EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems

Engineering and Physical Sciences Research Council

A Light Touch Approach to Teaching Transformers Multi-view Geometry

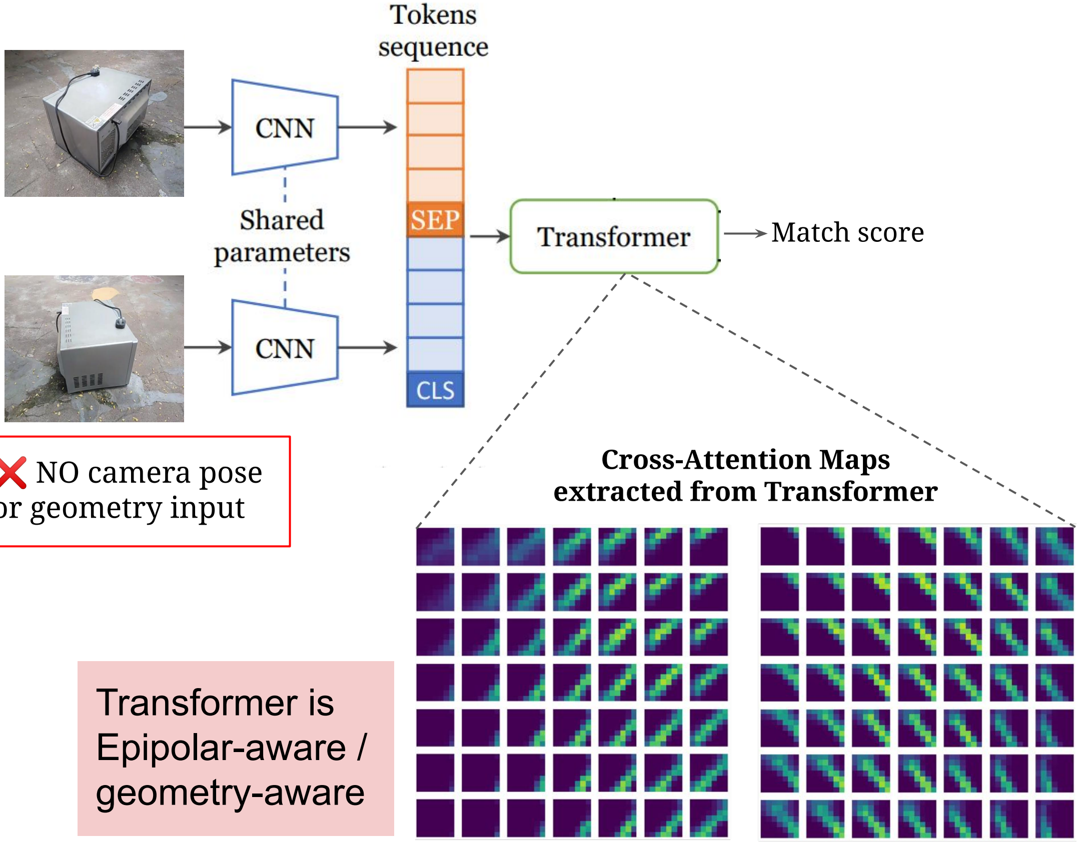
Yash Bhalgat, João F. Henriques, Andrew Zisserman

Visual Geometry Group, University of Oxford

CVPR 2023

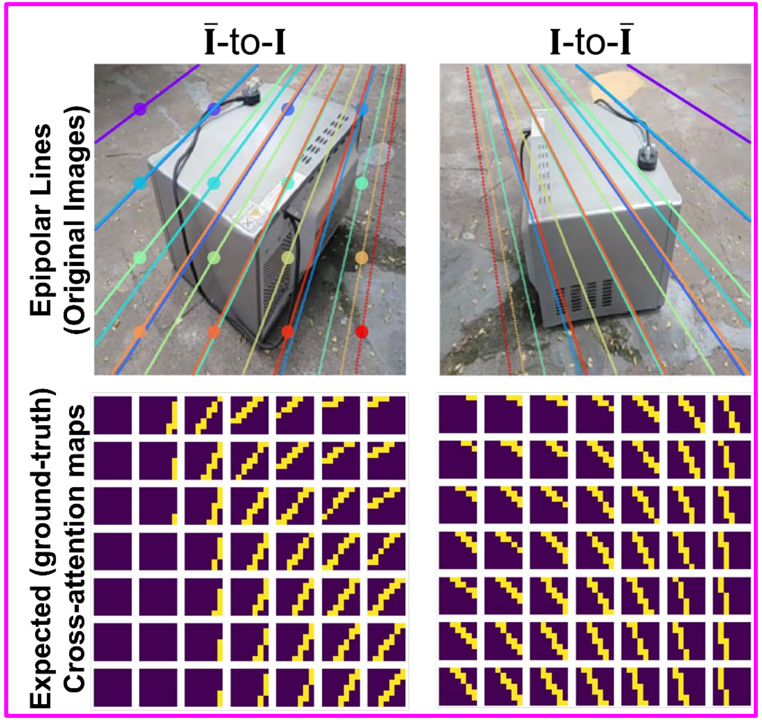
Poster Tag: TUE-PM-078

TL;DR (1/3) - **Geometry-aware Transformer**



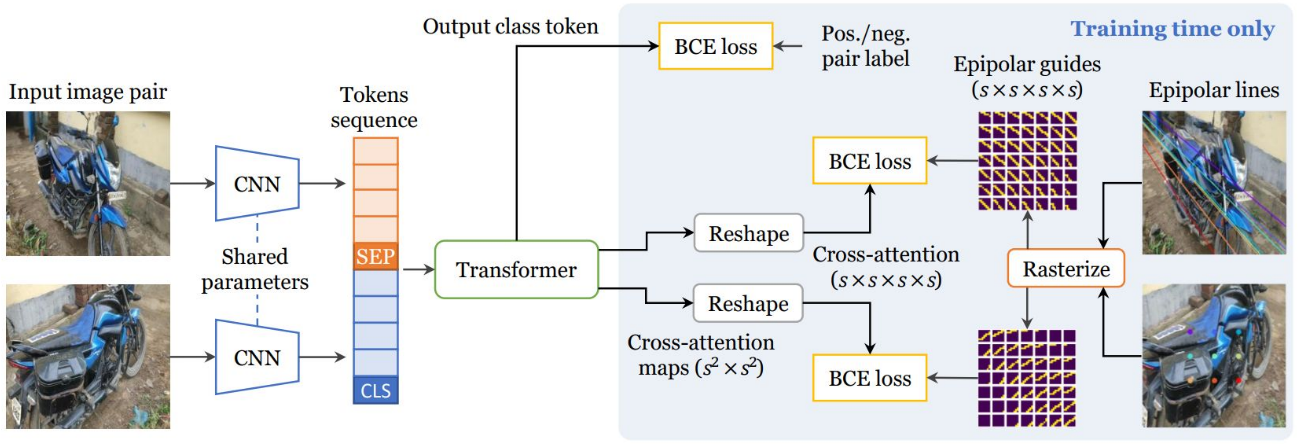
❌ NO camera pose or geometry input

Transformer is Epipolar-aware / geometry-aware



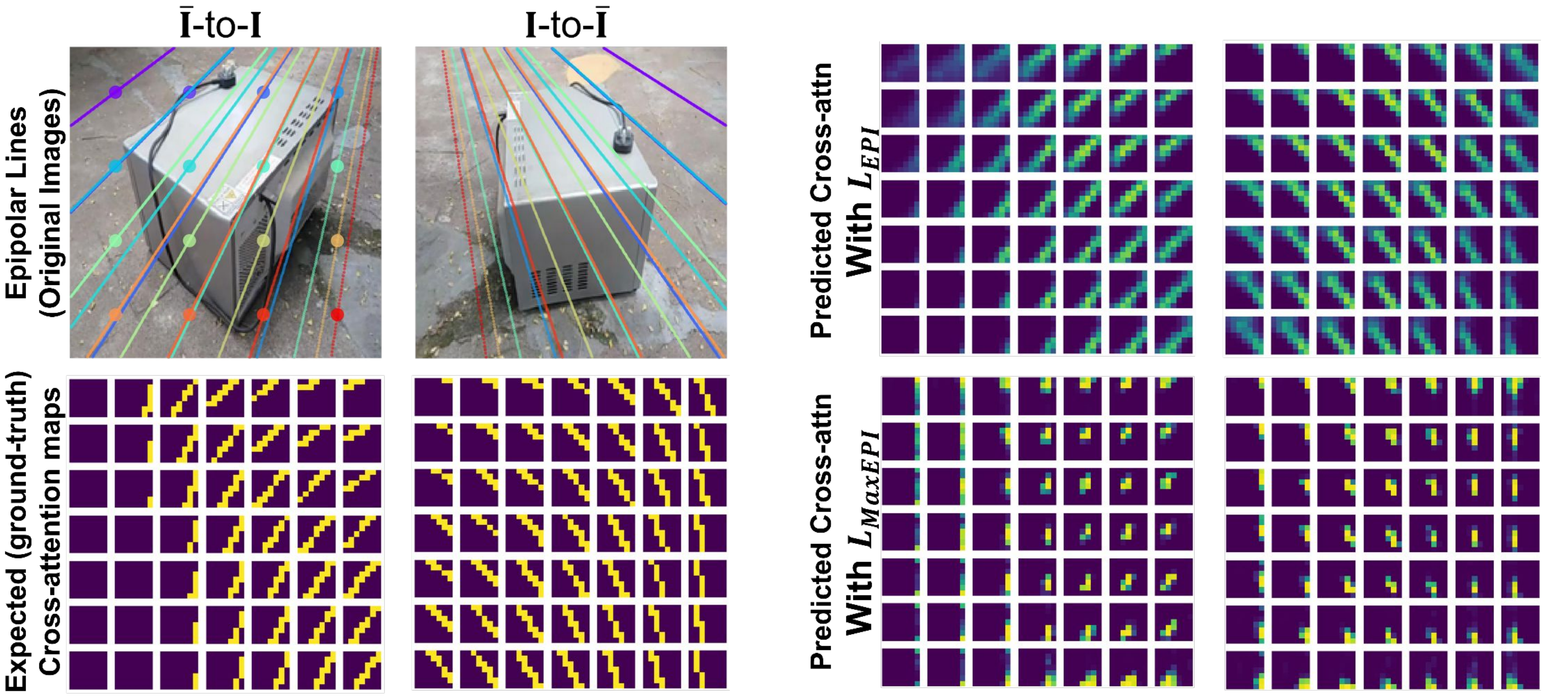
TL;DR (2/3) - **Epipolar-guided training**

- The world is inherently 3D and laws of projective geometry are a useful prior when dealing with images
- Vision Transformers (ViTs) can already search for matches (i.e. attend) across images, e.g. when used for retrieval. But **ViTs lack geometric priors.**
- **Can we keep ViT's flexibility, but add geometric priors for robustness?**



- We propose an ***Epipolar-guided training*** method to incorporate multi-view geometric priors into Transformers.
- Ground-truth pose or epipolar geometry is required *only during training*. During inference, the Transformer implicitly uses geometric reasoning in its predictions.

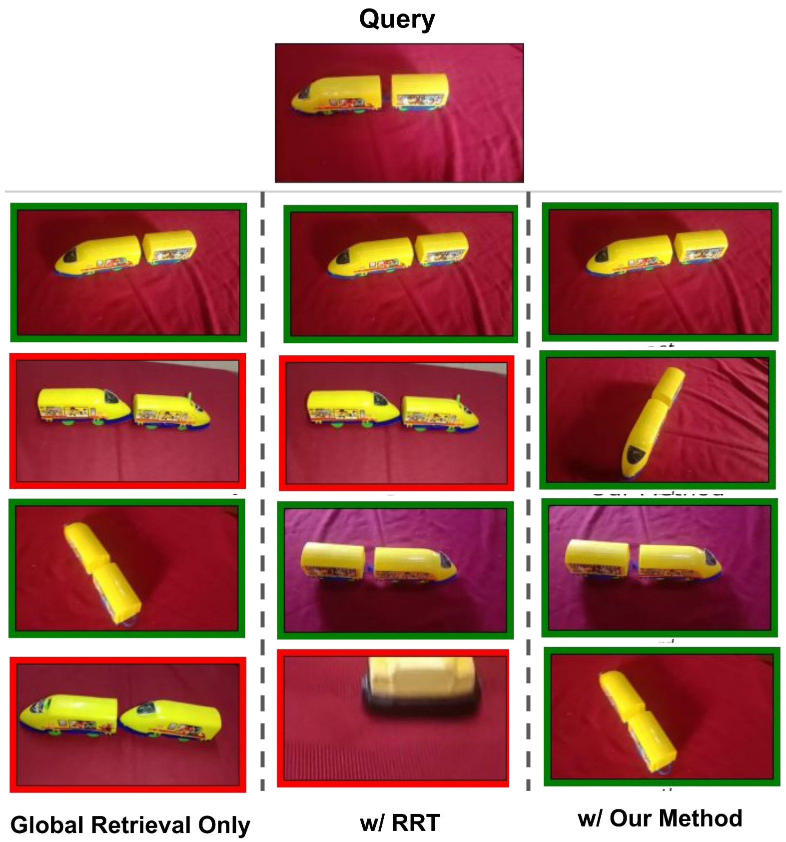
TLDR (2/3) - **What does the Transformer learn?**



Shown here: Predicted Cross-Attention maps for a test image pair (i.e. never seen in training) and without any input pose information. The Transformer implicitly estimates the epipolar geometry given 2 images and uses it for downstream predictions, e.g. for pose-invariant object retrieval.

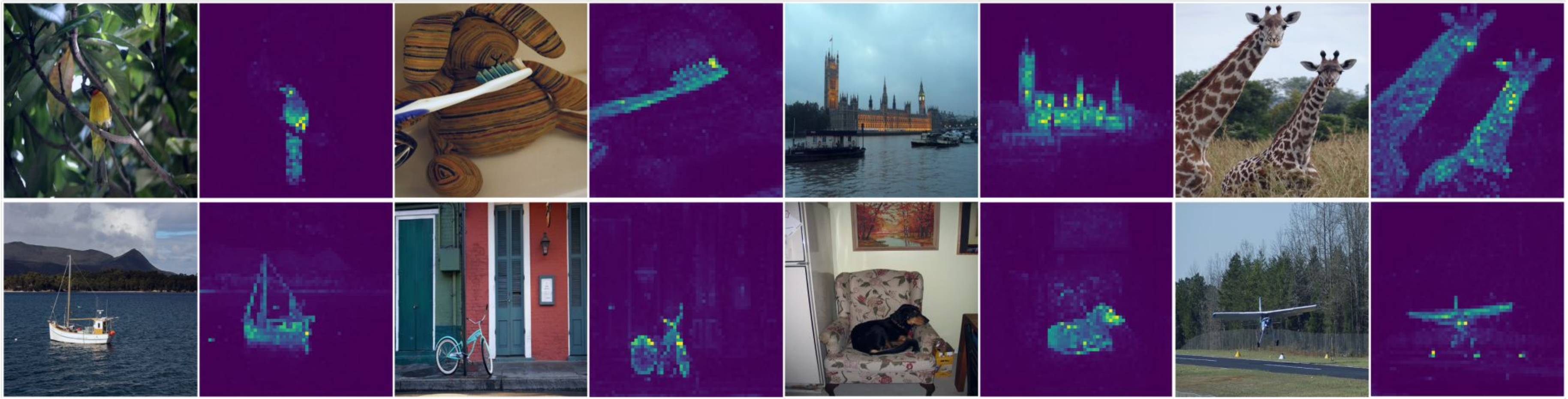
TLDR (3/3) - **State-of-the-art results in object retrieval**

Retrieval task: given a query image of an object, find other images of the same object in a large-scale dataset



Motivation

Vision Transformers (ViT): a success story



- Adopted Transformers after their success with natural language processing (e.g. GPT).
- Emergent property: **attends to objects** even without being explicitly supervised.

Caron et al., Emerging properties in self-supervised vision transformers, ICCV'21

Motivation

- The world is inherently 3D.
- There are *rigid laws* of projective geometry that are obeyed at all times.

Useful *prior information* to deal with ambiguity.

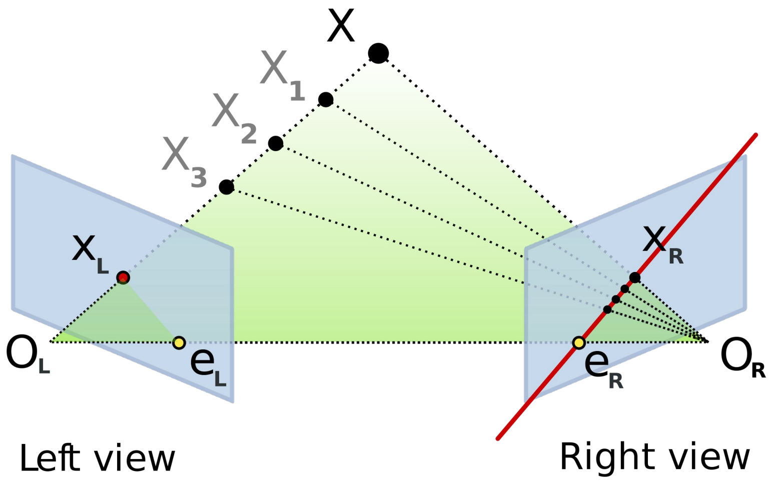
- However, the observed scenes and viewpoints can have *near-infinite variety*.
- Thus ViTs excel due to their immense flexibility, as they have no visual priors (unlike e.g. CNNs).

Can we keep ViT's flexibility, but add geometric priors for robustness?

Pose-invariant Image Retrieval

- One example where this can be useful is **image retrieval** from video or photos of a 3D environment.
- Given a query image (e.g. teddy bear, van), we would like to **re-identify** it in other images.
- If we know the camera poses, we can use **epipolar lines** to narrow down the search.

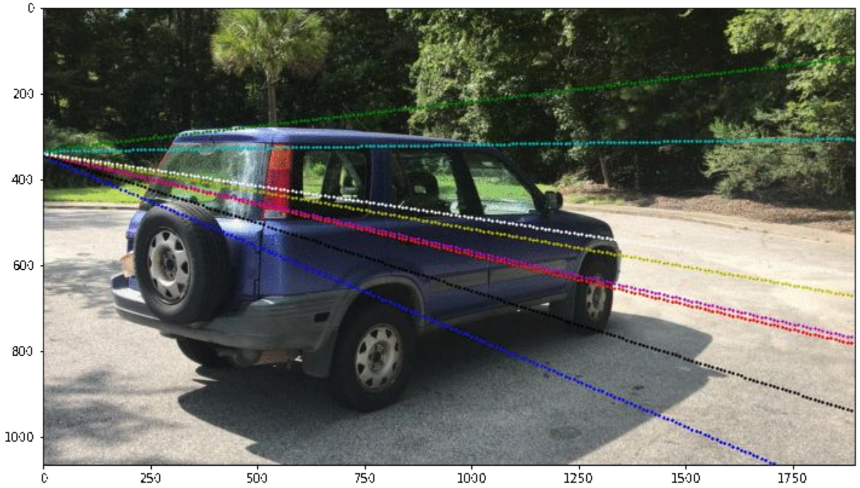
Epipolar Geometry



- Each point (e.g. X_L) in an image (left) projects into a **ray** in 3D space (varying depth, e.g. $X_1, X_2, \ldots$).
- Seen from another image (right), this 3D ray will appear as a **2D line** – an **epipolar line.**

Epipolar Geometry

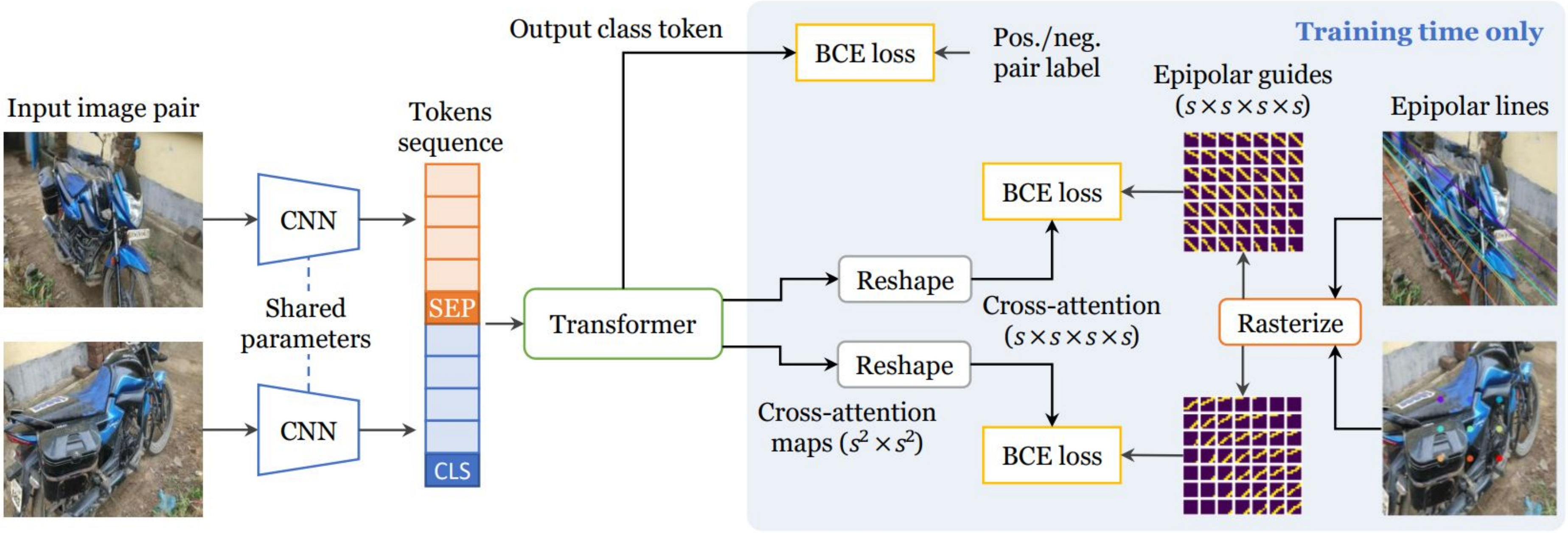
Randomly selected points in Image 1



Epipolar lines corresponding to the points in Image 1

- *Idea:* ViTs already **search for matches** (attend) across images when used for retrieval.
- Can we nudge them to do this search **only along epipolar lines**?

A Light Touch approach



- Local features extracted by a CNN are concatenated (along with CLS and SEP tokens) and input to a Transformer
- CLS token output is trained with BCE loss to predict if the input images match → Outputs score in [0.0, 1.0]
- Epipolar lines obtained with ground-truth pose information are rasterized into *s x s x s x s* tensors and used to supervise the Transformer's cross-attention maps using BCE losses

Proposed Epipolar Loss

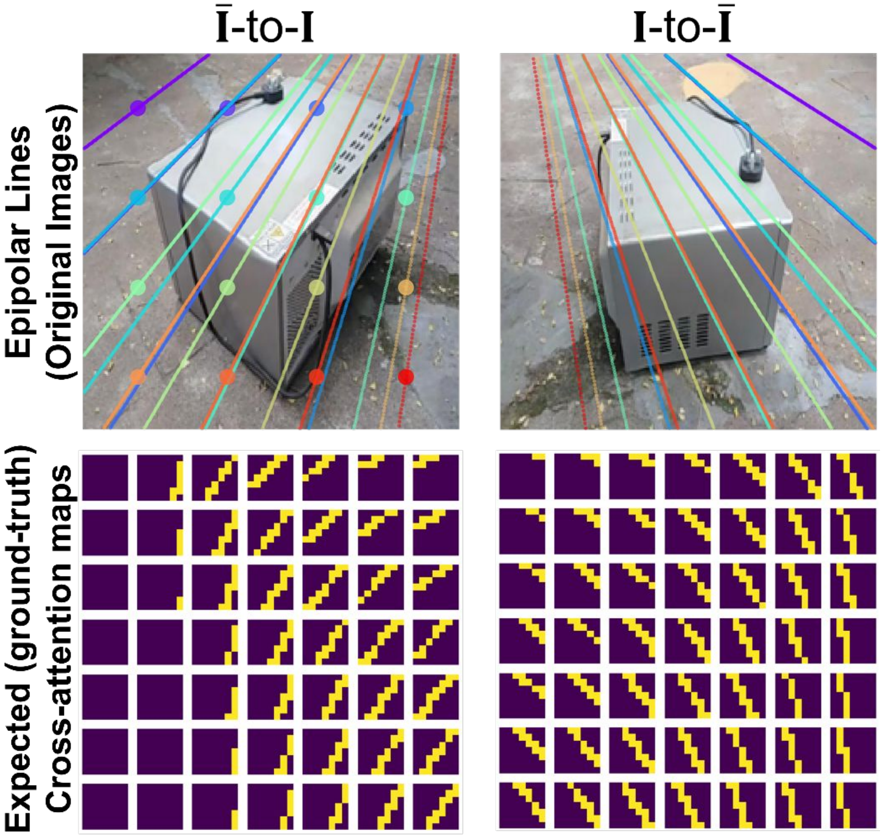
Epipolar Loss

$$L^{12}(i,j) = \mathrm{BCE}(\sigma(A^{12}(i,j)), \mathbb{1}(i,j))$$
$$L^{21}(i,j) = \mathrm{BCE}(\sigma(A^{21}(i,j)), \mathbb{1}(i,j))$$
$$L_{EPI} = \sum_{i=1}^{s^2}\sum_{j=1}^{s^2} L^{12}(i,j) + L^{21}(i,j)$$

- *$\{A^{12}, A^{21}\}$* are raw (i.e. before SoftMax) cross-attention maps from last layer

- $\mathbb{1}(i,j)$ is 1 if location *j* in other feature map lies on the epipolar line of location *i* in current map

Problem: encourages **many matches** in each line.



Proposed Max-Epipolar Loss

Max-Epipolar Loss

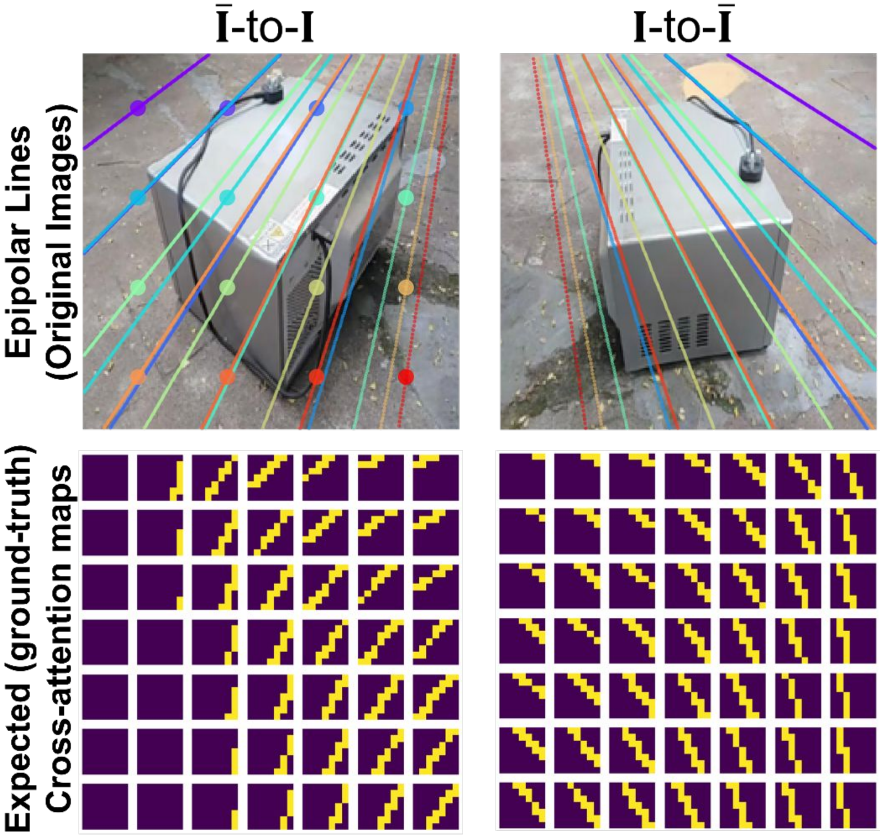
$$L_{MaxEPI} = L_{zero} + L_{max}$$

where

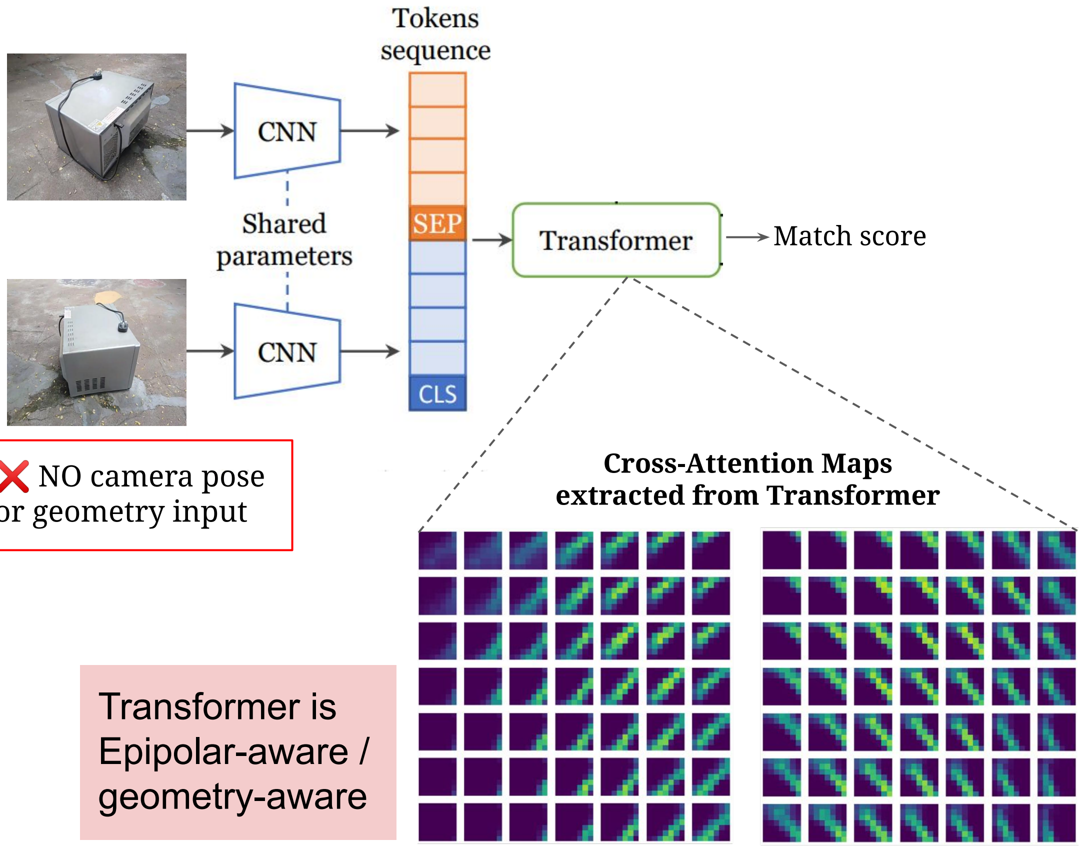
$$L_{max} = \sum_i \mathrm{BCE}\left(\max_{j \in e_i} \sigma(A(i,j)), 1\right)$$

$$L_{zero} = \sum_{\forall i,j,\mathbb{1}(i,j)=0} \mathrm{BCE}(\sigma(A(i,j)), 0)$$

- Not every point on epipolar line is a match in 3D
- L_{EPI} encourages every point on the epipolar line to have high attention
- L_{MaxEPI} *selects a point on the epipolar line with max cross-attention value and encourages cross-attention of that point to be high*

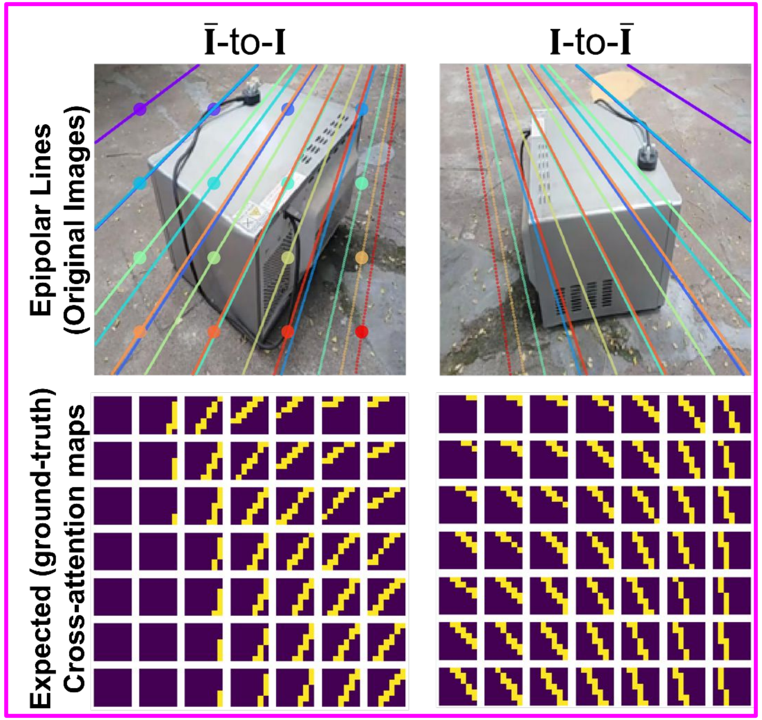


Inference using the geometry-aware Transformer



❌ NO camera pose or geometry input

Transformer is Epipolar-aware / geometry-aware

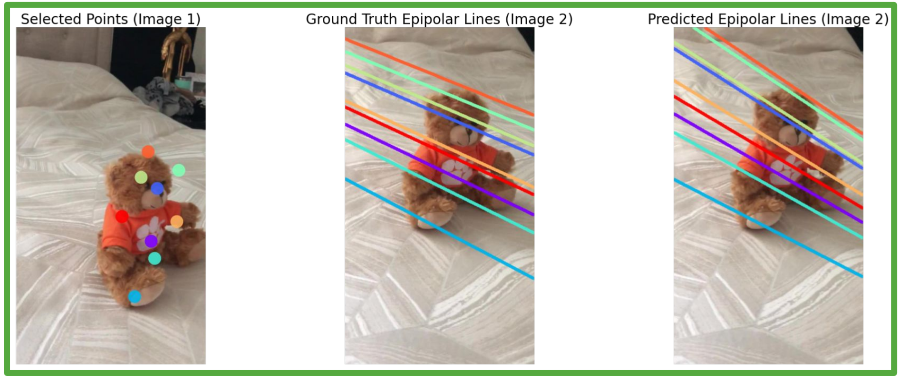


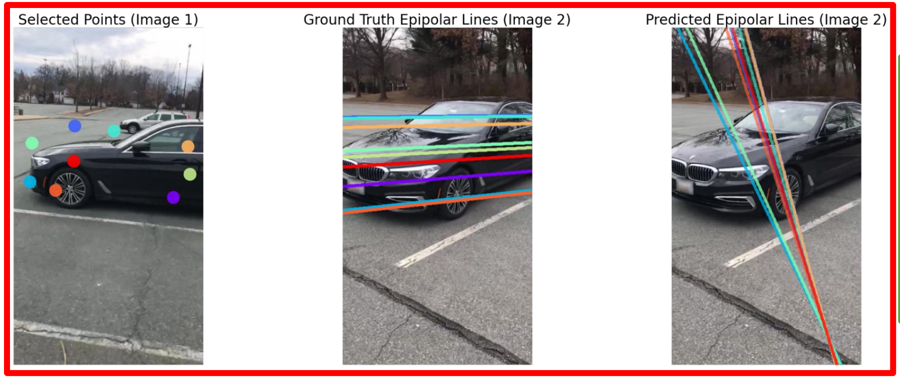
Computing Epipolar Geometry

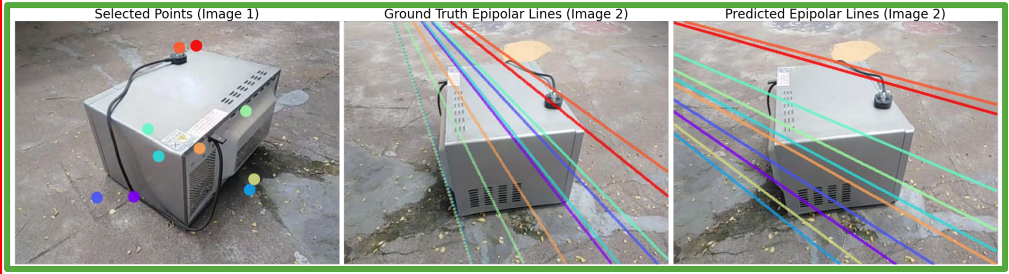
If GT pose isn't available, Fundamental Matrix can be estimated using

- Key-point matching with LoFTR
- Robust estimation with MAGSAC++

Fails to find good correspondences in 20% of cases







CO3D-Retrieve Benchmark

Built on top of [CO3Dv2] dataset

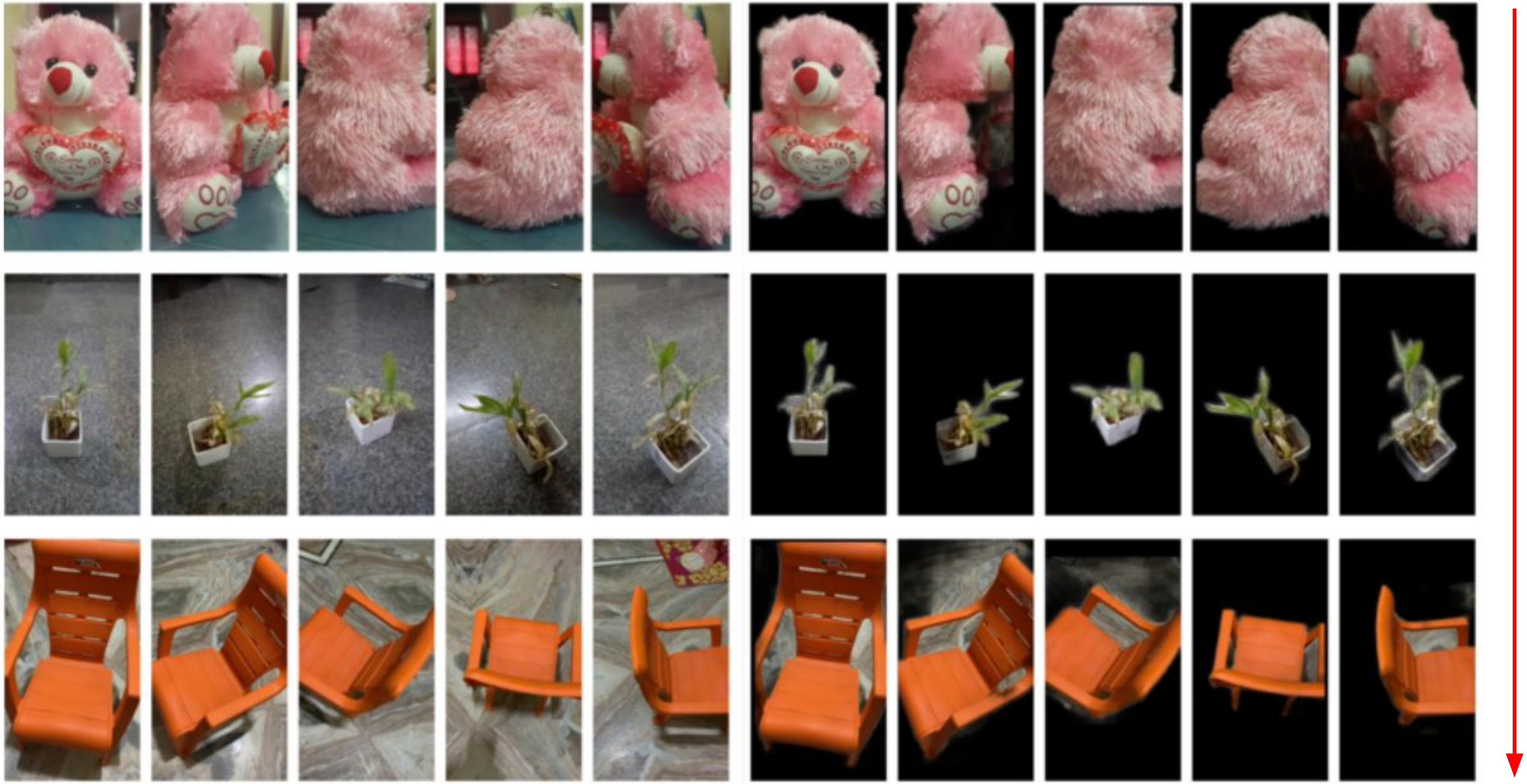
Dataset

- 5 frames per video
- Approx. maximum 144° separation between any two frames
- Total 181,857 images of 36,506 object instances
- Training set: 91,106 images of 18,241 object instances
- Testing set: 90,751 images from 18,265 object instances
- Set of objects in training and testing are non-overlapping

Retrieval setup

- Evaluate with each image as query
- Other images from same object are positives
- All images not of query object are negatives

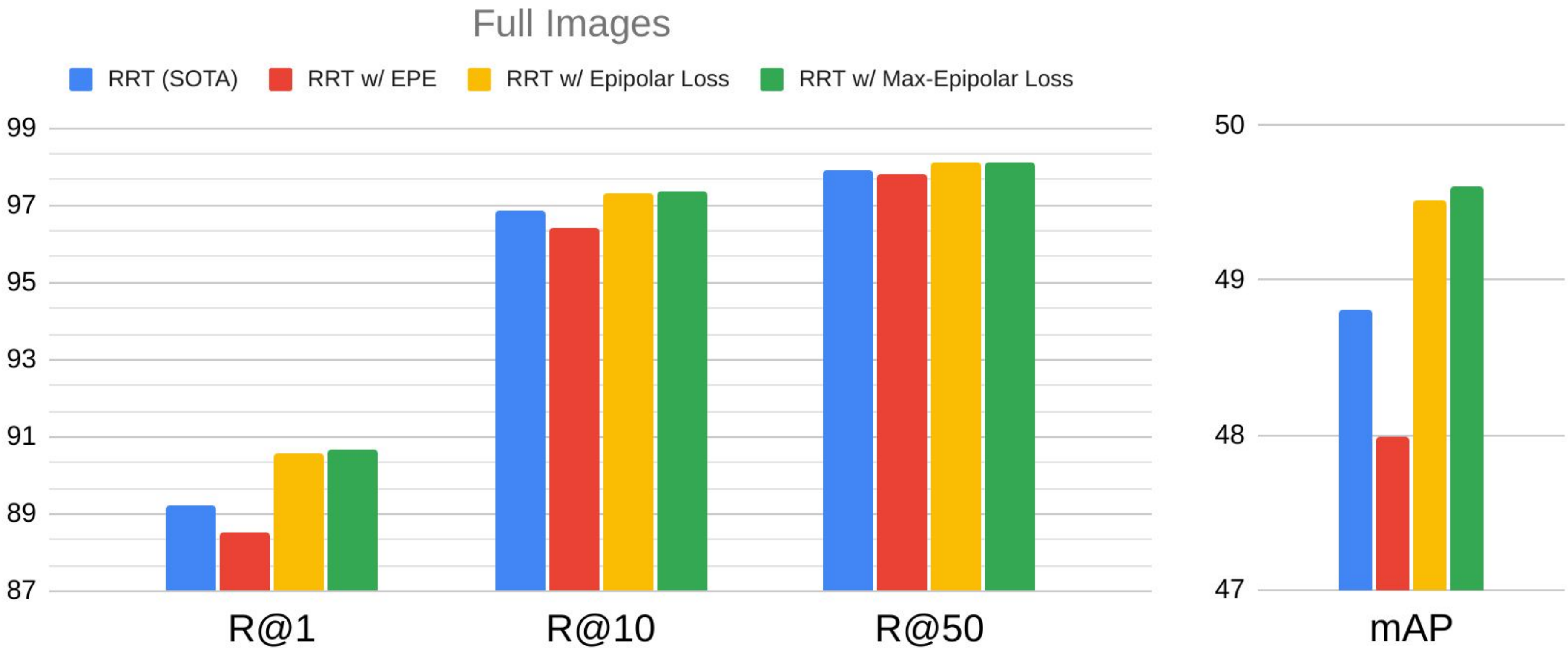
CO3D-Retrieve Benchmark



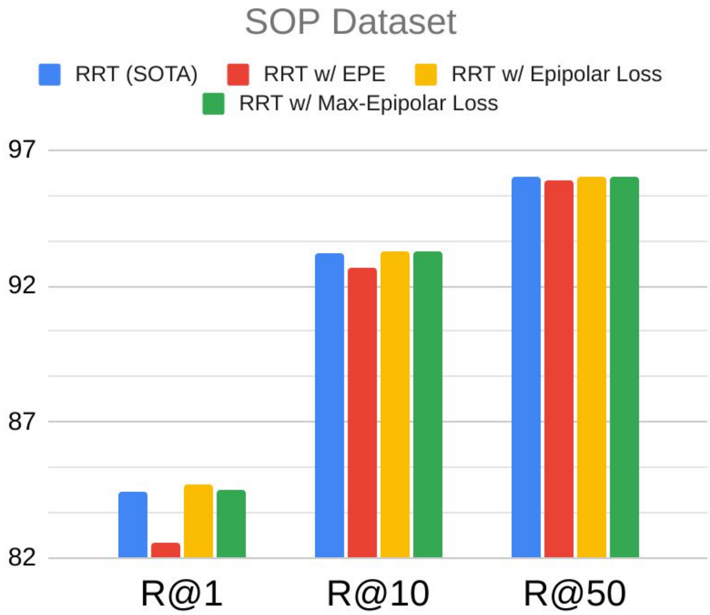
Low overlap

High overlap

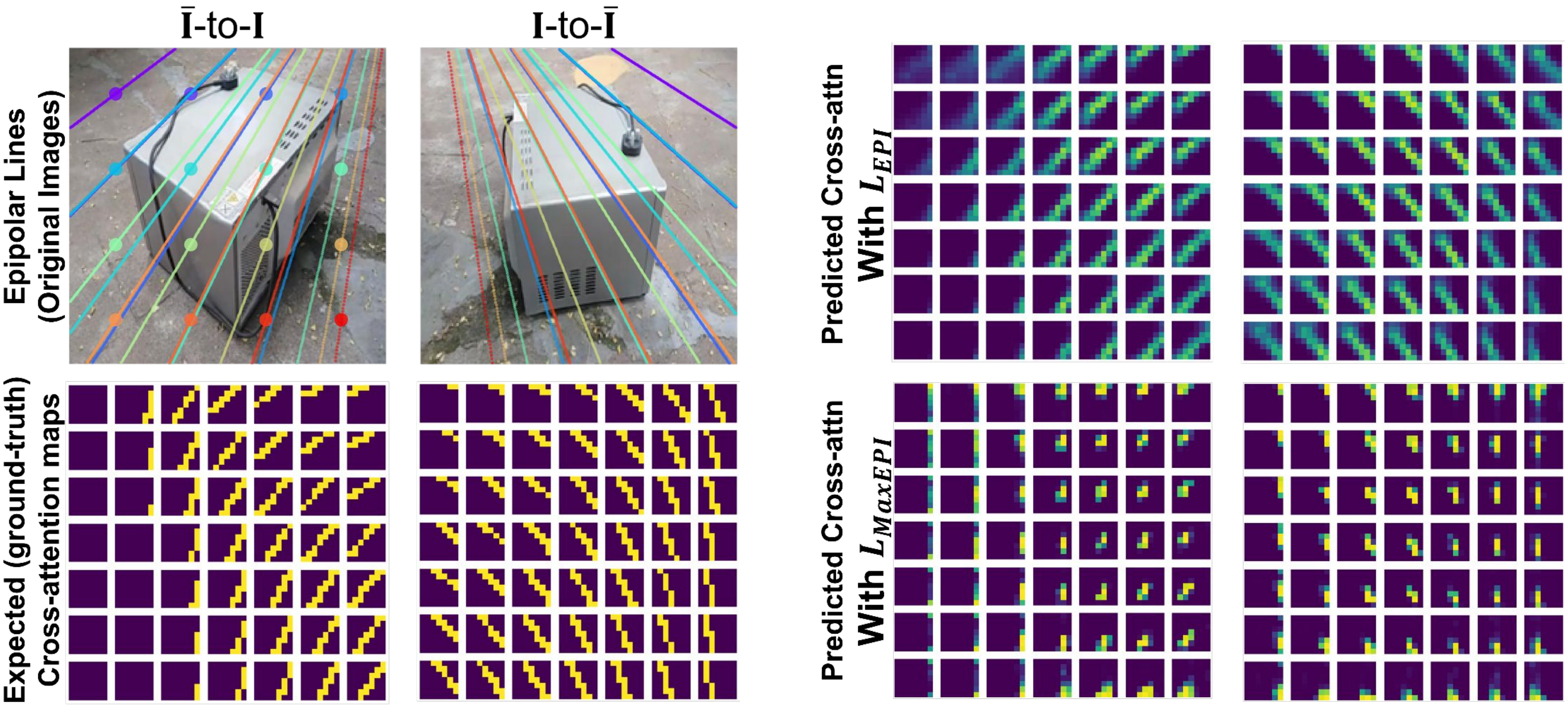
Performance on the CO3D-Retrieve benchmark



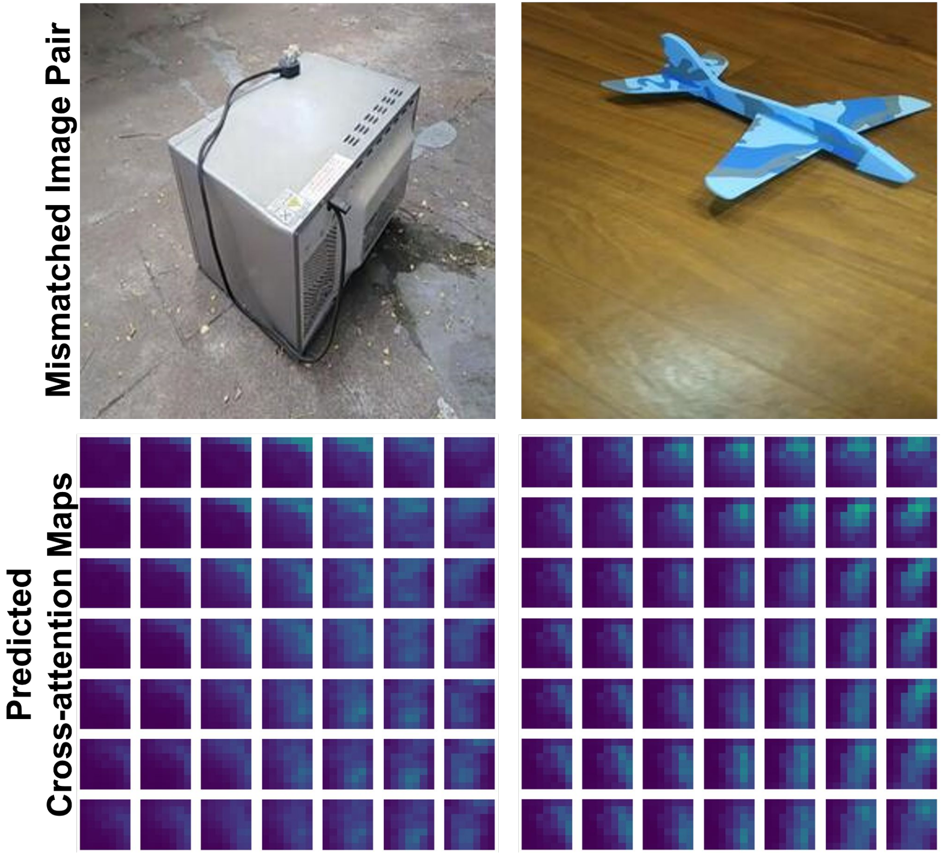
Performance on Stanford Online Products



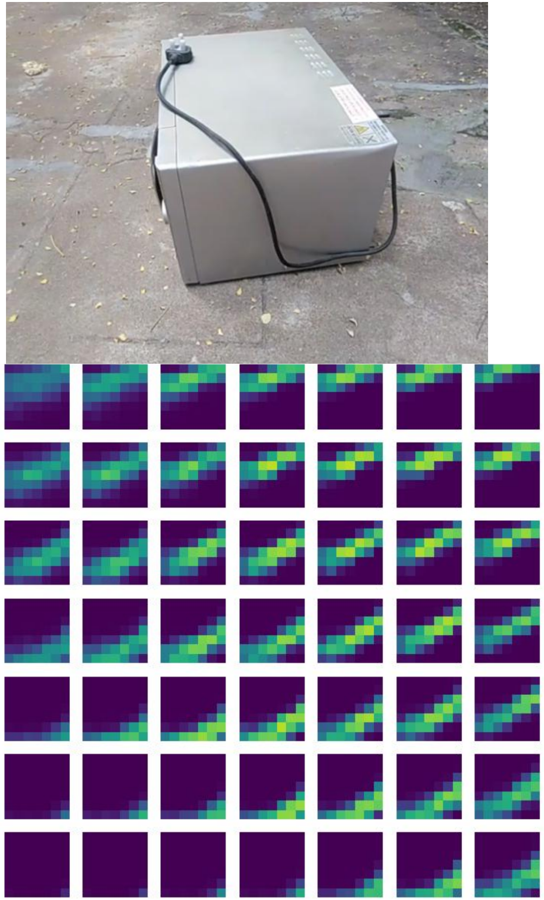
What does the Transformer learn?



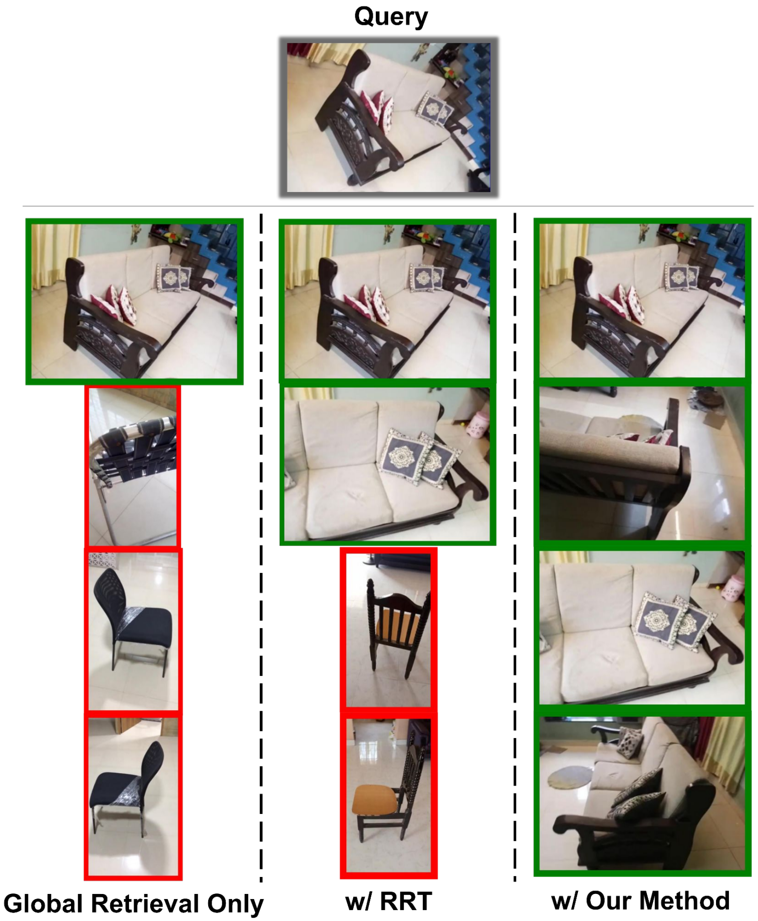
Predicted cross-attention with *mismatched* image pair



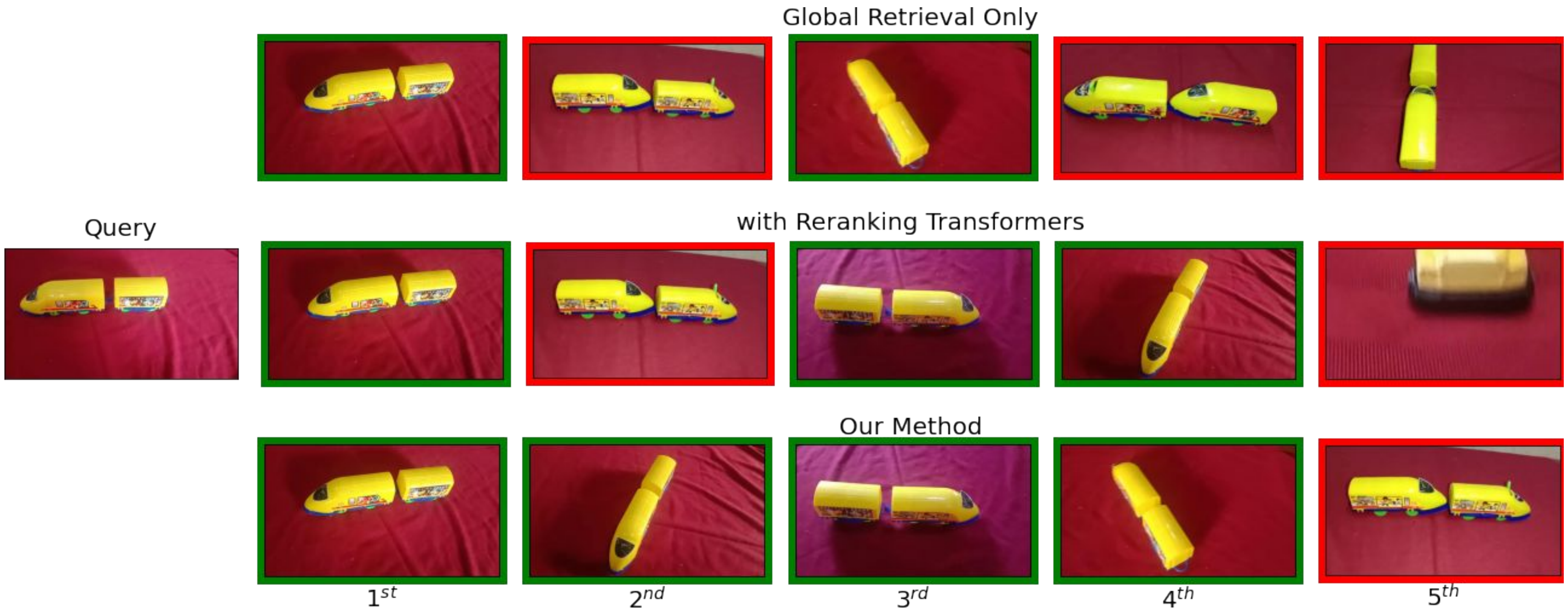
Predicted Epipolar Lines with camera movement



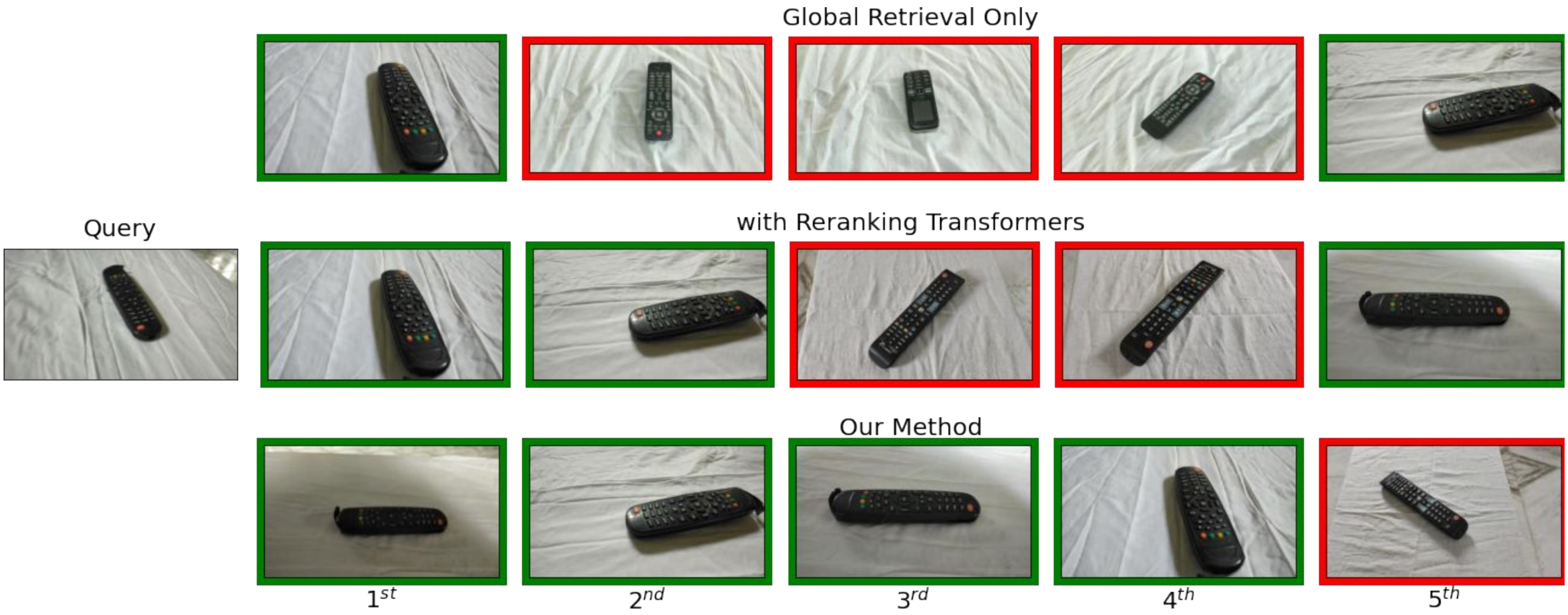
Qualitative Examples: CO3D-Retrieve



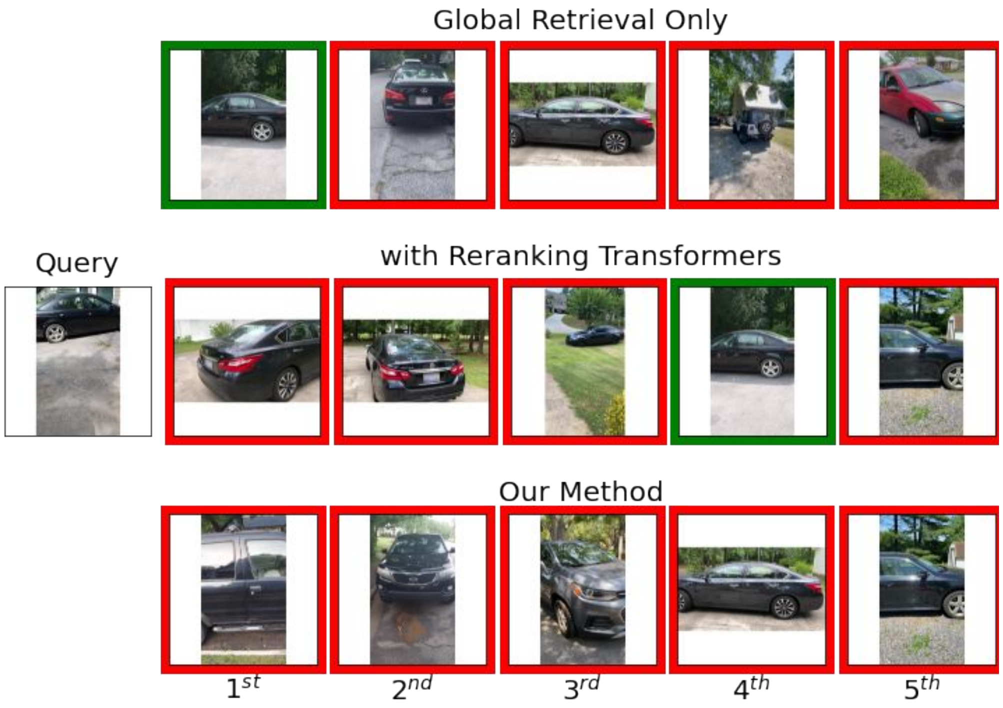
Qualitative Examples: CO3D-Retrieve



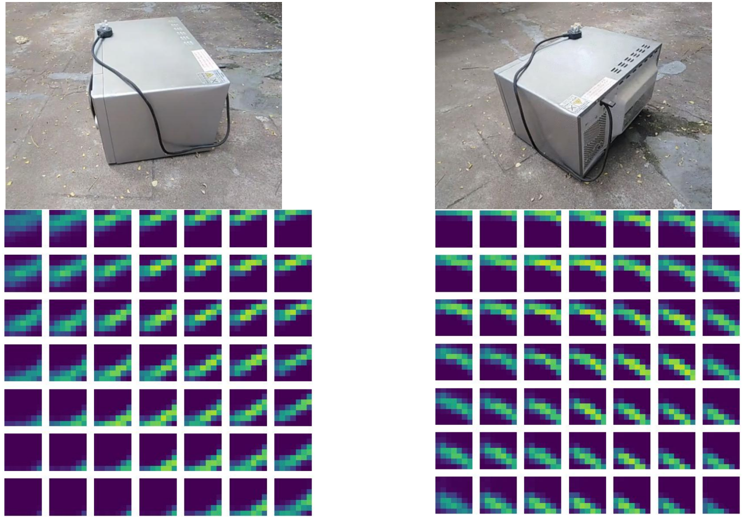
Qualitative Examples: CO3D-Retrieve



Some failure cases



Summary



- In this work we aimed to **teach multi-view geometry** to Transformer networks.
- We propose to do so with **epipolar guides** – a light touch approach.
- Ground-truth information (pose) is only needed **at training time**, not for inference.
- **Implicit loss functions** readily apply to existing architectures – no need to specialize.
- **State-of-the-art results** in object retrieval.
- Future work: other geometric relations or physical laws (e.g. Laws of motion).