2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

A Unified Framework for Piecewise Semantic Reconstruction in Dynamic Scenes via Exploiting Superpixel Relations

Yan Di, Henrique Morimitsu, Zhiqiang Lou and Xiangyang Ji

***Abstract*— This paper presents a novel framework for dense piecewise semantic reconstruction in dynamic scenes containing complex background and moving objects via exploiting superpixel relations. We utilize two kinds of superpixel relations: motion relations and spatial relations, each having three subcategories: *coplanar, hinge*, and *crack*. Spatial relations provide constraints on the spatial locations of neighboring superpixels and thus can be used to reconstruct dynamic scenes. However, spatial relations can not be estimated directly with epipolar geometry due to moving objects in dynamic scenes. We synthesize the results of semantic instance segmentation and motion relations to estimate spatial relations. Given consecutive frames, we mainly develop our method in five main stages: preprocessing, motion estimation, superpixel relation analysis, reconstruction and refinement. Extensive experiments on various datasets demonstrate that our method outperforms competitors in reconstruction quality. Furthermore, our method presents a feasible way to incorporate semantic information in Structure-from-Motion (SFM) based reconstruction pipelines.**

I. Introduction

Dense monocular reconstruction in complex dynamic scenes has been a challenging task in computer vision for many years. The main difficulty lies in the relative scale ambiguity (RSA) problem [1] between moving objects and the background, which is intractable with traditional SFM techniques [2]. Since reconstruction serves as an important step in many high-level applications (e.g., autonomous driving [3], robot navigation [4], scene understanding [5]), recently various reconstruction methods based on SFM or deep learning techniques were proposed and they have achieved preferable results on many challenging datasets, especially on the famous KITTI dataset [6], [4]. However, they are still unable to handle different types of more complex scenes.

Among deep learning-based methods, we mainly focus on unsupervised learning methods [7], [8], [9] that require no ground truth depth, since depth information is hard to collect in complex scenes. Similar to our method, Struct2depth [9] also takes semantic segmentation results as input. However, currently, unsupervised learning methods only succeed in traffic scenes, their performance is still far from ideal in more challenging scenes (e.g., MPI Sintel). Furthermore, these methods usually need complicated fine-tuning before being applied in different scenes, which limits their practical use. On the other hand, deep learning-based methods are adopted as viable options to complement traditional SFM approaches. For example, VSO [10] shows that important structural information can be introduced by using the results of semantic segmentation.

For SFM-based methods [11], [1], [12], they propose strong hypotheses to solve the RSA problem. Specifically, DMDE [11] proposes the ordering constraint that assumes foreground objects occlude the surrounding environment. S.Soup [1] utilizes the ARAP term under the assumption that the transformation between two frames is locally piecewise-rigid and globally as rigid as possible. MPDE [12] assumes that motion relations between neighboring superpixels indicate their spatial relations, and these spatial relations are then used to solve the RSA problem. However, the above hypotheses have many drawbacks. The ordering constraint in DMDE requires accurate motion segmentation, which is intractable in real complex scenes. The ARAP term in S.Soup is hard to optimize and may disappoint in scenes with substantial motion. The superpixel relation hypothesis proposed in MPDE is also imperfect, since it needs other hypotheses like the ordering constraint in DMDE as supplements to ensure that all superpixels are fully constrained.

In this paper, we propose to develop the idea of MPDE [12] by exploiting superpixel relations and incorporating semantic information in the SFM-based pipeline to obtain superior performance. We follow the definitions of superpixel relations in MPDE. Motion relation is determined by homography and spatial relation is determined by plane parameters. After estimating motion relations between neighboring superpixels, MPDE directly assumes that motion relations and spatial relations correspond one-to-one. However, in this paper, besides motion relations, we introduce semantic segmentation to help predict spatial relations. Meanwhile, semantic information also provides important prior knowledge in motion estimation, reconstruction and refinement. We then design a unified framework that takes RGB image sequences and semantic segmentation maps as inputs and outputs high-quality depth maps up to a global scale.

Our main contributions are summarized as follows:

- A unified piecewise dense reconstruction framework that exploits superpixel relations by incorporating semantic information in an SFM-based pipeline.
- An effective pixel-level refinement approach by making use of RGB images, superpixel relations and semantic information.

II. Related works

A. SFM-based methods

The typical reconstruction methods based on SFM are DMDE [11], S.Soup [1], MPDE [12]. DMDE first segments

*This work was supported by the National Key R&D Program of China under Grant No. 2018AAA0102800 and No. 2018AAA0102801.

The authors are with Department of Automation & BNRist, Tsinghua University, China.

optical flow field into a set of rigid parts and then reconstructs the whole scene with an ordering constraint that assumes moving objects occlude the surrounding environment. S.Soup introduces the ARAP term to avoid motion segmentation and outperforms DMDE on many challenging datasets. MPDE defines two kinds of superpixel relations: motion relation and spatial relation. Motion relations are used to predict spatial relations between neighboring superpixels.

B. Learning-based methods

We mainly focus on unsupervised learning methods that require no ground truth depth, since depth data is hard to collect in real complex scenes. [7] is the first successful unsupervised learning method in traffic scenes, which combines a depth net and a pose net with view synthesis. Subsequent papers usually follow the basic structure of [7] and design more elaborate loss functions or incorporate other relevant tasks. For example, [8] explores optical flow together with ego-motion, while [13] adds motion segmentation to the pipeline. [14] applies geometric constraints, while [15] utilizes edge detection. Struct2depth [9] takes RGB image sequences and semantic segmentation results as inputs, which is similar to our method. Up to now, unsupervised learning methods achieve preferable performance in traffic scenes. However, in more complex scenes like MPI Sintel, they still perform poorly.

C. Semantic SLAM methods

PDA [16] presents a probabilistic data association method for semantic SLAM. VSO [10] incorporates semantic segmentation results into an SLAM pipeline to enable medium-term tracking. Some methods apply high-level features, including lines [17], planes [18] and objects [19] to improve the performance. They usually apply object detection to improve camera pose estimation [19] and loop closures [16], [20]. Different from above methods, we provide a new approach to incorporate semantic segmentation into the SFM-based pipeline by exploiting superpixel relations. We solve the RSA problem [2] and reconstruct moving objects and the background simultaneously.

III. Framework

We first define several key symbols. For input images I_t and I_{t+1}, we aim to estimate a depth map D_t for I_t. We first estimate initial pixel matches $M = \{p_i, p_j\}$, where $p_i \in I_t$ and $p_j \in I_{t+1}$. Then we over-segment image I_t into non-overlapping superpixels $S = \{S_1, ..., S_i, ...S_n\}$, where n is the number of superpixels. For a superpixel S_i, $H_i \in \mathbb{R}^{3\times3}$ denotes its homography model, $\theta_i \in \mathbb{R}^3$ denotes its plane parameter and s_i denotes its scale. Given neighboring superpixels S_i and S_j, $R_e(i,j)$ denotes their motion relation and $R_s(i,j)$ denotes their spatial relation, B_{ij} denotes pixels on the shared boundary. For pixel p_i, P_i is its homogeneous representation. In this paper, the terms λ_* denote the weight of each energy term, τ_* denote threshold values, and ω_* denote weights in each term.

TABLE I

Criteria of motion and spatial relations.

Motion relations	Criterion
Coplanar	$\sum_{p\in S_i\cup S_j} \lvert H_iP - H_jP\rvert \approx 0$
Hinge	$\sum_{p\in B_{ij}} \lvert H_iP - H_jP\rvert \approx 0$
Crack	Otherwise

Spatial relations	Criterion
Coplanar	$\sum_{p\in S_i\cup S_j} \lvert \theta_iP - \theta_jP\rvert \approx 0$
Hinge	$\sum_{p\in B_{ij}} \lvert \theta_iP - \theta_jP\rvert \approx 0$
Crack	Otherwise

A. Overall workflow

As demonstrated in Fig. 1, our framework takes consecutive RGB images as inputs. Then we develop our method in five main steps: preprocessing, motion estimation, superpixel relations analysis, reconstruction and refinement. In the preprocessing step, we compute initial matches with a PatchMatch [21] method and apply over-segmentation with SLIC [22]. We adopt deep learning networks for semantic segmentation. In the motion estimation step, we jointly estimate homographies and motion relations. In the superpixel relation analysis step, we combine semantic information and motion relations to predict spatial relations. In the reconstruction step, we estimate plane parameters and assign a relative scale for each superpixel. We apply pixel-level refinement in the last refinement step. Note that we directly refine plane parameters instead of the depth map under the piecewise planar assumption.

B. Motion Estimation and Superpixel relations

The concept of superpixel relations is successfully used in [23], [24] to smooth optical flow, and then MPDE develops the idea to solve the RSA problem and reconstruct dynamic scenes. We follow the definition used in MPDE, where superpixel relation is subdivided into two kinds: motion relation and spatial relation, each having three subcategories: *coplanar, hinge* and *crack*. For convenience, we use $\{co, hi, cr\}$ to denote the three kinds of relations respectively. The detailed definitions are shown in Tab. I. In MPDE, since the authors aim to propose a pipeline requiring no priors, they simply assume that motion relations R_e and spatial relations R_s correspond one-to-one, in other words, $R_e = R_s$. However, this assumption is coarse and may fail in several cases. In this paper, we aim to predict accurate spatial relations by incorporating semantic segmentation information. We first construct a superpixel level graph $G = \{S, E, \omega\}$, where $E = \{(i,j)\}$, S_i and S_j are neighboring superpixels. $\omega = \{w_{ij}\}$ denotes the weight of each edge in E. Then we optimize the following energy function to jointly estimate homographies and motion relations,

$$\begin{aligned} E_s(\boldsymbol{H}, \boldsymbol{R_e}) = &\sum_{S_i\in\boldsymbol{S}} (E_{color}(H_i) + \lambda_{s_0} E_{semantic}(H_i)) \\ &+ \lambda_{s_1} \sum_{S_i\in\boldsymbol{S_u}} E_{epipolar}(H_i) \\ &+ \lambda_{s_2} \sum_{(i,j)\in\boldsymbol{E}} E_o(R_e(i,j)) \\ &+ \lambda_{s_3} \sum_{(i,j)\in\boldsymbol{E}} E_{pair}(H_i, H_j, R_e(i,j)). \end{aligned} \tag{1}$$

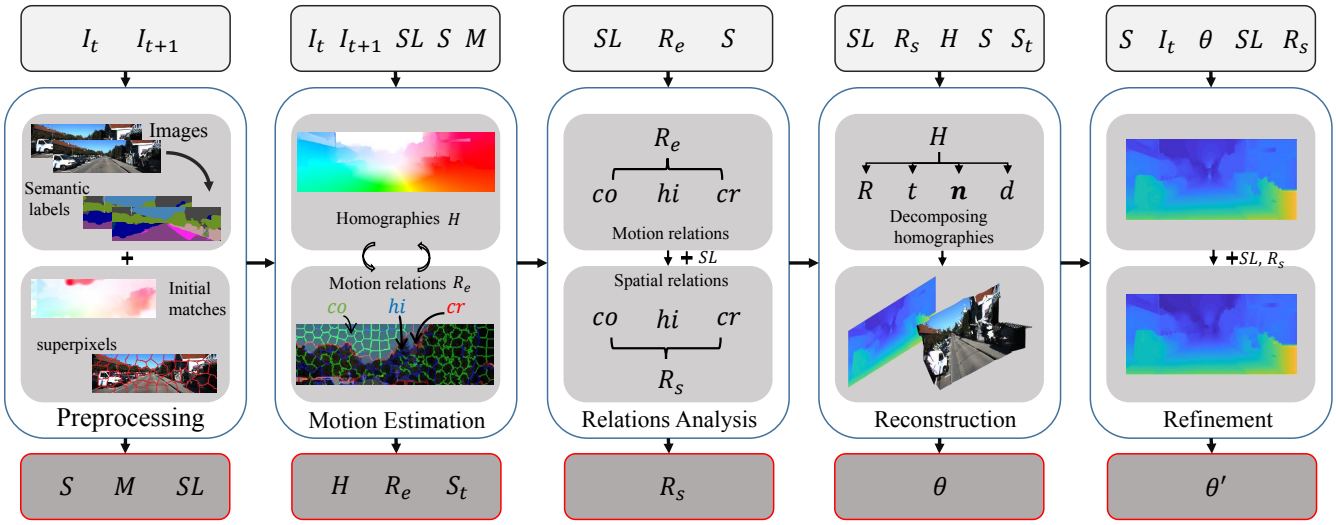

Fig. 1. The pipeline of the proposed method. Our pipeline consists of five main steps: preprocessing, motion estimation, relations analysis, reconstruction and refinement. The rounded rectangles in black borders list the inputs of each step. And the rounded rectangles in red borders list the outputs. SL: semantic labels. S_t: reliable background superpixels.

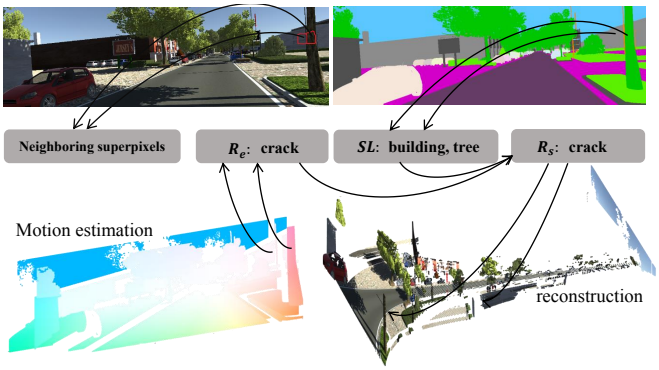

Fig. 2. This diagram demonstrates how to predict spatial relations R_s with R_e and semantic labels SL. Taking the marked superpixels as an example, their categories are *building* and *tree* and their motion relation is *crack*. Thus, considering both semantic and motion information, finally, we predict that the spatial relation between the two superpixels is *crack*.

where S_u denotes superpixels that belong to the background. We use the semantic segmentation results to directly obtain S_u.

In this energy, E_{color} is defined as,

$$E_{color} = \frac{1}{|S_i|} \sum_{p_j \in S_i} \rho(|I_t(p_j) - I_{t+1}(H_i P_j)|) \tag{2}$$

where $\rho(*)$ is a robust truncated function. The $E_{semantic}$ term is incorporated to ensure that corresponding pixels in the two frames have the same semantic label,

$$E_{semantic} = \frac{1}{|S_i|} \sum_{p_j \in S_i} \delta(SL_t(p_j) \neq SL_{t+1}(H_i P_j)) \tag{3}$$

where $SL_t(p_i)$ denotes the semantic label of pixel p_i in I_t. $\delta(*)$ is an indicator function defined as,

$$\delta(c) = \begin{cases} 0 & c \ is \ false \\ 1 & c \ is \ true. \end{cases} \tag{4}$$

The $E_{epipolar}$ term is defined on superpixels belonging to the background,

$$E_{epipolar} = \frac{1}{|S_i|} \sum_{p_j \in S_i} \rho\left(\frac{1}{Z_j} |(H_i P_j)^T F P_j|\right) \tag{5}$$

where F is the fundamental matrix estimated with all pixel matches in S_u by RANSAC [25]. Z_j is an adaptive normalization parameter defined in the form of Sampson distance. The E_o term is proposed to encourage a simpler explanation of the scenes,

$$E_o(R_e(i,j)) = \begin{cases} \omega_{cr1} & R_e(i,j) = cr, SL(S_i) = SL(S_j) \\ \omega_{cr2} & R_e(i,j) = cr, SL(S_i) \neq SL(S_j) \\ \omega_{hi} & R_e(i,j) = hi \\ \omega_{co} & R_e(i,j) = co \end{cases}, \tag{6}$$

where $SL(S_i)$ denotes the semantic label of superpixel S_i. $SL(S_i) = mode(SL(p_j))$, for all $p_j \in S_i$. In our experiment, we set $\omega_{cr1} > \omega_{cr2} > \omega_{hi} > \omega_{co} = 0$ under the assumption that the relations of superpixels belonging to the same category tend to be hinge or coplanar. The pairwise term E_{pair} encapsulates homographies H and motion relations R_e,

$$E_{pair}(H_i, H_j, R_e(i,j)) = \begin{cases} 0 & R_e(i,j) = cr \\ \frac{1}{|B_{ij}|} \sum_{p \in B_{ij}} |H_i \cdot P - H_j \cdot P| & R_e(i,j) = hi \\ \frac{1}{|S_i \cup S_j|} \sum_{p \in S_i \cup S_j} |H_i \cdot P - H_j \cdot P| & R_e(i,j) = co. \end{cases} \tag{7}$$

After determining motion relations R_e, we first select several reliable superpixels that belong to the background and fix their scales to be 1 in reconstruction. Different from the motion selection step in MPDE [12], we use the $E_{epipolar}$ term to select the superpixels. For superpixel S_i in S_u, if $E_{epipolar}(H_i) < \tau_0$, it is selected as a reliable background superpixel. We denote the reliable background superpixel set as S_t.

At last, we predict spatial relations. For neighboring superpixels with the same semantic label, their spatial relations and motion relations correspond one-to-one, which means

TABLE II

R_s PREDICTION WHEN SEMANTIC LABELS ARE DIFFERENT.

$SL(S_i)$	$SL(S_j)$	R_e	R_s
sky	-	-	*cr*
tree	*building*	-	*cr*
objects	*ground*	$R_e(i,j) = co \quad or \quad hi$	$R_e(i,j)$
objects	*ground*	$R_e(i,j) = cr$	*cr* or *hi*
otherwise		-	$R_e(i,j)$

$R_e = R_s$. For neighboring superpixels belonging to different categories, we analyze their categories and combine semantic and motion information to determine their spatial relations. In traffic scenes, the prediction rules are summarized in Tab. II. Note that when $SL(S_i) = objects$, $SL(S_j) = ground$ and $R_e(i,j) = cr$, we set $R_s = cr \quad or \quad hi$ according to semantic information. For example, superpixels on the bottom of vehicles are in *hinge* relations with superpixels on the surrounding ground, while superpixels on the top of vehicles are in *crack* relations. In non-traffic scenes, we use the last three rules in Tab. II.

C. Reconstruction

After estimating the spatial relations, we can then reconstruct the whole scene. By decomposing the homographies obtained in motion estimation, for superpixel S_i, we get rotation $R_i \in \mathbb{R}^{3\times3}$, translation $t_i \in \mathbb{R}^3$, plane norm $n_i \in \mathbb{R}^3$ and depth d_i, up to an unknown scale s_i. We optimize the following energy function in this step,

$$\begin{aligned} E_r(\boldsymbol{\theta}, \boldsymbol{s}) = \lambda_{r_1} \sum_{S_i \in \boldsymbol{S}} E_{fit}(\theta_i, s_i) + \lambda_{r_2} \sum_{S_i \in \boldsymbol{S}} E_{sim}(s_i) \\ + \sum_{(j,k)\in \boldsymbol{E}} E_{relations}(\theta_j, \theta_k) \\ + \lambda_{r_3} \sum_{(j,k)\in \boldsymbol{E_c}} E_{angle}(\theta_j, \theta_k), \\ \boldsymbol{s.t.} \quad s_i = 1, \forall S_i \in \boldsymbol{S}_t, \end{aligned} \tag{8}$$

The E_{fit} and E_{sim} are defined in the same way as in MPDE. The E_{fit} term fits a plane for each superpixel,

$$E_{fit}(\theta_i, s_i) = \sum_{p_i \in S_i} \rho(|\theta_i \cdot P_i - s_i d_i \boldsymbol{n_i}^T K^{-1} P_i|) \tag{9}$$

and the E_{sim} term favors simple explanations of the scene, $E_{sim}(s_i) = \delta(s_i \neq 1)$.

The $E_{relations}$ term is incorporated to solve the RSA problem and smooth the result with spatial relations,

$$E_{rel}(\theta_j, \theta_k) = \begin{cases} 0 & R_s(j,k) = cr \\ \frac{\omega_{hi}}{|B_{jk}|} \sum_{p\in B_{jk}} |\theta_j \cdot P - \theta_k \cdot P| & R_s(j,k) = hi \\ \frac{\omega_{co}}{|S_k \cup S_j|} \sum_{p\in S_k \cup S_j} |\theta_j \cdot P - \theta_k \cdot P| & R_s(j,k) = co, \end{cases} \tag{10}$$

where $\omega_{hi} < \omega_{co}$.

After presenting the main three terms, we can also incorporate several optional terms for specific scenes. For example, in traffic scenes, we can exploit perpendicular relations. In general, buildings and trees are perpendicular to the ground. Therefore, the plane norms of superpixels on buildings and trees are perpendicular to the plane norms of superpixels on the ground. We design E_{angle} under this assumption,

$$E_{angle}(\theta_i, \theta_j) = \left| \frac{(\theta_i K)(\theta_j K)^T}{\|\theta_i K\| \, \|\theta_j K\|} \right| \tag{11}$$

where $(i,j) \in E_c$. E_c is constructed with semantic information and contains all pairs of perpendicular superpixels.

D. Refinement

We refine the plane parameters under the piecewise planar assumption. For nearby pixels with similar appearances (color, gradient, etc.), we assume they are on the same 3D plane and share the same plane parameter. Then we follow FGS [26] to smooth the result based on the weighted least squares. The energy function is defined as follows,

$$E_{rf}(\theta') = \sum_p \left((\theta'_p - f(p))^2 + \lambda_r \sum_{q\in N(p)} \omega_{p,q} (\theta'_p - \theta'_q)^2 \right) \tag{12}$$

where $f(p) = \frac{1}{Z_p} \sum_{q\in\Omega_{W\times W}(p)} \omega_{p,q}(g)\theta_q$. $\Omega_{W\times W}(p)$ represents a square window whose center is p and sides are W in length. Z_p is a normalization parameter. For weighting parameter $\omega_{p,q}$, we combine RGB image I_t, semantic result SL and spatial relations R_s to define it,

$$\begin{aligned} \omega_{p,q} = \frac{1}{4} \left(\delta\left(SL(p) = SL(q)\right) + 1\right) * \left(\delta\left(R_s(i,j) \neq cr\right) + 1\right) \\ * exp(-|I_p - I_q|^2 / (2(\sigma_c)^2) \quad p \in S_i, q \in S_j \end{aligned} \tag{13}$$

where σ_c is a constant. The optimization method of Eq. 12 is introduced in FGS [26] in detail. After refinement, we obtain the final depth reconstructed result D_t of the input image I_t.

E. Optimization

For energy functions Eq.1 and Eq.8, we adopt an efficient coordinate descent method to optimize them. Taking Eq.1 as an example, we first initialize pixel matches by minimizing $E_{color} + E_{semantic}$ with PatchMatch [21]. We initialize homographies H, fundamental matrix F and R_e with these pixels matches. Then we iterate to optimize H, F and R_e as follows,

- Fix H and optimize R_e in a closed form.
- Fix R_e and optimize H with the fast propagation method, introduced in [27], [12].
- Update F with the latest H.

IV. EXPERIMENTS AND DISCUSSION

A. Experimental setup

In our experiments, we take two frames as inputs. We set superpixel size to be about 150 pixels per superpixel. We evaluate our method on MPI Sintel [28], KITTI [4], [3] and Virtual KITTI [29] datasets under popular evaluation metrics. For semantic segmentation, in traffic scenes including KITTI and Virtual KITTI, we adopt PSPNet [30] to predict semantic labels, while on MPI Sintel, we adopt the segmentation network in MRFlow [31]. For comparison, we select unsupervised learning methods SFMLearner [7], Geonet [8], DF-Net [13], Struct2depth [9] and traditional SFM-based methods DMDE [11], S.Soup [1] and MPDE

TABLE III

ESTIMATION ACCURACY OF MOTION RELATIONS AND DEPTH RELATIONS ON VIRTUAL KITTI DATASET.

		co	hi	cr
R_e	Ours	0.91	0.73	0.89
	MPDE	0.90	0.57	0.78
R_s	Ours	0.88	0.67	0.81
	MPDE	0.84	0.53	0.74

TABLE IV

PERFORMANCE COMPARISON ON KITTI DATASET.

	Abs Rel	RMSE	$\sigma < 1.25$
SFMLearner [7]	0.183	6.709	0.734
Geonet [8]	0.155	5.857	0.793
3DGeo [14]	0.159	5.912	0.784
Struct2depth [9]	**0.109**	4.750	0.874
USC-Net [33]	0.137	5.439	0.830
LR-consistency [34]	0.124	6.125	0.841
DF-Net [13]	0.150	5.507	0.806
DMDE [11]	0.148*	-	-
S.Soup [1]	0.126*	-	-
MPDE [12]	0.123	4.573	0.861
Ours	0.111	**4.537**	**0.891**

[12] as baseline methods. For learning-based methods, we directly use the publicly available codes and models in traffic scenes. On KITTI, we follow the widely-used eigen split [32]. On Virtual KITTI, we randomly select 120 pairs of images as the test set. Unlike in traffic scenes, on MPI Sintel, we finetune the models with a small split of the sequences and evaluate on the remaining part. We follow the split of S.Soup [1]. Since the authors of DMDE and S.Soup haven't released the source codes, we directly post the results from their papers, which are labelled with stars in Tab. IV, Tab. V and Tab. VI.

B. Accuracy of R_e and R_s

We conduct an experiment on Virtual KITTI dataset to evaluate the accuracy of our method on estimating R_e and predicting R_s. We first randomly select 60 pairs of images as the test set. We then estimate ground truth relations R_e^{gt} and R_s^{gt} with ground truth optical flow and depth maps. We compare the accuracy rate of our method and MPDE in Tab. III. As Tab. III shows, by incorporating semantic information, our method outperforms MPDE in both motion and spatial relations, which, to a large extent, explains why our method outputs superior results to MPDE.

C. Results on different benchmarks

We first evaluate our method on KITTI dataset. KITTI is an important real-world computer vision benchmark for multiple tasks. For depth estimation, it provides sparse LiDAR measurements as the ground truth depth.

We compare our method with top unsupervised learning and SFM-based methods in Tab. IV. As Tab. IV shows, our method achieves comparable performance to state-of-the-art methods. And although our method is slightly inferior to Struct2depth under *Abs Rel*, our method outperforms it on the other two metrics.

TABLE V

PERFORMANCE COMPARISON ON VIRTUAL KITTI DATASET.

	Abs Rel	RMSE	$\sigma < 1.25$
SFMLearner [7]	0.153	7.030	0.755
Geonet [8]	0.132	6.006	0.802
S.Soup [1]	0.105*	-	-
MPDE [12]	0.101	3.764	0.846
Ours	**0.096**	**3.105**	**0.880**

TABLE VI

PERFORMANCE COMPARISON ON MPI SINTEL DATASET.

	Abs Rel	$\sigma < 1.25$
SFMLearner [7]	0.473	0.207
Geonet [8]	0.440	0.277
DMDE [11]	0.297*	-
S.Soup [1]	0.167*	-
MPDE [12]	0.163	0.707
Ours	**0.152**	**0.753**

Virtual KITTI provides synthetic traffic videos for multiple vision tasks. We can directly evaluate our method with the ground truth depth maps. The evaluation results are listed in Tab. V. We can see that our method outperforms all other competitors. The results of state-of-the-art method Struct2depth [9] are not shown because it sometimes fails to output reasonable results on this dataset.

MPI Sintel is a famous dataset for evaluating optical flow and recently it started to provide perfect ground truth depth maps for the evaluation of depth estimation methods. Different from traffic scenes, MPI Sintel provides much more challenging scenes, e.g., a girl fighting with a dragon in a cave. This dataset is challenging due to large moving objects, complex scene structures, varying viewpoints, etc. We compare our method with competitors in Tab. VI. Due to the wide range of scene depth (e.g., in the $Mountain\_1$ sequence, the mountains are thousands of meters away), we manually select the depth range of each sequence in comparison for reliability. Before testing on MPI Sintel, all learning-based methods are fine-tuned, but they still output poor results. In fact, when fine-tuning, the loss function seldom converges. Our method outperforms all other methods under all considered metrics on this dataset.

D. Advantages and Disadvantages

Our method incorporates semantic segmentation into a traditional SFM-based reconstruction pipeline via exploiting superpixel relations. From the experiment results on different benchmarks, we summarize the advantages and disadvantages of our method over the competitors.

For supervised learning-based methods, the main advantage of our method is that we require no ground truth depth data. In dynamic scenes, especially like the scenes in MPI Sintel, ground truth depth data is much harder to collect than ground truth category labels. However, when data is available and sufficient, like in traffic scenes, supervised methods provide very accurate results, since they do not suffer from the scale-ambiguity problem.

For unsupervised learning methods, the main advantage

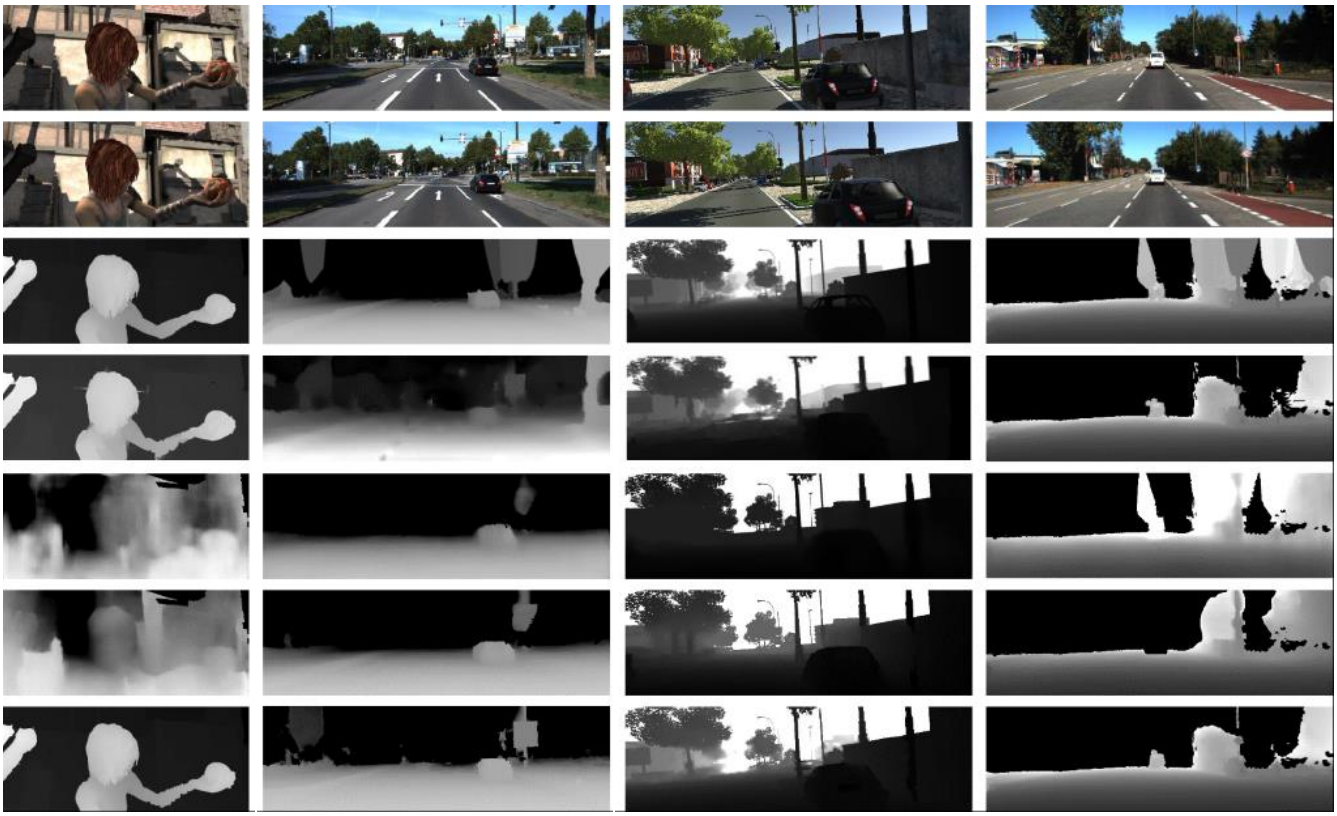

Fig. 3. Qualitative comparison on MPI Sintel, Virtual KITTI, and KITTI datasets. From Top to Bottom: input frame I_t, input frame I_{t+1}, ground truth depth, results of MPDE, DF-Net and USC-Net [33], our result. The first two columns show inverse depth maps while the last two show depth maps. Note that the results are adjusted for better visualization.

TABLE VII
PERFORMANCE COMPARISON WHEN TRAINING ON CITYSCAPES [35] AND EVALUATING ON KITTI.

	Abs Rel	RMSE	$\sigma < 1.25$
struct2depth [9]	0.152	5.557	0.796
Ours	0.111	4.537	0.891

of our method lies in that we can handle different dynamic scenes by adjusting several thresholds, while unsupervised learning methods usually need re-training or fine-tuning before being applied in a new dynamic scene. As shown in Tab. VII, when training on Cityscapes and evaluating on KITTI, the accuracy of Struct2depth decreases sharply while our method still outputs high-quality reconstruction results with the same parameters. Besides, current unsupervised learning methods mainly achieve great success in traffic scenes. In more challenging scenes, they still fail to output satisfying results, partly because the training data is insufficient.

The main disadvantage of our method is that we take 6 seconds to reconstruct a pair of images on a computer with i7-6700K, 16GB RAM, while learning-based methods usually run faster during inference. In the future, we plan to apply GPU acceleration in the implementation, and the running speed of our method can improve.

Comparing to SFM-based methods, typically DMDE, S.Soup, and MPDE, our method features higher-quality reconstruction and robustness to outliers. Note that although our method (6s) runs slower than learning-based methods (usually 0.05s), we outperform most SFM-based methods, like DMDE (60s) and S.Soup (600s). DMDE [11] depends on motion segmentation to figure out foreground parts, which is not trivial in dynamic scenes. S.Soup applies the ARAP term, which is hard to optimize and may disappoint in some cases with substantial motion. MPDE is the basis of our paper. In MPDE, motion relations and spatial relations are assumed to correspond one-to-one. However, this assumption is relatively coarse and needs other constraints as supplements. As shown in Tab. IV, Tab. V and Tab. VI, we outperform the above three methods on all benchmarks.

V. CONCLUSIONS

In this paper, we present a novel framework for dense semantic reconstruction in dynamic scenes, which consists of five main steps: preprocessing, motion estimation, superpixel relations analysis, reconstruction, and refinement. We incorporate semantic segmentation results into an SFM-based pipeline via exploiting superpixel relations, which provides a new insight into leveraging semantic information in traditional reconstruction techniques. Our framework outputs superior reconstruction results than competitors on several challenging datasets in a reasonable time, which demonstrates the great potential of our method in practical applications.

REFERENCES

[1] S. Kumar, Y. Dai, and H. Li, "Monocular dense 3D reconstruction of a complex dynamic scene from two perspective frames," in *ICCV*, 2017, pp. 4649–4657.

[2] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[3] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*. IEEE, 2012, pp. 3354–3361.

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[5] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys, "Joint 3D scene reconstruction and class segmentation," in *CVPR*, 2013, pp. 97–104.

[6] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*, 2012.

[7] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, vol. 2, no. 6, 2017, p. 7.

[8] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *CVPR*, vol. 2, 2018.

[9] V. Casser, S. Pirk, R. Majhourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," *national conference on artificial intelligence*, 2019.

[10] K.-N. Lianos, J. L. Schonberger, M. Pollefeys, and T. Sattler, "Vso: Visual semantic odometry," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 234–250.

[11] R. Ranftl, V. Vineet, Q. Chen, and V. Koltun, "Dense monocular depth estimation in complex dynamic scenes," in *CVPR*, 2016, pp. 4058–4066.

[12] Y. Di, H. Morimitsu, S. Gao, and X. Ji, "Monocular piecewise depth estimation in dynamic scenes by exploiting superpixel relations," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4363–4372.

[13] Y. Zou, Z. Luo, and J.-B. Huang, "DF-Net: unsupervised joint learning of depth and flow using cross-task consistency," in *ECCV*, 2018, pp. 36–53.

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

[15] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "LEGO: Learning edge with geometry all at once by watching videos," in *CVPR*, 2018, pp. 225–234.

[16] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1722–1729.

[17] D. G. Kottas and S. I. Roumeliotis, "Efficient and consistent vision-aided inertial navigation using line observations," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 1540–1547.

[18] S. Y. Bao and S. Savarese, "Semantic structure from motion," in *CVPR 2011*. IEEE, 2011, pp. 2025–2032.

[19] P. Gay, C. Rubino, V. Bansal, and A. Del Bue, "Probabilistic structure from motion with objects (psfmo)," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3075–3084.

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

[21] L. Bao, Q. Yang, and H. Jin, "Fast edge-preserving patchmatch for large displacement optical flow," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3534–3541.

[22] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, *et al.*, "SLIC superpixels compared to state-of-the-art superpixel methods," *TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[23] K. Yamaguchi, D. McAllester, and R. Urtasun, "Robust monocular epipolar flow estimation," in *CVPR*. IEEE, 2013, pp. 1862–1869.

[24] ——, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in *ECCV*. Springer, 2014, pp. 756–771.

[25] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[26] D. Min, S. Choi, J. Lu, B. Ham, K. Sohn, and M. N. Do, "Fast global image smoothing based on weighted least squares," *TIP*, vol. 23, no. 12, pp. 5638–5653, 2014.

[27] Y. Hu, Y. Li, and R. Song, "Robust interpolation of correspondences for large displacement optical flow," in *CVPR*, 2017.

[28] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *ECCV*, ser. Part IV, LNCS 7577, A. Fitzgibbon et al. (Eds.), Ed. Springer-Verlag, Oct. 2012, pp. 611–625.

[29] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *CVPR*, 2016.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[31] J. Wulff, L. Sevilla-Lara, and M. J. Black, "Optical flow in mostly rigid scenes," in *CVPR*, vol. 2, no. 3. IEEE, 2017, p. 7.

[32] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[33] J.-W. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," *arXiv preprint arXiv:1908.10553*, 2019.

[34] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[35] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.