C-3PO: Towards Rotation Equivariant Feature Detection and Description

Piyush Bagad Floor Eijkelboom Mark Fokkema Danilo de Goede Paul Hilders Miltiadis Kofinas

University of Amsterdam

ECCV TEL AVIV 2022

UNIVERSITY OF AMSTERDAM

Local Feature Matching

- Correspondence, correspondence and correspondence: Detect, describe and match!
- Promising outcomes with deep learning based methods that *learn* to recognise keypoints

Figure 1. Illustration of detecting, describing & matching keypoints.

R2D2: Reliable and Repeatable Detector and Descriptor

- CNN-based architecture that predicts (i) repeatability maps, (ii) reliability maps
- Simple architecture, modular code, easy to use!

Challenges

- Handling affine distortions such as rotations.

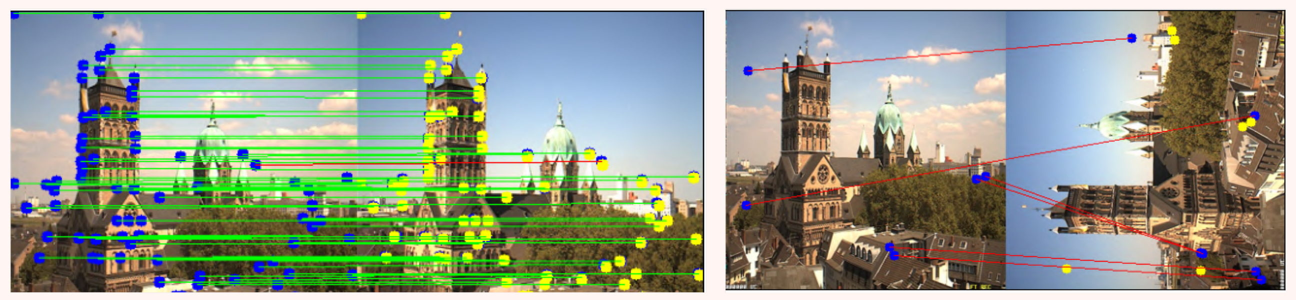


Figure 2. R2D2 finds it hard to maintain matching under rotations of target image.

- Inherent local nature of the task

HPatches Dataset: Overview

Figure 3. (Left) Inherent locality. (Right) HPatches dataset overview.

Geometry and steerability

To ensure the problem adheres to predefined notions of symmetry, we explicitly model the transformations g under which our problem should be symmetric as a group G. The signal s is called *equivariant* to G if applying a symmetry transformation $g \in G$ and then computing the signal in pixel x produces the same result as computing the signal s in x and then applying the transformation g:

$$\text{equivariance: } f(g.x) = g.(f(x)).$$

Rather than modeling a response for each group element (e.g. rotation by 90 degrees), we store the Fourier coefficients of an underlying Fourier basis over S^1 to the signal in order to store continuous responses:

$$s(x) \approx \sum_{n=-N}^{N} a_n e^{inx},$$

for some $N \in \mathbb{N}$. These features are learned using steerable kernels.

Methodology

In order to make local feature matching robust to rotations, we introduce geometric priors to the model directly. For this purpose, we propose **C-3PO**, a family of novel deep feature detection-and-description models based on steerable group convolutional networks.

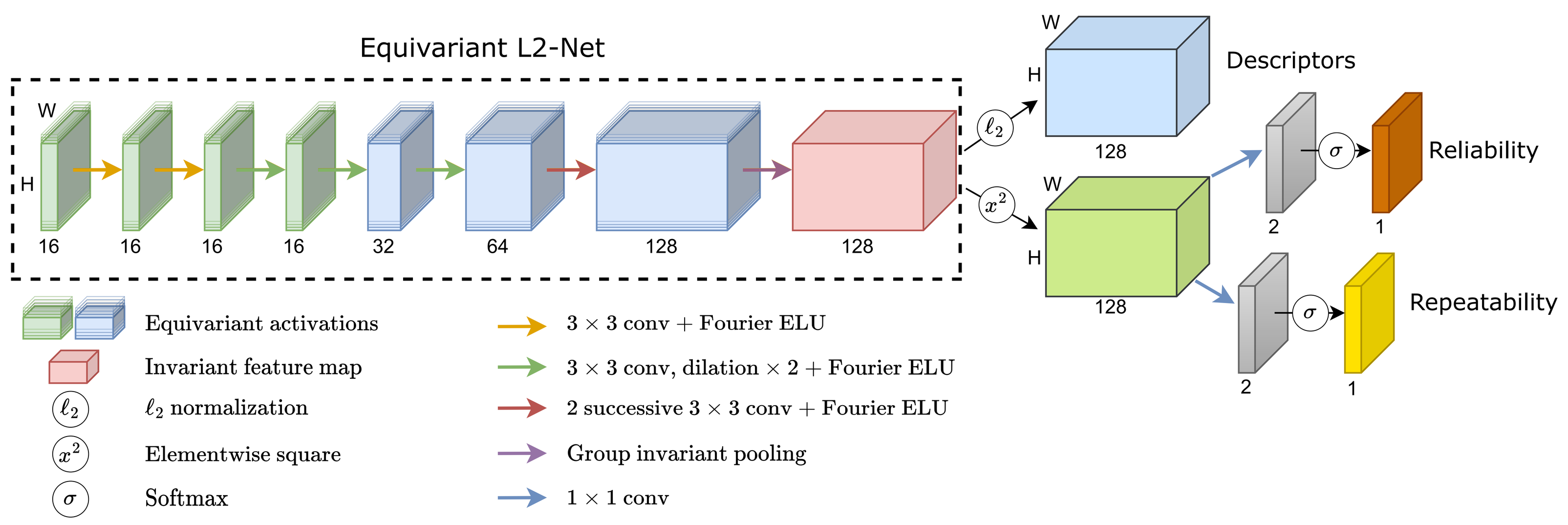


Figure 4. **C-3PO Network Architecture.** The network architecture of the SO(2) variant of C-3PO. The initial layers comprise an equivariant variant of L2-Net. In line with [?], the remaining part of the network consists of three heads outputting the feature descriptors, repeatability map, and reliability map.

We distinguish between three variants of C-3PO: the first two variants of C-3PO are based on the finite group C_n for $n \in \{4, 8\}$, and the last variant on the infinite group SO(2). While the input types of the first layers are equivalent for each variant, the intermediate signals transform according to the regular representations [?] of their respective group.

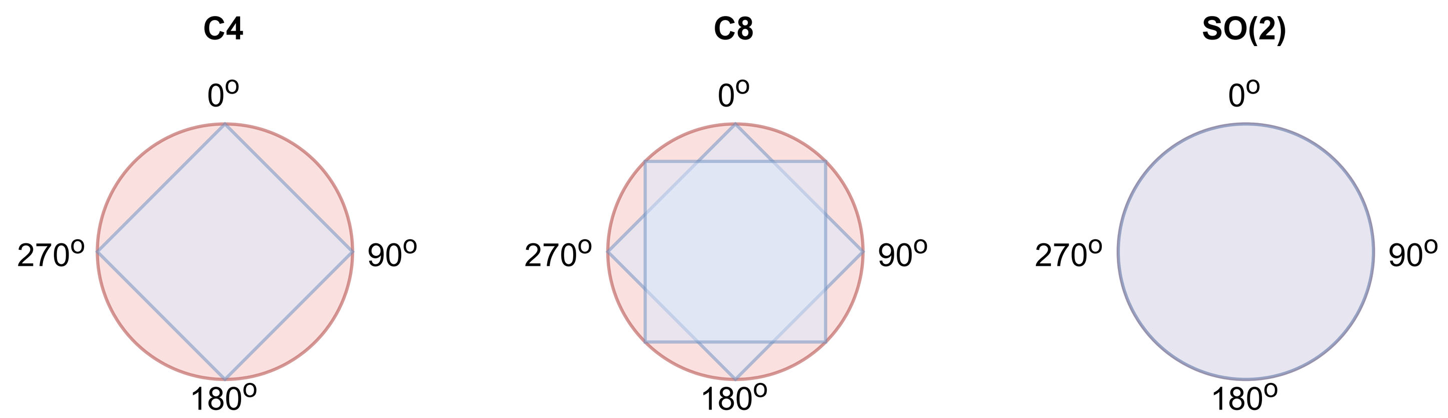


Figure 5. **Rotation groups.** A visualization of the finite C4 and C8, and the infinite SO(2) rotation groups.

Limitations

- Unusual behaviour at rotations of multiples of $\pi/4$: inherent locality?
- Introducing rotation equivariance comes at a price in terms of the number of parameters and inference time.
- Our study confines to pure convolution-based architecture. Equivariance for LoFTR?

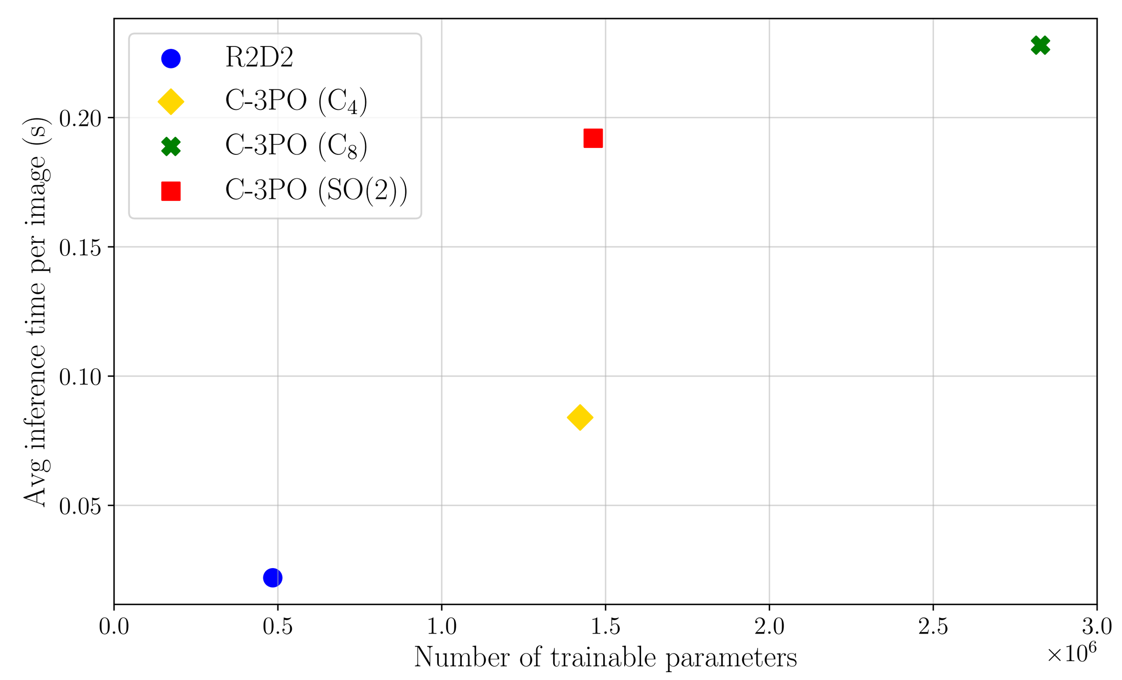


Figure 6. **Model Efficiency.** We show the computational efficiency of different network architectures, both in terms of inference time and number of trainable parameters.

Quantitative Results

To study the benefit of using rotation equivariant CNNs instead of standard CNNs, we compare performance in terms of *mean matching accuracy* (MMA) for input images from the HPatches dataset across rotations from 0° to 360° with an interval of 15°.

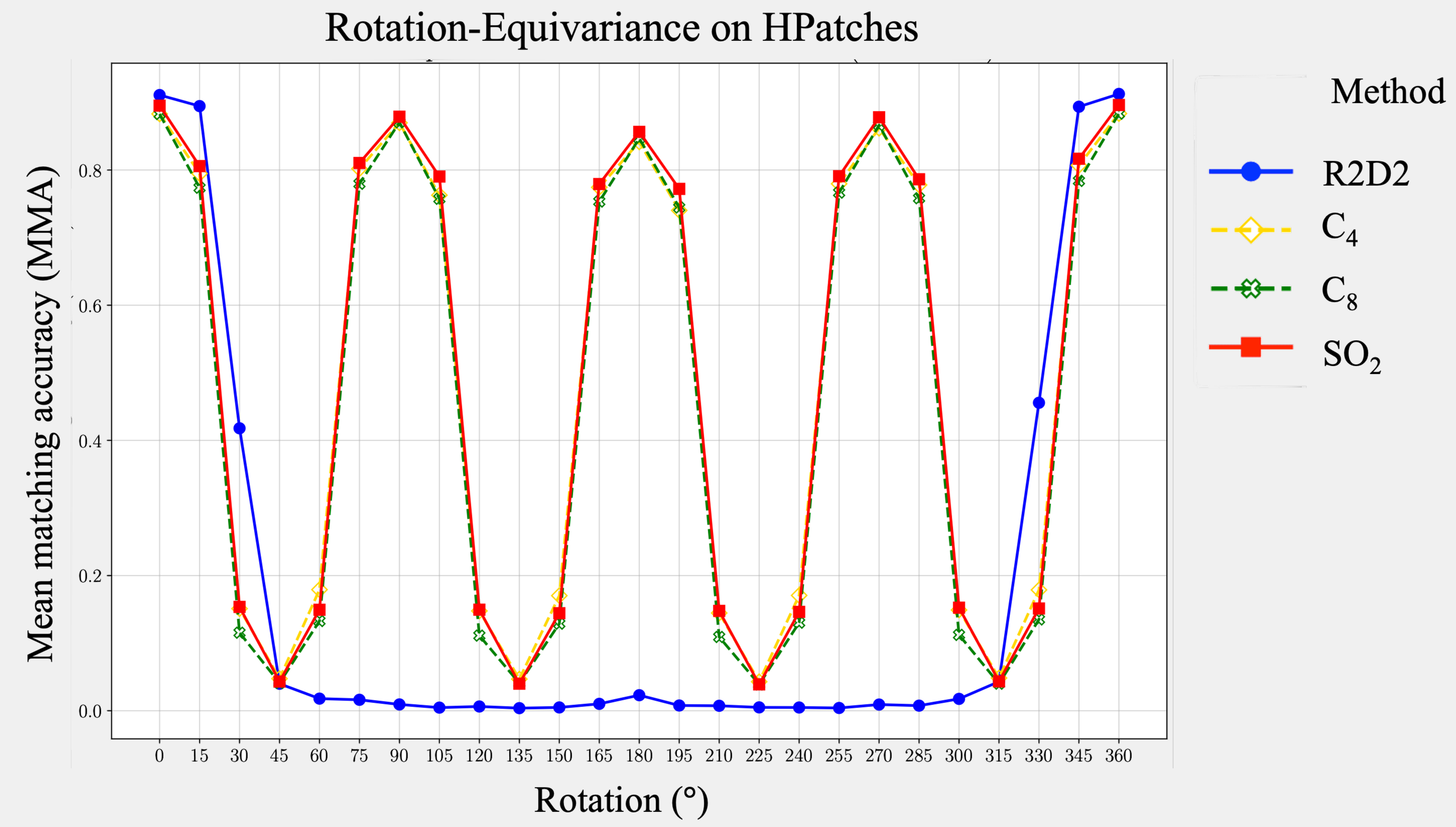


Figure 7. **Evaluation Rotation-Equivariance.** A comparison between R2D2 and the C-3PO models in terms of Mean Matching Accuracy.

Qualitative Analysis

To provide a more holistic understanding of the quantitative results, we show feature matching results on a sample image pair from the HPatches dataset.

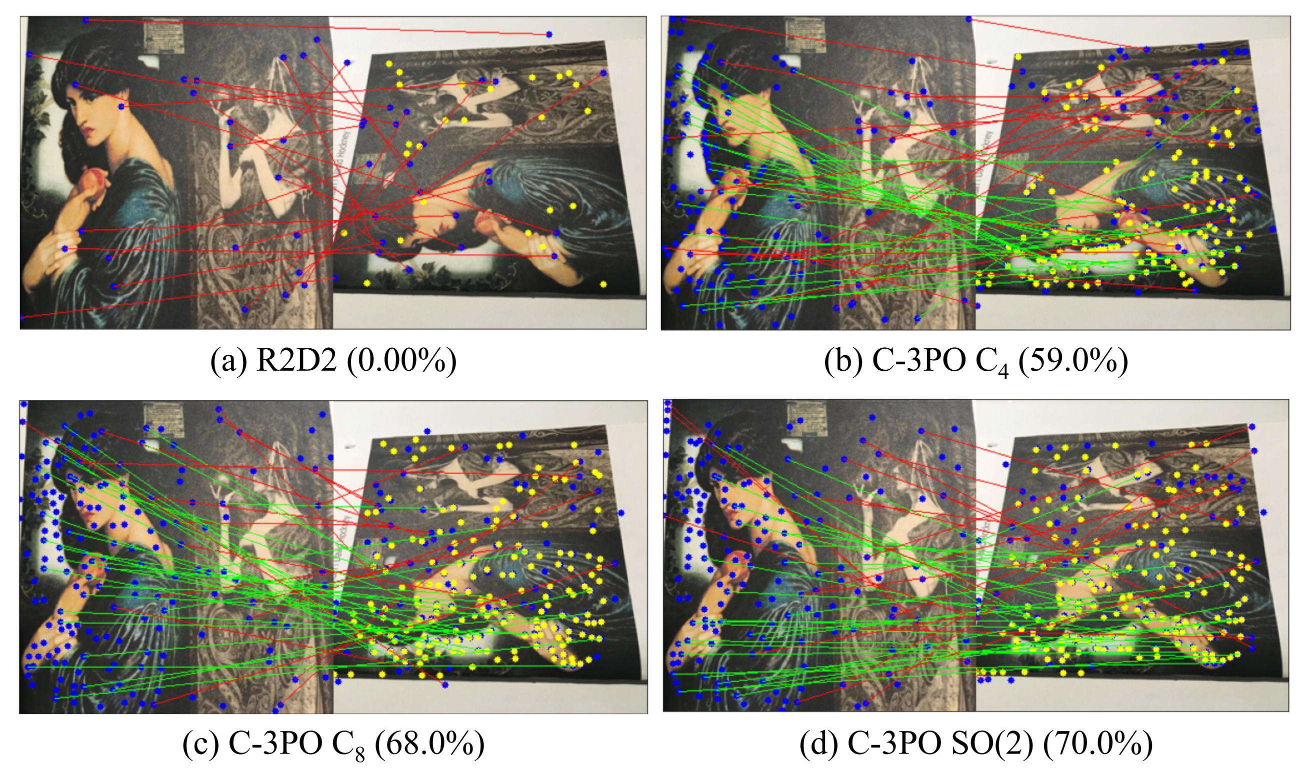
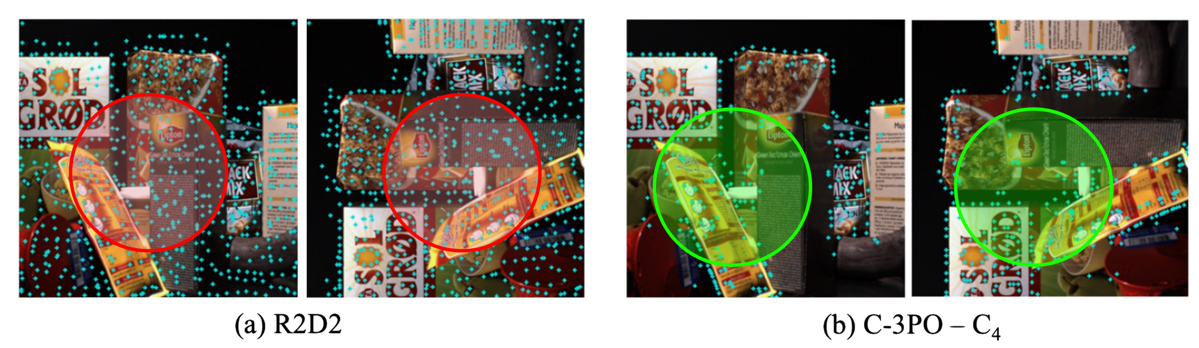


Figure 8. **Qualitative Matching Results.** Matches found for various models for a pair of images. The percentage in parenthesis shows the fraction of correct matches for each of the models for this particular image-pair. Blue points denote keypoint detected by the model. Yellow points on the target image denote the points in source image transformed by ground truth H. Correct matches are shown in green.

Equivariance → More robust keypoints?



SCAN ME

Acknowledgements: ELLIS Amsterdam & Master AI at UvA

piyush.bagad@student.uva.nl