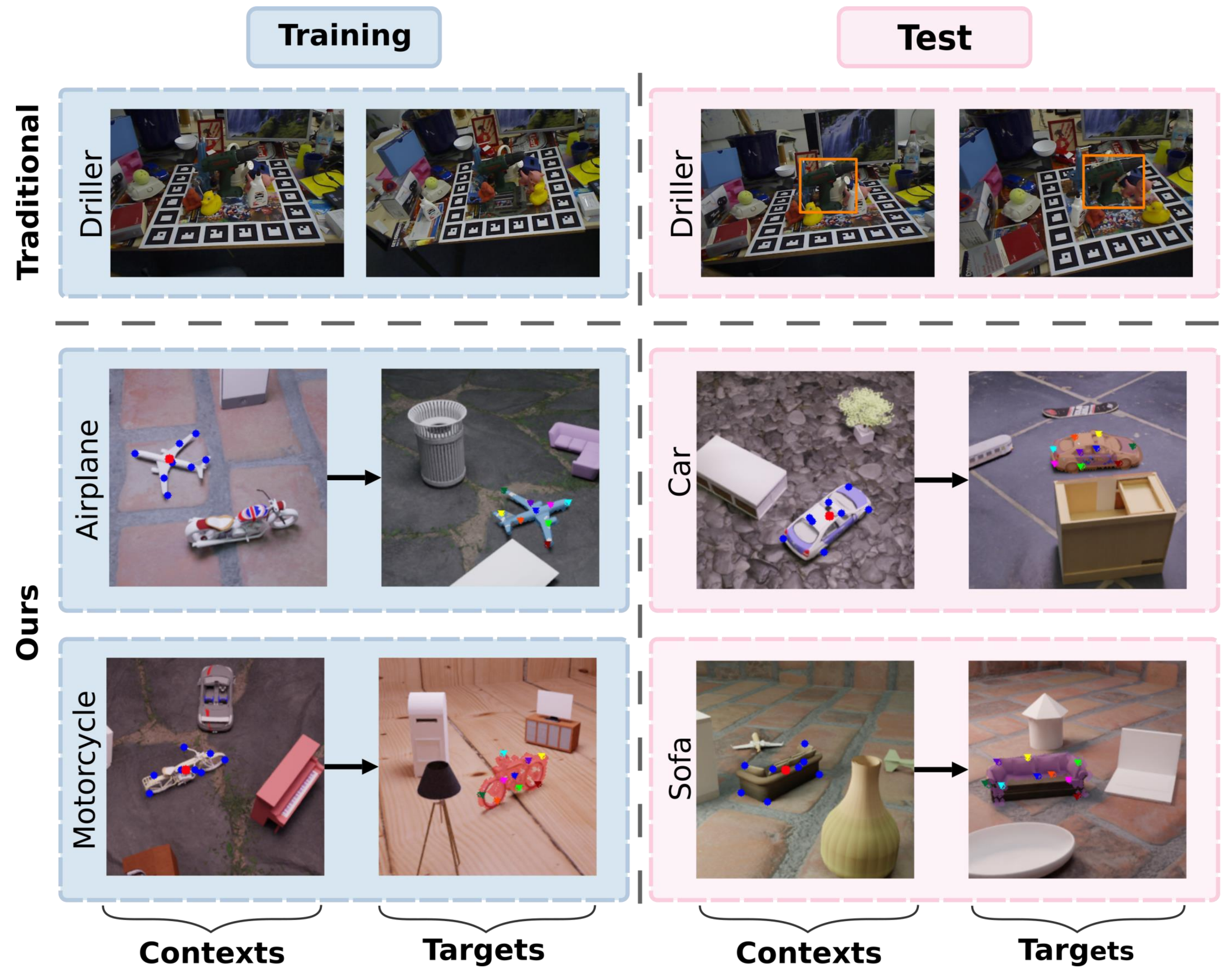
Category-Agnostic 6D Pose Estimation with Conditional Neural Processes

Yumeng Li*[1], Ning Gao*[1,2], Hanna Ziesche[1], Gerhard Neumann[2]

*Equal Contribution, [1]Bosch Center for Artificial Intelligence, [2]Autonomous Learning Robots, KIT

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

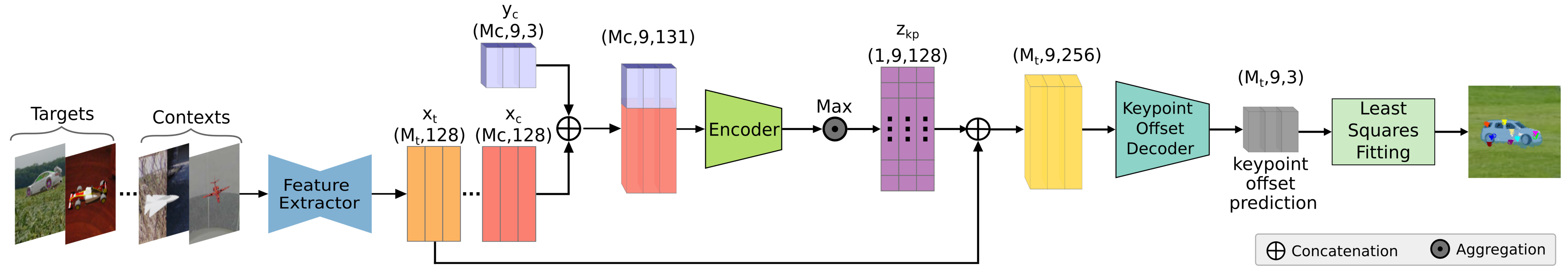
1 Contribution



Tradition Methods Instance-level 6D pose estimation methods are prone to overfit to specific objects and suffer from poor generalization.

Our Method We purpose a meta-learning based cross-category level 6D pose approach. The core idea lies in Conditional Neural Processes (CNPs) based meta-learner that extracts object-centric representations in a category-agnostic way.

2 Method Overview



Given an RGB-D image, the goal of 6D pose estimation is to compute rigid transformation $[R;t]$ from the object coordinates to the camera coordinates. We build on keypoint-based methods, which can be considered in three stages: (i) feature extraction, (ii) keypoint detection, (iii) pose fitting.

i. **Feature Extraction:** we rely on the fusion network FFB6D[1], where the output is per-point features of sampled seed points from the input depth images.

ii. **CNPs[2]-based Keypoint Detection:** the meta-learner encodes and aggregates the features of several context images into latent variables z_{kp}. The decoder predicts per-point offsets for each keypoint based on the target features and the keypoint latent variables z_{kp}.

$$r_i^u = h_\theta(x_{c,i} \oplus y_{c,i}^u),\ \ i = 1, ..., M_c,\ \ u = 1, ..., M_k,$$

$$z_{kp}^u = \max_{i=1}^{M_c}(r_i^u),\ \ u = 1, ..., M_k,$$

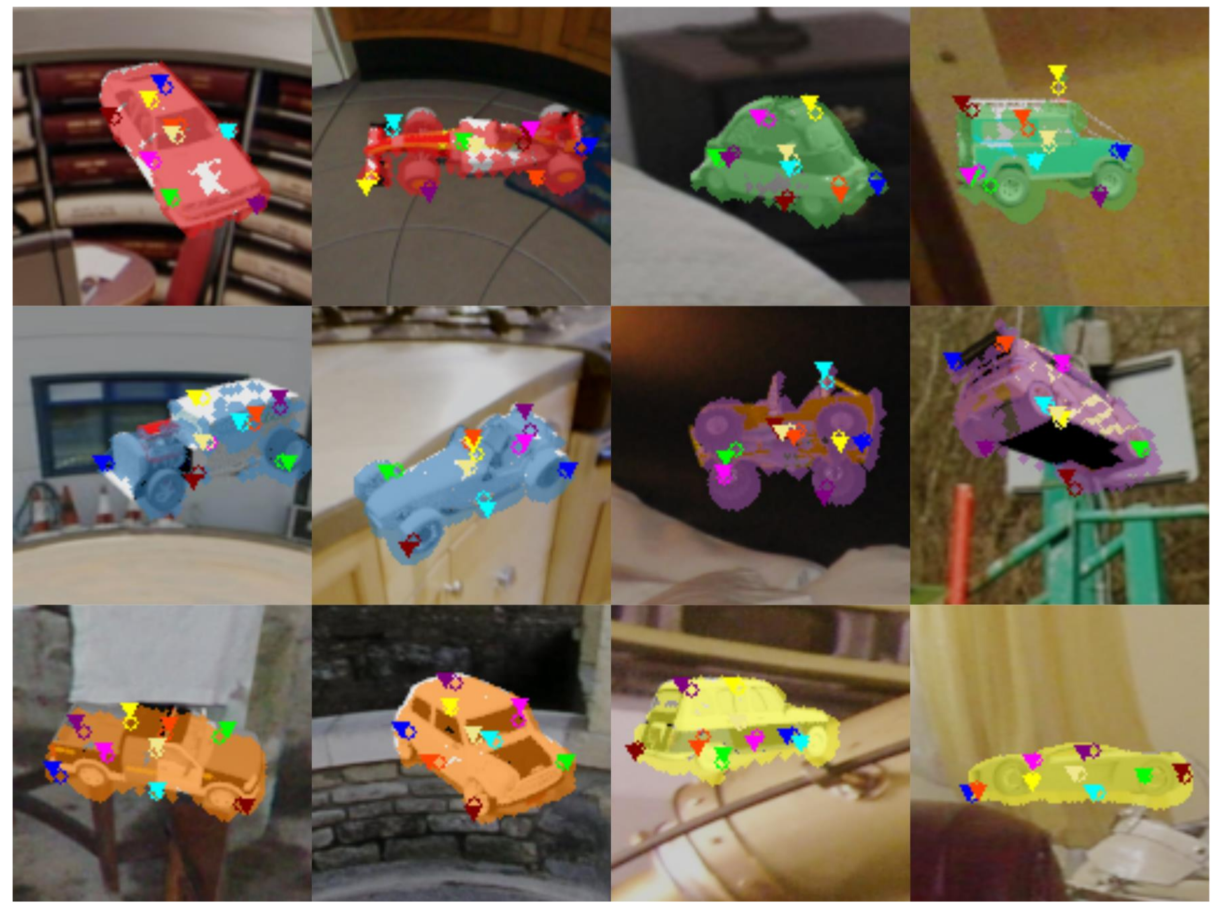
$$y_{of,i}^u = g_{\mathcal{K}}(x_{t,i} \oplus z_{kp}^u), i = 1, ..., M_t,\ \ u = 1, ..., M_k,$$

iii. **Pose Fitting:** given the pre-defined 3D keypoints in the object coordinates $\{p_i\}_{i=1}^{M_k}$ and keypoints prediction in the camera coordinates $\{p_i^*\}_{i=1}^{M_k}$, 6D pose can be computed by solving a least-squares fitting problem:

$$L_{lsf} = \sum_{i=1}^{M_k} \|p_i^* - (R \cdot p_i + t)\|^2 .$$

References: [1] Yisheng He, et al. "FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation." CVPR 2021
[2] Marta Garnelo, et al. "Conditional Neural Processes." ICML 2018

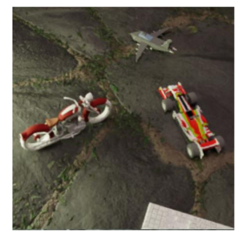
3 Results



To test the inter-category performance, the model is trained and tested on car category. The ADD-0.1d accuracy on 200 **unseen** new car objects reaches *96.7%*. The qualitative results are presented above.

Furthermore, we train our model on 20 categories. The ADD-0.1d accuracy of **novel** objects from trained intra-category and cross-category are *81.9%* and *59.0%* respectively.

4 Outlook



- ❑ Evaluation on photorealistic dataset without and with occlusion
- ❑ Evaluation on real-world dataset
- ❑ Improve rendering pipeline