Semantic-only Visual Odometry based on dense class-level segmentation

Howard Mahé*†, Denis Marraud* and Andrew I. Comport†
*Airbus Defence & Space, Elancourt, France – name.surname@airbus.com
†CNRS-I3S, Université Côte d'Azur, Sophia Antipolis, France – surname@i3s.unice.fr

***Abstract*—This paper proposes a novel approach called Semantic Visual Odometry (SemVO) which incorporates class-level consistency priors into the problem of 6-DoF Visual Odometry. Dense class-level labels are learnt for each pixel of the image using a CNN trained for semantic segmentation. A semantic error is formulated penalising the sum of squared differences (SSD) on class-level feature maps extracted from the decoder of a RefineNet. It will be shown how the proposed approach allows dense RGB-D camera tracking using solely a semantic error term. SemVO is evaluated on the ScanNet dataset and the results demonstrate how the number of classes affects performance. Results are also provided showing how best to fuse the new error function with classic dense photometric and geometric methods. Finally, it is demonstrated that SemVO improves over standard approaches for large camera motion applications.**

I. Introduction

Visual Odometry (VO), or visual ego-motion estimation, is the problem of determining the pose of a camera purely from vision. This technique is a key building block of modern SLAM (Simultaneous Localisation and Mapping) systems.

In recent years, Deep Convolutional Neural Networks (DCNN) have largely dominated most computer vision problems including perception tasks: image classification [1], object detection [2], semantic segmentation [3]–[5] ; and more recently even geometric tasks: dense depth estimation from single image [6], optical flow [6], relocalisation [7]. Recently, learning-based approaches have been applied to monocular VO by predicting the dense depth map from single image to retrieve the structure of the scene, reducing the problem to RGB-D image alignment. However, these methods were mainly benchmarked on constrained images from the KITTI dataset.

This paper focuses on building a semantic RGB-D VO for indoor environments. Even if some works have developed RGB-D VO based on semantic primitives such as points and line segments [8], planes [9] or objects [10], in this work, it is proposed to solve the problem of 6-DoF (degrees of freedom) ego-motion estimation by *directly* exploiting the high-level perceptual information learned by a CNN trained for semantic segmentation using a traditional dense geometric RGB-D VO.

The main contributions are threefold: 1) the 6-DoF camera pose estimation is reformulated using a novel semantic-only term, 2) the semantic SSD error is formulated on class-level feature maps from the decoder of a RefineNet and 3) it is shown how to best fuse the benefits of a semantic-only term with standard photometric and geometric terms for large camera motion applications.

The remainder of this paper is organised as follows. Section II reviews related work. A semantic error term is defined and a novel method is proposed for semantic-only dense RGB-D tracking in Section III. Fusing traditional joint photometric and geometric approaches with this semantic-only formulation is described in Section IV, followed by experimental results in Section V. Conclusions are drawn in Section VI.

II. Related work

A. Geometric RGB-D VO

Since low cost consumer colour and depth cameras (RGB-D) have been developed, their ability to provide reasonably accurate dense depth measurements makes them a good substitute to traditional stereo cameras for indoor applications.

Visual odometry (VO) has found much of its initial work grounded in geometric feature-based extraction and representations of the world [11], since they are sparse techniques which lend to computational efficiency.

With the advent of greater computing power, direct dense VO approaches have been proposed which purport to be more accurate and robust than sparse feature-based ones. These methods exploit all the pixels (dense) in the raw images (direct) and can be parallelisable for real-time operation. A first dense direct approach was proposed in [12] to use the dense stereo depth map to generate warped images and minimise the direct photometric error between intensity images for visual odometry and key-frame mapping. With the subsequent advent of commercial projective light RGB-D sensors, this approach was re-employed as seen in [13] and [14]. Similarly, Newcombe et al. [15] implemented KinectFusion, a depth-only frame-to-model tracking algorithm based on Iterative Closest Point (ICP). Tykkälä et al. [16] built upon these works and proposed a direct ICP bi-objective cost function which jointly minimises the photometric error and the geometric error balanced by a scaling factor λ.

B. Data-driven VO

In recent years, there has been a growing body of literature regarding data-driven VO, or learning-based VO, that formulates visual odometry as an end-to-end learning problem to regress relative camera poses. These approaches have made significant progress thanks to the ability of DCNN to cope with challenging environments. Kendall et al. [7] presented extensive work on PoseNet, a CNN-based camera re-localisation

approach that regresses the absolute 6-DoF pose of a camera in outdoor environments.

A series of works have also tackled end-to-end VO. Melekhov et al. [17] estimated the relative camera motion between unconstrained images pairs with a siamese CNN. Wang et al. [18] extend previous works encapsulating a CNN with stacked image pairs as input in a Recurrent Neural Network (RNN) that implicitly models sequential dynamics and relations. Ummenhofer et al. [6] proposed a stacked encoder-decoder network called DeMoN and exploit the natural regularisation of multi task learning for estimating simultaneously the camera's ego-motion, depth image, surface normals and optical flow. All methods so far require the ground-truth camera poses for conducting the supervised training. This suggests the need of an external motion tracking system or labelling images by SfM, which is expensive and labour-intensive.

By formulating a loss function to maximise photometric consistency between consecutive frames, the works of [19]–[21] implemented a view synthesis self-supervised training strategy for ego-motion and depth estimation. In SfM-Net [19], the authors proposed a modular framework that can be trained with various degrees of supervision using optional ground-truth camera motion or depth map. Unlike [19], [20] whose approaches are devised mainly for depth estimation and the authors give little attention to the performances on VO tasks, UnDeepVO [21] resolves difficult scale estimation problem by using stereo image pairs at training time. The loss function enforces both spatial geometric consistency between left-right pairs and also temporal geometric consistency between two consecutive monocular images. Their results seem quite promising even if the most accurate visual odometry approach on the KITTI odometry benchmark leader board[1] remains a direct stereo VO method.

Nevertheless, except for [6], all these methods produce a one shot camera motion estimate based on knowledge gained from a training set that cannot generalise to cover all possible variations present in any VO problem. Instead of replacing the motion estimator with a deep network, Peretroukhin and Kelly [22] use a Deep Pose Correction network (DPC-Net) to learn difficult-to-model corrections to a sparse stereo VO.

C. Geometric RGB-D VO with learned features

While data-driven methods struggle to attain the same performance as geometric VO, the authors believe there is still room for improving dense geometric VO with the representational power of deep networks.

The work most closely related to the present paper is by Czarnowski et al. [23]. This work also aims at developing a dense visual tracking approach based on a CNN image representation. Their system, however, differs in that it only tracks pure rotational motion and proposes to use multi-scale CNN features within a coarse-to-fine feature-based approach. Their rotation tracking results demonstrate robustness to varying lighting conditions and their pyramid of features (*semantic texture*) reached real-time performance using a GPU implementation to compute the image alignment on 15% of the 4227 feature maps in VGG16.

Compared to this work, the approach proposed here is significantly different since the aim is to perform full 6-DoF camera pose tracking (VO) using only high-level semantic segmentation classes rather than intermediary encoder features, all within a dense VO approach that takes advantage of recent advances in the field of dense CNN segmentation. Using class-level dense segmentation allows to decrease the computational cost, while still taking advantage a compact summary of all the intermediary encoder information. For example, using class-level semantics decreases the computational cost down to a single feature map alignment per pyramid level (typ. $\mathrm{N_P} = 3$) for the proposed approach while the *semantic texture* requires $2 \cdot 64$ to $3 \cdot 512$ for each pyramid level in a coarse to fine scheme.

III. Semantic-only RGB-D tracking

This section will first describe the proposed semantic-only RGB-D tracking based on class-level feature maps extracted from the decoder of a CNN trained for semantic segmentation.

A. Direct motion estimation framework

Consider a calibrated RGB-D sensor with a colour brightness function $\mathbf{I}:\Omega\rightarrow\mathbb{R}^{+};(\mathbf{p})\mapsto\mathbf{I}(\mathbf{p})$, a depth function $\mathbf{D}:\Omega\rightarrow\mathbb{R}^{+};(\mathbf{p})\mapsto\mathbf{D}(\mathbf{p})$ where $\Omega=[1,n]\times[1,m]\subset\mathbb{R}^2$. $(\mathbf{p}_1,\mathbf{p}_2,\ldots,\mathbf{p}_{nm})^\top\in\mathbb{R}^{nm\times 2}\subset\Omega$ are pixel locations within the image of dimension $n\times m$.
Similar to normal images, a feature map $\mathbf{F}_c$ of the ordered set of class-level feature maps $\mathbf{F}$ is defined as a function: $\mathbf{F}_c:\Omega^s\rightarrow\mathbb{R};(\mathbf{p})\mapsto\mathbf{F}_c(\mathbf{p})$ where $\Omega^s=[1,n/2^s]\times[1,m/2^s]\subset\mathbb{R}^2$ at scale $s\in\mathbb{N}^+$. The label prediction is also defined as a function $\widehat{\boldsymbol{\ell}}:\Omega^s\rightarrow[1;C];(\mathbf{p})\mapsto\widehat{\boldsymbol{\ell}}(\mathbf{p})=argmax(\{\mathbf{F}_c(\mathbf{p})\}_{c=1..C})$.

The pose of the camera is represented as the homogeneous pose matrix $\mathbf{T}\,(\mathbf{x})\in\mathbb{R}^{4\times4}$ which depends on a minimal parameterisation of 6 parameters defined here as the linear and angular velocity $\mathbf{x}=[v,\omega]^\top\in\mathbb{R}^6$, respectively. The homogeneous transformation matrix can be decomposed into rotational and translational components $\mathbf{T}(\mathbf{x})=(\mathbf{R}(\mathbf{x}),\mathbf{t}(\mathbf{x}))\in\mathbb{SE}(3)$. The relationship between both is given by the exponential map as $\mathbf{T}(\mathbf{x})=e^{[\mathbf{x}]_\wedge}$, with the operator $[\cdot]_\wedge$ defined as:

$$[\mathbf{x}]_\wedge=\begin{bmatrix}[\omega]_\times & v\\ 0 & 0\end{bmatrix} \tag{1}$$

where $[\cdot]_\times$ is the skew symmetric matrix operator.

Direct motion estimation is formulated as a minimization problem of solely a semantic cost function. Considering that all semantic errors $\mathbf{e}_{\boldsymbol{\mathcal{S}}i}$ are assumed independent and identically distributed (i.i.d.) and that the semantic errors are modelled as a Gaussian distribution, the estimation of the camera motion $\widehat{\mathbf{T}}(\mathbf{x})$ between a reference set of feature maps $\mathbf{F}^*$ and a current set $\mathbf{F}$ is obtained via a standard non-linear least-squares problem:

$$\underset{\mathbf{x}}{\operatorname{argmin}}\left\|\underbrace{\mathbf{F}^*-\mathbf{F}\Big(w\big(\widehat{\mathbf{T}}(\mathbf{x}),\mathbf{p}^*\big)\Big)}_{\mathbf{e}_{\boldsymbol{\mathcal{S}}}(\mathbf{x})}\right\|^2 \tag{2}$$

[1] http://www.cvlibs.net/datasets/kitti/eval_odometry.php

The superscript $*$ denotes reference measurements. The 3D points $\mathbf{P}_i=\begin{bmatrix}X_i & Y_i & Z_i\end{bmatrix}^\top \in \mathbb{R}^3$ are computed by back-projection

$$\mathbf{P}_i = \pi^{-1}(\mathbf{p}_i) = \mathbf{K}^{-1}\overline{\mathbf{p}_i}\ \mathbf{D}(\mathbf{p}_i) \tag{3}$$

where $\mathbf{K}\in\mathbb{R}^{3\times3}$ is the intrinsic camera matrix and $\overline{\mathbf{p}_i}\in\mathbb{R}^3$ are the homogeneous pixel coordinates obtained by projection as

$$\overline{\mathbf{p}_i} = \pi(\mathbf{P}_i) = 1/Z_i.\mathbf{K}\mathbf{P}_i \tag{4}$$

The inverse warping function $w(\cdot)$ projects the reference 3D points $\mathbf{P}_i^*$ transformed by $\mathbf{T}(\mathbf{x})$ onto the current frame at the warped pixel coordinates $\mathbf{p}_i^w = w(\mathbf{T}(\mathbf{x}), \mathbf{p}_i^*)$ (equ. 5).

$$\overline{\mathbf{p}_i^w} = \pi\big(\mathbf{R}(\mathbf{x})\pi^{-1}(\mathbf{p}_i^*) + \mathbf{t}(\mathbf{x})\big) \tag{5}$$

The closest feature map's value is found by bilinear interpolation of the current feature maps $\mathbf{F}$ at $\mathbf{p}_i^w$.

B. Semantic-only image alignment

The pose estimate $\widehat{\mathbf{T}}$ is computed at each iteration and is updated incrementally by a pose increment $\mathbf{T}(\mathbf{x})$ following an *inverse compositional* [24] update rule $\widehat{\mathbf{T}} \leftarrow \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})$ and the semantic error in (equ. 2) becomes:

$$\mathbf{e}_{\boldsymbol{\mathcal{S}}i,c}(\mathbf{x}) = \mathbf{F}^*_{i,c}\Big(w\big(\mathbf{T}(\mathbf{x}), \mathbf{p}_i^*\big)\Big) - \mathbf{F}_{i,c}\Big(w\big(\widehat{\mathbf{T}}, \mathbf{p}_i^*\big)\Big) \tag{6}$$

The cost function (equ. 6) is linearised and minimized around $\mathbf{x}=0$ using a first order Taylor expansion. This leads to a closed form solution:

$$\mathbf{x} = -\mathbf{H}^{-1}\mathbf{b}\ ;\ \mathbf{H} = \mathbf{J}_{\boldsymbol{\mathcal{S}}}(0)^\top \mathbf{J}_{\boldsymbol{\mathcal{S}}}(0)\ ;\ \mathbf{b} = \mathbf{J}_{\boldsymbol{\mathcal{S}}}(0)^\top \mathbf{e}_{\boldsymbol{\mathcal{S}}}(0) \tag{7}$$

where $\mathbf{J}_{\boldsymbol{\mathcal{S}}}$ represents the stacked Jacobian matrix obtained by derivation of the stacked semantic error $\mathbf{e}_{\boldsymbol{\mathcal{S}}}$ for all pixels nm through all C classes (or feature maps):

$$\begin{aligned} \mathbf{e}_{\boldsymbol{\mathcal{S}}i,c}(0) &= \mathbf{F}^*_{i,c} - \mathbf{F}_{i,c}\big(w(\widehat{\mathbf{T}}, \mathbf{p}_i^*)\big) \\ \mathbf{J}_{\boldsymbol{\mathcal{S}}i,c}(0) &= \nabla\mathbf{F}^*_{i,c}\ \frac{\partial w(\mathbf{T}(\mathbf{x}), \mathbf{p}_i^*)}{\partial \mathbf{x}}\Big|_{\mathbf{x}=0} \end{aligned} \tag{8}$$

The semantic Jacobian $\mathbf{J}_{\boldsymbol{\mathcal{S}}}(0)$ is calculated once for all iterations.

Two error function variants have been considered:

1) Minimisation across all class scores:

$$\begin{aligned} \mathbf{H} &= \sum_i^{nm}\sum_c^{C} \mathbf{J}_{\boldsymbol{\mathcal{S}}i,c}(0)^\top \mathbf{J}_{\boldsymbol{\mathcal{S}}i,c}(0) \\ \mathbf{b} &= \sum_i^{nm}\sum_c^{C} \mathbf{J}_{\boldsymbol{\mathcal{S}}i,c}(0)^\top \mathbf{e}_{\boldsymbol{\mathcal{S}}i,c}(0) \end{aligned} \tag{9}$$

In the first variant, it is assumed that the classification-score of a pixel to each class must contribute to the minimisation (i.e. each pixel has C scores). This means that not only will the score of the best class be used for pose estimation, but also, the score of the other classes that have been considered. In doing so this allows to account indirectly for the uncertainty when, for example, all classes have a similar score or alternatively when only one class has a high score. In order to define a valid error criterion on this basis, it is necessary to define an error function that is locally convex and which has zero error at the minimum. It is assumed here that the semantic *feature scores* of each pixel are invariant across different poses. This is similar to the Lambertian hypothesis formulated by photometric-based tracking, however, it obviously depends on the invariance of the classifier. This hypothesis has been verified experimentally under the small motion assumption and future work would be dedicated to developing pose-invariant scoring. Computationally, however, this dense approach leads to $\mathrm{C}\cdot\mathrm{N_P}$ alignments of class-level feature maps.

2) Minimisation using the best class:

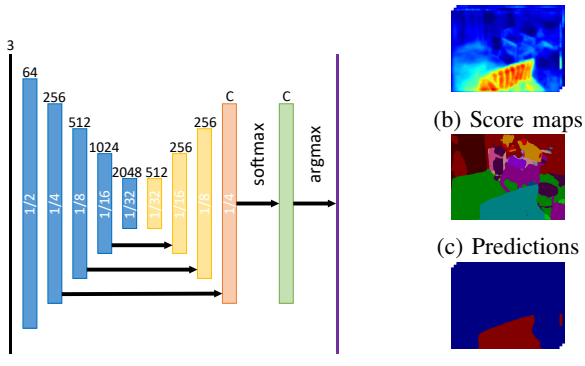
$$\begin{aligned} \mathbf{H} &= \sum_i^{nm} \mathbf{J}_{\boldsymbol{\mathcal{S}}i,\widehat{\ell}_i^*}(0)^\top \mathbf{J}_{\boldsymbol{\mathcal{S}}i,\widehat{\ell}_i^*}(0) \\ \mathbf{b} &= \sum_i^{nm} \mathbf{J}_{\boldsymbol{\mathcal{S}}i,\widehat{\ell}_i^*}(0)^\top \mathbf{e}_{\boldsymbol{\mathcal{S}}i,\widehat{\ell}_i^*}(0) \end{aligned} \tag{10}$$

where $\widehat{\ell}_i^*$ is the label prediction for the pixel i of the reference image. In the second variant, for a given pixel, only one score, from which the label is predicted, contributes to the minimisation. The invariance of the class label across various viewpoints is maintained as long as the classifier succeeds. An advantage of this approach is that it allows to pre-calculate the semantic error $\mathbf{e}_{\boldsymbol{\mathcal{S}}}$ and the semantic Jacobian $\mathbf{J}_{\boldsymbol{\mathcal{S}}}$ for class c only at pixels i such that $c = \widehat{\ell}_i^*$. This approach leads to the alignment of a single class-level feature map per pyramid level ($\mathrm{N_P}$) and is much more efficient that the first variant.

C. Feature maps selection

The semantic segmentation component has shown to provide relatively consistent label predictions $\widehat{\ell}$ across challenging conditions including intra- and inter-class variations, difficult lighting conditions and even viewpoint variations. Subsequently, the non-trivial assumption that class-level feature maps $\mathbf{F}_c$ are viewpoint-invariant is clear, yet dependent on the performance and subsequent validation of the classifier. Considering that the activation functions of RefineNet are $ReLU$, one can expect the score maps $\mathbf{F}_c$ and, to a lesser extent, the semantic errors $\mathbf{e}_{\boldsymbol{\mathcal{S}}}$, to have bigger values for some classes. This would naturally result in favouring the most confident classes over others.

In practice, semantic image alignment is performed with either with the scores (before *softmax*) of class-level feature maps (Fig. 1b) or with binary masks (Fig. 1d) which are a one-hot conversion of label predictions (Fig. 1c).



(a) Simplified RefineNet architecture

(b) Score maps

(c) Predictions

(d) Binary masks

Fig. 1: Semantic segmentation network architecture and an insight into the C class-level feature maps (orange) from the decoder (yellow), the label predictions (purple) and the corresponding binary prediction masks. Best viewed with colours.

Class-level feature maps (aka score maps), are extracted from the decoder of a RGB-RefineNet-101 [5]. RefineNet is considered as the current state-of-the-art neural network for semantic segmentation of RGB images. The encoder uses ResNets (here a ResNet-101) and the decoder is an improved version of basic multi-scale skip connections [4] called *multi-path refinement module* that heavily relies on residual connections with identity mappings [1]. We do not use expensive post-processing steps except a multi-scale evaluation (MSc eval). However, label smoothing by CRF [25] would certainly be beneficial. The predictions are refined to a final resolution of $1/4$, although they can be obtained at finer resolution using bilinear upsampling. Finally, the score maps are cropped with a border of 8 pixels in order to remove border effects in the CNN predictions.

RGB-RefineNet-101 was trained on **NYUDv2** dataset [26] for a C-class semantic segmentation task. The NYUDv2 dataset consists of 1449 RGB-D images of size 640x480 showing indoor scenes. The standard training/test split is used with 795/654 images respectively. It was assumed that the number of classes C and their definitions could be fine-tuned as an extra hyperparameter of the proposed method.

D. Semantic coarse to fine pyramid

The iterative semantic image alignment method is embedded in a coarse to fine multi-resolution pyramid scheme. This approach is commonly used by image alignment techniques to speed up the convergence and increase the size of the convergence domain. In the present context, the pyramid is computed on the semantic segmentation image instead of the input RGB image. Before alignment, a Gaussian pyramid of down-sampled versions of the semantic image is computed. Then, the optimisation proceeds by performing a number of iterations at each level, starting at the coarsest level of the pyramid to obtain an initial pose estimate before refining this estimate by propagating it down the pyramid levels until reaching the original image resolution.

IV. Hybrid Semantic RGB-D Tracking

A. Semantic tri-objective direct ICP

Inspired by [16], a tri-objective cost function is proposed to simultaneously minimise a semantic-only error $\mathbf{e}_{\boldsymbol{\mathcal{S}}}$ (equ. 6), a photometric error $\mathbf{e}_{\boldsymbol{\mathcal{I}}}$ (equ. 12) and a geometric error $\mathbf{e}_{\boldsymbol{\mathcal{G}}}$ (equ. 13) in order to draw advantages from each:

$$\underset{\mathbf{x}}{\operatorname{argmin}} \quad \lambda_{\boldsymbol{\mathcal{S}}}^2 \left\|\mathbf{e}_{\boldsymbol{\mathcal{S}}}\right\|^2 + \lambda_{\boldsymbol{\mathcal{I}}}^2 \left\|\mathbf{e}_{\boldsymbol{\mathcal{I}}}\right\|^2 + \left\|\mathbf{e}_{\boldsymbol{\mathcal{G}}}\right\|^2 \tag{11}$$

The photometric error $\mathbf{e}_{\boldsymbol{\mathcal{I}}}$ is defined as in [13]:

$$\mathbf{e}_{\boldsymbol{\mathcal{I}}_i}(\mathbf{x}) = \mathbf{I}_i^* \Big(w\big(\mathbf{T}(\mathbf{x}), \mathbf{p}_i^*\big)\Big) - \mathbf{I}_i \Big(w\big(\widehat{\mathbf{T}}, \mathbf{p}_i^*\big)\Big) \tag{12}$$

and the geometric error $\mathbf{e}_{\boldsymbol{\mathcal{G}}}$ as a point-to-plane ICP error with projective data association [27]:

$$\mathbf{e}_{\boldsymbol{\mathcal{G}}_i}(\mathbf{x}) = \Big(\widehat{\mathbf{R}}\mathbf{R}(\mathbf{x})\mathbf{N}_i^*\Big)^\top \Big(\mathbf{P}_i^m - \Pi_3 \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})\overline{\mathbf{P}_i^*}\Big) \tag{13}$$

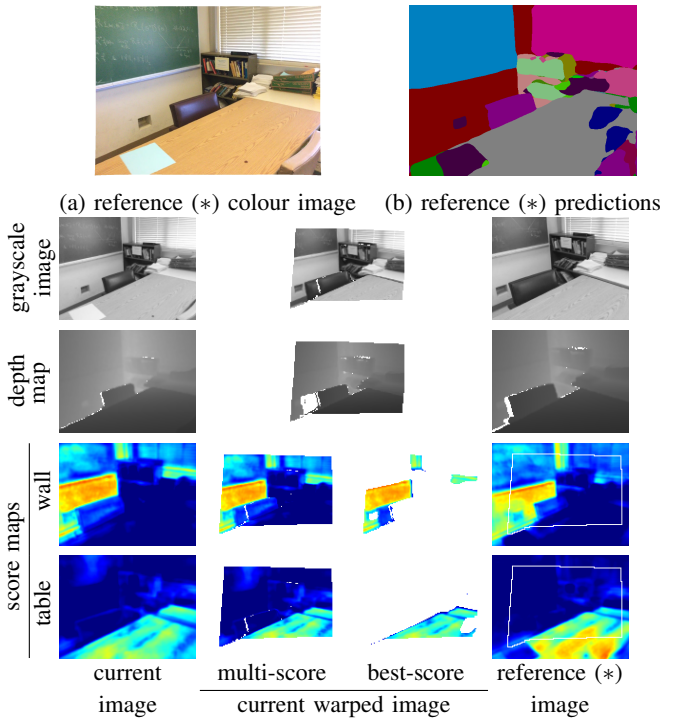
where $\Pi_3 = [\mathbf{1}, \mathbf{0}] \in \mathbb{R}^{3\times 4}$ is the projection matrix, $\mathbf{N}_i^* \in \mathbb{R}^3$ are the reference surface normals computed for each homogeneous 3D point $\overline{\mathbf{P}_i^*} \in \mathbb{R}^4$ using a local cross product on the image grid. The closest image intensity is found by bilinear interpolation of the current intensity function $\mathbf{I}$ at $\mathbf{p}_i^w$. The closest 3D point $\mathbf{P}_i^m$ is obtained by linearly interpolating the current depth map $\mathbf{D}$ at $\mathbf{p}_i^w$ and back-projecting it.

The non-linear least-squares tri-objective error $\mathbf{e} = [\lambda_{\boldsymbol{\mathcal{S}}}\mathbf{e}_{\boldsymbol{\mathcal{S}}}\ \lambda_{\boldsymbol{\mathcal{I}}}\mathbf{e}_{\boldsymbol{\mathcal{I}}}\ \mathbf{e}_{\boldsymbol{\mathcal{G}}}]^\top$ is iteratively minimised using a Gauss-Newton approach (equ. 7) where $\mathbf{J} = [\lambda_{\boldsymbol{\mathcal{S}}}\mathbf{J}_{\boldsymbol{\mathcal{S}}}\ \lambda_{\boldsymbol{\mathcal{I}}}\mathbf{J}_{\boldsymbol{\mathcal{I}}}\ \mathbf{J}_{\boldsymbol{\mathcal{G}}}]^\top$. $\mathbf{J}_{\boldsymbol{\mathcal{I}}}$ is the Jacobian matrix of the photometric error $\mathbf{e}_{\boldsymbol{\mathcal{I}}}$ calculated once for all iterations like $\mathbf{J}_{\boldsymbol{\mathcal{S}}}$ (equ. 8) and $\mathbf{J}_{\boldsymbol{\mathcal{G}}}$ is the standard point-to-plane ICP Jacobian matrix, calculated at each iteration.

B. Selection of scaling factors

Most hybrid approaches require scaling factors (here $\lambda_{\boldsymbol{\mathcal{S}}}$ and $\lambda_{\boldsymbol{\mathcal{I}}}$), either heuristic or automatic, which weight the contribution of the different errors. Two strategies have been selected here:

1) a constant λ fine-tuned empirically.
2) an adaptive λ inspired by [28] that varies using a sigmoid function which favours the semantic error $\mathbf{e}_{\boldsymbol{\mathcal{S}}}$ far from the solution and the photometric error $\mathbf{e}_{\boldsymbol{\mathcal{I}}}$ close to the minimum. Our experiments have shown that it is best to maintain the geometric error $\mathbf{e}_{\boldsymbol{\mathcal{G}}}$ all along.



(a) reference (∗) colour image (b) reference (∗) predictions

(c) Photometric term (row 1), geometric term (row 2) and semantic term (rows 3-4) with multi-score and best-score variants (col 2-3).

Fig. 2: Semantic image alignment of frames n and $n{+}k$ from ScanNet's *scene0030_00* where $n{=}1813$ and $k{=}20$. The current image $\{\mathbf{I}; \mathbf{D}; \mathbf{F}\}$ is iteratively warped minimising the tri-objective error $\mathbf{e}$ with the reference image $\{\mathbf{I}^*; \mathbf{D}^*; \mathbf{F}^*\}$. Here, the current warped image is obtained after 3 iterations.

V. Experiments

To show the effectiveness of the proposed approach, experiments have been carried out on the public ScanNet dataset [29] in order to firstly assess the quality of the semantic segmentation (Sec. V-A) and secondly fine-tune and evaluate the semantic image alignment for different parameter settings (Sec. V-B). This dataset contains various indoor environments captured with hand-held RGB-D sensors. For convenience, the full size resolution is set at 640×480 by down-sampling ScanNet's 1296×968 grayscale images using bicubic interpolation.

ScanNet. The ScanNet dataset [29] provides registered RGB-D images, ground-truth labels and computed camera poses estimated by a Structure from Motion (SfM) algorithm. The dataset provides 1513 sequences and 2.5M images with labels obtained by manually labeling 3D reconstructed meshes. The sequences acquired in this dataset benefit from a rich semantic environment, similar to the NYUDv2 scenes.

A. Semantic segmentation evaluation

The performance of the semantic segmentation network RefineNet-101, presented in Section III-C, is evaluated on four ScanNet sequences: *scene0002_00*, *scene0026_00*, *scene0030_00*, *scene0568_00*.

Metrics. Standard metrics [4] are defined as

- class accuracy: $\text{acc}_\text{i} = n_{ii}/t_i$
- mean class accuracy: $\text{macc} = (1/n_{cl}) \sum_i \text{acc}_\text{i}$
- Jaccard index: $\text{IoU}_\text{i} = n_{ii}/(t_i + \sum_j n_{ji} - n_{ii})$
- mean Jaccard index: $\text{mIoU} = (1/n_{cl}) \sum_i \text{IoU}_\text{i}$

where n_{cl} is the number of classes, n_{ij} is the number of pixels of class i classified as class j, and $t_i = \sum_j n_{ij}$ is the total number of pixels belonging to class i.

Transfer learning. We trained RGB-RefineNet-101 for the 40-class [2], the 13-class [3] and the 4-class [26] semantic segmentation tasks on the NYUDv2 dataset. Since NYUDv2 and ScanNet datasets have both been acquired with similar settings and environments, it is assumed that the training domain is close to the evaluation domain and consequently it was not fine-tuned on the ScanNet dataset. At test time, RefineNet was fed with 12k images from the four ScanNet sequences. The final predictions are inferred at a scale of $1/4$ producing feature maps at a resolution 160×120.
The results are shown in Table I. RefineNet-101 achieves quite good performance on most highly represented classes including *wall, floor, chair, sofa, table, blinds, sink*. These results confirm the benefit of semantic contribution in hybrid image alignment.

B. Semantic image alignment experiments

In all of the following experiments, image alignment is performed on image pairs n and $n+k$ for $k \in \{1, 2, 3, 6, 10, 15, 30\}$ skipping $k-1$ intermediate frames within the $\text{N}=5192$ frames of ScanNet's *scene0002_00* in order to cover a wide range of relative pose errors (**RPE**s) and simulate larger camera velocities. The proposed method is illustrated in Figure 2.

Several different image alignment methods are compared since the proposed error terms have different qualities and can be combined in various ways including: `Phot` for photometric term (equ. 12), `Geom` for geometric term (equ. 13), `MsSem` for the multi-score semantic term (equ. 9) and `Sem` for the best-score semantic term (equ. 10).

All tracking approaches are embedded into a coarse to fine scheme using the following $\text{N}_\text{P}=3$ pyramid levels of $\{\mathbf{I}; \mathbf{D}; \mathbf{F}\}$ images starting at scale $1/4$:

pyramid level	scale	resolution
0	1/4	160×120
1	1/8	80×60
2	1/16	40×30

Minimisation is iterated for a maximum of $[\lceil \text{K}/3 \rceil, \lceil \text{K}/2 \rceil, \text{K}]$ iterations in levels $[2, 1, 0]$ respectively, starting with the coarsest level, where K is the number of iterations at finer scale.

Metrics. Following [30], the relative pose error is defined between frames n and $n+k$ which measures the local accuracy of the motion estimation as

$$\mathbf{RPE}_n^k = \left(\mathbf{P}_n^{gt\,-1} \mathbf{P}_{n+k}^{gt}\right)^{-1} \left(\mathbf{P}_n^{est\,-1} \mathbf{P}_{n+k}^{est}\right) \tag{14}$$

From the $\text{M}=\text{N}-k$ individual **RPE**s along the sequence, the translational normalized root mean square error (nRMSE) is computed over all frame indices of the translational component

$$\text{nRMSE}\left(\mathbf{RPE}_{1:\text{N}}^k\right) = \left(\frac{1}{\text{M}} \sum_{n=1}^{\text{M}} \frac{\left\| trans(\mathbf{RPE}_n^k) \right\|^2}{\left\| trans(\mathbf{P}_n^{gt\,-1} \mathbf{P}_{n+k}^{gt}) \right\|^2} \right)^{1/2} \tag{15}$$

In case of divergence in the minimisation, the relative pose estimate $\mathbf{P}_n^{est\,-1} \mathbf{P}_{n+k}^{est}$ is set to the identity matrix.

Tuning the number of classes. Since RefineNet was trained for three different segmentation tasks, Fig. 3 plots the nRMSE of `Sem` method for score maps in input obtained with $\text{C}=\{4, 13, 40\}$ and for binary masks in input with $\text{C}=40$.

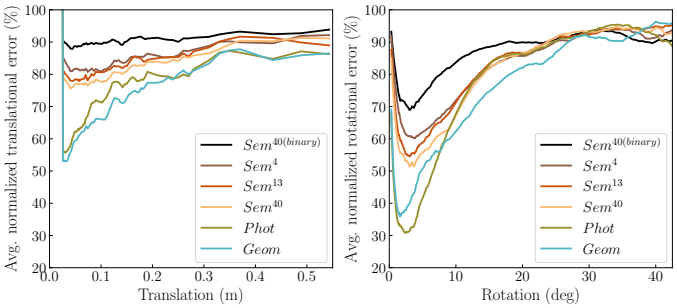


Fig. 3: Evaluation of `Sem` method with score maps obtained for $\text{C}=\{4, 13, 40\}$ semantic segmentation tasks. The rotational error (right) is nearly an image of the translational error (left). In practice, no loss in performance was noticed using the `Sem` method compared to the `MsSem` method.

`Sem`$^{40\ \text{(binary)}}$ performs quite bad compared to other `Sem` methods. This demonstrates that score maps contain more useful semantic information for VO than label predictions. Moreover, even if there is a gain in segmentation quality

	40-class task																												
Class	wall	floor	cabinet	chair	sofa	table	door	window	book shelf	picture	counter	blinds	desks	shelves	pillow	books	refrigerator	television	paper	shower curtain	box	toilet	sink	lamp	bag	other structure	other furniture	other prop	**mean**
Label	1	2	3	5	6	7	8	9	10	11	12	13	14	15	18	23	24	25	26	28	29	33	34	35	37	38	39	40	
acc_i (%)	74.4	66.7	8.2	55.0	49.1	68.3	19.6	9.6	43.5	58.5	39.7	72.4	0.2	2.2	39.4	64.4	5.5	46.4	40.3	12.3	25.6	69.3	76.0	2.3	17.8	27.1	3.1	59.0	37.7
IoU_i (%)	39.8	58.6	6.2	41.1	35.4	56.2	16.7	8.7	37.9	3.6	26.8	41.6	0.1	0.5	29.5	14.6	5.1	35.5	5.4	7.3	21.7	47.4	54.5	1.2	8.1	24.2	2.7	19.2	16.2

	13-class task												4-class task				
Class	books	chair	floor	furniture	objects	picture	sofa	table	tv	wall	window	**mean**	ground	structure	furniture	prop	**mean**
Label	2	4	5	6	7	8	9	10	11	12	13		1	2	3	4	
acc_i (%)	65.7	44.4	69.3	15.9	42.4	57.2	40.4	70.4	48.8	73.8	50.6	52.6	55.3	85.2	50.8	46.0	59.3
IoU_i (%)	15.9	32.7	57.9	14.6	26.5	2.3	27.4	47.3	28.4	51.9	20.7	25.0	52.1	54.4	42.8	25.3	43.6

TABLE I: Evaluation of RGB-RefineNet-101 averaged on four ScanNet sequences for 40-class [2], 13-class [3] and 4-class [26] segmentation tasks. A mIoU of 44.9%, 56.1% and 73.5% is reached on the NYUv2 test set, for these tasks respectively. For clarity, entries of classes that do not appear in considered sequences has been deleted.

when the number of classes C is smaller, the semantic tasks with C=$\{4, 13\}$ do not bring enough semantic information to improve semantic-only tracking, compared to 40-class task. Finally, Fig. 3 also shows that semantic-only tracking does not perform as good as Phot or Geom methods.

Tuning the scaling factors. In order to prove that semantic error can be useful in improving the standard Phot+Geom method, the scaling factor $\lambda_{\mathcal{I}}$ is set to 0.35 (best in practice) and $\lambda_{\mathcal{S}}$ is fine-tuned. Fig. 4 shows a plot of the translational nRMSE (col 1) of $\text{Sem}^{40}_{\lambda_\text{S}}\text{Phot}_{0.35}\text{Geom}$ method with varying $\lambda_{\mathcal{S}}$ settings. Col 2 shows the plot with the scaling factor $\lambda_{\mathcal{S}}$ which achieves the best translational error over different ground-truth translations. A sigmoid model was successfully fit to those points and thus verifies the hypothesis formulated in Section IV-B demonstrating that an adaptive λ strategy that varies using a sigmoid function can be employed.

Optimal adaptive scaling factors. The optimal adaptive lambda strategy is defined to be the one which always selects the best λ setting. Fig. 5 shows the plot of the nRMSE of $\text{Phot}_\text{opt}\text{Geom}$, $\text{Sem}^{40}_\text{opt}\text{Geom}$ and $\text{Sem}^{40}_\text{opt}\text{Phot}_{0.35}\text{Geom}$ approaches with this strategy. First, these results show that, in combination with Geom term, the Sem term performs better than Phot term far from the solution ($t{>}10cm$ or $\theta{>}8°$). Since a median ground-truth translation of $25cm$ and a standard deviation of $14cm$ at $k{=}30$ is observed, it can be concluded that the semantic term is more useful for applications such as frame-to-keyframe alignment or relocalisation. Second, it was found that the $\text{Sem}^{40}_\text{opt}\text{Phot}_{0.35}\text{Geom}$ method with adaptive $\lambda_{\mathcal{S}}$ outperforms the $\text{Phot}_\text{opt}\text{Geom}$ method, taking advantage of the semantic term far from the solution ($\{\lambda_{\mathcal{S}}{\approx}1/70; \lambda_{\mathcal{I}}{=}0.35\}$) and relying on the photometric term close to the solution ($\{\lambda_{\mathcal{S}}{\approx}0; \lambda_{\mathcal{I}}{=}0.35\}$).

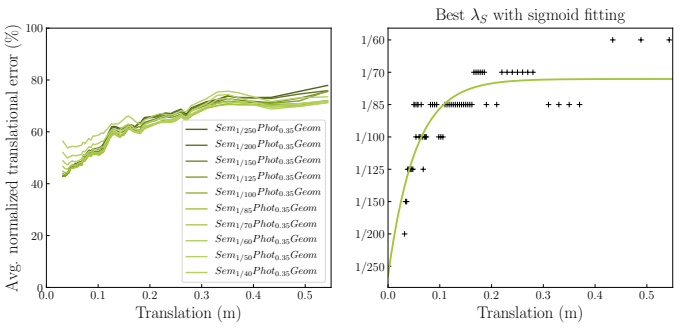


Fig. 4: Fine-tuning $\text{Sem}^{40}_{\lambda_\text{S}}\text{Phot}_{0.35}\text{Geom}$ approach with varying $\lambda_{\mathcal{S}}$ settings and a constant $\lambda_{\mathcal{I}}{=}0.35$.

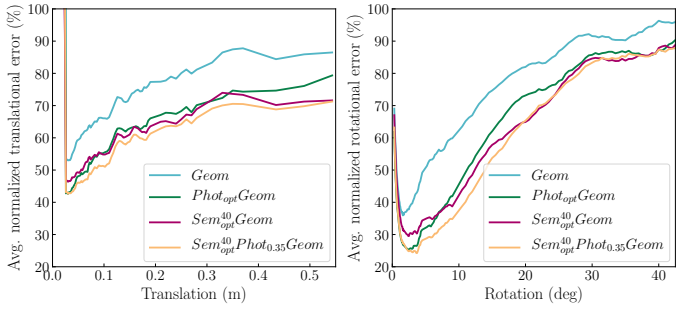


Fig. 5: Evaluation of multi-objective methods with optimal adaptive scaling factors.

VI. Conclusion

In this paper, a semantic-only visual odometry (SemVO) approach was proposed and implemented for full 6-DoF pose estimation using class-level feature maps from a deep neural network trained for semantic segmentation. It was demonstrated that dense semantic-only visual odometry is possible by assuming classification invariance across viewpoints. Various implementations were considered including a multi-score error function and a best-score error function. A tri-objective non-linear least-squares error function was also proposed to take advantage of the new semantic error term simultaneously with classic photometric and geometric terms. The impact of the number of classes on tracking performance was analysed along with optimal fusion parameters. Experiments on the ScanNet dataset confirm that the proposed semantic error term helps the minimisation to converge far from the solution. This condition could be verified in several applications including frame-to-keyframe alignment, tracking at low frame rate, high speed motion or even relocalisation.

In future works, the proposed semantic-based VO could be improved with a more consistent semantic segmentation exploiting recent advances by, firstly, the fusion of the RGB and depth channels for deep networks and, secondly, the combination of image and 3D point cloud semantic segmentation. Finally, it would be worth designing a complete semantic RGB-D SLAM with semantic-based tracking integrated in a consistent semantic mapping.

Acknowledgement

This work was supported by the H2020 Comanoid project (www.comanoid.eu).

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 564–571.

[3] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[5] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation," *arXiv preprint arXiv:1611.06612*, 2016.

[6] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," *arXiv preprint arXiv:1612.02401*, 2016.

[7] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[8] R. Gomez-Ojeda, J. Briales, and J. Gonzalez-Jimenez, "Pl-svo: Semi-direct monocular visual odometry by combining points and line segments," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 4211–4216.

[9] L. Ma, C. Kerl, J. Stückler, and D. Cremers, "Cpa-slam: Consistent plane-model alignment for direct rgb-d slam," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1285–1291.

[10] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1352–1359.

[11] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry," in *IEEE Conference on Computer Vision and Pattern Recognition,*, vol. 1, June 2004, pp. I–652–I–659 Vol.1.

[12] C. A.I., E. Malis, and P. Rives, "Accurate quadri-focal tracking for robust 3D visual odometry," in *IEEE International Conference on Robotics and Automation*, Rome, Italy, April 2007.

[13] C. Audras, A. I. Comport, M. Meilland, and P. Rives, "Real-time dense rgb-d localisation and mapping," in *Australian Conference on Robotics and Automation*, 2011.

[14] F. Steinbrücker, J. Sturm, and D. Cremers, "Real-time visual odometry from dense rgb-d images," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 719–722.

[15] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*. IEEE, 2011, pp. 127–136.

[16] T. Tykkälä, C. Audras, and A. I. Comport, "Direct iterative closest point for real-time visual odometry," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2050–2056.

[17] I. Melekhov, J. Kannala, and E. Rahtu, "Relative camera pose estimation using convolutional neural networks," *arXiv preprint arXiv:1702.01381*, 2017.

[18] S. Wang, R. Clark, H. Wen, and N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2043–2050.

[19] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, "Sfm-net: Learning of structure and motion from video," *arXiv preprint arXiv:1704.07804*, 2017.

[20] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," *arXiv preprint arXiv:1704.07813*, 2017.

[21] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," *arXiv preprint arXiv:1709.06841*, 2017.

[22] V. Peretroukhin and J. Kelly, "Dpc-net: Deep pose correction for visual localization," *arXiv preprint arXiv:1709.03128*, 2017.

[23] J. Czarnowski, S. Leutenegger, and A. Davison, "Semantic texture for robust dense tracking," *arXiv preprint arXiv:1708.08844*, 2017.

[24] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International journal of computer vision*, vol. 56, no. 3, pp. 221–255, 2004.

[25] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[26] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," *Computer Vision–ECCV 2012*, pp. 746–760, 2012.

[27] G. Blais and M. D. Levine, "Registering multiview range data to create 3d computer objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 820–824, 1995.

[28] L.-P. Morency and T. Darrell, "Stereo tracking using icp and normal flow constraint," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 4. IEEE, 2002, pp. 367–372.

[29] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[30] 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.