Supplementary Document for ClusterSLAM: A SLAM Backend for Simultaneous Rigid Body Clustering and Motion Estimation

Jiahui Huang[1] Sheng Yang[1,2] Zishuo Zhao[1] Yu-Kun Lai[3] Shi-Min Hu[1*]
[1]BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing
[2]Alibaba A.I. Labs, China [3]Cardiff University, UK
huang-jh18@mails.tsinghua.edu.cn, shengyang93fs@gmail.com, wingedkuriboh@126.com
LaiY4@cardiff.ac.uk, shimin@tsinghua.edu.cn

In this supplementary document, we first provide derivations to support Equations 1 - 3 in the main paper (Sec. 1), and then illustrate the detailed procedure of the Iterative Voting Consensus [1] we apply in Sec. 3.1.3 of the main paper (Sec. 2), as well as the sparsity patterns of the Hessian Matrix when using different solving strategies (Sec. 3). Finally, we provide sample frames of our dataset (Sec. 4) with our results for additional visual demonstrations on both the synthetic (Sec. 5) and the KITTI (Sec. 6) datasets.

1. Derivation of Equations 1 - 3

In this section we present the derivation of equations in the text. For simplicity, we denote the probabilistic density function of multivariate normal distribution [2] as:

$$\mathcal{N}(x;\mu,\Sigma) \triangleq \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp(-\frac{1}{2}\left\|x-\mu\right\|^2_\Sigma). \tag{1}$$

The definitions of notations above remain the same as in the main paper.

1.1. Equ. 1: Noise-aware distance term d^{ij}

We start by defining the affinity probability of landmark i and landmark j as the product of 3D geometric probability $\mathcal{P}_{3D}$ and vision based prior probability $\mathcal{P}_{2D}$ with the balance factor 2α:

$$\mathcal{P}^{ij} = \mathcal{P}_{3D}(\mathcal{P}_{2D})^{2\alpha}. \tag{2}$$

For $\mathcal{P}_{3D}$, we assume that the distance between the two endpoints should stay unchanged in different frames, *i.e.*, $l_t^{ij} = l_{t'}^{ij}, \forall t, t'$, and utilize MLE (Maximum Likelihood Estimate) to estimate such an optimal length l_*^{ij} considering the uncertainty of each l_t^{ij}:

$$\begin{aligned} l_*^{ij} &= \operatorname*{argmax}_l \exp(-\frac{1}{2}\sum_t \|l_t^{ij}-l\|^2_{\sigma_t^{ij}}) \\ &= \operatorname*{argmin}_l \sum_t \|l_t^{ij}-l\|^2_{\sigma_t^{ij}} \\ &= \sum_t (\frac{1}{\sigma_t^{ij}}\cdot l_t^{ij}) \Big/ \sum_t \frac{1}{\sigma_t^{ij}}. \end{aligned} \tag{3}$$

If the deviations of the observed lengths w.r.t. the optimal length conform to their noise distribution, the likelihood should be large and vice versa. Hence, the 3D geometric probability is summarized through the Maximum-a-Posteriori (MAP) over its all observations from these covisible frames, whose the total number is denoted as ψ^{ij}:

$$\mathcal{P}_{3D} = \prod_t \mathcal{N}(l_t^{ij}; l_*^{ij}, \sigma_t^{ij})^{\frac{1}{\psi^{ij}}}. \tag{4}$$

For the prior probability $\mathcal{P}_{2D}$, we define it as follows:

$$\mathcal{P}_{2D} = C_{2D} \exp\left(-\frac{1}{2}\max_t \|\mathbf{x}_t^i - \mathbf{x}_t^j\|^2_{\Sigma_t^{ij}}\right), \tag{5}$$

where C_{2D} is the normalizing constant for the probability; Σ_t^{ij} is the uncertainty of landmark distance in the image space and we assume this uncertainty stays unchanged in different frame t.

The noise-aware distance term d^{ij} is therefore taken as the negative logarithm of the affinity probability $\mathcal{P}^{ij}$ to avoid numerical underflow:

$$\begin{aligned} d^{ij} &= -\log \mathcal{P}^{ij} \\ &= -\log \mathcal{P}_{3D} - 2\alpha \log \mathcal{P}_{2D} \\ &= \frac{1}{2}\operatorname*{avg}_t \left(\left\|l_t^{ij} - l_*^{ij}\right\|^2_{\sigma_t^{ij}} + \log \sigma_t^{ij}\right) + \\ &\quad \alpha \max_t \left\|\mathbf{x}_t^i - \mathbf{x}_t^j\right\|^2_{\Sigma_t^{ij}} + \underbrace{\frac{1}{2}\log 2\pi + 2\alpha C_{2D}}_{\text{constant}, \forall i,j}. \end{aligned} \tag{6}$$

*corresponding author.

We ignore the trailing common constant in d^{ij} since it will not affect the result of the clustering algorithm which relies only on the relative ordering of values.

1.2. Equ. 2: Approximated variance σ_t^{ij}

The variance of length σ_t^{ij} is approximated by error propagation theorem [3], as the variance of $h(\mathbf{x})$ can be deducted by the variance of $\mathbf{x}$ denoted as Σ_x through a first-order approximation:

$$\Sigma_{h(\mathbf{x})} \approx \mathbf{J}_h \Sigma_{\mathbf{x}} \mathbf{J}_h^\top, \tag{7}$$

with $\mathbf{J}_h$ being the Jacobian of h w.r.t. $\mathbf{x}$. Such an approximation is the basis for both the stereo back-projection uncertainty and Equ. 2 of the main paper.

In Equ. 2 of the main paper, we define function h as the Euclidean distance function:

$$h\left(\begin{bmatrix}\mathbf{X}_t^{\times,i}\\ \mathbf{X}_t^{\times,j}\end{bmatrix}\right) = \left\|\mathbf{X}_t^{\times,i} - \mathbf{X}_t^{\times,j}\right\| = l_t^{ij}, \tag{8}$$

and its Jacobian can be computed as:

$$\mathbf{J}_h = \frac{1}{l_t^{ij}} \begin{bmatrix}\mathbf{X}_t^{\times,i} - \mathbf{X}_t^{\times,j}\\ \mathbf{X}_t^{\times,j} - \mathbf{X}_t^{\times,i}\end{bmatrix}^\top. \tag{9}$$

The covariance of argument $\mathbf{x}$ is:

$$\Sigma_{\mathbf{x}} = \begin{bmatrix}\Sigma_t^{\times,i} & \mathbf{0}\\ \mathbf{0} & \Sigma_t^{\times,j}.\end{bmatrix} \tag{10}$$

We can then obtain $\sigma_t^{ij} = \Sigma_h$ by substituting Equ. 9 and Equ. 10 into Equ. 7.

1.3. Equ. 3: Transformation estimation

The noise-aware pose estimation method assume the frame-to-model transformation $\mathbf{T}_t^{\mathbf{qc}}$ will perfectly align $\mathbf{X}_t^{\mathbf{c},i}$ and $\hat{\mathbf{X}}_{t-1}^{\mathbf{q},i}$ with tolerable noise Σ_g^i:

$$\begin{aligned}\mathbf{T}_t^{\mathbf{qc}} &= \underset{\mathbf{T}}{\operatorname{argmax}} \prod_i \sum_{g\in\mathcal{G}_{t-1}^{\mathbf{q},i}} \mathcal{N}\left(\mathbf{T}\mathbf{X}_t^{\mathbf{c},i} - \hat{\mathbf{X}}_{t-1}^{\mathbf{q},i}; \mathbf{0}, \Sigma_g^i\right)\\ &= \underset{\mathbf{T}}{\operatorname{argmax}} \prod_i \max_{g\in\mathcal{G}_{t-1}^{\mathbf{q},i}} \mathcal{N}\left(\mathbf{T}\mathbf{X}_t^{\mathbf{c},i} - \hat{\mathbf{X}}_{t-1}^{\mathbf{q},i}; \mathbf{0}, \Sigma_g^i\right)\\ &= \underset{\mathbf{T}}{\operatorname{argmin}} \sum_i \min_{g\in\mathcal{G}_{t-1}^{\mathbf{q},i}} \left(\frac{1}{2}\left\|\mathbf{T}\mathbf{X}_t^{\mathbf{c},i} - \hat{\mathbf{X}}_{t-1}^{\mathbf{q},i}\right\|_{\Sigma_g^i}^2 - \mathbf{C}_g^i\right)\end{aligned} \tag{11}$$

The covariance Σ_g^i is derived from the uncertainty of observation Σ_t^i and the fused uncertainty of registered model $\mathcal{G}_{t-1}^{\mathbf{q},i}$ using error propagation theorem of Gaussian distribution and Gaussian mixture distribution [3].

2. Detailed Algorithm of Consensus Clustering

We detail how the Iterative Voting Consensus algorithm is applied to our method in Alg. 1. For more analysis on the performance of this algorithm we refer readers to [1]. The desired number of the consensus clustering K is selected as the sum of cluster numbers in all chunks for outdoor cases and the maximum of cluster numbers among all chunks for indoor cases. θ is initialized randomly, the probability of cluster $\mathbf{q} = \mathbf{0}$ (the cluster with static landmarks) is set to 80% while the other clusters are chosen uniformly. Clusters with too few landmarks (≤ 2) are then pruned so the total number of clusters is controllable.

Algorithm 1 Iterative Voting Consensus

Input: a set of landmarks $\bigcup_i \mathbf{X}^i$ to be classified, a set of clusters $\bigcup_m \theta_m$ obtained from all the chunks and the desired number of the consensus clustering K.

Output: Consensus clustering θ with K clusters.

Initialize θ as described in text;

repeat

Let $\mathcal{X}_{\mathbf{q}} = \{i|\theta(i) = \mathbf{q}\}$ be the $\mathbf{q}$-th cluster;

Compute the representation center for each cluster: $y_{\mathcal{X}_{\mathbf{q}}} = \{majority\{(\mathcal{X}_{\mathbf{q}})_1\}, \ldots, majority\{(\mathcal{X}_{\mathbf{q}})_m\}\}$, where $\{(\mathcal{X}_{\mathbf{q}})_m\}$ is the set of clustering results of $\mathcal{X}_{\mathbf{q}}$ by θ_m;

for all landmark i **do**

Re-assign $\theta(i) \leftarrow \operatorname{argmin}_{\mathbf{q}} D(y^i, y_{\mathcal{X}_{\mathbf{q}}})$, where $D(y^i, y_{\mathcal{X}_{\mathbf{q}}})$ is the Hamming distance between vector y^i and $y_{\mathcal{X}_{\mathbf{q}}}$, where only valid values in the sparse vector y^i are counted in the distance;

until θ does not change.

3. Hessian Matrix Sparsity Pattern

We plot the sparsity pattern of Hessian matrices of *decoupled* and *fully-decoupled* optimization methods in Figure 1. We define N as the number of landmarks (# Landmarks), T as the number of frames (# Frames) and Q as the number of clusters (# Clusters). *Decoupled* optimization method essentially solves Q sub-problems and the size for each of them is $(N/Q + T)^2$ while *fully-coupled* optimization method solves one problem with size $(N+TQ)^2$ which is much larger. However, the above analysis does not reflect the sparse nature of graph-based optimization and may not precisely characterize experimental results. For empirical study of the comparison, please refer to Sec. 4.3-F of the paper.

4. Dataset

The synthetic dataset used for evaluation is named and shown in Fig. 2. All indoor dataset has the prefix `SUNCG`

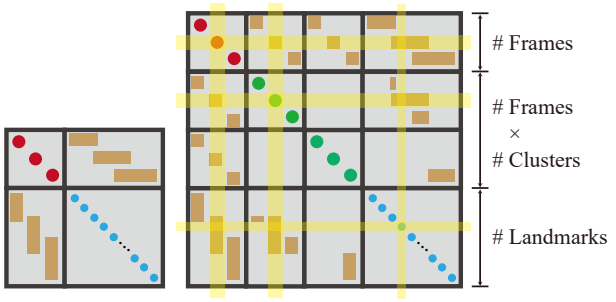

Figure 1. Comparison of sizes and sparsity patterns of Hessian matrix between *decoupled* and *fully-coupled*. Red: camera pose block. Green: cluster motion block. Blue: landmark position block. The Brown rectangles show a possible configuration of the sparsity pattern of observations. The intersection of the transparent yellow rectangles denotes the 9 Hessian blocks filled for one observation using Gauss-Newton method.

and the outdoor dataset has the common prefix `CARLA`. Each dataset is captured using an ideally-synchronized stereo camera and only images from the left camera are shown.

5. Results on Synthetic Dataset

Please refer to Figs. 3-4. As a result, we obtain better coincidence with the ground-truth and outperform other compared methods. Some trajectories from the ground-truth are longer than our prediction, which is due to insufficient observations of the corresponding clusters.

6. Results on KITTI dataset

Figs. 5-7 show results of our algorithm on KITTI sequences using batch input. We are showing the raw output without post-processing smoothing step mentioned in the paper.

References

[1] Nam Nguyen and Rich Caruana. Consensus clusterings. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 607–612. IEEE, 2007. 1, 2

[2] Wikipedia contributors. Multivariate normal distribution — Wikipedia, the free encyclopedia, 2019. [Online; accessed 23-March-2019]. 1

[3] Wikipedia contributors. Propagation of uncertainty — Wikipedia, the free encyclopedia, 2019. [Online; accessed 23-March-2019]. 2

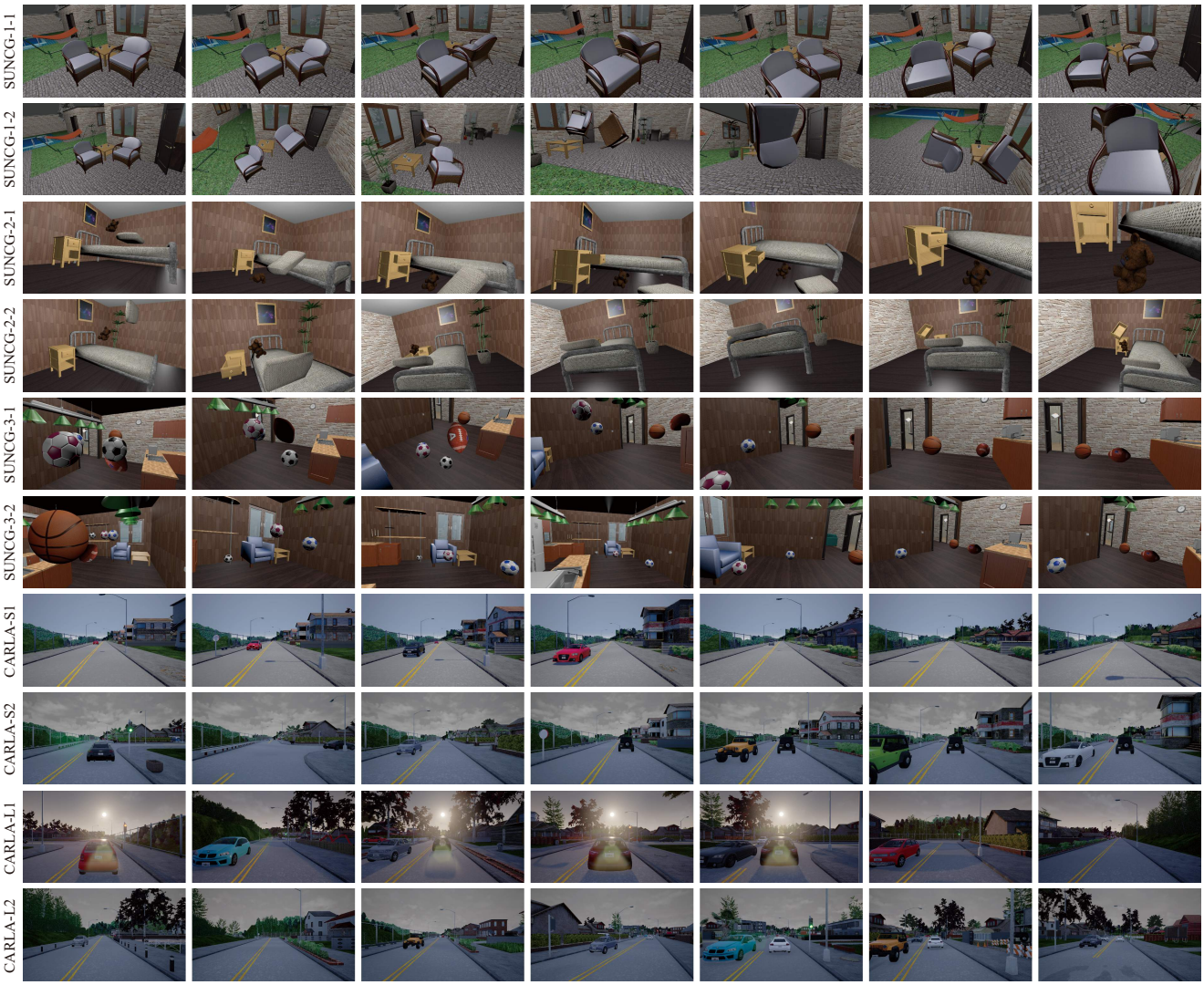

Figure 2. Sampled frames from the synthetic dataset used in our experiments.

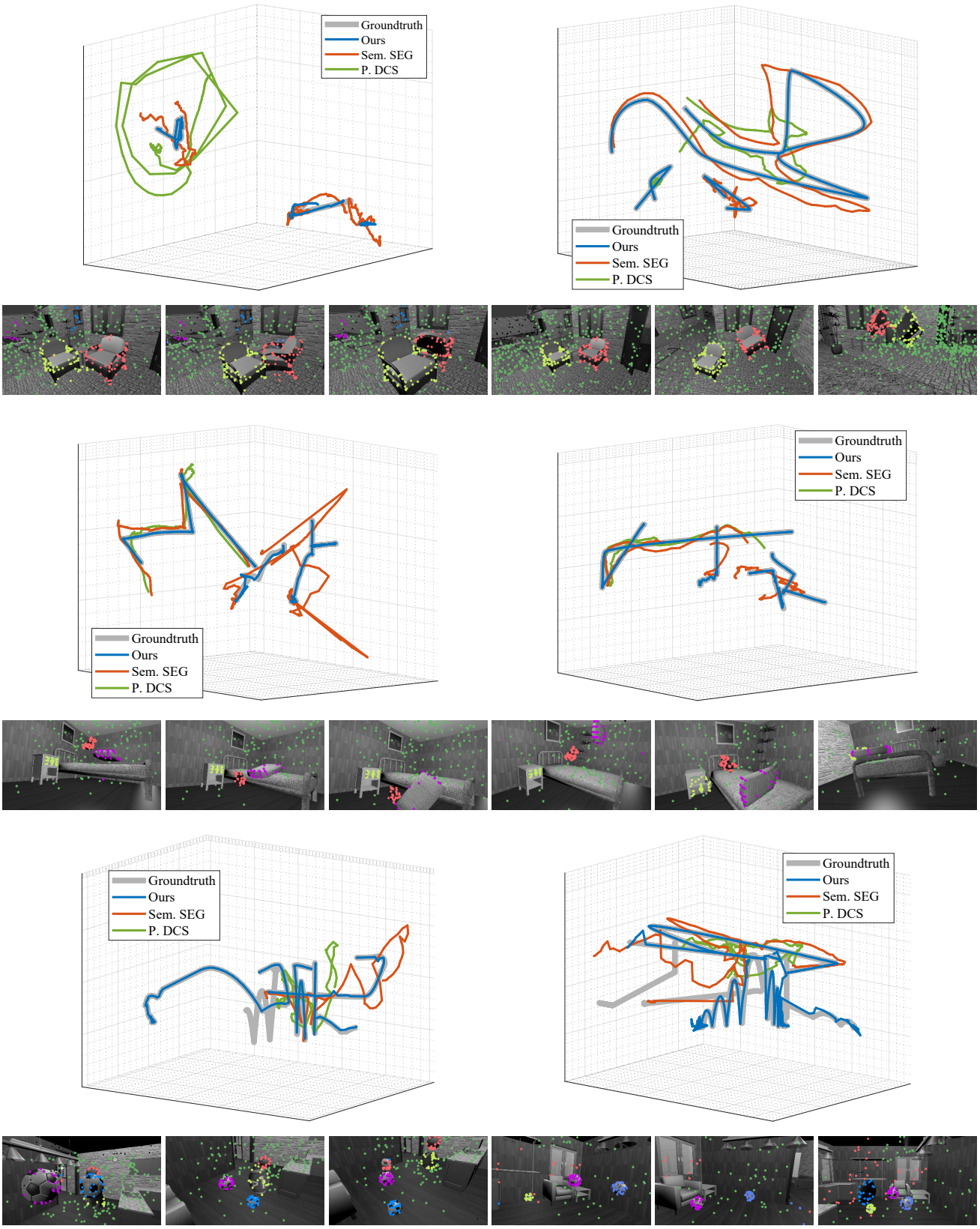

Figure 3. Results on SUNCG Indoor dataset. Corresponding datasets (enumerated in row-major order) are `SUNCG-1-1`, `SUNCG-1-2`, `SUNCG-2-1`, `SUNCG-2-2`, `SUNCG-3-1`, `SUNCG-3-2`, respectively. The three pictures below each trajectory plot show sampled frames with landmark classification overlayed.

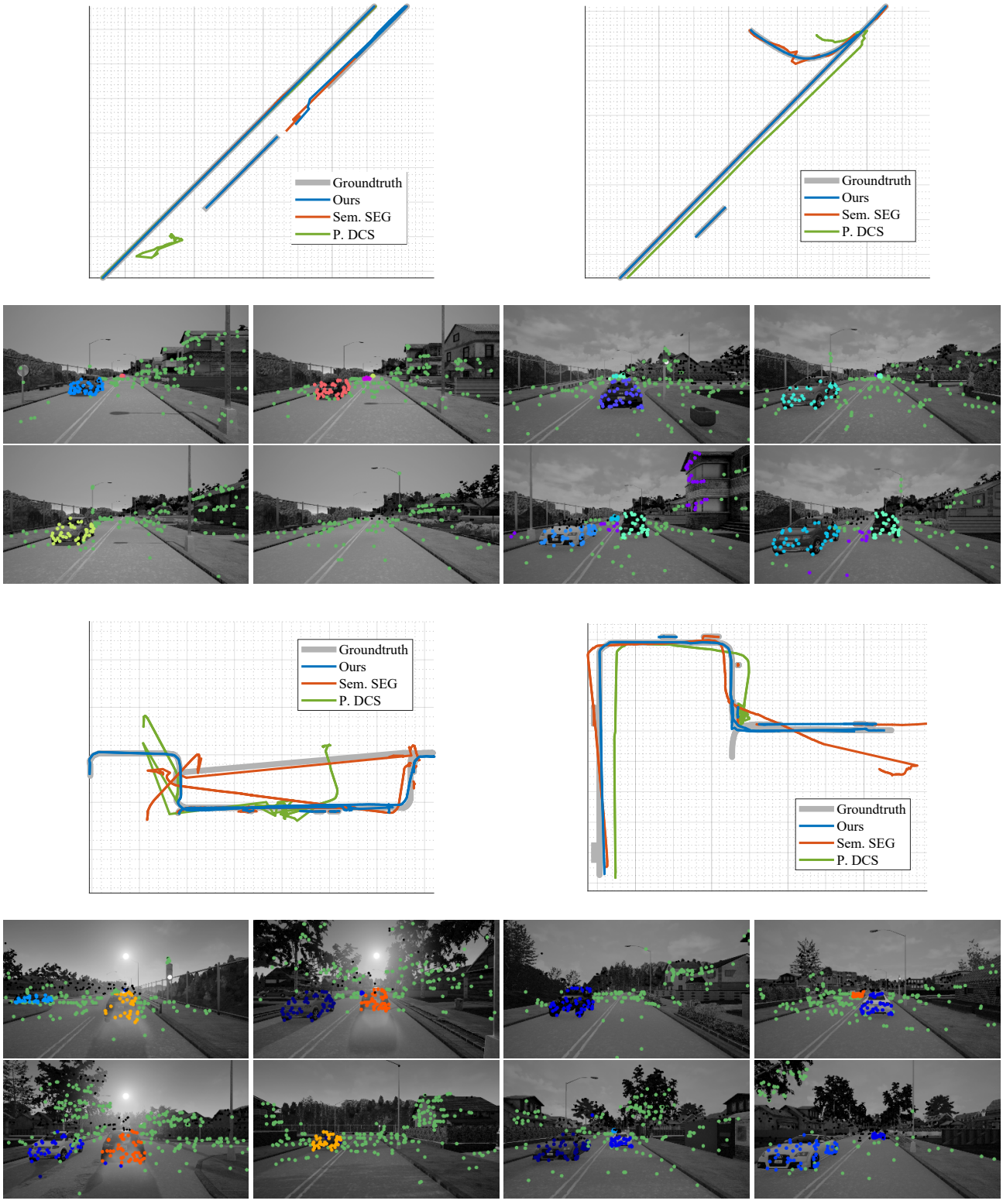

Figure 4. Results on CARLA Outdoor dataset. Corresponding datasets (enumerated in row-major order) are `CARLA-S1`, `CARLA-S2`, `CARLA-L1`, `CARLA-L2`, respectively. The four pictures below each trajectory plot show sampled frames with landmark classification overlayed.

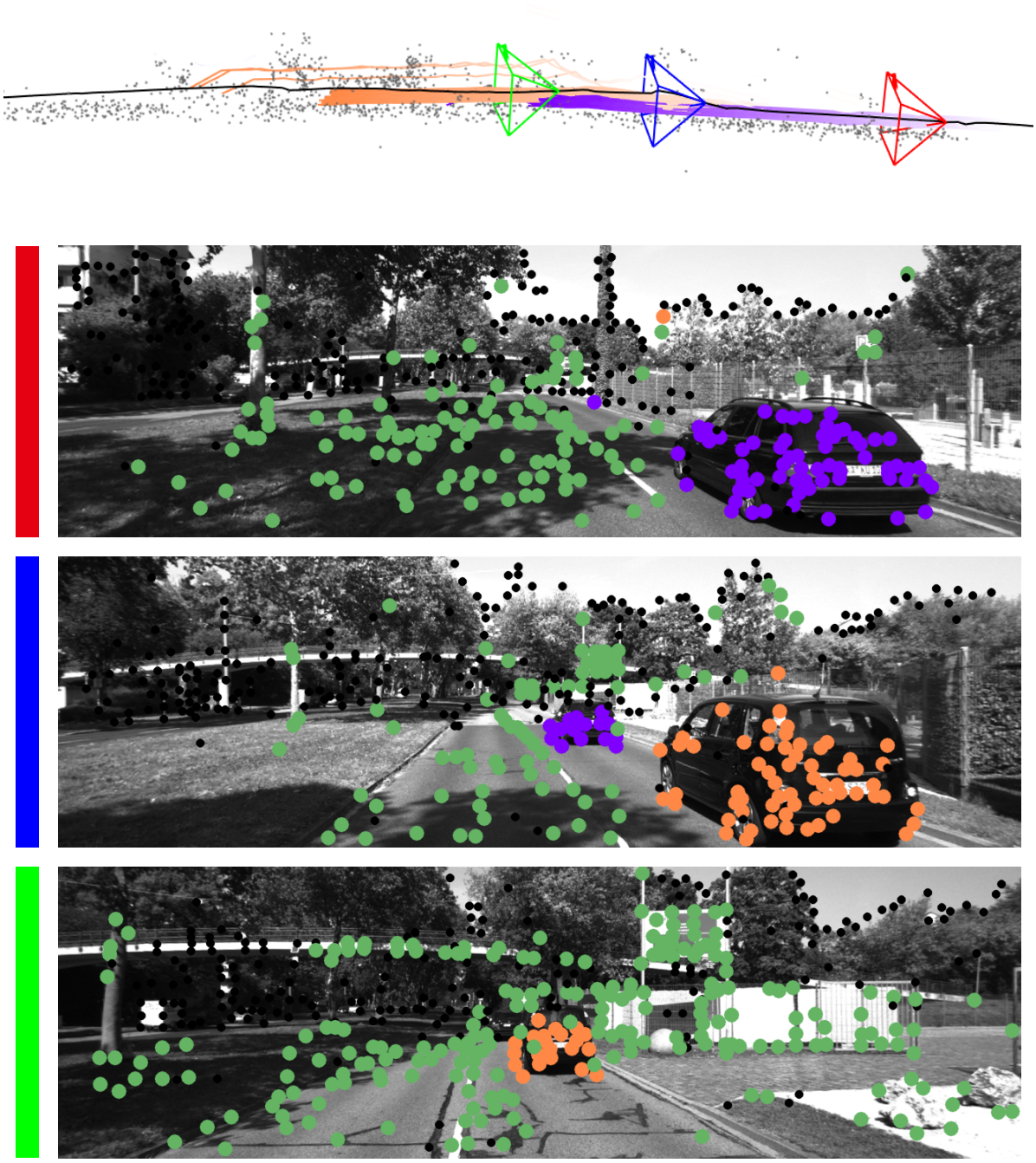

Figure 5. Results on KITTI 0013 sequence. The color of the bar on the left of each picture corresponds to the color of camera in the trajectory.

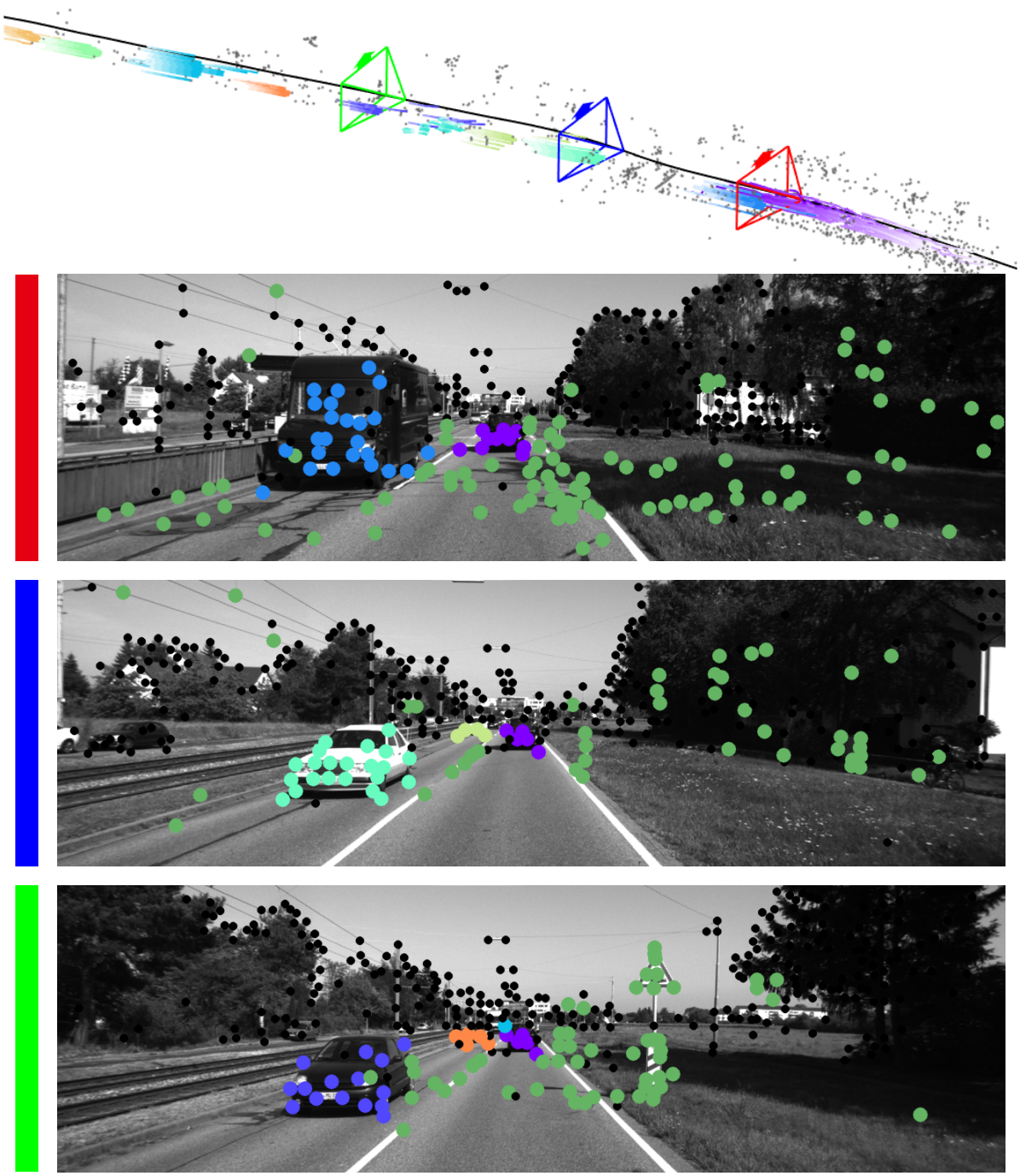

Figure 6. Results on KITTI 0015 sequence. The color of the bar on the left of each picture corresponds to the color of camera in the trajectory.

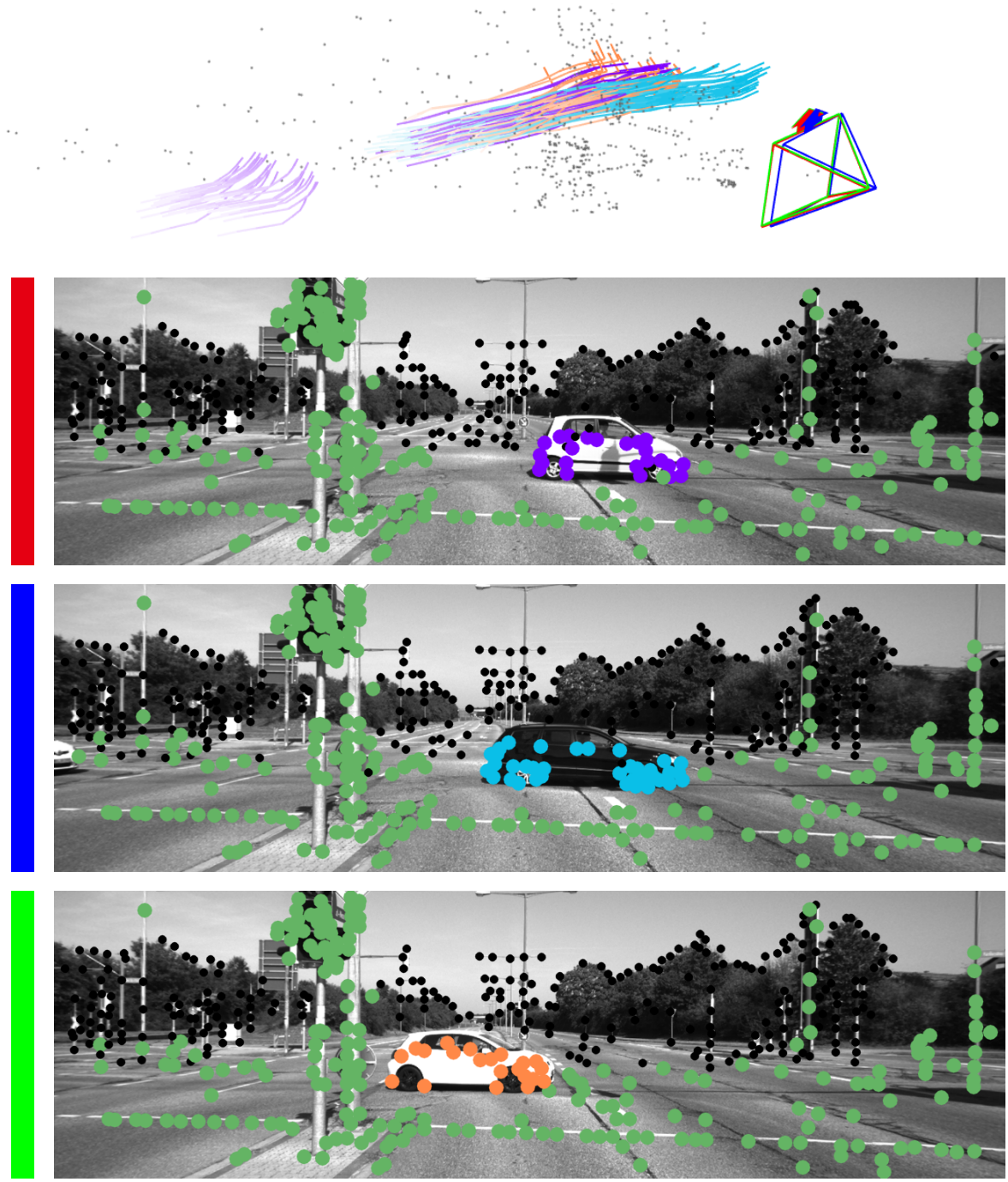

Figure 7. Results on KITTI 0017 sequence. The color of the bar on the left of each picture corresponds to the color of camera in the trajectory.