# PerspectiveNet: 3D Object Detection from a Single RGB Image via Perspective Points

Siyuan Huang, Yixin Chen, Tao Yuan, Siyuan Qi, Yixin Zhu, Song-Chun Zhu

University of California, Los Angeles

NEURAL INFORMATION PROCESSING SYSTEMS

## Task

Detect objects and recover the **3D object bounding boxes** from a single RGB image.

## Motivation

- **Layered representation** of computer vision by David Marr. For example, the early vision (texture, texton, primal sketch), mid-level vision ($2\frac{1}{2}$D), and high-level vision (primitive-based 3D).
- **3D prior and 3D geometry constraint** that can facilitate the 3D learning from a single image.

## Problems

- Could an intermediate representation bridge the huge gap between 2D image and 3D world?
- Is such an intermediate representation a better and more invariant prior compared to the priors obtained directly from specific tasks?
- How to incorporate the intermediate representation into an efficient end-to-end training framework?

We propose the **perspective points** as an intermediate representation to solve these problems.

## Contribution

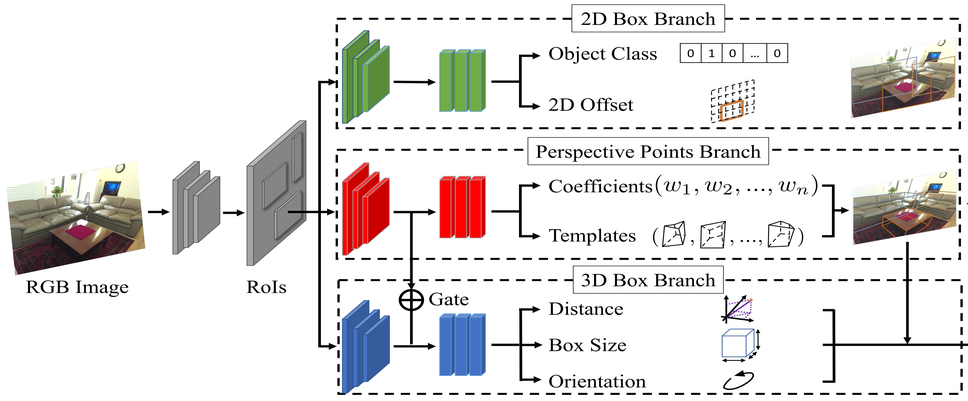
1. Perspective points bridge the gap between 2D and 3D bounding boxes without utilizing any extra category-specific 3D shape priors.
2. Devise a template-based method to efficiently and robustly estimate the perspective points.
3. Consistency between the 2D perspective points and 3D bounding boxes can be maintained by a differentiable projective function. The entire framework is end-to-end trainable.
4. Our method significantly outperforms previous methods.

## Overview

(a) 2D Bounding Boxes (b) 2D Perspective Points (c) 3D Bounding Boxes

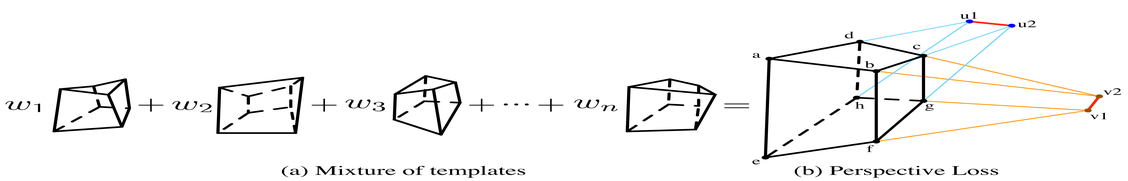
The proposed PerspectiveNet utilizes (b) the 2D perspective points as the intermediate representation to bridge the gap. The perspective points are the 2D perspective projection of the 3D bounding box corners, containing rich 3D information.

## Framework




For each proposed box, its RoI feature is fed into three network branches to predict: (i) the object class and the 2D box offset, (ii) 2D perspective templates (projected 3D box corners and object center) and the corresponding coefficients, and (iii) the 3D box size, orientation, and its distance from the camera.

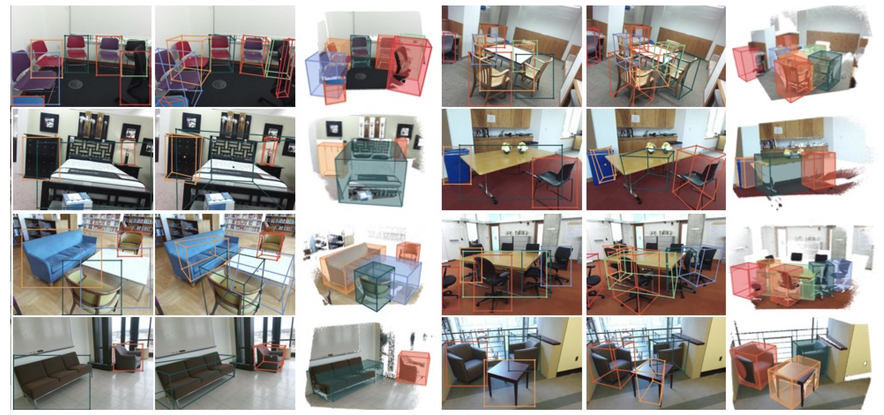
## Template-based Regression




(a) Mixture of templates (b) Perspective Loss

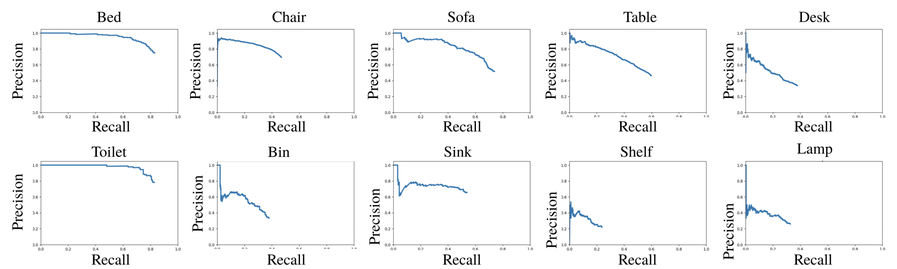
(a) The perspective points are estimated by a mixture of templates through a linear combination. Each template encodes geometric cues including orientations and viewpoints. (b) The perspective loss enforces each set of 2D perspective points to be the perspective projection of a (vertical) 3D cuboid.

## Qualitative Results



## Heatmaps vs. Templates

## Precision-Recall Curves




## Quantitative Results

| | bed | chair | sofa | table | desk | toilet | bin | sink | shelf | lamp | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3DGP [49] | 5.62 | 2.31 | 3.24 | 1.23 | - | - | - | - | - | - | - |
| HoPR [38] | 58.29 | 13.56 | 28.37 | 12.12 | 4.79 | 16.50 | 0.63 | 2.18 | 1.29 | 2.41 | 14.01 |
| CooP [36] | 63.58 | 17.12 | 41.22 | 26.21 | 9.55 | 58.55 | 10.19 | 5.34 | 3.01 | 1.75 | 23.65 |
| Ours (w/o. cam) | 71.39 | 34.94 | 55.63 | 34.10 | 14.23 | 73.73 | 17.47 | 34.41 | 4.21 | 9.54 | 34.96 |
| Ours (full) | **79.69** | **40.42** | **62.35** | **44.12** | **20.19** | **81.22** | **22.42** | **41.35** | **8.29** | **13.14** | **39.09** |