



# Coupled Iterative Refinement for 6D Multi-Object Pose Estimation

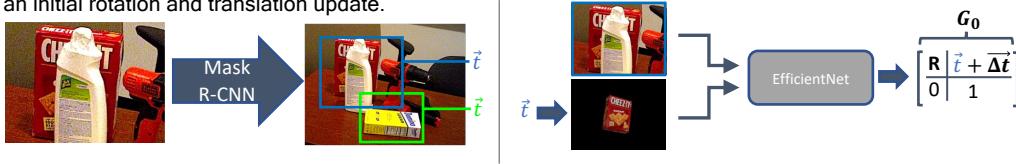
Lahav Lipson, Zachary Teed, Ankit Goyal and Jia Deng, Princeton University

Code: [github.com/princeton-vl/Coupled-Iterative-Refinement](https://github.com/princeton-vl/Coupled-Iterative-Refinement)



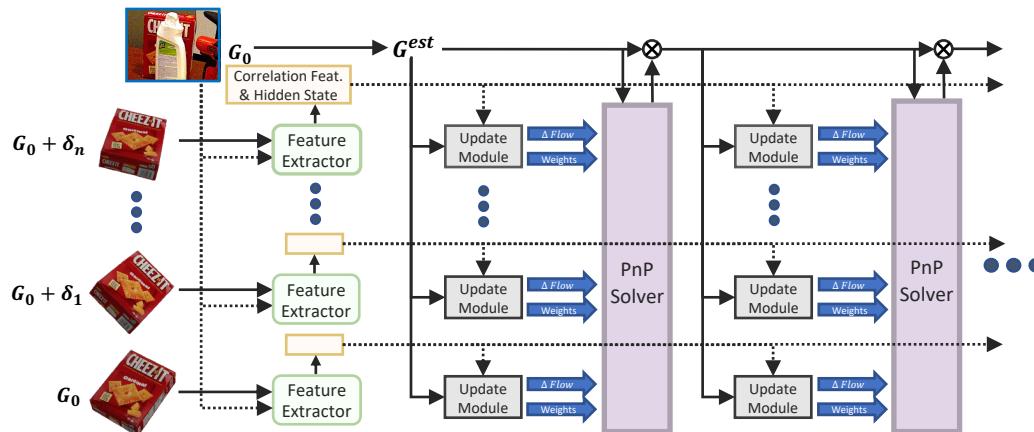
## Stages 1 & 2: Detection and Pose Initialization

We use a Mask R-CNN to produce a set of bounding boxes and labels for object candidates. We then render an image of the object at a translation inferred from the bounding box. Lastly, we directly regress an initial rotation and translation update.



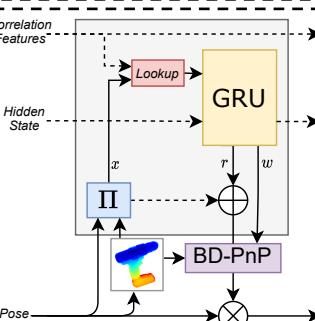
## Stage 3: Iterative Pose Updates

To refine a pose estimate, we use the pose to induce optical flow between the input image and the object rendered at, and around, our current pose estimate  $G_0$ . An update module predicts revisions to the optical flow and pixelwise confidence weights. We then solve for a pose update which explains these flow revisions and confidence weights. This entire process is repeated until the pose converges to a good solution.



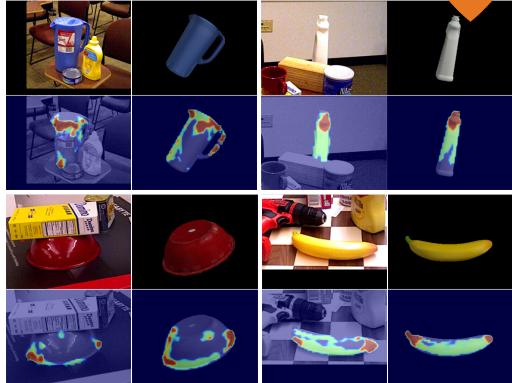
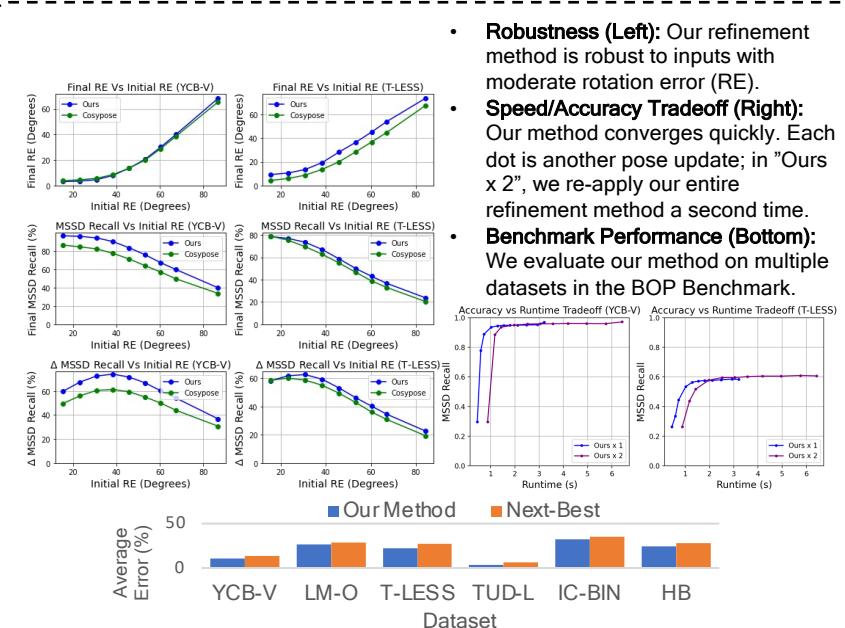
## Coupled Refinement

The current pose estimate induces a flow-field  $x$  between the image and render. We "lookup" correlation features using this flow, from which a GRU predicts flow revisions  $r$  and confidence weights  $w$ . The solver produces a pose update which explains the revisions and weights. This pose update is applied using retraction of the SE3 manifold.

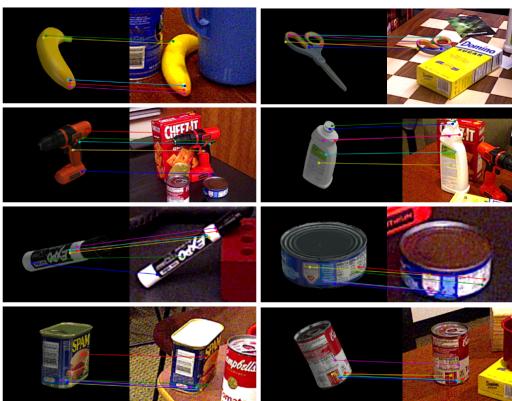


## The Solver

The solver (purple) minimizes the Mahalanobis distance defined by  $r$  and  $w$ . The solver also tries to find a pose which aligns the render depth and sensor depth. If the sensor depth is not provided, it is generated from the current pose estimate (See Figure ←). Our solver is fully differentiable, which enables our method to learn  $w$ .



**Predicted confidence weights.** The "Hot" pixels indicate surface features that are highly weighted in the pose optimization step. Our method has low confidence over texture-less regions and high confidence over textured ones, over thin structures, on edges, and on occluded areas.



**Predicted high-confidence matches.** In this figure, we apply non-max suppression to the confidence weights using a 5-pixel radius and then visualize the most confident predicted correspondences.