ObjectMatch: Robust Registration using Canonical Object Correspondences

Can Gümeli   Angela Dai   Matthias Nießner
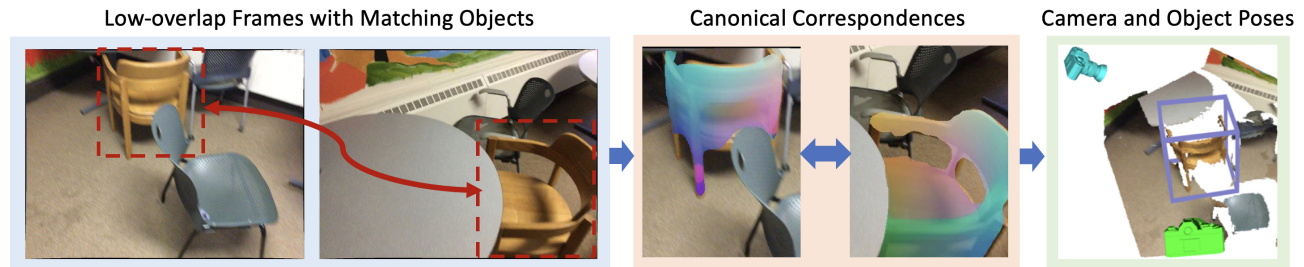
Technical University of Munich

Figure 1. Modern camera pose estimation relies on feature matching between overlapping frames - in this work, we present ObjectMatch to find correspondences between frames with little or no overlap by predicting semantic mappings through canonical object correspondences. The images above share no direct overlap, yet our method establishes indirect correspondences, thus enabling a successful registration.

Abstract

We present ObjectMatch[1], a semantic and object-centric camera pose estimator for RGB-D SLAM pipelines. Modern camera pose estimators rely on direct correspondences of overlapping regions between frames; however, they cannot align camera frames with little or no overlap. In this work, we propose to leverage indirect correspondences obtained via semantic object identification. For instance, when an object is seen from the front in one frame and from the back in another frame, we can provide additional pose constraints through canonical object correspondences. We first propose a neural network to predict such correspondences on a per-pixel level, which we then combine in our energy formulation with state-of-the-art keypoint matching solved with a joint Gauss-Newton optimization. In a pairwise setting, our method improves registration recall of state-of-the-art feature matching, including from 24% to 45% in pairs with 10% or less inter-frame overlap. In registering RGB-D sequences, our method outperforms cutting-edge SLAM baselines in challenging, low-frame-rate scenarios, achieving more than 35% reduction in trajectory error in multiple scenes.

---

[1]https://cangumeli.github.io/ObjectMatch/

1. Introduction

RGB-D registration and 3D SLAM has been a fundamental task in computer vision, with significant study and enabling many applications in mixed reality, robotics, and content creation. Central to both state-of-the-art traditional and learning-based camera pose estimation is establishing correspondences between points in input frames. However, correspondence estimation remains quite challenging when there is little or no overlap between frames.

In contrast, humans can easily localize across these challenging scenarios by leveraging additional semantic knowledge – in particular, by further localizing at the level of objects and identifying matching objects between views. For instance, when observing a chair from the back and the side (e.g., in Figure 1), view overlap is minimal (or even no view overlap), resulting in failed registration from keypoint matching. However, the semantic knowledge of the chair and its object pose nonetheless enables humans to estimate the poses from which the front and side views were taken. Thus, we propose to take a new perspective on camera pose estimation and imbue camera registration with awareness of such semantic correspondences between objects for robust performance in these challenging scenarios.

To this end, we propose ObjectMatch, a new paradigm for camera pose estimation leveraging canonical object correspondences in tandem with local keypoint correspondences between views. This enables significantly more robust registration under a variety of challenging scenar-

ios, including low view overlap. For a sequence of input frames, ObjectMatch learns to semantically identify objects across frames, enabling a compact, global parameterization of 9-DoF object poses. Object correspondences are established through predicting normalized object coordinates [43], dense correspondences from object pixels to a canonically oriented space for each object. We then formulate a joint camera and object pose optimization that constrains object correspondences indirectly, operating irrespective of the shared visibility of image regions. Our approach is complementary to state-of-the-art SLAM methods, and we leverage our energy formulation to complement state-of-the-art keypoint matching [10, 11, 36] in a joint Gauss-Newton optimization.

Our method outperforms strong baselines in both pairwise registration and registration of RGB-D frame sequences. In pairwise registration of challenging ScanNet [9] image pairs, we improve pose recall from 24% to 45% when the overlap is below 10%. On sequence registration of room-scale RGB-D scenes, our method outperforms various strong baselines in difficult, low-frame-rate settings in several TUM-RGBD [39] and ScanNet [9] scenes, reducing the trajectory error by more than 35% in multiple challenging scenes.

To sum up, our main contributions include:

- An object-centric camera pose estimator that can handle low-overlap frame sets via indirect, canonical object correspondences established with predicted dense, per-pixel normalized object coordinates.

- A joint energy formulation that leverages semantic object identification and dense, normalized object coordinates corresponding to canonical object geometries.

- Our semantic grounding of object correspondences enables significantly more robust registration in low-overlap and low-frame-rate cases. ObjectMatch improves over state of the art from 24% to 45% registration recall of $\leq 10\%$ overlap frame pairs and achieves over 35% trajectory error reduction in several challenging sequences.

2. Related Work

RGB-D Registration and SLAM. In recent years, there have been many advances in indoor RGB-D reconstruction. Earlier RGB-D fusion approaches focus on frame-to-model camera tracking [20, 29]. To handle the loop closures better, more recent SLAM systems introduce explicit strategies or global optimization methods for handling loop closures through global optimization [5, 10, 28, 37, 44] to fix tracking errors. More recently, deep learning techniques have been applied to registration and SLAM scenarios, with methods ranging from geometric point cloud registration [6, 19, 31] as well as neural field based SLAM techniques [18, 40, 48].

Despite all the successes in RGB-D SLAM and registration, the task is still challenging since incomplete loop closures observed via low-overlap frames cannot be handled, and most SLAM methods require a very high overlap between consecutive frames to track cameras accurately.

Feature Matching. Modern RGB(-D) camera pose estimators rely on a feature-matching backbone. Classical global registration techniques [5,46] use FPFH features [34] over point cloud fragments. On the other hand, many global RGB-D SLAM techniques rely on sparse color features [10, 28]. While being successful in many scenarios, conventional feature matching often fails when the interframe overlap is low. Therefore, deep learning techniques have been utilized for predicting overlapping regions based on geometry or color. On the geometric side, Deep Global Registration [6] predicts overlapping point features using nearest neighbor search over learned geometric features [7]. Methods such as PREDATOR [19] and Geometric Transformer [31] use attention mechanisms to target overlapping regions for registration. In the domain of color features, SuperPoint and SuperGlue [11, 36] build a formative approach in GNN-based keypoint feature matching. Methods such as LoFTR [42] introduce more dense and accurate sub-pixel level matching. Despite being very successful in handling wide-baseline scenarios, learned feature matching still requires a significant amount of shared visibility and geometric overlap.

Camera Pose Estimation with Semantic Cues. Several methods have been developed to incorporate semantic priors to improve low-overlap registration. PlaneMatch [38] proposed coplanarity priors for handling loop closures and low-overlap cases. Our method instead leverages object-centric constraints, exploiting the power of semantic object recognition. Another related direction is feature hallucination by leveraging image and object semantics. NeurHal [13] focuses on correspondence hallucination using image inpainting and outpainting, formulating a PnP optimization over hallucinated matches. Virtual Correspondence (VC) [24] introduces a human-centric approach that leverages hallucinated object volumes to form virtual epipolar constraints between frames. In contrast, we use indirect instead of direct correspondences that do not require hallucinated object volumes or image regions. Furthermore, our method works on a diverse set of furniture categories while VC focuses on humans. Pioneered by SLAM++ [35], there is also a rich literature of object-centric SLAM solutions, e.g., [26, 41]. Such SLAM methods leverage local, per-frame poses of objects to establish constraints; instead, we develop a global object pose optimization that is more robust against occluded viewpoints.

Object Pose Estimation using Normalized Object Coordinates. 3D object pose estimation has been widely stud-
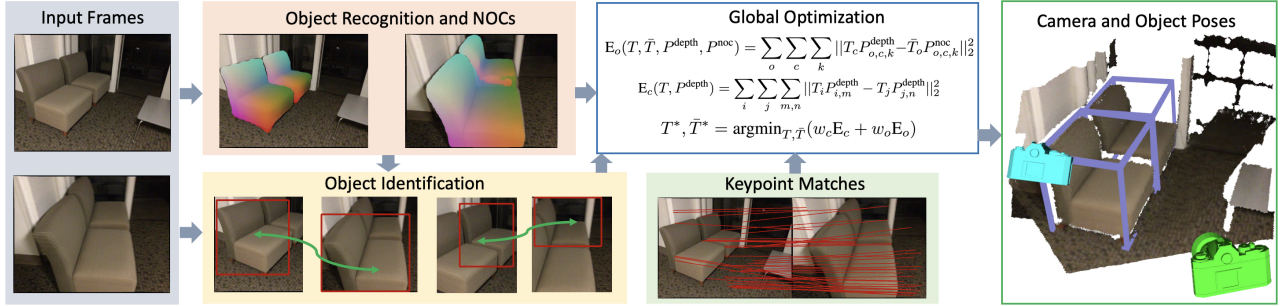
Figure 2. Overview of our approach to incorporate object correspondence grounding in global pose estimation. From a set of input RGB-D frames, ObjectMatch predicts object instances for each frame with dense normalized object correspondences. The predicted object instances are used to identify objects across frames, forming indirect object correspondences. We combine object correspondences with SuperGlue [11, 36] keypoint matches in a joint energy optimization that yields both camera and object poses in a global registration.

ied from RGB and RGB-D inputs. Normalized Object Co-ordinate Space (NOCS) [43] was proposed to form dense correspondences from input RGB-D frames to canonical object geometries, enabling better generalization than direct regression of object poses. End2End SOCs [2] formulated a NOC-based approach for CAD retrieval and alignment to 3D scans, using a differentiable Procrustes algorithm. To enable CAD alignment to single RGB images, ROCA [14] leveraged NOC space in combination with predicted depths, formulating a robust differentiable Procrustes optimization [14]. Seeing Behind Objects [27] further leveraged NOC correspondences both to obtain local object poses and object completion for re-identification for RGB-D multi-object tracking. Wide Disparity Re-localization [26] uses the NOC predictions from [43] to construct an object-level map for re-localization in SLAM. In contrast to these approaches that focus on individual object poses, we use NOC correspondences directly in a multi-frame, global camera, and object pose optimization.

3. Method

3.1. Problem Setup

Given K RGB-D frames $\{(I_1^c, I_1^d), ..., (I_K^c, I_K^d)\}$, we aim to optimize their 6-DoF camera poses $T_c = \{T_2, ..., T_K\}$, assuming the first frame is the reference, i.e., $T_1 = \mathbb{I}$. A 6-DoF camera pose T_i is represented by Euler angles γ and translations t, $T_i = (\gamma_x, \gamma_y, \gamma_z, t_x, t_y, t_z)$.

We also parameterize global, 9-DoF object poses, $\bar{T}_o = (\gamma_x, \gamma_y, \gamma_z, t_x, t_y, t_z, s_x, s_y, s_z)$, comprising 6-DoF angles and translations, and 3-DoF anisotropic scales s.

We formulate a joint energy optimization of the form:

$$T^*, \bar{T}^* = \text{argmin}_{T, \bar{T}}(E_c(T, M) + E_o(T, \bar{T}, N)) \quad (1)$$

where M are inter-frame feature matches and N are intra-frame canonical object correspondences established with normalized object coordinates (NOCs) that densely map
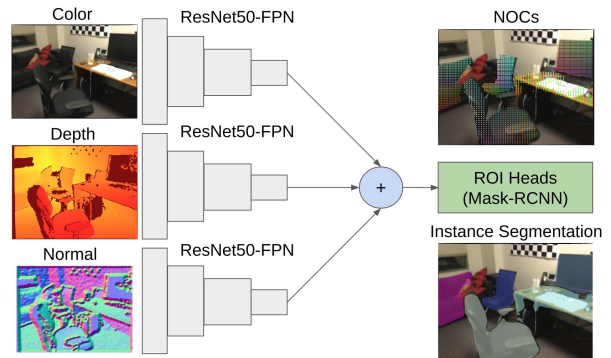

Figure 3. Multi-modal object recognition and NOC prediction. Our ResNet-FPN [16, 21] backbone takes color, reversed jet colored depth [12], and 2x downsampled colored 3D depth normals, and produces multi-scale features by averaging different input encodings. From the obtained features, our method recognizes objects and predicts NOCs for each object, based on a Mask-RCNN [15]-style prediction.

to the canonical space of an object in $[-0.5, 0.5]^3$. E_c is the feature-matching energy function, and E_o is our object-centric energy function. Since robust feature matching and optimization are readily available off the shelf [10, 11, 31, 36], our method focuses on building the object-centric energy E_o. To this end, we need two function approximators, realized via deep neural networks: (1) a learned model for object recognition and NOC prediction, and (2) a learned model for object identification. The realization of these networks is described in Sections 3.2 and 3.3, respectively, and energy function Eq. 1 in Section 3.4. An overview of our approach is visualized in Figure 2.

3.2. Predicting Object Correspondences

To obtain object constraints in our final optimization, we recognize objects via object detection and instance segmentation, and predict object correspondences as dense NOCs [43] for each object, as shown in Figure 3. We build on a Mask-RCNN [15] with ResNet50-FPN [16, 21] back-

bone, pre-trained on ImageNet [33] and COCO [22].

To input the depth of the RGB-D frames, we propose a modified, multi-modal ResNet50-FPN [12, 16, 21] backbone. Our backbone takes 480x640 color, 480x640 reverse jet-colored depth, and 240x320 colored depth normals as input. We average the resulting FPN features to a single feature pyramid:

$$G = \frac{\text{FPN}^c(I^c) + \text{FPN}^d(I^d) + \text{U}(\text{FPN}^n(I^n))}{3}, \quad (2)$$

where I^c, I^d, I^n are color, depth, and normal images, $\text{FPN}^c, \text{FPN}^d, \text{FPN}^n$ are the corresponding ResNet50-FPN backbones, and U is an upsampling operator to match the normal features' spatial size with others. This enables fine-tuning the pre-trained bounding box, class, and instance segmentation heads of Mask-RCNN [15], while also exploiting depth information. We use symmetrically structured FPNs, all pre-trained on ImageNet and COCO as initialization, but without any parameter sharing.

To obtain object correspondences, we establish mappings from detected objects to their canonical spaces, in the form of dense NOCs. That is, for each pixel in the object's instance mask, we predict 3D points P_o^{noc} using a fully convolutional network:

$$P_o^{\text{noc}} = \text{FCN}(G_o), \ \ p \in [-0.5, 0.5]^3 \ \forall \ p \in P_o^{\text{noc}}. \quad (3)$$

We optimize an ℓ_1 loss L_{noc} using ground-truth NOCs $P^{\text{noc-gt}}$,

$$L_{\text{noc}} = \sum_o \sum_i ||P_{o,i}^{\text{noc}} - P_o^{\text{noc-gt}}||_1. \quad (4)$$

Since symmetric objects [1, 43] induce ambiguities in NOCs (e.g., a round table), we classify symmetry type of objects (round, square, rectangle, non-symmetric), $c_{\text{sym}} = \text{MLP}^{\text{sym}}(G_o)$, optimized using a cross-entropy loss L_{sym}. We also make L_{noc} symmetry aware, taking the minimum ℓ_1 difference over the set of correct NOCs [43]. We use non-symmetric objects during inference to avoid inconsistent NOCs across views.

In addition to NOCs, we also regress anisotropic 3D object scales s_o using a fully connected network, $s_o = \text{MLP}^{\text{scale}}(G_o)$, and optimize s_o with an ℓ_1 loss L_{scale}. The object scale enables holistic object pose understanding within each frame and helps to filter potential object matches across views using scale consistency.

Finally, to make our NOC-depth correspondences least-squares friendly for our desired optimization, we also introduce a per-frame differentiable Procrustes objective L_{proc}, using a differentiable Kabsch solver [32] to obtain local object rotations and translations:

$$R_o^*, t_o^* = \text{argmin}_{R_o, t_o} (\sum_i ||R_o(P_{o,i}^{\text{noc}} \odot s_o) + t_o - P_{o,i}^{\text{depth}}||_2^2) \quad (5)$$
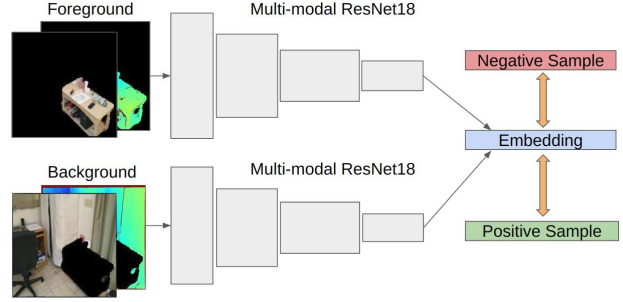

Figure 4. Our foreground/background metric-learning encoder for object matching, inspired by re-OBJ [3]. Using the detected and segmented objects from the model in Section 3.2, we encode foreground and background regions of objects, using light-weight, multi-modal ResNet18 encoders on the RGB-D features.

for each object o, where P_o^{depth} are back-projected input RGB-D depths corresponding to the object's predicted instance mask in its region of interest, and $\odot$ denotes element-wise multiplication. We train the local object poses with

$$L_{\text{proc}} = w_r \sum_o ||R_o^* - R_o^{\text{gt}}||_1 + w_t \sum_o ||t_o^* - t_o^{\text{gt}}||_2^2. \quad (6)$$

Our full loss used for training is then

$$L = L_m + w_n L_{\text{noc}} + w_s L_{\text{scale}} + w_{\text{sym}} L_{\text{sym}} + w_p L_{\text{proc}}, \quad (7)$$

where L_m is the sum of Mask-RCNN losses [15] and w are scalar weights balancing additional losses.

Implementation. We use the augmented 400k ScanNet train image split for training [9, 14], with Scan2CAD labels of the most common 9 furniture categories following [1, 14, 25]. We train a standard Detectron2 Mask-RCNN pipeline [15, 45] with 1k warm-up iterations, 0.003 base learning rate, and learning rate decays at 60k and 100k iterations with a total of 120k training iterations.

3.3. Matching Object Instances

In our global pose optimization formulation, the relation between frames is formed via a global identification of objects across frames. To enable such identification without any heuristic spatial assumptions, we train a simple metric learner that ranks object semantic similarities across frames. Our object matching is shown in Figure 4.

We characterize objects for matching by their respective RGB-D features in the object instance mask, in addition to the global context in the 5-times upscaled object bounding box. All inputs are resized to 224x224 and input to a lightweight ResNet18 [16] backbone pre-trained on ImageNet [33].

Similar to object detection, we employ two backbones for color and colored depth inputs. We omit normal input in object matching, as it empirically did not provide any

benefit. For each input modality, we train two ResNet18 backbones for masked and inverse-masked crops, namely foreground (object) and background (context) encodings,

$$e = \mathrm{MLP}([\mathrm{RN}^c(F_c), \mathrm{RN}^c(B_c)] + [\mathrm{RN}^d(F_d), \mathrm{RN}^d(B_d)])$$
(8)

where RN are ResNet18s, MLP is a fully connected network, F, B are foreground and background crops for color (c) and depth (d), and e is the object embedding vector.

Given an anchor object embedding e_a, a correctly matching object embedding e_p, and a negative example e_n, we train our metric learning using a triplet margin loss [17]:

$$L_{\mathrm{tri}} = \max(\mathrm{d}(e_a, e_p) - \mathrm{d}(e_a, e_n) + 1.0, 0),$$
(9)

where d is the ℓ_2 distance.

We only consider triplets from the same category, as the object recognition pipeline provides classification. At inference time, we match the best instances using the Hungarian algorithm and apply a threshold $d(e_i, e_j) < \alpha$ for the matching object pairs from the same class. This semantic matching can be scaled to multiple frames via, e.g., object tracking with re-identification, or in our case, a simple pose graph optimization over frame pairs.

Implementation. We implement the identification network using PyTorch [30] and train it on ScanNet data [9]. We train the network for 100k iterations with a batch size of 8 using a momentum optimizer with a learning rate 1e-4 and momentum 0.9.

3.4. Energy Optimization

We realize the joint energy minimization in Eq. 1 using keypoint and NOC constraints.

Using the predicted NOC constraints with back-projected depths, we can re-write the E_o in Eq. 1 as:

$$\mathrm{E}_o(T, \bar{T}, P^{\mathrm{depth}}, P^{\mathrm{noc}}) = \sum_o \sum_c \sum_k ||T_c P^{\mathrm{depth}}_{o,c,k} - \bar{T}_o P^{\mathrm{noc}}_{o,c,k}||_2^2$$
(10)

where $\bar{T}$, T represent 9-DoF and 6-DoF camera and object transformations, respectively, and subscripts o, c, k correspond to objects, cameras (frames), and points (pixels) within the frames, respectively. Here, object indices are determined by the object identification, and global object poses $\bar{T}$ indirectly constrain frames to each other without any explicit inter-frame feature matching.

In many cases, object constraints alone may not be sufficient to optimize the camera pose (e.g., frames may not share matching objects together). However, our object-based optimization is fully complementary to classical feature matching, and we thus formulate our objective in combination with feature-matching constraints E_c:

$$\mathrm{E}_c(T, P^{\mathrm{depth}}) = \sum_i \sum_j \sum_{m,n} ||T_i P^{\mathrm{depth}}_{i,m} - T_j P^{\mathrm{depth}}_{j,n}||_2^2 \quad (11)$$

Our method is agnostic to the feature matcher, whether classical or learning-based. In this work, we experiment with two different keypoint matching systems to realize E_c, namely SuperGlue [11, 36] and Geometric Transformer [31], both offering state-of-the-art indoor feature matching.

With both object and feature-matching constraints, we realize the desired joint energy formulation as

$$T^*, \bar{T}^* = \mathrm{argmin}_{T, \bar{T}}(w_c \mathrm{E}_c + w_o \mathrm{E}_o).$$
(12)

where w_c, w_o weight feature-matching, and object energies.

Since non-linear least squares problems can be sensitive to outliers, we additionally employ outlier removal. Similar to BundleFusion [10], we apply Kabsch filtering to both intra-frame and keypoint constraints, using the matching depth-NOC and depth-depth correspondences, respectively. That is, we iteratively solve an orthogonal Procrustes problem and only keep the correspondences that have lower optimization errors. We use a liberal 20cm threshold to handle wide-baseline frames. Objects and inter-frame matches are rejected if the number of NOCs is < 15 and the number of keypoints is < 5, respectively.

To solve this least-squares problem, we use a Gauss-Newton optimizer. To handle outliers during this global optimization, we remove $> 15cm$ error residuals during optimization. As global optimization can produce coarse results, we further apply an ICP refinement.

To handle SLAM-style sequence registration, we use a state-of-the-art global pose graph optimization [5, 47] over pairwise frame registration, using a single hierarchy level for simplicity. However, our method could be scaled to multi-frame optimization via object tracking and subdivided into multiple hierarchy levels to handle very large-scale scenes.

4. Experiments

4.1. Pairwise Registration

Evaluation Dataset. We evaluate pairwise camera pose estimation results on 1569 challenging and diverse frame pairs from the validation set of ScanNet [9]. Unlike previous works [19, 36], our evaluation includes a significant number of low-overlap pairs, with over 15% of pairs having $\leq 10\%$ overlap.

Optimization. To establish object correspondences across frame pairs in Eq. 10, we use the top-1 matching object with embedding distance threshold 0.05. In the absence of both keypoint matches and object matches, we consider object matches with distance < 0.15. Only objects of the same class label and predicted maximum scale ratio < 1.5 can match. We determine the best matching object based on the number of NOC-depth constraints after the filtering described in Section 3.4. We combine it with keypoint constraints and refine the results with ICP.

Classical Baselines. We compare our method against various hand-crafted feature-matching and registration baselines. These baselines include BundleFusion keypoint optimization with SIFT descriptors and Kabsch filtering [10, 23], (SIFT + BF), and two geometric global registration baselines, Fast Global Registration (Fast GR) [46] and Global Registration (GR) [5], based on approximate and exact RANSAC over FPFH [34] features, respectively. As a post-processing step, all methods are refined using ICP. We use Open3D [47] implementations for global registration baselines as well as ICP post-processing of all methods. Our custom implementation of SIFT + BF uses the SIFT implementation of OpenCV [4].

Learned Baselines. We also compare with pose estimators using learned models for feature or object matching: Geometric Transformer (GeoTrans) [31], which performs dense geometric feature matching, and SG + BF which leverages the powerful learned SuperGlue [36] and SuperPoint [11] feature matching approaches in combination with BundleFusion [10] Kabsch filtering and 3D Gauss-Newton optimization. We also use our network predictions to create a 3D object tracking baseline, *Object Track*, mimicking tracking-based object-SLAM systems. Object Track obtains local object poses and object matches from Sections 3.2 and 3.3, but instead uses relative local object poses to get the camera pose instead of a global energy optimization. All methods are refined using ICP as a post-processing step.

Evaluation Metrics. We use the Pose Recall metric for pose evaluation, following previous RGB-D pairwise registration works [6, 19]. To comprehensively capture performance, we employ several thresholds for absolute translation (in cm) and rotation (in °). Since absolute translation is more difficult than rotation, its thresholds are twice the angle thresholds, following [6].

Quantitative Results. In Table 1 and Table 2, we evaluate on ScanNet [9] validation frame pairs, measuring recall in various thresholds and overlap levels. Combined with state-of-the-art feature matching in a joint optimization, our method outperforms both classical and learning-based methods. Furthermore, the gap increases with lower overlap, as shown in Table 2 and Figure 6, since our method can leverage object-based correspondences to estimate alignment in the absence of keypoint matches. We also show that our global optimization (Ours (w/o keypoints)) significantly outperforms naive object-based tracking (Object Track), demonstrating the efficacy of our global optimization formulation. Our approach is designed to complement keypoint matching, notably improving the state of the art in combination with state-of-the-art geometric and RGB-D feature matching.

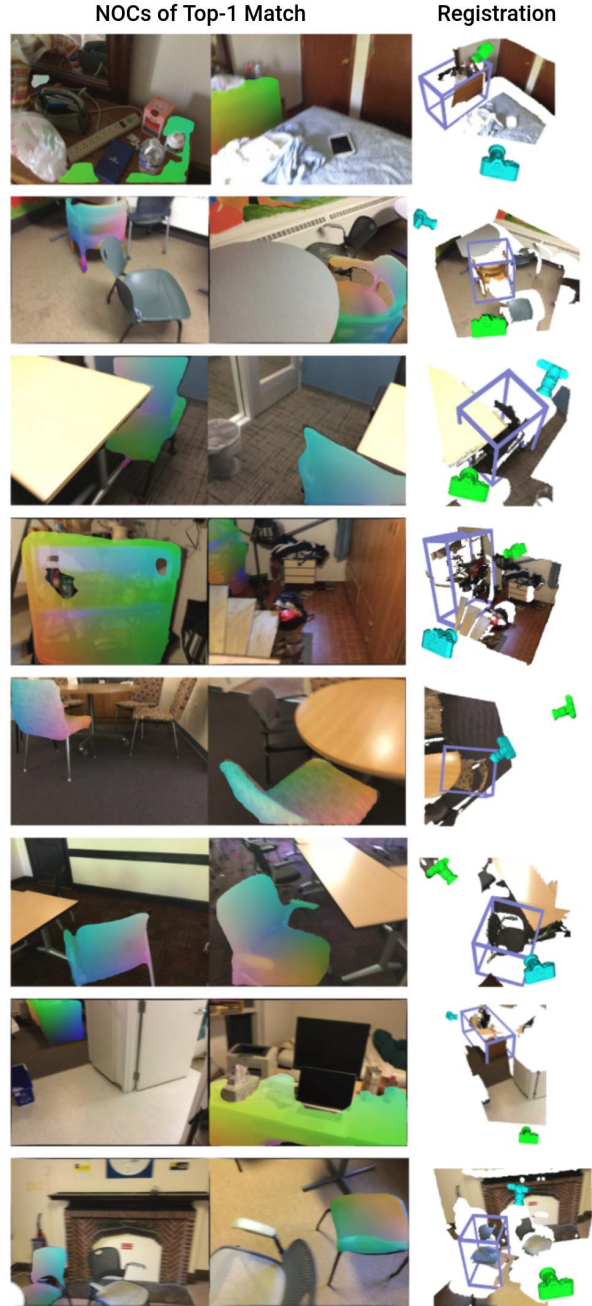**Analysis of Performance by Overlap.** In Table 2 and

Figure 5. Low-overlap registration on ScanNet [9], where traditional feature matching fails. Predicted NOC correspondences are visualized, along with object box and camera poses of the left and right images in green and blue, respectively.

Figure 6, we show 30cm, 15° recall performance at different geometric overlap rates, using the radius-based geometric overlap percentage measurement from [19] with 1cm threshold. While all methods' performances decrease with overlap, our method retains significantly more robustness. Classical keypoint matching (SIFT + BF, Fast GR,

Method	Pose Recall by Threshold		
	5°, 10cm	10°, 20cm	15°, 30cm
SIFT + BF [10, 23]	16.83	18.42	20.40
Fast GR [34, 46]	39.07	42.96	44.68
GR [5]	43.08	47.67	50.03
Object Track	37.41	43.21	45.19
GeoTrans [31]	67.11	74.57	76.93
SG + BF [10, 11, 36]	71.13	80.11	81.71
Ours (w/o keypoints)	56.41	63.67	66.99
Ours (w/ GeoTrans)	68.32	76.80	79.35
Ours (w/ SG + BF)	**74.25**	**84.58**	**87.06**

Table 1. Pose Recall results on ScanNet [9] validation images. Combined with state-of-the-art feature matching, our method outperforms various classical and learning-based baselines. Our approach complements both Geometric Transformer [31] (Ours (w/ GeoTrans)) and SuperGlue [36] (Ours (w/ SG + BF)) feature matches, notably improving pose recall.

Method	Recall by Overlap %		
	≤ 10	(10, 30)	≥ 30
SIFT + BF [10, 23]	1.30	2.20	33.59
Fast GR [34, 46]	0.00	13.97	69.04
GR [5, 34]	0.65	20.36	74.62
Object Track	8.44	31.54	58.86
GeoTrans [31]	22.08	64.27	93.11
SG + BF [10, 11, 36]	24.03	73.45	95.95
Ours (w/o keypoints)	32.47	53.29	80.31
Ours (w/ GeoTrans)	35.06	65.47	94.42
Ours (w/ SG + BF)	**45.45**	**81.44**	**97.16**

Table 2. Recall at 15°, 30cm by overlap percentage. Our method significantly outperforms strong baselines in challenging low-overlap frame pairs, almost doubling recall for overlap ≤10%.
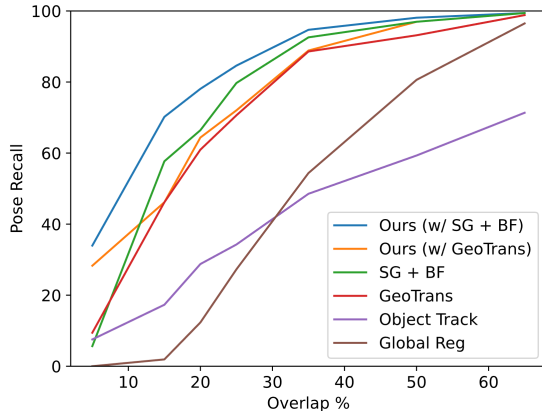

Figure 6. Change of pose performance by overlap percentage on ScanNet [9] validation pairs. Our method is significantly more robust against decreasing overlap compared to various classical and learned baselines.

SIFT) suffer strongly in low-overlap scenarios, which are challenging for hand-crafted descriptors. The learned approaches of GeoTrans and SG+BF maintain some performance with decreasing overlap but still suffer strongly. Our

incorporation of object-based reasoning effectively complements keypoint matching while significantly improving robustness to low overlap, obtaining a performance improvement from 24.03% to 45.45% for overlap ≤ 10%.

Qualitative Results. Figure 5 shows examples of low-overlap registration on ScanNet [9] frame pairs. Due to the minimal overlap, traditional keypoint matching cannot find sufficient correspondences, while our object grounding enables accurate camera pose estimation through indirect object constraints.

4.2. Registration of RGB-D SLAM Sequences

We further demonstrate the effectiveness of our method in an RGB-D SLAM setting, registering sequences of frames in room-scale scenes. In particular, we evaluate challenging, low-frame-rate settings that reflect common non-expert capture such as fast camera motion.

For optimization, we use a global pose graph optimization over pairwise registration similar to the multi-way global registration of [5, 6, 47], as it offers an off-the-shelf, simple, and robust solution. We refer to our supplementary material for further optimization details.

Datasets and Evaluation Metrics. We evaluate our method on TUM-RGBD [39] and ScanNet [9] scenes. For ScanNet, we use a set of 12 scenes from the validation and test sets, having a wide range of sizes, environments, and camera trajectories. We sample every 30th frame for TUM-RGBD (1Hz) and every 20th frame in ScanNet (1.5Hz). All methods use the same hyperparameters for each dataset, respectively. To evaluate registration quality, we use the standard root mean squared trajectory error used for TUM-RGBD evaluation [39].

Baselines. We implement various popular feature descriptors from cutting-edge SLAM systems in global pose graph optimization setups and thoroughly tune the hyperparameters of each method to support low-frame-rate sequences. We compare against Redwood [5], a global multi-way pose graph that uses FPFH [34] features. We also deploy robust RGB-D SIFT features used in BundleFusion [10, 23] in a global pose graph optimization, creating the SIFT + BF baseline. Finally, we use the state-of-the-art SuperGlue matches of SuperPoint features with BundleFusion's sparse 3D optimization, SG + BF, as described in Section 4.1.

Quantitative Results. In Table 3, we compare our method to state-of-the-art traditional and learned SLAM approaches. Across both TUM-RGBD and ScanNet validation and test scenes [9, 39], our object-grounded approach enables improved camera pose estimation. In particular, for the larger-scale ScanNet scenes, ObjectMatch achieves notable reductions in trajectory error, of 38%, 49%, 50%, and 44% in scenes 0423, 0430, 0718, and 0773.

Method	TUM Scene @ 1Hz					ScanNet Scene @ 1.5Hz											
	fr1/desk	fr1/xyz	fr2/xyz	fr2/desk	fr3/office	0011	0081	0169	0207	0423	0430	0461	0494	0718	0760	0773	0805
SIFT + BF [10,23]	30.60	3.49	2.73	157.02	38.01	222.49	108.03	126.57	104.64	160.69	248.00	20.90	33.98	81.48	38.03	22.47	86.80
Redwood [5,34,47]	3.10	1.97	1.88	140.17	175.64	164.01	50.96	47.24	47.17	88.43	156.76	10.47	9.77	133.53	6.55	43.19	253.39
SG + BF [10,11,36]	3.01	1.92	1.81	5.30	5.14	13.88	6.77	8.09	5.62	7.67	35.97	3.31	7.16	13.42	6.37	12.43	12.98
Ours (w/ SG + BF)	**3.01**	**1.92**	**1.81**	**5.26**	**4.92**	**12.53**	**5.73**	**6.85**	**5.32**	**4.75**	**18.23**	**2.67**	**6.85**	**6.70**	**5.88**	**7.00**	**10.86**

Table 3. ATE RMSE values (cm) on room-scale TUM-RGBD [39] and ScanNet [9] test scenes. We evaluate at challenging 1 FPS and 1.5 FPS by sampling every 30th and 20th frame from TUM-RGBD and ScanNet, respectively. ObjectMatch outperforms strong baselines leveraging classical and learning-based SLAM systems.
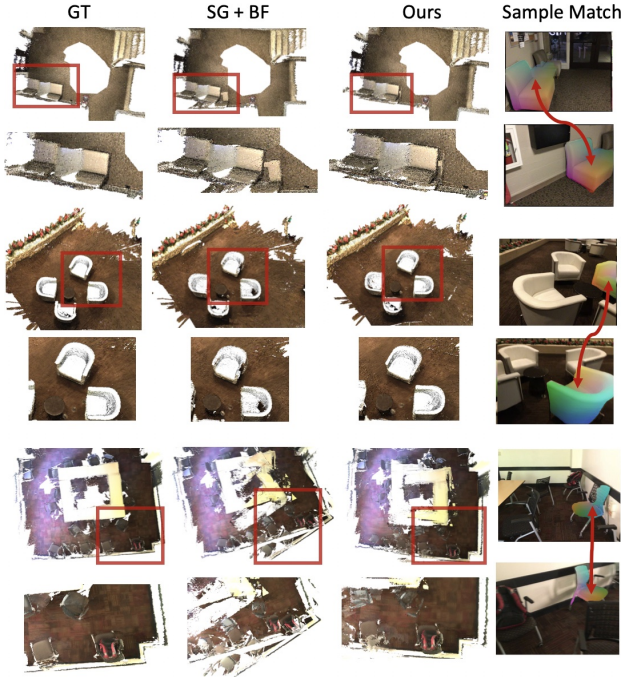

Figure 7. Qualitative comparison for registration of ScanNet scenes [9], in comparison with only feature matching SG + BF [10, 11, 36]. All results are visualized by volumetric fusion of the respective RGB-D frames with the estimated poses. Our optimized poses produce more accurate, clean, and consistent reconstruction, due to incorporating the object-based information in our global optimization. Example loop closures with matching object correspondences are shown on the right.

Qualitative Results. In Figure 7, we show ScanNet [9] scene reconstructions using estimated camera poses. RGB-D frames are fused using a scalable TSDF volumetric fusion [8,47]. Our approach complements state-of-the-art feature matching [10, 11, 36], enabling more accurate and consistent reconstructions using our indirect correspondences, in comparison to feature matching alone. We refer to the supplemental for additional qualitative sequence registration results.

Limitations and Future Work. ObjectMatch shows the capability of incorporating object semantics into global pose optimization; however, various limitations remain. In particular, our approach leverages higher-level seman-

tic correspondence given by objects for registration; however, in scenarios where object matches do not exist in frame views, we cannot leverage this constraint and instead use only feature matches. Further leveraging background information regarding shared structures seen from different views would provide additional semantic correspondence information to further complement registration. Additionally, while estimating canonical correspondences with objects can provide significant view information, object alignments can be somewhat coarse, while reasoning at a finer-grained level of object parts would provide more precise reasoning. Finally, we believe future work can reduce our method's system complexity by adopting a joint multi-frame SLAM optimization with fewer hyperparameters as well as a unified end-to-end architecture for learning object correspondences and identification together.

5. Conclusion

We have presented ObjectMatch which introduces a new paradigm for incorporating object semantics into camera pose registration. We leverage indirect, canonical object correspondences established with normalized object coordinates for camera pose optimization aspects in RGB-D registration and SLAM. To obtain these correspondences, we propose two multi-modal neural networks for object recognition with normalized object coordinate prediction and object identification. ObjectMatch operates in tandem with state-of-the-art feature matching in a joint Gauss-Newton optimization, with its object grounding enabling registration to handle frame pairs with very low to no shared visibility. As a result, ObjectMatch significantly improves the state-of-the-art feature matching in both pairwise and sequence registration of RGB-D frames, particularly in the challenging low-overlap regime. Overall, we hope our method opens up new possibilities in the context of leveraging semantic information for camera pose estimation.

References

[1] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2614–2623, 2019. 4

[2] Armen Avetisyan, Angela Dai, and Matthias Nießner. End-to-end cad model retrieval and 9dof alignment in 3d scans. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pages 2551–2560, 2019. 3

[3] Vaibhav Bansal, Stuart James, and Alessio Del Bue. re-obj: Jointly learning the foreground and background for object instance re-identification. In *International Conference on Image Analysis and Processing*, pages 402–413. Springer, 2019. 4

[4] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 6

[5] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5556–5565, 2015. 2, 5, 6, 7, 8

[6] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2514–2523, 2020. 2, 6, 7

[7] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *ICCV*, 2019. 2

[8] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312. ACM, 1996. 8

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 4, 5, 6, 7, 8

[10] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 2, 3, 5, 6, 7, 8

[11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2, 3, 5, 6, 7, 8

[12] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687. IEEE, 2015. 3, 4

[13] Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. Visual correspondence hallucination. In *International Conference on Learning Representations*, 2022. 2

[14] Can Gümeli, Angela Dai, and Matthias Nießner. Roca: Robust cad model retrieval and alignment from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4022–4031, 2022. 3, 4

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 4

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4

[17] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015. 5

[18] Jiahui Huang, Shi-Sheng Huang, Haoxuan Song, and Shi-Min Hu. Di-fusion: Online implicit 3d reconstruction with deep priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8932–8941, 2021. 2

[19] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4267–4276, June 2021. 2, 5, 6

[20] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 2

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3, 4

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[23] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 6, 7, 8

[24] Wei-Chiu Ma, Anqi Joyce Yang, Shenlong Wang, Raquel Urtasun, and Antonio Torralba. Virtual correspondence: Humans as a cue for extreme-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15924–15934, 2022. 2

[25] Kevis-Kokitsi Maninis, Stefan Popov, Matthias Niesser, and Vittorio Ferrari. Vid2cad: Cad model alignment using multi-view constraints from videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[26] Yuhang Ming, Xingrui Yang, and Andrew Calway. Object-augmented rgb-d slam for wide-disparity relocalisation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2203–2209. IEEE, 2021. 2, 3

[27] Norman Muller, Yu-Shiang Wong, Niloy J Mitra, Angela Dai, and Matthias Nießner. Seeing behind objects for 3d multi-object tracking in rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6071–6080, 2021. 3

[28] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 2

[29] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):1–11, 2013. 2

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[31] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11143–11152, 2022. 2, 3, 5, 6, 7

[32] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 4

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 4

[34] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009. 2, 6, 7, 8

[35] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1352–1359, 2013. 2

[36] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2, 3, 5, 6, 7, 8

[37] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 134–144, 2019. 2

[38] Yifei Shi, Kai Xu, Matthias Niessner, Szymon Rusinkiewicz, and Thomas Funkhouser. Planematch: Patch coplanarity prediction for robust rgb-d reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–766, 2018. 2

[39] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012. 2, 7, 8

[40] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 2

[41] Edgar Sucar, Kentaro Wada, and Andrew Davison. NodeSLAM: Neural object descriptors for multi-view shape reconstruction. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. 2

[42] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2

[43] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 2, 3, 4

[44] T Whelan, S Leutenegger, RF. Salas-Moreno, B Glocker, and AJ. Davison. Elasticfusion: Dense slam without a pose graph. In *Robotics: Science and Systems*, 2015. 2

[45] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 4

[46] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *European conference on computer vision*, pages 766–782. Springer, 2016. 2, 6, 7

[47] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 5, 6, 7, 8

[48] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 2