C*: Cross-modal Simultaneous Tracking And Rendering for 6-DoF Monocular Camera Localization Beyond Modalities

Shuji Oishi[1], Yasunori Kawamata[2], Masashi Yokozuka[1], Kenji Koide[1], Atsuhiko Banno[1], and Jun Miura[2]

***Abstract*—We present a monocular camera localization technique for a three-dimensional prior map. Visual localization has been attracting considerable attention as a lightweight and widely available localization technique for any mobilities; however, it still suffers from appearance changes and a high computational cost. With a view to achieving robust and real-time visual localization, we first reduce the localization problem to alternate local tracking and occasional keyframe rendering by following a simultaneous tracking and rendering algorithm. At the same time, by using an information-theoretic metric denoted normalized information distance in the local tracking, we developed a 6-DoF localization method robust to intensity variations between modalities and varying sensor properties. We quantitatively evaluated the accuracy and robustness of our method using both synthetic and real datasets and achieved reliable and practical localization even in the case of extreme appearance changes.**

***Index Terms*—Localization, Multi-Modal Perception, Visual Tracking**

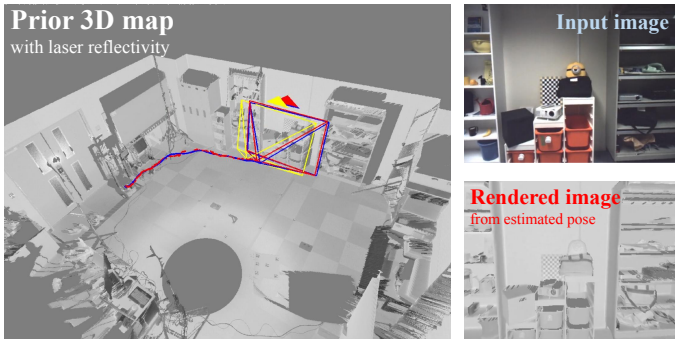


Fig. 1. 6-DoF visual localization beyond modalities. See the attached video or `https://youtu.be/PW_DlMiH-_w` for more information.

I. Introduction

ACCURATE and robust localization is a key technology for any autonomous mobilities, and it has become increasingly important to achieve autonomous navigation. Given an accurate pose, a mobile robot can plan a path toward the destination and navigate itself using feedback control, thereby resulting in an efficient and safe autonomous system.

Various approaches have been proposed toward achieving reliable localization. Although using the global positioning system would be one of the most common approaches, it can be easily disrupted in several regions, such as urban and indoor environments, because of multi-path effects and sky-view obstruction, while autonomous service robots are in high demand in such regions.

Thus far, successful autonomous mobilities have adopted three-dimensional (3D) map-based approaches that estimate the robot pose by aligning the local observation to the previously constructed 3D map. Fortunately, performing the 3D modeling of a real site or building has become easy, as 3D reconstruction technologies, such as Simultaneous Localization And Mapping (SLAM) [1] [2] and Structure from Motion [3], have evolved, and the reconstructed map can be used as a prior for localization. In the robotics community, LiDAR-based approaches that utilize point cloud registration techniques, such as iterative closest points (ICP) and normal distributions transform, have achieved highly accurate and robust localization by estimating the pose at which the local observation and prior map are geometrically consistent with each other. However, LiDAR-based approaches significantly increase the sensor cost.

As an alternative to them, visual localization has been attracting considerable attention to address the above-mentioned issues. It infers the pose of an agile monocular camera in a known 3D map and, therefore, substantially reduces the sensor cost compared with that in LiDAR-based approaches, thereby resulting in lightweight and widely available localization for any mobilities. It can be categorized as either indirect methods that use image descriptor matching [4] or direct methods that compare the appearance (pixel intensities) in each view [5]. Although these approaches achieve high localization performances, they inherently perform in the limited cases wherein sensors with the same properties are used for mapping and localization to ensure that the local observation is consistent with the appearance of the prior unless we *overcome* the difference of sensor characteristics or modalities.

Manuscript received: February 21, 2020; Revised May 15, 2020; Accepted June 23, 2020.

This paper was recommended for publication by Sven Behnke upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by JSPS KAKENHI(Grant Number 18K18072), research grant by Naito Science & Engineering Foundation, and a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

[1]Mobile Robotics Research Team, National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki, Japan (e-mail: {shuji.oishi, yokotsuka-masashi, k.koide, atsuhiko.banno}@aist.go.jp)

[2]Active Intelligent Systems Laboratory, Toyohashi University of Technology, Aichi, Japan (e-mail: kawamata@aisl.cs.tut.ac.jp, jun.miura@tut.jp)

Digital Object Identifier (DOI): see top of this page.

With a view to achieving easy and reliable pose tracking, we are interested in monocular camera localization for both different sensing modalities and appearance changes between mapping and localization phases. The problem, however, is what is refer to as "cross-modal / long-term visual localization," which has been a widely studied topic. Inspired by simultaneous tracking and rendering (STAR) [5], reducing the localization problem to successive local tracking against synthetic keyframes, we progressed toward achieving accurate and robust visual localization, even for different sensor properties. Specifically, by employing an information-theoretic metric in the local tracking to estimate the intensity relationship between the monocular camera image and appearance of the prior 3D map, we developed cross-modal STAR, C^*, as depicted in Fig.1.

II. RELATED WORK

Indirect visual localization first constructs a database of image descriptors with their 3D locations as a prior map and, subsequently, estimates the camera pose via feature-point matching [4] [6]. Specifically, it uses image descriptors or pre-trained visual words to make the 3D feature points discriminative and match feature points extracted in the current frame against the database. According to the correspondences it localizes the camera pose by minimizing 2D reprojection errors. While descriptors can be repeatably matched even when slight illumination changes occur, they become less discriminative when viewpoint changes significantly. This is because the visual feature associated with each map point is usually extracted from a local intensity pattern in a 2D image obtained from a certain viewpoint, thereby resulting in unstable localization for large viewpoint changes.

Direct visual localization achieves stable tracking by performing a pixel-wise comparison between incoming camera images and a set of colored 3D points [7] [5]. Compared with feature-based methods, direct methods achieve more robust pose estimation even for significant viewpoint changes. They perform the pose estimation by relying only on map point colors (intensities), which results in low viewpoint dependency. However, in direct methods, it is implicitly assumed that the appearance is captured using the same sensor and remains over time; therefore, the performance can easily worsen because of illumination changes or different sensor property/modality, thereby restricting their applications to only limited situations.

In both indirect and direct methods, the localization accuracy can be remarkably reduced or the localization can fail if the appearance excessively changes because of some reason. To develop visual localization that is robust to extreme appearance changes, several approaches utilize geometric information. Caselitz *et al.* [8] employed ORB-SLAM [9] to generate 3D points of the scene from image sequences and tracked the camera pose by aligning the points against the prior map with 7-DoF ICP. Although this approach uses a monocular camera, it results in point cloud registration, which is independent of appearance changes, thereby achieving robust localization. However, 3D points reconstructed using indirect visual SLAM are inherently sparse, and 7-DoF ICP with sparse point clouds may result in rough localization. Neubert *et al.* [10] developed an appearance-independent localization by defining a likelihood function for a particle filter that strongly assumed the co-occurrence of geometric and texture edges. By successively updating particles according to the likelihood function, it estimated the optimal pose from which the edges in a virtual depth image captured were highly correlated to those in the current camera image. Similar methods that minimize the distance between edges in input and reference images by assuming the co-occurrence between different modalities have been proposed [11] [12] [13]; however, the assumption should be carefully designed to ensure the co-occurrence depending on individual cases, otherwise the similarity evaluation may fail.

Deep learning-based approaches to visual localization have also come into fashion. For example, Kendall and Cipolla [14] developed an end-to-end robust pose estimation from a monocular camera image. They proposed novel loss functions based on scene geometry, which allows simultaneous position and rotation learning, and achieved highly robust localization under different lighting conditions and appearance changes. However, as discussed in [15], data-driven visual localization tends to be less accurate than classical indirect approaches due to the difficulty in training the pose regression layer well. Focusing on this, [15] developed several approaches with hand-crafted or data-driven models for computing essential matrices; however, they still lag behind classical approaches in terms of accuracy.

To perform accurate localization even in case of illumination changes or across different sensor modalities, some research works used information-theoretic metrics. Wolcott *et al.* [16] achieved localization on a road map captured using LiDAR by maximizing the mutual information (MI) between in-vehicle camera views and virtually rendered images; however, only 3-DoF pose estimation was performed. Pascoe *et al.* [17] [18] also proposed a 6-DoF pose estimation by employing normalized information distance (NID) [19] and demonstrated a promising localization performance in aligning images from different modalities; however, the tracking rate was only 2 Hz. Information-theoretic metrics can compare intensity distributions instead of intensities or intensity patterns, and do not expect any specific types of coincidence or mapping. This allows robust similarity evaluation beyond modalities; however, simultaneously, the computation is relatively demanding.

Inspired by the above-mentioned works, we developed a cross-modal monocular camera tracking algorithm toward achieving robust and practical localization. The main contribution of this paper is a new pipeline that combines STAR and NID; The combination of keyframe-based localization and robust local tracking achieves efficient and reliable localization even under extreme appearance changes, as demonstrated in Section IV.

III. PROPOSED METHOD

A. Overview

Figure 2 provides an overview of the proposed method. Given an initial camera pose in a prior 3D map, we track the

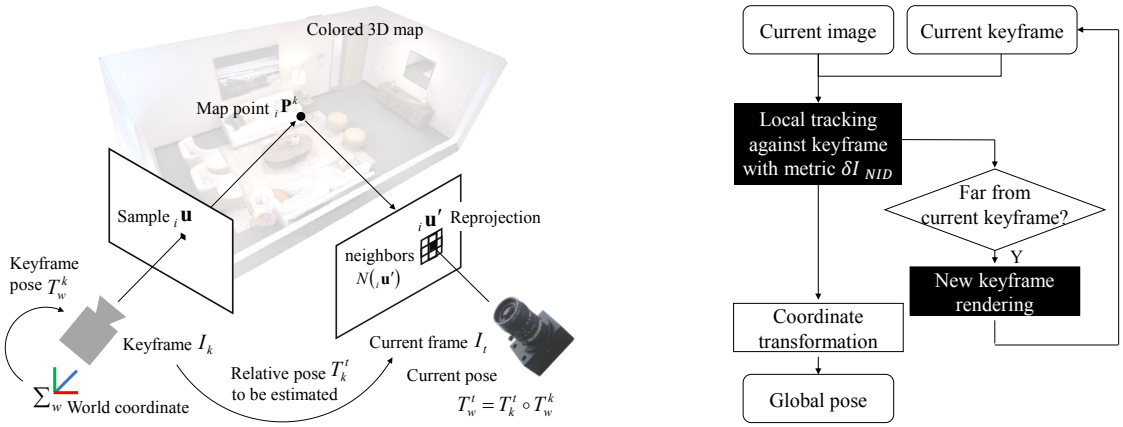


Fig. 2. Overview of the proposed method.

current camera pose T_t^W by comparing the appearances of the 3D prior map with a monocular image I_t. Using STAR [5], we split the problem into the following two separate tasks: One deals with relative pose estimation between the current frame and a synthetic keyframe, while the other renders as necessary the synthetic keyframe, which comprises a virtual monocular image and its depth map, to generate a local view (map) of the prior 3D map from the given pose. The alternate tracking and rendering relieve us from the frequent re-rendering process in each iteration, as in [18]. It also enables us to estimate the 6-DoF camera pose even from a large-scale 3D map by reducing the localization problem to successive local tracking, thereby avoiding error accumulation in pose estimation over time.

To perform localization against a keyframe, we leverage an information-theoretic metric for ensuring robustness. Because prior 3D maps can be captured using different sensors, such as a digital camera with different photometric characteristics and LiDAR, conventional photometric error minimization may fail to estimate the accurate camera pose. Therefore, we use NID [19] to compare images with different properties or modalities. NID is a true metric that is independent of the number of samples, unlike MI [20], and it is effective for aligning images under different conditions, as employed in pose estimation in a 3D textured prior map [17] [18]. Benefitting from the above-mentioned characteristics, we develop a STAR-based robust and practical localization technique by leveraging NID.

B. Simultaneous Tracking and Rendering

1) Local tracking: Following prior work [5] [21], we formulate the visual localization problem. Given a monocular image I_t and a synthetic keyframe one I_k rendered at a pose T_W^k, the local tracking problem is defined as $\mathbb{SE}(3)$ optimization with regard to the relative motion $\mathbf{T}_k^t$, by comparing I_t with I_k in terms of a certain similarity metric δI. Specifically, because the keyframe image I_k has the corresponding depth map D_k, each pixel ${}_i\mathbf{u} = ({}_iu, {}_iv) \in I_k$ can be back-projected to a 3D point ${}_i\mathbf{P}^k = ({}_ix, {}_iy, {}_iz)$ in the keyframe coordinate by using the camera projection model $\pi : \mathbb{R}^3 \mapsto \mathbb{R}^2$ as follows:

$$ {}_i\mathbf{P}^k = \pi^{-1}\left({}_i\mathbf{u}, D_k({}_i\mathbf{u})\right), \tag{1} $$

where $\pi^{-1} : \mathbb{R}^2 \mapsto \mathbb{R}^3$ denotes the inverse projection function. The 3D point ${}_i\mathbf{P}^k$ is mapped to the current frame coordinate ${}_i\mathbf{u}'$ by both relative transformation ${}_i\mathbf{P}^t = \mathbf{T}_k^t \, {}_i\mathbf{P}^k$ and subsequent camera projection π as follows:

$$ {}_i\mathbf{u}' = \pi\left({}_i\mathbf{P}^t\right). \tag{2} $$

Notably, the intrinsic camera parameters in the projection π are obtained via a calibration process, and we use the same camera model to render the synthetic keyframes in this paper.

Given the relative pose $\mathbf{T}_k^t$, the 2D–3D and 3D–2D projections enable us to find the pixel-wise correspondences between the current image I_t and keyframe image I_k by observing the same 3D point. Therefore, the most probable relative pose $\hat{\mathbf{T}}_k^t \in \mathbb{SE}(3)$ is obtained by minimizing the sum of per-pixel differences as follows:

$$ \begin{aligned} \hat{\mathbf{T}}_k^t &= \arg\min_{\mathbf{T}_k^t} \delta I\left(I_t, I_k\right) \\ &= \arg\min_{\mathbf{T}_k^t} \sum_{{}_i\mathbf{u}\in\Omega_k} \delta I_i\left(I_t({}_i\mathbf{u}'), I_k({}_i\mathbf{u})\right) \\ &= \arg\min_{\mathbf{T}_k^t} \sum_{{}_i\mathbf{u}\in\Omega_k} \delta I_i\left(I_t(\pi(\mathbf{T}_k^t \cdot \pi^{-1}({}_i\mathbf{u}, D_k({}_i\mathbf{u})))), I_k({}_i\mathbf{u})\right), \end{aligned} \tag{3} $$

where $\Omega_k \subset I_k$ denotes sample pixels that are to be projected onto the current camera image I_t. In our method, all the pixels in I_k with valid depth values are used as the samples Ω_k for performing dense localization.

For example, photometric registration solves the problem by defining the metric δI as a photometric error δI_{photo}; The difference in the intensity domain is directly measured as follows:

$$ \delta I_{photo}\left(I_t, I_k\right) \equiv \sum_{{}_i\mathbf{u}\in\Omega_k} \sigma\left(I_t({}_i\mathbf{u}') - I_k({}_i\mathbf{u})\right) \tag{4} $$

where σ denotes a robust kernel to suppress the effect of outliers, such as t-distribution and Huber kernel [5].

2) Keyframe selection: Because the accuracy of pose tracking against a keyframe deteriorates as the view overlap decreases, we render a new keyframe as necessary to sustain the local tracking. The idea of keyframe selection is common in not only visual localization but also visual SLAM context.

Inspired from [22], we create a new keyframe for the most recently tracked poses for ensuring conservative keyframe update independent of co-visibility [5] or average scene depth [23]. Specifically, the following distance is defined for the exponential map $\xi \in \mathbb{R}^6$ of the Lie algebra $\mathfrak{se}(3)$ corresponding to the relative position and orientation $\mathbf{T}_k^t$:

$$dist\left(\mathbf{T}_k^t\right) \equiv \xi^T \mathbf{W} \xi, \tag{5}$$

where the first three elements of ξ represent a translation whereas the latter three elements represent a rotation, and $\mathbf{W}$ denotes a diagonal weighting matrix to perform keyframe selection that is sensitive to certain movements. When the distance exceeds the threshold τ_{dist}, a new keyframe is generated using the most recent camera pose.

The weights and threshold were empirically determined so that each keyframe could sufficiently overlap with the incoming camera images; however, simultaneously, the rendering process could not be performed frequently, as it incurs high computational cost.

C. Robust metric leveraging information theory

The similarity metric δI should be carefully chosen to compare images with different sensor properties or modalities. In this study, because we aim to localize a monocular camera in a prior 3D map reconstructed using a variety of sensors such as LiDAR and different digital cameras, the appearance difference significantly matters. Because photometric error (see Eq.4), which directly evaluates the similarity of camera images and a 3D prior map in the intensity domain, has an underlying assumption that the appearances are consistent, it is not robust to appearance differences/changes, thereby resulting in tracking failure. Therefore, following [17] [18], we employ NID to stably localize a monocular camera for different modalities.

1) Computation of joint probability and NID: The NID of two discrete random variables, the current frame I_t and keyframe I_k, is given as follows:

$$\delta I_{NID}\left(I_t, I_k\right) \equiv \frac{H(I_t, I_k) - I(I_t; I_k)}{H(I_t, I_k)}, \tag{6}$$

where $H(I_t, I_k)$ and $I(I_t; I_k)$, respectively, denote the joint entropy and MI of samples Ω_k, and they are calculated as follows:

$$\begin{aligned} \mathrm{H}\left(I_t\right) &= -\sum_{x=1}^{n} p_t(x) \log\left(p_t(x)\right), & (7)\\ \mathrm{H}\left(I_t, I_k\right) &= -\sum_{x=1}^{n}\sum_{y=1}^{n} p_{t,k}(x,y) \log\left(p_{t,k}(x,y)\right), & (8)\\ I(I_t; I_k) &= H(I_t) + H(I_k) - H(I_t, I_k), & (9) \end{aligned}$$

where $p_{t,k}$ denotes the joint probability obtained using an $n \times n$-dimensional histogram, and the marginal probabilities p_t and p_k are derived from $p_{t,k}$.

To compute NID, we must obtain $p_{t,k}$ from samples Ω_k. We implemented a weighted voting approach based on [24] (see Fig.3), wherein a 2D cubic B-spline was employed to make the joint probability C^2 continuous for performing gradient-based optimization as follows:

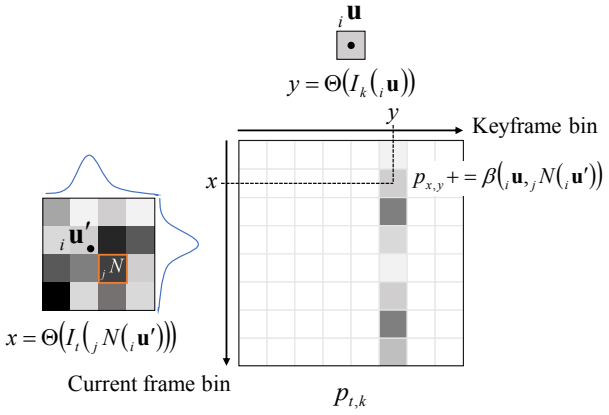


Fig. 3. Joint probability calculation via spatially weighted voting using 2D cubic B-spline.

$$\begin{aligned} p_{t,k}(x,y) &= \frac{1}{\|\Omega_k\|} \sum_{{}_i\mathbf{u}\in\Omega_k} \beta\left({}_i\mathbf{u}, {}_jN\left({}_i\mathbf{u}'\right)\right), & (10)\\ x &= \Theta\left(I_t\left({}_jN\left({}_i\mathbf{u}'\right)\right)\right), & (11)\\ y &= \Theta\left(I_k\left({}_i\mathbf{u}\right)\right), & (12) \end{aligned}$$

where Θ returns a histogram bin corresponding to the input intensity, and $\beta\left({}_i\mathbf{u}, {}_jN\left({}_i\mathbf{u}'\right)\right)$ denotes the B-spline coefficient that weights the contribution of each sample ${}_i\mathbf{u}$ over the 4×4 local support region ${}_jN({}_i\mathbf{u}')$ centered at the reprojection ${}_i\mathbf{u}'$. Notably, the coefficients are normalized as $\sum_j \beta\left({}_i\mathbf{u}, {}_jN\left({}_i\mathbf{u}'\right)\right) = 1, \forall {}_i\mathbf{u}$ to ensure that each sample ${}_i\mathbf{u}$ contributes equally to the histogramming.

2) Computation of derivatives for optimization: Differentiating Eq.6 with respect to the relative transform $\mathbf{T}_k^t$ yields the Jacobian of NID, thereby allowing us to seek the optimal transformation as $\hat{\mathbf{T}}_k^t$, which minimizes the NID between the current frame and keyframe:

$$\frac{d\delta I_{NID}}{d\mathbf{T}_k^t} = \frac{\left(\frac{dH_{t,k}}{d\mathbf{T}_k^t} - \frac{dI_{t,k}}{d\mathbf{T}_k^t}\right) H_{t,k} - \left(H_{t,k} - I_{t;k}\right) \frac{dH_{t,k}}{d\mathbf{T}_k^t}}{H_{t,k}^2}, \tag{13}$$

where

$$\begin{aligned} \frac{dH_{t,k}}{d\mathbf{T}_k^t} &= -\sum_{{}_i\mathbf{u}\in\Omega_k} \frac{dp_{t,k}}{d\mathbf{T}_k^t} \left(1 + \log\left(p_{t,k}\right)\right), & (14)\\ \frac{dI_{t;k}}{d\mathbf{T}_k^t} &= \sum_{{}_i\mathbf{u}\in\Omega_k} \frac{dp_{t,k}}{d\mathbf{T}_k^t} \left(1 + \log\left(\frac{p_{t,k}}{p_t}\right)\right). & (15) \end{aligned}$$

The primary computation involved in the differentiation is finding the derivative of the joint distribution $p_{t,k}$. Notably, the joint distribution can be differentiated using the chain-rule as follows:

$$\begin{aligned} \frac{dp_{t,k}(x,y)}{d\mathbf{T}_k^t} &= \frac{1}{\|\Omega_k\|} \sum_{{}_i\mathbf{u}\in\Omega_k} \frac{d\beta\left({}_i\mathbf{u}, {}_jN\left({}_i\mathbf{u}'\right)\right)}{d\mathbf{T}_k^t}\\ &= \frac{1}{\|\Omega_k\|} \sum_{{}_i\mathbf{u}\in\Omega_k} \frac{d\beta}{d_i\mathbf{u}'} \frac{d_i\mathbf{u}'}{d_i\mathbf{T}_k^t}, \end{aligned} \tag{16}$$

TABLE I
PARAMETERS USED IN THE EXPERIMENTS.

Parameter	value
Camera-image size	800×600
NID-histogram bins n	16
Keyframe-selection threshold τ_{dist}	0.01
Keyframe-selection weights $\mathbf{W}$	diag(0.1, 0.1, 0.1, 1.0, 1.0, 1.0)

where $\frac{\partial_i \mathbf{u}'}{\partial_i \mathbf{T}_k^t}$ denotes the differential of the reprojection ${}_i\mathbf{u}'$ with respect to the relative pose $\mathbf{T}_t^k$ [25] calculated using focal lengths f_x and f_y as follows:

$$\frac{d_i \mathbf{u}'}{d\mathbf{T}_k^t} = \begin{bmatrix} \frac{-f_x}{{}_iz'} & 0 & \frac{{}_iu'}{{}_iz'} & \frac{{}_iu'_i v'}{f_x} & \frac{-f_x^2 - {}_iu'^2}{f_x} & {}_iv' \\ 0 & \frac{-f_y}{{}_iz'} & \frac{{}_iv'}{{}_iz'} & \frac{f_y^2 + {}_iv'^2}{f_y} & -\frac{{}_iu'_i v'}{f_y} & -{}_iu' \end{bmatrix}. \tag{17}$$

According to Eq.13, we perform a gradient-based optimization to determine the most probable relative pose $\mathbf{T}_t^k$. Specifically, the Broyden–Fletcher–Goldfarb–Shannon (BFGS) algorithm implemented in Ceres Solver [26] was adopted for optimizing the local tracking.

IV. EXPERIMENTS

A. Setup

In the following experiments, we used the parameters listed in Table I. Each experiment was performed using a desktop PC equipped with an Intel Core i7-6850K and a GeForce GTX1080, and the NID-histogram calculation was parallelized using CUDA to reduce the computation time.

In addition, we implemented competitive localization methods that use the same STAR framework but different cost functions. Specifically, we chose a method based on photometric errors [5] and one based on edge distances [11] as baselines to compare the localization performances under the same conditions. Both of them use pixels with sufficient gradient; thus, we extract high-gradient pixels with a Sobel filter and image thresholding. In addition, to reproduce [11], a distance transform is applied to the edge image to evaluate edge distances.

B. Quantitative evaluation in the Replica Dataset

To quantitatively evaluate the accuracy and robustness of our method, we used a set of photo-realistic 3D models, namely the Replica Dataset [27]. It provides the 3D models of various indoor environments, and each of them has not only high-quality texture but also semantic labels according to object categories.

We selected "room0" and "apartment0" from the dataset and generated camera-image sequences along predefined trajectories (see Fig.4(a)(i)), which included both translational and rotational motions. Given the initial pose, we successively estimated the camera poses from the sequences by using each localization technique to evaluate the accuracy against the ground truth. In addition, as depicted in the first row in Fig. 4, we used two types of map representations, texture and semantics, and also deteriorated the trajectory image sequences

TABLE II
QUANTITATIVE-EVALUATION RESULTS: RMS TRANSLATION ERROR [M], ROTATION ERROR [DEG], AND LOCALIZATION-SUCCESS RATIO (THE RATIO OF NUMBER OF FRAMES WITH ESTIMATION ERRORS WITHIN 1 M TO THAT OF TOTAL FRAMES) [%].

Index	Map / Image	Photometric [5]	Edge [11]	Our method
I. room0 in the Replica dataset				
R.1	Texture / Camera (Fig.4(a) / Fig.4(c))	0.0148 [m] 0.241 [deg] **100** [%]	0.0139 [m] 0.276 [deg] **100** [%]	**0.00490** [m] **0.0763** [deg] **100** [%]
R.2	Texture / Blurred (Fig.4(a) / Fig.4(d))	0.170 [m] 2.78 [deg] **100** [%]	0.540 [m] 7.95 [deg] 9.5 [%]	**0.0644** [m] **1.06** [deg] **100** [%]
R.3	Texture / Overexposed (Fig.4(a) / Fig.4(e))	0.502 [m] 5.60 [deg] 1.3 [%]	0.736 [m] 14.5 [deg] 1.8 [%]	**0.0128** [m] **0.232** [deg] **100** [%]
R.4	Texture / Occlusion (Fig.4(a) / Fig.4(f))	0.365 [m] 4.63 [deg] 3.6 [%]	- [m] - [deg] 0 [%]	**0.00390** [m] **0.0574** [deg] **100** [%]
R.5	Texture / Salt&Pepper (Fig.4(a) / Fig.4(g))	0.567 [m] 13.1 [deg] 13.7 [%]	0.545 [m] 8.03 [deg] 9.8 [%]	**0.00550** [m] **0.0677** [deg] **100** [%]
R.6	Texture / Underexposed (Fig.4(a) / Fig.4(h))	- [m] - [deg] 0 [%]	- [m] - [deg] 0 [%]	**0.00770** [m] **0.1250** [deg] **100** [%]
R.7	Semantics / Camera (Fig.4(b) / Fig.4(c))	- [m] - [deg] 0 [%]	**0.0249** [m] **0.490** [deg] **100** [%]	0.0710 [m] 1.38 [deg] **100** [%]
II. apartment0 in the Replica dataset				
A.1	Texture / Camera (Fig.4(i) / Fig.4(k))	**0.0117** [m] 0.233 [deg] **100** [%]	0.0203 [m] 0.418 [deg] **100** [%]	0.0147 [m] **0.176** [deg] **100** [%]
A.2	Texture / Blurred (Fig.4(i) / Fig.4(l))	0.193 [m] 4.47 [deg] 22.5 [%]	0.490 [m] 8.14 [deg] 5.2 [%]	**0.0537** [m] **0.821** [deg] **100** [%]
A.3	Texture / Overexposed (Fig.4(i) / Fig.4(m))	0.676 [m] 8.38 [deg] 4.5 [%]	0.402 [m] 3.51 [deg] 0.7 [%]	**0.0141** [m] **0.175** [deg] **100** [%]
A.4	Texture / Occlusion (Fig.4(i) / Fig.4(n))	- [m] - [deg] 0 [%]	- [m] - [deg] 0 [%]	**0.0123** [m] **0.132** [deg] **100** [%]
A.5	Texture / Salt&Pepper (Fig.4(i) / Fig.4(o))	- [m] - [deg] 0 [%]	0.514 [m] 8.56 [deg] 5.3 [%]	**0.00410** [m] **0.0544** [deg] **100** [%]
A.6	Texture / Underexposed (Fig.4(i) / Fig.4(p))	- [m] - [deg] 0 [%]	- [m] - [deg] 0 [%]	**0.0130** [m] **0.185** [deg] **100** [%]
A.7	Semantics / Camera (Fig.4(j) / Fig.4(k))	0.510 [m] 5.82 [deg] 0.5 [%]	**0.171** [m] **2.68** [deg] 32.2 [%]	0.195 [m] 3.23 [deg] **49.7** [%]
III. An indoor environment scaned with a LiDAR				
L.1	Texture / Camera (Fig.4(q) / Fig.4(s))	0.0696 [m] 1.88 [deg] 33.7 [%]	0.0487 [m] 1.35 [deg] 32.5 [%]	**0.0196** [m] **0.966** [deg] **100** [%]
L.2	Texture / Dim light (Fig.4(q) / Fig.4(t))	0.235 [m] 8.43 [deg] 3.3 [%]	0.387 [m] 7.52 [deg] 18.7 [%]	**0.0359** [m] **0.806** [deg] **100** [%]
L.3	Texture / Obstacle (Fig.4(q) / Fig.4(u))	0.0432 [m] 0.896 [deg] **100** [%]	0.0422 [m] 0.915 [deg] 55.6 [%]	**0.0191** [m] **0.417** [deg] **100** [%]
L.4	Reflectance / Camera (Fig.4(r) / Fig.4(s))	0.786 [m] 19.0 [deg] 6.6 [%]	0.119 [m] 3.12 [deg] 7.6 [%]	**0.0589** [m] **1.60** [deg] **100** [%]

to examine robustness against extreme appearance changes or cross-modality.

First, we performed localization experiments in the textured 3D map (see Fig.4(a)) on the basis of the image sequences (see Fig.4(c)–(h)) by assuming different sensor properties in both mapping and localization phases. The translational and rotational RMS errors are depicted in Fig. 5(a)–(f) and Table II(R.1)–(R.6). Each method performed satisfactorily and achieved high accuracies upon performing localization using the image sequence of the original texture. However, when the appearance changed because of blur, over/under exposure, occlusion, or salt-and-pepper noise, tracking with the conventional methods [5] [11] immediately failed. However, our method robustly tracked the agile monocular camera even under extreme appearance changes.

Next, similar experiments were performed using the 3D map colored by semantics (see Fig.4(b)) and the texture image sequence (see Fig.4(c)), assuming the case of different modalities. We simply assigned to each map point a bin number as a color corresponding to the object category because the joint entropy is independent of how the mapping is performed. Figure5(g) and Table II(R.7) show the errors

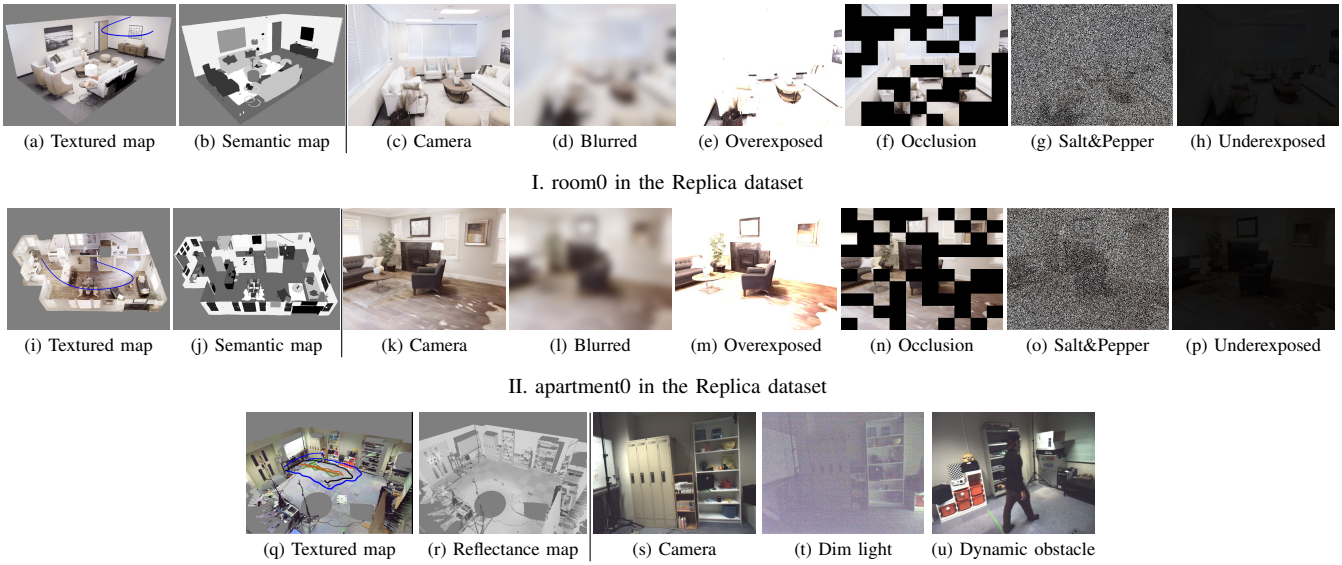


Fig. 4. Various 3D maps and deteriorated image sequences on the trajectories, which are illustrated in (a), (i), and (q). Notably, we captured the image sequences using a hand-held camera in the indoor evaluation; therefore, multiple trajectories are depicted in (q) (blue for (s), orange for (t), green for (u), and black for (r)).

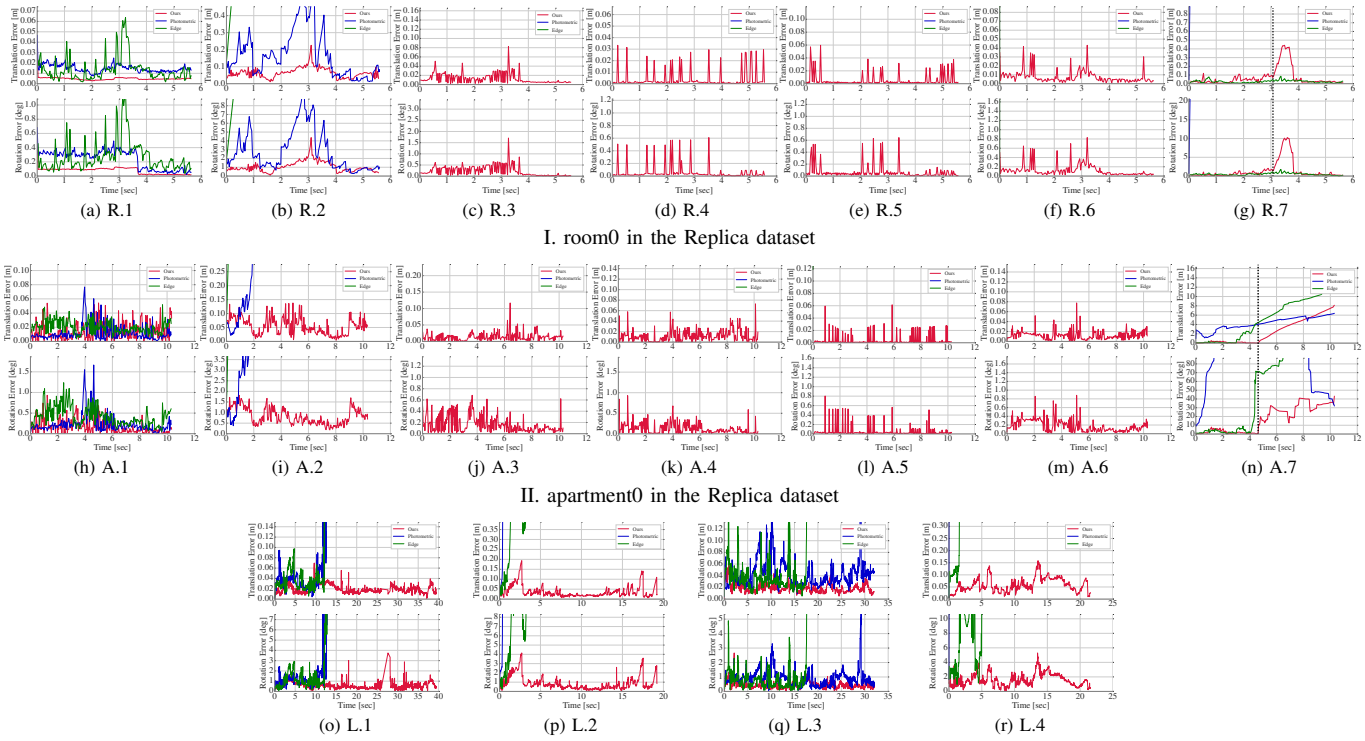


Fig. 5. Translational and rotational RMS errors: Overall, compared with the conventional methods, C^* provided more accurate and robust localization results. Refer to Table II for the map and image sequence used in each experiment index.

of estimates against the ground truth. As in the previous experiment, significant appearance changes made it difficult for the photometric-based method [5] to evaluate the difference between the input and rendered images, thereby resulting in immediate tracking failure. However, edge-based method [11] and the proposed method successfully estimated the camera poses by respectively determining the co-occurrence of edges in both domain, and semantic labels denoted as brightness and texture intensities by leveraging NID. Notably, the localization accuracy of our method reduced at the point indicated by the broken lines in Fig.5(g), and that will be discussed later in Section V.

In addition, we performed similar experiments using another 3D map and image sequences (see Fig.4(i)–(p)), and the

Fig. 6. Images rendered from the poses estimated using C^* based on the corresponding input images in the indoor evaluation. C^* successfully localized the monocular camera even in cases with appearance changes due to sensor properties, lighting conditions, a dynamic obstacle, and different modalities.

corresponding results are shown in Fig.5(h)–(n) and Table II(A.1)–(A.7).

The localization frequency of C^* was approximately 12.5 Hz, which achieved faster localization than that achieved by 2 Hz employed in the previous NID tracking [17] [18], owing to the STAR algorithm, which enabled real-time and robust visual localization. In our implementation, the average time of keyframe update including rendering and the associated data processing was, respectively, 9.36[ms] and 16.3[ms] in room0 and apartment0, and we updated keyframes only 34 and 56 times while the number of total frames on each trajectory was 170 and 289.

C. Evaluation in a LiDAR map

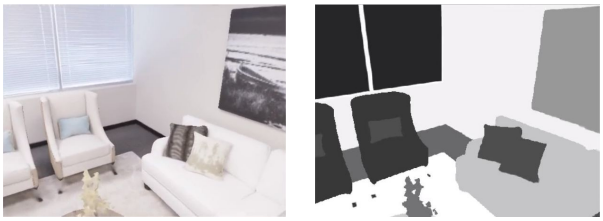
To evaluate our method in a real environment, we constructed a 3D map using a LiDAR Focus3D (FARO Technologies, Inc.), which can capture not only the RGB colors but also near-infrared intensity (laser reflectance) of each 3D point. In addition, we used a digital camera, Flea3 (FLIR Systems, Inc.), to capture image sequences to be tested for 6-DoF localization (see Fig.4(q)–(u)). Notably, the spectral sensitivity and wavelength are different for both the LiDAR and digital cameras, thereby enabling us to examine the localization tolerance against different sensor properties or modalities. Moreover, we estimated the ground truth poses of an agile camera by using OptiTrack (NaturalPoint, Inc.) and, subsequently, compared the estimates from each method with the ground truth in the same manner as in the previous evaluation.

Figure 5(o)–(r) and Table II(L.1)–(L.4) show the experimental results for the indoor environment. In addition, Fig. 6 depicts both the input camera images and the images rendered from the estimated poses in each trial. The results agree with the previous evaluations performed using the Replica dataset.

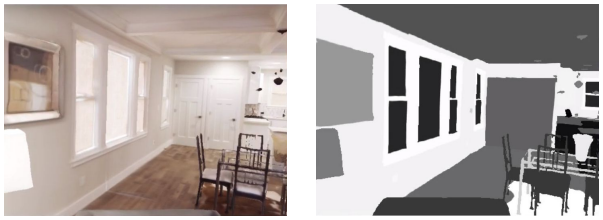
V. CONCLUSIONS

We proposed a cross-modal monocular camera localization method, C^*. Our method achieves highly reliable local tracking using NID while saving the cost of rendering and the associated data processing by occasionally updating keyframes. We proved the validity of the proposed method via experiments with a realistic synthesized dataset and LiDAR scan.

Future work includes improving the robustness of our method. In conclusion, although C^* stably localized the



(a) Around the black broken line illustrated in Fig.5(g)



(b) Around the black broken line illustrated in Fig.5(n)

Fig. 7. Pairs of input and rendered images just before the unstable localization of C^* in R.7 and A.7 : Several semantic colors have almost the same RGB colors (white), thereby preventing the NID cost from shaping a *ravine* appropriate for optimization.

camera poses in most of the experiments, it could not localize accurately or not accomplish to track a trajectory with the semantic image sequences generated in the room0 and apartment0 datasets, as indicated by the broken lines in Fig.5(g)(n). This implies the limitation of C^*: As depicted in Fig.7, when intensity co-occurrence, here the co-occurrence of texture and semantics, was not clearly observed, the NID could not satisfactorily shape the converge basin, thereby resulting in unstable local tracking. As Jeong *et al.* [28] suggested, formulating a cost function as the sum of edge distance, NID, and other available costs might be a solution to obtain a better converge basin for global optimization. However, to apply our visual localization to autonomous navigation, we must simultaneously consider the computational cost for real-time applications, and thus the future work also includes more efficient cost evaluation and keyframe update [29].

REFERENCES

[1] J. Zhang and S. Singh, "LOAM: Lidar Odometry and Mapping in Real-time," in *Robotics: Science and Systems Conference*, July 2014.

[2] M. Yokozuka, S. Oishi, S. Thompson, and A. Banno, "VITAMIN-E: visual tracking and MappINg with extremely dense feature points," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9641–9650.

[3] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[4] L. Liu, H. Li, and Y. Dai, "Efficient Global 2D-3D Matching for Camera Localization in a Large-Scale 3D Map," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2391–2400.

[5] K. Ok, W. N. Greene, and N. Roy, "Simultaneous tracking and rendering: Real-time monocular localization for MAVs," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 4522–4529.

[6] A. Sujiwo, E. Takeuchi, L. Y. Morales, N. Akai, Y. Ninomiya, and M. Edahiro, "Localization based on multiple visual-metric maps," in *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2017, pp. 212–219.

[7] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J. J. Leonard, and J. McDonald, "Real-time large scale dense rgb-d slam with volumetric fusion," *The International Journal of Robotics Research*, vol. 34, pp. 598–626, April 2015.

[8] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular camera localization in 3D LiDAR maps," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 1926–1931.

[9] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[10] P. Neubert, S. Schubert, and P. Protzel, "Sampling-based methods for visual navigation in 3D maps by synthesizing depth images," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 2492–2498.

[11] K. Qiu, T. Liu, and S. Shen, "Model-Based Global Localization for Aerial Robots Using Edge Alignment," *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1256–1263, 2017.

[12] D. Wong, Y. Kawanishi, D. Deguchi, I. Ide, and H. Murase, "Monocular localization within sparse voxel maps," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 499–504.

[13] Y. Lu, J. Huang, Y. Chen, and B. Heisele, "Monocular localization in urban environments using road markings," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 468–474.

[14] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5974–5983.

[15] Q. Zhou, T. Sattler, M. Pollefeys, and L. Leal-Taix, "To Learn or Not to Learn: Visual Localization from Essential Matrices," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[16] R. W. Wolcott and R. M. Eustice, "Visual localization within lidar maps for automated urban driving," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 176–183.

[17] G. Pascoe, W. Maddern, and P. Newman, "Robust Direct Visual Localisation using Normalised Information Distance," in *The British Machine Vision Conference (BMVC)*, 2015, pp. 70.1–70.13.

[18] G. Pascoe, W. Maddern, A. D. Stewart, and P. Newman, "Farlap: Fast robust localisation using appearance priors," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 6366–6373.

[19] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitanyi, "The similarity metric," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, Dec 2004.

[20] N. X. Vinh, J. Epps, and J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," *The Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.

[21] F. Steinbruecker, J. Sturm, and D. Cremers, "Real-Time Visual Odometry from Dense RGB-D Images," in *Workshop on Live Dense Reconstruction with Moving Cameras at the Intl. Conf. on Computer Vision (ICCV)*, 2011.

[22] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-Scale Direct Monocular SLAM," in *European Conference on Computer Vision*, 2014, pp. 834–849.

[23] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast Semi-Direct Monocular Visual Odometry," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[24] G. Pascoe, W. Maddern, M. Tanner, P. Pinies, and P. Newman, "NID-SLAM: Robust Monocular SLAM Using Normalised Information Distance," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1446–1455.

[25] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, Oct 1996.

[26] S. Agarwal, K. Mierle, and Others, "Ceres solver," http://ceres-solver.org (Accessed September, 15, 2019).

[27] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, "The Replica Dataset: A Digital Replica of Indoor Spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[28] J. Jeong, L. Y. Cho, and A. Kim, "Road is Enough! Extrinsic Calibration of Non-overlapping Stereo Camera and LiDAR using Road Information," *arXiv preprint arXiv:1902.10586*, 2019.

[29] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A Platform for Embodied AI Research," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.