# $\mathrm{P}^2$Net: Patch-match and Plane-regularization for Unsupervised Indoor Depth Estimation


Zehao Yu*[1,2], Lei Jin*[1,2], and Shenghua Gao†[1,3]

[1] ShanghaiTech Univsertiy
[2] DGene Inc
[3] Shanghai Engineering Research Center of Intelligent Vision and Imaging
{yuzh,jinlei,gaoshh}@shanghaitech.edu.cn


`https://github.com/svip-lab/Indoor-SfMLearner`


**Abstract.** This paper tackles the unsupervised depth estimation task in indoor environments. The task is extremely challenging because of the vast areas of non-texture regions in these scenes. These areas could overwhelm the optimization process in the commonly used unsupervised depth estimation framework proposed for outdoor environments. However, even when those regions are masked out, the performance is still unsatisfactory. In this paper, we argue that the poor performance suffers from the non-discriminative point-based matching. To this end, we propose $\mathrm{P}^2$Net. We first extract points with large local gradients and adopt patches centered at each point as its representation. Multiview consistency loss is then defined over patches. This operation significantly improves the robustness of the network training. Furthermore, because those textureless regions in indoor scenes (*e.g.*, wall, floor, roof, *etc.*) usually correspond to planar regions, we propose to leverage superpixels as a plane prior. We enforce the predicted depth to be well fitted by a plane within each superpixel. Extensive experiments on NYUv2 and ScanNet show that our $\mathrm{P}^2$Net outperforms existing approaches by a large margin.



**Keywords:** Unsupervised Depth estimation, Patch-based Representation, Multiview Photometric Consistency, Piece-wise Planar Loss


## 1 Introduction

Depth estimation, as a fundamental problem in computer vision, bridges the gap between 2D images and 3D world. Lots of supervised depth estimation methods [7, 10, 30] have been proposed with the recent trend in convolution neural networks (CNNs). However, capturing a large number of images in different scenes with accurate ground truth depth requires expensive hardware and time [4, 15, 38, 41, 43]. To overcome the above challenges, another line of work [14, 16, 46, 55] focuses on unsupervised depth estimation that only uses either stereo videos or monocular videos as training data. The key supervisory signal in these work


∗ Equal Contribution
† Corresponding author

is the appearance consistency between the real view and the view synthesized based on the estimated scene geometry and ego-motion of the camera. Bilinear interpolation [20] based warping operation allows the training process to be fully differentiable.

While recent works of unsupervised depth estimation [50, 54, 56] have demonstrated impressive results on outdoor datasets, the same training process may easily collapse [53] on indoor datasets such as NYUv2 [41] or ScanNet [4]. The primary reason is that indoor environments contain large non-texture regions where the photometric consistency (the main supervisory signal in unsupervised learning) is unreliable. In such regions, the predicted depth might decay to infinite, while the synthesized view still has a low photometric error. Similar problems [16, 17, 32, 50] are also observed on outdoor datasets, especially in road regions. While the propotion of such regions is small on outdoor datasets, which would only lead to degradation in performance, the large non-texture regions on indoor scenarios can easily overwhelm the whole training process.

An intuitive try would be to mask out all the non-texture regions during the loss calculation. However, as the experimental results will demonstrate, merely ignoring the gradients from these non-texture regions still leads to inferior results. The reason is that we are minimizing per pixel (point) based multi-view photometric consistency error in the training process, where each point should be matched correctly across different views. Such point-based representation is not discriminative enough for matching in indoor scenes, since many other pixels in images could have the same intensity values. This operation could easily result in false matching. Taking inspiration from traditional multi-view stereo approaches [12, 39] that represent a point with a local patch, we propose to replace point-based representation with a patch-based representation to increase the discriminative ability in the matching process. Specifically, points with large local gradients are selected as our keypoints. We assume the same depth for pixels within a local window around every keypoint. We then project these local patches to different views with the predicted depth map and camera motion, and minimize multi-view photometric consistency error over the patches. Compared to point-based representation, our patch-based solution leads to a more distinctive characterization that produces more representative gradients with a wider basin of convergence.

Finally, to handle the rest large non-texture regions in indoor scenes, we draw inspiration from the recent success of work [11, 29, 51] that leverages the plane prior for indoor scene reconstruction. We make the assumption that homogeneous-colored regions, for example, walls, can be approximated with a plane. Here we adopt a similar strategy with the previous work [2, 3] that approximates the planar regions with superpixels. Specifically, we first extract planar regions by superpixels [9], then use a planar consistency loss to enforce the predicted depth in these regions can be well fitted by a plane, *i.e.*, low plane-fitting error within each superpixel. This allows our network to produce a more robust result.

Compared with MovingIndoor [53], a pioneer work on unsupervised indoor depth estimation that requires to first establish sparse correspondences between consecutive frames, and then propagates the sparse flows to the entire image, our P$^2$Net is direct, and no pre-matching process is required. Therefore, there is no concern for falsely matched pairs that might misguide the training of the network. Further, the supervisory signal of MovingIndoor [53] comes from the consistency between the synthesized optical flow and the predicted flow of the network. Such indirect supervision might also lead to a sub-optimal result. Our P$^2$Net instead supervises the network from two aspects: local patches for textured regions and planar consistency for the non-texture regions.

Our contributions can be summarized as follows: i) we propose to extract discriminative keypoints with large local gradients and use patches centered at each point as its representation. ii) patch-match: A patch-based warping process that assumes the same depth for pixels within a local patch is proposed for a more robust matching. iii) plane-regularization: we propose to use superpixels to represent those homogeneous-texture or non-texture piece-wise planar regions and regularize the depth consistency within each superpixel. On the one hand, our P$^2$Net leverages the discriminative patch-based representation that improves the matching robustness. On the other hand, our P$^2$Net encodes the piece-wise planar prior into the network. Consequently, our approach is more suitable for indoor scene depth estimation. Extensive experiments on widely-used indoor datasets NYUv2 [41] and ScanNet [4] demonstrate that P$^2$Net outperforms state-of-the-art by a large margin.

## 2 Related Work

### 2.1 Supervised Depth Estimation

A vast amount of research has been done in the field of supervised depth estimation. With the recent trend in convolution neural networks (CNNs), many different deep learning based approaches have been proposed. Most of them frame the problem as a per-pixel regression problem. Particularly, Eigen et al. [5] propose a multi-scale coarse-to-fine approach. Laina et al. [25] improve the performance of depth estimation by introducing a fully convolutional architecture with several up-convolution blocks. Kim et al. [22] use conditional random fields to refine the depth prediction. Recently, Fu et al. [10] treat the problem from an ordinal regression perspective. With a carefully designed discretization strategy and an ordinal loss, their method is able to achieve new state-of-the-art results in supervised depth estimation. Other work focus on combining depth estimation with semantic segmentation [21, 52] and surface norm estimation [6, 34]. Yin et al. [49] show that high-order 3D geometric constraints, the so-called virtual normal, can further improve depth prediction accuracy. However, all of these methods rely on vast amounts of labeled data, which is still a large cost in both hardware and time.

### 2.2 Unsupervised Depth Estimation

Unsupervised learning of depth estimation has been proposed to ease the demand for large-scale labeled training data. One line of work exploits stereo images or videos [46, 14, 16] as training data and trains a network to minimize the photometric error between synthesized view and real view. Godard et al. [16] introduce a left-right disparity consistency as regularization. Another line of work learns depth from monocular video sequences. Zhou et al. [55] introduce a separate network to predict camera motion between input images. Their method learns to estimate depth and ego-motion simultaneously. Later work also focus on joint-learning by minimizing optical flow errors [37, 50], or combining SLAM pipelines into deep networks [40, 44]. However, none of the above approaches produce satisfactory results on indoor datasets. MovingIndoor [53] is the first to study unsupervised depth estimation in indoor scenes. The authors propose an optical flow estimation network, SFNet, initialized with sparse flows from matching results of SURF [1]. The dense optical flows are used as the supervisory signal for the learning of the depth and pose. By contrast, we propose to supervise the training with a more discriminative patch-based multi-view photometric consistency error and regularize the depth within homogeneous-color regions with a planar consistency loss. Our method is direct, and no pre-matching process is required. Therefore, there is no concern for falsely matched pairs that might misguide the training of the network.

### 2.3 Piece-wise Planar Scene Reconstruction

Piece-wise planar reconstruction is an active research topic in multi-view 3D reconstruction [11, 13], SLAM [2, 3] and has drawn increasing attention recently [29, 48, 51, 28]. Traditional methods [12, 13] generate plane hypotheses by fitting planes to triangulated 3D points, then assign hypotheses to each pixel via a global optimization. Concha and Civera [2, 3] used superpixels [9] to describe non-texture region in a monocular dense SLAM system. Their method has shown impressive reconstruction results. Raposo et al. [36] proposed $\pi$Match, a vSLAM pipeline with plane features to for a piecewise planar reconstruction. In their more recent work [35], they recovered structure and motion from planar regions and combined these estimations into stereo algorithms. Together with Deep CNNs, Liu et al. [29] learn to infer plane parameters and associate each pixel to a plane in a supervised manner. Yang and Zhou [48] learn a similar network with only depth supervision. Following work [28, 51] further formulate the planar reconstruction problem as an instance segmentation problem and have shown significant improvements. Inspired by these work, we incorporate the planar prior for homogeneous-color regions into our unsupervised framework and propose a planar consistency loss to regularize the depth map in such regions in the training phrase.

## 3 Method

### 3.1 Overview

Our goal is to learn a depth estimator for indoor environments with only monocular videos. Following recent success on unsupervised depth estimation [55], our

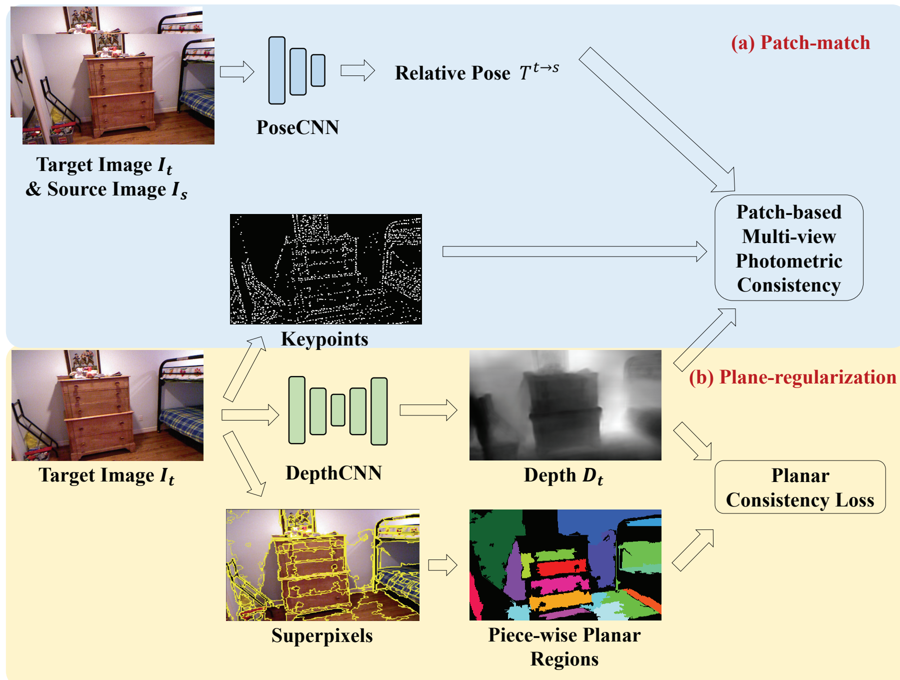


Fig. 1: Overall network architecture. Given input images, DepthCNN predicts the corresponding depth for the target image $I_t$, PoseCNN outputs the relative pose from the source to the target view. Our $P^2$Net consists of two parts: a) **Patch-match Module**: We warp the selected pixels along with their local neighbors with a patch-based warping module. b) **Plane-regularization Module**: We enforce depth consistency in large superpixel regions.

$P^2$Net contains two learnable modules: DepthCNN and PoseCNN. DepthCNN takes a target view image $I_t$ as input and outputs its corresponding depth $D_t$. PoseCNN takes a source view image $I_s$ and a target view image $I_t$ as input and predicts the relative pose $T_{t \to s}$ between two consecutive frames. A commonly used strategy is to first synthesize a novel view $I'_t$ with the predicted depth map $D_t$ and camera motion $T_{t \to s}$, and minimize the photometric consistency error between the synthesized view $I'_t$ and its corresponding real view $I_t$. However, the training process soon collapses when directly applying this strategy to indoor scenarios.

Our observation is that the large non-texture regions in indoor scenes might easily overwhelm the whole training process. Therefore, we propose to select representative keypoints that have large local variances. However, representing a point with a single intensity value, as done in previous unsupervised learning frameworks [16, 17], is non-discriminative and may result in false matching. To address this problem, we propose a **Patch-match Module**, a patch-based representation that combines a point with the local window centered at that point to increase their discriminative abilities and minimize patch-based multi-view photometric consistency error. To handle the large non-texture regions, we propose a **Plane-regularization Module** to extract homogeneous-color regions using large superpixels and enforce that the predicted depth map within a superpixel may be approximated by a plane. The overview of our P$^2$Net is depicted in Fig. 1.

### 3.2 Keypoints Extraction

Different from outdoor scenes, the large proportion of the non-texture regions in indoor scenes can easily overwhelm the training process, leading to trivial solutions where DepthCNN always predicts an infinity depth, and PoseCNN always gives an identity rotation. Thus, only points within textured regions should be kept in the training process to avoid the network being stuck in such trivial results. Here, we adopt the points selection strategy from Direct Sparse Odometry (DSO) [8] for its effectiveness and efficiency. Points from DSO are sampled from pixels that have large intensity gradients. Examples of extracted DSO keypoints are shown in Fig. 3.

A critical benefit of our direct method over matching based approaches [53] is that we do not need to pre-compute the matching across images, which itself is a challenging problem. As a result, our points need to be extracted from the target image once only. No hand-crafted descriptor for matching is needed.

### 3.3 Patch-based Multi-view Photometric Consistency Error

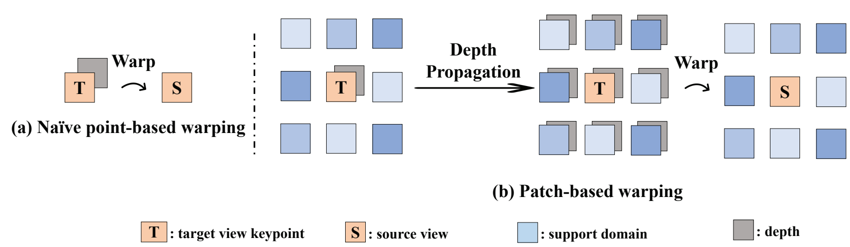


Fig. 2: Two types of warping operations. a) Naive point-based warping. b) Our proposed patch-based warping. Note that we are defining pixels over its support domain and warp the entire window. Combining support domains into the pixel leads to more robust representations. Best viewed in color.

With the extracted keypoints from the previous step, we can simply define a photometric consistency error by comparing the corresponding pixels' values. However, such point-based representation is not representative enough and may easily cause false matching because there are many pixels with the same intensity values in an image. In traditional sparse SLAM pipelines [8], to overcome the above challenge, a support domain $\Omega_{p_i}$ is defined over each point $p_i$'s local window. Photometric loss is then accumulated over each support domain $\Omega_{p_i}$ instead of a single isolated point. This operation would lead to more robust results as the extracted keypoints combined with their support domains are becoming much more unique.

Inspired from the above operation, here we propose a patch-based warping process as in Fig. 2. Specifically, we extract DSO keypoints $p_i^t$ from the target view $t$, the original point-based warping process first back-projects the keypoints to the source view $I_s$ with:

$$p_i^{t \longrightarrow s} = KT^{t \longrightarrow s} D(p_i) K^{-1} p_i^t \tag{1}$$

where $K$ denotes the camera intrinsic parameters, $T^{t \longrightarrow s}$ the relative pose between the source view $I_s$ and the target view $I_t$, and $D(p_i)$ the depth of point $p_i$.Then we sample the intensity values with bilinear interpolation [20] at $p_i^{t \longrightarrow s}$ in the source view.

On the contrast, our approach assumes a same depth within each pixel's local window $\Omega_{p_i}^t$. Then, for every extracted keypoint, we warp the point together with its local support region $\Omega_{p_i}^t$ with the exact same depth. Our warping process can thus be described as :

$$\Omega_{p_i}^{t \longrightarrow s} = KT^{t \longrightarrow s} D(p_i) K^{-1} \Omega_{p_i}^t \tag{2}$$

where $\Omega_{p_i}^t$ and $\Omega_{p_i}^{t \longrightarrow s}$ denotes the support domains of the point $p_i$ in the target view and the source view, respectively. From a SLAM perspective, we characterize each point over its support region, such patch-based approaches makes the representation of each point more distinctive and robust. From a deep learning perspective, our operation allows a larger region of valid gradients compared to the bilinear interpolation with only four nearest neighbors as in Equation (1).

Given a keypoint $p = (x, y)$, we define its support region $\Omega_p$ over a local window with size $N$ as:

$$\Omega_p = \{(x + x_p, y + y_p), x_p \in \{-N, 0, N\}, y_p \in \{-N, 0, N\}\} \tag{3}$$

$N$ is set to 3 in our experiments. Following recent work [17], we define our patch-based multi-view photometric consistency error as a combination of an L1 loss and a structure similarity loss SSIM [45] over the support region $\Omega_{p_i}$:

$$L_{SSIM} = SSIM(I_t \left[\Omega_{p_i}^t\right], I_s \left[\Omega_{p_i}^{t \longrightarrow s}\right]) \tag{4}$$

$$L_{L1} = ||I_t \left[\Omega_{p_i}^t\right] - I_s \left[\Omega_{p_i}^{t \longrightarrow s}\right] ||_1 \tag{5}$$

$$L_{ph} = \alpha L_{SSIM} + (1 - \alpha) L_{L1} \tag{6}$$

where $I_t[p]$ denotes pixel values at $p$ in image $I_t$ via a bilinear interpolation, and $\alpha = 0.85$ a weighting factor. Note that when more than one source images are used in the photometric loss, we follow [17] to select the one with the minimum $L_{ph}$ for robustness purpose.

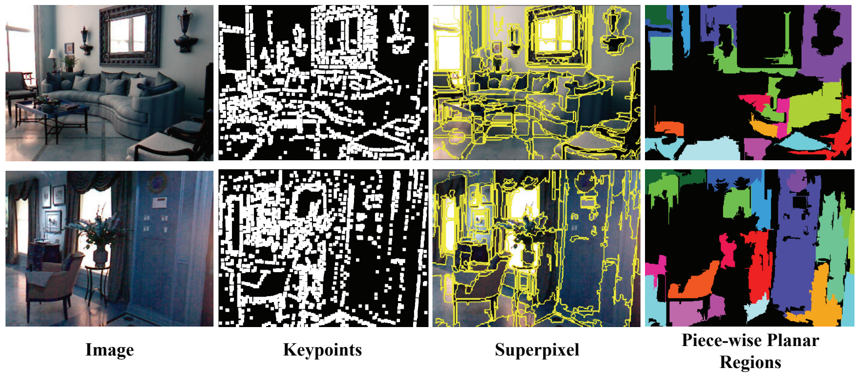


Fig. 3: Examples of input images, their corresponding keypoints, superpixels and piece-wise planar regions obtained from large superpixels.

### 3.4 Planar Consistency Loss

Finally, to further constrain the large non-texture regions in indoor scenes, we propose to enforce piecewise planar constraints into our network. Our assumption is that, most of the homogeneous-color regions are planar regions, and we can assume a continuous depth that satisfies the planar assumptions within these regions. Following representative work on reconstruction of indoor scenes [3, 2], we adopt the Felzenszwalb superpixel segmentation [9] in our approach. The segmentation algorithm follows a greedy approach and segments areas with low gradients, and hence produces more planar regions. Examples with images, superpixels segmentation and piece-wise planar regions determined by superpixels, are demonstrated in Figure 3. We can see that our assumption is reasonable, since indoor scenes generally consists of many man-made objects, like floor, walls, roof, *etc*. Further, previous work also shows the good performance of indoor scene reconstruction with a piece-wise planar assumption in [28, 29, 51].

Specifically, given an input image $I$, we first extract superpixels from the image and only keep regions larger than 1000 pixels. An intuition is that the planar regions, like walls, floor, the surface of a table, are more likely to be within a larger area. Given an extracted superpixel $SPP_m$ and its corresponding depth $D(p_n)$ from an image, where $p_n$ enumerates all the pixels within $SPP_m$, we first backproject all the points $p_n$ back to 3D space,

$$p_n^{3D} = D(p_n)K^{-1}p_n, p_n \subseteq SPP_m \tag{7}$$

where $p_n^{3D}$ denotes the corresponding point of $p_n$ in 3D world. We define the plane in 3D following [29, 51] as

$$A_m^\top p_n^{3D} = \mathbf{1} \tag{8}$$

where $A_m$ is plane parameter of $SPP_m$.

We use a least square method to fit the plane parameters $A_m$. Mathematically, we form two data matrices $Y_m$ and $P_n$, where $Y_m = \mathbf{1} = \left[1\ 1\ ...\ 1\right]^\top$, $P_n = \left[p_1^{3D}\ p_2^{3D} ...\ p_n^{3D}\right]^\top$:

$$P_n A_m = Y_m \tag{9}$$

Then $A_m$ can be computed with a closed-form solution:

$$A_m = \left(P_n^\top P_n + \epsilon E\right)^{-1} P_n^\top Y_m. \tag{10}$$

where $E$ is an identity matrix, and $\epsilon$ a small scalar for numerical stability. After obtaining the plane parameters, We can then retrieve our fitted planar depth for each pixel within the superpixel $SPP_m$ as $D^{'}(p_n) = (A_m^\top K^{-1} p_n)^{-1}$. We then add another constraint to enforce a low plane-fitting error within each superpixel:

$$L_{spp} = \sum_{m=1}^{M} \sum_{n=1}^{N} |D(p_n) - D^{'}(p_n)| \tag{11}$$

Here $M$ denotes the number of superpixels, and $N$ number of pixels in each superpixel.

### 3.5 Loss Function

We also adopt an edge-aware smoothness term $L_{sm}$ over the entire depth map as that in [16, 17]:

$$L_{sm} = |\partial_x d_t^*|\, e^{-|\partial_x I_t|} + |\partial_y d_t^*|\, e^{-|\partial_y I_t|}, \tag{12}$$

where $\partial_x$ denotes the gradients along the $x$ direction, $\partial_y$ along the $y$ direction and $d_t^* = d_t / \overline{d_t}$ is the normalized depth.

Our overall loss function is defined as :

$$L = L_{ph} + \lambda_1 L_{sm} + \lambda_2 L_{spp} \tag{13}$$

where $\lambda_1$ is set to 0.001, $\lambda_2$ is set to 0.05 in our experiments.

## 4 Experiments

### 4.1 Implementation Details

We implement our solution under the PyTorch [33] framework. Following the pioneer work on unsupervised depth estimation in outdoor scenes, we use the same encoder-decoder architecture as that in [17] with separate ResNet18s [18] pretrained on ImageNet as our backbones, the same PoseCNN as that in [17]. Adam [23] is adopted as our optimizer. The network is trained for a total of 41 epochs with a batch size of 12. Initial learning rate is set to $1e-4$ for the first 25 epochs. Then we decay it once by 0.1 for the next 10 epochs. We adopt random flipping and color augmentation during training. All images are resized to 288 $\times$

| Methods | Supervised | rms ↓ | rel ↓ | log10 ↓ | $\delta < 1.25$ ↑ | $\delta < 1.25^2$ ↑ | $\delta < 1.25^3$ ↑ |
|---|---|---|---|---|---|---|---|
| Make3D [38] | ✓ | 1.214 | 0.349 | - | 0.447 | 0.745 | 0.897 |
| Liu et al. [31] | ✓ | 1.200 | 0.350 | 0.131 | - | - | - |
| Ladicky et al. [24] | ✓ | 1.060 | 0.335 | 0.127 | - | - | - |
| Li et al. [26] | ✓ | 0.821 | 0.232 | 0.094 | 0.621 | 0.886 | 0.968 |
| Liu et al. [30] | ✓ | 0.759 | 0.213 | 0.087 | 0.650 | 0.906 | 0.976 |
| Li et al. [27] | ✓ | 0.635 | 0.143 | 0.063 | 0.788 | 0.958 | 0.991 |
| Xu et al. [47] | ✓ | 0.586 | 0.121 | 0.052 | 0.811 | 0.954 | 0.987 |
| DORN [10] | ✓ | 0.509 | 0.115 | 0.051 | 0.828 | 0.965 | 0.992 |
| Hu et al. [19] | ✓ | 0.530 | 0.115 | 0.050 | 0.866 | 0.975 | 0.993 |
| PlaneNet [29] | ✓ | 0.514 | 0.142 | 0.060 | 0.827 | 0.963 | 0.990 |
| PlaneReg [51] | ✓ | 0.503 | 0.134 | 0.057 | 0.827 | 0.963 | 0.990 |
| MovingIndoor [53] | × | 0.712 | 0.208 | 0.086 | 0.674 | 0.900 | 0.968 |
| Monov2 [17] | × | 0.617 | 0.170 | 0.072 | 0.748 | 0.942 | 0.986 |
| $P^2$Net (3 frames) | × | **0.599** | **0.159** | **0.068** | **0.772** | **0.942** | **0.984** |
| $P^2$Net (5 frames) | × | 0.561 | 0.150 | 0.064 | 0.796 | 0.948 | 0.986 |
| $P^2$Net (5 frames PP) | × | **0.553** | **0.147** | **0.062** | **0.801** | **0.951** | **0.987** |
| ResNet18 | ✓ | 0.591 | 0.138 | 0.058 | 0.823 | 0.964 | 0.989 |

Table 1: Performance comparison on the NYUv2 dataset. We report results of depth supervised approaches in the first block, plane supervised results in the second block, unsupervised results in the third and fourth block, and the supervised upper bound of our approach denoted as ResNet18 in the final block. PP denotes the final result with left-right fliping augmentation in evaluation. Our approach achieves state-of-the-art performance among the unsupervised ones. ↓ indicates the lower the better, ↑ indicates the higher the better.

384 pixels during training. Predicted depth are up-sampled back to the original resolution during testing. Since unsupervised monocular depth estimation exists scale ambiguity, we adopt the same median scaling strategy as that in [17, 55] for evaluation. A larger baseline is also beneficial for training, and we use a 3-frame (one target frame, 2 source frames) input in our ablation experiments and report the final results with a 5-frame (one target frame, 4 source frames) input. Besides the standard DSO keypoints, we also draw points randomly to have a fixed number of 3K points from one image.

### 4.2 Datasets

We evaluate our $P^2$Net on two publicly available datasets of indoor scenes, including NYU Depth V2 [41] and ScanNet [4].

**NYU Depth V2.** NYU Depth V2 consists of a total 582 indoor scenes. We adopt the same train split of 283 scenes following previous work on indoor depth estimation [53] and provide our results on the official test set with the standard depth evaluation criteria. We sample the training set at 10 frames interval as our target views and use $\pm 10$, $\pm 20$ frames as our source views. This leaves us around 20K unique images, a number much less than the 180K images used in the previous work of unsupervised indoor depth estimation [53]. We undistort the input image as in [42] and crop 16 black pixels from the border region.

We compare with MovingIndoor [53], the pioneer work on unsupervised indoor depth estimation and Monov2 [17], a state-of-the-art unsupervised depth estimation method on outdoor datasets. Quantitative results are provided in Table 1. Our method achieves the best result. We further provide some visualization of our predicted depth in Fig. 4. GeoNet collapsed during training as we inspected. Compared to MovingIndoor [53], our method preserves much more details owing to the patch-based multi-view consistency module. A supervised upper bound, denoted as ResNet18, is also provided here by replacing the backbone network in [19] with ours.

Results for surface normal estimation are provided in Tab. 2. We compare with other methods that fits norm from the point clouds. Not only is our result the best among the unsupervised ones, it is also close to supervised results like DORN [10]. We visualize some results of our method for surface normal estimation in Fig. 5.

**ScanNet.** ScanNet [4] contains around 2.5M images captured in 1513 scenes. While there is no current official train/test split on ScanNet for depth estimation, we randomly pick 533 testing images from diverse scenes. We directly evaluate our models pretrained on NYUv2 under a transfer learning setting to test the generalizability of our approach. We showcase some of the prediction results in Fig. 4. We achiever better result as reported in Tab. 3.

| Methods | Supervised | Mean ↓ | 11.2° ↑ | 22.5° ↑ | 30° ↑ |
|---|---|---|---|---|---|
| GeoNet [34] | ✓ | 36.8 | 15.0 | 34.5 | 46.7 |
| DORN [10] | ✓ | 36.6 | 15.7 | 36.5 | 49.4 |
| MovingIndoor [53] | × | 43.5 | 10.2 | 26.8 | 37.9 |
| Monov2 [17] | × | 43.8 | 10.4 | 26.8 | 37.3 |
| $P^2$Net (3 frames) | × | 38.8 | 11.5 | 31.8 | 44.8 |
| $P^2$Net (5 frames) | × | 36.6 | 15.0 | 36.7 | 49.0 |
| $P^2$Net (5 frames pp) | × | **36.1** | **15.6** | **37.7** | **50.0** |

Table 2: Surface normal evaluation on NYUv2. PP denotes the final result with left-right fliping augmentation in evaluation.

### 4.3 Ablation Experiments

**Patch-match and Plane-regularization.** For our baseline, we first calculate the variance within a local region for each pixel. This servers as our texture/non-texture region map. Photometric loss is directly multiplied by the map. This represents the most straightforward case when only point-based supervision is provided. We report the numbers in the first row of Tab. 4. Then we add our

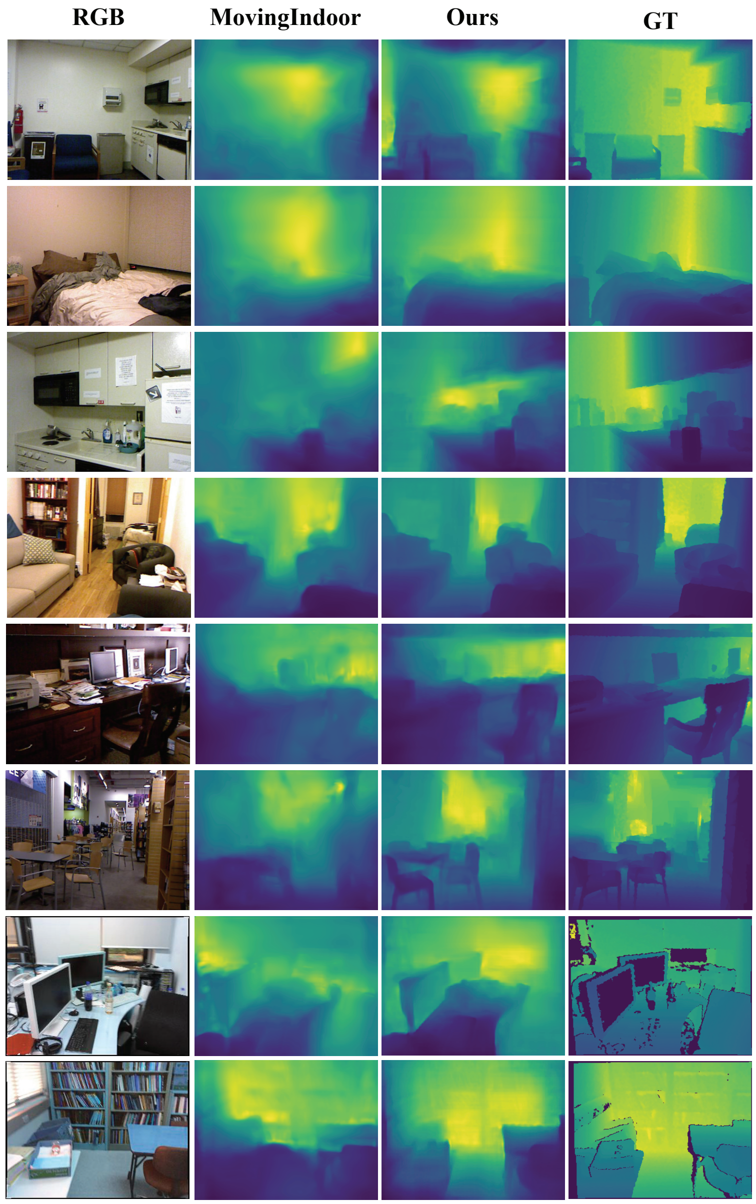


Fig. 4: Depth visualization on NYUv2 (first 6 rows) and ScanNet (last 2 rows). We trained our model on NYUv2 and directly transfer the weights to ScanNet without fine-tunning. From left right: input image, results of MovingIndoor [53], our results and ground truth depth. GeoNet would collapse on indoor datasets due to the large non-texture regions. Compared to MovingIndoor [53], our methods preserve more details.

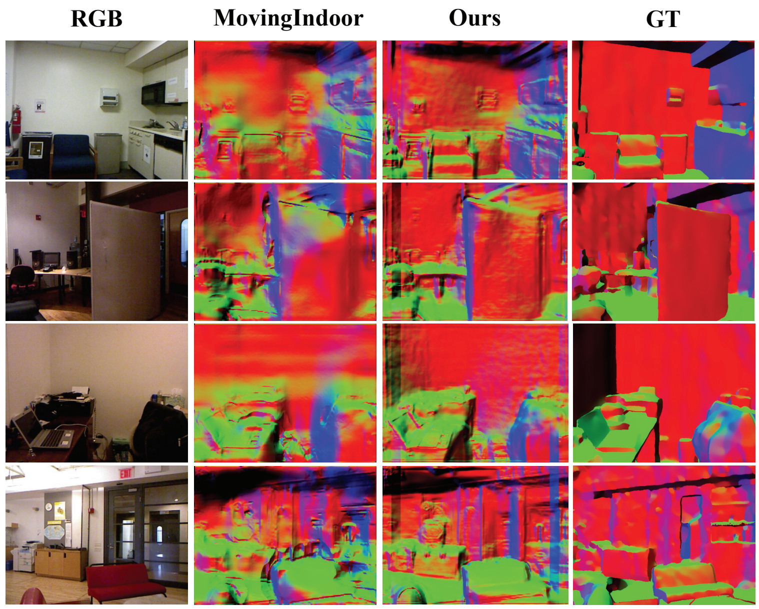


Fig. 5: Visualization of fitted surface norm from 3D point clouds on the NYUv2 dataset. From left to right: input image, results of MovingIndoor [53], ours and ground truth normal. Our method produces more smooth results in planar regions.

proposed Patch-match module and report the results in the second line, the Plane-regularization module in the fourth line. Experiments demonstrate the effectiveness of our proposed modules.

**Different keypoint types.** Here, we demonstrate that our method is not limited to some specific type of keypoint detectors. We replace DSO with a blob region detector SURF [1]. We achieve similar results as reported in line two and three in Tab. 4.

**Camera pose.** Following previous work [42] on predicting depth from videos, we provide our camera pose estimation results on the ScanNet dataset, consisting a total of 2000 pairs of images from diverse scenes. Note that since our method is monocular, there exists scale ambiguity in our predictions. Hence, we follow [42] and rescale our translation during evaluation. Results are reported in Tab. 5. Our method performs better than MovingIndoor [53].

**Results on outdoor scenes.** Here we also provide our results on the KITTI benchmark in Tab. 6. We trained and evaluated our results on the same subset as in [17]. Our method outperforms another unsupervised indoor depth estimation approach MovingIndoor. Different from indoor scenes, the main challenge in outdoor scenes are moving objects (like cars) and occlusions, which seldom occur in indoor scenes. Our method does not take such priors into consideration. On the contrast, Monov2 is specially designed to handle these cases.

| Methods | rms ↓ | rel ↓ | log10 ↓ | $\delta < 1.25$ ↑ | $\delta < 1.25^2$ ↑ | $\delta < 1.25^3$ ↑ |
|---|---|---|---|---|---|---|
| MovingIndoor [53] | 0.483 | 0.212 | 0.088 | 0.650 | 0.905 | 0.976 |
| Monov2 [17] | 0.458 | 0.200 | 0.083 | 0.672 | 0.922 | 0.981 |
| P$^2$Net | **0.420** | **0.175** | **0.074** | **0.740** | **0.932** | **0.982** |

Table 3: Performance comparison on transfer learning. Results are evaluated directly with NYUv2 pretrained models on ScanNet. Our model still achieves the best result.

| Keypoint | Patch Match | Plane Regularization | rms ↓ | rel ↓ | $\delta < 1.25$ ↑ | $\delta < 1.25^2$ ↑ | $\delta < 1.25^3$ ↑ |
|---|---|---|---|---|---|---|---|
| - | | | 0.786 | 0.240 | 0.628 | 0.884 | 0.962 |
| DSO | ✓ | | 0.612 | 0.166 | 0.758 | 0.945 | 0.985 |
| SURF | ✓ | | 0.622 | 0.169 | 0.750 | 0.941 | 0.986 |
| DSO | ✓ | ✓ | **0.599** | **0.159** | **0.772** | **0.942** | **0.984** |

Table 4: Ablation study of our proposed module on the NYUv2 dataset.

| Method | rot(deg) | tr(deg) | tr(cm) |
|---|---|---|---|
| Moving [53] | 1.96 | 39.17 | 1.40 |
| Monov2 [17] | 2.03 | 41.12 | **0.83** |
| P$^2$Net | **1.86** | **35.11** | 0.89 |

Table 5: Results on camera pose.

| Method | rel↓ | rms ↓ | $\delta < 1.25$ ↑ |
|---|---|---|---|
| Moving [53] | 0.130 | 5.294 | - |
| P$^2$Net | **0.126** | **5.140** | **0.862** |
| Monov2 [17] | 0.115 | 4.863 | 0.877 |

Table 6: Results on KITTI.

# 5 Conclusion

This paper propose P$^2$Net that leverages patches and superpixels for unsupervised depth estimation task in indoor scenes. Extensive experiments validate the effectiveness of our P$^2$Net. Here for simplicity we adopt the fronto-parallel assumption. One possible solution could be to first pretrain the network and calculate normal from depth. Then we can combine normal into the training process.


# Acknowledgements

The work was supported by National Key R&D Program of China (2018AAA0100704), NSFC #61932020, and ShanghaiTech-Megavii Joint Lab. We would also like to thank Junsheng Zhou from Tsinghua University for detailed commonts of reproducing his work and some helpful discussions.

## References


1. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: ECCV (2006)
2. Concha, A., Civera, J.: Using superpixels in monocular slam. In: ICRA (2014)
3. Concha, A., Civera, J.: Dpptam: Dense piecewise planar tracking and mapping from a monocular sequence. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5686–5693. IEEE (2015)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
5. Eigen, D., Puhrsch, C., Fergus, R.: Prediction from a single image using a multi-scale deep network. In: NIPS (2014)
6. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
7. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS. pp. 2366–2374 (2014)
8. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE transactions on pattern analysis and machine intelligence **40**(3), 611–625 (2017)
9. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International journal of computer vision **59**(2), 167–181 (2004)
10. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: CVPR (2018)
11. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Manhattan-world stereo. In: CVPR (2009)
12. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE transactions on pattern analysis and machine intelligence **32**(8), 1362–1376 (2009)
13. Gallup, D., Frahm, J.M., Pollefeys, M.: Piecewise planar and non-planar stereo for urban scene reconstruction. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2010)
14. Garg, R., BG, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: ECCV (2016)
15. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
16. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
17. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: ICCV (2019)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
19. Hu, J., Ozay, M., Zhang, Y., Okatani, T.: Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In: WACV (2019)
20. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: NIPS (2015)
21. Jiao, J., Cao, Y., Song, Y., Lau, R.: Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In: ECCV (2018)
22. Kim, S., Park, K., Sohn, K., Lin, S.: Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In: ECCV (2016)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

24. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: CVPR (2014)
25. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 3DV (2016)
26. Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: CVPR (2015)
27. Li, J., Klein, R., Yao, A.: A two-streamed network for estimating fine-scaled depth maps from single rgb images. In: ICCV (2017)
28. Liu, C., Kim, K., Gu, J., Furukawa, Y., Kautz, J.: Planercnn: 3d plane detection and reconstruction from a single image. In: CVPR (2019)
29. Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y.: Planenet: Piece-wise planar reconstruction from a single rgb image. In: CVPR (2018)
30. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE transactions on pattern analysis and machine intelligence **38**(10), 2024–2039 (2015)
31. Liu, M., Salzmann, M., He, X.: Discrete-continuous depth estimation from a single image. In: CVPR (2014)
32. Luo, C., Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R., Yuille, A.: Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. arXiv preprint arXiv:1810.06125 (2018)
33. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NIPS (2019)
34. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: CVPR (2018)
35. Raposo, C., Antunes, M., Barreto, J.P.: Piecewise-planar stereoscan: Sequential structure and motion using plane primitives. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(8), 1918–1931 (2018)
36. Raposo, C., Barreto, J.P.: pimatch: Monocular vslam and piecewise planar reconstruction using fast plane correspondences. In: ECCV (2016)
37. Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X., Zha, H.: Unsupervised deep learning for optical flow estimation. In: AAAI (2017)
38. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence **31**(5), 824–840 (2008)
39. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: ECCV (2016)
40. Shi, Y., Zhu, J., Fang, Y., Lien, K., Gu, J.: Self-supervised learning of depth and ego-motion with differentiable bundle adjustment. arXiv preprint arXiv:1909.13163 (2019)
41. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
42. Teed, Z., Deng, J.: Deepv2d: Video to depth with differentiable structure from motion. arXiv preprint arXiv:1812.04605 (2018)
43. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., Shakhnarovich, G.: DIODE: A Dense Indoor and Outdoor DEpth Dataset. CoRR **abs/1908.00463** (2019)
44. Wang, C., Miguel Buenaposada, J., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: CVPR (2018)

45. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
46. Xie, J., Girshick, R., Farhadi, A.: Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In: ECCV (2016)
47. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: CVPR (2017)
48. Yang, F., Zhou, Z.: Recovering 3d planes from a single image via convolutional neural networks. In: ECCV (2018)
49. Yin, W., Liu, Y., Shen, C., Yan, Y.: Enforcing geometric constraints of virtual normal for depth prediction. In: ICCV (2019)
50. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: CVPR (2018)
51. Yu, Z., Zheng, J., Lian, D., Zhou, Z., Gao, S.: Single-image piece-wise planar 3d reconstruction via associative embedding. In: CVPR (2019)
52. Zhang, Z., Cui, Z., Xu, C., Jie, Z., Li, X., Yang, J.: Joint task-recursive learning for semantic segmentation and depth estimation. In: ECCV (2018)
53. Zhou, J., Wang, Y., Qin, K., Zeng, W.: Moving indoor: Unsupervised video depth learning in challenging environments. In: ICCV (2019)
54. Zhou, J., Wang, Y., Qin, K., Zeng, W.: Unsupervised high-resolution depth learning from videos with dual networks. In: ICCV (2019)
55. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)
56. Zou, Y., Luo, Z., Huang, J.B.: Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In: ECCV (2018)