# Fusing Local Similarities for Retrieval-based 3D Orientation Estimation of Unseen Objects

Chen Zhao, Yinlin Hu, Mathieu Salzmann

EPFL

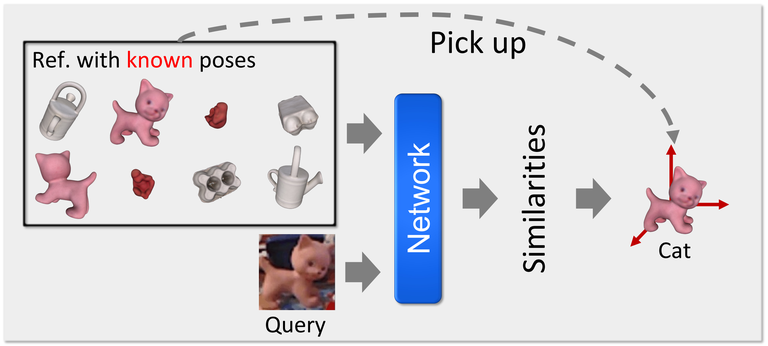
ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022 October 23-27, 2022

## What are Unseen Objects?

- Never observed during training

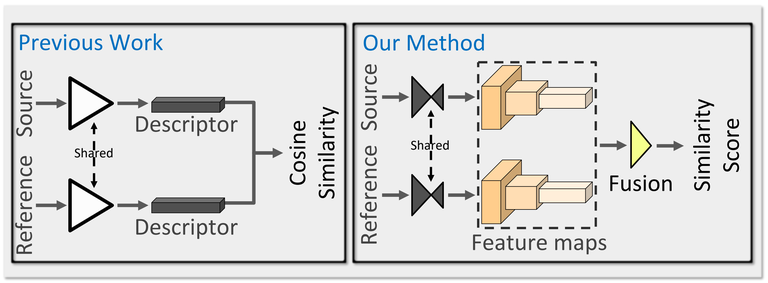

## 3D Orientation Estimation

- Given a previously unseen object, we predict the category label and 3D orientation by using a retrieval-based method.



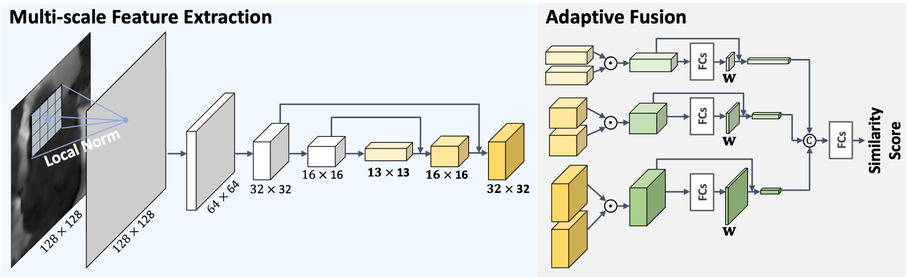

## Global Sim vs. Local Sim

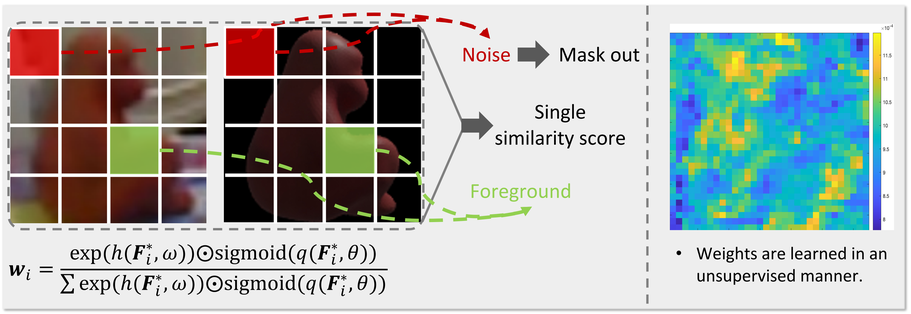
- We prevent the network from learning object-specific features by computing multi-scale local similarities between the query image and reference images.




## Network Architecture

- Given a pair of images, we use a siamese network to extract multi-scale local features.
- We present an adaptive fusion module to convert local feature similarities to a single image similarity score.




$$w_i = \frac{\exp(h(\boldsymbol{F}_i^*, \omega)) \odot \text{sigmoid}(q(\boldsymbol{F}_i^*, \theta))}{\sum \exp(h(\boldsymbol{F}_i^*, \omega)) \odot \text{sigmoid}(q(\boldsymbol{F}_i^*, \theta))}$$



- Weights are learned in an unsupervised manner.

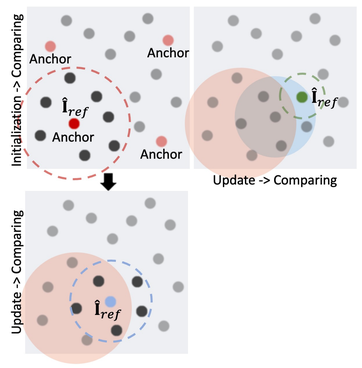
## Fast Retrieval

- A naïve image retrieval strategy compares query with every reference. Given N objects with R references each, the cost of O(NR) quickly becomes unaffordable as N and R increase.
- We design a fast retrieval strategy, which is around 60 times faster than the naïve one.

**Algorithm 1:** Fast Retrieval

**Input:** $\mathbf{I}_{src}, \mathcal{I}_{ref}, \mathcal{R}_{ref} = \{\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_R\}, k_{ac}, R$
**Output:** $\hat{\mathbf{I}}_{ref}, \mathbf{R}_{src}$

Sample $k_{ac}$ anchors from $\mathcal{I}_{ref}$ using FPS;
Estimate similarities using Eq. 2;
Initialize $\hat{\mathbf{I}}_{ref}$ as the most similar anchor;
$j = 1$;
**repeat**
  Define a search space around $\hat{\mathbf{I}}_{ref}$ with a radius of $\lfloor R/2^j \rfloor$;
  Compute anchors using FPS;
  Estimate similarities using Eq. 2;
  Update $\hat{\mathbf{I}}_{ref}$;
  $j$++;
**until** $\hat{\mathbf{I}}_{ref}$ *converges*;
Determine $\mathbf{R}_{src}$ as $\hat{\mathbf{R}}_{ref} \in \mathcal{R}_{ref}$.




## Quantitative Results

- We conduct experiments on three datasets, LineMOD, LineMOD-Occluded, and T-LESS.
- For LineMOD and LineMOD-O, we split images to three groups according to the contained objects. We use two groups as training data and the other one as testing data.
- For T-LESS, we test the methods using the models pretrained on LineMOD.

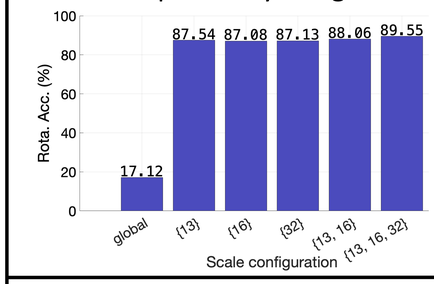
| Methods | Rotation Accuracy (%) on LineMOD | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Split #1 | | Split #2 | | Split #3 | | Mean | |
| | Seen | **Unseen** | Seen | **Unseen** | Seen | **Unseen** | Seen | **Unseen** |
| HOG | 38.89 | 40.17 | 28.21 | 30.74 | 31.02 | 28.48 | 32.71 | 33.13 |
| LD | 94.50 | 8.63 | 89.57 | 12.47 | 91.47 | 5.22 | 91.85 | 8.77 |
| NetVLAD | **100.00** | 36.11 | 98.66 | 20.33 | 99.35 | 23.38 | 99.34 | 26.61 |
| PFS | **100.00** | 6.31 | 99.19 | 6.65 | **99.46** | 5.54 | **99.55** | 6.17 |
| MPE | 91.94 | 38.96 | 66.47 | 41.46 | 87.72 | 61.62 | 82.04 | 47.35 |
| GDR-Net | 99.89 | 4.61 | **99.28** | 4.82 | 99.31 | 5.02 | 99.49 | 4.82 |
| **Ours** | 97.49 | **89.55** | 94.90 | **79.04** | 93.67 | **75.96** | 95.35 | **81.52** |

| Methods | Rotation Accuracy (%) on LineMOD-O | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Split #1 | | Split #2 | | Split #3 | | Mean | |
| | Seen | **Unseen** | Seen | **Unseen** | Seen | **Unseen** | Seen | **Unseen** |
| HOG | 0.60 | 0.60 | 0.18 | 0.18 | 5.25 | 5.25 | 2.01 | 2.01 |
| LD | 32.21 | 6.25 | 26.56 | 3.26 | 24.57 | 4.57 | 27.78 | 4.69 |
| NetVLAD | 51.60 | 24.32 | 42.20 | 18.05 | 36.56 | 18.84 | 43.45 | 20.40 |
| PFS | **71.40** | 6.25 | **60.88** | 13.15 | **54.67** | 4.68 | **62.32** | 8.73 |
| MPE | 40.47 | 22.56 | 27.31 | 5.20 | 35.06 | 18.22 | 34.28 | 15.33 |
| GDR-Net | 63.37 | 3.12 | 55.31 | 2.97 | 49.91 | 2.39 | 56.20 | 2.83 |
| **Ours** | 64.92 | **60.75** | 56.51 | **52.41** | 52.47 | **37.85** | 57.97 | **50.34** |

| Methods | Rotation Accuracy (%) on T-LESS | | | | | | |
|---|---|---|---|---|---|---|---|
| | HOG | LD | NetVLAD | PFS | MPE | GDR-Net | **Ours** |
| Acc. (%) | 74.22 | 24.19 | 56.46 | 17.92 | 66.88 | 11.89 | **78.73** |

## Ablation Studies

- The accuracy significantly decreases when local similarities are replaced by the global similarity in our framework.




- Greedy Search vs. our Fast Retrieval

| Methods | Greedy Search | **Fast Retrieval** |
|---|---|---|
| Acc. (%) | 95.93 | 89.55 |
| Time (s) | 30.74 | **0.42** |

## Contact Info

**E-mail:** chen.zhao@epfl.ch;
**Homepage:** https://sailor-z.github.io