PlaneMatch: Patch Coplanarity Prediction for Robust RGB-D Reconstruction

Yifei Shi[1,2], Kai Xu[1,2⋆], Matthias Nießner[3], Szymon Rusinkiewicz[1], and Thomas Funkhouser[1,4]

[1] Princeton University
[2] National University of Defense Technology
[3] Technical University of Munich
[4] Google

Abstract. We introduce a novel RGB-D patch descriptor designed for detecting coplanar surfaces in SLAM reconstruction. The core of our method is a deep convolutional neural network that takes in RGB, depth, and normal information of a planar patch in an image and outputs a descriptor that can be used to find coplanar patches from other images. We train the network on 10 million triplets of coplanar and non-coplanar patches, and evaluate on a new coplanarity benchmark created from commodity RGB-D scans. Experiments show that our learned descriptor outperforms alternatives extended for this new task by a significant margin. In addition, we demonstrate the benefits of coplanarity matching in a robust RGBD reconstruction formulation. We find that coplanarity constraints detected with our method are sufficient to get reconstruction results comparable to state-of-the-art frameworks on most scenes, but outperform other methods on established benchmarks when combined with traditional keypoint matching.

Keywords: RGB-D registration, co-planarity, loop closure

1 Introduction

With the recent proliferation of inexpensive RGB-D sensors, it is now becoming practical for people to scan 3D models of large indoor environments with hand-held cameras, enabling applications in cultural heritage, real estate, virtual reality, and many other fields. Most state-of-the-art RGB-D reconstruction algorithms either perform frame-to-model alignment [1] or match keypoints for global pose estimation [2]. Despite the recent progress in these algorithms, registration of hand-held RGB-D scans remains challenging when local surface features are not discriminating and/or when scanning loops have little or no overlaps.

An alternative is to detect planar features and associate them across frames with coplanarity, parallelism, and perpendicularity constraints [3–9]. Recent work has shown compelling evidence that planar patches can be detected and tracked

⋆ Corresponding author: kevin.kai.xu@gmail.com

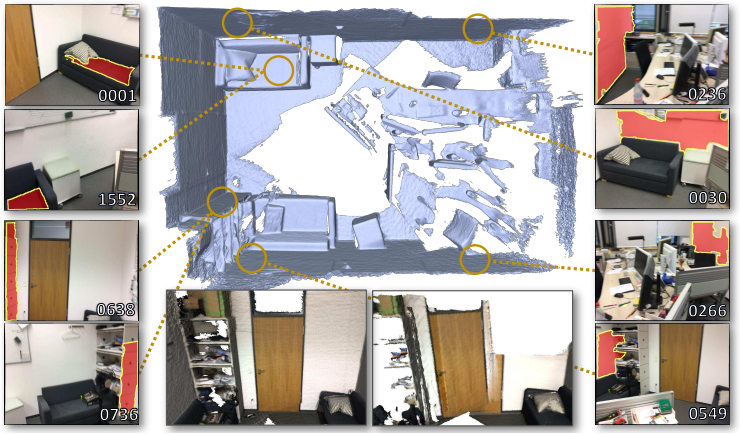


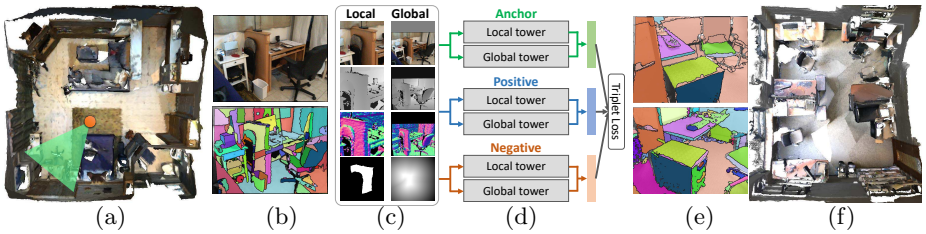
Fig. 1: Scene reconstruction based on coplanarity matching of patches across different views (numbers indicate frame ID) for both overlapping (left two pairs) and non-overlapping (right two pairs) patch pairs. The two pairs to the right are long-range, without overlapping. The bottom shows a zoomed-in comparison between our method (left) and key-point matching based method [2] (right).

robustly, especially in indoor environments where flat surfaces are ubiquitous. In cases where traditional features such as keypoints are missing (e.g., wall), there seems tremendous potential to support existing 3D reconstruction pipelines.

Even though coplanarity matching is a promising direction, current approaches lack strong per-plane feature descriptors for establishing putative matches between disparate observations. As a consequence, coplanarity priors have only been used in the context of frame-to-frame tracking [3] or in post-process steps for refining a global optimization [4]. We see this as analogous to the relationship between ICP and keypoint matching: just as ICP only converges with a good initial guess for pose, current methods for exploiting coplanarity are unable to initialize a reconstruction process from scratch due to the lack of discriminative coplanarity features.

This paper aims to enable global, *ab initio* coplanarity matching by introducing a discriminative feature descriptor for planar patches of RGB-D images. Our descriptor is learned from data to produce features whose L2 difference is predictive of *whether or not two RGB-D patches from different frames are coplanar*. It can be used to detect pairs of coplanar patches in RGB-D scans *without an initial alignment*, which can be used to find loop closures or to provide coplanarity constraints for global alignment (see Figure 1).

A key novel aspect of this approach is that it focuses on detection of coplanarity rather than overlap. As a result, our plane patch features can be used to discover long-range alignment constraints (like "loop closures") between distant, non-overlapping parts of the same large surface (e.g., by recognizing carpets on



(a) (b) (c) (d) (e) (f)

Fig. 2: An overview of our method. We train an embedding network (c-d) to predict coplanarity for a pair of planar patches across different views, based on the co-planar patches (b) sampled from training sequences with ground-truth camera poses (a). Given a test sequence, our robust optimization performs reconstruction (f) based on predicted co-planar patches (e).

floors, tiles on ceilings, paneling on walls, etc.). In Figure 1, the two patch pairs shown to the right helped produce a reconstruction with globally flat walls.

To learn our planar patch descriptor, we design a deep network that takes in color, depth, normals, and multi-scale context for pairs of planar patches extracted from RGB-D images, and predicts whether they are coplanar or not. The network is trained in a self-supervised fashion where training examples are automatically extracted from coplanar and noncoplanar patches from ScanNet [10].

In order to evaluate our descriptor, we introduce a new coplanarity matching datasets, where we can see in series of thorough experiments that our new descriptor outperforms existing baseline alternatives by significant margins. Furthermore, we demonstrate that by using our new descriptor, we are able to compute strong coplanarity constraints that improve the performance of current global RGB-D registration algorithms. In particular, we show that by combining coplanarity and point-based correspondences reconstruction algorithms are able to handle difficult cases, such as scenes with a low number of features or limited loop closures. We outperform other state-of-the-art algorithms on the standard TUM RGB-D reconstruction benchmark [11]. Overall, the research contributions of this paper are:

- A new task: predicting coplanarity of image patches for the purpose of RGB-D image registration.
- A self-supervised process for training a deep network to produce features for predicting whether two image patches are coplanar or not.
- An extension of the robust optimization algorithm [12] to solve camera poses with coplanarity constraints.
- A new training and test benchmark for coplanarity prediction.
- Reconstruction results demonstrating that coplanarity can be used to align scans where keypoint-based methods fail to find loop closures.

2 Related Work

RGB-D Reconstruction: Many SLAM systems have been described for reconstructing 3D scenes from RGB-D video. Examples include KinectFusion [13, 1], VoxelHashing [14], ScalableFusion [15], Point-based Fusion [16], Octrees on CPU [17], Elastic Fusion [18], Stereo DSO [19], Colored Registration [20], and Bundle Fusion [2]. These systems generally perform well for scans with many loop closures and/or when robust IMU measurements are available. However, they often exhibit drift in long scans when few constraints can be established between disparate viewpoints. In this work, we detect and enforce coplanarity constraints between planar patches to address this issue as an alternative feature channel for global matching.

Feature Descriptors: Traditionally, SLAM systems have utlized *keypoint* detectors and descriptors to establish correspondence constraints for camera pose estimation. Example keypoint descriptors include SIFT [21], SURF [22], ORB [23], etc. More recently, researchers have learned keypoint descriptors from data – e.g., MatchNet [24], Lift [25], SE3-Nets [26], 3DMatch [27], Schmidt et al. [28]. These methods rely upon repeatable extraction of keypoint positions, which is difficult for widely disparate views. In contrast, we explore the more robust method of extracting planar patches without concern for precisely positioning the patch center.

Planar Features: Many previous papers have leveraged planar surfaces for RGB-D reconstruction. The most commmon approach is to detect planes in RGB-D scans, establish correspondences between matching features, and solve for the camera poses that align the corresponding features [29–36]. More recent approaches build models comprising planar patches, possibly with geometric constraints [4, 37], and match planar features found in scans to planar patches in the models [4–8]. The search for correspondences is often aided by hand-tuned descriptors designed to detect overlapping surface regions. In contrast, our approach finds correspondences between *coplanar patches* (that may not be overlapping); we learn descriptors for this task with a deep network.

Global Optimization: For large-scale surface reconstruction, it is common to use off-line or asynchronously executed global registration procedures. A common formulation is to compute a pose graph with edges representing pairwise transformations between frames and then optimize an objective function penalizing deviations from these pairwise alignments [38–40]. Recent methods [12, 41] use indicator variables to identify loop closures or matching points during global optimization using a least-squares formulation. We extend this formulation by setting indicator variables for individual coplanarity constraints.

3 Method

Our method consists of two components: 1) a deep neural network trained to generate a descriptor that can be used to discover coplanar pairs of RGB-D

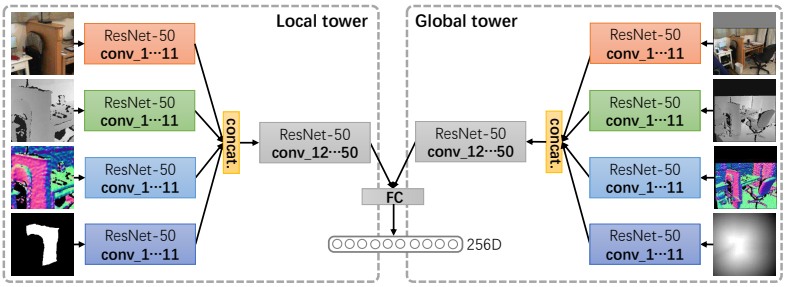


Fig. 3: Network architecture of the local and global towers. Layers shaded in the same color share weights.

patches without an initial registration, and 2) a global SLAM reconstruction algorithm that takes special advantage of detected pairs of coplanar patches.

3.1 Coplanarity Network

Coplanarity of two planar patches is by definition geometrically measurable. However, for two patches that are observed from different, yet unknown views, whether they are coplanar is not determinable based on geometry alone. Furthermore, it is not clear that coplanarity can be deduced solely from the local appearance of the imaged objects. We argue that the prediction of coplanarity across different views is a structural, or even semantic, visual reasoning task, for which neither geometry nor local appearance alone is reliable.

Humans infer coplanarity by perceiving and understanding the structure and semantics of objects and scenes, and contextual information plays a critical role in this reasoning task. For example, humans are able to differentiate different facets of an object, from virtually *any* view, by reasoning about the structure of the facets and/or by relating them to surrounding objects. Both involve inference with a context around the patches being considered, possibly at multiple scales. This motivates us to learn to predict cross-view coplanarity from appearance and geometry, using *multi-scale contextual information*. We approach this task by learning an embedding network that maps coplanar patches from different views nearby in feature space.

Network Design: Our coplanarity network (Figure 2 and 3) is trained with triplets of planar patches, each involving an anchor, a coplanar patch (positive) and a noncoplanar patch (negative), similar to [42]. Each patch of a triplet is fed into a convolutional network based on ResNet-50 [43] for feature extraction, and a triplet loss is estimated based on the relative proximities of the three features. To learn coplanarity from both appearance and geometry, our network takes multiple channels as input: an RGB image, depth image, and normal map.

We encode the contextual information of a patch at two scales, local and global. This is achieved by cropping the input images (in all channels) to rectangles of 1.5 and 5 times the size of the patch's bounding box, respectively. All

cropped images are clamped at image boundaries, padded to a square, and then resized to 224×224. The padding uses 50% gray for RGB images and a value of 0 for depth and normal maps; see Figure 3.

To make the network aware of the region of interest (as opposed to context) in each input image, we add, for each of the two scales, an extra binary mask channel. The local mask is binary, with the patch of interest in white and the rest of the image in black. The global mask, in contrast, is continuous, with the patch of interest in white and then a smooth decay to black outside the patch boundary. Intuitively, the local mask helps the network distinguish the patch of interest from the close-by neighborhood, e.g. other sides on the same object. The global mask, on the other hand, directs the network to learn global structure by attending to a larger context, with importance smoothly decreasing based on distance to the patch region. Meanwhile, it also weakens the effect of specific patch shape, which is unimportant when considering global structure.

In summary, each scale consists of RGB, depth, normal, and mask channels. These inputs are first encoded independently. Their feature maps are concatenated after the 11-th convolutional layer, and then pass through the remaining 39 layers. The local and global scales share weights for the corresponding channels, and their outputs are finally combined with a fully connected layer (Figure 3).

Network Training: The training data for our network are generated from datasets of RGB-D scans of 3D indoor scenes, with high-quality camera poses provided with the datasets. For each RGB-D frame, we segment it into planar patches using agglomerative clustering on the depth channel. For each planar patch, we also estimate its normal based on the depth information. The extracted patches are projected to image space to generate all the necessary channels of input to our network. Very small patches, whose local mask image contains less than 300 pixels with valid depths, are discarded.

Triplet Focal Loss: When preparing triplets to train our network, we encounter the well-known problem of a severely imbalanced number of negative and positive patch pairs. Given a training sequence, there are many more negative pairs, and most of them are too trivial to help the network learn efficiently. Using randomly sampled triplets would overwhelm the training loss by the easy negatives.

We opt to resolve the imbalance issue by dynamically and discriminatively scaling the losses for hard and easy triplets, inspired by the recent work of focal loss for object detection [44]. Specifically, we propose the *triplet focal loss*:

$$L_{\text{focal}}(x_a, x_p, x_n) = \max\left(0, \frac{\alpha - \Delta d_{\text{f}}}{\alpha}\right)^{\lambda}, \tag{1}$$

where x_a, x_p and x_n are the feature maps extracted for anchor, positive, and negative patches, respectively; $\Delta d_{\text{f}} = d_{\text{f}}(x_n, x_a) - d_{\text{f}}(x_p, x_a)$, with d_{f} being the L2 distance between two patch features. Minimizing this loss encouranges the anchor to be closer to the positive patch than to the negative in descriptor space, but with less weight for larger distances.

See Figure 4, left, for a visualization of the loss function with $\alpha = 1$. When $\lambda = 1$, this loss becomes the usual margined loss, which gives non-negligible loss

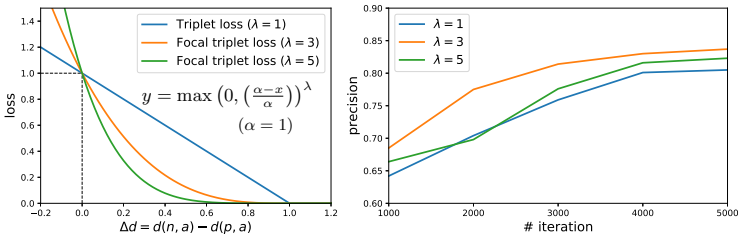


Fig. 4: Visualization and comparison (prediction accuracy over #iter.) of different triplet loss functions.

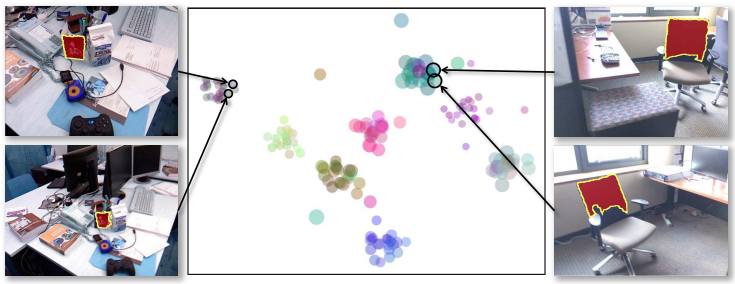


Fig. 5: t-SNE visualization of coplanarity-based features of planar patches from different views. Ground-truth coplanarity (measured by mutual RMS point-to-plane distance) is encoded by color and physical size of patches by dot size.

to easy examples near the margin α. When $\lambda > 1$, however, we obtain a focal loss that down-weights easy-to-learn triplets while keeping high loss for hard ones. Moreover, it smoothly adjusts the rate at which easy triplets are down-weighted. We found $\lambda = 3$ to achieve the best training efficiency (Figure 4, right). Figure 5 shows a t-SNE visualization of coplanarity-based patch features.

3.2 Coplanarity-Based Robust Registration

To investigate the utility of this planar patch descriptor and coplanarity detection approach for 3D reconstruction, we have developed a global registration algorithm that estimates camera poses for an RGB-D video using pairwise constraints derived from coplanar patch matches in addition to keypoint matches.

Our formulation is inspired by the work of Choi et al. [12], where the key feature is the robust penalty term used for automatically selecting the correct matches from a large pool of hypotheses, thus avoiding iterative rematching as in ICP. Note that this formulation does not require an initial alignment of camera poses, which would be required for other SLAM systems that leverage coplanarity constraints.

Given an RGB-D video sequence $\mathcal{F}$, our goal is to compute for each frame $i \in \mathcal{F}$ a camera pose in the global reference frame, $\mathbf{T}_i = (\mathbf{R}_i, \mathbf{t}_i)$, that brings

them into alignment. This is achieved by jointly aligning each pair of frames $(i, j) \in \mathcal{P}$ that were predicted to have some set of coplanar patches, Π_{ij}. For each pair $\pi = (p, q) \in \Pi_{ij}$, let us suppose w.l.o.g. that patch p is from frame i and q from j. Meanwhile, let us suppose a set of matching key-point pairs Θ_{ij} is detected and matched between frame i and j. Similarly, we assume for each point pair $\theta = (\mathbf{u}, \mathbf{v}) \in \Theta_{ij}$ that key-point $\mathbf{u}$ is from frame i and $\mathbf{v}$ from j.

Objective Function: The objective of our coplanarity-based registration contains four terms, responsible for coplanar alignment, coplanar patch pair selection, key-point alignment, and key-point pair selection:

$$E(T, s) = E_{\text{data-cop}}(T, s) + E_{\text{reg-cop}}(s) + E_{\text{data-kp}}(T, s) + E_{\text{reg-kp}}(s). \tag{2}$$

Given a pair of coplanar patches predicted by the network, the *coplanarity data term* enforces the coplanarity, via minimizing the point-to-plane distance from sample points on one patch to the plane defined by the other patch:

$$E_{\text{data-cop}}(T, s) = \sum_{(i,j)\in\mathcal{P}} \sum_{\pi\in\Pi_{ij}} w_\pi \, s_\pi \, \delta^2(\mathbf{T}_i, \mathbf{T}_j, \pi), \tag{3}$$

where δ is the *coplanarity distance* of a patch pair $\pi = (p, q)$. It is computed as the root-mean-square point-to-plane distance over both sets of sample points:

$$\delta^2 = \frac{1}{|\mathcal{V}_p|} \sum_{\mathbf{v}_p \in \mathcal{V}_p} d^2(\mathbf{T}_i \mathbf{v}_p, \phi_q^G) + \frac{1}{|\mathcal{V}_q|} \sum_{\mathbf{v}_q \in \mathcal{V}_q} d^2(\mathbf{T}_j \mathbf{v}_q, \phi_p^G),$$

where $\mathcal{V}_p$ is the set of sample points on patch p and d is point-to-plane distance:

$$d(\mathbf{T}_i \mathbf{v}_p, \phi_q^G) = (\mathbf{R}_i \mathbf{v}_p + \mathbf{t}_i - \mathbf{p}_q) \cdot \mathbf{n}_q.$$

$\phi_q^G = (\mathbf{p}_q, \mathbf{n}_q)$ is the plane defined by patch q, which is estimated in the *global* reference frame using the corresponding transformation $\mathbf{T}_j$, and is updated in every iteration. s_π is a control variable (in $[0, 1]$) for the selection of patch pair π, with 1 standing for selected and 0 for discarded. w_π is a weight that measures the confidence of pair π's being coplanar. This weight is another connection between the optimization and the network, besides the predicted patch pairs themselves. It is computed based on the feature distance of two patches, denoted by $d_{\text{f}}(p, q)$, extracted by the network: $w_{(p,q)} = e^{-d_{\text{f}}^2(p,q)/(\sigma^2 d_{\text{fm}}^2)}$, where d_{fm} is the maximum feature distance and $\sigma = 0.6$.

The *coplanarity regularization term* is defined as:

$$E_{\text{reg-cop}}(s) = \sum_{(i,j)\in\mathcal{P}} \sum_{\pi\in\Pi_{ij}} \mu \, w_\pi \, \Psi(s_\pi), \tag{4}$$

where the penalty function is defined as $\Psi(s) = \left(\sqrt{s} - 1\right)^2$. Intuitively, minimizing this term together with the data term encourages the selection of pairs incurring a small value for the data term, while immediately pruning those pairs whose data term value is too large and deemed to be hard to minimize. w_π is defined the same as before, and μ is a weighting variable that controls the emphasis on pair selection.

The *key-point data term* is defined as:

$$E_{\text{data-kp}}(T, s) = \sum_{(i,j)\in\mathcal{P}} \sum_{\theta\in\Theta_{ij}} s_\theta \, ||\mathbf{T}_i\mathbf{u} - \mathbf{T}_j\mathbf{v}||, \tag{5}$$

Similar to coplanarity, a control variable s_θ is used to determine the selection of point pair θ, subjecting to the *key-point regularization term*:

$$E_{\text{reg-kp}}(s) = \sum_{(i,j)\in\mathcal{P}} \sum_{\theta\in\Theta_{ij}} \mu\,\Psi(s_\theta), \tag{6}$$

where μ shares the same weighting variable with Equation (4).

Optimization: The optimization of Equation (2) is conducted iteratively, where each iteration interleaves the optimization of transformations T and selection variables s. Ideally, the optimization could take every pair of frames in a sequence as an input for global optimization. However, this is prohibitively expensive since for each frame pair the system scales with the number of patch pairs and key-point pairs. To alleviate this problem, we split the sequence into a list of overlapping fragments, optimize frame poses within each fragment, and then perform a final global registration of the fragments, as in [12].

For each fragment, the optimization takes all frame pairs within that fragment and registers them into a rigid point cloud. After that, we take the matching pairs that have been selected by the intra-fragment optimization, and solve the inter-fragment registration based on those pairs. Inter-fragment registration benefits more from long-range coplanarity predictions.

The putative matches found in this manner are then pruned further with a rapid and approximate RANSAC algorithm applied for each pair of fragments. Given a pair of fragments, we randomly select a set of three matching feature pairs, which could be either planar-patch or key-point pairs. We compute the transformation aligning the selected triplet, and then estimate the "support" for the transformation by counting the number of putative match pairs that are aligned by the transformation. For patch pairs, alignment error is measures by the root-mean-square closest distance between sample points on the two patches. For key-point pairs, we simply use the Euclidean distance. Both use the same threshold of 1cm. If a transformation is found to be sufficiently supported by the matching pairs (more than 25% consensus), we include all the supporting pairs into the global optimization. Otherwise, we simply discard all putative matches.

Once a set of pairwise constraints have been established in this manner, the frame transformations and pair selection variables are alternately optimized with an iterative process using Ceres [45] for the minimization of the objective function at each iteration. The iterative optimization converges when the relative value change of each unknown is less than 1×10^{-6}. At a convergence, the weighting variable μ, which was initialized to 1m in the beginning, is decreased by half and the above iterative optimization continues. The whole process is repeated until μ is lower than 0.01m, which usually takes less than 50 iterations. The supplementary material provides a study of the optimization behavior, including convergence and robustness to incorrect pairs.

4 Results and Evaluations

4.1 Training Set, Parameters, and Timings

Our training data is generated from the ScanNet [10] dataset, which contains 1513 scanned sequences of indoor scenes, reconstructed by BundleFusion [2]. We adopt the training/testing split provided with ScanNet and the training set (1045 scenes) are used to generate our training triplets. Each training scene contributes $10K$ triplets. About $10M$ triplets in total are generated from all training scenes. For evaluating our network, we build a *coplanarity benchmark* using 100 scenes from the testing set. For hierarchical optimization, the fragment size is 21, with a 5-frame overlap between adjacent fragments. The network training takes about 20 hours to converge. For a sequence of $1K$ frames with 62 fragments and 30 patches per frame, the running time is 10 minutes for coplanarity prediction (0.1 second per patch pair) and 20 minutes for optimization (5 minutes for intra-fragment and 15 minutes for inter-fragment).

4.2 Coplanarity Benchmark

We create a benchmark **COP** for evaluating RGB-D-based coplanarity matching of planar patches. The benchmark dataset contains $12K$ patch pairs with ground-truth coplanarity, which are organized according to the physical size/area of patches (COP-S) and the centroid distance between pairs of patches (COP-D). COP-S contains $6K$ patch pairs which are split uniformly into three subsets with *decreasing* average patch size, where the patch pairs are sampled at random distances. COP-D comprises three subsets (each containing $2K$ pairs) with *increasing* average pair distance but uniformly distributed patch size. For all subsets, the numbers of positive and negative pairs are equal. Details of the benchmark are provided in the supplementary material.

4.3 Network Evaluation

Our network is the first, to our knowledge, that is trained for coplanarity prediction. Therefore, we perform comparison against baselines and ablation studies. See visual results of coplanarity matching in the supplementary material.

Comparing to Baseline Methods: We first compare to two hand-crafted descriptors, namely the color histogram within the patch region and the SIFT feature at the patch centroid. For the task of key-point matching, a commonly practiced method (e.g., in [46]) is to train a neural network that takes image patches centered around the key-points as input. We extend this network to the task of coplanarity prediction, as a non-trivial baseline. For a fair comparison, we train a triplet network with ResNet-50 with only one tower per patch taking three channels (RGB, depth, and normal) as input. For each channel, the image is cropped around the patch centroid, with the same padding and resizing scheme as before. Thus, no mask is needed since the target is always at the image center.

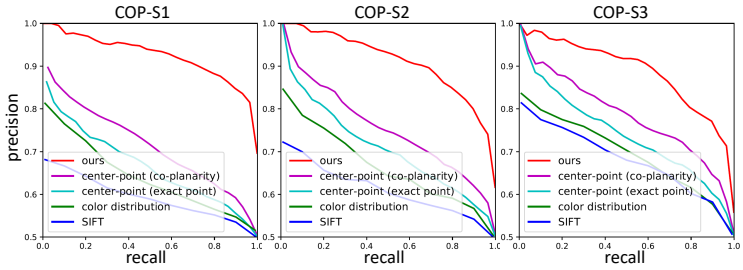


Fig. 6: Comparing to baselines including center-point matching networks trained with coplanarity and exact point matching, respectively, SIFT-based point matching and color distribution based patch matching.

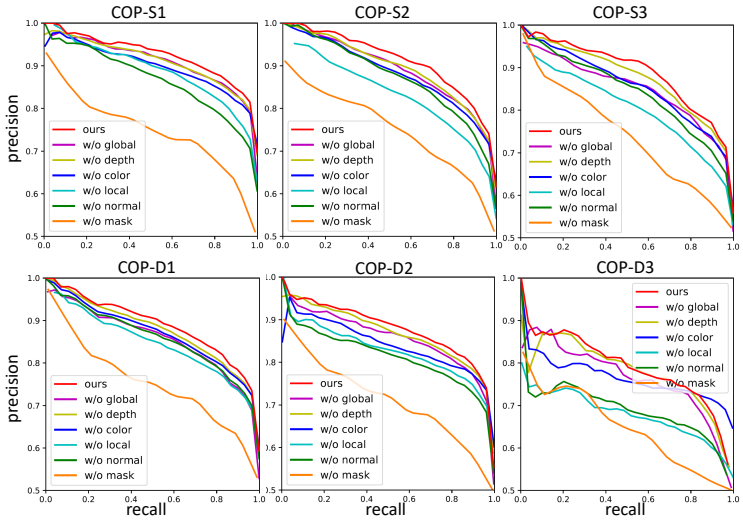


Fig. 7: Ablation studies of our coplanarity network.

We train two networks with different triplets for the task of 1) exact center point matching and 2) coplanarity patch matching, respectively.

The comparison is conducted over COP-S and the results of precision-recall are plotted in Figure 6. The hand-crafted descriptors fail on all tests, which shows the difficulty of our benchmark datasets. Compared to the two alternative center-point-based networks (point matching and coplanarity matching), our method performs significantly better, especially on larger patches.

Ablation Studies: To investigate the need for the various input channels, we compare our full method against that with the RGB, depth, normal, or mask input disabled, over the COP benchmark. To evaluate the effect of multi-scale context, our method is also compared to that without local or global channels. The PR plots in Figure 7 show that our full method works the best for all tests.

From the experiments, several interesting phenomena can be observed. First, the order of overall importance of the different channels is: mask > normal > RGB > depth. This clearly shows that coplanarity prediction across different views can neither rely on appearance or geometry alone. The important role of masking in concentrating the network's attention is quite evident. We provide a further comparison to justify our specific masking scheme in the supplementary material. Second, the global scale is more effective for bigger patches and more distant pairs, for which the larger scale is required to encode more context. The opposite goes for the local scale due the higher resolution of its input channels. This verifies the complementary effect of the local and global channels in capturing contextual information at different scales.

4.4 Reconstruction Evaluation

Quantitative Results: We perform a quantitative evaluation of reconstruction using the TUM RGB-D dataset by [11], for which ground-truth camera trajectories are available. Reconstruction error is measured by the absolute trajectory error (ATE), i.e., the root-mean-square error (RMSE) of camera positions along a trajectory. We compare our method with six state-of-the-art reconstruction methods, including RGB-D SLAM [47], VoxelHashing [14], ElasticFusion [18], Redwood [12], BundleFusion [2], and Fine-to-Coarse [4]. Note that unlike the other methods, Redwood does not use color information. Fine-to-Coarse is the most closely related to our method, since it uses planar surfaces for structurally-constrained registration. This method, however, relies on a good initialization of camera trajectory to bootstrap, while our method does not. Our method uses SIFT features for key-point detection and matching. We also implement an enhanced version of our method where the key-point matchings are pre-filtered by BundleFusion (named 'BundleFusion+Ours').

As an ablation study, we implement five baseline variants of our method. 1) 'Coplanarity' is our optimization with only coplanarity constraints. Without key-point matching constraint, our optimization can sometimes be under-determined and needs reformulation to achieve robust registration when not all degrees of freedom (DoFs) can be fixed by coplanarity. The details on the formulation can be found in the supplementary material. 2) 'Keypoint' is our optimization with only SIFT key-point matching constraints. 3) 'No D. in RANSAC' stands for our method where we did not use our learned patch descriptor during the voting in frame-to-frame RANSAC. In this case, any two patch pairs could cast a vote if they are geometrically aligned by the candidate transformation. 4) 'No D. in Opt' means that the optimization objective for coplanarity is not weighted by the matching confidence predicted by our network (w_π in Equation (3) and (4)). 5) 'No D. in Both' is a combination of 3) and 4).

Table 1 reports the ATE RMSE comparison. Our method achieves state-of-the-art results for the first three TUM sequences (the fourth is a flat wall). This is achieved by exploiting our long-range coplanarity matching for robust large-scale loop closure, while utilizing key-point based matching to pin down the possible free DoFs which are not determinable by coplanarity. When being

Method	fr1/desk	fr2/xyz	fr3/office	fr3/nst
RGB-D SLAM	2.3	**0.8**	3.2	1.7
VoxelHashing	2.3	2.2	2.3	8.7
Elastic Fusion	2.0	**1.1**	1.7	1.6
Redwood	2.7	9.1	3.0	192.9
Fine-to-Coarse	5.0	3.0	3.9	3.0
BundleFusion	1.6	**1.1**	2.2	**1.2**
Ours	**1.4**	**1.1**	**1.6**	1.5
BundleFuison+Ours	**1.3**	**0.8**	**1.5**	**0.9**

(a) Comparison to alternatives.

Method	fr1/desk	fr2/xyz	fr3/office	fr3/nst
No D. in RANSAC	9.6	4.8	12.6	2.3
No D. in Opt.	4.8	2.7	2.5	1.9
No D. in Both	18.9	8.3	16.4	2.4
Key-point Only	5.6	4.4	5.2	2.6
Coplanarity Only	2.5	2.1	3.7	–
Ours	**1.4**	**1.1**	**1.7**	**1.5**

(b) Comparison to baselines.

Table 1: Comparison of ATE RMSE (in cm) with alternative and baseline methods on TUM sequences [11]. Colors indicate the best and second best results.

combined with BundleFusion key-points, our method achieves the best results over all sequences. Therefore, our method complements the current state-of-the-art methods by providing a means to handle limited frame-to-frame overlap.

The ablation study demonstrates the importance of our learned patch descriptor in our optimization – i.e., our method performs better than all variants that do not include it. It also shows that coplanarity constraints alone are superior to keypoints only for all sequences except the flat wall (fr3/nst). Using coplanar and keypoint matches together provides the best method overall.

Qualitative Results: Figure 8 shows visual comparisons of reconstruction on sequences from ScanNet [10] and new ones scanned by ourselves. We compare reconstruction results of our method with a state-of-the-art key-point based method (BundleFusion) and a planar-structure-based method (Fine-to-Coarse). The low frame overlap makes the key-point based loop-closure detection fail in BundleFusion. Lost tracking of successive frames provides a poor initial alignment for Fine-to-Coarse, causing it to fail. In contrast, our method can successfully detect non-overlapping loop closures through coplanar patch pairs and achieve good quality reconstructions for these examples without an initial registration. More visual results are shown in the supplementary material.

Effect of Long-Range Coplanarity. To evaluate the effect of long-range coplanarity matching on reconstruction quality, we show in Figure 9 the reconstruction results computed with all, half, and none of the long-range coplanar pairs predicted by our network. We also show a histogram of coplanar pairs survived the optimization. From the visual reconstruction results, the benefit of long-range coplanar pairs is apparent. In particular, the larger scene (bottom) benefits more from long-range coplanarity than the smaller one (top). In Figure 8, we also give the number of non-overlapping coplanar pairs after optimization, showing that long-range coplanarity did help in all examples.

5 Conclusion

We have proposed a new planar patch descriptor designed for finding coplanar patches without a priori global alignment. At its heart, the method uses a deep

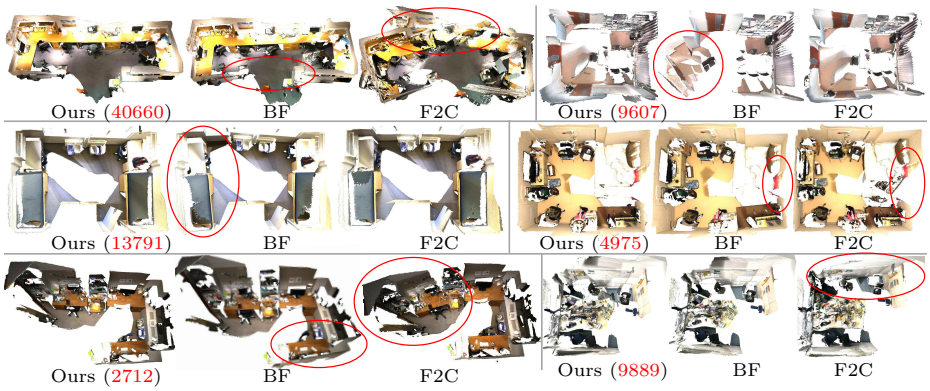


Fig. 8: Visual comparison of reconstructions by our method, BundleFusion (BF) [2], and Fine-to-Coarse (F2C) [4], on six sequences. Red ellipses indicate parts with misalignment. For our results, we give the number of long-range coplanar pairs selected by the optimization.

Fig. 9: Reconstruction results with 100% (left column), 50% (middle) and 0% (right) of long-range coplanar pairs detected, respectively. The histograms of long-range coplanar patch pairs (count over patch distance ($1 \sim 5$m)) are given.

network to map planar patch inputs with RGB, depth, and normals to a descriptor space where proximity can be used to predict coplanarity. We expect that deep patch coplanarity prediction provides a useful complement to existing features for SLAM applications, especially in scans with large planar surfaces and little inter-frame overlap.

Acknowledgement We are grateful to Min Liu, Zhan Shi, Lintao Zheng, and Maciej Halber for their help on data preprocessing. We also thank Yizhong Zhang for the early discussions. This work was supported in part by the NSF (VEC 1539014/ 1539099, IIS 1421435, CHS 1617236), NSFC (61532003, 61572507, 61622212), Google, Intel, Pixar, Amazon, and Facebook. Yifei Shi was supported by the China Scholarship Council.

References

1. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. In: Proc. UIST. (2011) 559–568
2. Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C.: BundleFusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. ACM Trans. on Graph. **36**(3) (2017) 24
3. Zhang, Y., Xu, W., Tong, Y., Zhou, K.: Online structure analysis for real-time indoor scene reconstruction. ACM Transactions on Graphics (TOG) **34**(5) (2015) 159
4. Halber, M., Funkhouser, T.: Fine-to-coarse global registration of rgb-d scans. arXiv preprint arXiv:1607.08539 (2016)
5. Lee, J.K., Yea, J.W., Park, M.G., Yoon, K.J.: Joint layout estimation and global multi-view registration for indoor reconstruction. arXiv preprint arXiv:1704.07632 (2017)
6. Ma, L., Kerl, C., Stückler, J., Cremers, D.: Cpa-slam: Consistent plane-model alignment for direct rgb-d slam. In: Robotics and Automation (ICRA), 2016 IEEE International Conference on, IEEE (2016) 1285–1291
7. Trevor, A.J., Rogers, J.G., Christensen, H.I.: Planar surface slam with 3d and 2d sensors. In: Robotics and Automation (ICRA), 2012 IEEE International Conference on, IEEE (2012) 3041–3048
8. Zhang, E., Cohen, M.F., Curless, B.: Emptying, refurnishing, and relighting indoor spaces. ACM Transactions on Graphics (TOG) **35**(6) (2016) 174
9. Huang, J., Dai, A., Guibas, L., Nießner, M.: 3dlite: Towards commodity 3d scanning for content creation. ACM Transactions on Graphics 2017 (TOG) (2017)
10. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. (2017)
11. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: Proc. IROS. (Oct. 2012)
12. Choi, S., Zhou, Q.Y., Koltun, V.: Robust reconstruction of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 5556–5565
13. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion: Real-time dense surface mapping and tracking. In: Proc. ISMAR. (2011) 127–136
14. Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M.: Real-time 3d reconstruction at scale using voxel hashing. ACM TOG (2013)
15. Chen, J., Bautembach, D., Izadi, S.: Scalable real-time volumetric surface reconstruction. ACM TOG **32**(4) (2013) 113
16. Keller, M., Lefloch, D., Lambers, M., Izadi, S., Weyrich, T., Kolb, A.: Real-time 3d reconstruction in dynamic scenes using point-based fusion. In: Proc. 3DV, IEEE (2013) 1–8
17. Steinbruecker, F., Sturm, J., Cremers, D.: Volumetric 3d mapping in real-time on a cpu. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), Hongkong, China (2014)
18. Whelan, T., Leutenegger, S., Salas-Moreno, R.F., Glocker, B., Davison, A.J.: ElasticFusion: Dense SLAM without a pose graph. In: Proc. RSS, Rome, Italy (July 2015)

19. Wang, R., Schwörer, M., Cremers, D.: Stereo dso: Large-scale direct sparse visual odometry with stereo cameras. arXiv preprint arXiv:1708.07878 (2017)
20. Park, J., Zhou, Q.Y., Koltun, V.: Colored point cloud registration revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 143–152
21. Lowe, D.G.: Object recognition from local scale-invariant features. In: Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Volume 2., Ieee (1999) 1150–1157
22. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. Computer vision–ECCV 2006 (2006) 404–417
23. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: Computer Vision (ICCV), 2011 IEEE international conference on, IEEE (2011) 2564–2571
24. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: Matchnet: Unifying feature and metric learning for patch-based matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3279–3286
25. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: European Conference on Computer Vision, Springer (2016) 467–483
26. Byravan, A., Fox, D.: Se3-nets: Learning rigid body motion using deep neural networks. In: Robotics and Automation (ICRA), 2017 IEEE International Conference on, IEEE (2017) 173–180
27. Zeng, A., Song, S., Niessner, M., Fisher, M., Xiao, J., Funkhouser, T.: 3DMatch: Learning local geometric descriptors from rgb-d reconstructions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1802–1811
28. Schmidt, T., Newcombe, R., Fox, D.: Self-supervised visual descriptor learning for dense correspondence. IEEE Robotics and Automation Letters **2**(2) (2017) 420–427
29. Concha, A., Civera, J.: Dpptam: Dense piecewise planar tracking and mapping from a monocular sequence. In: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, IEEE (2015) 5686–5693
30. Dou, M., Guan, L., Frahm, J.M., Fuchs, H.: Exploring high-level plane primitives for indoor 3d reconstruction with a hand-held rgb-d camera. In: Asian Conference on Computer Vision, Springer (2012) 94–108
31. Hsiao, M., Westman, E., Zhang, G., Kaess, M.: Keyframe-based dense planar slam. In: Proc. International Conference on Robotics and Automation (ICRA), IEEE. (2017)
32. Pietzsch, T.: Planar features for visual slam. KI 2008: Advances in Artificial Intelligence (2008) 119–126
33. Proença, P.F., Gao, Y.: Probabilistic combination of noisy points and planes for rgb-d odometry. arXiv preprint arXiv:1705.06516 (2017)
34. Salas-Moreno, R.F., Glocken, B., Kelly, P.H., Davison, A.J.: Dense planar slam. In: Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on, IEEE (2014) 157–164
35. Taguchi, Y., Jian, Y.D., Ramalingam, S., Feng, C.: Point-plane slam for hand-held 3d sensors. In: Robotics and Automation (ICRA), 2013 IEEE International Conference on, IEEE (2013) 5182–5189
36. Weingarten, J., Siegwart, R.: 3d slam using planar segments. In: Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on, IEEE (2006) 3062–3067

37. Stuckler, J., Behnke, S.: Orthogonal wall correction for visual motion estimation. In: Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on, IEEE (2008) 1–6
38. Grisetti, G., Kummerle, R., Stachniss, C., Burgard, W.: A tutorial on graph-based slam. IEEE Intelligent Transportation Systems Magazine **2**(4) (2010) 31–43
39. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In: In the 12th International Symposium on Experimental Robotics (ISER, Citeseer (2010)
40. Zhou, Q.Y., Koltun, V.: Dense scene reconstruction with points of interest. ACM Transactions on Graphics (TOG) **32**(4) (2013) 112
41. Zhou, Q.Y., Park, J., Koltun, V.: Fast global registration. In: European Conference on Computer Vision, Springer (2016) 766–782
42. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 815–823
43. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
44. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. arXiv preprint arXiv:1708.02002 (2017)
45. Agarwal, S., Mierle, K.: Ceres solver: Tutorial & reference. Google Inc **2** (2012) 72
46. Chang, A., Dai, A., Funkhouser, T., , Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: Proceedings of the International Conference on 3D Vision (3DV). (2017)
47. Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., Burgard, W.: An evaluation of the rgb-d slam system. In: Robotics and Automation (ICRA), 2012 IEEE International Conference on, IEEE (2012) 1691–1696