Unsupervised Learning of Monocular Depth and Ego-Motion in Outdoor/Indoor Environments

Ruipeng Gao, Xuan Xiao, Weiwei Xing, *Member, IEEE*, Chi Li, and Lei Liu

***Abstract*—Visual-based unsupervised learning [1]–[3] has emerged as a promising approach in estimating monocular depth and ego-motion, avoiding intensive efforts on collecting and labeling the ground truth. However, they are still restrained by the brightness constancy assumption among video sequences, especially susceptible with frequent illumination variations or nearby textureless surroundings in indoor environments. In this article, we selectively combine the complementary strength of visual and inertial measurements, i.e., videos extract static and distinct features while inertial readings depict scale-consistent and environment-agnostic movements, and propose a novel unsupervised learning framework to predict both monocular depth and ego-motion trajectory simultaneously. This challenging task is solved by learning both forward and backward inertial sequences to eliminate inevitable noises, and reweighting visual and inertial features via gated neural networks in various environments or with user-specific moving dynamics. In addition, we also employ structure cues to produce scene depths from a single image and explore structure consistency constraints to calibrate the depth estimates in indoor buildings. Experiments on the outdoor KITTI data set and our dedicated indoor prototype reveal that our approach consistently outperforms the state of the art on both depth and ego-motion estimates. To the best of our knowledge, this is the first work to fuse visual and inertial data without any supervision signals for monocular depth and ego-motion estimation, and our solution remain effective and robust even in textureless indoor scenarios.**

***Index Terms*—Ego-motion, monocular depth, structure cues, unsupervised learning, visual-inertial fusion.**

I. Introduction

INFERRING monocular depth and ego-motion is the basis for novel AR/VR features in various Internet of Things (IoT) systems. For example, it produces high-quality depth information to inexpensively complement LIDAR sensors used in self-driving cars [4]. It also enables 3-D object detection as a new smart payment tool on mobile phones [5] and guides users to explore indoor locations and avoid obstacles when playing the Pokémon GO [6]. With more than a decade of research on computer vision since the Structure from Motion (SfM) [7], recent approaches have started to exploit unsupervised deep learning methods [3], [8], [9] to estimate both depth and ego-motion from monocular videos. Although they have explored a variety of object masks and loss functions between consecutive frames to improve the accuracy, one major obstacle to its ubiquitous availability is the image outliers in specific environments, e.g., with frequent illumination variations, nearby occluded/moving objects, and capturing textureless surfaces. Thus, its robustness is inevitably low, especially in indoor buildings, such as manmade tunnels and parking structures.

Additionally, when using the ego-motion network to track a trajectory, it requires a consistent scale over the entire video sequence. However, vision-based deep models are always learned via pairwise stitching, causing the per-frame scale ambiguity and producing individual scale factors on different snippets, thus impeding wide adoptions for long-time location inference.

Inspired by the complementary strengths from different sensing modalities, we explore a visual-inertial fusion approach that enables the environment robustness of monocular depth estimates and the scale consistence of ego-motion trajectories. Especially, visual ones can capture accurate translation measurements in areas with distinct appearances and proper illuminations, and can be exploited to calibrate inertial noises; while the inertial sensor is proprioceptive thus can produce scale-consistent and environment-agnostic measurements, and its much higher sampling rate ($\sim$ 100 Hz) is suitable to track fast movements, such as braking and turning. Thus, we aim to devise an automatically reweighting strategy, which selectively fuses visual and inertial features according to dynamic environmental conditions and their specific sensing properties.

The realization of such benefits, however, turns out to be a nontrivial journey. First, due to inevitable noises and bias in inertial measurements, we need to calibrate raw inertial readings and extract the most effective and robust motion features, instead of directly conducting integrations for trajectory tracing. Second, there are numerous extreme environmental conditions (e.g., illumination variations, occluded/moving objects, and textureless surfaces) and user-specific moving dynamics (e.g., fast rotation, sudden brakes, and bump jolting) in the real-world scenarios, thus our model should automatically select and reweigh the optimal motion features from different sensing modalities. Third, when driving in indoor environments, such as manmade tunnels and parking structures, our surroundings are mainly comprised of white walls and floors/ceilings with few distinct decorations; thus, even the

Manuscript received 18 October 2021; revised 6 January 2022; accepted 7 February 2022. Date of publication 16 February 2022; date of current version 24 August 2022. This work was supported in part by the Beijing NSF under Grant L192004; in part by NSFC under Grant 62072029 and Grant 61876018; and in part by the DiDi Research Collaboration Plan. *(Corresponding author: Ruipeng Gao.)*

Ruipeng Gao, Xuan Xiao, and Weiwei Xing are with the School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: rpgao@bjtu.edu.cn; xiaoxuan@bjtu.edu.cn; wwxing@bjtu.edu.cn).

Chi Li and Lei Liu are with the Maps and Public Transportation Department, DiDi Corporation, Beijing 100089, China (e-mail: lichi@didiglobal.com; liuleifrey@didiglobal.com).

Digital Object Identifier 10.1109/JIOT.2022.3151629

deep neural networks (DNNs) cannot extract sufficient visual features to align and match similar images.

In this article, we propose a selective visual-inertial fusion framework with unsupervised learning. It exploits DNNs to calibrate inertial readings and extract the most effective inertial motion features, with a monocular camera as the supervised signal. The inertial features further enhance the ego-motion network with scale-consistent trajectories and produce more precise warped frames to supervise the depth network in extreme environments. In addition, for vehicles in manmade indoor buildings, we also explore the principle of structure cues to infer multiple geometric scenes from a single image, and derive them to calibrate the depth estimates. To the best of our knowledge on monocular depth and ego-motion estimation, this article is the first to combine visual and inertial observations within one unsupervised learning framework, and our solution remains robust even in textureless indoor environments.

Specifically, we make the following contributions.

1) We devise a recurrent neural network to eliminate inertial outliers and extract the most effective motion features from inertial readings. It is trained only with the unlabeled monocular video sequences within our unsupervised learning framework.
2) We propose a generic sensor fusion strategy based on gated neural networks, which selectively combines complementary sensing modalities and produces environment-robust and scale-consistent ego-motion trajectories. It, in turn, calibrates the transformation of warped images, thus improving the accuracy of monocular depth estimates.
3) We explore the principle of indoor structure cues from a single image and produce corresponding scene depth information of each pixel point. We further exploit the scene depth map to calibrate indoor depth estimates by DNNs.
4) We perform extensive evaluations on KITTI data sets for outdoor scenarios and develop our dedicated prototype for indoor scenarios. Our approach consistently outperforms the state of the art on both depth and ego-motion estimates. We also demonstrate the effectiveness of our ego-motion network for long-time vehicle tracking.

II. Related Work

Supervised Depth Estimation: Learning-based approaches have shown significant effectiveness to predict the depth of each pixel on a colorful image. They explore various supervision signals to infer the depth, including depth sensors [10]–[12], objects with known size [13], sparse ordinal depths [14], matched appearances [9], and unpaired synthetic depth measurements [15]. However, one major obstacle to ubiquitous usability is the lack of ground-truth depth information and environmental annotations. Mayer *et al.* [16] investigated the potential of using synthetic training data, which cannot involve every situation in the real world. Some recent work [17], [18] employ conventional SfM to yield sparse training data, where the SfM and learning platforms are always separated. Thus, the supervised approaches are still susceptible to camera fluctuations and external disturbances, especially in complex and dynamic driving scenarios.

Unsupervised/Semisupervised Depth Estimation: While supervised approaches have demonstrated promising results, the expense of gathering sufficient ground-truth data prevents them from ubiquitous practical use. Unsupervised/semisupervised approaches are explored to solve this problem. Garg *et al.* [19] first investigated the geometry constraints between stereo-cameras as supervision for training a self-supervised network for depth estimation. Compared to stereo-cameras, monocular videos are more prevalent in daily life thus are more preferred for depth estimation. Zhou *et al.* [8] leveraged a sequence of image frames taken by a monocular camera and use the constraints from adjacent frames as supervision. However, due to the camera movements, the moving objects in the scene can easily impact the estimation performance. To solve this problem, several work [2], [3] leverage image masks to get rid of the noisy and useless parts. In order to improve the accuracy of depth prediction, monodepth2 [1] effectively selects the pixels which are more suitable for loss computation, and UnVIO [20] employ a loss item on 3-D geometric consistency which is extremely time-consuming ($10\times$ more) to construct the point cloud thus we remove it for efficiency. Since monocular video-based depth estimation is closely related to the camera movement, jointly estimating both depth and ego-motion is a more nature way to improve the performance. To improve the accuracy of pose estimation, DF-VO [21] explores an integrating deep learning method with epipolar geometry.

Visual-Inertial Sensor Fusion: Visual-inertial odometry (VIO) has been successfully used for camera pose tracking which enabled a lot of augmented reality (AR) and navigation applications. There are many famous VIO algorithms which have fused vision and inertial data with either filters or optimization frameworks to improve the tracking accuracy, e.g., MSCKF [22], VINS [23], ROVIO [24], and OKVIS [25]. Such algorithms take the similar idea of leveraging visual and inertial data as our work, but leverage them in a deterministic way, e.g., only using inertial measurement unit (IMU) data for the scale factor of whole scene ([26]). In addition, they always rely on handcrafted features, but naively using all features may lead to incorrect feature extraction or matching, e.g., crippling the entire system in low-light conditions or with excess inertial noises. Some latest approaches have used deep learning techniques for more robust data fusion, e.g., the selective fusion [27] and VINet [28]. While VIO algorithms work well for ego-motion estimation, they only produce very sparse depth samples, which are far from a complete depth map of the scene. Our approach selectively fuses visual and inertial data for depth and ego-motion estimation simultaneously, which can not only provide robust ego-motion estimation as traditional VIO algorithms but also generate depth estimations at a more fine-grained level.

III. Overview

In this article, we aim to learn monocular depth and ego-motion via a selective visual-inertial fusion perspective. Our

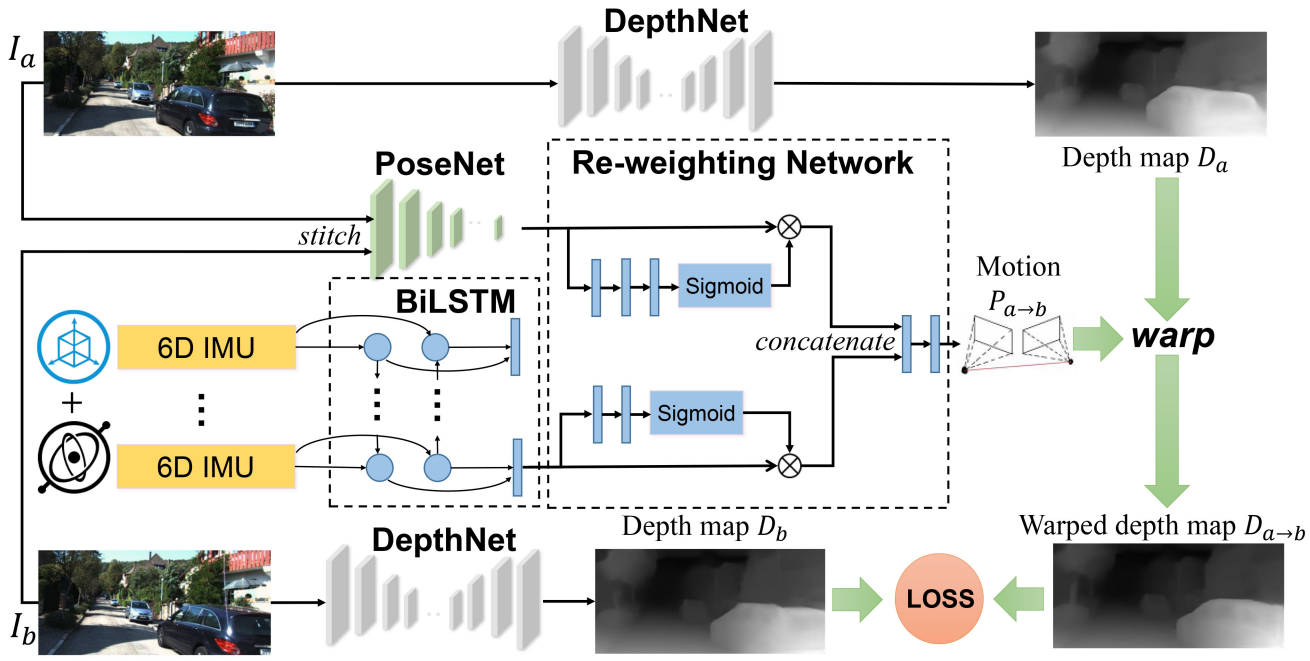

Fig. 1. Overview of our unsupervised learning framework. Given two consecutive frames I_a and I_b, the DepthNet [29] predicts individual depth maps D_a and D_b, while the PoseNet [8] and the BiLSTM [30] encode visual and inertial motion features between two frames, respectively. Our Reweighting Network selectively fuses motion features from two modalities, and produces 6-D ego-motion estimates $P_{a\to b}$. During training, we warp both frame I_a and its depth map D_a to the other pose, and compute loss without any supervision signals.

TABLE I
NOTATIONS

Symbols	Descriptions
$< I_a, I_b >$	Two consecutive image frames
$< D_a, D_b >$	Estimated depth maps for image pair $< I_a, I_b >$
$\mathbf{x}_{a:b}$	6D inertial sequence between two frames I_a and I_b
$P_{a\to b}$	6D ego-motion transformation from frame I_a to frame I_b
$I_{a\to b}$	The warped image of frame I_a according to $P_{a\to b}$
$D_{a\to b}$	The warped image of depth map D_a according to $P_{a\to b}$
D'_b	The interpolated image of depth map D_b to match $D_{a\to b}$

model inputs consist of unlabeled videos from a monocular camera and inertial measurements from an IMU, including 3-axis accelerations from an accelerometer and 3-axis angular rates from a gyroscope. Fig. 1 depicts a modular overview on our unsupervised learning framework with four components: 1) DepthNet [29] as the monocular depth network; 2) PoseNet [8] as the visual-based motion encoder; 3) BiLSTM [30] as the inertial-based motion encoder; and 4) a reweighting network for selective feature fusion.

Table I demonstrates the nation of symbols used in our model. Especially, given two consecutive frames I_a and I_b, the DepthNet predicts their individual depth map D_a and D_b. In the meanwhile, the PoseNet and the BiLSTM encode visual and inertial motion features between two frames, respectively. Our reweighting network selectively fuses motion features from two modalities and produces 6-D ego-motion estimates $P_{a\to b}$ for image warping. During training, we warp both frame I_a and its depth map D_a to the other pose and compute the corresponding geometry consistency loss (i.e., the depth inconsistency between $D_{a\to b}$ and the interpolated D'_b), photometric loss (i.e., the color inconsistency between $I_{a\to b}$ and I_b), depth smoothness loss, and our structure consistency loss to constrain the model. This is an unsupervised learning framework without any supervision signals.

IV. Motion Feature Extraction and Fusion

In this section, we present how to extract the most effective and robust motion features from visual and inertial readings, respectively. We also propose a selective feature fusion mechanism to combine their complementary strengths in different environments or with user-specific moving dynamics.

A. Visual-Based Motion Encoder

Given two monocular images $\mathbf{I}_a$ and $\mathbf{I}_b$, we stack them and employ the PoseNet [8] as our visual-based motion encoder. It consists of seven stride-2 convolutions followed by a 1×1 convolution with $6(N-1)$ output channels, corresponding to three Euler angles and 3-D translations for 6-D ego-motion representation. All layers are followed by a ReLU activation function except for the last one. We denote the visual-based motion feature as $\mathbf{f}_V$

$$\mathbf{f}_V = \text{PoseNet}(\mathbf{I}_a, \mathbf{I}_b). \tag{1}$$

B. Inertial-Based Motion Encoder

Compared with a camera, the proprioceptive inertial sensor produces high-frequency, environment-agnostic, and scale-consistent motion measurements, especially over a long trajectory. However, inertial readings are easily plagued by heavy noises from a commodity IMU, causing unbounded tracking errors through double integrations. Thus, we adopt a bidirectional long short-term memory (LSTM) network

as the inertial-based feature encoder. Intuitively, its bidirectional structure captures both forward and backward motion transformations between two poses, which is consistent with warping pairwise images back and forth. We denote the inertial sequence between each two images as $\mathbf{x}_{a:b}$, and extract its corresponding motion feature $\mathbf{f}_I$ as

$$\mathbf{f}_I = \text{BiLSTM}(\mathbf{x}_{a:b}). \tag{2}$$

C. Selective Feature Fusion

Intuitively, visual and inertial measurements have complementary strengths. Videos are suitable to predict accurate translations in static and distinct areas under proper illuminations, while inertial data provide high-frequency, scale-consistent, and environment-agnostic motion estimations.

Therefore, we explore a visual-inertial reweighting network to selectively fuse them according to various environmental conditions and user-specific moving dynamics. Especially, we apply gated operations by a sigmoid nonlinearity to produce dynamic confidence on each modality, i.e.,

$$w_I = \text{Sigmoid}_I(\mathbf{f}'_I), w_V = \text{Sigmoid}_V(\mathbf{f}'_V) \tag{3}$$

where $\mathbf{f}'_I = \text{FC}(\mathbf{f}_I)$ and $\mathbf{f}'_V = \text{FC}(\mathbf{f}_V)$, namely, the transformed inertial and visual motion features after several FC layers.

Finally, we multiply individual motion feature with its corresponding weight, concatenate the reweighted features among two sensing modalities, and deploy two FC layers to predict the relative 6-D ego-motion, i.e.,

$$P_{a\to b} = \text{FC}([w_I * \mathbf{f}_I; w_V * \mathbf{f}_V]). \tag{4}$$

Fig. 2 shows the learned weights of both visual and inertial features along an example trajectory, and we have found several interesting observations.

1) Inertial features contribute more in tracking fast rotation movements (e.g., turning), deriving from high sampling rates and precise rotation measurements from the gyroscope.
2) The inertial weight increases for low-speed tracking, and decreases with sharp translation movements (e.g., speeding and braking), due to the low-quality of a commodity accelerometer.
3) The visual weight remains at a low level with frequent illumination variations, which violate the brightness constancy assumption. These demonstrate the effectiveness of our selective fusion approach.

V. Unsupervised Learning

Inspired by other unsupervised learning approaches [2], [3], [8], [31], we exploit the principle of projection geometry and warp each frame and its depth map into a related pose, and design multiple loss items to constrain the model training without any supervision signals.

A. Projection Geometry

Given one frame I_a, we predict its corresponding depth map D_a via the DepthNet (shown in Fig. 1), thus each pixel on

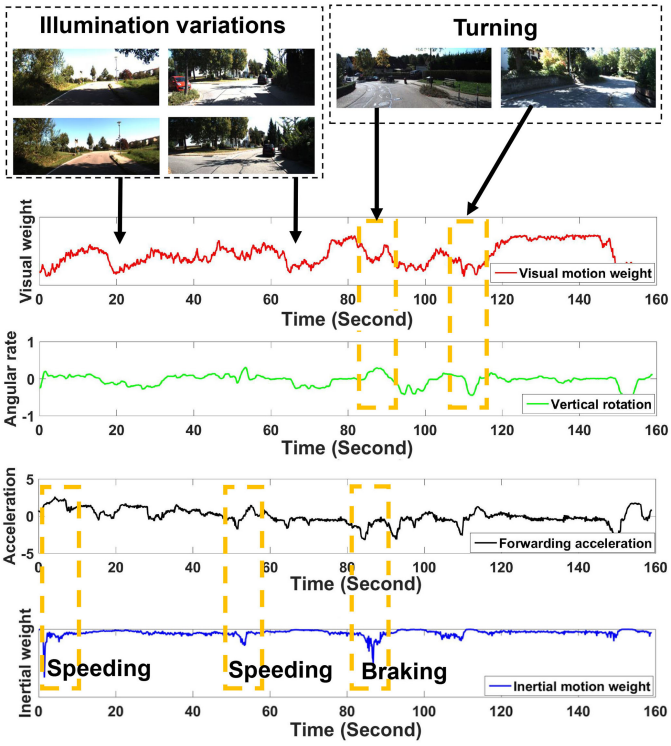

Fig. 2. Selective visual-inertial fusion examples for monocular depth and ego-motion estimation. Top to bottom: visual weights, vehicle's angular rate, vehicle's forwarding accelerations, and inertial weights. Illumination variations and turning actions decrease visual weights, while sharp speeding/braking reduce inertial weights.

frame I_a can be projected back into the 3-D scene, i.e.,

$$Q(i,j) = K^{-1} \cdot D_a(i,j) \cdot [i,j,1]^T \tag{5}$$

where $Q(i,j)$ denotes the projected 3-D point cloud for each pixel on the frame, and K is the camera intrinsic matrix which is learned by camera calibration.

Next, once the camera's movement $P_{a\to b}$ to another frame I_b is acquired by the motion network, we reproject the 3-D point cloud onto this related pose, thus transforming the depth map D_a to a warped depth map $D_{a\to b}$, i.e.,

$$\begin{aligned} D_{a\to b}(\hat{i},\hat{j}) \cdot [\hat{i},\hat{j},1]^T &= K \cdot P_{a\to b} \cdot Q(i,j) \\ &= K \cdot P_{a\to b} \cdot \left(K^{-1} \cdot D_a(i,j) \cdot [i,j,1]^T\right) \end{aligned} \tag{6}$$

where $[\hat{i},\hat{j},1]^T$ denotes the transformed pixel on warped depth map $D_{a\to b}$. This equation also ensures $[\hat{i},\hat{j},1]^T$ as homogeneous coordinates on the image.

B. Loss Computation

First, (6) requires the geometry consistency between two frames, i.e., they correspond to the same point cloud in 3-D scene; thus, the warped depth map should be consistent with the interpolated one. Following [3], we define the *geometry consistency loss* among all pixel coordinates:

$$L_{gc} = \sum \frac{|D_{a\to b} - D'_b|}{D_{a\to b} + D'_b} \cdot M_{gc} \tag{7}$$

where M_{gc} denotes the validity mask matrix [2], and D'_b is the interpolated depth.

Second, based on the well-known brightness constancy assumption [32], we summarize color constancy errors between a frame and a warped image which is transformed via motion estimation from a related frame. The image *reconstruction loss* among all pixel coordinates is formulated as

$$L_{\text{rec}} = \sum \lambda_1 \|(I_{a\to b} - I_b) \cdot M_{\text{rec}}\|_1 + \lambda_2 (1 - \text{SSIM}(I_{a\to b}, I_b)) \cdot M_{\text{rec}} \quad (8)$$

where M_{rec} is the weight mask for active objects which is computed analytically from the predicted depth and ego-motion. SSIM denotes the structural similarity metric [33], and (λ_1, λ_2) are weight parameters. Especially, we set $\lambda_1 = 0.15$ and $\lambda_2 = 0.85$ in our model.

In addition, a *depth smoothness loss* has been widely used to regularize depth estimations [19], since it tolerates sharp variations at pixels in the depth map which are consistent with the original frame. For the depth map D of image I, it is formulated among all pixel coordinates

$$L_{ds} = \sum \|\partial_x D\| e^{-\|\partial_x I\|} + \|\partial_y D\| e^{-\|\partial_y I\|}. \quad (9)$$

In sum, our overall loss function in outdoor scenarios is defined as

$$L_{\text{total}} = \alpha L_{gc} + \beta L_{\text{rec}} + \gamma L_{ds} \quad (10)$$

which α, β, and γ are hyperparameters. We set $\alpha = 0.5$, $\beta = 1$, and $\gamma = 0.1$ during training.

VI. Structure Cue

The projection geometry largely relies on distinct feature points in surrounding area, thus is not suitable in featureless indoor environments with similar decorations and white walls, e.g., tunnels, underground parking structures, and multilevel overpasses. In addition, since GPS receptions are obscured or even blocked indoors, existing scene optimization methods (e.g., bundle adjustment [34] and loop closure detection [35]) cannot be launched. We observe that majority indoor environments are comprised of manmade buildings, which usually follow the Manhattan world assumption[1] with many structure cues in such environments, e.g., two-side walls are parallel and they both are perpendicular to ceilings and floors. In this section, we exploit such structure cues to infer the geometric scene from a single image and use it to calibrate the learned depth estimates by DepthNet.

A. Geometric Reasoning from Single Image

Line Segments and Vanishing Points: We use the Canny edge detector [36] to extract line segments on each image, and follow Rother [37] to estimate three orthogonal vanishing points correlated to the image. Especially, it adopts a random sample consensus (RANSAC) solution to fine tune the orthogonality under optimization; thus, clusters line segments into three categories oriented to three vanishing points [marked as green, red, and blue on Fig. 3(a)], and removes line segments in other directions as outliers (marked as yellow). This clustering algorithm is based on the Manhattan world assumption, which is suitable for majority manmade buildings. We denote

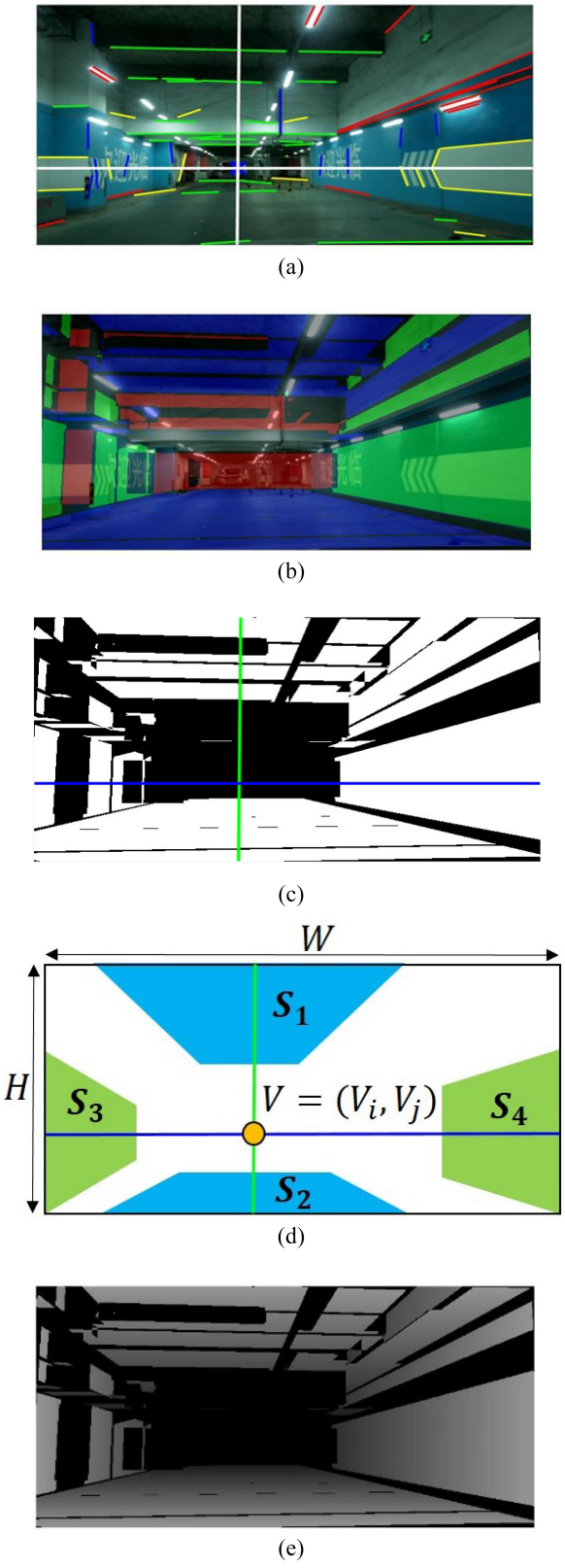

Fig. 3. Geometric reasoning from a single image. (a) Line segment extraction and clustering. (b) Orientation map. (c) Orientation mask and structure lines. (d) Four major planes with respective depth weight. (e) Scene depth map.

[1]Scenes are built on a cartesian grid which leads to regularities in the image edge gradient statistics, which is suitable for majority manmade buildings.

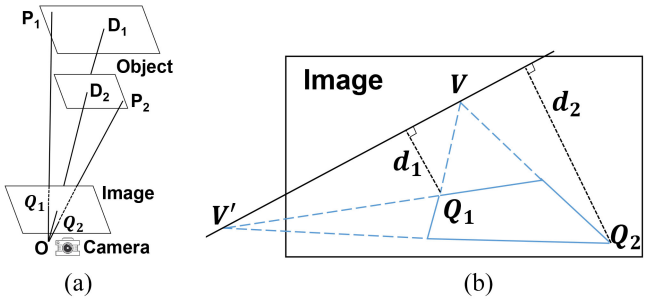

Fig. 4. Illustration for scene depth by structure cues. (a) 3-D view, where O is a camera on the vehicle, P_1 and P_2 are two practical points and they are projected to Q_1 and Q_2 on image plane. (b) Image view, where V and V' are two vanishing points, and VV' forms the structure line.

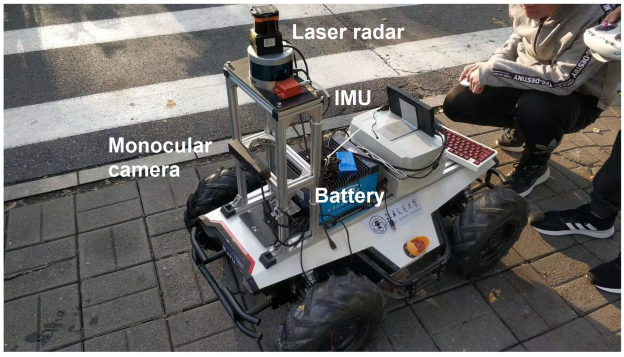

Fig. 5. Our dedicated prototype for data collection in indoor buildings.

the center vanishing point as $V = (V_i, V_j)$ on the image, i.e., the intersection point for all red lines on Fig. 3(a).

Orientation Map: As shown in Fig. 3(b), we leverage a geometric reasoning algorithm [38] to generate the corresponding orientation image based on these oriented line segments, i.e., the orientation of each pixel is perpendicular to the orientation of two nearby line segments. For example, the horizontal floor surface is produced by a green line above it and two red lines to its two sides. Intuitively, the orientation map represents the orientation consistency in pixel regions, thus restricting the depth estimates with plane constraints under projective geometry. In addition, we observe that there exists blank pixels with no orientations in the orientation map, which denote some uncertain areas in other orientations against the Manhattan world assumption; thus, we formulate a corresponding orientation mask [Fig. 3(c)] to record the orientation confidence only for horizontal and width planes (i.e., removing vertical and uncertain planes).

Scene Depth Map: For a typical image taken in manmade environments, its orientation map usually produces four major planes [Fig. 3(d)], including one ceiling plane S_1 (up), one floor plane S_2 (down), and two surrounding walls S_3 (left) and S_4 (right) at both sides.

1) According to the principle of the projective geometry, the scene depth for image points on a plane is inversely proportional to its respective distance to the corresponding structure line (Fig. 4), i.e., $|OD_1|/|OD_2| = (d_2/d_1)$ where V and V' are two vanishing points and VV' is connected as the structure line. Thus, it establishes the scene depth relations on each plane.
2) We aim to connect all planes in an image to calculate its depth estimates within a uniform scale. Intuitively, we observe that the position of center vanishing point V reflects the 2-D rotation of a camera. For example, if the camera's orientation is rotated upward and to the right, the vanishing point V will accordingly move in the inverse direction on the image, i.e., toward the left bottom corner. Therefore, the camera captures more and closer views on top and right regions by this rotation event. For each image point p_k (within the orientation mask), we define its weighted scene depth s_k as

$$s_k = \frac{w_{S_k}}{\text{dist}(p_k, l_k)} \tag{11}$$

where $S_k \in \{S_1, S_2, S_3, S_4\}$ denotes the plane index that p_k belongs to, l_k denotes its corresponding vanishing line, and dist() computes the distance from a pixel point to an image line. w_{S_k} represents the weight of plane S_k based on the center vanishing point V, i.e.,

$$w_{S_1} = \frac{V_j}{H}, w_{S_2} = \frac{H - V_j}{H}, w_{S_3} = \frac{V_i}{W}, w_{S_4} = \frac{W - V_j}{W} \tag{12}$$

where $W \times H$ denotes the image resolution. Finally, we produce the scene depth map for each image [Fig. 3(e)].

B. Structure Consistency

The scene depth map constrains the structure consistency of pixel depths on four major planes, thus we further use it to calibrate depth estimates from the DepthNet. Since there is a depth scale factor between scene depth map and DepthNet, we aim to minimize their variations with a uniform scale for all pixel points p_k on the image, thus the *structure consistency loss* is defined as

$$L_s = \min_{\eta} \sum_{p_k \in I} M_k \cdot \|\widehat{d}_k - \eta s_k\|^2 \tag{13}$$

where M denotes the orientation map of image I, $\widehat{d}_k$ is the estimated depth for pixel point p_k via the DepthNet, s_k is its corresponding scene depth, and η denotes the optimal scale factor for this image.

Finally, for indoor environments with structure cues, we add the structure consistency loss item into total loss computation, i.e.,

$$L_{\text{total}} = \alpha L_{gc} + \beta L_{\text{rec}} + \gamma L_{ds} + \delta L_s \tag{14}$$

with $\delta = 0.1$ during our training process.

VII. Evaluation

A. Implementation

Outdoor Data Set: For outdoor driving scenarios, the KITTI data set [39] is the most common benchmark to evaluate the accuracy of depth and ego-motion. It also provides both monocular videos and inertial data with accurate time stamps. To compare with prior work, we transform the image resolution as 416×128. In addition, the labeled 3-D point clouds

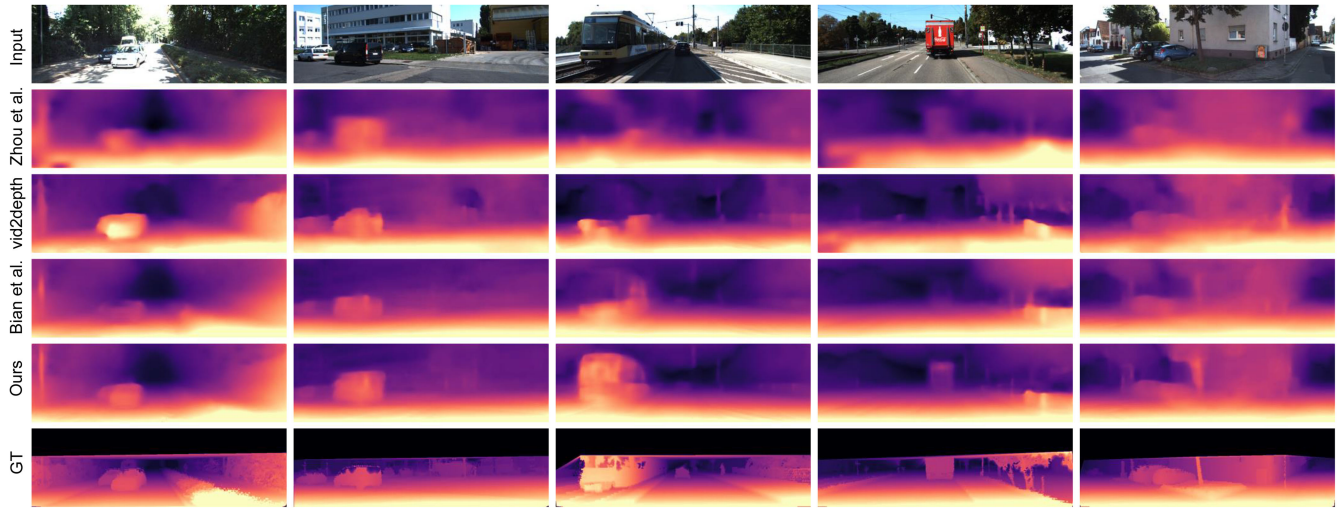

Fig. 6. Sample depth estimates from KITTI database, produced by Zhou *et al.* [8], vid2depth [2], Bian *et al.* [3], our approach, and the ground truth from 3-D point cloud.

and vehicle poses in KITTI are only used as ground-truth signals for evaluation. We follow Eigen [10] to split the data set: 40 238 frames with corresponding inertial readings for training, 4448 for validation, and 697 for test.

Indoor Prototype: Although there are several public data sets for indoor Unmanned Ground Vehicles (UGV), they either do not gather inertial readings (e.g., NYU Depth V2 [40]) or only focus on fine-grained objects instead of large-scale hallways and lobbies (e.g., EuRoc MAV Data set [41]). Thus, we develop a dedicated UGV prototype (Fig. 5) based on NVIDIA Jetson Xavier NX with a monocular camera (10 Hz) and a commonly-used IMU (100 Hz). We use it to collect data in a 80 m$\times$50 m teaching building in our campus, and attach a laser radar (10 Hz) to measure the ground truth of scene depth. The image resolution is set as 480×288, and the ratio for training/validation/test split is $8:1:1$, i.e., there are 13 769 frames for training, 739 for validation, and 739 for test.

Data Augmentation: For videos, we augment each frame with random scaling, cropping, and horizontal flips to increase the image diversity. For inertial readings in KITTI data set, although there is a "synced" data set, its visual and inertial measurements are recorded at the same sampling rate (10 Hz) which cannot capture fast movements, thus we choose inertial measurements from the "unsynced" data set (100 Hz) and align them with the image time stamp. In addition, in order to ensure ten IMU samples between two consecutive frames, we linearly interpolate vacant samples and randomly remove redundant ones. Note that since we have horizontally flipped some random frames for augmentation, we also conduct flipping operations for corresponding inertial sequences.

Training Details: We implemented our system using the TensorFlow framework [42]. We modify the DispNet [29] with single-scale supervision as our depth network for KITTI data set, and with four scales in indoor scenarios. We also employ the PoseNet [8] without mask prediction branch as our visual-based motion encoder. In order to improve data utility, pairwise frames are trained both forward and backward in PoseNet. We leverage batch normalization and adopt the Adam optimizer [43] ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) at learning rate of $2e-4$. The training batch size is set as 4. Table II depicts the detailed settings of key parameters.

TABLE II
KEY PARAMETERS

Components	Parameters&Values
DepthNet [29]	All convolutional layers are followed by ReLU activation, except for the prediction layers as $1/(10 * sigmoid(x) + 0.1)$
PoseNet [8]	7 convolutions at the stride of 2, followed by a 1×1 convolution
BiLSTM	Input size of $10 * 6$, and hidden units of 32
Re-weighting Network	3 image FC layers (units of 128, 32, and 1) and 2 inertial FC layers (units of 32 and 1)
Structure cues	Minimum line length of 20 pixels and maximum line segment gap of 5 pixels
Loss (Equation 14)	$\alpha = 0.5$, $\beta = 1$, $\gamma = 0.1$, and $\delta = 0.1$
Optimizer	Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) at learning rate of $2e-4$ and batch size of 4

B. Evaluation of Outdoor Depth Estimation

Table III quantitatively compares the accuracy of monocular depth estimates on the outdoor KITTI data set with recent approaches and ours. The image resolution is commonly set as 416×128. The results have demonstrated that our method consistently produces the top two results over all metrics. This indicates that inertial measurements used in motion networks can indirectly improve depth estimates without involving loss computation. In addition, our visual-inertial fusion strategy can be effectively and favorably intergraded with other unsupervised learning systems for UGVs. Fig. 6 further depicts several example images for comparison. We observe that our method improves the smoothness of depth estimates regardless of illumination variations.

C. Evaluation of Outdoor Motion Estimation

We compare our method on KITTI data set with two representative SLAM frameworks, i.e., ORB-SLAM2 [35] and

TABLE III
QUANTITATIVE RESULTS ON MONOCULAR DEPTH ESTIMATION ON THE OUTDOOR KITTI DATA SET. THE COMMON IMAGE RESOLUTION IS SET AT 416×128 FOR BOTH OUR APPROACH AND THE STATE OF THE ART

Method	Error ↓				Accuracy ↑		
	AbsRel	SqRel	RMS	RMSlog	< 1.25	$< 1.25^2$	$< 1.25^3$
Zhou *et al.* [8]	0.208	1.768	6.856	0.283	0.678	0.885	0.957
vid2depth [2]	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Wang *et al.* [44]	0.151	1.257	5.583	0.228	**0.810**	0.936	0.974
GeoNet-VGG [31]	0.164	1.303	6.090	0.247	0.765	0.919	0.968
GeoNet-Resnet [31]	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Bian *et al.* [3]	0.149	**1.137**	5.771	0.230	0.799	0.932	0.973
Ours (videos(416×128) + IMU, direct)	0.151	1.170	5.692	0.227	0.795	0.932	0.974
Ours (videos(416×128) + IMU, selective)	**0.146**	1.148	**5.527**	**0.220**	0.804	**0.937**	**0.976**
Ours (videos(832×256) + IMU, selective)	**0.138**	**1.092**	**5.414**	**0.215**	**0.825**	**0.940**	**0.976**

TABLE IV
QUANTITATIVE RESULTS ON TRACKING TRAJECTORIES IN THE KITTI ODOMETRY DATA SET

Method	Average translation errors t_{err} (%)		Average rotation errors r_{err} ($^\circ/100m$)		Absolute trajectory errors ATE (m)	
	Seq. 09	Seq. 10	Seq. 09	Seq. 10	Seq. 09	Seq. 10
ORB-SLAM	21.44	5.88	**0.87**	**0.71**	109.71	**11.80**
VINS	**3.16**	**2.85**	9.52	9.21	**26.99**	72.43
Zhou *et al.* 2017	9.15	15.27	15.87	34.29	326.57	125.22
Bian *et al.* 2019 (416×128)	9.50	13.47	6.03	7.72	93.41	42.86
Ours (direct fusion)	9.71	13.99	4.89	6.58	66.87	77.36
Ours (selective fusion)	**8.22**	**12.17**	**4.79**	**5.31**	**30.88**	**31.02**

VINS [23] (without loop closure detection). They both employ the optimization strategy (e.g., bundle adjustment [34]) to improve global accuracy. We also report results from two recent works [3], [8] with scale adjustments. Since the ground truth of sequences 11–20 in KITTI data set are not released, we can only use the first ten sequences for both training and validation, thus we follow Zhou *et al.* [8] to split the data set, i.e., training the network with $00-08$ trajectory sequences and testing with $09-10$ sequences. We also remove the first five frames (~ 0.5 s) since they always cause initialization failures in ORB-SLAM.

Table IV shows the average translation errors (t_{err}), average rotation errors (r_{err}), and absolute trajectory errors (ATEs) among all frames (not only five frames). Note that the results are slightly lower than the official report of KITTI which is based on stereo videos, and its unit of r_{err} is $^\circ/1$ m while ours is $^\circ/100$ m. The results indicate that our approach outperforms other unsupervised learning methods over all metrics, especially reducing the ATE to a much lower level. When compared with ORB-SLAM, we produce significantly better results on sequence 09 and slightly large errors on sequence 10. Thus, our visual-inertial fusion strategy improves the accuracy and robustness of the motion network. Some example trajectories from different methods are shown in Fig. 7 for comparison, and ours are very similar with the ground truth.

In addition, Table V further depicts the ATE results over five frames. Our method outperforms all other monocular depth estimation approaches, and remains comparable even to the state-of-the-art SLAM systems.

D. Ablation Study

Image Resolutions: The last two rows in Table III shows the accuracy of depth estimates over images with two resolutions. It indicates that our approach derives slightly lower errors and nearly the same accuracy at a larger 832×256 image resolution. Thus, our method remain the robustness even at small resolutions.

TABLE V
FIVE FRAMES ATE ON THE KITTI ODOMETRY SPLIT, WITH AVERAGE ATE AND STANDARD DEVIATION, BOTH IN METERS

Method	Seq. 09	Seq. 10
ORB-SLAM	0.014 ± 0.008	0.012 ± 0.011
VINS	$\mathbf{0.014 \pm 0.007}$	$\mathbf{0.011 \pm 0.009}$
Zhou *et al.*	0.021 ± 0.017	0.020 ± 0.015
MonoDepth2	0.017 ± 0.008	0.015 ± 0.010
Ours	$\mathbf{0.015 \pm 0.008}$	$\mathbf{0.013 \pm 0.008}$

TABLE VI
ALL FRAMES ATE WITH CORRUPTIONS ON VIDEO QUALITY

Videos	Original	Bright	Dim	None
Basic (Bian *et al.* [3])	93.41m	145.98m	158.59m	255.68m
Ours (direct)	66.87m	76.05m	84.27m	243.56m
Ours (selective)	**30.88m**	**33.55m**	**55.30m**	**136.01m**

TABLE VII
EFFECTIVENESS OF USING BILSTM FOR INERTIAL LEARNING

Method	Error ↓				Accuracy ↑
	AbsRel	SqRel	RMS	RMSlog	< 1.25
FC	0.151	1.171	5.625	0.225	0.800
CNN	0.155	1.218	5.731	0.227	0.795
LSTM	0.151	**1.148**	5.536	0.223	0.800
BiLSTM	**0.146**	**1.148**	**5.527**	**0.220**	**0.804**

Corrupting Video Quality: In order to evaluate our robustness on video quality, we manually generate three categories of videos from the original KITTI data set, with bright videos (contrast $-$ 50%, brightness $+$ 100), dim videos (contrast $-$ 50%, brightness $-$ 30), and all black videos (shown in Fig. 8). Table VI shows the ATEs over different fusion strategies. Compared with the basic motion network [3], our visual-inertial fusion approach is effective to produce accurate

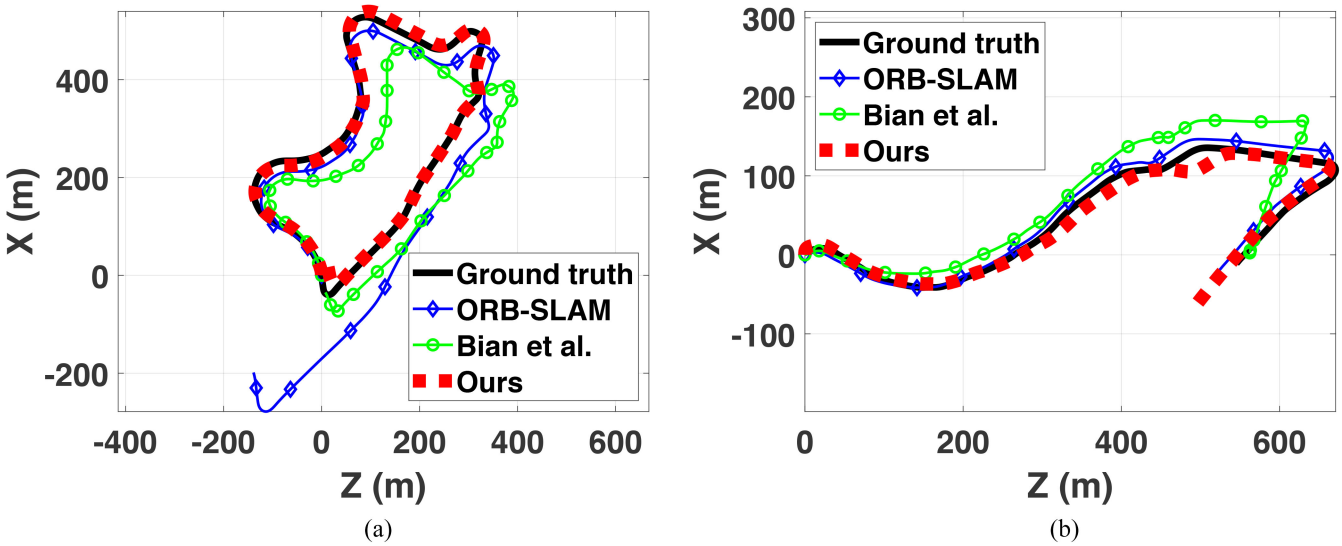

Fig. 7. Trajectory estimates on the KITTI odometry data set. (a) Seq. 09. (b) Seq. 10.

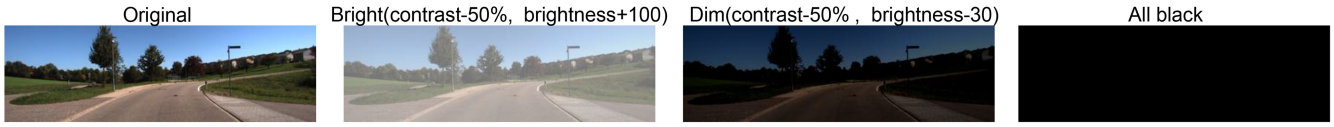

Fig. 8. Corruption examples on video quality.

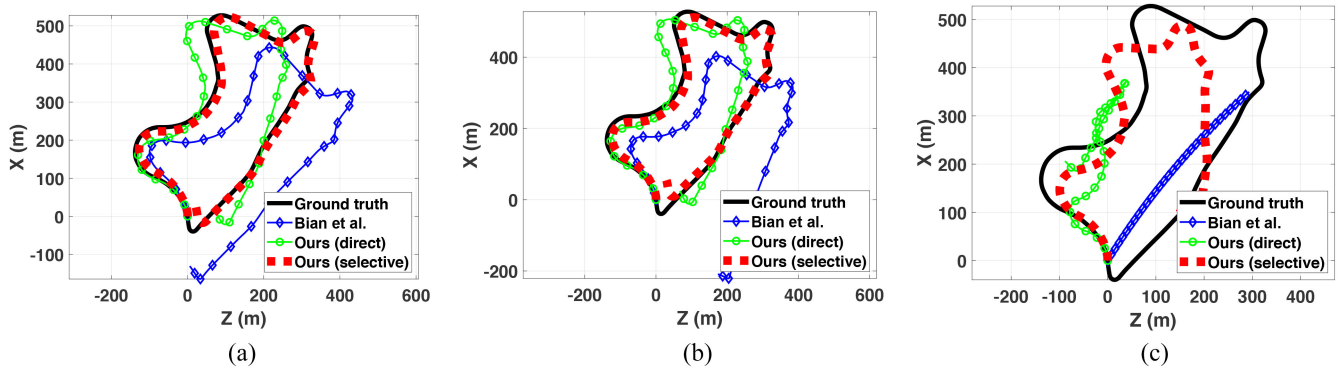

Fig. 9. Trajectory estimates of Seq. 09 with corruptions on video quality. (a) Bright video. (b) Dim video. (c) No video.

TABLE VIII
QUANTITATIVE RESULTS ON INDOOR MONOCULAR DEPTH ESTIMATION, COLLECTED BY OUR DEDICATED PROTOTYPE IN A CAMPUS BUILDING

Method	Error ↓				Accuracy ↑		
	AbsRel	SqRel	RMS	RMSlog	< 1.25	$< 1.25^2$	$< 1.25^3$
Zhou *et al.* [8]	0.296	1.047	2.788	0.381	0.506	0.782	0.910
Bian *et al.* [3]	0.416	4.818	3.453	0.422	0.524	0.795	0.915
Ours without IMU data	0.305	1.236	2.691	0.362	0.551	0.809	0.914
Ours without structure cues	**0.287**	**0.836**	**2.411**	0.365	0.510	0.788	0.920
Ours	0.288	1.065	2.579	**0.350**	**0.557**	**0.820**	**0.922**

and useable trajectories in all cases, even without video inputs (all black images). Fig. 9 further depicts qualitative results on sequence 09.

Motion Network: Table VII depicts the effectiveness of using Bi-LSTM to extract motion features from IMU data. Compared with FC, CNN, and LSTM for inertial learning, the BiLSTM achieves the least errors and best accuracy on monocular depth estimation.

Reweighting Network: We have evaluated the effectiveness of our reweighting network in Tables III, IV, and VI (the last two rows in each table). The tag "direct fusion" denotes the fusion scheme without reweighting network, i.e., concatenating visual and inertial features and fusing them with FC networks. From the three tables, we observe that our reweighting network (with tag "selective fusion") improves the accuracy of both monocular depth estimation and trajectory

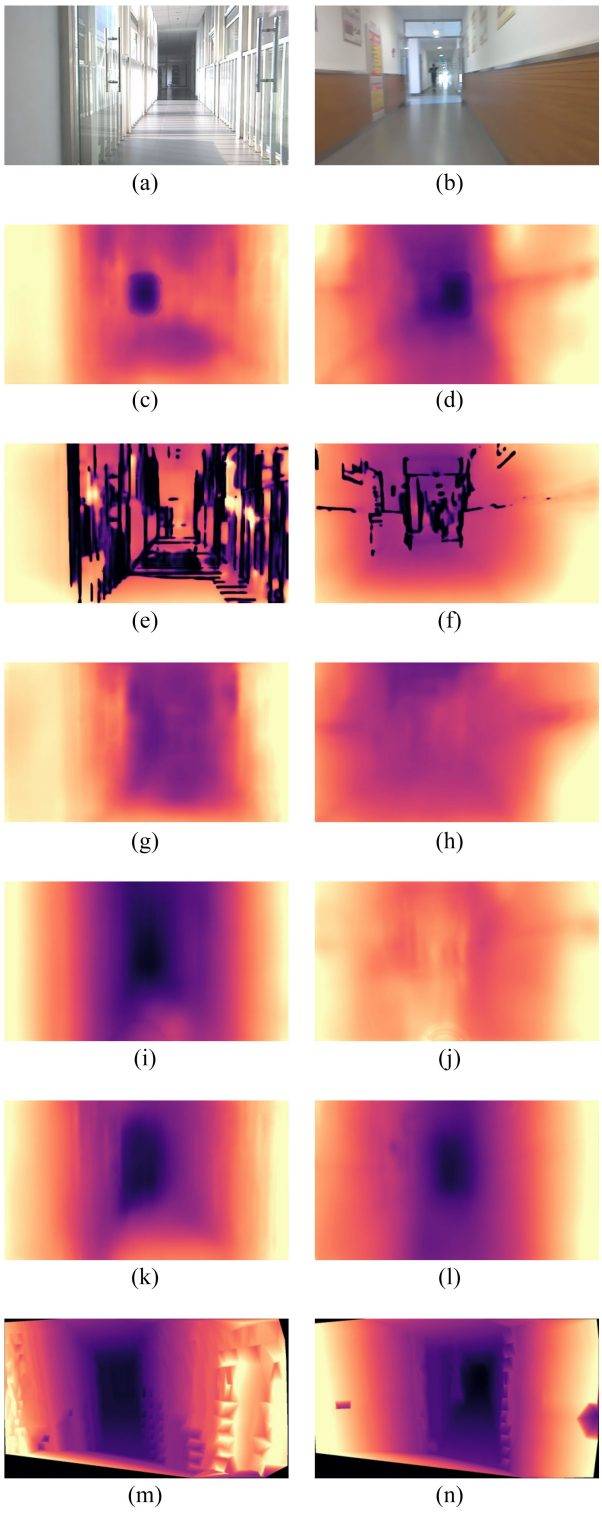

Fig. 10. Examples of monocular depth estimation in an indoor building. The first column is for a long corridor with glass walls and the second column is with blank walls. (a) Original image #1. (b) Original image #2. (c) Depth map from Zhou *et al.* [8]. (d) Depth map from Zhou *et al.* [8]. (e) Depth map from Bian *et al.* [3]. (f) Depth map from Bian *et al.* [3]. (g) Ours without structure cues. (h) Ours without structure cues. (i) Ours without IMU. (j) Ours without IMU. (k) Ours. (l) Ours. (m) Ground truth by laser radar. (n) Ground truth by laser radar.

tracking, and it also enhances the environmental robustness in extreme illumination conditions. The reason is that although inertial data is not involved in calculating loss functions, it can calibrates the warped image thus improves depth estimates through the unsupervised learning framework.

E. Evaluation of Indoor Depth Estimation

Since current public indoor UGV data sets cannot satisfy our requirements (e.g., gathering both monocular images and inertial readings in large-scale indoor buildings), we develop our dedicated prototype and collect data in a 80 m$\times$50 m campus building. The ground truth is recorded via a laser radar to measure surrounding point clouds. The image resolution is set as 480×288 and we adopt a commodity IMU (Wit-motion JY61P at \$10.0) to collect inertial readings at 100 Hz.

Table VIII demonstrates the accuracy of indoor monocular depth estimation. Compared with two recent approaches ([8] and [3]) designed for outdoor monocular depth learning, our method produces slightly lower errors and much higher accuracy. This is because the structure cues calibrate pixel depth estimates with scene constrains, thus eliminate outliers with extreme errors and raise the accuracy. In addition, although our IMU is cheap with low-quality measurements, it is still effective to reduce all types of errors.

In addition, Fig. 10 illustrate two example images on a long corridor, one with glass walls and the other with blank walls. We observe that the depth map by Zhou *et al.* [8] is easily disarranged by illumination variations from sunlight or lamplight, and there are obvious black stripes on the depth map by Bian *et al.* [3] due to the depth uncertainty on areas with similar appearances. In comparison, ours match the ground truth well without sharp depth mutations on floors and walls.

VIII. Conclusion

In this article, we proposed a novel unsupervised learning framework to infer monocular depth and ego-motion from a visual-inertial fusion perspective. It enables environment-agnostic depth estimates and scale-consistent motion trajectories by selectively combining different sensing modalities within a reweighting network. We also explored the principle and effectiveness of structure cues to calibrate depth estimates in indoor buildings. Extensive experiments on the KITTI data set and our dedicated prototype demonstrate our effectiveness and robustness compared with the state of the art. To the best of our knowledge, this was the first approach to fuse visual and inertial data for monocular depth and ego-motion estimation in an unsupervised manner, and remain suitable for indoor environments.

References

[1] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 3828–3838.

[2] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints," in *Proc. IEEE CVPR*, 2018, pp. 5667–5675.

[3] J. Bian *et al.*, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," in *Proc. NeurIPS*, 2019, p. 4.

[4] A. Cipolletta *et al.*, "Energy-quality scalable monocular depth estimation on low-power CPUs," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 25–36, Jan. 2022.

[5] J. Li *et al.*, "Monocular 3D object detection based on depth guided local convolution for smart payment in D2D systems," *IEEE Internet Things J.*, early access, Nov. 16, 2021, doi: 10.1109/JIOT.2021.3128440.

[6] "Pokémon Go." 2021. [Online]. Available: https://www.pokemon.com/us/app/pokemon-go/

[7] S. Ullman, "The interpretation of structure from motion," *Proc. Royal Soc. London B. Biol. Sci.*, vol. 203, no. 1153, pp. 405–426, 1979.

[8] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE CVPR*, 2017, pp. 1851–1858.

[9] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proc. IEEE CVPR*, 2018, pp. 340–349.

[10] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2366–2374.

[11] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

[12] K. Sartipi, T. Do, T. Ke, K. Vuong, and S. I. Roumeliotis, "Deep depth estimation from visual-inertial SLAM," in *Proc. IEEE IROS*, 2020, pp. 10038–10045.

[13] Y. Wu, S. Ying, and L. Zheng, "Size-to-depth: A new perspective for single image depth estimation," 2018, *arXiv:1801.04461*.

[14] W. Chen, Z. Fu, D. Yang, and J. Deng, "Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 1288–1296.

[15] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proc. IEEE CVPR*, 2018, pp. 2800–2810.

[16] N. Mayer *et al.*, "What makes good synthetic training data for learning disparity and optical flow estimation?" *Int. J. Comput. Vis.*, vol. 126, pp. 942–960, Apr. 2018.

[17] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from Internet photos," in *Proc. IEEE CVPR*, 2018, pp. 2041–2050.

[18] M. Klodt and A. Vedaldi, "Supervising the new with the old: Learning SFM from SFM," in *Proc. ECCV*, 2018, pp. 1–16.

[19] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. ECCV*, 2016, pp. 1–16.

[20] P. Wei, G. Hua, W. Huang, F. Meng, and H. Liu, "Unsupervised monocular visual-inertial odometry network," in *Proc. IJCAI*, 2020, pp. 2347–2354.

[21] H. Zhan, C. S. Weerasekera, J. Bian, and I. Reid, "Visual odometry revisited: What should be learnt?" 2019, *arXiv:1909.09803*.

[22] A. Z. Zhu, N. Atanasov, and K. Daniilidis, "Event-based visual inertial odometry," in *Proc. IEEE CVPR*, 2017, pp. 5391–5399.

[23] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[24] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2015, pp. 298–304.

[25] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[26] A. Wong, X. Fei, S. Tsuei, and S. Soatto, "Unsupervised depth completion from visual inertial odometry," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1899–1906, Apr. 2020.

[27] C. Chen *et al.*, "Selective sensor fusion for neural visual-inertial odometry," in *Proc. IEEE CVPR*, 2019, pp. 10534–10543.

[28] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "VINet: Visual-inertial odometry as a sequence-to-sequence learning problem," *arXiv:1701.08376*, 2017.

[29] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE CVPR*, 2016, pp. 4040–4048.

[30] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of LSTM and BiLSTM in forecasting time series," in *Proc. IEEE Int. Conf. Big Data*, 2019, pp. 3285–3292.

[31] Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE CVPR*, 2018, pp. 1983–1992.

[32] H. W. Haussecker and D. J. Fleet, "Computing optical flow with physical models of brightness variation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 661–673, Jun. 2001.

[33] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE CVPR*, 2017, pp. 270–279.

[34] C. Engels, H. Stewénius, and D. Nistér, "Bundle adjustment rules," in *Photogrammetric Comput. Vision*, vol. 2, no. 32, pp. 1–6, 2006.

[35] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[36] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[37] C. Rother, "A new approach for vanishing point detection in architectural environments," in *Proc. BMVC*, 2000, pp. 382–391.

[38] D. C. Lee, M. Hebert, and T. Kanade, "Geometric reasoning for single image structure recovery," in *Proc. IEEE CVPR*, 2009, pp. 2136–2143.

[39] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[40] P. K. N. Silberman, D. Hoiem, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. ECCV*, 2012, pp. 746–760.

[41] M. Burri *et al.*, "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, 2016.

[42] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[44] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *Proc. IEEE CVPR*, 2018, pp. 2022–2030.

Ruipeng Gao received the B.S. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2010, and the Ph.D. degree from the Peking University, Beijing, in 2016.

He was a Visiting Scholar with Purdue University, West Lafayette, IN, USA, in 2019. He is currently an Associated Professor with the School of Software Engineering, Beijing Jiaotong University, Beijing. His research interests include mobile computing and applications, Internet of Things, and intelligent transportation systems.

Xuan Xiao received the B.S. degree in network engineering from China University of Mining and Technology, Xuzhou, China, in 2016. He is currently pursuing the Ph.D. degree in software engineering from Beijing Jiaotong University, Beijing, China.

His research interests include mobile computing and applications, and Internet of Things.

Weiwei Xing (Member, IEEE) received the B.S. degree in computer science and technology and the Ph.D. degree in signal and information processing from Beijing Jiaotong University, Beijing, China, in 2001 and 2006, respectively.

She was a Visiting Scholar with the University of Pennsylvania, Philadelphia, PA, USA, from 2011 to 2012. She is currently a Professor with the School of Software Engineering, Beijing Jiaotong University. Her research interests include computer vision, pattern recognition, and intelligent transportation algorithms and applications.

Chi Li received the B.S. degree from Nanchang Hangkong University, Nanchang, China, in 2014, and the M.S. degree from Beihang University, Beijing, China, in 2017.

He is currently a Location Algorithm Engineer with DiDi Company, Beijing. His research interests include inertial navigation, integrated navigation, and machine learning.

Lei Liu received the B.S. degree from Shandong University, Jinan, China, in 2010, and the M.S. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2013.

He is currently an Expert Algorithm Engineer with DiDi Company, Beijing, in charge of positioning and POI recommendation.