# Supplementary Material:
# A Light Touch Approach to Teaching Transformers Multi-view Geometry

Yash Bhalgat    João F. Henriques    Andrew Zisserman
Visual Geometry Group
University of Oxford
{yashsb,joao,az}@robots.ox.ac.uk

## List of Figures

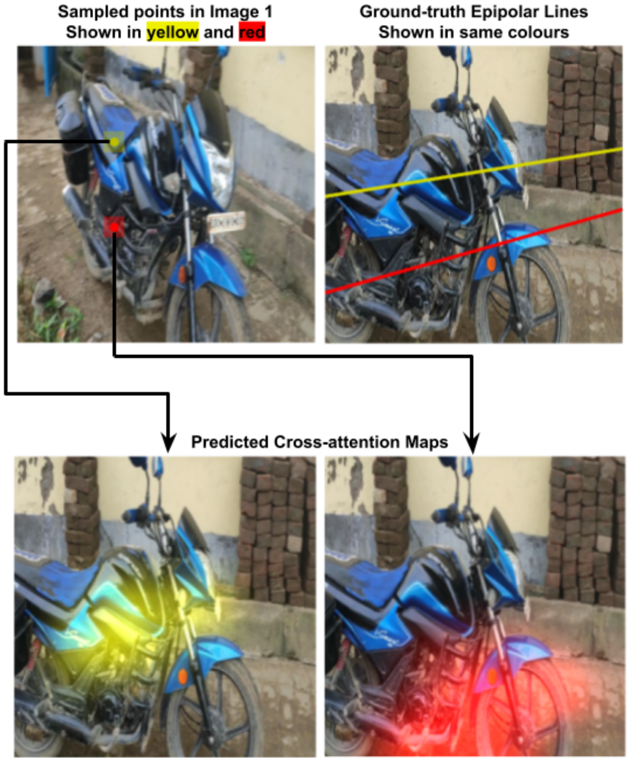

Sampled points in Image 1
Shown in yellow and red
Ground-truth Epipolar Lines
Shown in same colours
Predicted Cross-attention Maps

Figure 1. Cross-attention maps, extracted from Transformer model trained with Max-Epipolar Loss, overlaid on the image. Note: the $14 \times 14$ -sized attention maps are bilinearly upsampled to the size of the image.

## 1. Do cross-attention maps from Max-Epipolar Loss correspond to true matching points?

We overlay the cross-attention maps, extracted from the Transformer [4] model trained with Max-Epipolar Loss ($L_{MaxEPI}$), on the original images to see if the location of highest attention coincides with the position of the actual matching point. Fig. 1 shows one such visual illustration. We can see that the peaks of the attention maps *loosely* coincide with the actual matching points on the corresponding ground-truth epipolar lines.

## 2. Cross-attention Visualization for mismatched image pair

Similar to Fig. 4 in the main paper, we visualize the cross-attention maps for a pair of mismatched images, as shown in Fig. 2. These cross-attention maps are extracted from a RRT [4] model trained with Epipolar Loss ($L_{EPI}$).

## 3. Qualitative examples

In Figures 4-14, we provide a few qualitative results on the CO3D-Retrieve benchmark (Figures 4-8) and the Stanford Online Products [2] dataset (Figures 9-14). We visually compare the top-5 retrievals obtained with the Global Retrieval (R50) model, Reranking Transformer [4] model and a reranking model trained with our Epipolar Loss. We also accompany each example with its corresponding Precision-Recall curve, which provides a more detailed perspective on the retrieval performance.

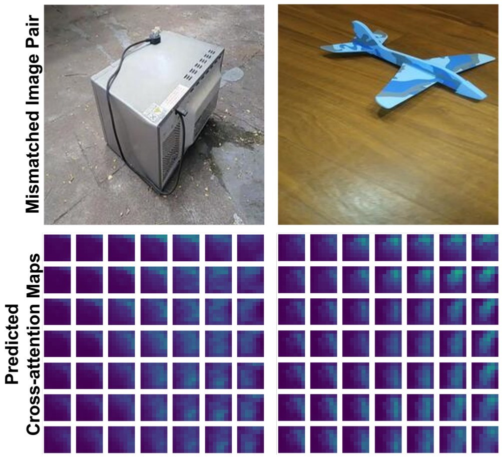


Figure 2. Predicted attention maps for a non-matching test image pair. A valid epipolar geometry does not exist for this pair, hence the model predicts *diffuse* attention maps.

In the CO3D-Retrieve benchmark, the *maximum* number of reference images per query is 4. So, for all the examples shown, the Precision-Recall curve for our method saturates at Precision $= 1.0$ since the top-4 retrievals are correct.

## 4. Visualization of Attention Maps

For the sake of clarity, we describe, with an example, how the cross-attention map predicted by our Transformer model (trained with Epipolar Loss) contains information about the true epipolar geometry between the input image pair. Figure 15 shows such an example, where we select two points in a $7 \times 7$ grid (because the feature map extracted by our backbone is spatially $7 \times 7$) of the first image and show the actual (ground-truth) as well as the predicted epipolar lines in the other image.

## 5. Implementation details CO3D-Retrieve

For experiments with the CO3D-Retrieve benchmark, the global-retrieval-only model (R50 (trained) as described in Sec. 5.1) is trained for 50 epochs with the Adam optimizer and a learning rate that starts at 0.0001 and decays exponentially by a factor of 10 every 20 epochs. The Reranking Transformer head is trained on top of this trained global model, by either freezing or finetuning the global model, with or without the Epipolar Loss. When training without the Epipolar Loss, the model is trained using a SGD optimizer with an initial learning rate of $5 \times 10^{-5}$ decayed exponentially by a factor of 10 over 40 epochs. When training with the Epipolar Loss, the above procedure is fol-

| Dataset Name | Same Category Retrieval | Full Dataset Retrieval |
|---|---|---|
| CO3D-Retrieve | 52.03 | 49.52 |
| SOP [2] | 38.61 | 37.25 |

Table 1. Comparison of mAP computed while ranking only the images from the same category as the query image.

lowed without Epipolar Loss for 20 epochs and the model is trained for an additional 20 epochs with Epipolar Loss with a learning rate of $10^{-6}$. As mentioned in the main paper, the hyperparameters we use for our experiments with SOP [2] are the same as [4], except that we use 40 epochs (instead of 100 in [4]) when training with the Epipolar Loss with a constant learning rate of $10^{-4}$.

## 6. mAP analysis with same category retrieval

We empirically observe that a majority of high-ranked (i.e. top-5) false positives are images from the same category. We conduct an experiment where we compute the mAP while ranking only the images from the same category as the query image and ignoring out-of-class images. As shown in Tab. 1, we get a higher mAP when retrieval is only performed on the same category images. This means there are confusing images from *outside* the categories, albeit a small fraction compared to intra-class.

Further per-category analysis with our method reveals that the *top 5* categories with the highest proportions of intra-class false positives (in descending order) are `banana`, `suitcase`, `laptop`, `keyboard`, `umbrella` in CO3D-Retrieve dataset and `fan`, `cabinet`, `mug`, `coffee_maker`, `kettle` in SOP [2].

## 7. Breakdown of $R@K$ based on fraction of overlapping pixels

The proposed CO3D-Retrieve benchmark includes large variations in viewing angle among images of the same instance. We conduct an analysis to understand how our proposed method performs under a range of pose variations. To do this, we first compute an ”Overlap Score (OS)” for each instance using ground-truth point-clouds available in CO3D-Retrieve. Large pose-variations lead to a low OS. We divide the query-set into 10 bins uniformly between OS=0.2 to 0.8 and compute the $R@1$ for each bin. These limits of OS are chosen because there are very few ($< 1\%$) instance with an OS beyond the $[0.2, 0.8]$ range. Fig. 3 shows $R@1$ for our proposed method and RRT [4] with respect to the OS. We can see that the R@1 of our method drops by $5.5\%$ from highest to lowest OS, while RRT [4] drops by $10.6\%$. In conclusion, our Epipolar Loss is useful in extreme viewpoint changes.

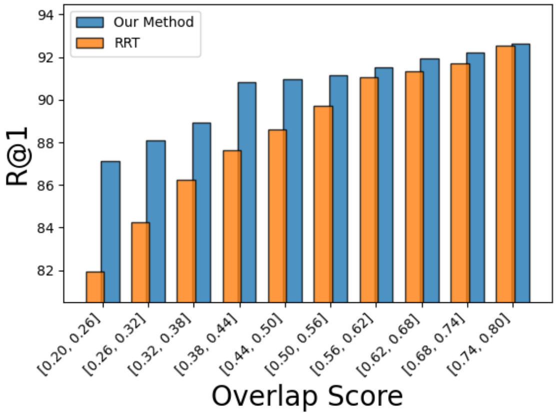


Figure 3. Breakdown of $R@1$ according to the Overlap Score of instances in the CO3D-Retrieve benchmark. The query set is divided into bins based on OS; these bins are shown on the x-axis.

| Model | Local features resolution | $R@1$ | $R@10$ | $R@50$ |
|---|---|---|---|---|
| Original | 7×7 | 90.57 | 97.33 | 98.10 |
| High-res | 14×14 | 90.71 | 97.42 | 98.15 |

Table 2. Comparison with transformer model trained on $448 \times 448$ images. Both models are trained with $L_{EPI}$. “Original” corresponds to the result in Table 1 of main paper.

## 8. High-resolution results

In our experiments throughout the paper, the local features tensor obtained from Resnet50 backbone has a spatial resolution of 7×7. We train another transformer model with $L_{EPI}$ on input images of size $448 \times 448$ so that we obtain $14 \times 14$ local features. By doing this, we can obtain higher resolution ($14 \times 14 \times 14 \times 14$) cross-attention maps, as shown in Fig. 16. Tab. 2 shows the performance achieved by the high resolution transformer model.

## 9. Failure Cases

It is important to look at the cases where our proposed method fails to retrieve good matches and analyze them for further improvement. Figures 17 and 18 show a few such examples for CO3D-Retrieve and SOP [2] respectively. We see that a common failure scenario for our method is when the query image is a close-up of the object (Fig. 17 (c,d) and Fig. 18 (c,d)) or repetitive patterns in objects such as keyboards (Fig. 17 (d)). A critical future direction for our work is to make the model robust to these scenarios.

## 10. Quality of Epipolar Geometry with LoFTR/MAGSAC++ method

During training, when the ground-truth epipolar geometry is not available, we use a *pseudo*-geometry predicted using a pretrained LoFTR [3] model for matching and MAGSAC++ [1] for robust optimization. The quality of the predicted epipolar geometry depends on the quality and number of matches obtained by the LoFTR model. In Figure 19, we show two examples demonstrating the success and failure cases of this method.

## References

[1] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. Magsac++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1304–1312, 2020. 1, 4, 12

[2] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 1, 2, 3, 7, 8, 11

[3] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 1, 4, 12

[4] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12105–12115, 2021. 2, 3

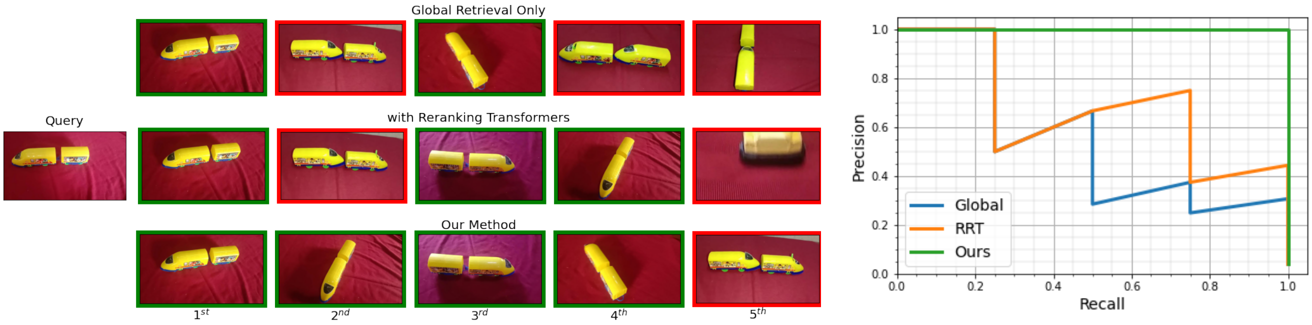


Figure 4. CO3D-Retrieve: Example 1.

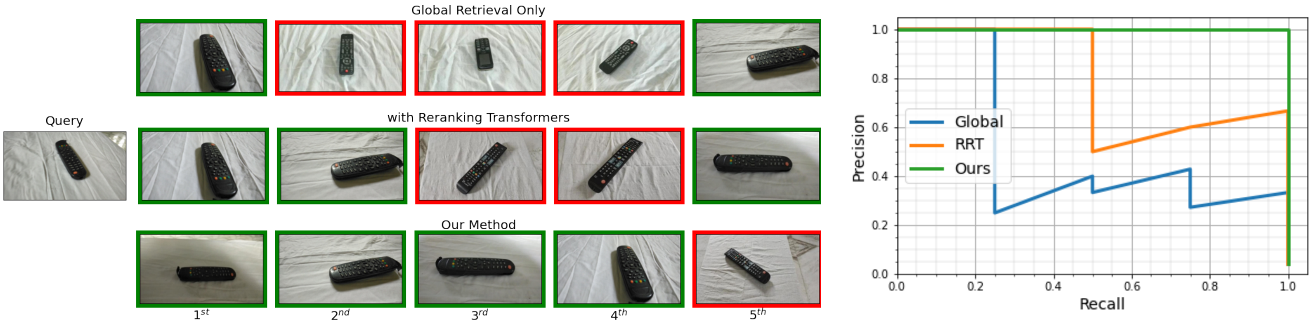


Figure 5. CO3D-Retrieve: Example 2.

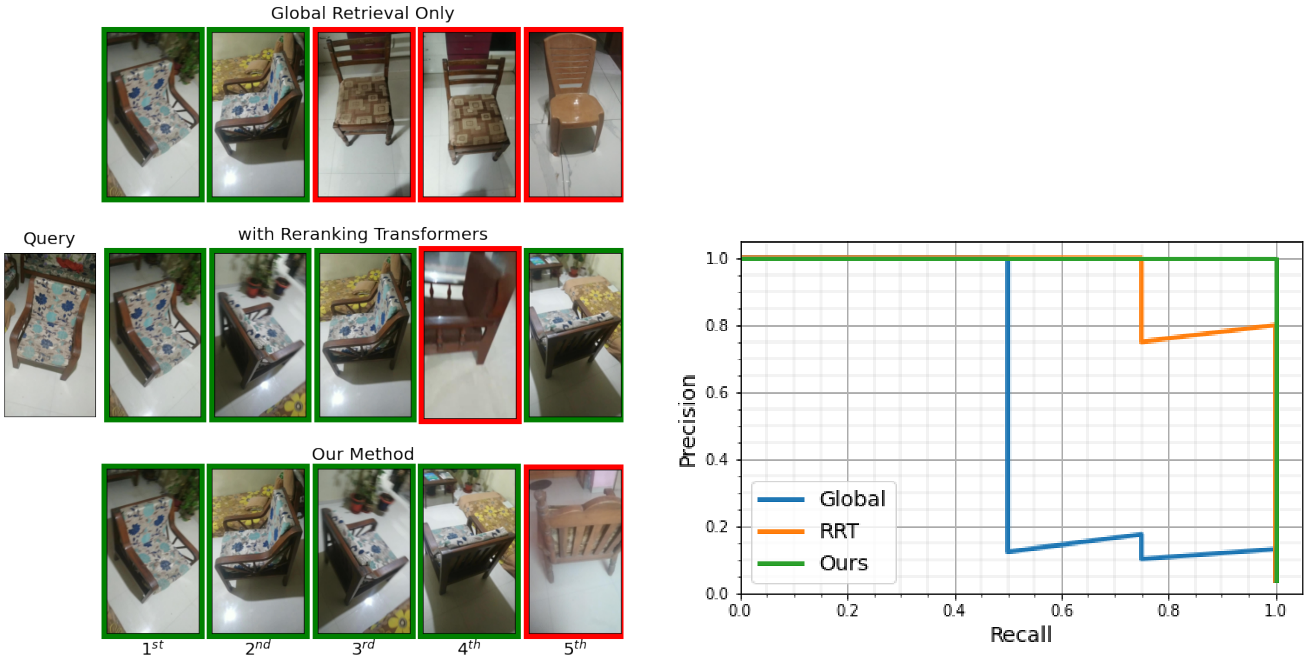


Figure 6. CO3D-Retrieve: Example 3.

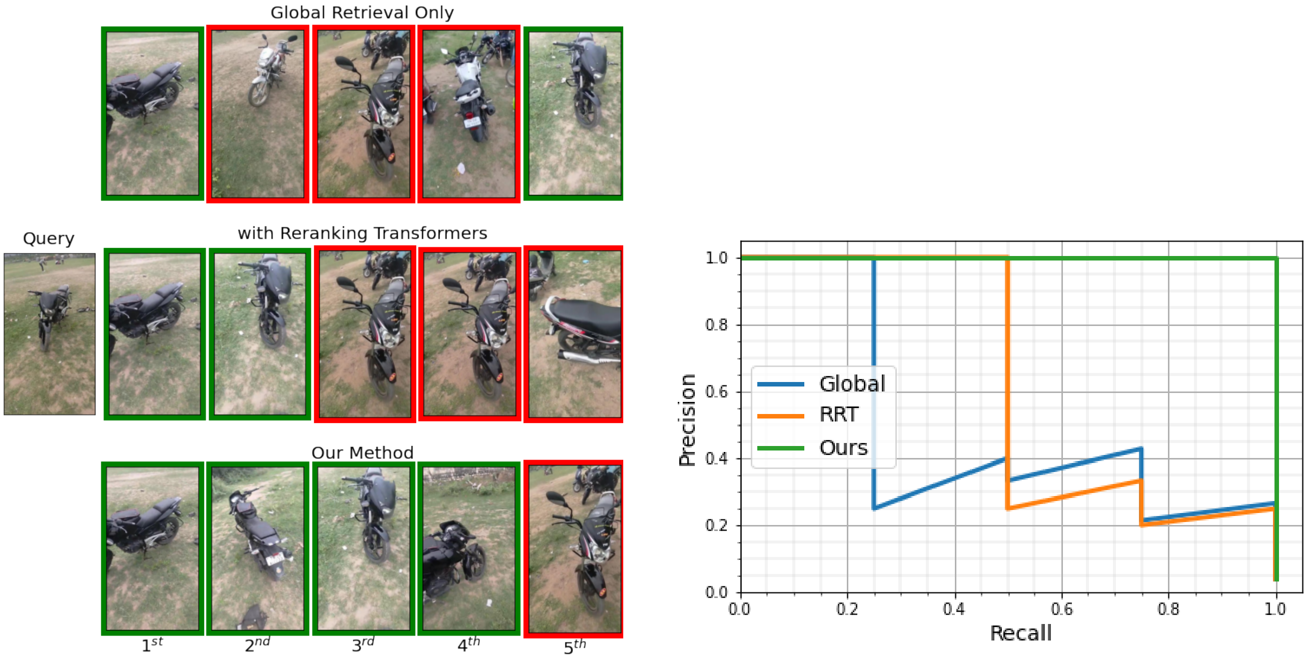


Figure 7. CO3D-Retrieve: Example 4.

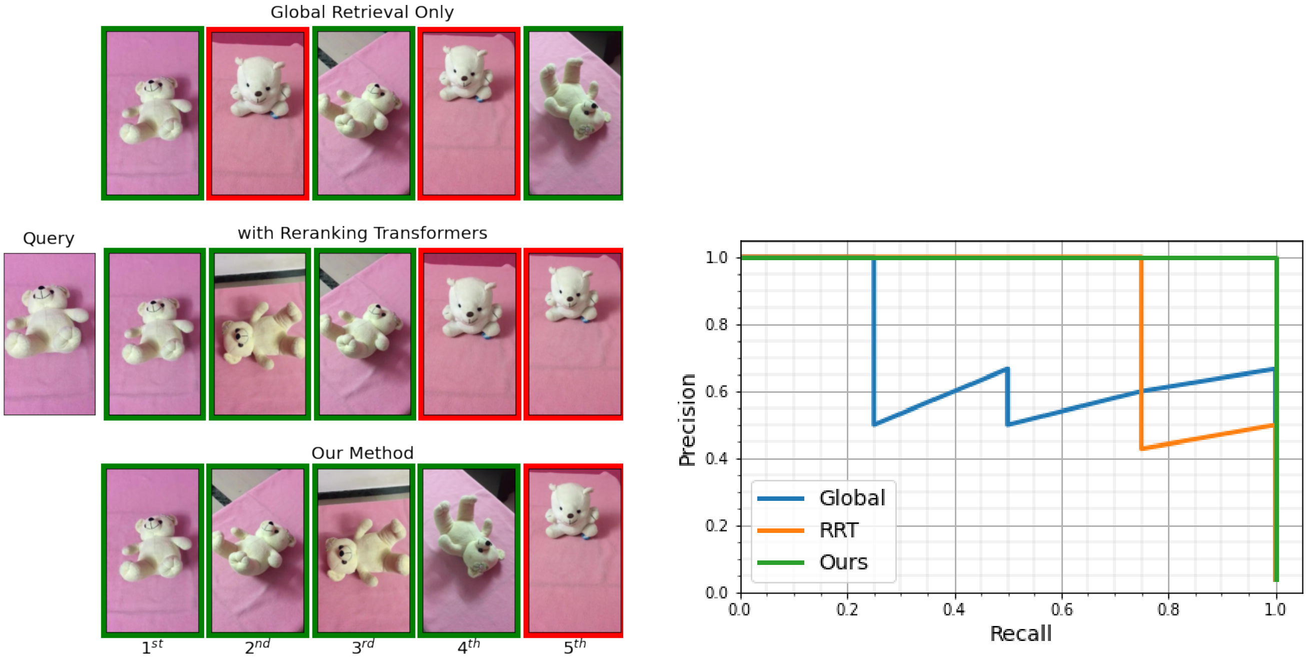


Figure 8. CO3D-Retrieve: Example 5.

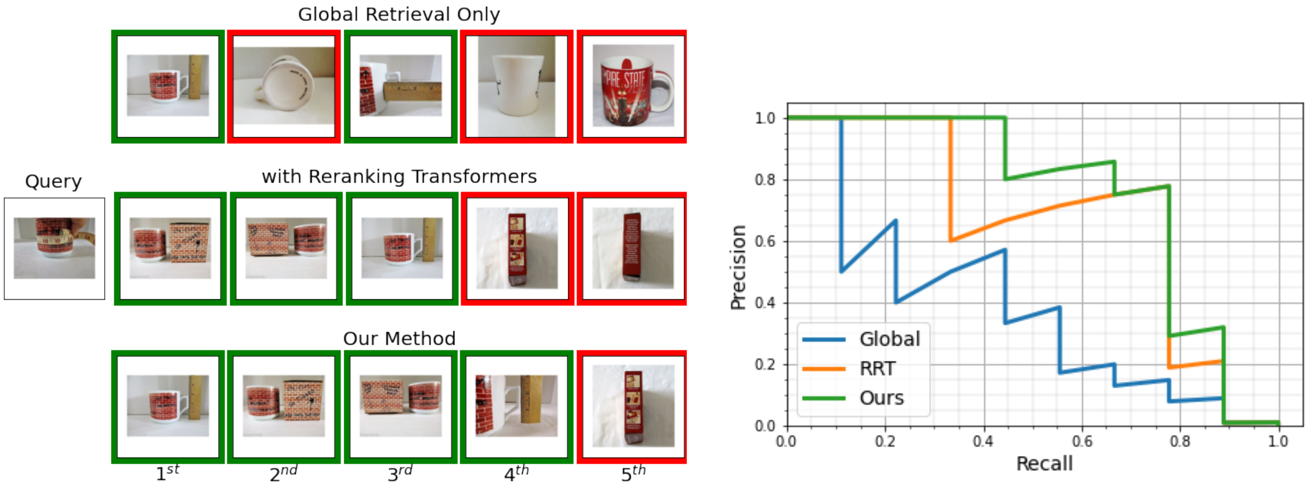


Figure 9. SOP [2] dataset. Example 1.

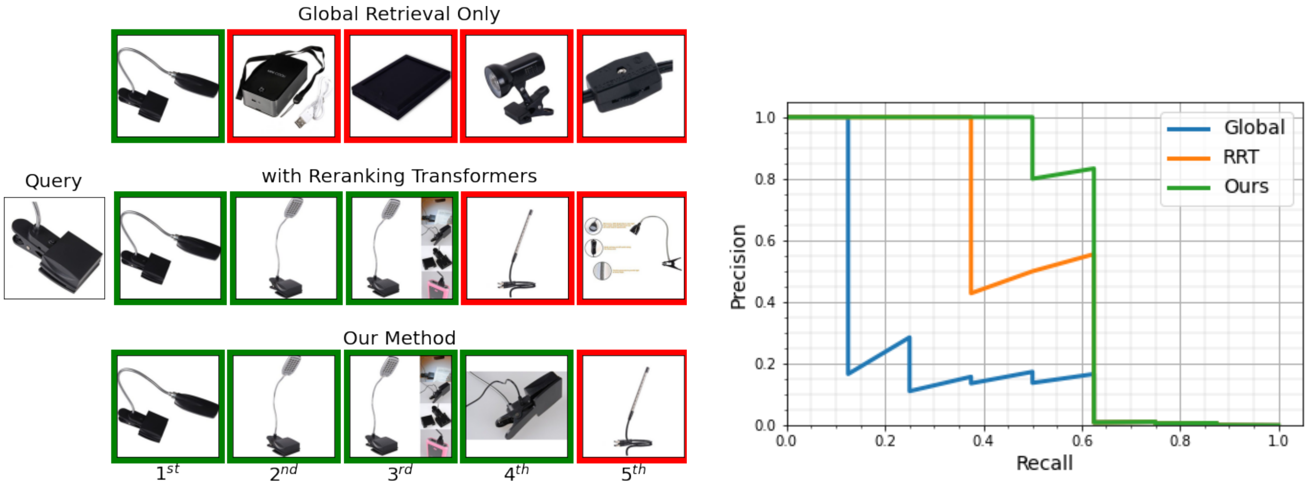


Figure 10. SOP [2] dataset. Example 2.

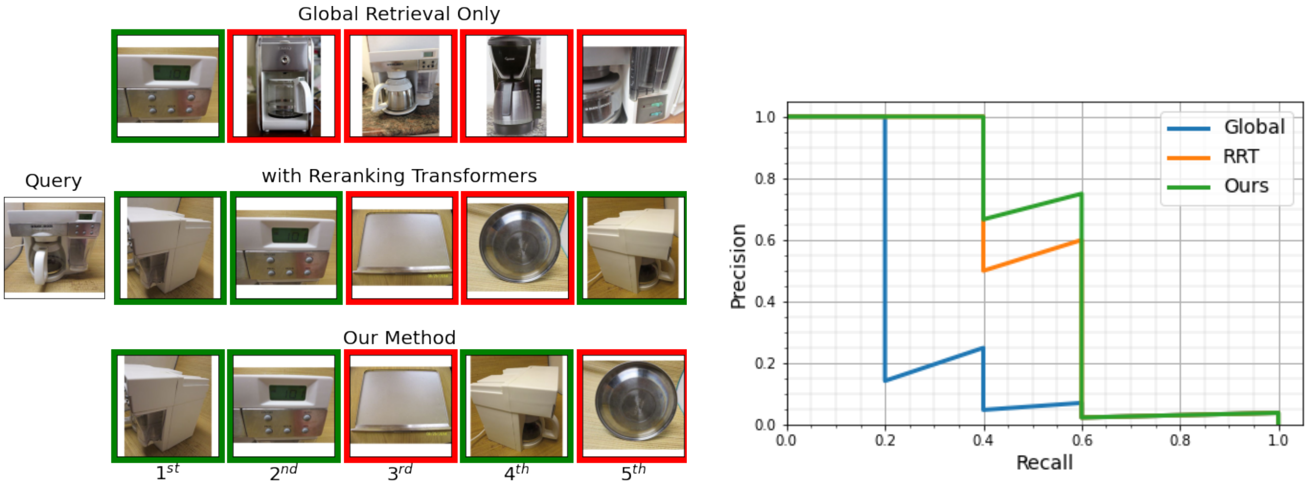


Figure 11. SOP [2] dataset. Example 3.

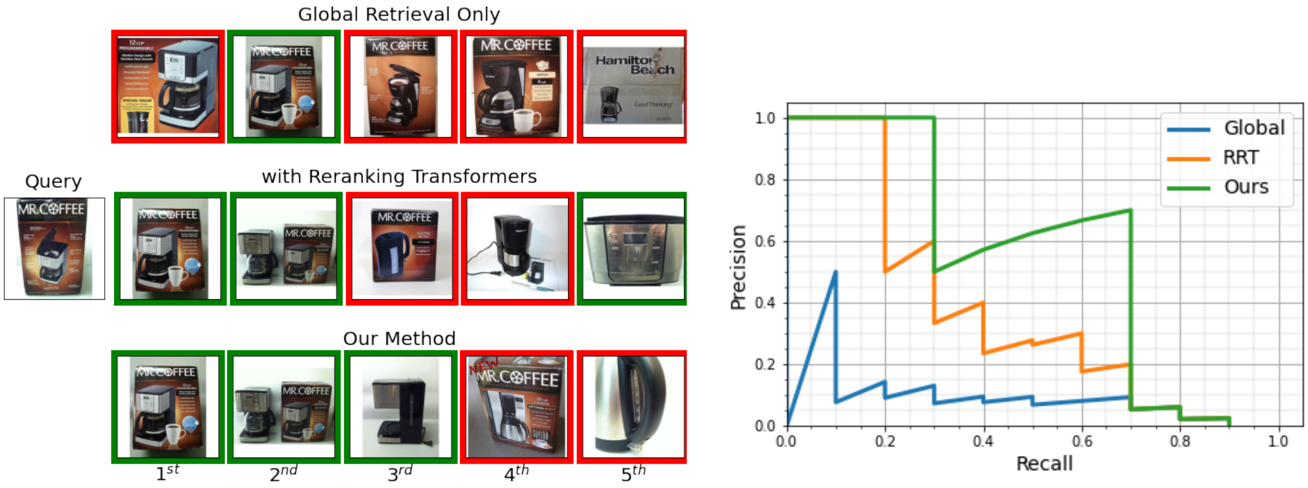


Figure 12. SOP [2] dataset. Example 4.

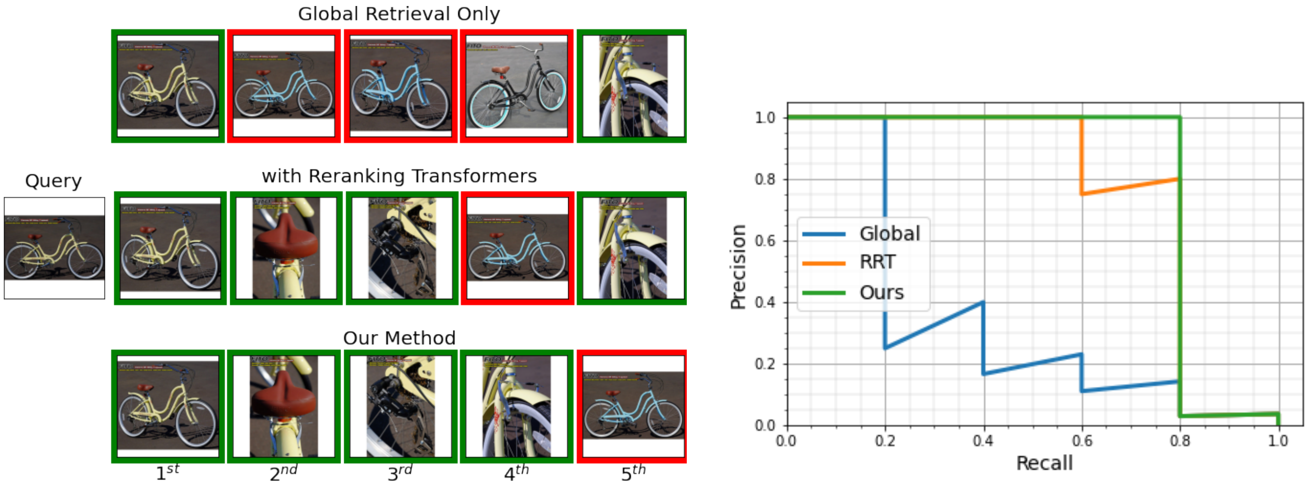


Figure 13. SOP [2] dataset. Example 5.

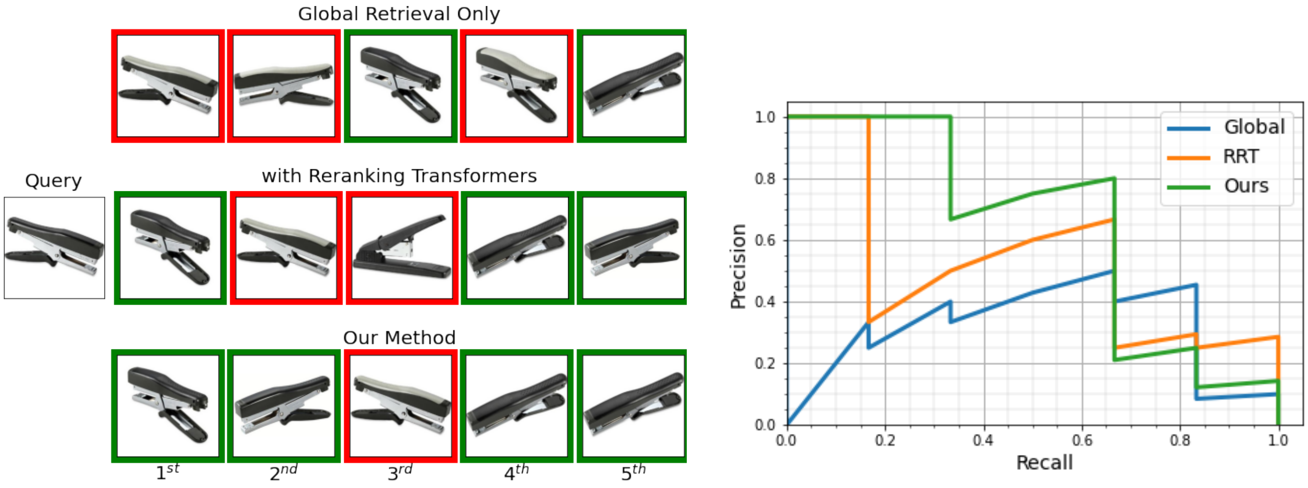
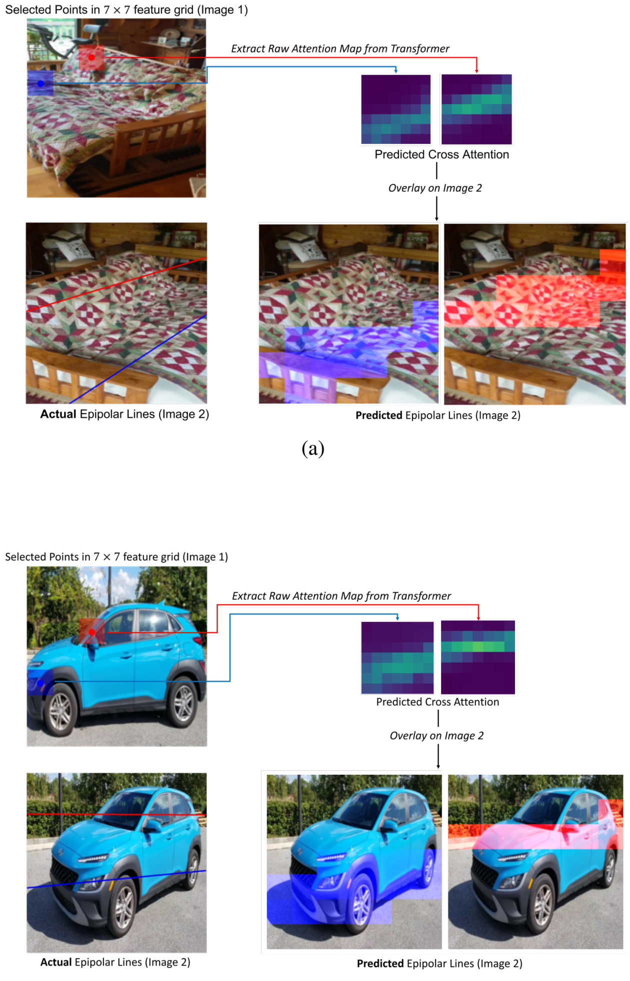


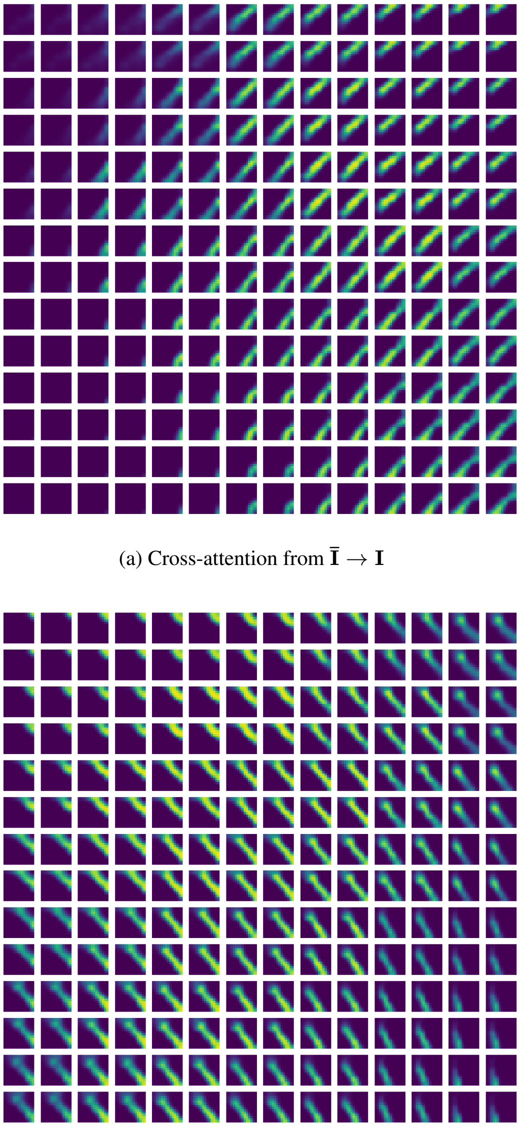
Figure 14. SOP [2] dataset. Example 6.


Selected Points in 7 × 7 feature grid (Image 1)
Extract Raw Attention Map from Transformer
Predicted Cross Attention
Overlay on Image 2
Actual Epipolar Lines (Image 2)
Predicted Epipolar Lines (Image 2)
Selected Points in 7 × 7 feature grid (Image 1)
Extract Raw Attention Map from Transformer
Predicted Cross Attention
Overlay on Image 2
Actual Epipolar Lines (Image 2)
Predicted Epipolar Lines (Image 2)

(a)

(b)

Figure 15. Examples showing how the cross-attention map predicted by our transformer model (trained with Epipolar Loss) contains information about the true epipolar geometry. Red and Blue colours are used to show the two selected points and their corresponding actual vs predicted epipolar lines.



(a) Cross-attention from $\bar{\mathbf{I}} \rightarrow \mathbf{I}$

(b) Cross-attention from $\mathbf{I} \rightarrow \bar{\mathbf{I}}$

Figure 16. Cross-attention maps extracted from the transformer model trained with $448 \times 448$ input images. Due to the higher input resolution, the cross-attention maps are obtained at a higher resolution of $14 \times 14 \times 14 \times 14$.

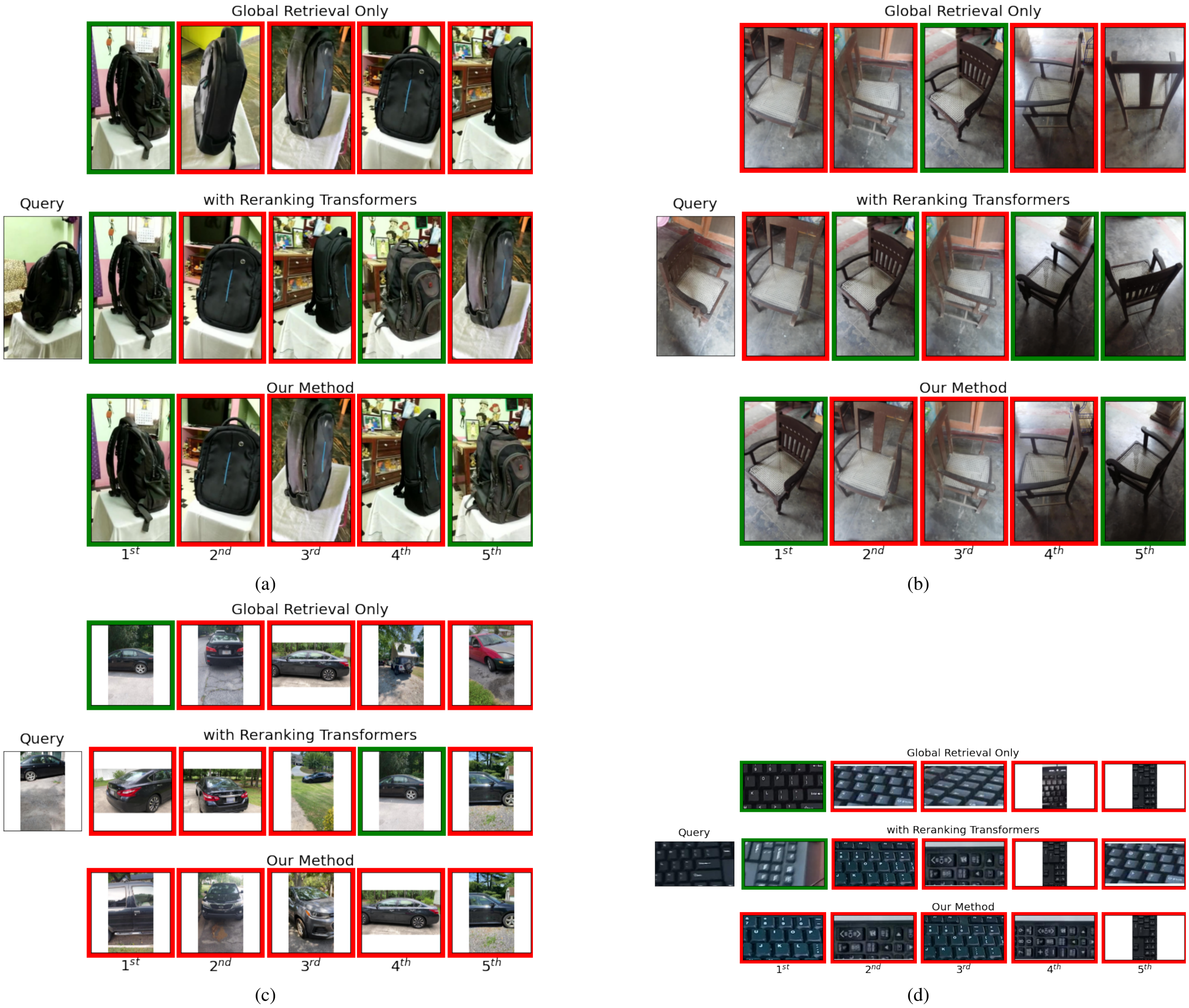
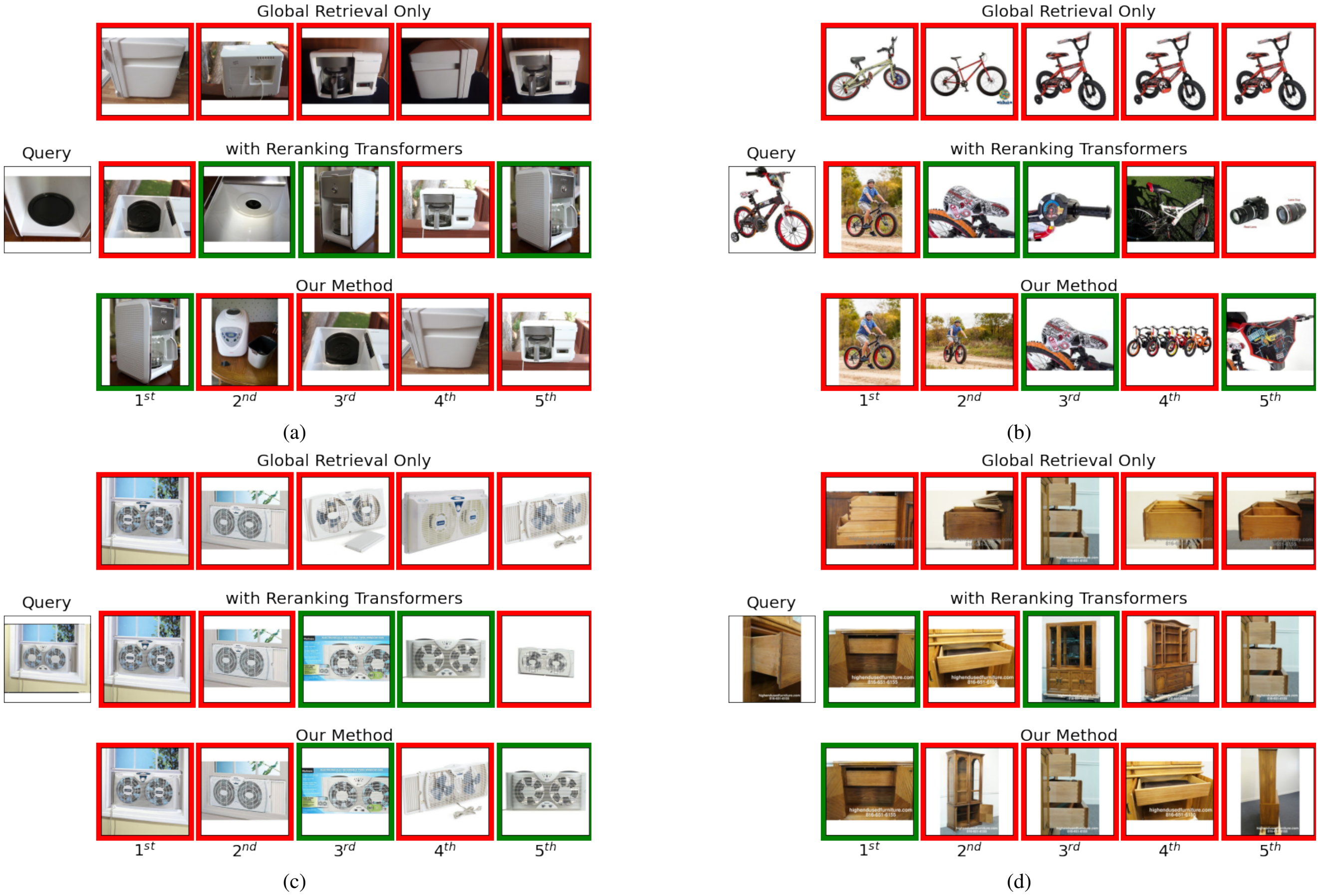


Figure 17. Failure cases from the CO3D-Retrieve dataset.



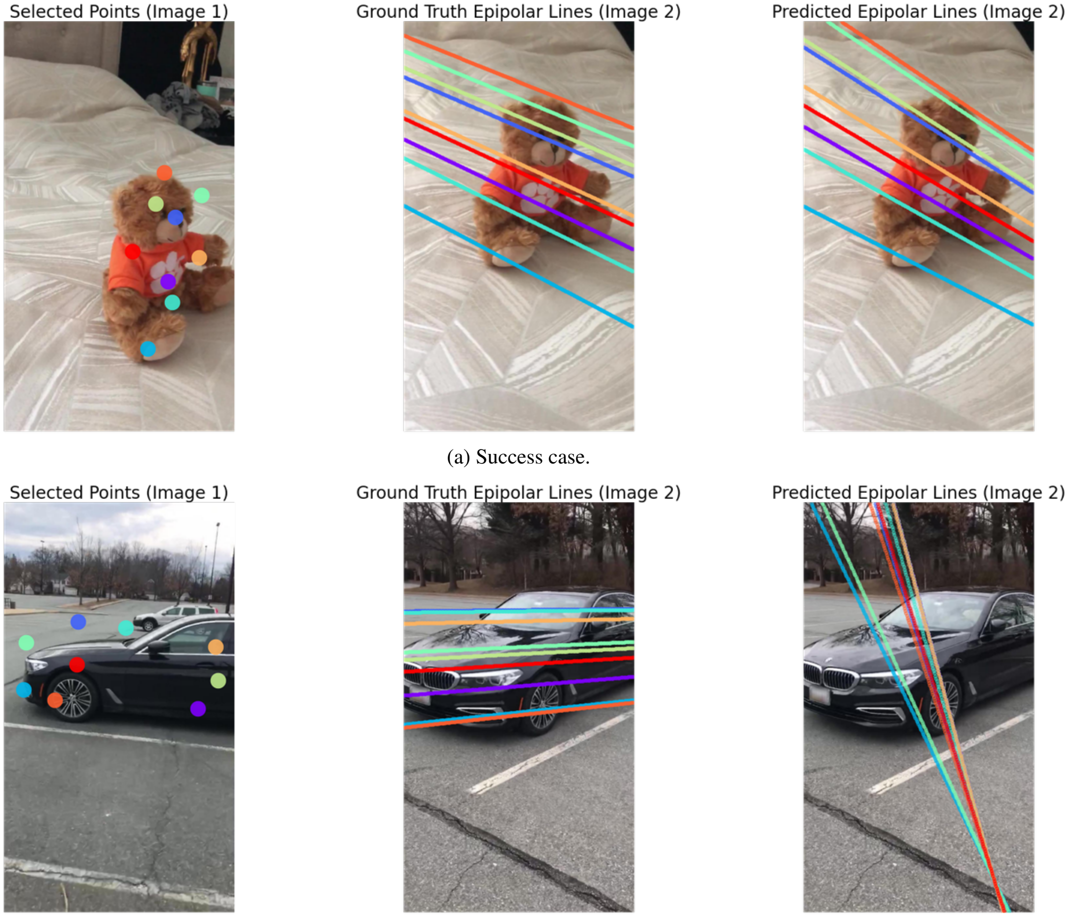
(a)

(b)

(c)

(d)

Figure 18. Failure cases from the SOP [2] dataset.


Selected Points (Image 1)
Ground Truth Epipolar Lines (Image 2)
Predicted Epipolar Lines (Image 2)
Selected Points (Image 1)
Ground Truth Epipolar Lines (Image 2)
Predicted Epipolar Lines (Image 2)

(a) Success case.

(b) Failure case.

Figure 19. Qualitative examples demonstrating the Epipolar geometry predicted using a pretrained LoFTR [3] for matching and MAGSAC++ [1] for robust optimization.