DeepRob

Seminar 4
Dense Descriptors, Category-level Representations
University of Michigan and University of Minnesota

This Week: Rigid Body Objects

- Seminar 3: Object Pose, Geometry, SDF, Implicit Surfaces
    1. SUM: Sequential scene understanding and manipulation, Sui et al., 2017
    2. iSDF: Real-Time Neural Signed Distance Fields for Robot Perception, Oriz et al., 2022

- Seminar 4: Dense Descriptors, Category-level Representations
    1. Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation, Florence et al., 2018
    2. Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation, Wang et al., 2019
    3. kPAM: KeyPoint Affordances for Category-Level Robotic Manipulation, Manuelli et al., 2019
    4. Single-Stage Keypoint-Based Category-Level Object Pose Estimation from an RGB Image, Lin et al., 2022

Today: Dense Descriptors, Category-level Representations

- Seminar 3: Object Pose, Geometry, SDF, Implicit Surfaces
  1. SUM: Sequential scene understanding and manipulation, Sui et al., 2017
  2. iSDF: Real-Time Neural Signed Distance Fields for Robot Perception, Oriz et al., 2022

- Seminar 4: Dense Descriptors, Category-level Representations
  1. Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation, Florence et al., 2018
  2. Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation, Wang et al., 2019
  3. kPAM: KeyPoint Affordances for Category-Level Robotic Manipulation, Manuelli et al., 2019
  4. Single-Stage Keypoint-Based Category-Level Object Pose Estimation from an RGB Image, Lin et al., 2022

Single-Stage Keypoint-Based Category Level Object Pose Estimation from an RGB Image

By: Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A. Vela, Stan Birchfield

Presented by: Brandon Apodaca, Yu Zhu

Autonomous Robotics

How can a robot autonomously set goals and formulate plans to achieve them?

1. Identify objects and their poses in the environment
2. Create a goal and formulate a plan
3. Execute plan

Semantic Scene Understanding

Instance-level:

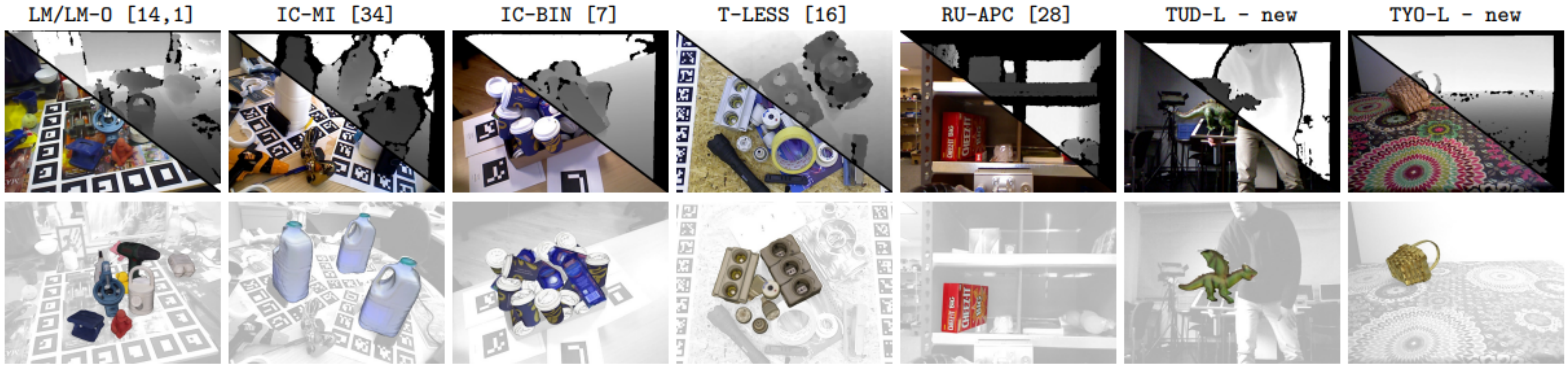
- Determine specific objects
- Not easily scalable
- Require large number of detectors

Category-level:

- Generalized object identification
- 3D CAD models are not required

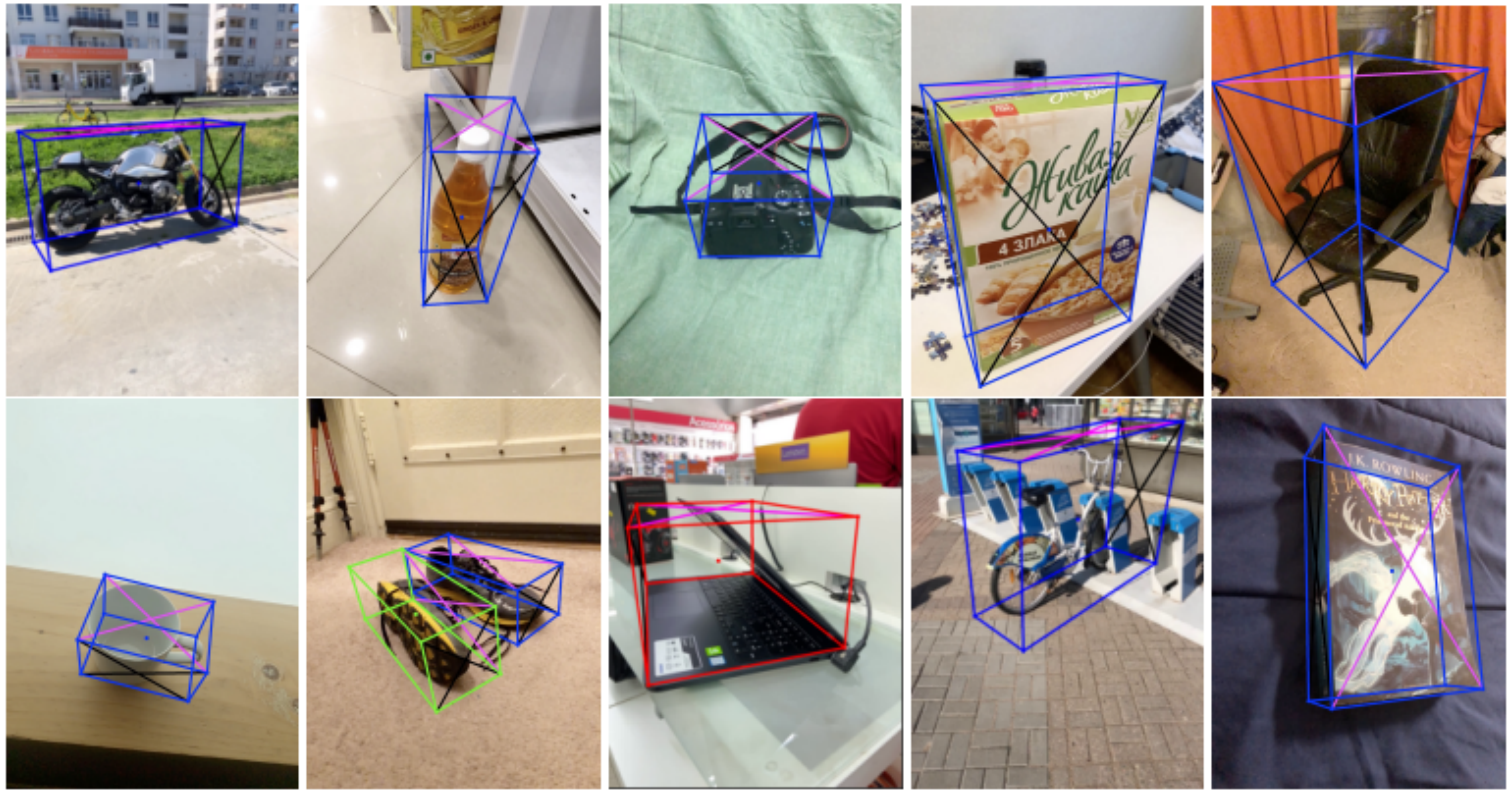
Existing Pose Estimation Methods: Instance-Level

- Template matching methods align known 3D CAD models to observed 3D point clouds [1] or 2D images [2]
- Regression-based methods establish 2D-3D correspondence by regressing the 6 DoF pose [3] or predict the image coordinate of projected keypoints [4]

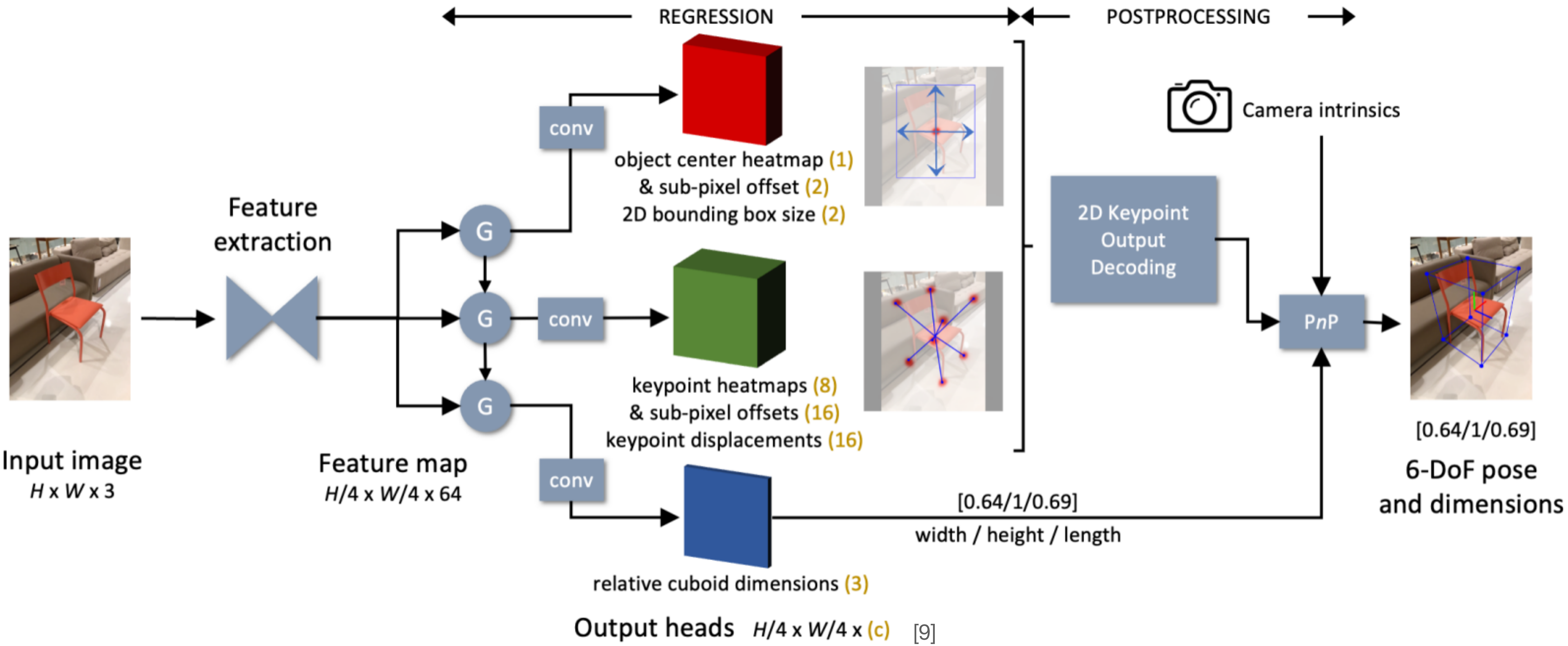


Existing Pose Estimation Methods: Categroy-Level

- Normalized coordinate space (NOCS) requires 3D meshes for training [5]
- Other methods reply on RGBD image [6] to match features
- Existing monocular methods have room for improvement [7, 8]



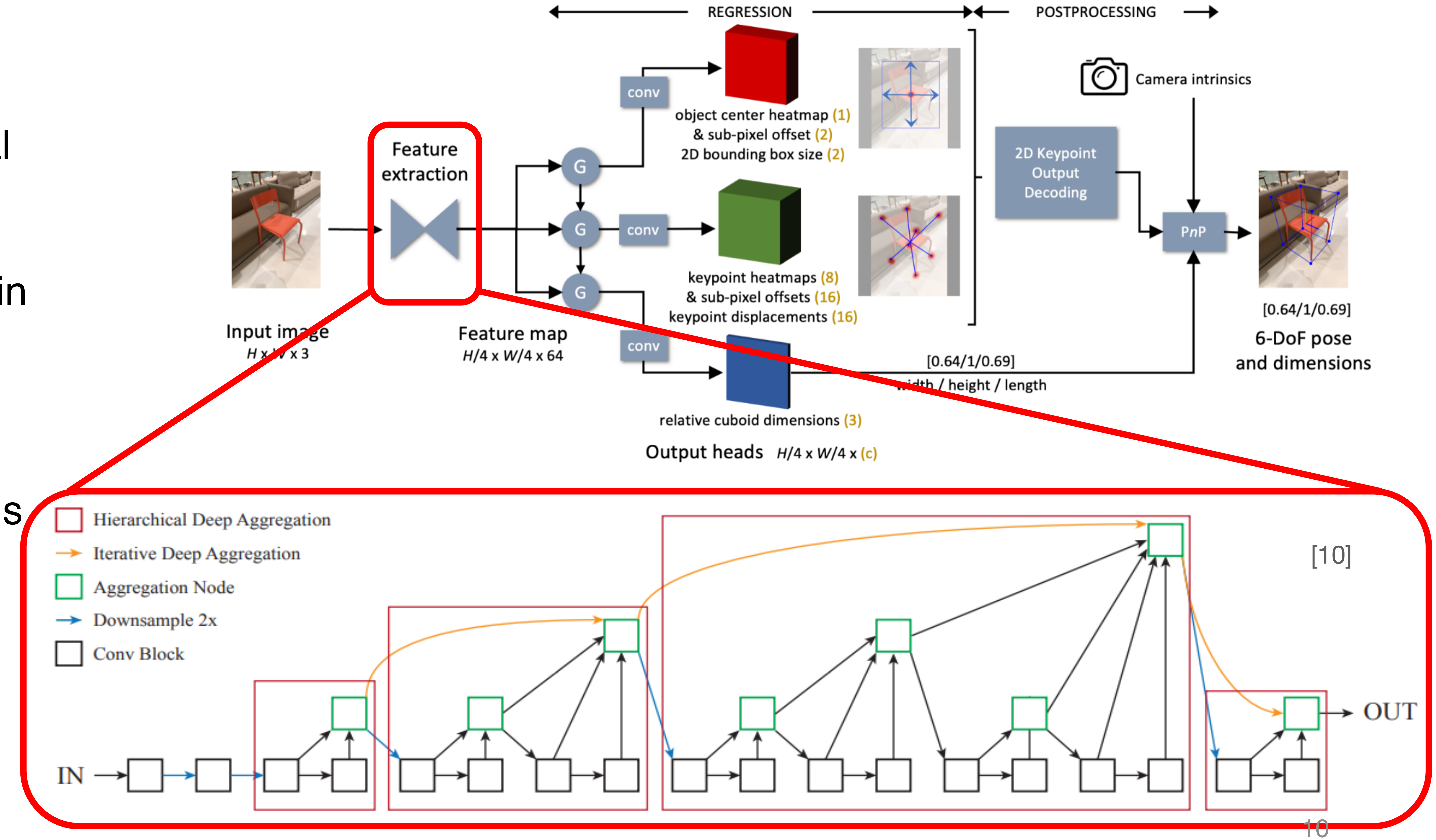
Complete Network



[9]

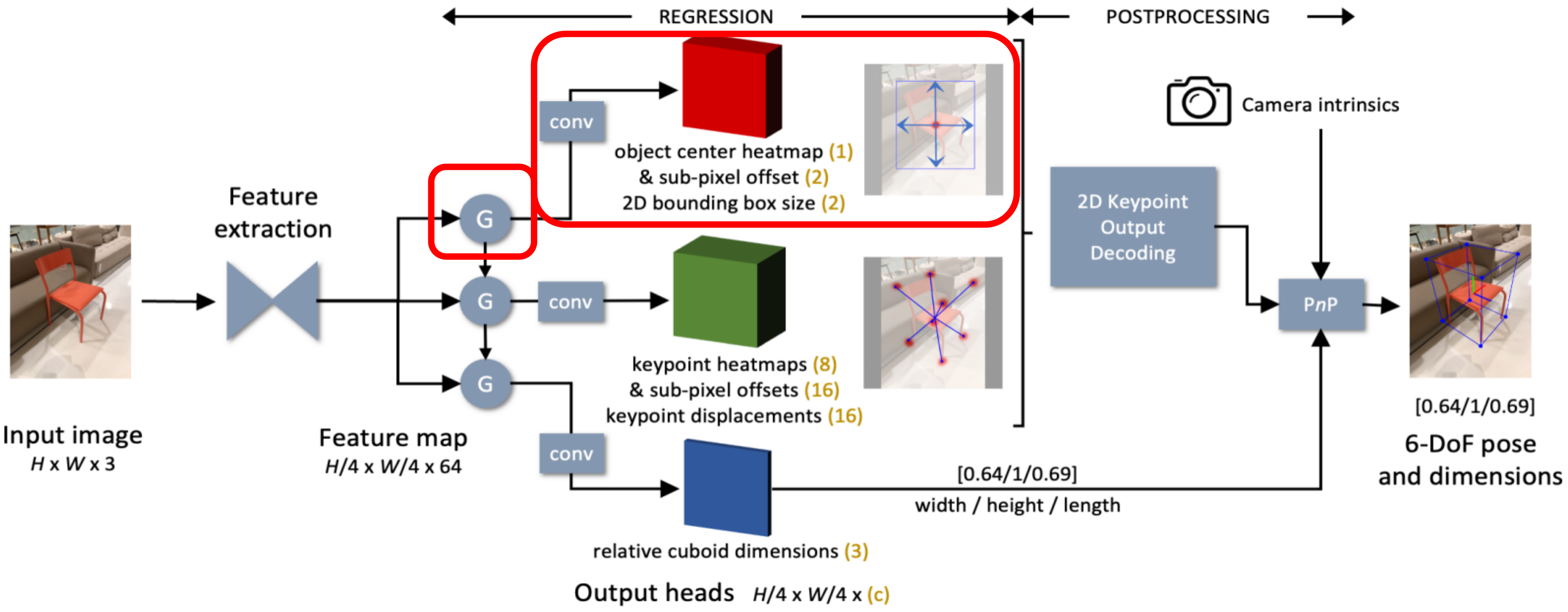
Feature Extraction via DLA-34

- Produce multiple intermediate feature maps of different spatial resolutions
- Iterative connections join neighboring stages to refine representation
- Hierarchical connections to better propagate features and gradients



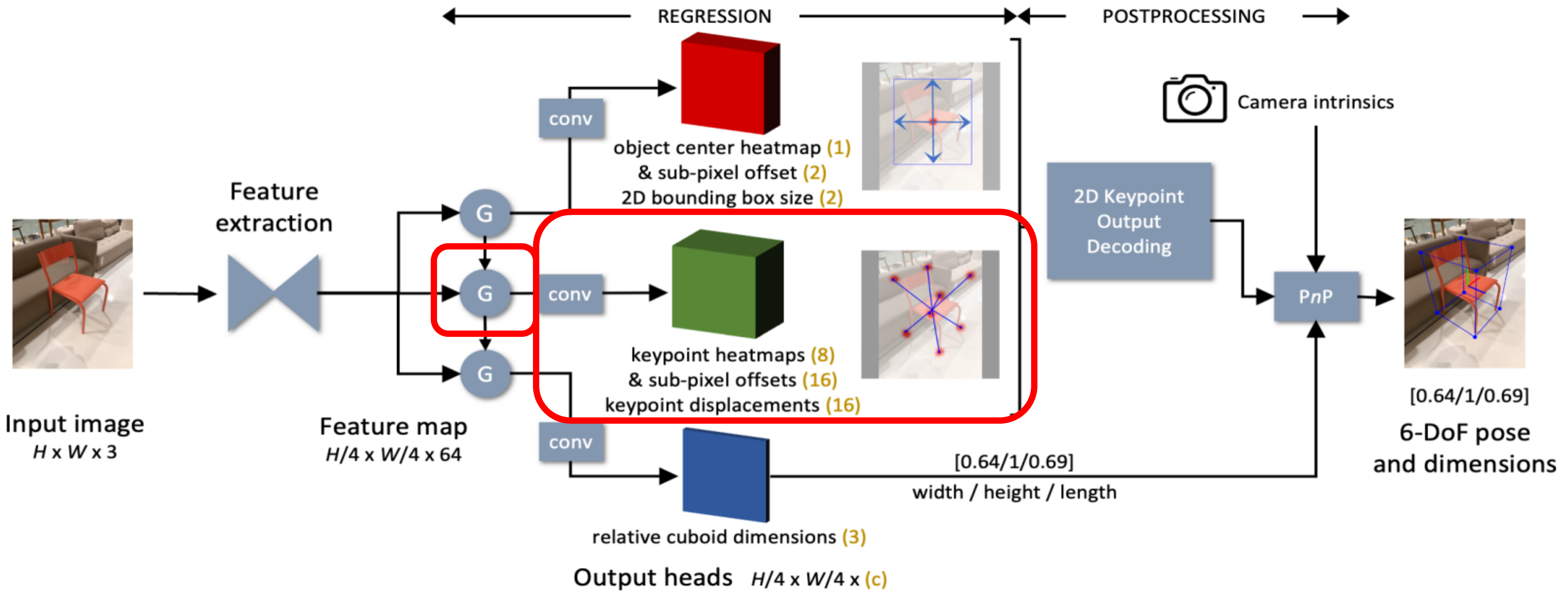
Object Detection Branch

- Generate a heatmap to indicate the centroid of objects
- Output the object center sub-pixel offset to reduce discretization error



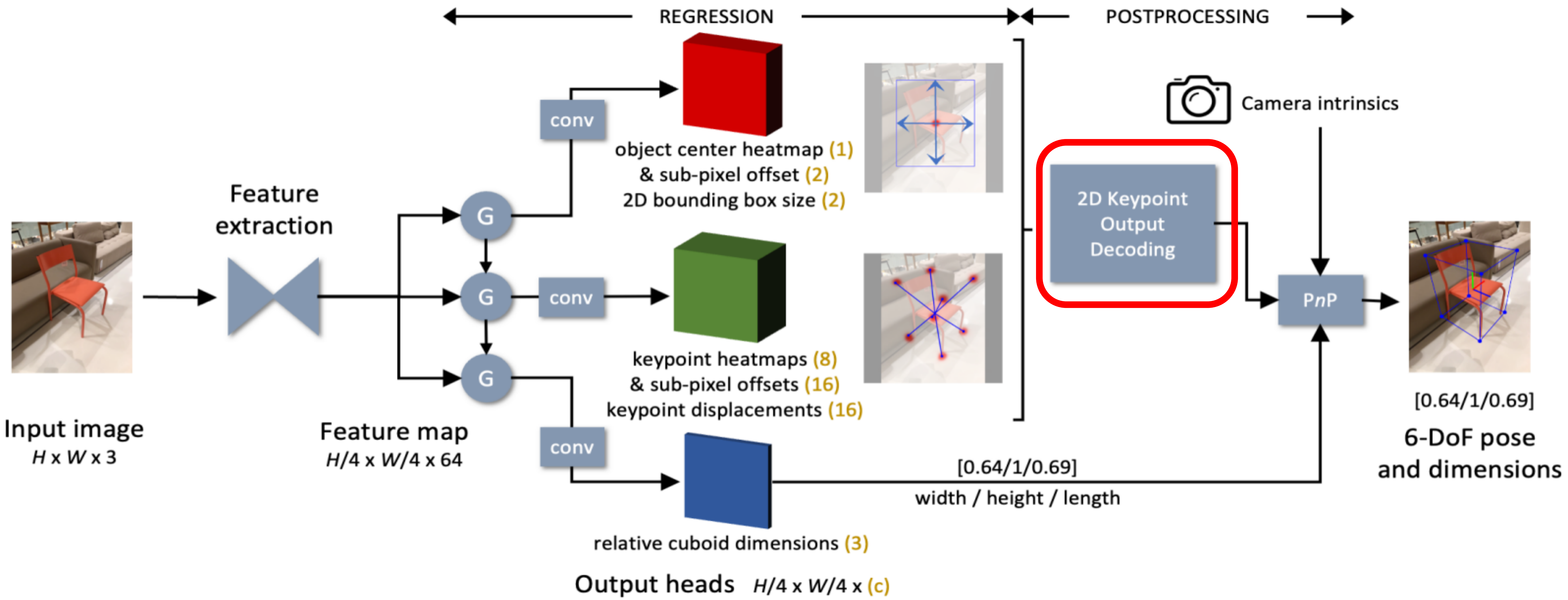
Keypoint Detection Branch

- 8 Keypoint heatmaps to indicate the location of keypoints
- Output the keypoint sub-pixel offset to mitigate discretization error
- Generate displacement vectors from bounding box center point



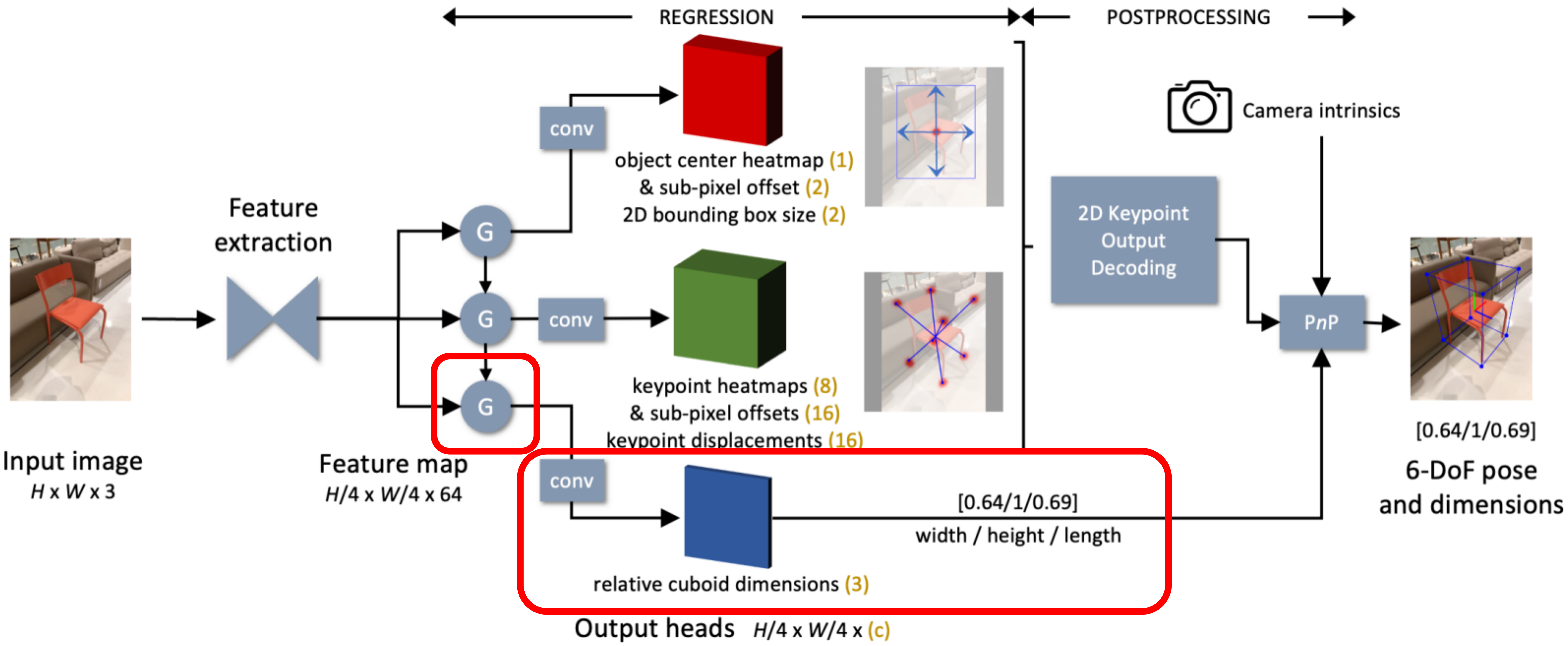
2D Keypoint Output Decoding

- Find high confidence peaks in heatmaps to determine object center or keypoints
- Displacement-based keypoints are given by 2D x-y displacements under the center point
- Sub-pixel offsets to adjust the keypoint locations



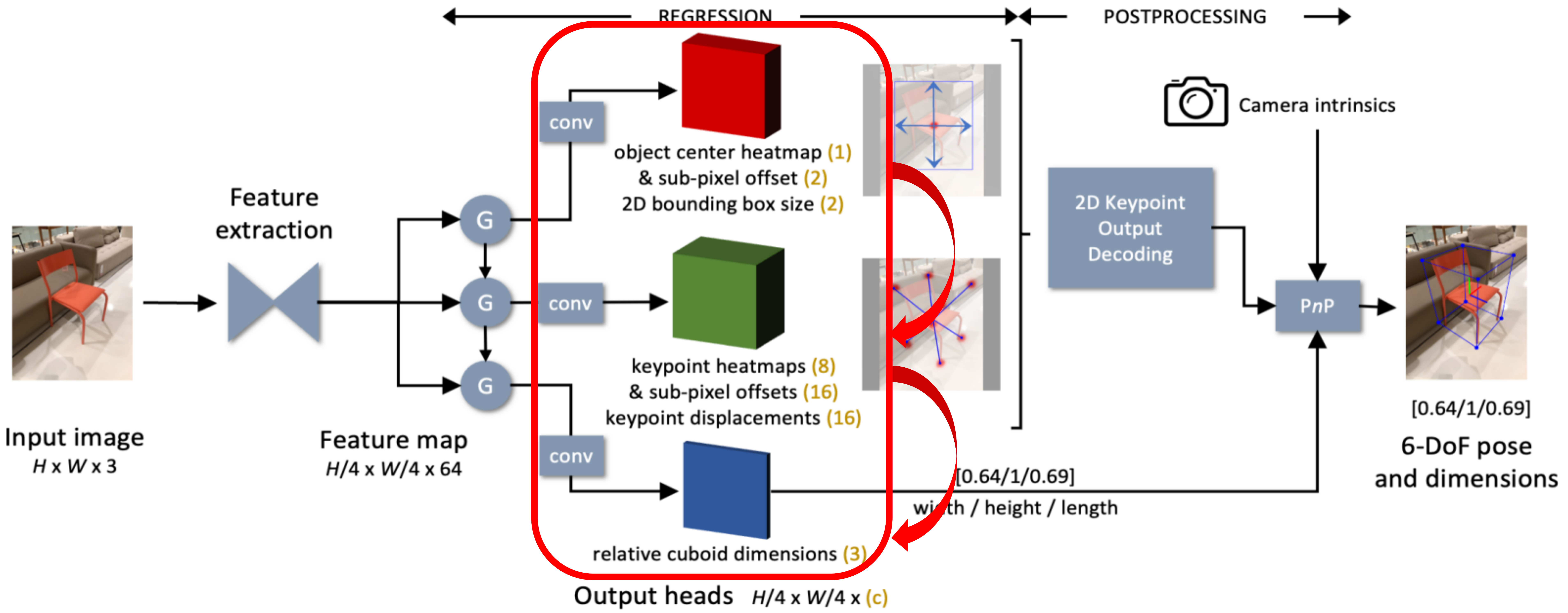
Cuboid Dimensions Branch

- Output cuboid aspect ratio (x/y, 1, z/y) with y axis being the up axis

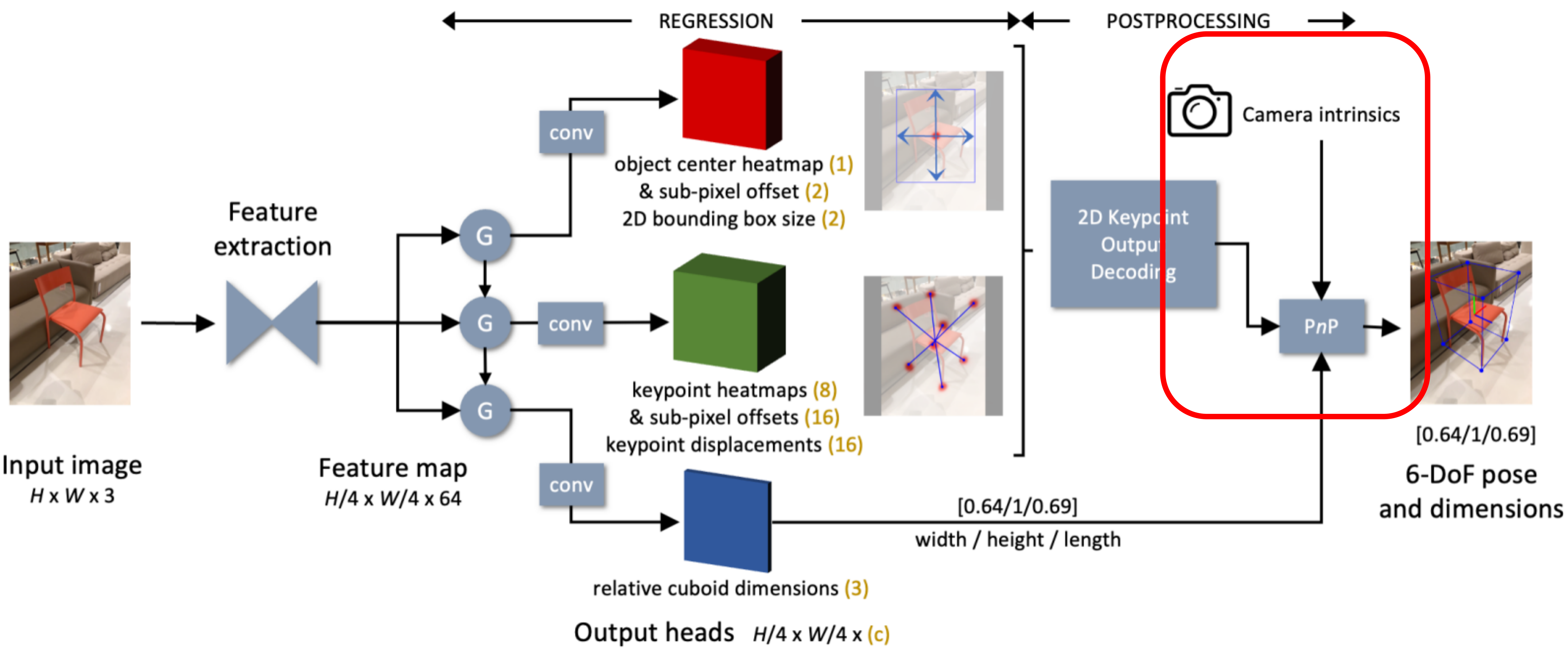


convGRU for Sequential Feature Association

- Motivation: to help the prediction of last group (dimension branch)
- Use a recurrent neural network for propagating information from earlier task



Off-the-shelf PnP algorithm yields 6-DoF pose



Loss Functions

- Penalty-reduced focal loss for center point and keypoint heatmaps:

$$\mathcal{L}_{\mathrm{p}} = \frac{-1}{N} \sum_{ij} \begin{cases} (1-\hat{Y}_{ij})^{\alpha} \log(\hat{Y}_{ij}) & \text{if } Y_{ij} = 1 \\ (1-Y_{ij})^{\beta} (\hat{Y}_{ij})^{\alpha} \log(1-\hat{Y}_{ij}) & \text{otherwise} \end{cases}$$

$$\alpha = 2, \beta = 4$$

- L1 center sub-pixel offset and keypoint sub-pixel offset loss:

$$\mathcal{L}_{\mathrm{off}} = \frac{1}{N} \sum_{p} \left\| \hat{O}_{\tilde{p}} - \left(\frac{p}{R} - \tilde{p} \right) \right\| \qquad \tilde{p} = \left\lfloor \frac{p}{R} \right\rfloor$$

- Overall loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{all}} = & \lambda_{\mathrm{p}_{cen}} \mathcal{L}_{\mathrm{p}_{cen}} + \lambda_{\mathrm{off}} \mathcal{L}_{\mathrm{off}} + \lambda_{\mathrm{bbox}} \mathcal{L}_{\mathrm{bbox}} \\ & + \lambda_{\mathrm{p}_{key}} \mathcal{L}_{\mathrm{p}_{key}} + \lambda_{\mathrm{offkey}} \mathcal{L}_{\mathrm{offkey}} \\ & + \lambda_{\mathrm{dis}} \mathcal{L}_{\mathrm{dis}} + \lambda_{\mathrm{dim}} \mathcal{L}_{\mathrm{dim}} \end{aligned}$$

$$\lambda_{\mathrm{p}_{cen}} = \lambda_{\mathrm{off}} = \lambda_{\mathrm{p}_{keu}} = \lambda_{\mathrm{offkey}} = \lambda_{\mathrm{dis}} = \lambda_{\mathrm{dim}} = 1, \; \lambda_{\mathrm{bbox}} = 0.1.$$

Results

Metrics:

- 3D Intersection over Union (IoU) with a threshold of 50%
- Mean normalized distance between the projections of 3D bounding box keypoints
- Viewpoint error of azimuth (lateral angle) with a threshold of 15° and elevation (vertical angle) with a threshold of 10°

Objectron Dataset performance compared against:

- MobilePose
- A two-stage network

Results

- Significantly outperform MobilePose
- Two-stage method falls behind on 3D IoU metric due to its failure for end-to-end training and taking dimensions into account

POSE ESTIMATION COMPARISON ON THE OBJECTRON TEST SET [15].

Stage	Method	Bike	Book	Bottle*	Camera	Cereal_box	Chair	Cup*	Laptop	Shoe	Mean
					Average precision at 0.5 3D IoU (↑)						
One	MobilePose [14]	0.3109	0.1797	0.5433	0.4483	0.5419	0.6847	0.3665	0.5225	0.4171	0.4461
Two	Two-stage [15]	0.6127	0.5218	0.5744	**0.8016**	0.6272	**0.8505**	0.5388	0.6735	0.6606	0.6512
One	Ours	**0.6419**	**0.5565**	**0.8021**	0.7188	**0.8211**	0.8471	**0.7704**	**0.6766**	**0.6618**	**0.7218**
					Mean pixel error of 2D projection of cuboid vertices (↓)						
One	MobilePose [14]	0.1581	0.0840	0.0818	0.0773	0.0454	0.0892	0.2263	0.0736	0.0655	0.1001
Two	Two-stage [15]	**0.0828**	**0.0477**	0.0405	**0.0449**	**0.0337**	**0.0488**	0.0541	**0.0291**	**0.0391**	**0.0467**
One	Ours	0.0872	0.0563	**0.0400**	0.0511	0.0379	0.0594	**0.0376**	0.0522	0.0463	0.0520
					Average precision at 15° azimuth error (↑)						
One	MobilePose [14]	0.4376	0.4111	0.4413	0.5293	0.8780	0.6195	0.0893	0.6052	0.3934	0.4894
Two	Two-stage [15]	0.8234	0.7222	0.8003	0.8030	**0.9404**	**0.8840**	0.6444	**0.8561**	0.5860	0.7844
One	Ours	**0.8622**	**0.7323**	**0.9561**	**0.8226**	0.9361	0.8822	**0.8945**	0.7966	**0.6757**	**0.8398**
					Average precision at 10° elevation error (↑)						
One	MobilePose [14]	0.7130	0.6289	0.6999	0.5233	0.8030	0.7053	0.6632	0.5413	0.4947	0.6414
Two	Two-stage [15]	**0.9390**	**0.8616**	0.8567	0.8437	**0.9476**	**0.9272**	0.8365	**0.7593**	0.7544	**0.8584**
One	Ours	0.9072	0.8535	**0.8881**	**0.8704**	0.9467	0.8999	**0.8562**	0.6922	**0.7900**	0.8560

Ablation Experiment

DIFFERENT STRATEGIES FOR 2D KEYPOINT OUTPUT DECODING (AVERAGE PRECISION AT 0.5 3D IOU METRIC (↑)).

Strategy	w/o add. proc.	Bike	Book	Bottle*	Camera	Cereal_box	Chair	Cup*	Laptop	Shoe	Mean
Displacement	✓	0.6254	0.5263	0.7917	**0.7191**	0.8115	**0.8492**	0.7553	0.6737	**0.6688**	0.7134
Heatmap	✓	0.5788	0.5539	0.7970	0.7035	0.8138	0.8260	0.7626	0.6124	0.6079	0.6951
Distance [16]	✗	0.6305	0.5436	0.7837	0.7111	0.8044	0.8460	0.7640	0.6692	0.6529	0.7117
Sampling [38]	✗	0.6279	0.5516	0.7873	0.7182	0.8134	0.8466	0.7687	0.6751	0.6641	0.7170
Disp. + Heatmap	✓	**0.6419**	**0.5565**	**0.8021**	0.7188	**0.8211**	0.8471	**0.7704**	**0.6766**	0.6618	**0.7218**

DIFFERENT STRATEGIES FOR COMPUTING CUBOID DIMENSIONS.

Method	Mean cuboid dimension error (↓)				Average precision at 0.5 3D IoU (↑)			
	Book	Laptop	Others	Mean	Book	Laptop	Others	Mean
Keypoint lifting [14] (no dim. pred.)	-	-	-	-	0.3999	0.5159	0.6540	0.6104
Estimated dim. (w/o convGRU)	0.8474	0.9124	**0.2434**	0.3849	0.5401	0.6378	**0.7528**	0.7164
Estimated dim. (w/ convGRU)	**0.7440**	**0.6799**	0.2475	**0.3507**	**0.5565**	**0.6766**	0.7519	**0.7218**
Ground truth dim. (oracle)	*0*	*0*	*0*	*0*	*0.6955*	*0.6942*	*0.7907*	*0.7694*

Conclusions

Primary Contributions:

1. Detect unseen objects from known category and estimate their poses from a monocular RGB input
2. Incorporate convGRU feature association to improve the accuracy of scale estimation
3. Prediction of relative dimension of 3D bounding cuboid for category-level pose estimation

Future work:

1. Incorporate shape geometry embeddings
2. Leverage different backbone networks
3. Use iteration to refine results

References

[1] Zeng, Andy, et al. "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge." *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017.

[2] Li, Yi, et al. "Deepim: Deep iterative matching for 6d pose estimation." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[3] Xiang, Yu, et al. "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes." arXiv preprint arXiv:1711.00199 (2017).

[4] Rad, Mahdi, and Vincent Lepetit. "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth." Proceedings of the IEEE international conference on computer vision. 2017.

[5] Wang, He, et al. "Normalized object coordinate space for category-level 6d object pose and size estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[6] Chen, Dengsheng, et al. "Learning canonical shape space for category-level 6d object pose and size estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[7] Hou, Tingbo, et al. "MobilePose: Real-time pose estimation for unseen objects with weak shape supervision." arXiv preprint arXiv:2003.03522 (2020).

[8] Ahmadyan, Adel, et al. "Objectron: A large scale dataset of object-centric videos in the wild with pose annotations." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[9] Lin, Yunzhi, et al. "Single-stage keypoint-based category-level object pose estimation from an RGB image." 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022.

[10] Yu, Fisher, et al. "Deep layer aggregation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

Thank you

Next Time: Object Tracking

- Seminar 5: Recurrent Networks and Object Tracking
  1. DeepIM: Deep Iterative Matching for 6D Pose Estimation, Li et al., 2018
  2. PoseRBPF: A Rao-Blackwellized Particle Filter for 6D Object Pose Tracking, Deng et al., 2019
  3. 6-PACK: Category-level 6D Pose Tracker with Anchor-Based Keypoints, Wang et al., 2020
  4. XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model, Cheng and Schwing, 2022
- Seminar 6: Visual Odometry and Localization
  1. Backprop KF: Learning Discriminative Deterministic State Estimators, Haarnoja et al., 2016
  2. Differentiable Particle Filters: End-to-End Learning with Algorithmic Priors, Jonschkowski et al., 2018
  3. Multimodal Sensor Fusion with Differentiable Filters, Lee et al., 2020
  4. Differentiable SLAM-net: Learning Particle SLAM for Visual Navigation, Karkus et al., 2021

DeepRob

Seminar 4
Dense Descriptors, Category-level Representations
University of Michigan and University of Minnesota