Supplementary Material for EpiMask: Leveraging Epipolar Distance Based Masks in Cross-Attention for Satellite Image Matching

Rahul Deshmukh Aditya Chauhan Avinash Kak

deshmuk5@purdue.edu chauha35@purdue.edu kak@purdue.edu

Purdue University, West Lafayette

1. Overview

In this supplementary material, we provide additional details related to EpiMask. In Sec. 2, we describe the model architecture and hyperparameters. In Sec. 3, we include an additional ablation study that could not be presented in the main manuscript due to page limitations. In Sec. 4, we visualize the model's self-attention, cross-attention, and confidence matrices. In Sec. 5, we present comprehensive quantitative and qualitative results, along with training and validation losses and metrics. Finally, in Sec. 6, we provide a fine-grained analysis of all ablation studies.

2. Architectural Details

2.1. Encoder Decoder

Our encoder is based on the Swin Transformer foundation model trained on high-resolution aerial imagery from the Satlas Pretrain dataset [1]. The encoder produces multi-scale feature maps at $1/4$, $1/8$, $1/16$, and $1/32$ of the input resolution, with channel dimensions 128, 256, 512, and 1024, respectively as shown in Fig. 1.

During training, the encoder is kept frozen and adapted using LoRA [3] by inserting low-rank adapters to all linear layers of self-attention and reduction modules of the Swin Transformer blocks. We experimented with two configurations, LoRA-16 and LoRA-32. The configuration details are summarized in Table 1.

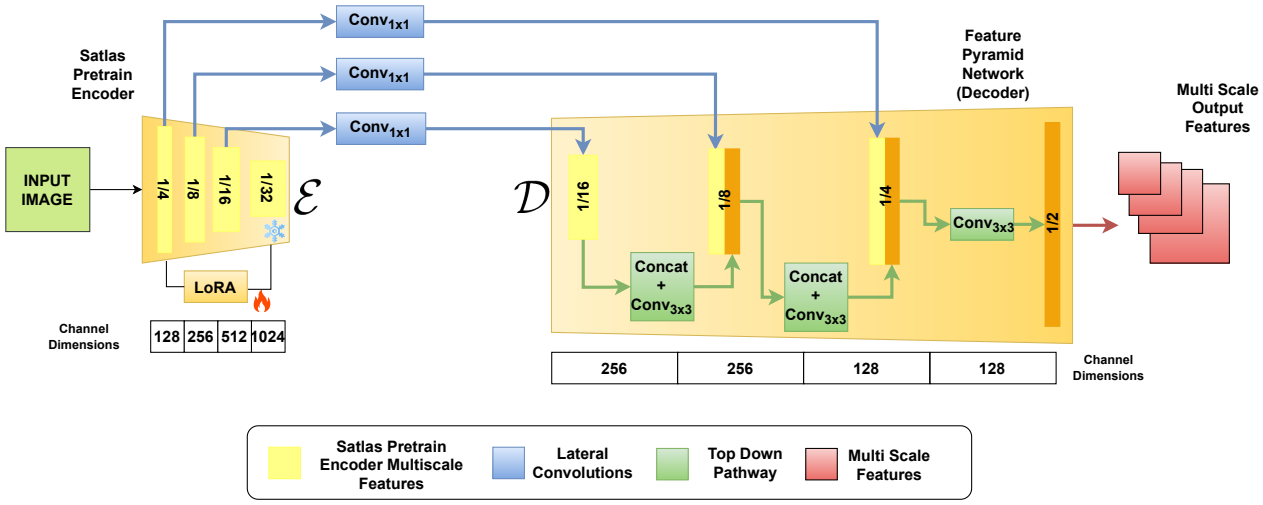

Figure 1. Feature extractor architecture used in EpiMask-HR

As shown in Fig. 1, the decoder is a Feature Pyramid Network (FPN) [6]. We take Swin feature maps at $1/32$, $1/16$, $1/8$,

LoRA Config	Rank (r)	Alpha (α)
LoRA-16	16	8
LoRA-32	32	16

Table 1. LoRA configurations used to adapt the Satlas Pretrain encoder.

and $1/4$ with channel dimensions $\{1024, 512, 256, 128\}$, and apply 1×1 lateral convolutions to project them to a channel dimensions of the FPN at that level. A top-down pathway upsamples higher-level features and fuses them with lateral features using *concatenation followed by convolution* skip-fusion which carries out concatenation of features along the channel dimension followed by a 3×3 convolution. This produces FPN outputs at $1/16$, $1/8$, $1/4$, and $1/2$ resolutions with channel dimensions $\{256, 256, 128, 128\}$.

For **EpiMask-HR** variant, we use the $1/4$ and $1/2$ resolution FPN maps with output channel dimensions $\{d_c{=}128, d_f = 128\}$ for (F_c, F_f) respectively. Whereas for the **EpiMask-LR** variant, we use the $1/8$ and $1/2$ resolution FPN maps with output channel dimension $\{d_c{=}256, d_f{=}128\}$ for (F_c, F_f) respectively.

2.2. Transformer

Coarse-Level Masked Attention Transformer: The coarse transformer operates on the coarse encoder-decoder output features (F_c). For transformer layers in EpiMask-HR, we use a model embedding dimension of $d_c = 128$, while for the transformer layers in EpiMask-LR we use $d_c = 256$. In both cases we use $N_h^c = 8$ attention heads. The architecture consists of $N_c = 8$ coarse transformer layers with interleaved linear self-attention and masked cross-attention layers. Each masked cross-attention layer consists of the following sequence of layers: (1) Projection layers for query, key, value; (2) Multi-headed cross-attention with epipolar-distance based masks; (3) Layer-Norm; (4) Feed Forward with two-layer MLP with hidden dimension of $2{*}d_c$ and ReLU activation; (5) Final LayerNorm and residual connections. The input embeddings for the coarse transformer are obtained by flattening the coarse encoder-decoder output feature map (F_c^L, F_c^R) into a sequence of shape $[B,\ p^2/r_c^2,\ d_c]$ and adding 2D sinusoidal positional encodings.

For the cross-attention masking, we adopt a warm-up strategy where no masking is applied during the first $N_m = 5$ epochs, after which a linearly decreasing epipolar band width (b) is applied across the N_c masked attention layers. The mask band width is linearly decreased from $b = p$ to $b = \gamma p$ across the masked attention layers. The EpiMask-HR variants are trained with $p = 336$ and $\gamma = \{0.4, 0.6\}$, whereas the EpiMask-HR variants are trained with $p = 448$ and $\gamma = \{0.4, 0.6\}$.

Finally, the coarse matching confidence matrix is obtained using a dual-softmax operator based matching layer with epipolar-distance based masking. To obtain a set of coarse correspondences, we threshold the confidence matrix with a confidence threshold of $\delta_c = 0.3$.

Fine-Level Transformer: The fine-level transformer module is the same as LoFTR [7] module, which uses a transformer model embedding dimension of $d_f = 128$ with $N_h^f = 8$ heads and a two-layer transformer self-attention followed by cross-attention. Both self-attention and cross-attention use linear attention [4]. The feed-forward network in transformers has hidden dimension $2 * d_f$ and ReLU activation. Fine-level embeddings are extracted from the higher-resolution FPN feature maps (F_f) centered around each coarse match such that for every coarse correspondence, we crop a local $w \times w$ (with $w = 5$) window in the fine feature maps. We also concatenate the corresponding coarse features to the fine features within each window before feeding them into the fine-level transformer.

2.3. Training Hyperparameters

Image-Patch and Batch Size: We follow the SatDepth preprocessing protocol and use $p \times p$ square image patches. For EpiMask-LR we use $p = 448$ patches in order to be consistent with SatDepth [2]. For EpiMask-HR we could not fit a patch size of $p = 448$ on our compute and rather use $p = 336$ patches. During training, we could fit a batch size of $B = 2$ and $B = 1$ per gpu for the EpiMask-LR and EpiMask-HR variants respectively.

Optimizer: We optimize all trainable parameters using the `AdamW` optimizer with weight decay $\lambda_w = 0.1$ and PyTorch-default betas ($\beta_1{=}0.9, \beta_2{=}0.999$).

Learning-Rate and Warm-Up: We use a canonical learning rate of $lr = 8 \times 10^{-3}$ for a reference batch size of B_{ref}=64, and scale it linearly with the actual batch size. For our configuration this yields $lr_{\text{true}} = 5 \times 10^{-4}$. A linear warm-up schedule is applied for the first 30,000 optimizer steps, during which the learning rate increases from a small fraction ($0.1 * lr_{\text{true}}$) to lr_{true}.

Learning-Rate Scheduler: After warm-up we employ a 'MultiStep-LR' scheduler at the epoch level. The learning rate is decayed by $\gamma_{lr} = 0.5$ at epochs $\{8, 12, 16, 20, 24\}$, and remains constant between milestones. This schedule is used for both EpiMask-HR and EpiMask-LR.

Gradient Accumulation and Clipping: We accumulate gradients over $grad\_accum$=8 batches before each optimizer step and apply global gradient clipping with a threshold of $\lambda_c = 0.5$ to stabilize training, particularly in the early stages when the epipolar-aware transformer layers are still adapting.

Model Size and FLOPs: We report the total and trainable number of parameters, along with the FLOPs for both the EpiMask-HR and EpiMask-LR variants, in Tab. 2.

Params / FLOPs	EpiMask-HR	EpiMask-LR
# Total Params (M)	103	107
# Trainable Params (M)	15.3	19.2
FLOPs @ p=336 (GMACs)	96.35	93.21
FLOPs @ p=448 (GMACs)	174.85	165.04

Table 2. Model parameter size and FLOPs

3. Training Strategy Ablation

In the main manuscript, we presented five ablation studies evaluating different components of our model. Here, we include the final ablation study analyzing the impact of our training strategy on overall performance. We compare two approaches: (1) Single-Stage Training and (2) Two-Stage Training. In the single-stage setup, LoRA layers are applied to the pretrained encoder from the beginning, and the entire model (including the decoder and the coarse- and fine-level transformers) is trained jointly from scratch. In contrast, the two-stage setup begins by training the model without LoRA layers to obtain stable weights for the decoder and both transformer modules. In the second stage, we initialize the model with the stage-one weights and introduce learnable LoRA layers into the pretrained encoder for fine-tuning. We present the performance comparison for the two strategies in Fig. 2. The two-stage strategy performs better than the single-stage strategy.

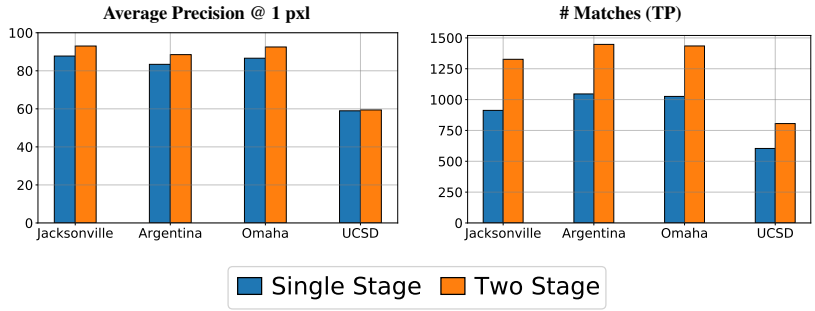

Figure 2. Average precision and number of true-positive matches for all testing AOIs for models trained with single-stage and two-stage training strategies.

4. Understanding EpiMask

To better understand the internal workings of EpiMask, we visualize the self-attention maps (Fig. 4) and masked cross-attention maps (Fig. 5) across all layers and heads of our coarse-level masked attention transformer. We also show the coarse-level matching confidence matrix in Fig. 3.

From Fig. 4, we observe that different attention heads across the self-attention layers specialize in diverse pattern, some attend locally, while others capture long-range dependencies. In the case of masked cross-attention (Fig. 5), earlier layers often do not focus on the true corresponding region. However, by the final layer, five out of eight heads converge to the correct region, indicating progressive refinement. Finally, the matching confidence visualization (Fig. 3) shows that the model assigns the highest matching score (Top-1) to the true correspondence, while the remaining high-confidence candidates (up to Top-20) lie along the epipolar line as reasonable alternatives.

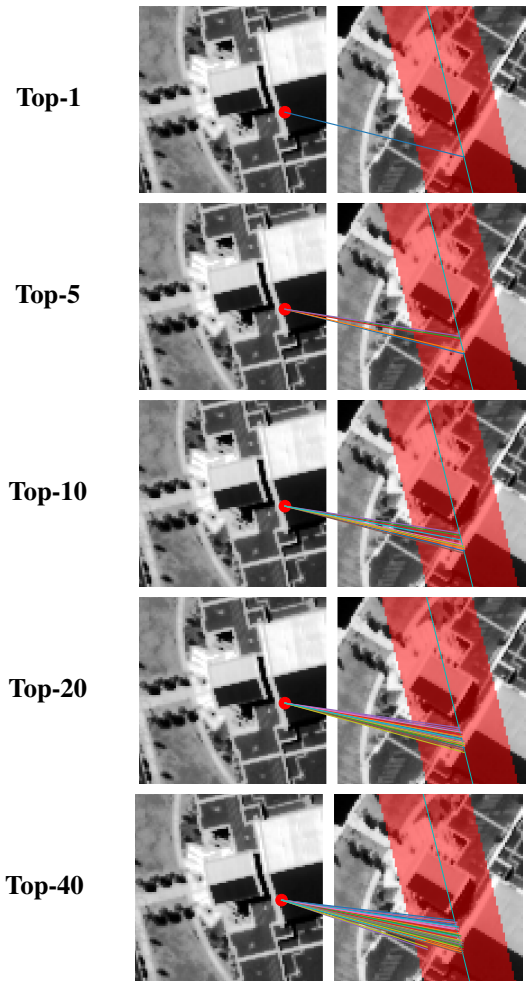

Figure 3. Visualization of Top-K confidence scores in the masked coarse matching module. Shown are Top-K matching points (selected by sorting the confidence scores and taking the K highest) in the right image patch corresponding to the query pixel in the left image patch. In each figure, the query pixel is shown as red dot in the left image patch with corresponding epipolar line (cyan line) and mask (red band) in the right image patch. Top-K matching locations to the query pixel are displayed using lines originating from the query pixel. We observe that the model assigns the highest matching score (Top-1) to the true correspondence, while the remaining high-confidence candidates (up to Top-20) lie along the epipolar line as reasonable alternatives.

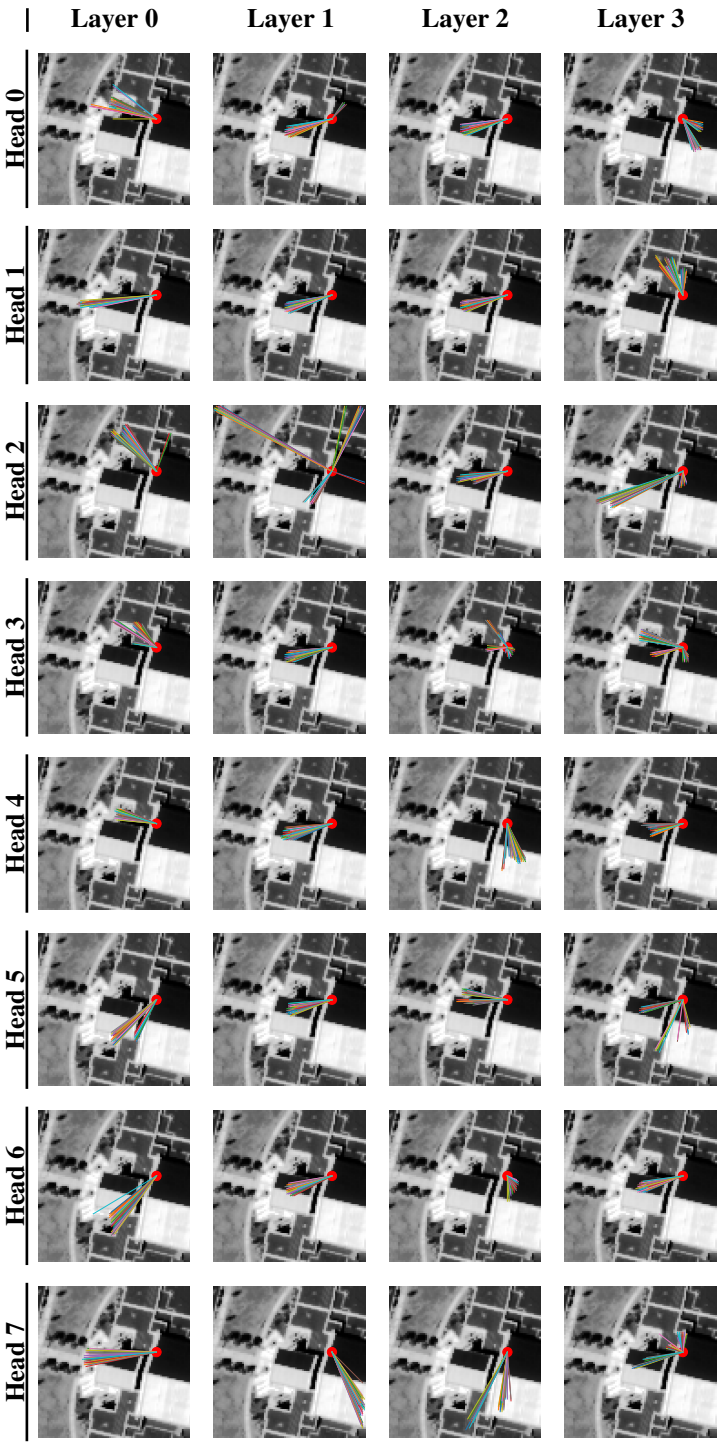

Figure 4. Visualization of self-attention maps in the coarse transformer. Shown are attention maps for each self-attention layer and individual heads within the multi-headed attention. In each figure, the query pixel is shown as red dot and top 40 attention locations are displayed using lines originating from the query pixel. We observe that different attention heads in the self-attention layers exhibit different attentional behaviors: While some attend locally (*e.g*. Layer-0 and Head-0) others capture long-range dependencies (*e.g*. Layer-1 and Head-2).

Figure 5. Visualization of masked-cross-attention maps in the coarse transformer. Shown are attention maps for each masked cross-attention layer and individual heads within the multi-headed attention. In each figure, the query pixel is shown as red dot in the left image patch with corresponding epipolar line (cyan line) and mask (red band) in the right image patch. Top 40 attention locations corresponding to the query pixel are displayed using lines originating from the query pixel. We observe that the earlier layers often do not focus on the true corresponding region. However, by the final layer (Layer-3), five out of eight heads (Heads-0,1,4,5,6) attend to the correct region.

5. Experimental Results

In the main manuscript, due to page limitations we presented quantitative results only for Jacksonville and San Fernando testing AOIs. Here we present comprehensive quantitative aggregated results for each testing AOI in Fig. 6. We also present the quantitative results for 'Simulated-Rotation' experiment for all testing AOIs in Fig. 7. Following SatDepth [2] we present qualitative results for large view-angle and time differences in Figs. 8 and 9 respectively. Finally, we present the plots for training and validation losses and metrics for our four model configurations in Fig. 10.

Jacksonville

	Method	Pose estimation AUC ↑ @5°	@10°	@20°	Precision ↑ @1px	# Matches ↑ (TP)
SatDepth[2]	SIFT + satCAPS [8]	38.49	43.26	50.94	10.67	21
	satDualRC-Net [5]	41.19	47.57	56.31	19.94	40
	satLoFTR [7]	78.48	87.02	92.30	54.87	108
	satMatchFormer [9]	81.37	89.01	93.56	61.96	124
Ours	EpiMask-LR$_{\gamma=0.6}$	89.32	93.67	96.13	62.38	1027
	EpiMask-LR$_{\gamma=0.4}$	89.51	93.69	96.06	61.95	1007
	EpiMask-HR$_{\gamma=0.6}$	91.53	94.81	96.66	81.29	1187
	EpiMask-HR$_{\gamma=0.4}$	**92.66**	**95.57**	**97.14**	**83.32**	**1286**

San Fernando

	Method	Pose estimation AUC ↑ @5°	@10°	@20°	Precision ↑ @1px	# Matches ↑ (TP)
SatDepth[2]	SIFT + satCAPS [8]	36.75	40.09	45.69	7.89	16
	satDualRC-Net [5]	40.57	46.77	55.58	15.88	32
	satLoFTR [7]	53.60	62.96	71.34	42.58	71
	satMatchFormer [9]	54.57	64.15	72.68	39.83	73
Ours	EpiMask-LR$_{\gamma=0.6}$	79.62	87.92	93.11	39.23	134
	EpiMask-LR$_{\gamma=0.4}$	80.38	88.58	93.52	45.60	**148**
	EpiMask-HR$_{\gamma=0.6}$	85.92	91.08	94.17	72.78	56
	EpiMask-HR$_{\gamma=0.4}$	**87.52**	**92.73**	**95.80**	**70.57**	113

Omaha

	Method	Pose estimation AUC ↑ @5°	@10°	@20°	Precision ↑ @1px	# Matches ↑ (TP)
SatDepth[2]	SIFT + satCAPS [8]	69.42	71.77	75.24	10.64	21
	satDualRC-Net [5]	71.10	74.35	78.56	28.31	57
	satLoFTR [7]	82.41	87.36	91.18	46.85	90
	satMatchFormer [9]	81.83	86.65	90.49	44.10	86
Ours	EpiMask-LR$_{\gamma=0.6}$	93.92	96.34	97.87	63.59	1132
	EpiMask-LR$_{\gamma=0.4}$	94.09	96.44	97.91	62.88	1119
	EpiMask-HR$_{\gamma=0.6}$	94.91	96.86	98.09	77.51	1239
	EpiMask-HR$_{\gamma=0.4}$	**95.67**	**97.44**	**98.49**	**80.91**	**1319**

UCSD

	Method	Pose estimation AUC ↑ @5°	@10°	@20°	Precision ↑ @1px	# Matches ↑ (TP)
SatDepth[2]	SIFT + satCAPS [8]	68.68	70.47	73.25	7.86	16
	satDualRC-Net [5]	70.85	73.98	78.06	24.68	49
	satLoFTR [7]	76.82	80.52	83.94	38.00	66
	satMatchFormer [9]	80.10	83.96	87.15	34.75	63
Ours	EpiMask-LR$_{\gamma=0.6}$	91.55	94.02	95.99	40.10	642
	EpiMask-LR$_{\gamma=0.4}$	**93.23**	**95.78**	**97.48**	**47.88**	692
	EpiMask-HR$_{\gamma=0.6}$	93.04	95.12	96.56	47.73	701
	EpiMask-HR$_{\gamma=0.4}$	91.78	93.63	95.20	47.20	**735**

Figure 6. Weighted average of Precision, Pose error, and number of True Positive (TP) matches over all testing image patches for all testing AOIs of SatDepth.

Figure 7. Comparison of all model average precision and number of true-positive matches for all testing AOIs of SatDepth in the simulated rotation experiment. EpiMask consistently achieves the highest precision and detects the largest number of matches w.r.t varying view-angle differences α^v and track-angle differences α^t.

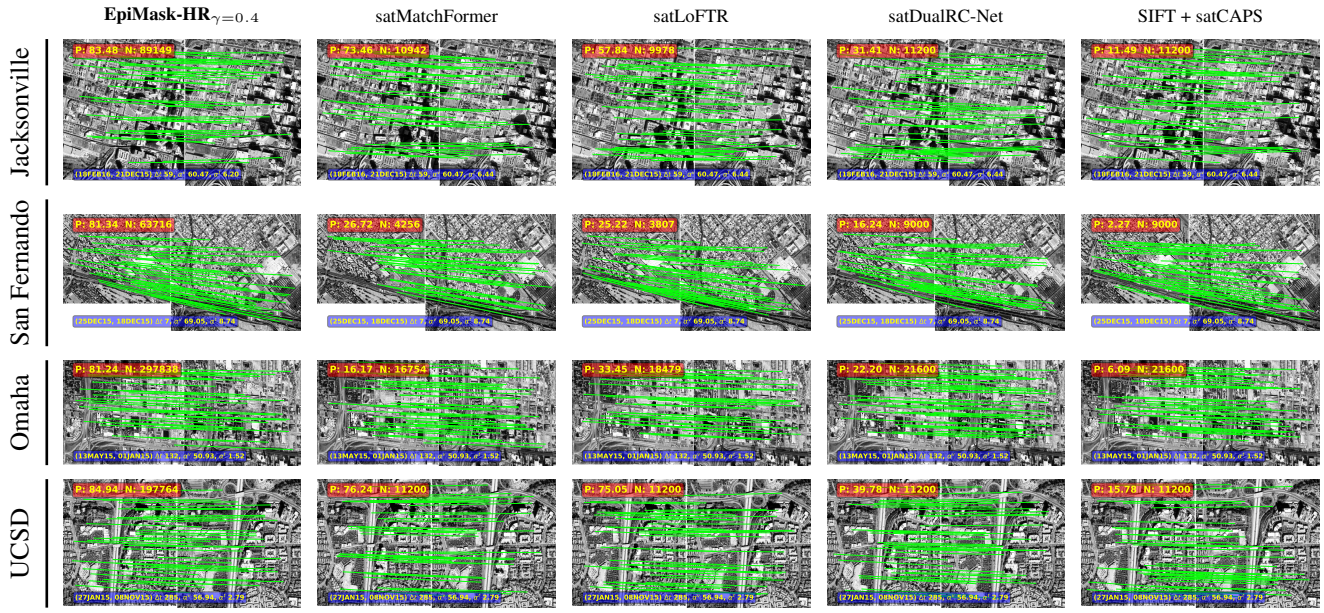

Figure 8. Qualitative comparison of our model results against other models for large view-angle difference (α^v) – our model has the highest precision score. Precision (P) and number of matches (N) are displayed at the top of each plot. Image pair names, time difference (Δt), view-angle difference (α^v), and track-angle difference (α^t) are displayed at the bottom. The green lines depict 40 randomly chosen true matches.

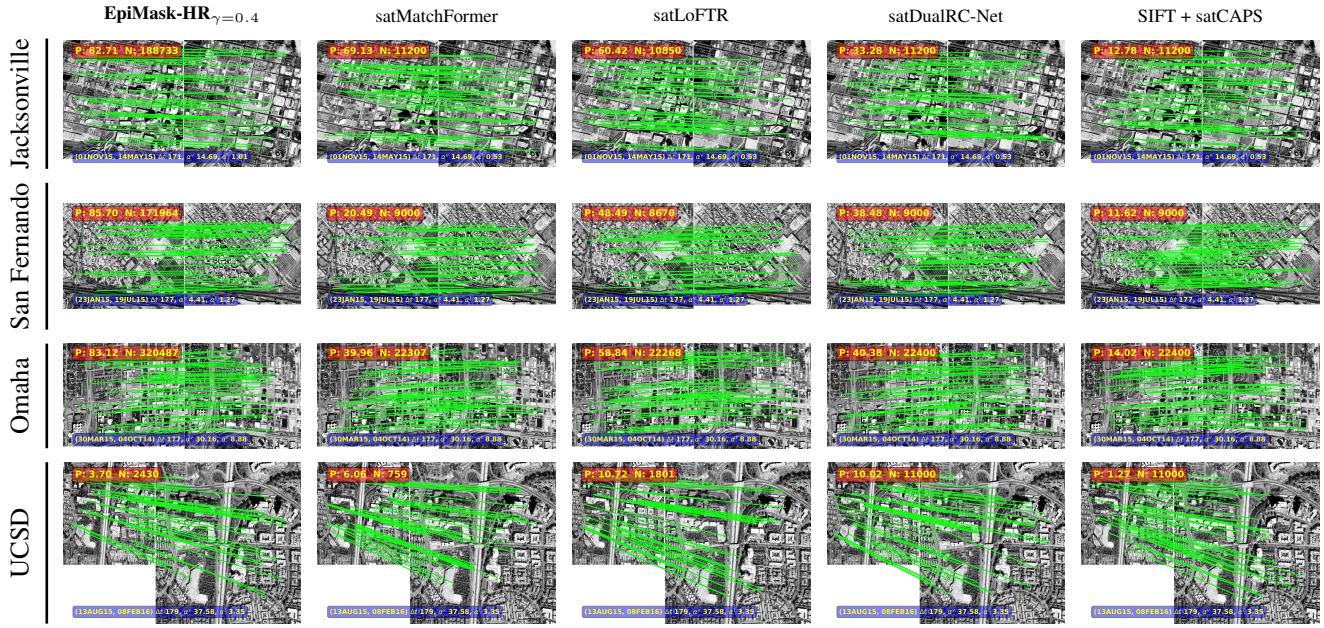

Figure 9. Qualitative comparison of our model results against other models for large time difference (Δt) – our model has the highest precision score. Precision (P) and number of matches (N) are displayed at the top of each plot. Image pair names, time difference (Δt), view-angle difference (α^v), and track-angle difference (α^t) are displayed at the bottom. The green lines depict 40 randomly chosen true matches.

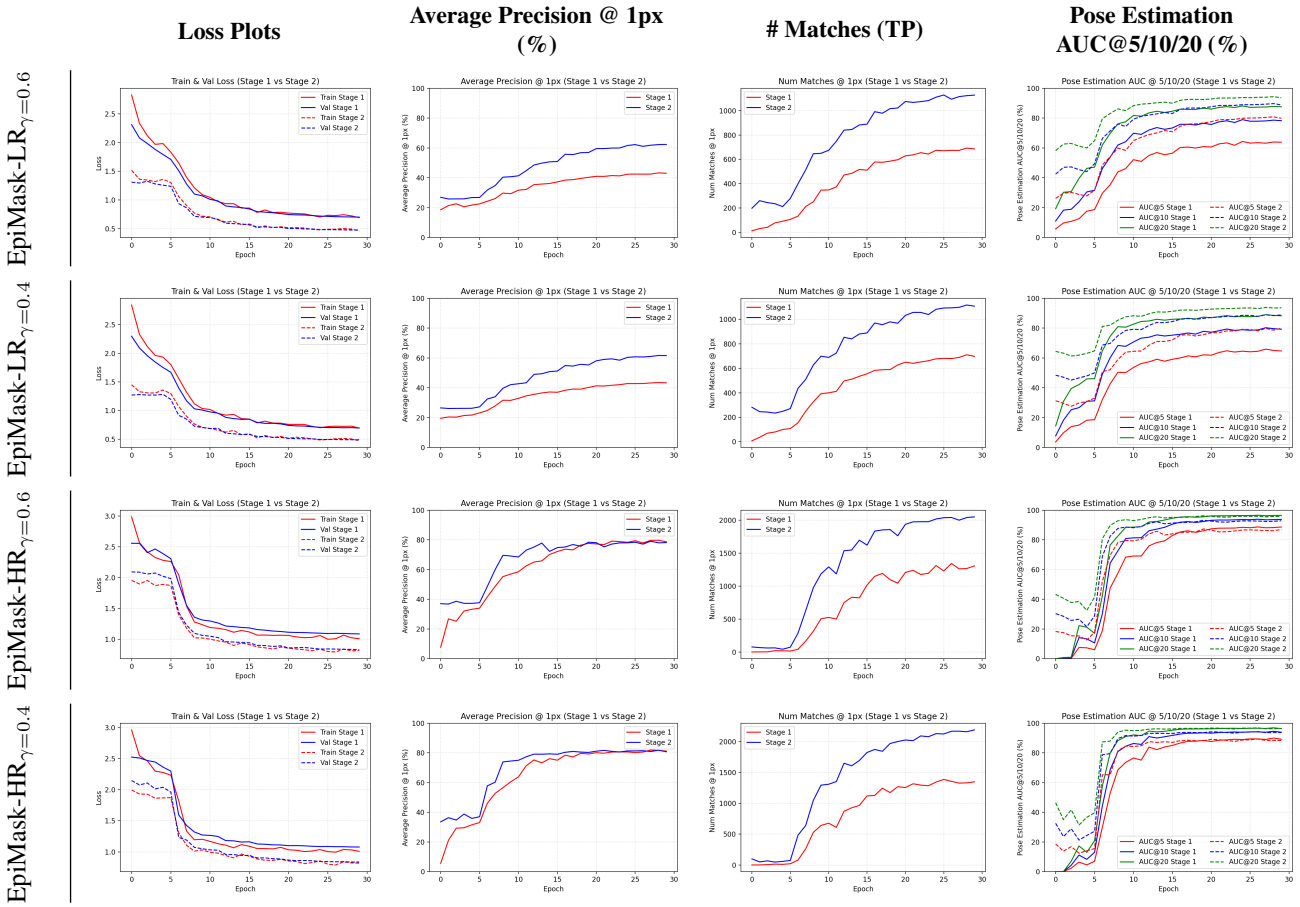

Figure 10. Plots for loss and metrics for our four model configurations. The first column shows the training and validation loss per epoch for both stages of training. The other columns show metrics (Precision, Number of true-positive matches, and pose estimation AUC) per epoch on the validation set for both stages of training.

6. Ablation Studies

In the main manuscript we presented averaged results for each testing AOIs, where the average was over all view and track angles. In the following sub-sections we present a fine-grained analysis for each ablation study w.r.t. different ranges for view angle and track angle differences.

6.1. Resolution Ablation

In the main manuscript we presented results for the resolution ablation for all testing AOIs averaged over all angles. In this section we present a fine-grained analysis over different ranges of view angle (α^v) and track angle differences (α^t) (see Fig. 11). The High-Res configuration performs the best.

Figure 11. Average precision and number of true positive matches for resolution ablation.

6.2. Attention Mask Width Ablation

In the main manuscript we presented results for the attention mask width ablation for all testing AOIs averaged over all angles. In this section we present a fine-grained analysis over different ranges of view angle (α^v) and track angle differences (α^t) (see Fig. 12). As shown in Fig. 12, performance remains largely unchanged, indicating that the model inherently focuses on geometrically consistent regions within the epipolar band. This robustness suggests that fine-grained tuning of mask width is unnecessary. In practice, narrower masks may be preferred for newer satellites with accurate pose metadata, while wider masks can better handle older sensors with noisier estimates of camera pose.

Figure 12. Average precision and number of true positive matches for attention mask width ablation.

6.3. Positional Encoding Ablation

In the main manuscript we presented results for the positional encoding ablation for all testing AOIs averaged over all angles. In this section we present a fine-grained analysis over different ranges of view angle (α^v) and track angle differences (α^t) (see Fig. 13). As shown in Fig. 13, average precision remains similar, but positional encodings consistently increase true-positive matches.

Figure 13. Average precision and number of true positive matches for positional encoding ablation.

6.4. Fine-Tuning Ablation

In the main manuscript we presented results for the fine-tuning ablation for all testing AOIs averaged over all angles. In this section we present a fine-grained analysis over different ranges of view angle (α^v) and track angle differences (α^t) (see Fig. 14). As shown in Fig. 14, LoRA significantly increases the number of true positives, indicating that lightweight fine-tuning effectively adapts the pretrained backbone to our matching task. Increasing the rank from 16 to 32 provides negligible gains, suggesting that a moderate rank of 16 strikes a good balance between parameter efficiency and adaptation quality.

Figure 14. Average precision and number of true positive matches for fine-tuning ablation.

6.5. Feature Extractor Ablation

In the main manuscript we presented results for the feature extractor ablation for all testing AOIs averaged over all angles. In this section we present a fine-grained analysis over different ranges of view angle (α^v) and track angle differences (α^t) (see Fig. 15). As shown in Fig. 15, the concatenation followed by convolution strategy performs better compared to naïve element-wise addition.

Figure 15. Average precision and number of true positive matches for feature extractor ablation.

6.6. Training Strategy Ablation

Earlier in Sec. 3 we presented results for the training strategy ablation for all testing AOIs averaged over all angles. In this section we present a fine-grained analysis over different ranges of view angle (α^v) and track angle differences (α^t) (see Fig. 16). As shown in Fig. 16, the two-stage training strategy performs better than the single stage.

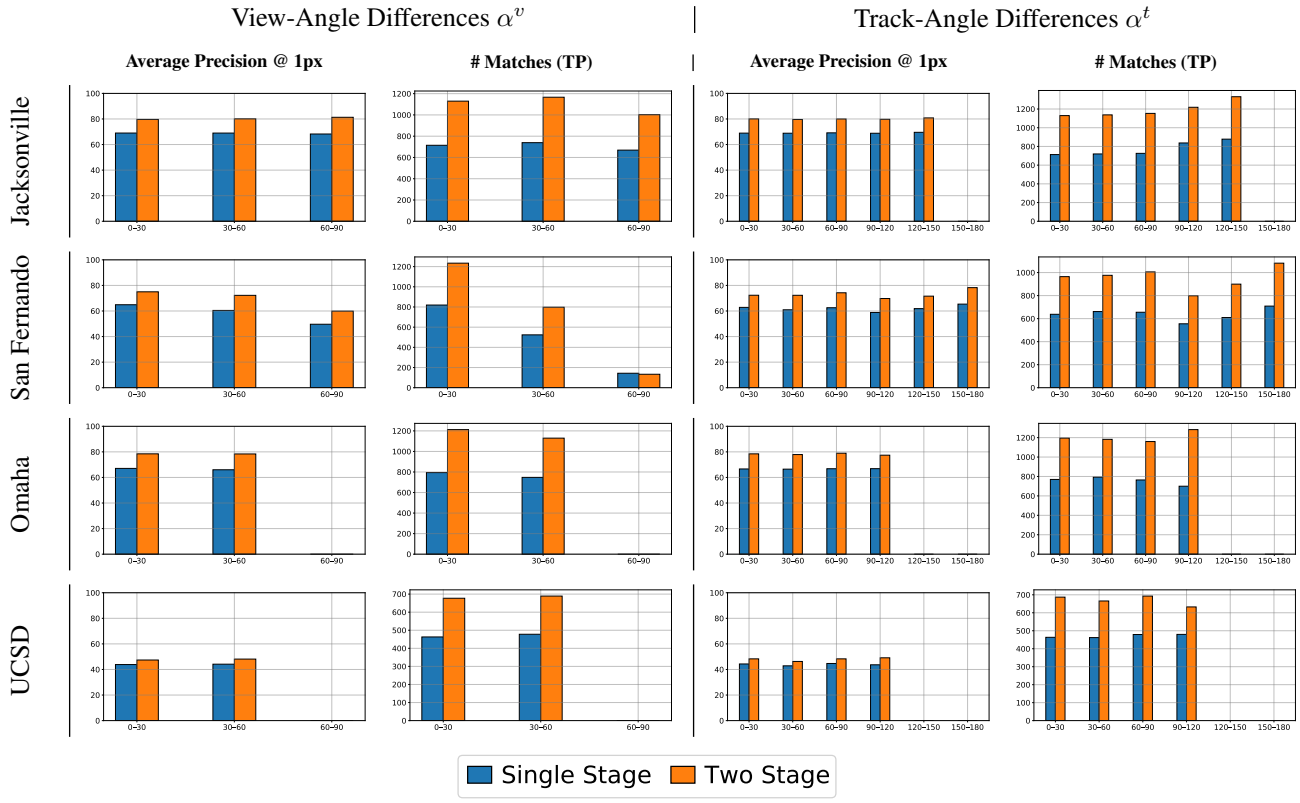

Figure 16. Average precision and number of true positive matches for training strategy ablation.

References

[1] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of Intl. Conf. on Computer Vision (ICCV)*, 2023. 1

[2] Rahul Deshmukh and Avinash C. Kak. SatDepth: A Novel Dataset for Satellite Image Matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 19:894–903, 2026. 2, 7

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proceedings of Intl. Conf. on Learning Representations (ICLR)*, 2022. 1

[4] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. 2020. 2

[5] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-Resolution Correspondence Networks. In *Proceedings of Conf. on Neural Information Processing Systems (NeurIPS)*, 2020. 7

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[7] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 7

[8] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 7

[9] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Proceedings of the Asian Conference on Computer Vision*, 2022. 7