ is refined to $\widetilde{F}_k$ via an attention block. In the decoder, $\widetilde{F}_4$ is upsampled and concatenated with $\widetilde{F}_3$, followed by two 3×3 convolutions; this process repeats until a full-resolution keypoint heatmap is reconstructed. By inserting LAM at multiple scales, MAResU-Net fuses global context and local detail, making it well-suited for SAR–EO registration [9].

### 4.5. Improving on MAResU-Net

With MAResU-Net as a foundation, we explored architectural variants to improve performance.

**ResU-Net 18/34/50** We varied the ResNet backbone depth within ResU-Net. ResU-Net-18 uses ResNet-18's four stages (18 conv layers), reducing parameters (approx. 11M) and mitigating overfitting [7]. ResU-Net-34 employs ResNet-34's stages (34 layers, approx. 21M params), balancing expressivity and generalization[7]. ResU-Net-50 uses ResNet-50's bottleneck blocks (50 layers, approx. 34 M params), providing richer features at the expense of overfitting and longer training [7]. Empirically, ResU-Net-34 achieved the best trade-off between accuracy and stability [7].

**Cross-Modal Fusion (MAResU-Net + CrossAttn)** In addition, we focused on building a model that could fuse the two modalities with cross-attention. Our CrossAttn variants enable SAR and EO features to inform each other at every scale, rather than processing them in isolation. At each encoder stage, we combine the SAR and EO feature channels into a single representation. We then separate that representation back into SAR and EO streams and let each stream "attend" to the other—so SAR features learn which EO patterns are most relevant, and vice versa. The two attended outputs are merged and passed forward, allowing the network to discover shared structures (such as edges or corners) that appear in both modalities. This cross-modal attention encourages the model to build representations that bridge the gap between SAR's backscatter patterns and EO's visual cues, improving keypoint matching compared to attending within each modality alone.

### 4.6. Implementation Details & Training

Model code draws from the following libraries: `torch` (core PyTorch[11]), `torch.nn` (layers such as `Module`, `Conv2d`, `Softmax`, `Parameter`), `torchvision.models` (pretrained ResNet backbones[1]), `torch.nn.functional` (aliased as `F` for activations and pooling). Data loading and tile extraction were adapted from the SpaceNet 9 repository (TensorFlow and PyTorch). We reused the sponsor's KeypointArchitecture and SiameseU-Net code with minor edits. MAResU-Net's attention blocks were reimplemented in PyTorch—following [9]. CrossAttn code (feature splitting, dual-attention passes, and fusion) was written in PyTorch as part of this project.

## 5. Experiments, Results and Discussion

### 5.1. Evaluation Metrics

We evaluated model performance using both quantitative registration metrics and qualitative assessments. In particular, our primary quantitative metrics were:

**End-Point Error (EPE)**

$$\mathrm{EPE} \;=\; \frac{1}{N}\sum_{i=1}^{N}\bigl\|\mathbf{p}_i^{\mathrm{pred}} - \mathbf{p}_i^{\mathrm{gt}}\bigr\|_2, \tag{1}$$

where $\mathbf{p}_i^{\mathrm{pred}}$ is the predicted keypoint location (in pixel coordinates) and $\mathbf{p}_i^{\mathrm{gt}}$ is the ground-truth location, and $N$ is the number of keypoints in the test set [15].

**Percentage of Correct Keypoints (PCK) at Threshold $t$**

$$\mathrm{PCK}(t) \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\bigl(\|\mathbf{p}_i^{\mathrm{pred}} - \mathbf{p}_i^{\mathrm{gt}}\|_2 < t\bigr), \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function. We report PCK at $t = 10$ pixels (PCK(10)) and $t = 20$ pixels (PCK(20)).

### 5.2. Overfitting Analysis & Mitigation

Overfitting was a significant concern given the limited number of unique AOIs (only three training scenes). Although each tile was large (512×512 pixels), the risk of memorizing scene-specific textures remained high. We implemented several strategies to reduce overfitting:

- **Data Augmentation:** We applied random flips and 90° rotations to each tile, ensuring the model learned rotationally invariant keypoint features rather than scene-specific patterns.
- **Validation-Based Checkpointing:** During training, we regularly evaluated performance on a held-out set of validation tiles. Whenever validation metrics improved, we saved a "best model" checkpoint, preventing later epochs from overfitting to the training tiles.

## 5.3. Initial Trials

To identify the most promising architecture, we trained each of the three base implementations—KeyPointArchitecture, SiameseUNet, and MAResU-Net—for 50 epochs under seven learning rates: $1\times10^{-2}$, $5\times10^{-3}$, $1\times10^{-3}$, $5\times10^{-4}$, $1\times10^{-4}$, $5\times10^{-5}$, and $1 \times 10^{-5}$. All models used the same fixed train/validation split, mean-squared error (MSE) on predicted heatmaps as the loss function, and reported tie-point EPE (in pixels) on the validation set after 50 epochs. Data augmentation and cross-validation were not applied at this stage. As shown in Figure 3, MAResU-Net consistently outperformed the other baselines, achieving end-point errors below 80 px at a learning rate of $5 \times 10^{-5}$, compared to around 110 px for KeyPointArchitecture and SiameseUNet_Pixelwise. These preliminary results motivated us to refine the MAResU-Net architecture.

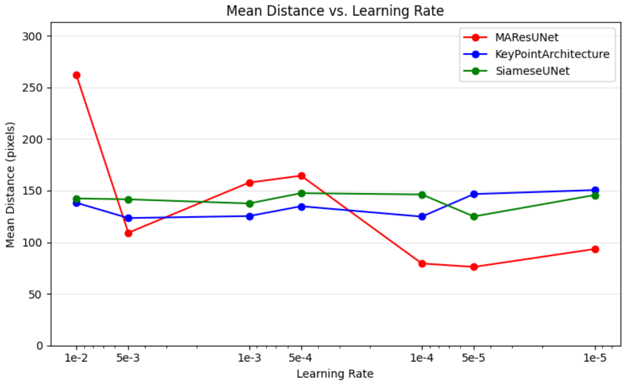

Figure 3. Validation EPE (pixels) after 50 epochs for each baseline at various learning rates. MAResU-Net (red) shows superior performance, especially at lower learning rates.

## 5.4. MAResU-Net Variants

Having selected MAResU-Net as our foundation, we compared three encoder backbone depths—ResUNet-18, ResUNet-34, and ResUNet-50—under the same training settings. Each model was trained for 50 epochs using AdamW, a batch size of 4, and learning rates swept in $\{5 \times 10^{-4}, 1 \times 10^{-4}, 5 \times 10^{-5}\}$. Figure 4 plots validation EPE versus learning rate for each variant. ResUNet-34 achieved the lowest RMSE at $1\times10^{-4}$, balancing representational capacity and generalization.

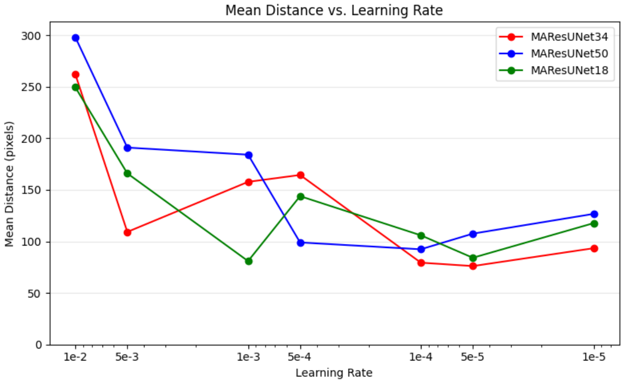

Figure 4. Validation EPE (pixels) after 50 epochs for MAResU-Net variants (ResUNet-18, ResUNet-34, ResUNet-50) across three learning rates. ResUNet-34 (red) performs best at $5 \times 10^{-5}$.

## 5.5. Cross-Attention Variants

Next, we evaluated cross-attention extensions of MAResU-Net. We implemented CrossAttn_v3 variants based on ResUNet-18, ResUNet-34, and ResUNet-50, each trained for 50 epochs with learning rates in $1 \times 10^{-2}$, $5 \times 10^{-3}$, $1\times10^{-3}$, $5\times10^{-4}$, $1\times10^{-4}$, $5\times10^{-5}$,. All models used the same batch size (4) and optimizer (AdamW). Figure 5 shows that MAResU-Net34_CrossAttn achieved the lowest validation RMSE ($\approx$ 80 px at $5 \times 10^{-5}$), outperforming its non-attention counterpart and other depth variants. This confirmed that cross-modal fusion at each skip connection yields more accurate keypoint predictions.

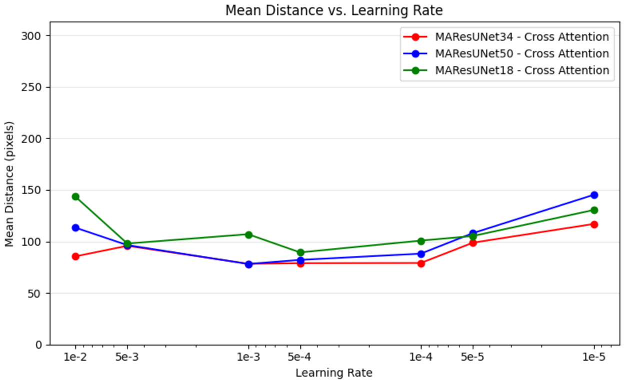

Figure 5. Validation EPE (pixels) after 50 epochs for cross-attention variants. MAResU-Net34_CrossAttn (red) outperforms other depths and non-attention baselines at $5 \times 10^{-5}$.

## 5.6. Final Training

For the final evaluation, we trained the selected architectures—KeyPointArchitecture, SiameseUNet_Pixelwise, MAResU-Net34, and MAResU-Net34_CrossAttn—using five-fold cross-validation and on-the-fly data augmentation (random flips and rotations). Each fold was trained for 50 epochs with the learning rate that produced the best validation results in earlier trials (e.g., $5 \times 10^{-5}$ for MAResU-Net34_CrossAttn). We saved the best model checkpoint per fold and computed test-set metrics (end-point error and PCK) for each. Table 1 and 2 reports the average and standard deviation across folds, demonstrating that MAResU-

Net34_CrossAttn achieves the lowest mean end-point error and highest PCK.

### 5.7. Quantitative Results

Through the final training, the MAResU-Net (with Cross Attention) outperforms the baseline implementations. The EPE across each of the models is shown in Table 1. Additionally, the Percentage of Correct Keypoints is shown in Table 2.

Table 1. Final test-set performance: End-Point Error (EPE) (mean ± std across 5 folds).

| Model | EPE (px) ↓ |
|---|---|
| KeyPointArchitecture | $155.3 \pm 16.4$ |
| SiameseUNet_Pixelwise | $144.6 \pm 4.0$ |
| MAResU-Net34 | $105.3 \pm 8.8$ |
| **MAResU-Net34_CrossAttn** | $\mathbf{88.5 \pm 9.4}$ |

Table 2. Final test-set performance: PCK(10) and PCK(20) (mean ± std across 5 folds).

| Model | PCK(10) ↑ | PCK(20) ↑ |
|---|---|---|
| KeyPointArchitecture | $.49\% \pm .68\%$ | $1.48\% \pm 1.03\%$ |
| SiameseUNet_Pixelwise | $.49\% \pm .68\%$ | $2.47\% \pm 2.31\%$ |
| MAResU-Net34 | $0.74\% \pm 0.68\%$ | $3.70\% \pm 1.51\%$ |
| **MAResU-Net34_CrossAttn** | $\mathbf{2.22\% \pm 1.83\%}$ | $\mathbf{4.69\% \pm 1.35\%}$ |

### 5.8. Qualitative Results

Qualitative evaluation revealed that model performance varies significantly across different landscape types. In urban regions—dense with features such as building corners, roads, and other man-made structures—both SAR and EO imagery contain strong, complementary cues, and the model often achieves sub-pixel or single-pixel accuracy. Conversely, in flat or sparsely textured areas (e.g., deserts, agricultural fields), SAR backscatter offers few distinctive reflections, leading to large alignment errors.

### 5.9. Discussion

The MAResU-Net with cross-attention achieved a 43% reduction in error compared to the KeyPointArchitecture and a significant improvement over the baseline MAResU-Net. The following sections analyze its strengths and shortcomings.

**Failure Case Analysis:** In regions with minimal texture, SAR imagery struggles to provide reliable keypoint cues, since SAR relies on surface geometry to generate strong reflections. Figure 6 shows an outskirt tile with very low contrast: the EO image clearly displays faint road lines, whereas the SAR image is nearly featureless. In these cases, our model's predicted keypoint can be off by up to $\sim 300$ pixels. To improve performance in such settings, future work might involve annotating even minor SAR features (e.g., subtle elevation changes or low-contrast edges).

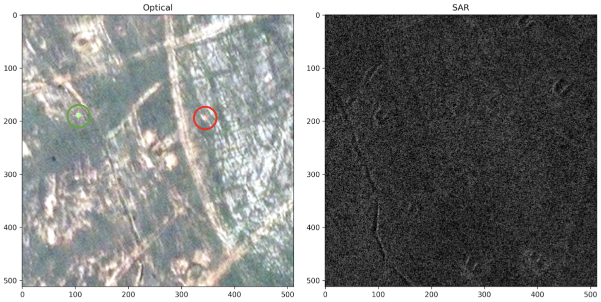

Figure 6. Failure case in a flat region with low contrast. EO (left) shows faint road lines, but SAR (right) lacks distinguishable features, resulting in a large End-Point Error ($\approx$ 300 px).

**Success Case Analysis:** By contrast, in dense urban environments, the model can lock onto sharply defined corners of buildings and road intersections. Figure 7 illustrates an urban tile where the predicted keypoint aligns within 2–3 pixels of the ground truth (EPE $\approx$ 2 px). Such high-precision localization demonstrates that, given sufficient distinctive features in both SAR and EO, the network effectively leverages cross-modal cues. With larger training volumes covering a broader range of urban layouts, we expect similarly high accuracy across more AOIs.

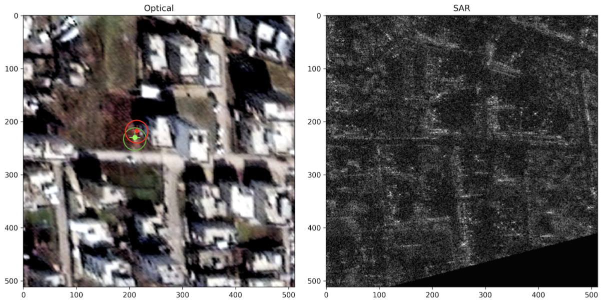

Figure 7. Success case in an urban region. The model matches building corners between EO (left) and SAR (right), achieving an End-Point Error of $\approx$ 2 px.

## 6. Conclusions and Future Work

In this report, we demonstrated that integrating cross-modal attention into Residual U-Net architectures yields substantial gains in keypoint matching accuracy for SAR–EO registration tasks. On our SpaceNet 9 test folds, the MAResU-Net34_CrossAttn model achieved an average end-point error (EPE) of $88.5 \pm 9.4$px, representing a 43% reduction compared to the KeyPointArchitecture baseline

($155.3 \pm 16.4$px). This improvement underscores the value of explicitly allowing SAR features to attend to EO features (and vice versa) at multiple scales, which helps the network learn shared structural cues that are otherwise difficult to capture in single-stream or within-modality attention models.

Despite these gains, several failure modes remain. Cross-attention models still struggle in large, featureless regions, where both SAR and EO modalities provide limited distinctive cues. In such regions, the network's heatmap predictions can become ambiguous, leading to large localization errors. Moreover, because our training set comprises only 400 tile pairs from three AOIs, the model occasionally overfits to sensor-specific noise patterns or scene-specific terrain features. In future work, incorporating additional scenes from varied environments and landscapes would help. We also expect that deeper architectures (e.g., ResUNet-50) will benefit more from larger datasets, as their increased capacity can then be fully exploited without overfitting.

## 7. Contributions and Acknowledgments

This project was not a part of any research outside of CS231N. This paper drew off two code repositories. The SpaceNet 9 Challenge organizers provided starter code. That starter code is only available to challenge participants. The original code will been provided in a folder labeled ”Challenge-Provided Resources.” Additionally, this starter code drew from the MAResUNet Github repository. That repository can be found here.

## References

[1] Torchvision: Datasets, transforms, and models for computer vision. `https://github.com/pytorch/vision`, 2016. Accessed: 2025-06-04.

[2] H. Chen, Z. Qi, and Z. Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:5607514, 2022.

[3] L. Chen, S. Wang, and Y. Zhou. Isift: Iterative sift and ransac for cross-modal image registration. In *International Conference on Pattern Recognition (ICPR)*, pages 678–685, 2021.

[4] A. Dahal, S. A. Murad, and N. Rahimi. Heuristical comparison of vision transformers against convolutional neural networks for semantic segmentation on remote sensing imagery. *IEEE Sensors Journal*, 25(10):17364–17372, 2025.

[5] R. C. Daudt, B. Le Saux, and A. Boulch. Fully convolutional siamese networks for change detection. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, pages 4273–4278, 2020.

[6] R. Hansch, M. Schmitt, J. Shermeyer, M. Bosch, R. Hinsdale, and D. Hoover. Introducing spacenet 9 – cross-modal satellite imagery registration for natural disaster responses. In *Proceedings of IGARSS 2024*. IEEE, 2024.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[8] U. M. Inc. Umbra sar imagery dataset for spacenet 9. `https://www.umbra.com/spacenet9`, 2020.

[9] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang. Multistage attention resu-net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 2021.

[10] X. Ma, X. Zhang, M.-O. Pun, and M. Liu. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:5403215, 2024.

[11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[12] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[13] S. C. Team. Keypointarchitecture: Baseline keypoint detection for cross-modal registration, 2025. Accessed: 2025-05-15.

[14] M. Technologies. Maxar open data program: Earthquake relief data. `https://www.maxar.com/open-data`, 2020.

[15] V. J. Traver and R. Paredes. Study of convolutional neural networks for global parametric motion estimation on log-polar imagery. *IEEE Access*, PP:1–1, 08 2020.

[16] P. Wang, Y. Liu, X. Liang, D. Zhu, X. Gong, Y. Ye, H. F. Lee, and B. Huang. Cirsm-net: A cyclic registration network for sar and optical images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:5610619, 2025.

[17] Y. Xiao, T. Xu, X. Yu, Y. Fang, and J. Li. A lightweight fusion strategy with enhanced interlayer feature correlation for small object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:4708011, 2024.

[18] H. Yan, A. Ma, and Y. Zhong. Progressive symmetric registration for multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 63:5600317, 2025.

[19] J. Yu, C. Zhang, and X. Li. Roadcross: Robust feature matching at road intersections for sar–optical registration. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171:45–58, 2021.

[20] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du. Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:5605415, 2023.