Monocular Dense 3D Reconstruction of a Complex Dynamic Scene from Two Perspective Frames

Suryansh Kumar[1] Yuchao Dai[1,2] Hongdong Li[1,3]
[1]Australian National University, Canberra, Australia
[2]Northwestern Polytechnical University, Xi'an, China
[3]Australia Centre for Robotic Vision

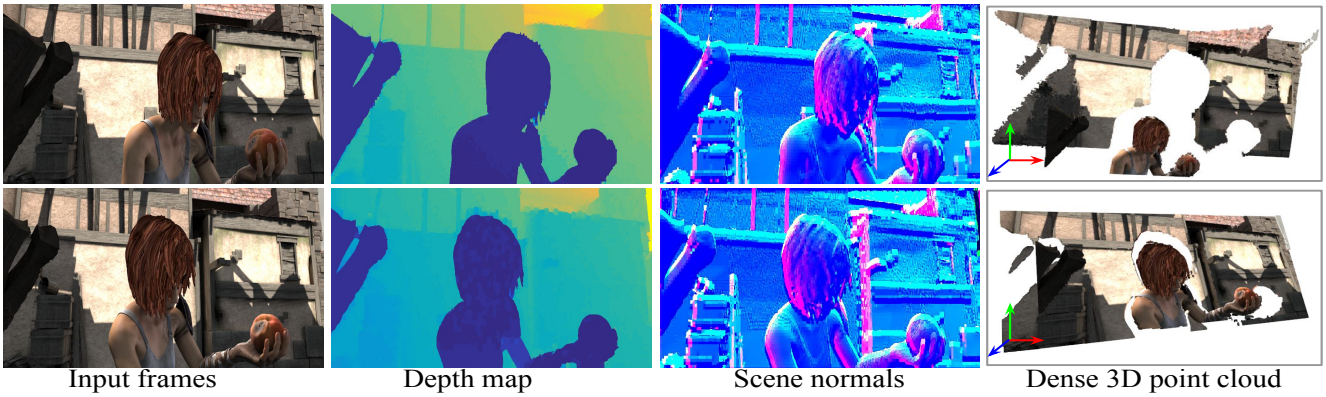


Figure 1: Dense 3D reconstruction of a complex dynamic scene from two perspective frames using our method. Here, both the subject and the camera are moving with respect to each other. (MPI Sintel [5] alley_1 frame 1 and 10).

Abstract

This paper proposes a new approach for monocular dense 3D reconstruction of a complex dynamic scene from two perspective frames. By applying superpixel over-segmentation to the image, we model a generically dynamic (hence non-rigid) scene with a piecewise planar and rigid approximation. In this way, we reduce the dynamic reconstruction problem to a "3D jigsaw puzzle" problem which takes pieces from an unorganized "soup of superpixels". We show that our method provides an effective solution to the inherent relative scale ambiguity *in structure-from-motion. Since our method does not assume a template prior, or per-object segmentation, or knowledge about the rigidity of the dynamic scene, it is applicable to a wide range of scenarios. Extensive experiments on both synthetic and real monocular sequences demonstrate the superiority of our method compared with the state-of-the-art methods.*

1. Introduction

Accurate recovery of dense 3D structure of dynamic scenes from images has many applications in motion capture [19], robot navigation[11], scene understanding [12], computer animation [5] *etc*. In particular, the proliferation of monocular camera in almost all modern mobile devices has elevated the demand for sophisticated dense reconstruction algorithm. When a scene is rigid, its 3D reconstruction can be estimated using conventional rigid-SfM (structure-from-motion) techniques [13]. However, real-world scenes are more complex containing not only rigid motions but also non-rigid deformations, as well as their combination. For example, a typical outdoor traffic scene consists of both multiple rigid motions of vehicles, and non-rigid motions of pedestrians *etc*. Therefore, it is highly desirable to develop a *unified* monocular 3D reconstruction framework that can handle generic (complex and dynamic) scenes.

To tackle the problem of monocular 3D reconstruction for dynamic scenes, a straightforward idea is to first pre-segment the scene into different regions, each corresponding to a single rigidly moving object or a rigid part of an object, then apply rigid-SfM technique to each of the regions. This idea of object-level motion segmentation has been used in previous work for non-rigid reconstruction [22][23], and for scene-flow estimation [20]. Russel *et al*. [24] proposed to simultaneously segment a dynamic scene into its con-

stituent objects and reconstruct a 3D model of the scene. Ranftl *et al*. [22] developed a two-stage pipeline (segmentation and then reconstruction) for monocular dynamic reconstruction. However, in a general dynamic setting, the task of densely segmenting rigidly moving objects or parts is not trivial. Consequently, inferring motion models for deforming shapes becomes very challenging. Furthermore, the success of object-level segmentation builds upon the assumption of multiple rigid motions, which fails to handle more general scenarios such as *e.g*. when the objects themselves are nonrigid or deformable.

This motivates us to ask a natural question: "*Is object-level motion segmentation essential for the dense 3D reconstruction of a complex dynamic scene?*". In this paper, we will justify our stance by proposing an approach that is free from object-level motion segmentation. We develop a unified method that is able to recover a dense and detailed 3D model of a complex dynamic scene, from its two perspective images, without assuming motion types or segmentation. Our method is built upon two basic assumptions about the scene, which are: 1) the deformation of the scene between two frames is *locally-rigid,* but *globally as-rigid-as-possible*, 2) the structure of the scene in each frame can be approximated by a *piecewise planar*. We call our new algorithm the *SuperPixelSoup* algorithm, for reasons that will be made clear in Section 2. Fig-1 shows some sample 3D reconstruction by our proposed method.

The main contributions of this work are:

1. We present a unified framework for dense two-frame 3D reconstruction of a complex dynamic scene, which achieves state-of-the-art performance.

2. We propose a new idea to resolve the inherent *relative scale ambiguity* for monocular 3D reconstruction by exploiting the as-rigid-as-possible (ARAP) constraint.

1.1. Related work

For brevity, we give a brief review only to previous works for monocular dynamic reconstruction that are mostly related to our work. The linear low-rank model has been used for dense nonrigid reconstruction. Garg *et al*. [10] solved the task with an orthographic camera model assuming feature matches across multiple frames. Fayad *et al*. [7] recovered deformable surfaces with a quadratic approximation, again from multiple frames. Taylor *et al*. [25] proposed a piecewise rigid solution using locally-rigid SfM to reconstruct a soup of rigid triangles. While their method is conceptually similar to ours, there are major differences: 1) We achieve *two-view* dense reconstruction while they need multiple views(N $\geq$ 4); 2) We use the *perspective camera model* while they rely on an orthographic camera model. Many real-world images such as a typical driving scene (*e.g*., KITTI) cannot be well explained by orthographic projection; 3) We solve the relative scale indeterminacy issue, which is an inherent ambiguity for 3D reconstruction under perspective projection, while Taylor *et al*.'s method does not suffer from this, at the cost of being restricted to the orthographic camera model. Russel *et al*. [24] and Ranftl *et al*. [22] used object-level segmentation for dense dynamic reconstruction. In contrast, our method is free from object segmentation, hence circumvents the difficulty associated with motion segmentation in a dynamic setting. The template-based approach is yet another method for deformable surface reconstruction. Yu *et al*. [29] proposed a direct approach to capturing dense, detailed 3D geometry of generic, complex non-rigid meshes using a single RGB camera. While it works for generic surfaces, the need of a template prevents its wider application to more general scenes. Wang [28] introduced a template-free approach to reconstruct a poorly-textured, deformable surface. However, its success is restricted to a single deforming surface rather than the entire dynamic scene. Varol *et al*. [27] reconstructed deformable surfaces based on a piecewise reconstruction, by assuming overlapping pieces.

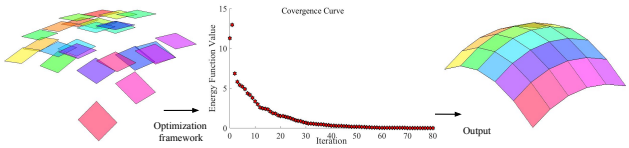


Figure 2: Reconstructing a 3D surface from a soup of un-scaled superpixels via solving a 3D Superpixel Jigsaw puzzle problem.

2. Overview of the proposed method

In this section, we present a high-level overview of our "SuperPixel Soup" algorithm for dense 3D scene reconstruction of a complex dynamic scene from two frames.

Given two perspective images (denoted as the reference image $\mathbf{I}$ and the next image $\mathbf{I}'$) of a generally dynamic scene, our goal is to recover the dense 3D structure of the scene. We first pre-segment the image into superpixels, then model the deformation of the scene by the union of piecewise rigid motions of its superpixels. Specifically, we divide the overall non-rigid reconstruction into small rigid reconstruction for each individual superpixel, followed by an assembly process which glues all these local individual reconstructions in a globally coherent manner. While the concept of the above divide-and-conquer procedure looks simple, there is, however, a fundamental difficulty (of *relative scale indeterminacy*) in its implementation. Relative scale indeterminacy refers to the well-known fact that using a moving camera one can only recover the 3D structure up to an unknown scale. In our method, the individual rigid reconstruction of each superpixel can only be determined up to an unknown scale, the assembly of the entire

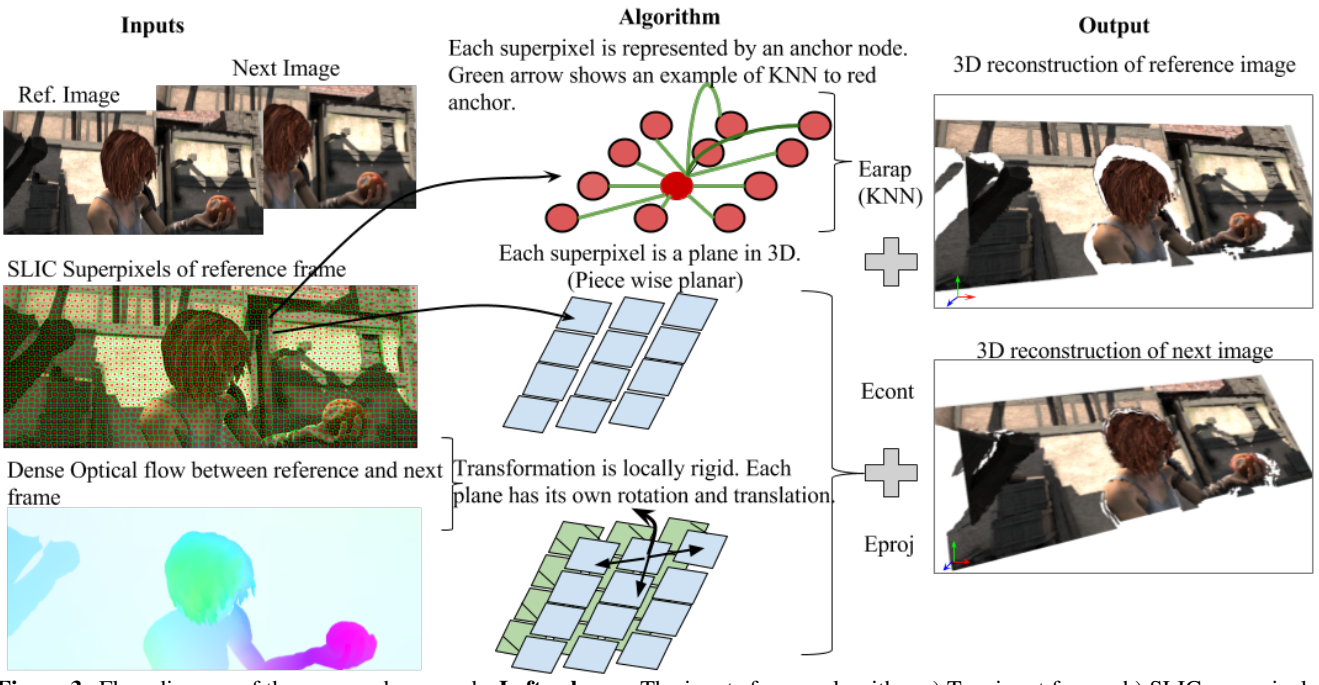


Figure 3: Flow diagram of the proposed approach. **Left column:** The inputs for our algorithm a) Two input frames b) SLIC superpixels [1] of the reference frame c) Dense optical flow between two frames. **Middle column:** Each individual superpixel is represented by an anchor node (in dark red). Every anchor node constrains the motion of K other anchor node (E_{arap}) in both frames. The depth continuity term ($E_{\textbf{cont}}$) is defined only for neighboring superpixels that shares the common boundary. **Right column:** The dense 3D point clouds of the reference frame and the next frame, where each individual plane in the next frame is related to the reference frame via a rigid motion.

non-rigid scene is only possible if and only if these relative scales among the superpixels are solved –which is, however, a challenging open task itself.

In this paper, we show how this can be done, under two very mild assumptions (about the dynamic scene and about the deformation). Specifically, these assumptions are:

- Basic Assumption-1: The transformation (*i.e.* deformation) between the two frames are locally *piecewise-rigid*, and globally **as rigid as possible**. In other words, the deformation is not arbitrary but rather regular in terms of rigidity.
- Basic Assumption-2: The 3D scene surface to be reconstructed is **piecewise-smooth** (or moreover, **piecewise-planar**) in both frames.

Under these assumptions, our method solves the unknown relative scales and obtains a globally-coherent dense 3D reconstruction of a complex dynamic (hence generally non-rigid) scene from its two perspective views.

Intuitively, our new method can be understood as the following process: Suppose every individual superpixel corresponds to a small planar patch moving rigidly in 3D space. Since the correct scales for these patches are not determined, they are floating in 3D space as a set of unorganized superpixel soup. Our method then starts from finding for each superpixel an appropriate scale, under which the entire set of superpixels can be assembled (glued) together coherently, forming a piecewise smooth surfaces, *as if* playing the game of "3D jig-saw puzzle". Hence, we call our method the "SuperPixel Soup" algorithm (see Figure 2 for a conceptual visualization).

The overall procedure of our method is presented in Algorithm-1.

Algorithm 1 : SuperPixel Soup

Input: Two monocular image frames and dense optical flow correspondences between them.
Output: 3D reconstruction of both image.
1. Divide the image into N superpixel and construct a K-NN graph to represent the entire scene as a graph $G(V, E)$ defined over superpixels §4.
2. Employ the two-view epipolar geometry to recover the rigid motion and 3D geometry for each 3D superpixel.
3. Optimize the proposed energy function to assemble (or glue) and align all the reconstructed superpixels ("3D Superpixel Jigsaw Puzzle").

3. Problem Statement

To implement the above idea of piecewise rigid reconstruction, we first partition the reference image **I**

into superpixels $\{\mathbf{s}_1, \mathbf{s}_2, .., \mathbf{s}_i, .., \mathbf{s}_N\}$, where each superpixel $\mathbf{s}_i$ is parametrized by its boundary pixels $\{\mathbf{x}_{bi} = [u_{bi}, v_{bi}, 1]^T \, | b = 1, ..., B_i\}$ in the image plane. We further define an *anchor point* $\mathbf{x}_{ai}$ for each superpixel, as the centroid point of the superpixel. Such a superpixel partition of the image plane naturally induces a piecewise planar segmentation of the corresponding 3D scene surface. We call each of the 3D segments as a 3D superpixel, and denote its boundary coordinates (in 3D space) as $\{\mathbf{S}_i\}$ in capital $\mathbf{S}$. Although *surfel* is perhaps a better term, we nevertheless call it "3D superpixel" for the sake of easy exposition. We further assume each 3D superpixel is a small 3D *planar patch*, parameterized by surface normal $\mathbf{n}_i \in \mathbb{R}^3$, 3D anchor-point $\mathbf{X}_{ai}$, and 3D boundary-points $\{\mathbf{X}_{bi}\}$ (*i.e.* these are the pre-images of $\mathbf{x}_{ai}$ and $\{\mathbf{x}_{bi}\}$). We assume every 3D superpixel $\mathbf{s}_i$ moves rigidly according $\mathbf{M}_i = \begin{pmatrix} \mathbf{R}_i & \lambda_i \hat{\mathbf{t}}_i \\ \mathbf{0} & 1 \end{pmatrix} \in \mathrm{SE}(3)$, where $\mathbf{R}_i$ represents rotation, $\hat{\mathbf{t}}_i$ is the translational direction, and λ_i is the unknown scale.

Now we are in a position to state the problem in a more precise way: Given two intrinsically calibrated perspective images $\mathbf{I}$ and $\mathbf{I}'$ of a generally dynamic scene and the corresponding dense correspondences, *i.e.*, optical flow field, our task is to reconstruct a piecewise planar approximation of the dynamic scene surface. We need a *dense* flow field, but do not require it to be perfect because it is only used to initialize our algorithm, and as the algorithm runs, the final flow field will be refined. The deformable scene surface in the reference frame (*i.e.*, $\mathbf{S}$) and the one in the second frame (*i.e.*, $\mathbf{S}'$) are parametrized by their respective 3D superpixels $\{\mathbf{S}_i\}$ and $\{\mathbf{S}'_i\}$, where each $\mathbf{S}_i$ is described by its surface normal $\mathbf{n}_i$ and an anchor point $\mathbf{X}_{ai}$. Any 3D plane can be determined by an anchor point $\mathbf{X}_{ai}$ and a surface normal $\mathbf{n}_i$. If one is able to estimate all the 3D anchor points and all the surface normals, the problem is solved.

4. Solution

Build a K-NN graph. We identify a 3D superpixel by its anchor point. The distance between two 3D superpixels is defined as the Euclidean distance between their anchor points in 3D space.

By connecting K nearest neighbors, we build a K-NN graph G(V,E) (*e.g.* as illustrated in Fig. 3 and Fig. 4). The graph vertices are anchor points, connecting with each other via graph edges. Overloading notation, we let $\mathbf{X}_{ai} = [X_{ai}, Y_{ai}, Z_{ai}]^T$ represent 3D world coordinates of the i-th superpixel. Suppose that we know the perfect $\mathbf{M}_i$, $\mathbf{n}_i$ for each individual $\mathbf{S}_i$, then $\mathbf{S}$ can be mapped to $\mathbf{S}'$ by moving each individual superpixel based on its corresponding locally rigid motion. The world and the image coordinates in the subsequent frames can be inferred by $\mathbf{X}'_{ai} = \mathbf{M}_i \mathbf{X}_{ai}$ and $\mathbf{s}'_i = \mathbf{K}\left(\mathbf{R}_i - \frac{\mathbf{t}_i \mathbf{n}_i^T}{d_i}\right)\mathbf{K}^{-1}\mathbf{s}_i$, where the latter represents a plane-induced homography [13], with d_i as the depth of the plane.

As-Rigid-As-Possible (ARAP) Energy Term. Our new method is built upon the idea that the correct scales of 3D superpixels can be estimated by enforcing prior assumptions that govern the deformation of the dynamic surface. Specifically, we require that, locally, the motion that each 3D-superpixel undergoes is rigid, and globally the entire dynamic scene surface must move as-rigid-as-possible (ARAP). In other words, while the dynamic scene is globally non-rigid, its deformation must be *regular* in the sense that it deforms as rigidly as possible. To implement this idea, we define an ARAP-energy term as:

$$E_{\text{arap}} = \sum_{i=1}^{N} \sum_{k \in \mathcal{N}_i} w_1(\mathbf{s}_{ai}, \mathbf{s}_{ak}) \|\mathbf{M}_i - \mathbf{M}_k\|_F + w_2(\mathbf{s}_{ai}, \mathbf{s}_{ak}) . \Big| \|\mathbf{X}_{ai} - \mathbf{X}_{ak}\|_2 - \|\mathbf{X}'_{ai} - \mathbf{X}'_{ak}\|_2 \Big|_1 . \tag{1}$$

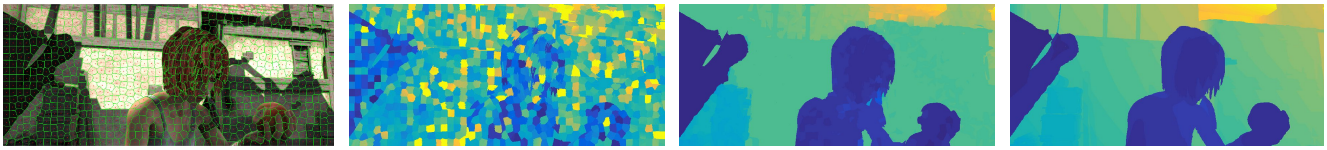
Here, the first term favors smooth motion between local neighbors, while the second term encourages inter-node distances between the anchor node and its K nearest neighbor nodes (denoted as $k \in \mathcal{N}_i$) to be preserved before and after motion (hence as-rigid-as-possible). We define the weighting parameters as:

$$w_1(\mathbf{s}_{ai}, \mathbf{s}_{ak}) = w_2(\mathbf{s}_{ai}, \mathbf{s}_{ak}) = \exp(-\beta \|\mathbf{s}_{ai} - \mathbf{s}_{ak}\|). \tag{2}$$

These weights are set to be inversely proportional to the distance between two superpixels. This is to reflect our intuition that, the further apart two superpixels are, the weaker the E_{arap} energy is. Although there may be redundant information in these two terms, we keep both nonetheless for the sake of flexibility in algorithm design. Note that, this term is only defined over anchor points, hence it enforces no depth smoothness along boundaries. The weighting term in E_{arap} advocates the local rigidity by penalizing over the distance between anchor points. This allows immediate neighbors to have smooth deformation over time. Also, note that E_{arap} is generally *non-convex*.

Planar Re-projection Energy Term. With the assumption that each superpixel represents a plane in 3D, it must satisfy corresponding planar reprojection error in 2D image space. This reprojection cost reflects the average dissimilarity in the optical flow correspondences across the entire superpixel due to its motion. Therefore, it helps us to constrain the surface normals, rotation and translation direction such that they obey the observed planar homography in the image space.

$$E_{\text{proj}} = \sum_{i=1}^{N} \frac{w_3}{|\mathbf{s}_i|} \sum_{j=1}^{|\mathbf{s}_i|} \|(\mathbf{s}_i^j)' - \mathbf{K}\left(\mathbf{R}_i - \frac{\mathbf{t}_i \mathbf{n}_i^T}{d_i}\right)\mathbf{K}^{-1}(\mathbf{s}_i^j)\|_F . \tag{3}$$



a) b) c) d)

Figure 4: a) Superpixelled reference image b) Individual superpixel depth with arbitrary scale (*unorganised superpixel soup*) c) recovered depth map using our approach (*organised superpixel soup*) d) ground-truth depth map.

where $|\mathbf{s}_i|$ represents the total number of pixel inside the i^{th} superpixel [1].

3D Continuity Energy Term. To favor a continuous/smooth surface reconstruction, we require two neighboring superpixels to have a smooth transition at their boundaries. We define a 3D continuity energy term as:

$$E_{\text{cont}} = \sum_{i=1}^{N} \sum_{k \in \mathcal{N}_i} w_4(\mathbf{s}_{bi}, \mathbf{s}_{bk}) \Big(\|\mathbf{X}_{bi} - \mathbf{X}_{bk}\|_F + \rho(\|\mathbf{X}'_{bi} - \mathbf{X}'_{bk}\|_F) \Big). \tag{4}$$

This term ensures the 3D coordinates across superpixel boundaries to be continuous in both frames. The neighboring relationship in E_{cont} is different from E_{arap} term. Here, the neighbors share common boundaries with each other. For each boundary pixel of a given superpixel, we consider its 4-connected neighboring pixels. w_4 is a trade-off scalar, which is defined as:

$$w_4(\mathbf{s}_{bi}, \mathbf{s}_{bk}) = \exp(-\beta \|\mathtt{I}(\mathbf{s}_{bi}) - \mathtt{I}(\mathbf{s}_{bk})\|_F), \tag{5}$$

i.e. weighting the inter-plane transition by the color difference. Here, subscript 'bi' and 'bk' indicate that the involved pixels shares the common boundary ('b') between i^{th} and k^{th} superpixel in the image space. ρ is a truncation function defined as $\rho = \min(., \sigma)$ to allow piecewise discontinuities. Here, β is a trade-off constant chosen empirically.

Combined Energy Function. Recall that our goal is to estimate piecewise rigid motion $(\mathbf{R}_i, \mathbf{t}_i)$, depth d_i, surface normal $\mathbf{n}_i$ and scale λ_i for each planar superpixel in 3D, given initialization. The key is to estimate the unknown relative scale λ_i. We solve this by minimizing the following energy function $E = E_{\text{arap}} + \alpha_1 E_{\text{proj}} + \alpha_2 E_{\text{cont}}$, namely,

$$\min_{\lambda_i, \mathbf{n}_i, d_i, \mathbf{R}_i, \mathbf{t}_i} E = E_{\text{arap}} + \alpha_1 E_{\text{proj}} + \alpha_2 E_{\text{cont}}, \quad \text{s. t.} \sum_{i=1..N} \lambda_i = 1, \lambda_i > 0. \tag{6}$$

The last equality constraint fixes the unknown freedom of a global scale. $\lambda_i > 0$ enforces the cheriality constraint [13].

[1] For brevity, we slightly abuse notation; both terms in Eq:-3 represent inhomogeneous image coordinate.

Optimization. The above energy function (Eq.- 6) is non-convex. We first solve the relative scales λ_i efficiently by minimizing the ARAP term in Eq.-(1) using interior-point methods [4]. Although the solutions found by the interior point method are at best local minimizers, empirically they appear to give good 3D reconstructions. In our experiments, we initialized all λ_i with an initial value of $\frac{1}{N}$.

Assigning superpixels to a set of planes can lead to non-smooth blocky effect at their boundaries. To smooth these blocky effects, we employ a refinement step to optimize over the surface normals, rotations, translations, and depths for all 3D superpixels using Eq.- 3 and Eq.-4. We solved the resultant discrete-continuous optimization with the Max-Product Particle Belief propagation (MP-PBP) procedure by using the TRW-S algorithm [16]. In our implementation, we generated 50 particles as proposals for the unknown parameters. Repeating the above strategy for 5-10 iterations, we obtained a smooth and refined 3D structure of the dynamic scene.

Implementation details. We partitioned a reference image into about 1,000-2,000 superpixels [1]. We used a state-of-the optical flow algorithm [3] to compute dense correspondences across two frames. Parameters like α_1, α_2, β, σ were tuned differently for different datasets. However, $\beta = 3$ and $\sigma = 15$ are fixed for all our tests on MPI Sintel and on VKITTI. To initialize the iteration, local rigid motion is estimated using traditional SfM pipeline [13]. Our current implementation in C++/MATLAB takes around 10-12 minutes to converge for images of size 1024×436 on a regular desktop with Intel core i7 processor.

5. Experiments

We evaluated the performance of our method both qualitatively and quantitatively on various bedatasets that contain dynamic objects: the KITTI dataset [11], the virtual KITTI [9], the MPI Sintel [5] and the YouTube-Objects [21]. We also tested our method on some commonly used non-rigid deformation data: *Paper*, *T-shirts* and *Back* sequence [27] [26][10]. Example images and our reconstruction results are illustrated in Fig. 5.

Evaluation Metrics: For quantitative evaluation, the errors are reported in *i.e*. mean relative error (MRE), defined

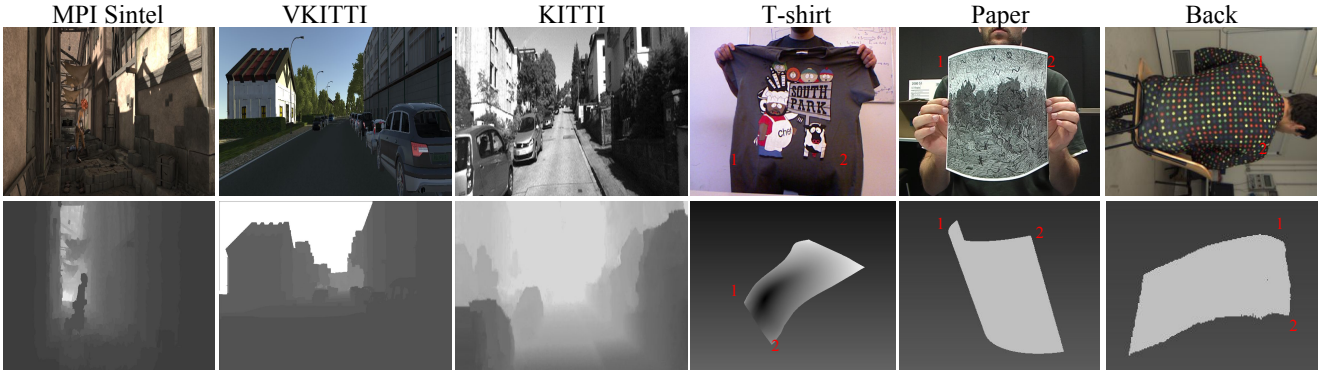


Figure 5: 3D reconstruction and depth map obtained using our algorithm on different benchmarking datasets. The first three columns demonstrate the reconstruction of the entire scene that is composed of rigid and complex motion. The last three columns show the accurate reconstruction of deformable objects on real non-rigid benchmark datasets.

as $\frac{1}{P}\sum_{i=1}^{P}|z_{gt}^{i}-z_{est}^{i}|/z_{gt}^{i}$. Here, z_{est}^{i}, z_{gt}^{i} denotes the estimated, and ground-truth, depth respectively with P as the total number of 3D points. The error is computed after rescaling the recovered shape properly, as the reconstruction is only made up to an unknown global scale. We used MRE for the sake of consistency with previous work [22]. Quantitative evaluations for the YouTube-Objects dataset and the Back dataset are missing because for them no ground-truth results are provided.

Baseline Methods: The performance of our presented method is compared to several monocular dynamic reconstruction methods, which include the Block Matrix Method (BMM) [6], Point Trajectory Approach (PTA) [2], and Low-rank Reconstruction (GBLR) [8]), Depth Transfer (DT) [15], and (DMDE) [22]. [2]

In Fig-(6) we show the recovered depth map along with scene surface normals. These results highlight the effectiveness of our method in handling diverse scenarios.

MPI Sintel: This dataset is derived from an animation movie with complex dynamic scenes. It contains highly dynamic sequences with large motions and significant illumination changes. It is a challenging dataset particularly for the piece-wise planar assumption due to the presence of many small and irregular shapes in the scene. We selected 120 pairs of images to test our method, which includes alley_1, ambush_4, mountain_1, sleeping_1 and temple_2. Fig-8(a) gives quantitative comparisons against several other competing methods. As observed in the figure, our method outperforms all the competing methods on all the testing sequences shown here.

Virtual KITTI: The Virtual KITTI dataset contains computer rendered photo-realistic outdoor driving scenes which resemble the KITTI dataset. The advantage of using this dataset is that it provides perfect ground-truths for many measurements. Furthermore, it helps to simulate algorithm related to dense reconstruction with noise free and distortion-free images, facilitating quick experimentation. We selected 120 images from 0001_morning, 0002_morning, 0006_morning and 0018_morning. The results obtained are shown in Figure 8(a). Again, our method outperforms all the competing methods with a clear margin on all the test sequences.

KITTI: We tested real KITTI to evaluate our method's performance for noisy real-world sequences. We used the KITTI's sparse LiDAR points as the 3D ground-truth for evaluation. We also used other sequences for qualitative analysis (see Figure 5). Figure 8(b) demonstrates the obtained depth accuracy. Our method achieves the best performance for all the testing sequences.

YouTube-Objects: We tested our method on sequences from the Youtube-Objects Dataset [21]. These are community-contributed videos downloaded from the YouTube. Due to the lack of ground truth 3D reconstruction, we only show the results in Fig. 7 visually.

Non-rigid datasets (Paper, T-shirt, Back): We benchmarked our method in commonly used deformable object sequences, namely, Kinect_Paper and Kinect_Tshirt [26]. Table-1 presents the mean depth error obtained on these sequences. Note that all the benchmarking non-rigid structure-from-motion methods reported in Table-1 (GLRT [8], BMM [6], and PTA [2]) used multi-frame while our method only used two frames. Qualitative results are demonstrated in Fig. 5.

Comparison: Table 1 provides a statistical comparison between our method and other competing methods. It shows that our method delivers consistently superior reconstruction accuracy on these benchmarking datasets, even better than those methods which use multiple image frames.

Effect of K: Under our method, the ARAP energy term

[2] We did not compare our method with [24] due to the code provided by the authors of [24] crashed unexpectedly on several of the test sequences.

[3] Intrinsic matrix was obtained through personal communication.

[3] Intrinsic matrix for the Back sequence is not available with dataset. We made an approximate estimation.

Figure 6: Depth map and scene normals on MPI and VKITTI dataset.

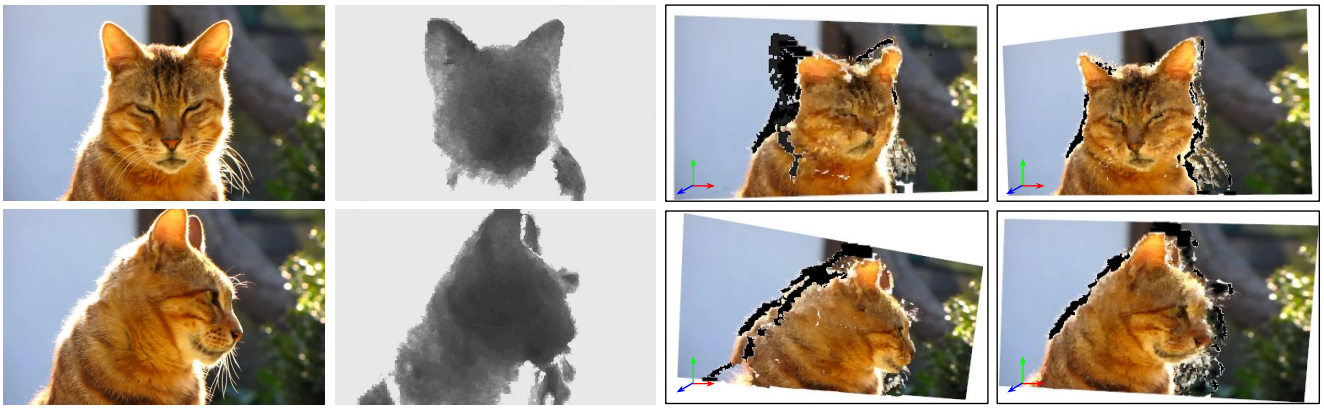


Figure 7: Depth and 3D reconstruction results for the cat sequence taken from YouTube-Objects Dataset[21][3]. For this experiment, we used 10,000 superpixels.

Method → (Method type)	DT [15] (Single frame)	GLRT [8] (Multi-frame)	BMM [6] (Multi-frame)	PTA [2] (Multi-frame)	DMDE [22] (Two-frame)	Ours (Two-frame)
MPI Sintel	0.4833	0.4101	0.3121	0.3177	0.297	**0.1669**
Virtual KITTI	0.2630	0.3237	0.2894	0.2742	-	**0.1045**
KITTI	0.2703	0.4112	0.3903	0.4090	0.148	**0.1268**
kinect_paper	0.2040	0.0920	**0.0322**	0.0520	-	0.0476
kinect_tshirt	0.2170	0.1030	0.0443	**0.0420**	-	0.0480

Table 1: Performance Comparison: This table lists the MRE errors. For DMDE [22] we used its previously reported result as its implementation is not available publicly.

is evaluated within K nearest neighbors, different K may have a different effect on the resultant 3D reconstruction. We conducted an experiment to analyze the effect of varying K on the MPI Sintel dataset and the results are illustrated in Fig. 9. With the increase of K, the recovered scene becomes more rigid, as the neighborhood size increases.

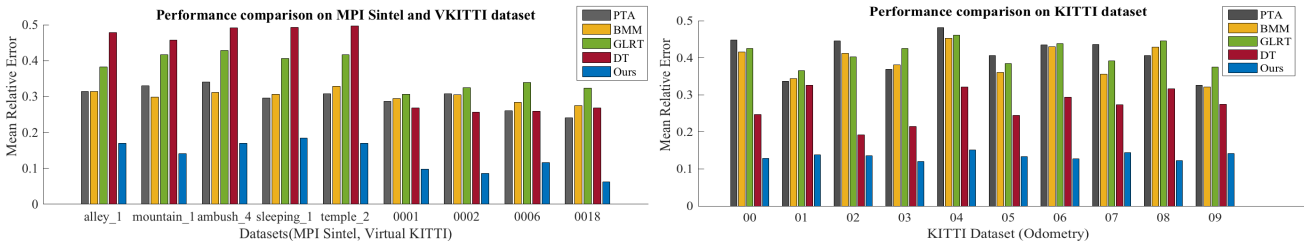
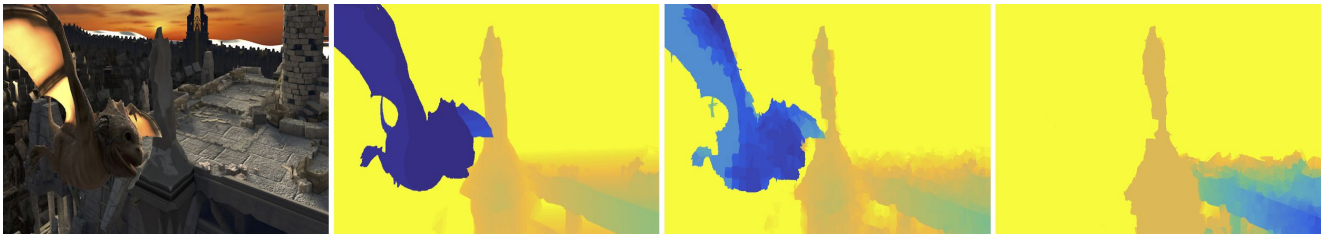


Figure 8: Quantitative evaluation on benchmark datasets. The depth error is calculated by adjusting the numerical scale of obtained depth map to ground-truth value, to account for global scale ambiguity. (a)-(b) comparison on MPI, Virtual KITTI and KITTI dataset. PTA [2], BMM [6], GLRT[8], DT [15]. These numerical values show the fidelity of reconstruction that can be retrieved on benchmark datasets using our formulation.



a) b) c) d)

Figure 9: Effect of parameter K in building the K-NN graph. Our algorithm results in good reconstruction if a suitable K is chosen, in accordance with levels of complexity of the dynamic scene. b) Ground-truth depth-map (scaled for illustration purpose). c) when K=4, a reasonable reconstruction is obtained. d) when K=20, regions tend to grow bigger. (Best viewed in color.)

When k=20, the dragon region was absorbed into the sky region, which results in an incorrect reconstruction. In most of our experiments, we used a K in the range of $15 - 20$, which achieved satisfactory reconstructions.

Our approach may disappoint if the neighboring relations between superpixels do not hold in the successive frame due to the substantial motion. A couple of examples for such situations are discussed and shown in the supplementary material for better understanding. Furthermore, we encourage the readers to go through the supplementary material for few more analysis and possible future works.

6. Conclusion

To reconstruct a dense 3D model of a complex, dynamic, and generally non-rigid scene from its two images captured by an arbitrarily-moving monocular camera is often considered as a very challenging task in Structure-from-Motion. In contrast, the reconstruction of a rigid and stationary scene from two views is a mature and standard task in 3D computer vision, which can be solved easily if not trivially.

This paper has demonstrated that such a dense 3D reconstruction of dynamic scenes is, in fact possible, provided that certain prior assumptions about the scene geometry and about the dynamic deformation of the scene are satisfied. Specifically, we only require that 1) the dynamic scene to be reconstructed is piecewise planar, and 2) the deformation itself between the two frames is locally-rigid but globally as-rigid-as-possible. Both assumptions are mild and realistic, commonly satisfied by real-world scenarios. Our new method dubbed as *the SuperpixelSoup algorithm* is able to solve such a challenging problem efficiently, leading to accurate and dense reconstruction of complex dynamic scenes. We hope in theory our method offers a valuable new insight to monocular reconstruction, and in practice, it provides a promising means to perceive a complex dynamic environment by using a single monocular camera. Finally, we want to stress that the rigidity assumption (and the ARAP constraint) used by the paper is a powerful tool in multi-view geometry research—careful investigation of which may open up new opportunities in the development of advanced techniques for 3D reconstruction.

Acknowledgment

This work was supported in part by Australian Research Council (ARC) grants (DE140100180, DP120103896, LP100100588, CE140100016), Australia ARC Centre of Excellence Program on Robotic Vision, NICTA (Data61) and Natural Science Foundation of China (61420106007). We thank the AC and the reviewers for their invaluable suggestions and comments.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 3, 5

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1442–1456, 2011. 6, 7, 8

[3] C. Bailer, B. Taetz, and D. Stricker. Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4015–4023, 2015. 5

[4] H. Y. Benson, R. J. Vanderbei, and D. F. Shanno. Interior-point methods for nonconvex nonlinear programming: Filter methods and merit functions. *Computational Optimization and Applications*, 23(2):257–272, 2002. 5

[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision*, pages 611–625. Springer, 2012. 1, 5

[6] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014. 6, 7, 8

[7] J. Fayad, L. Agapito, and A. Del Bue. Piecewise quadratic reconstruction of non-rigid surfaces from monocular sequences. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *European Conference on Computer Vision*, pages 297–310, 2010. 2

[8] K. Fragkiadaki, M. Salas, P. Arbeláez, and J. Malik. Grouping-based low-rank trajectory completion and 3d reconstruction. In *Advances in Neural Information Processing Systems*, pages 55–63, 2014. 6, 7, 8

[9] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. *arXiv preprint arXiv:1605.06457*, 2016. 5

[10] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1272–1279, 2013. 2, 5

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Rob. Res.*, 32(11):1231–1237, Sept. 2013. 1, 5

[12] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3d scene reconstruction and class segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 1

[13] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 4, 5

[14] T. Igarashi, T. Moscovich, and J. F. Hughes. As-rigid-as-possible shape manipulation. In *ACM transactions on Graphics (TOG)*, volume 24, pages 1134–1141. ACM, 2005.

[15] K. Karsch, C. Liu, and S. B. Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2144–2158, 2014. 6, 7, 8

[16] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1568–1583, 2006. 5

[17] S. Kumar, Y. Dai, and H. Li. Multi-body non-rigid structure-from-motion. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 148–156. IEEE, 2016.

[18] S. Kumar, Y. Dai, and H. Li. Spatio-temporal union of subspaces for multi-body non-rigid structure-from-motion. *Pattern Recognition*, 71:428–443, 2017.

[19] M. Loper, N. Mahmood, and M. J. Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):220, 2014. 1

[20] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. 1

[21] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3282–3289. IEEE, 2012. 5, 6, 7

[22] R. Ranftl, V. Vineet, Q. Chen, and V. Koltun. Dense monocular depth estimation in complex dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4058–4066, 2016. 1, 2, 6, 7

[23] C. Russell, R. Yu, and L. Agapito. Video pop-up: Monocular 3d reconstruction of dynamic scenes. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *European Conference on Computer Vision*, pages 583–598, Cham, 2014. Springer International Publishing. 1

[24] C. Russell, R. Yu, and L. Agapito. Video pop-up: Monocular 3d reconstruction of dynamic scenes. In *European conference on computer vision*, pages 583–598, 2014. 1, 2, 6

[25] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2761–2768. IEEE, 2010. 2

[26] A. Varol, M. Salzmann, P. Fua, and R. Urtasun. A constrained latent variable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2248–2255. Ieee, 2012. 5, 6

[27] A. Varol, M. Salzmann, E. Tola, and P. Fua. Template-free monocular reconstruction of deformable surfaces. In *International Conference on Computer Vision*, pages 1811–1818. IEEE, 2009. 2, 5

[28] X. Wang, M. Salzmann, F. Wang, and J. Zhao. Template-free 3d reconstruction of poorly-textured nonrigid surfaces. In *European Conference on Computer Vision*, pages 648–663, 2016. 2

[29] R. Yu, C. Russell, N. D. Campbell, and L. Agapito. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *IEEE International Conference on Computer Vision*, pages 918–926. IEEE, 2015. 2