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, 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 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 rely on 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 inter-frame 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-
ied from RGB and RGB-D inputs. Normalized Object Coordinate 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 $T_e = \{T_2, ..., T_K\}$, assuming the first frame is the reference, i.e., $T_1 = I$. A 6-DoF camera pose $T_i$ is represented by Euler angles $\gamma$ 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, $T_o = (\gamma_x, \gamma_y, \gamma_z, t_x, t_y, t_z, s_x, s_y, s_z)$, comprising 6-DoF angles and translations, and 3-DoF anisotropic scales $s$.

We formulate a joint energy optimization of the form:

$$T^*, \bar{T}^* = \arg \min_{T, \bar{T}} (E_c(T, M) + E_o(T, \bar{T}, N))$$  \hspace{1cm} (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 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:

\[ G = \frac{\text{FPN}^{c}(I^{c}) + \text{FPN}^{d}(I^{d}) + U(\text{FPN}^{n}(I^{n}))}{3}, \]

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:

\[ P_{o}^{noc} = \text{FCN}(G_{o}), \quad p \in [-0.5, 0.5]^{3} \forall p \in P_{o}^{noc}. \]

We optimize an \( \ell_{1} \) loss \( L_{noc} \) using ground-truth NOCs \( P_{noc\text{-}gt} \),

\[ L_{noc} = \sum_{o} \sum_{i} \left\| P_{o,i}^{noc} - P_{o,i}^{noc\text{-}gt} \right\|_{1}. \]

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), \( c_{sym} = \text{MLP}^{sym}(G_{o}) \), optimized using a cross-entropy loss \( L_{sym} \). We also make \( L_{noc} \) symmetry aware, taking the minimum \( \ell_{1} \) 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, \( s_{o} = \text{MLP}^{scale}(G_{o}) \), and optimize \( s_{o} \) with an \( \ell_{1} \) 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:

\[ R_{o,i}^{*}, t_{o,i}^{*} = \arg\min_{R_{o,i}, t_{o,i}} \left( \sum_{o} \left\| R_{o} (P_{o,i}^{noc} \odot s_{o}) + t_{o} - P_{o,i}^{depth} \right\|_{2}^{2} \right) \]

for each object \( o \), where \( P_{depth}^{o} \) 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

\[ L_{proc} = w_{r} \sum_{o} \left\| R_{o}^{*} - R_{o}^{gt} \right\|_{1} + w_{t} \sum_{o} \left\| t_{o}^{*} - t_{o}^{gt} \right\|_{2}. \]

Our full loss used for training is then

\[ L = L_{m} + w_{n} L_{noc} + w_{s} L_{scale} + w_{sym} L_{sym} + w_{p} L_{proc}, \]

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,
\[ e = \text{MLP}([\text{RN}^c(F_c), \text{RN}^c(B_c)] + [\text{RN}^d(F_d), \text{RN}^d(B_d)]) \]
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]:
\[ L_{\text{tri}} = \max(d(e_a, e_p) - d(e_a, e_n) + 1, 0), \]
where \( d \) is the \( \ell_2 \) 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:
\[ E_o(T, \bar{T}, P^{\text{depth}}, P^{\text{noc}}) = \sum_{o} \sum_{c} \sum_{k} ||T_o P^{\text{depth}}_{o,c,k} - \bar{T}_o P^{\text{noc}}_{o,c,k}||^2 \]
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 \( E_c \):
\[ E_c(T, P^{\text{depth}}) = \sum_{i} \sum_{j} \sum_{m,n} ||T_i P^{\text{depth}}_{i,m,n} - T_j P^{\text{depth}}_{j,m,n}||^2 \]

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 \( E_o \), 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
\[ T^*, \bar{T}^* = \arg\min_{T, \bar{T}} (w_o E_o + w_c E_c). \]
where \( w_o, w_c \) 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 [33] 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, 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,
Pose Recall by Threshold

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

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%.

### 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
---|---|---
SIFT + BF [10, 23] | fr1/desk 30.60 0011 | fr2/desk 108.03 0081 |
Redwood [5, 34, 47] | fr1/xyz 3.49 0169 | fr2/xyz 2.73 0207 |
SG + BF [10, 11, 36] | fr1/office 157.02 0423 | fr3/office 38.01 0430 |
Ours (w/ SG + BF) | 0011 0461 0494 | 0081 0494 0718 |
| 0169 0430 | 0169 0430 |
| 0207 0461 | 0207 0461 |
| 0423 0494 | 0423 0494 |
| 0011 0081 0169 0207 0423 0430 0461 0494 0718 0760 0773 0805 | 0011 0081 0169 0207 0423 0430 0461 0494 0718 0760 0773 0805 |

**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 semantic 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.

**Acknowledgements.** This work was supported by the ERC Starting Grant Scan2CAD (804724), the Bavarian State Ministry of Science and the Arts coordinated by the Bavarian Research Institute for Digital Transformation (bidt), the German Research Foundation (DFG) Grant “Making Machine Learning on Static and Dynamic 3D Data Practical”, and the German Research Foundation (DFG) Research Unit “Learning and Simulation in Visual Computing.”
References


