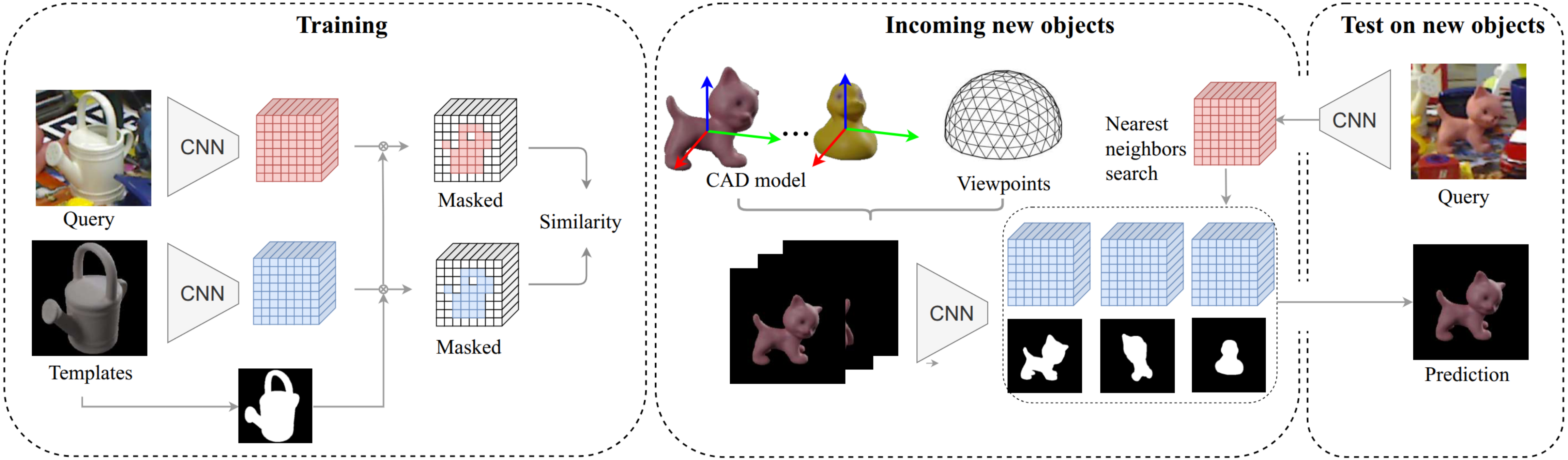
Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects and Robustness to Occlusions

Complete Proposed Process

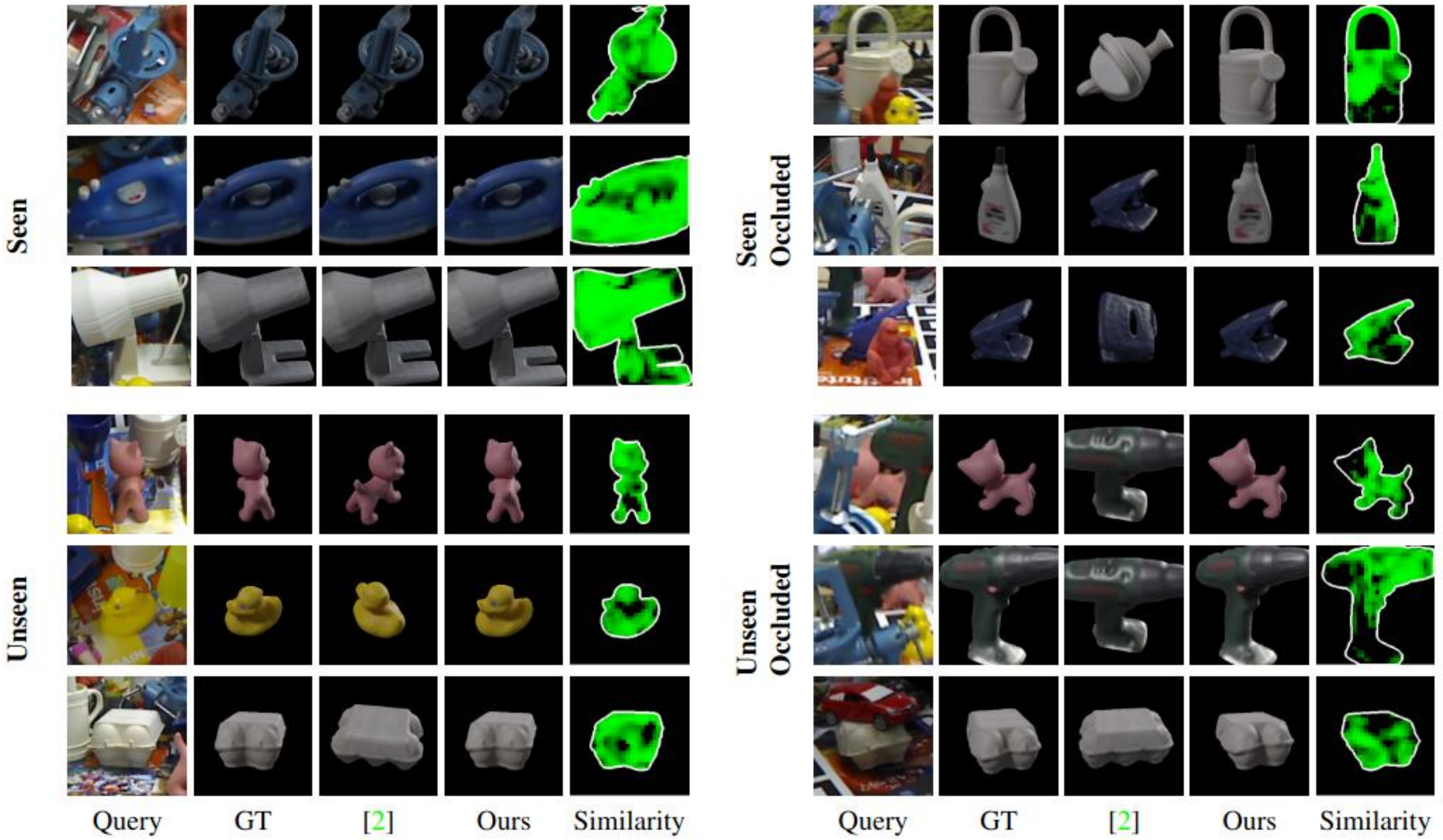


- Poses are extracted by matching an image to a large dataset of templates allowing masking of tested images
- Resulting template matches are then scored for similarity to the initial image allowing for more precise pose estimation
- Proposal aims to have a model which can be trained fast yet still work with occlusions and on objects with very different geometry

The Problem of 6D pose

- Goal is to find the pose of a specific object just based on the RGB image and corresponding CAD model
- Images are taken from a much larger cluttered image, and the CAD model is used to create a set of templates with known poses
- Templates are extracted following the guidelines provided by the respective dataset, and images are cropped to a uniform size

Improvements and Goal

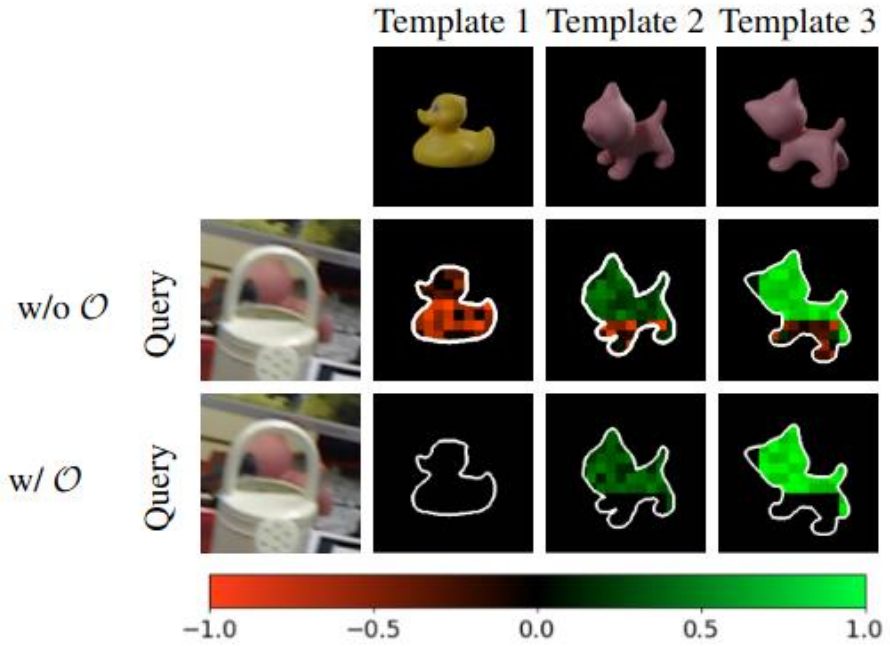


- Train a model to recognize local features instead of global features unlike predecessors
- This change will allow the model to work even on unseen images and corresponding templates
- The resulting model will also be more robust to occlusions and cluttered background

Method (Loss Function and Similarity)

$$\mathcal{L} = -\sum_{i=1}^{N} \log \frac{\exp\left(\mathrm{sim}(\overline{\mathbf{q}}_i, \overline{\mathbf{t}}_i)/\tau\right)}{\sum_{k=1}^{N} 1_{[k \neq i]} \exp\left(\mathrm{sim}(\overline{\mathbf{q}}_i, \overline{\mathbf{t}}_k)/\tau\right)},$$

$$\mathrm{sim}^*(\overline{\mathbf{q}}, \overline{\mathbf{t}}) = \frac{1}{|\mathcal{M}|} \sum_l \mathcal{M}^{(l)} \mathcal{O}^{(l)} \mathcal{S}\left(\overline{\mathbf{q}}^{(l)}, \overline{\mathbf{t}}^{(l)}\right)$$



- InfoNCE loss function takes into account positive and negative pairs, and adds a temperature parameter (0.1)
- Similarity function is modified from cosine similarity to include a template mask and occlusion mask
- Template mask is used to remove the background and occlusion mask is used to get rid of large areas where the image is dissimilar

Experimental - Setup

LINEMOD (LM) & Occlusion-LINEMOD (O-LM):

- **3 Splits for generalization testing:**
  - Split #1: Unseen = **Ape**, Benchvise, Camera, **Can**
  - Split #2: Unseen = **Cat, Driller, Duck, Eggbox**
  - Split #3: Unseen = **Glue, Holepuncher**, Iron, Lamp, Phone
- **Templates:** 301 per object
- **Metric:** Acc15 (correct object + pose error < 15°)

T-LESS:

- **Training:** Objects 1-18
- **Testing:** All 30 objects (including unseen 19-30)
- **Templates:** 92,232 (dense) or 21,672 (coarse) per object
- **Metric:** Recall at VSD < 0.3

Qualitative Results

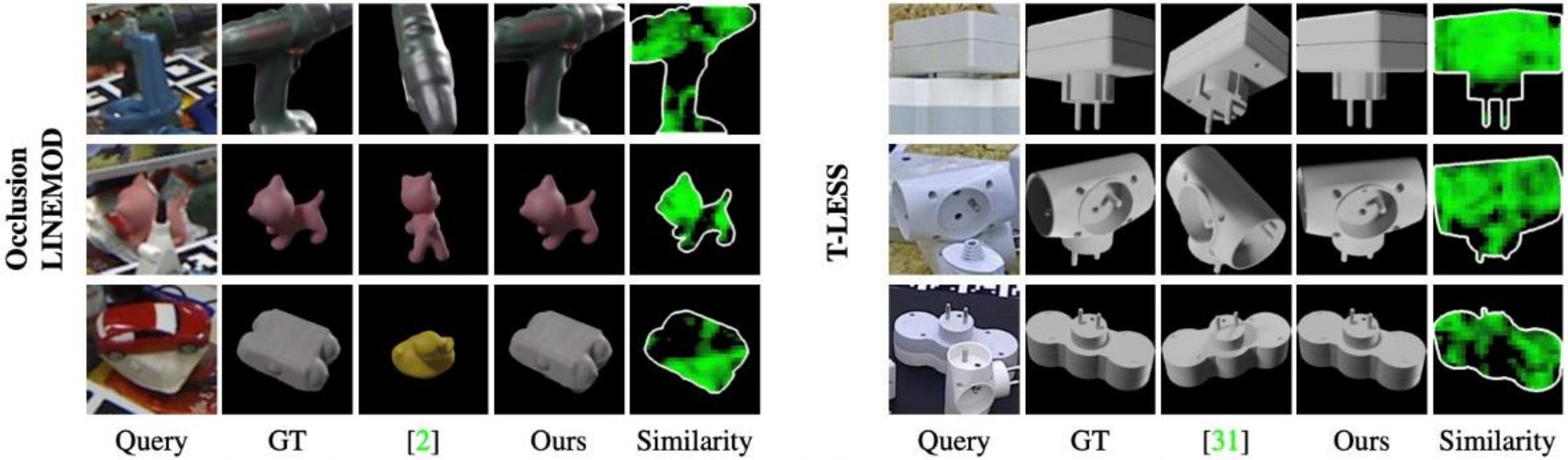


Figure 5: **Qualitative results on unseen objects** of Occlusion-LINEMOD (left) and T-LESS (right). Our method retrieves the correct template and pose while [2, 31] fails on unseen objects, particularly in the presence of occlusion.

Quantitative Results

Method	Backbone	Features	Loss	Seen LM				Seen O-LM				Unseen LM				Unseen O-LM			
				#1	#2	#3	Avg.	#1	#2	#3	Avg.	#1	#2	#3	Avg.	#1	#2	#3	Avg.
[39]	Base [39]	Global	[39]	87.0	83.1	85.1	85.0	19.2	23.1	15.0	19.1	13.2	15.5	18.2	15.2	9.3	5.1	5.1	6.5
[39]	Base [39]	Global	Eq. (2)	95.2	95.3	95.4	95.3	19.6	25.3	16.1	20.3	13.3	17.0	20.5	16.9	8.2	6.4	6.7	7.1
[2]	Base [39]	Global	[2]	89.2	85.4	83.3	86.3	18.3	21.9	17.6	19.5	14.1	16.3	19.7	16.7	8.2	7.5	7.6	7.8
[2]	Base [39]	Global	Eq. (2)	96.3	95.2	96.5	96.0	18.3	23.1	15.8	19.1	11.5	17.7	17.2	15.5	7.1	6.5	6.5	6.7
Ours	Base [39]	Local	[39]	84.8	85.5	86.3	85.5	50.1	51.3	42.2	47.9	69.6	63.2	46.2	59.7	35.3	34.3	44.2	37.9
Ours	Base [39]	Local	Eq. (2)	95.6	96.9	92.0	94.8	68.9	71.0	57.7	65.8	78.8	82.5	64.1	75.1	42.2	57.1	59.8	53.0
[39]	ResNet50 [11]	Global	Eq. (2)	98.8	96.9	98.8	98.1	66.7	73.2	62.7	67.5	42.2	43.7	49.4	45.1	22.3	22.5	45.9	29.9
[2]	ResNet50 [11]	Global	Eq. (2)	96.9	97.1	94.5	96.1	63.6	71.8	58.9	64.7	39.9	44.9	48.3	44.3	15.5	21.8	50.2	29.1
Ours	ResNet50 [11]	Local	Eq. (2)	**99.3**	**99.0**	**99.2**	**99.1**	**77.3**	**84.1**	**76.8**	**79.4**	**94.4**	**97.4**	**88.7**	**93.5**	**71.4**	**72.7**	**85.3**	**76.3**

Table 2: **Comparison of our method with [39] and [2]** on seen and unseen objects of LM and O-LM under the three different splits detailed at the beginning of Section 4.1. We report Acc15 ↑, the accuracy of predicting correctly the object identity *and* its pose with an error less than 15 degrees. We are on par on the "easy" case and outperform them by a large margin on the 3 other configurations. Using the InfoNCE loss rather than the loss from [2] brings some improvement, but the main improvement comes from our approach based on local features.

Limitations/Possible Improvements

Figure 7: The “Cat” object is often barely visible in the test images of Occluded-LINEMOD, resulting in large errors.

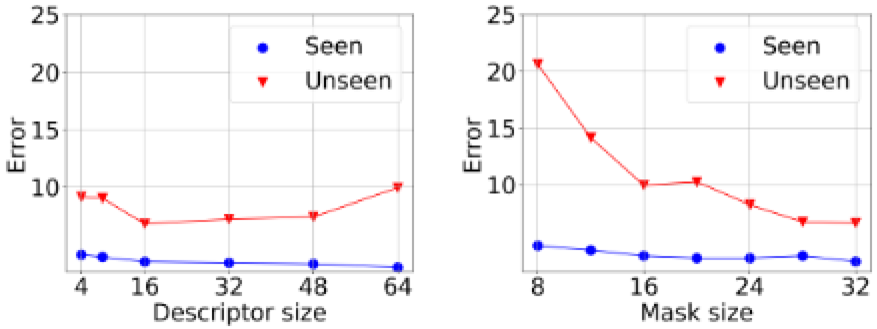


Figure 6: **Influence of the local feature dimension C and of the resolution of the local features and masks.** Using a good resolution is much more important than using high-dimensional local features as this allows discarding background more precisely when computing the similarity score.

- Extremely heavy occlusions remain challenging as seen above
- Above outline is not an output of the paper

Demo