



4D Temporally Coherent Multi-Person Semantic Reconstruction and Segmentation

Armin Mustafa¹ · Chris Russell¹ · Adrian Hilton¹

Received: 11 December 2019 / Accepted: 2 February 2022 / Published online: 28 April 2022
© The Author(s) 2022

Abstract

We introduce the first approach to solve the challenging problem of automatic 4D visual scene understanding for complex dynamic scenes with multiple interacting people from multi-view video. Our approach simultaneously estimates a detailed model that includes a per-pixel semantically and temporally coherent reconstruction, together with instance-level segmentation exploiting photo-consistency, semantic and motion information. We further leverage recent advances in 3D pose estimation to constrain the joint semantic instance segmentation and 4D temporally coherent reconstruction. This enables per person semantic instance segmentation of multiple interacting people in complex dynamic scenes. Extensive evaluation of the joint visual scene understanding framework against state-of-the-art methods on challenging indoor and outdoor sequences demonstrates a significant ($\approx 40\%$) improvement in semantic segmentation, reconstruction and scene flow accuracy. In addition to the evaluation on several indoor and outdoor scenes, the proposed joint 4D scene understanding framework is applied to challenging outdoor sports scenes in the wild captured with manually operated wide-baseline broadcast cameras.

Keywords Dynamic 4D reconstruction · Segmentation · Scene understanding · Sports

1 Introduction

With the advent of autonomous vehicles and rising demand for immersive content in augmented and virtual reality, understanding dynamic scenes with multiple interacting people has become increasingly important. Understanding refers to reconstructing, segmenting and temporally aligning the reconstructions over time. In this paper we propose a framework for 4D dynamic scene understanding with multiple people in the scene from multi-view videos to address this demand. By “4D Scene understanding” we refer to a unified framework that describes: 3D modelling; motion/flow estimation; and semantic instance segmentation on a per frame basis for an entire sequence. Recent advances in pose esti-

mation (Cao et al. 2017; Tome et al. 2017) and recognition (He et al. 2017; Xie et al. 2016; Chen et al. 2016) using deep learning have achieved excellent performance for complex images. We exploit these advances to obtain 3D human-pose and an initial semantic instance segmentation from multiple view videos to bootstrap the detailed 4D understanding and modelling of complex dynamic scenes captured with multiple static or moving cameras (see Fig. 1). Joint 4D reconstruction allows us to understand how people move and interact, giving contextual information in general scenes.

Existing multi-task methods for scene understanding perform per frame joint reconstruction and semantic instance segmentation from a single image (Kendall et al. 2017), showing that joint estimation of both reconstruction and segmentation can improve the quality of each task. Other methods have fused semantic segmentation with reconstruction (Mustafa and Hilton 2017) or flow estimation (Sevilla-Lara et al. 2016) demonstrating significant improvement in both semantic segmentation and reconstruction/scene flow. Hence, we exploit the advantages of performing joint optimization in this paper to understand dynamic scenes with multiple interacting people by simultaneous reconstruction, flow and segmentation estimation from multiple view video.

Communicated by Mei Chen.

✉ Armin Mustafa
a.mustafa@surrey.ac.uk
Chris Russell
chris.russell@surrey.ac.uk
Adrian Hilton
a.hilton@surrey.ac.uk

¹ Centre for Vision, Speech and Signal Processing (CVSSP),
University of Surrey, Guildford GU27XH, UK



Fig. 1 Joint 4D semantic instance segmentation and reconstruction exploiting 3D human-pose of interacting people in dynamic scenes. Shades of pink in segmentation represents instances of people. Colour assigned to reconstruction of frame 80 is reliably propagated to frame 120 using proposed temporal coherence (Color figure online)

The first category of methods in joint estimation for dynamic scenes generate segmentation and reconstruction from multi-view video (Mustafa et al. 2016) and monocular video (Floros and Leibe 2016; Larsen et al. 2007) without any output scene flow estimate. The second category of methods segment and estimates motion in 2D (Sevilla-Lara et al. 2016), or give spatio-temporal aligned segmentation (Chiu and Fritz 2013; Luo et al. 2015; Djelouah et al. 2016) from multiple views without inferring the shape of the objects. The third category of methods in 4D temporally coherent reconstruction either align meshes using correspondence information between consecutive frames (Zanfir and Sminchisescu 2015) or extract the scene flow by estimating the pairwise surface correspondence between reconstructions at successive frames (Wedel et al. 2011; Basha et al. 2010). However methods in all of these three categories do not exploit semantic information of the scene, as seen in Table 1. The fourth category of joint estimation methods exploit semantic information by introducing joint semantic segmentation and reconstruction for general dynamic scenes (Hane et al. 2016; Xie et al. 2016; Kundu et al. 2014; Ulusoy et al. 2017; Mustafa and Hilton 2017) and street scenes (Engelmann et al. 2016; Vineet et al. 2015). However these methods give per-frame semantic segmentation and reconstruction with no motion estimate. This leads to unaligned geometry, pixel level incoherence in both segmentation and reconstruction for dynamic sequences and does not work for complex scenes with multiple interacting people such as stadium sports. Other methods for semantic video segmentation classify objects exploiting spatio-temporal semantic information (Tsai et al. 2016; Luo et al. 2015; Chiu and Fritz 2013) but do not perform reconstruction. Hence none of the existing methods in the literature give 4D temporally coherent reconstruction and

instance segmentation on complex dynamic scenes with multiple interacting people. We address this gap in the literature by proposing a novel framework for joint multi-view 4D temporally coherent reconstruction, semantic instance segmentation and flow estimation for general dynamic scenes automatically without any manual intervention.

Methods in the literature have exploited human-pose information to improve results in semantic segmentation (Xia et al. 2017) and reconstruction (Huang et al. 2017). However existing joint estimation/ optimization methods for dynamic scenes (with multiple people) do not exploit human-pose information often detecting interacting people as a single object (Mustafa and Hilton 2017). Table 1 shows a comparison between the tasks performed by state-of-the-art methods. In addition to performing joint multi-person 4D temporally coherent reconstruction and semantic instance segmentation, we exploit advances in 3D human-pose estimation to propose the first approach for 4D (3D in time) human-pose based scene understanding of general dynamic scenes with multiple interacting dynamic objects (people) with complex non-rigid motion. 3D human-pose estimation makes full use of multi-view information and is used as a prior to constrain the shape, segmentation and motion in space and time in the joint scene understanding estimation to improve the results on challenging scenes in the wild including sports. Sports reconstruction presents a challenging problem with a small number (6–12) of independently manually operated panning and zooming broadcast cameras, sparsely located around the stadium to cover a large area with multiple players. This results in multiple view wide-baseline capture at different resolutions with motion blur due to player and camera motion. The framework enables high-quality reconstruction and semantic instance segmentation for multi-player occlusions in sports captured from wide-baseline moving cameras, overcoming limitation of previous multiple view reconstruction algorithms. The contributions of the paper are:

- 4D scene understanding for multiple interacting people in dynamic scenes from multi-view video.
- Joint instance-level segmentation, temporally coherent reconstruction and scene flow with human-pose priors.
- Robust 4D temporal coherence and per-pixel semantic coherence for dynamic scenes containing interactions.
- An extensive performance evaluation against 15 state-of-the-art methods demonstrating improved semantic segmentation, reconstruction and motion estimation.

This paper is an extended version of ICCV 2019 paper (Mustafa et al. 2019), which includes detailed information about the method, ablation studies, performance evaluation on multi-person benchmarks and results on challenging sports datasets.

2 Related Work

Humans extract rich information from the world around them, and for autonomous machines (self-driving cars, robots) to navigate safely around people, machines must be able to perceive the scenes as humans do. Scene understanding refers to the simultaneous extraction of 3D reconstruction, semantic information of objects and motion estimation, illustrated in Fig. 1. Scene understanding has become increasingly popular in the past 5 years and it brings machines one step closer to understand the real to human level, machine perception of the real-world. This section provides a review of recent notable methods for scene understanding tasks (individually and joint) for single-view (Table 2) and multi-view (Table 3) video.

2.1 Scene Understanding for Single-View Video-Table 2

Semantic segmentation Fully Convolutional Network (Shelhamer et al. 2015) extract deep per-pixel CNN features followed by the classification of each pixel in the image for semantic segmentation. Deeplabv3+ (Chen and Zhu 2018) improved results by using an encoder-decoder architecture with Resnet and atrous spatial pyramid pooling to classify each pixel. Mask RCNN gives semantic instance segmentation on challenging scenes (He et al. 2017) by using a Region Proposal Network that shares full-image convolutional features with the detection network and adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. An improved mask is predicted in Li et al. (2020) by effectively combining instance level information with semantic information with lower level fine-granularity. A flow alignment module is proposed in Chen et al. (2020) to learn Semantic Flow between feature maps of adjacent levels, and broadcast high-level features to high resolution features for improved semantic segmentation.

Depth estimation Fusion4D (Dou et al. 2016) introduced a method for real-time online reconstruction for a video sequence from RGB image, depth and high-quality segmentation as input (Dou et al. 2016) and is restricted to relative simple indoor scenes. The proposed method only need RGB images as input and works for crowded indoor and outdoor scenes with multiple people. A single multi-scale convolution network architecture was proposed for depth prediction and semantic labelling in Eigen and Fergus (2015). Unsupervised monocular depth estimation was performed by learning depth from a stereo pair in Godard et al. (2017). Traditional SFM was used in a self-supervised method (Klodt and Vedaldi 2018) to learn and predict depth from monocular video. A recent method (Wang et al. 2020) decomposes a scene into semantic segments and then predicts a scale and shift invariant depth map for each semantic segment in a canonical space from a single image (Wang et al. 2020). An optimization based depth estimation method was proposed in Rossi et al. (2020) exploiting the underlying piece-wise planarity of scenes and other depth estimation method (Rodriguez and Mikolajczyk 2020) from a single image. This bridges the domain gap by leveraging semantic predictions and low-level edge features to provide guidance for the target domain.

Motion estimation One of the first methods to construct CNNs capable of estimating optical flow as a supervised learning task was introduced in Dosovitskiy et al. (2015). CNN was proposed to estimate human flow fields specifically from pairs of images in Ranjan et al. (2018). Deep Epipolar Flow (Zhong et al. 2019) was used for unsupervised flow estimation introducing global epipolar constraints into network learning. A transformer encoder-decoder type network was proposed along with a memory-based dictionary, which

Table 1 Comparison of tasks state-of-the-art methods are solving against the proposed method

	Semantic	Segment	Instance	3D	Motion	Pose
Kendall et al. (2017); Ulusoy et al. (2017); Engelmann et al. (2016)	✓	✓	✓	✓	×	×
Sevilla-Lara et al. (2016)	✓	✓	✓	×	✓	×
Mustafa and Hilton (2017); Hane et al. (2016); Kundu et al. (2014)	✓	✓	×	✓	×	×
Xia et al. (2017)	✓	✓	✓	×	×	✓
Huang et al. (2017)	×	×	×	✓	✓	✓
Floros and Leibe (2016)	✓	✓	×	✓	✓	×
Larsen et al. (2007); Roussos et al. (2012)	×	×	✓	✓	✓	×
Mustafa et al. (2016)	×	✓	×	✓	✓	×
Tsai et al. (2016); Luo et al. (2015); Chiu and Fritz (2013)	✓	✓	×	×	✓	×
Proposed	✓	✓	✓	✓	✓	✓

Table 2 Methods for scene understanding tasks for single-view video

Semantics	Depth
Badrinarayanan et al. PAMI 2015 Shelhamer et al. (2015)	Eigen & Fergus ICCV 2015 Eigen and Fergus (2015)
Chen et al. ECCV 2018 Chen and Zhu (2018)	Godard et al. CVPR 2017 Godard et al. (2017)
He et al. CVPR 2018 He et al. (2017)	Klodt and Vedaldi ECCV 2018 Klodt and Vedaldi (2018)
Motion	Joint
Dosovitskiy et al. ICCV 2015 Dosovitskiy et al. (2015)	Kendall et al. CVPR 2018 Kendall et al. (2018)
Ranjan et al. BMVC 2018 Ranjan et al. (2018)	Ranjan et al. CVPR 2019 Ranjan et al. (2019)
Zhong et al. CVPR 2019 Zhong et al. (2019)	Chen et al. CVPR 2019 Chen et al. (2019)

aims to preserve the global motion patterns in training data to improve flow estimation for humans (Cai et al. 2020).

Scene understanding Simultaneous semantic instance segmentation and depth estimation was proposed (Kendall et al. 2018) from a single view video exploiting uncertainties in multi-task learning framework. Unsupervised methods for joint depth, flow and motion segmentation; and joint depth and semantic segmentation from a monocular video were proposed in Ranjan et al. (2019) and Chen et al. (2019) respectively. A recent method (Zeng and Gevers 2020) reconstructs and semantically segments 3D indoor scenes from a single panorama image, however this method only works for static scenes.

All of these method either perform a single task (reconstruction, segmentation or flow estimation) or the joint scene understanding methods work for a single view video only. However the proposed method solves multiple tasks together giving a full scene understanding from multiple view videos by jointly estimating semantic instance segmentation, depth and motion exploiting human pose information.

2.2 Scene Understanding for Multi-View Videos-Table 3

Segmentation Co-temporal multi-view segmentation was proposed in Djelouah et al. (2016) with no semantic infor-

mation. A multi-view semantic segmentation network was designed in Guerry et al. (2017) for the consistent labelling of static scenes. Semantic information across space and time was used in a joint framework (Mustafa and Hilton 2017) for multi-view semantic reconstruction of dynamic scenes. Abhijit et al. (2020) fuses features from multiple per view predictions on 3D mesh vertices to predict mesh semantic segmentation labels for 3D semantic segmentation, but this method works only for static scenes.

Reconstruction Temporally coherent reconstruction was obtained in Mustafa et al. (2016) from multi-view videos. An end-to-end deep learning architecture was introduced in Yao et al. (2018) for depth map inference from multi-view images. Trager et al. (2019) defined a new characterization of multi-view geometry by proposing a coordinate-free description of Carlsson-Weinshall duality. A recent approach estimates high fidelity 3D human pose and volumetric reconstruction from multiple camera views by using a dual loss in a generative adversarial network (Gilbert et al. 2020). Another deep learning approach (Bi et al. 2020) reconstructs scene appearance from unstructured images captured under collocated point lighting using reflectance volumes. However both of these approaches give per frame reconstruction which are unaligned in time without any semantic information.

Motion estimation Limited methods have been proposed in multi-view motion estimation. The first-ever method to

Table 3 Methods for scene understanding tasks for multi-view video

Semantics	Depth
Djelouah et al. 3DV 2016 Djelouah et al. (2016)	Mustafa et al. CVPR 2016 Mustafa et al. (2016)
Guerry et al. ICCV 2017 Guerry et al. (2017)	Yao et al. ECCV 2018 Yao et al. (2018)
Mustafa et al. CVPR 2017 Mustafa and Hilton (2017)	Trager et al. CVPR 2019 Trager et al. (2019)
Motion	Joint
Szeliski et al. CVPR 1999 Szeliski (1999)	Hane et al. PAMI 2016 Hane et al. (2016)
Mustafa et al. 3DV 2017 Mustafa et al. (2017)	Dai and Niesser ECCV 2018 Dai and Nießner (2018)
Lai et al. CVPR 2019 Lai et al. (2019)	Mustafa et al. ICCV 2019 Mustafa et al. (2019)

estimate motion and stereo from multi-view images was proposed by Szeliski (1999). Scene flow was obtained from multiple light-field images in Mustafa et al. (2017) exploiting epipolar constraints. Recently a network to jointly learn spatio-temporal correspondence for stereo matching and flow estimation was introduced (Lai et al. 2019).

Scene understanding Multi-view scene understanding for static scenes was introduced in Hane et al. (2016) through a joint formulation of depth and semantics. 3D semantic scene segmentation of indoor RGB-D environments was performed in Dai and Nießner (2018) using a joint 3D multi-view prediction network.

All of the above methods either focus on a single task on segmentation, reconstruction and flow estimation exploiting multiple views or work for static scenes giving per frame reconstructions and segmentation unaligned in time. The proposed method performs joint semantic instance segmentation, 4D reconstruction and motion estimation of dynamic scenes with multiple interacting people in the scenes addressing the gap in the literature for full scene understanding from multi-view videos. Also most of the methods explained above use deep learning based approach to solve the reconstruction, flow estimation and segmentation, but the proposed method is an optimization framework which does not need any ground-truth data for training or require no manual intervention for 4D temporally coherent semantic reconstruction of dynamic scenes.

3 Joint 4D Dynamic Scene Understanding

Overview:

This section describes our approach to joint 4D scene understanding, with different stages shown in Fig. 2. The overview of the proposed method is as follows:

- **Input** The input to the joint optimisation is multi-view video. The proposed algorithm requires synchronised cameras, however it works for all datasets the datasets which are either synchronised through audio information (Hasler et al. 2009) or time code generator. Slight errors that are introduced through audio synchronisation are handled well with the proposed method. More details on the datasets are given in the Experiments section.
- **Initial Semantic Instance Segmentation - Sect. 3.1:** Initial semantic labels are estimated for each pixel in the image per-view using state-of-the-art semantic instance segmentation (He et al. 2017). An initial reconstruction is obtained for each object in the scene combining the initial semantic instance segmentation with the sparse reconstruction (Mustafa and Hilton 2017). Semantic information for each view is combined

with sparse 3D feature correspondence between views to obtain an initial semantic 3D reconstruction. This initial reconstruction is inaccurate due to the errors in the per-view semantic information which is combined across views.

- **Key-frame Detection-Sect. 3.2:** To achieve stable long-term 4D understanding a set of unique key-frames are detected exploiting multi-view information for final temporally coherent 4D reconstruction, key-frames are detected for the entire sequence exploiting shape, 3D pose and semantic information.
- **3D Human Pose Estimation and Estimation of Sparse Temporal Tracks - Section 3.3:** 3D human pose is estimated for each person in the scene to constraint the joint per-view optimization to estimate semantic instance segmentation, motion and 3D reconstruction. Sparse temporal feature tracks are obtained per view between key-frames to initialise the joint estimation. This allows robust 4D understanding in the presence of large non-rigid motion between frames.
- **Joint Estimation of Semantic Instance and Shape - Section 3.3:** The initial reconstruction and semantic instance segmentation is refined for each object instance per-view through novel joint optimisation of segmentation, shape, and motion constrained by 3D human-pose. Key-frames are used to introduce robust temporal coherence in the joint estimation across long-sequences with large non-rigid deformation. Per-view information is merged into a single 3D model using Poisson surface reconstruction (Kazhdan et al. 2006).
- **4D Scene Understanding - Section 3.4:** The process is repeated for the entire sequence and is combined across views and in time to obtain temporally coherent 4D semantic reconstruction for dynamic scenes. Depth, motion and semantic instance segmentation is combined across views between frames for 4D temporally coherent reconstruction and dense per-pixel semantic coherence for final 4D understanding of scenes. Figure 2 shows segmentation, reconstruction and tracking of both static and dynamic objects in the scene.

3.1 Initial Semantic Instance Segmentation

Existing methods for semantic segmentation do not give instance level segmentation of the scene. Previous approaches for semantic segmentation either segment the image followed by a per-segment object category classification (Mostajabi et al. 2015; Gupta et al. 2014), which can lead to propagation of errors from segmentation or give deep per-pixel CNN features followed by per-pixel classification in the image (Farabet et al. 2013; Hariharan et al. 2015), leading to segmentations with fuzzy boundaries and spatially disjointed regions or predict semantic segmentation from

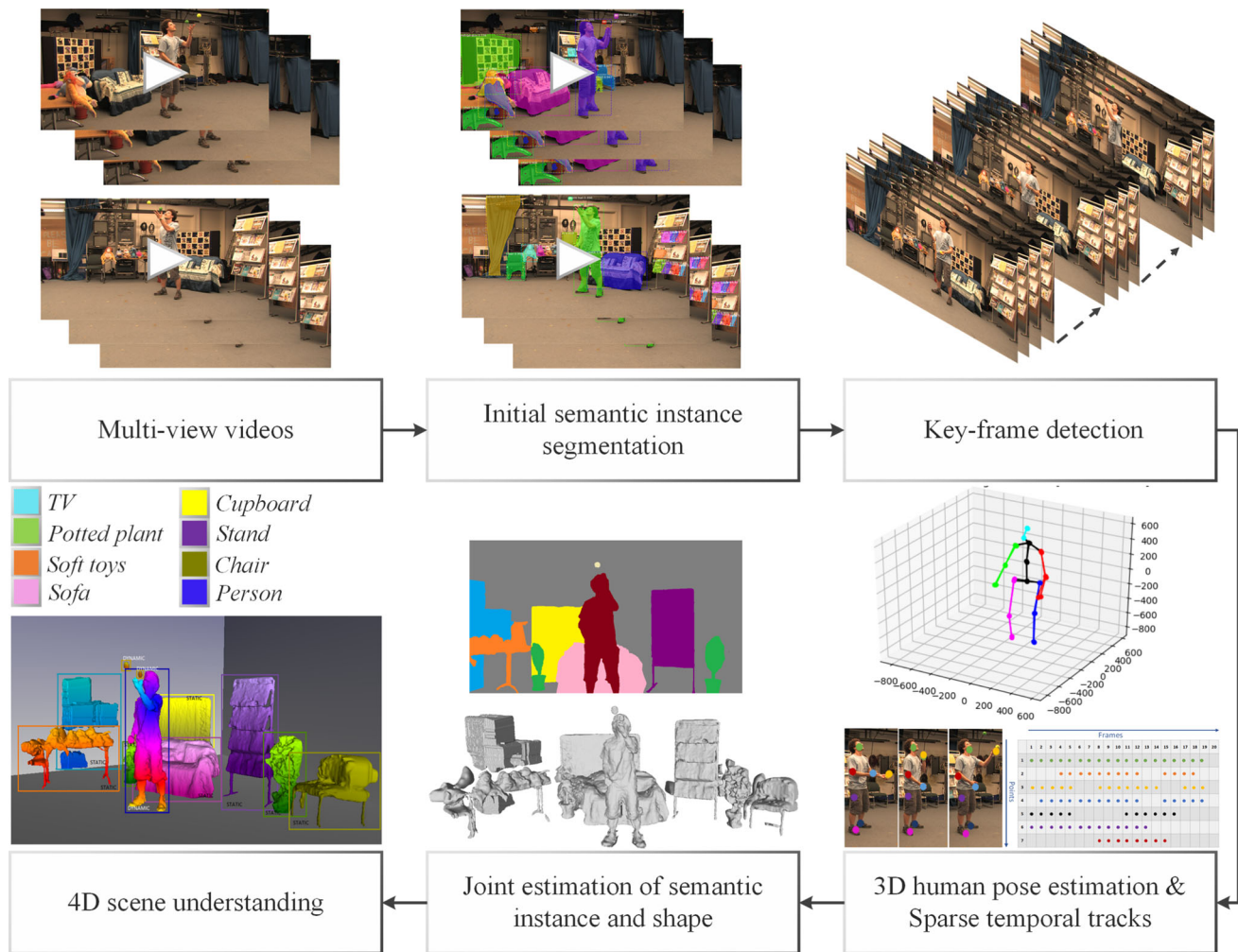


Fig. 2 4D dynamic scene understanding framework for multiple interacting people in the scene from multi-view video

raw pixels (Shelhamer et al. 2015) followed by conditional random fields (Kundu et al. 2016; Zheng et al. 2015). To address these issues, methods proposed semantic segmentation prediction from the raw pixels (Shelhamer et al. 2015) followed by conditional random fields (Kundu et al. 2016; Zheng et al. 2015) to improve segmentation. However none of these methods give instance segmentation of the scene. A recent state-of-the-art method (He et al. 2017) gives a good estimate of initial semantic instance segmentation masks (probability estimates of various classes at each pixel) from complex single images. We employ this state-of-the-art semantic instance segmentation method (He et al. 2017) to predict initial semantic unary potentials using pre-trained parameters on MS-COCO (Lin et al. 2014) and PASCAL VOC12 (Everingham et al. 2012) for each view. However this pre-segmentation can be replaced with any state-of-the-art methods as the framework refines the semantic labels and it is not sensitive to errors in the initialisation. Poor quality of initial semantics will increase computation cost.

3.2 Key-Frame Detection

Previous work (Newcombe et al. 2015; Mustafa et al. 2017) showed that sparse key-frames allow robust long-term correspondence for 4D reconstruction. In this work we introduce the additional use of pose in the detection and sparse temporal feature correspondence across key-frames to prevent the accumulation of errors in long sequences. Key-frame detection is used to improve the long term temporal coherence in the proposed joint semantic instance segmentation and 4D reconstruction. The 3D meshes are aligned for frames in between two key-frames K_i and K_{i+1} and between key-frames N_K to obtain full 4D scene reconstruction for the sequence. N_K is the total number of key-frames in the sequence. 4D scene alignment between key-frames is explained in Sect. 3.4.

Key-frames are detected exploiting similarity between frames, 3D pose and shape. Distance between 2 frames is

also taken into account to estimate key-frames. All the metrics used to estimate key-frames are defined below:

3.2.1 Sparse Correspondence Metric ($M_{i,j}^c$)

This measures appearance similarity between frames for each object, defined as the ratio of the number of sparse temporal correspondences Q to the total number of features R . SFD features are detected for each temporal frame and brute force matching (Mustafa et al. 2019) is performed to estimate the correspondences. The term is defined below:

$$M_{i,j}^c = \frac{2Q_{i,j}^c}{R_i^c + R_j^c}$$

where $Q_{i,j}^c$ are the number of sparse temporal correspondences between frame i and j for view c , R_i^c are the number of total features for frame i , view c and R_j^c are the number of total features for frame j , view c .

3.2.2 3D Pose Metric ($P_{i,j}^c$)

3D human poses are estimated for each time frame (Tomè et. 2018) and this metric measures the distance between the regularised human-pose:

$$P_{i,j}^c = \frac{\|P_i - P_j\|}{P_{max}^c}$$

where $j > i$ and P_{max}^c is the maximum change of pose between frames for view c . P_{max}^c is calculated by measuring the distance between regularised poses for 20 frames and choosing the maximum value. This term ensures that the distance of poses between key-frames is limited.

3.2.3 Semantic Metric ($L_{i,j}^c$)

This term checks the semantic similarity between two frames by comparing the number of pixels with the same semantic labels. An affine warp (Evangelidis and Psarakis 2008) is used to align semantic regions to measure semantic similarity between two frames. The metric is defined as the ratio of the number of pixels with the same class label $z_{i,j}^c$ to the pixels in the segmented region $y_{i,j}^c$:

$$L_{i,j}^c = \frac{z_{i,j}^c}{y_{i,j}^c}$$

3.2.4 Distance Metric ($D_{i,j}^c$)

This metric measures the distance between frames and makes sure that the distance between two key-frames is not large as

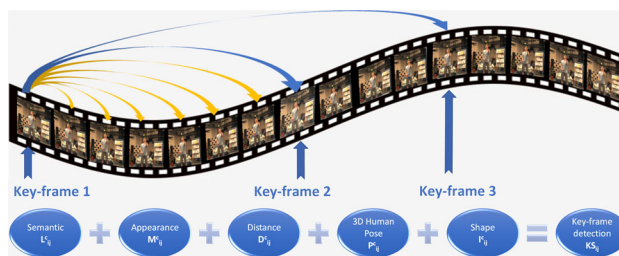


Fig. 3 An illustration of key-frame detection and matching across a short sequence for stable long-term temporal coherence

it will introduce errors in the final reconstruction and segmentation. The term is defined as:

$$D_{i,j}^c = \frac{j - i}{D_{max}^c}$$

where $j > i$ and D_{max}^c is the maximum number of frames between key-frames for view c . This term ensures that the distance between two key-frames does not exceed D_{max}^c . This is set to 100 throughout this work.

3.2.5 Shape Metric ($I_{i,j}^c$)

It is defined as the ratio of the intersection of the aligned segmentation or silhouette (Evangelidis and Psarakis 2008) (h) to the union of the area (a):

$$I_{i,j}^c = \frac{h_{i,j}^c}{a_{i,j}^c}$$

This give a measure of shape or silhouette overlap for an object between frames i and j for view c . The silhouette are projection of initial coarse 3D reconstruction in each view.

All these metrics defined above are combined to estimate keyframes using the *Key-frame similarity metric*, which is defined as:

$$K S_{i,j} = 1 - \frac{1}{5N_v} \sum_{c=1}^{N_v} (M_{i,j}^c + L_{i,j}^c + D_{i,j}^c + P_{i,j}^c + I_{i,j}^c) \tag{1}$$

Key-frame detection exploits sparse correspondence ($M_{i,j}^c$), pose ($P_{i,j}^c$), shape ($I_{i,j}^c$), semantic label ($L_{i,j}^c$) and distance ($D_{i,j}^c$) information across views N_v between frame i and j for each object in view c , to improve the long-term temporal coherence of the proposed method, using similar frames across the sequence, illustrated in Fig. 3. All frames with similarity $K S_{i,j} > 0.75$ in a sequence are selected as key-frames defined as $K = \{K^1, K^2, \dots, K^{N_k}\}$ where N_k is the number of key-frames. We also define another term N_f^i , which is the number of frames between K_i and K_{i+1} .

3.3 Joint Per-View Optimisation

Sparse reconstruction is obtained for each frame from multiple views using Colmap (Schönberger and Frahm 2016; Schönberger et al. 2016). The multi-view cameras are synchronised either directly or in post-processing using the audio information. The sparse point cloud is clustered in 3D (Rusu 2009) with each cluster representing a unique foreground object. The per-view semantic instance segmentation obtained in the previous step is combined across views with sparse reconstruction to obtain an initial coarse reconstruction \mathcal{R} for each person in the frame, where i represents different number of objects for each frame (Mustafa and Hilton 2017). This initial semantic coarse reconstruction \mathcal{R} is refined through a joint scene understanding optimization. The optimization is performed per-view to obtain depth, semantic segmentation and flow for each view.

3.3.1 Spatio-Temporal Coherence in the Optimisation

Constraints are applied on the spatial and temporal neighborhood to enforce consistency in the appearance, semantic label, 3D human pose and motion across views and time.

Spatial coherence Multi-view spatial coherence is enforced in the optimisation such that the motion, shape, appearance, 3D pose and class labels are consistent across views using an 8-connected spatial neighbourhood ψ_S for each camera view such that the set of pixel pairs $(p; q)$ belong to the same frame.

Temporal coherence Temporal coherence is enforced in the joint optimisation by enforcing coherence across key-frames (Sect. 3.2) to handle large non-rigid motion and to reduce errors in sequential alignment for long sequences in the 4D scene understanding. Sparse temporal feature correspondences are used for key-frame detection and robust initialisation of the joint optimisation. They measure the similarity between frames and unlike optical flow are robust to large motions and visual ambiguity. To achieve robust temporal coherence in the 4D scene understanding framework for large non-rigid motion, sparse temporal feature correspondences in 3D are obtained across the sequence.

The temporal neighbourhood is defined for each frame between its respective key-frames. Sparse temporal correspondence tracks define the temporal neighbourhood $\psi_T = \{(p, q) \mid q = p + e_{i,j}\}$; where $j = \{t - 1, t + 1\}$, $e_{i,j}$ is the displacement vector from image i to j , p and q are pixels in the image.

3.3.2 Joint Optimisation

The goal of the joint estimation is to refine initial semantic instance segmentation and reconstruction by assigning a label from a set of classes obtained from initial seman-

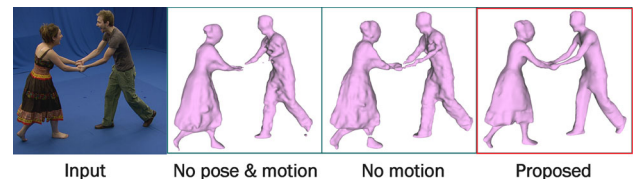


Fig. 4 Comparison of reconstruction without pose and motion in the optimisation framework, proposed result is best

tic instance segmentation $\mathcal{L} = \{l_1, \dots, l_{|\mathcal{L}|}\}$ ($|\mathcal{L}|$ is the total number of classes), a depth value from a set of depth values $\mathcal{D} = \{d_1, \dots, d_{|\mathcal{D}|-1}, \mathcal{U}\}$ (each depth value is sampled on the ray from camera and \mathcal{U} is an unknown depth value to handle occlusions), and a motion flow field $\mathcal{M} = \{m_1, \dots, m_{|\mathcal{M}|}\}$ simultaneously for the region \mathcal{R} of each object per view. $|\mathcal{M}|$ is the set of pre-defined discrete flow-fields for pixel $p = (x, y)$ in image I by $m = (\delta x, \delta y)$ in time and for each view.

Joint semantic instance segmentation, reconstruction and motion estimation is achieved by global optimisation of a cost function over unary E_{unary} and pairwise E_{pair} terms, defined as:

$$\begin{aligned} E(l, d, m) &= E_{unary}(l, d, m) + E_{pair}(l, d, m) \\ E_{unary} &= \lambda_d E_d(d) + \lambda_a E_a(l) + \lambda_{sem} E_{sem}(l) + \lambda_f E_f(m) \\ E_{pair} &= \lambda_s E_s(l, d) + \lambda_c E_c(l) + \lambda_r E_r(l, m) + \lambda_p E_p(l, d, m) \end{aligned} \quad (2)$$

where, d is the depth, l is the class label, and m is the motion at pixel p . Novel terms are introduced for flow E_f , motion regularisation E_r and human-pose E_p costs, explained in Sects. 3.3.4 and 3.3.3 respectively. Results of the joint optimisation with and without pose (E_p) and motion (E_f, E_r) information are presented in Fig. 4, showing the improvement in results. Ablation analysis on individual costs in Sect. 4 demonstrates the improvement in performance with the novel introduction of motion and pose constraints in the joint optimisation. Standard unary terms for depth (E_d), semantic (E_{sem}), and appearance (E_a) costs, explained in Sect. 3.3.6. Standard pairwise terms colour contrast (E_c) is used to assist segmentation and smoothness (E_s) cost ensures that depth varies smoothly in a neighbourhood.

3.3.3 Human-Pose Constraints $E_p(l, d, m)$

We use 3D human-pose to constrain joint optimisation and improve the flow, reconstruction and instance segmentation, in both 2D and 3D for dynamic scenes with multiple interacting people (see Fig. 1). 3D human-pose is used as it is consistent across multiple views unlike 2D human-pose. A state-of-the-art method for 3D human-pose estimation from

multiple cameras (Tomè et. 2018) is used in the paper. Previous work on 3D pose estimation (Tome et al. 2017) iteratively builds a 3D model of human-pose consistent with 2D estimates of joint locations and prior knowledge of natural body pose. In Tomè et. (2018), multiple cameras are used when estimating the 3D model; this then feeds back into new estimates of the 2D joint locations in each image. This approach allows us to take full advantage of 3D estimates of pose, consistent across all cameras when finding fine grained 2D correspondences between images, and leading to more life-like, vivid human reconstructions.

Initial semantic reconstruction is updated if the 3D pose of the person lies outside the region \mathcal{R} by dilating the boundary to include the missing joints. This allows for more robust and complete reconstruction and segmentation. We use a standard set of 17 joints (Tomè et. 2018) defined as \mathcal{B} . A circle \mathcal{C}_i is placed around the joint position in 2D and a sphere \mathcal{S}_i is placed around the joint position in 3D based on the confidence map to identify the nearest neighbour vertices for every joint b_i .

$$E_p(l, d, m) = \sum_{b_i \in \mathcal{B}} \lambda_{2d} e_{2d}(l, m) + \lambda_{3d} e_{3d}(d) \tag{3}$$

$$e_{2d}(l, m) = e_{2d}^L(l) + e_{2d}^S(l) + e_{2d}^M(m)$$

$$e_{3d}(d) = e_{3d}^M(d) + e_{3d}^S(d), \text{ if } d_p \neq \mathcal{U} \text{ else } 0$$

where e_{2d} enforces pose constraint in 2D domain and e_{3d} enforces human pose constrain in 3D domain and λ_{2d} and λ_{3d} are weighting terms. e_{2d} comprises of semantic e_{2d}^S , motion e_{2d}^M and segmentation e_{2d}^L constraints and e_{3d} includes motion e_{3d}^M and semantic e_{3d}^S constraint.

3D shape term This term constrains the reconstruction in 3D such that the neighbourhood points around the joints do not move far from the respective joints, and is defined as:

$$e_{3d}^S(d) = \exp\left(-\frac{1}{|\sigma_{S_D}|} \sum_{\Phi(p) \in \mathcal{S}_i} \|O\|_F^2\right)$$

where $\Phi(p)$ is the 3D projection of pixel p . The Frobenius norm $\|O\|_F = \|\left[\Phi(p) \ b_i\right]\|_F$ is applied on the 3D points in all directions to obtain the ‘net’ motion at each pixel within \mathcal{S}_i (sphere around the joint position in 3D) and $\sigma_{S_D} = \left\langle \frac{\|O\|_F^2}{\vartheta_{\Phi(p), b_i}} \right\rangle$, with the operator $\langle \cdot \rangle$ denoting the mean computed in \mathcal{S}_i .

3D motion term This enforces as rigid as possible (Sorkine and Alexa 2007) constraints on 3D points in the neighbourhood of each joint b_i in space and time. An optimal rotation matrix R_i is estimated for each b_i by minimising the energy defined as:

$$e_{3d}^M(d) = \sum_{\Phi(p) \in \mathcal{S}_i} \left\| \left(b_i^{t+1} - \Phi(p)^{t+1} \right) - R_i \left(b_i^t - \Phi(p)^t \right) \right\|_2^2 + \lambda_{3d}^p \left\| p - e_{3d}^M \right\|_2^2$$

λ_{3d}^p is the weighing constant. This term ensures that each joint does not move too far away from the original position.

2D term 3D poses are back-projected in each view to constrain per view appearance (e_{2d}^L), semantic segmentation (e_{2d}^S) and motion estimation (e_{2d}^M) in 2D. If $p \in \mathcal{C}_i$,

$$e_{2d}^L(l) = \exp\left(-\sum_{p \in \psi_S} \sum_{p \in \psi_T} \frac{\|I(\Pi(b_i)) - I(p)\|^2}{|\sigma_{S_L}|}\right)$$

$$e_{2d}^S(l) = \exp\left(-\sum_{p \in \psi_S} \sum_{p \in \psi_T} \frac{\|\Pi(b_i) - p\|^2}{|\sigma_{S_S}|}\right)$$

$$e_{2d}^M(m) = \exp\left(-\sum_{p \in \psi_S} \sum_{k \in \psi_T} \frac{\|\vartheta_{p, \Pi(b_i^k)} - \vartheta_{p+m_p, \Pi(b_i^{k+1})}\|^2}{|\sigma_{S_M}|}\right)$$

where, Π is the back-projection of 3D poses to 2D, N_{pose} is the number of nearest neighbours, $\sigma_{S_L} = \left\langle \frac{\|\Pi(b_i) - q\|^2}{\vartheta_{\Pi(b_i), q}} \right\rangle$ and, σ_{S_S} and σ_{S_M} is defined similarly, and $\vartheta_{\Pi(b_i), q}$ is the Euclidean distance between pixel $\Pi(b_i)$ and q . Similarly other $\vartheta_{p, \Pi(b_i^k)}$ and $\vartheta_{p+m_p, \Pi(b_i^{k+1})}$ denotes the Euclidean distances between other pixels in 2D. $e_{2d}^L(l)$ and $e_{2d}^S(l)$ ensures that the pixels around projected 3D pose $\Pi(b_i)$ have the same semantic label and appearance across views (ψ_S) and time (ψ_T) thereby ensuring spatio-temporal appearance and semantic consistency respectively.

3.3.4 Motion Constraints- $E_f(m)$ and $E_r(l, m)$

Flow term This term is obtained by integrating the sum of three penalisers over the reference image domain inspired from (Tao et al. 2012), defined as:

$$E_f(p, m_p) = e_F^T(p, m_p) + e_F^V(p, m_p) + e_F^S(p, m_p)$$

where, $e_F^T(p, m_p) = \sum_{i=1}^{N_v} \|(I_i(p, t) - I_i(p + m_p, t + 1))\|^2$ penalises deviation from the brightness constancy assumption in a temporal neighbourhood for the same view; $e_F^V(p, m_p) = \sum_{t \in \psi_T} \sum_{i=2}^{N_v} \|(I_1(p, t) - I_i(p + m_p, t))\|^2$ penalises deviation in appearance from the brightness constancy assumption between the reference view and other views at other time instants; and $e_F^S(p, m_p) = 0$ if $p \in N$ otherwise ∞ which forces the flow to be close to nearby sparse temporal correspondences. $I_i(p, t)$ is the intensity at point p at time t in camera i . The flow vector m is located within a window from a sparse constraint at p and it forces

the flow to approximate the sparse 2D temporal correspondences.

Motion regularisation term This penalises the absolute difference of the flow field to enforce motion smoothness and handle occlusions in areas with low confidence (Tao et al. 2012).

$$E_r(l, m) = \sum_{p, q \in N_p} \|\Delta m\|^2 \lambda_r^L e_r^L(p, q, m_p, m_q, l_p, l_q) + \lambda_r^A e_r^A(p, q, m_p, m_q, l_p, l_q)$$

where $\Delta m = m_p - m_q$ and;

$$e_r^X = \begin{cases} \text{mean}_{q \in N_p} E_X(q, m_q) - \min_{q \in N_p} E_X(q, m_q), & \text{if } \forall_{l_p=l_q} \\ 0, & \text{otherwise} \end{cases}$$

We compute e_R^L (semantic regularisation) and e_R^A (appearance regularisation) as the minimum subtracted from the mean energy within the neighbourhood search window N_p for each pixel p . λ_r^L and λ_r^A are constants, computed empirically.

The motion term in the proposed framework is not tailored to human motion. Results are shown for human motion because of its higher complexity, which makes the method more generalizable to different types of motion. We can easily handle linear motion of rigid objects (like cars).

3.3.5 Long-term Temporal Coherence

Sparse temporal correspondences The sparse 3D points projected in all views are matched between frames N_f^i and key-frames across the sequence using nearest neighbour matching (Mustafa et al. 2019) followed by a symmetry test which employs forward and backward match consistency by performing two-way matching to remove the inconsistent correspondences. This gives sparse temporal feature correspondence tracks per frame for each object: $F_i^c = \{f_1^c, f_2^c, \dots, f_{R_i^c}^c\}$, where $c = 1$ to N_v . R_i^c are the 3D points visible at each frame i . Exhaustive matching is performed, such that each frame is matched to every other frame to handle appearance, reappearance and disappearance of points between frames.

Key-frame detection Features at view c frame i , F_i^c are matched to features at view c to frames $j = \{i + 1, \dots, N_f^i\}$ to give correspondences for all the frames N_f^i with key-frame K_i . The corresponding joint locations from the 3D pose are back-projected in each view and added to sparse temporal tracks in between key-frames. Any new point-tracks are added to the list of point tracks for key-frame K_i . More details on key-frame detection are provided in Sect. 3.2.

3.3.6 Unary Terms - $E_{unary}(l, d, m)$

Depth term This gives a measure of photo-consistency between views $E_d(d) = \sum_{p \in \psi_S} e_d(p, d_p)$, defined as:

$$e_d(p, d_p) = \begin{cases} M(p, q) = \sum_{i \in \mathcal{O}_k} m(p, q), & \text{if } d_p \neq \mathcal{U} \\ M_{\mathcal{U}}, & \text{if } d_p = \mathcal{U} \end{cases}$$

where $M_{\mathcal{U}}$ is the fixed cost of labelling pixel unknown and q denotes the projection of the hypothesised point P (3D point along the optical ray passing through pixel p located at a distance d_p from the camera) in an auxiliary camera. \mathcal{O}_k is the set of the k most photo-consistent pairs with reference camera and $m(p, q)$ is inspired from (Mustafa et al. 2016).

Appearance term This term is computed using the negative log likelihood (Boykov and Kolmogorov 2004) of the colour models (GMMs with 10 components) learned from the initial semantic mask in the temporal neighbourhood ψ_T and the foreground markers obtained from the sparse 3D features for the dynamic objects. It is defined as:

$$E_a(l) = \sum_{p \in \psi_T} \sum_{p \in \psi_S} -\log P(I_p | l_p)$$

where $P(I_p | l_p = l_i)$ denotes the probability of pixel p belonging to layer l_i .

Semantic term This term is based on the probability of the class labels at each pixel based on Chen et al. (2016), defined as:

$$E_{sem}(l) = \sum_{p \in \psi_T} \sum_{p \in \psi_S} -\log P_{sem}(I_p | l_p)$$

where $P_{sem}(I_p | l_p = l_i)$ denotes the probability of pixel p being in layer l_i in the reference image obtained from initial semantic instance segmentation (He et al. 2017).

3.3.7 Pairwise Terms - $E_{pair}(l, d, m)$

There are two pairwise terms in the joint per-view optimization - smoothness and contrast. These terms are inspired from Guillemaut and Hilton (2010), which includes a proof as to how these pairwise terms satisfy the regularity condition required for graph-cut optimisation via alpha-expansion (Boykov and Kolmogorov 2004).

Smoothness term This term ensures that depth labels vary smoothly within a neighbourhood and is defined as:

$$E_s(l, d) = \lambda_s^t \sum_{p, q \in \psi_T} e_s(l_p, d_p, l_q, d_q, d_{max}^t) + \lambda_s^s \sum_{p, q \in \psi_S} e_s(l_p, d_p, l_q, d_q, d_{max}^s)$$

$$e_s(l_p, d_p, l_q, d_q, d_{max}) = \begin{cases} \min(|d_p - d_q|, d_{max}), & \text{if } l_p = l_q \text{ and } d_p, d_q \neq \mathcal{U} \\ 0, & \text{if } l_p = l_q \text{ and } d_p, d_q = \mathcal{U} \\ d_{max}, & \text{otherwise} \end{cases}$$

where, d_{max}^s avoids over-penalising large discontinuities for spatial smoothness and is set to 50 times the size of the depth sampling step. d_{max}^t ensures smoothness in time over the temporal neighbourhood and is twice the value of d_{max}^s to allow large movement in the object.

Contrast term This term is defined as:

$$E_c(l) = \sum_{p,q \in \psi_T} e_c(p, q, l_p, l_q, \sigma_\alpha^t, \vartheta_{p,q}^t, \sigma_\beta^t) + \sum_{p,q \in \psi_S} e_c(p, q, l_p, l_q, \sigma_\alpha^s, \vartheta_{p,q}^s, \sigma_\beta^s)$$

$$e_c(p, q, l_p, l_q, \sigma_\alpha, \vartheta, \sigma_\beta) = \mu(l_p, l_q) \times \left(\lambda_{ca} e^{-\frac{\|B(p)-B(q)\|^2}{2(\sigma_\alpha)^2(\vartheta)^2}} + \lambda_{cl} e^{-\frac{\|L(p)-L(q)\|^2}{2(\sigma_\beta)^2}} \right)$$

where $\mu(l_p, l_q) = 1$ if $(l_p = l_q)$ otherwise 0 and $\vartheta_{p,q}$ is the euclidean distance between p and q . ‘Bilateral’ kernel B forces pixels with similar colour and position to have similar labels and the Gaussian kernel L enforces spatial smoothness, with $\sigma_\alpha = \left\langle \frac{\|B(p)-B(q)\|^2}{\vartheta_{p,q}^2} \right\rangle$ and σ_β controlling the scale of these kernels, where the operator $\langle \cdot \rangle$ denotes the mean computed across the neighbourhoods ψ_S and ψ_T for spatial and temporal contrast respectively.

The proposed joint optimization is inspired from previous work (Guillemaut and Hilton 2010) which perform joint segmentation and reconstruction to achieve a globally consistent solution by performing the joint optimization per-view and by initializing the reconstruction with a reliable visual hull which is obtained using per-view segmentation which is taken as input. In the proposed method we obtain a globally consistent solution by performing joint per-view optimization on a reliable initial coarse reconstruction which is obtained by combining semantic instance segmentation with sparse reconstruction. Global optimisation of Equation 2 is performed per-view over all terms simultaneously, subject to each pixel p in the region \mathcal{R} using the α -expansion algorithm by iterating through the set of labels in $\mathcal{L} \times \mathcal{D} \times \mathcal{M}$ Boykov et al. (2001). Each label $\mathcal{L}, \mathcal{D}, \mathcal{M}$ is initialised before: \mathcal{L} is initialised using the initial semantic segmentation obtained in Sect. 3.1; \mathcal{D} is initialised using the depth of the initial coarse reconstruction estimate \mathcal{R} , such that the each d_i is obtained by sampling the optical ray from the camera within the region \mathcal{R} . The ray is sampled by a factor of 50 to calculate each d_i as in Mustafa et al. (2016); and \mathcal{M} is initial-

ized using discrete flow fields as in Tao et al. (2012); Menze et al. (2015). Each iteration is solved by graph-cut using the min-cut/max-flow algorithm (Boykov and Kolmogorov 2004). Convergence is achieved in 7-8 iterations.

3.4 4D Scene Understanding

The final 4D scene model fuses the semantic instance segmentation, depth information and dense flow across views and in time between frames (N_f^i) and key-frames (K_i). The initial instance segmentation, human pose and motion information for each object is combined to obtain final instance segmentation of the scene. The per-view depth maps obtained by optimizing Equation 2 for each camera view are combined across views using Poisson surface reconstruction (Kazhdan et al. 2006) to obtain a mesh for each object in the scene. For sports sequence with large calibration errors (1-2 pixels) each view-dependent 2.5D foreground scene representation is converted into a regular mesh with vertices defined by image pixel locations. Vertex connectivity is decided based on the layer segmentation and thresholding of the angle separating the line segment connecting 3D surface points defined by pairs of neighbouring pixels and the optical ray passing through the midpoint of the pixel pair (a threshold of 80 deg is used). This allows pixel belonging to different layers or located at a depth discontinuity to be correctly converted into separate mesh components.

The 3D meshes for each object per frame are combined with per-view motion estimates obtained by optimizing Equation 2 to get 4D temporally coherent meshes for each person in the scene. The most consistent motion information from all views for each 3D point is used to estimate correspondences between two frames. This is combined with spatial semantic instance information to give per-pixel semantic and temporal coherence. Appearing, disappearing, and reappearing regions are handled by using the sparse temporal tracks and their respective motion estimate. The dense flow and semantic instance segmentation together with 3D models of each object in the scene gives the final 4D understanding of the scenes. Examples are shown in Figs. 1 and 5 on two datasets, where objects are coloured in one key-frame and colours are propagated reliably between frames and key-frames across the sequence for robust 4D scene modelling.

The proposed method handles multiple people, appearing, disappearing and re-appearing in the scene. The method labels and tracks all static and dynamic objects in the scene. Multiple people and objects are identified using the initial semantic instance segmentation together with the clustering of the sparse reconstruction at each time frame. Object tracking and re-appearance is handled using the sparse temporal feature tracks and proposed dense flow. Exhaustive matching between all frames enables object re-identification. The pose constraints are only used for the human class and for

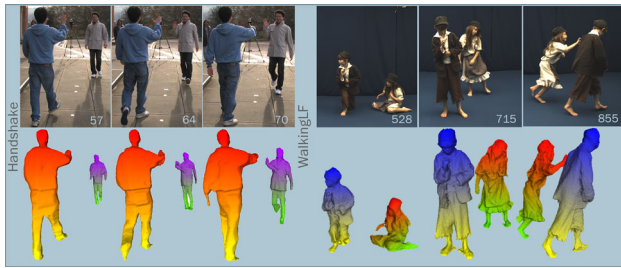


Fig. 5 Example of 4D scene reconstruction for one indoor and one outdoor dataset. The first frame is uniquely coloured for each dataset and the colours are propagated using proposed motion estimation



Fig. 6 Proposed semantic instance segmentation for Juggler2 outdoor dataset with three people

other classes $E - E_p$ is minimized allowing us to work with different objects. An example is shown for Juggler2 dataset in Fig. 6 with 3 humans and an object entering the scene.

4 Results and Evaluation

Joint semantic instance segmentation, reconstruction and flow estimation (Sect. 3) is evaluated quantitatively and qualitatively against 15 state-of-the-art methods on a variety of

publically available multi-view indoor and outdoor dynamic scene datasets, detailed in Table 4. Juggler2 and Magician datasets are synchronised using audio information and the rest of the datasets are synchronised using time code generator. A list of tasks performed by each state-of-the-art method is illustrated in Table 5.

Algorithm parameters listed in Table 6 are the same for all outdoor datasets, and for indoor datasets parameters depend on the number of cameras (N_v). Pairwise costs are constant $\lambda_p = 0.9$, $\lambda_c = \lambda_s = \lambda_r = 0.5$ for all datasets. The parameters defined in Table 6 cover all possibilities of datasets (indoor, outdoor, different number of views). The change in parameters does not drastically affect the performance. We used indoor parameters (row 2 in table) for outdoor dataset Juggler2. This reduces the reconstruction performance by 3%, segmentation by 4% and motion performance by 2%.

Due to the low resolution of objects in the sports dataset (people are only 30-70 pixel in height) and high the calibration errors (1-2 pixels), the parameters above could not be used for the proposed framework. The pairwise costs are as follows: $\lambda_p = 2$, $\lambda_c = \lambda_s = \lambda_r = 1.1$ and the unary costs are shown in the bottom row of Table 6.

4.1 Reconstruction Evaluation

The proposed approach is compared against state-of-the-art approaches for semantic co-segmentation and reconstruction (SCSR) Mustafa and Hilton (2017), piecewise scene flow (PRSM) Vogel et al. (2015), multi-view stereo (SMVS) Langguth et al. (2016), and deep learning based stereo approaches (LocalStereo) Taniai et al. (2018). Since PRSM

Table 4 Properties of all datasets: N_v is the number of views, L is the sequence length, KF gives number of key-frames, and Tracks gives the number of sparse temporal correspondence tracks averaged over the

Datasets	Resolution	N_v	Baseline	L	KF	Tracks
Handshake Kim et al. (2012)	1920 × 1080	8(all S)	15° – 30°	125	15	1945
Meetup Guillemaut and Hilton (2010)	1920 × 1080	16(all S)	25° – 35°	100	9	1341
Juggler2 Ballan et al. (2010)	960 × 544	6(all M)	15° – 45°	300	16	1278
Handstand Vlasic et al. (2008)	1600 × 1200	8(all S)	25° – 45°	174	12	1056
Rachel yyy (xxx)	3840 × 2160	16(all S)	20° – 30°	270	15	1978
Juggler1 yyy (xxx)	1920 × 1080	8(2 M)	15° – 30°	253	17	2083
Dance xxx (yyy)	780 × 582	8(all S)	35° – 45°	60	7	732
Magician Ballan et al. (2010)	960 × 544	6(all M)	15° – 45°	300	10	1312
Human3.6 Ionescu et al. (2014)	1000 × 1000	4(all S)	25° – 30°	250	14	994
MagicianLF Mustafa et al. (2017)	2048 × 2048	25(all S)	5° – 8°	350	5	1312
WalkLF Mustafa et al. (2017)	2048 × 2048	20(all S)	5° – 8°	221	7	1934
Football yyy (xxx)	1920 × 1080	8(4 S)	15° – 40°	200	8	1443
Wembley Guillemaut and Hilton (2010)	1920 × 1080	5(4 S)	25° – 45°	100	6	1885
Soccer yyy (xxx)	720 × 288	15(all S)	20° – 30°	100	5	1767

entire sequence for each object (S stands for static cameras and M for moving cameras)

Table 5 Illustration of tasks performed by state-of-the-art methods which are evaluated in Sect. 4 for segmentation, reconstruction and motion estimation against the proposed method

	Semantic	Segment	Instance	3D	Motion
SCV Tsai et al. (2016)	✓	✓	×	×	✓
SCSR Mustafa and Hilton (2017)	✓	✓	×	✓	×
JSR Guillemaut and Hilton (2010)	×	✓	×	✓	×
Dv3+ Chen and Zhu (2018)	✓	✓	×	×	×
MRCNN He et al. (2017)	✓	✓	✓	×	×
PSP Zhao et al. (2017)	✓	✓	×	×	×
CRF RNN Zheng et al. (2015)	✓	✓	×	×	×
Segnet Badrinarayanan et al. (2017)	✓	✓	×	×	×
RTSeg Siam et al. (2018)	✓	✓	×	×	×
PRSM Vogel et al. (2015)	×	×	×	✓	✓
LocalStereo Tani ai et al. (2018)	×	×	×	✓	×
SMVS Langguth et al. (2016)	×	×	×	✓	×
DCflow Xu et al. (2017)	×	×	×	×	✓
Deepflow Weinzaepfel et al. (2013)	×	×	×	×	✓
4DMatch Mustafa et al. (2016)	×	×	×	×	✓
Proposed	✓	✓	✓	✓	✓

Table 6 Parameters for all datasets. I is Indoor

	λ_d	λ_a	λ_{sem}	λ_f	λ_s^t/λ_s^s	$\lambda_{ca}/\lambda_{cl}$	λ_r^L/λ_r^C	$\lambda_{2d}/\lambda_{3d}$
Outdoor	1.2	0.5	0.5	0.4	1.0	5.0	0.6	7.5
$I, N_v < 6$	1.0	0.7	0.5	0.6	0.4	5.0	0.4	7.5
$I, 6 \leq N_v < 20$	1.0	0.7	0.2	0.4	0.4	5.0	0.4	5.0
$I, N_v \geq 20$	1.0	1.0	0.5	0.5	0.2	5.0	0.4	5.0
Sports	2.5	0.8	0.5	0.8	2.0	6.0	0.3	10

Vogel et al. (2015) and LocalStereo Tani ai et al. (2018) methods only work for 2 views/stereo pair of images, we divide the cameras in pairs and stereo is estimated for each pair.

The per-view depth maps for each camera view are combined across views using Poisson surface reconstruction (Kazhdan et al. 2006) to obtain a mesh for each object in the scene in a similar way to the proposed method. Default parameters are used to run both of these methods. The other state-of-the-art methods SMVS (Langguth et al. 2016) and SCSR (Mustafa and Hilton 2017) are multi-view approaches, where code available online is used to estimate the per-frame reconstruction using default parameters. Qualitative comparison with proposed method is shown in Fig. 7.

Pre-trained parameters were used for LocalStereo and per-view depth maps were fused using Poisson reconstruction. The quality of surface obtained using proposed method is improved compared to state-of-the-art methods. In contrast to previous approaches, limbs of people are reliably reconstructed because of the exploitation of human-pose and temporal information (motion) in the joint optimisation.

For quantitative comparison to state-of-the-art methods, we project the reconstruction onto different views and compute the projection errors shown in Table 7. A significant

improvement is obtained in projected surface completeness with the proposed approach. Further quantitative evaluation of the surface obtained using state-of-the-art methods is shown in Fig. 8. The reconstructions shown for Handstand are compared against the proposed method and the errors are colour coded, with red showing the maximum error.

4.2 Segmentation Evaluation

Our approach is evaluated against a variety of state-of-the-art multi-view (SCV Tsai et al. (2016), SCSR Mustafa and Hilton (2017), and JSR Guillemaut and Hilton (2010)) and single-view (Dv3+ Chen and Zhu (2018), MRCNN He et al. (2017), PSP Zhao et al. (2017), CRF RNN Zheng et al. (2015), and Segnet Badrinarayanan et al. (2017)) segmentation methods, shown in Fig. 9. For fair evaluation against single-view semantic segmentation methods, multi-view consistency is applied for segmentation estimated from each view to obtain multi-view consistent semantic segmentation using dense multi-view correspondence. Colour and visualizations in the results are kept from the original papers and default parameters are used for state-of-the-art methods.

Table 7 Reconstruction evaluation: Projection error across views against state-of-the-art methods, LS is LocalStereo, $P_P = E - E_p$, $P_M = E - E_f - E_r$, $P_{PM} = E - E_f - E_r - E_p$, $P_S = E - E_{sem}$ and $P_{PS} = E - E_{sem} - E_p$, where E is defined in Equation 2

Methods	Handshake	Handstand	Rachel	Juggler1	Juggler2	Magician	Dance	Meetup	Human3.6	MagicianLF	WalkLF
PRSM Vogel et al. (2015)	1.56	1.79	1.51	1.57	1.68	1.72	1.79	1.98	2.01	1.59	1.41
LS Taniati et al. (2018)	1.24	1.38	1.15	1.21	1.18	1.33	1.46	1.47	1.64	1.20	1.23
SMVS Langguth et al. (2016)	0.84	0.97	0.73	0.75	0.85	0.92	0.85	0.96	1.19	0.94	0.88
SCSR Mustafa and Hilton (2017)	0.70	0.84	0.67	0.69	0.73	0.78	0.77	0.87	0.92	0.77	0.71
P_{PS}	0.73	0.87	0.65	0.70	0.71	0.75	0.74	0.88	0.90	0.78	0.70
P_{PM}	0.71	0.85	0.64	0.68	0.69	0.73	0.72	0.85	0.87	0.75	0.68
P_P	0.57	0.71	0.56	0.59	0.61	0.64	0.62	0.75	0.77	0.67	0.63
P_S	0.59	0.69	0.59	0.57	0.63	0.66	0.60	0.73	0.76	0.65	0.60
P_M	0.55	0.68	0.55	0.54	0.59	0.61	0.59	0.74	0.73	0.62	0.59
Proposed	0.46	0.55	0.47	0.49	0.51	0.53	0.55	0.57	0.60	0.49	0.44

Bold values indicate best performing method

Table 8 Segmentation comparison against state-of-the-art methods using the Intersection-over-Union metric

Methods	Handshake	Handstand	Rachel	Juggler1	Juggler2	Magician	Dance	Meetup	Human3.6	MagicianLF	WalkLF
CRFRNN Zheng et al. (2015)	62.7	55.8	61.6	40.5	68.7	52.4	49.3	41.1	42.9	60.8	63.6
Segnet Badrinayanan et al. (2017)	47.9	51.1	55.2	45.1	61.9	55.3	53.9	43.9	49.4	59.3	65.9
JSR Guillemaut and Hilton (2010)	67.8	58.7	58.4	56.2	66.0	61.3	57.9	50.2	53.4	62.3	68.9
SCV Tsai et al. (2016)	56.4	52.6	48.8	49.5	59.1	59.2	56.7	42.0	49.1	58.2	65.7
Dv3+ Chen and Zhu (2018)	63.8	58.9	64.0	48.8	69.7	58.9	57.6	48.4	54.8	69.6	69.1
MRCNN He et al. (2017)	65.2	59.6	67.4	50.3	70.5	60.5	58.7	47.2	53.4	69.5	70.2
PSP Zhao et al. (2017)	74.7	64.5	75.5	67.9	81.2	73.4	71.5	62.6	65.3	74.6	82.5
SCSR Mustafa and Hilton (2017)	81.8	75.2	78.4	81.4	89.3	88.2	85.1	78.9	70.4	82.2	86.7
P_{PM}	85.7	75.9	78.6	81.8	89.6	88.5	85.5	79.2	70.6	82.9	87.5
P_P	86.3	77.4	80.7	82.6	90.1	89.1	87.6	80.8	76.3	86.1	89.3
P_M	87.6	79.1	81.7	83.5	90.5	89.6	86.4	81.9	75.4	85.2	88.1
Proposed	89.6	83.3	85.8	88.2	91.1	90.9	88.5	84.7	81.1	89.4	91.8

Bold values indicate best performing method



Fig. 7 Reconstruction evaluation against existing methods. Two different views of 3D model are shown for proposed method

Only MRCNN and the proposed approach gives instance segmentation.

Quantitative evaluation against state-of-the-art methods is measured by *Intersection-over-Union* with ground-truth, shown in Table 8. Ground-truth is available on-line for most of the datasets and obtained by manual labelling for other datasets. Pre-trained parameters were used for semantic segmentation methods. The semantic instance segmentation results from the joint optimisation are significantly better compared to the state-of-the-art methods ($\approx 20 - 40\%$).

Semantic segmentation comparison results against CRF RNN (Zheng et al. 2015), Segnet (Badrinarayanan et al. 2017), PSP (Zhao et al. 2017) are shown in Fig. 10 on four datasets. Ground-truth segmentation comparison is shown in Fig. 11 against JSR (Guillemaut and Hilton 2010) and SCSR Mustafa and Hilton (2017). The red and green regions highlight the error, green regions are present in segmentation but not ground-truth and red regions are present in ground-truth but not the segmentation.

4.3 Motion Evaluation

Flow from the joint estimation is evaluated against state-of-the-art methods: (a) Dense flow algorithms DCflow (Xu et al.

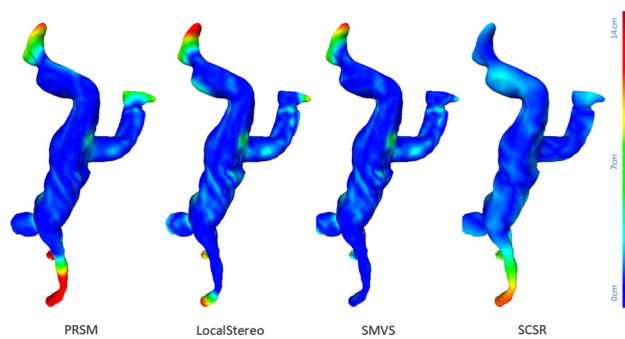


Fig. 8 Comparison of reconstruction obtained using state-of-the-art methods against proposed method

2017) and Deepflow Weinzaepfel et al. (2013); (b) Scene flow methods PRSM (Vogel et al. 2015); and (c) Non-sequential alignment of partial surfaces 4DMatch (Mustafa et al. 2016) (requires a prior 3D mesh of the object as input for 4D reconstruction).

Per-view motion estimate from the proposed method is compared with each of the state-of-the-art methods. All of the methods we have compared with DCflow, Deepflow, PRSM and 4DMatch estimate flow for each camera which makes a fair comparison and default parameters are used for all state-of-the-art methods.

The key-frames of sequences are coloured using the unique color scheme shown in Fig. 12 and the colour is propagated using the motion estimate from the joint optimisation throughout the sequence. With accurate motion estimates the colors should propagate reliably across the sequence. The red regions in 2D dense flow in Fig. 12 are the regions for which reliable correspondences are not found. This demonstrates improved performance using the proposed method. The colours in the 4D alignment in Fig. 13 are not reliably propagated by DCFlow for limbs.

We also compare the silhouette overlap error (S_e) across frames, key-frames and views to evaluate long-term temporal coherence in Table 9 for all datasets. This is defined as $S_e = \frac{1}{N_v N_k N_f} \sum_{i=1}^{N_k} \sum_{j=1}^{N_f} \sum_{c=1}^{N_v} \frac{\text{Area of intersection}}{\text{Area of semantic segmentation}}$. Dense flow in time is used to obtain the propagated mask for each image. The propagated mask is overlapped with semantic segmentation at each time instant to evaluate the accuracy of the propagated mask. The lower the S_e the better. Our approach gives the lowest error demonstrating higher accuracy compared to the state-of-the-art methods.

We evaluate the temporal coherence across the Meetup sequence, by evaluating the variation in appearance for each scene point between frames and between key-frames and frames for state-of-the-art methods. The metric is defined as: $\sqrt{\frac{\Delta r^2 + \Delta g^2 + \Delta b^2}{3}}$, where Δ is the difference operator. Evaluation shown in Table 10 against state-of-the-art methods

demonstrates the stability of long term temporal tracking for proposed method (the lower the error the better).

4.4 Ablation Study on Equation 2

We perform an ablation study on Equation 2, such that we remove motion E_f , E_r , pose E_p and semantic E_{sem} constraints from the equation, defining $P_M = E - E_f -$

E_r , $P_P = E - E_p$, $P_{PM} = E - E_f - E_r - E_p$, $P_S = E - E_{sem}$ and $P_{PS} = E - E_{sem} - E_p$. Reconstruction, flow and semantic segmentation is obtained with removed constraints, and the results are shown in Tables 7, 9 and 8 respectively. The proposed approach gives best performance with joint pose, motion and semantic constraints.

Based on this ablation analysis, it is shown that the contribution of each term is task dependent. **Reconstruction** Pose

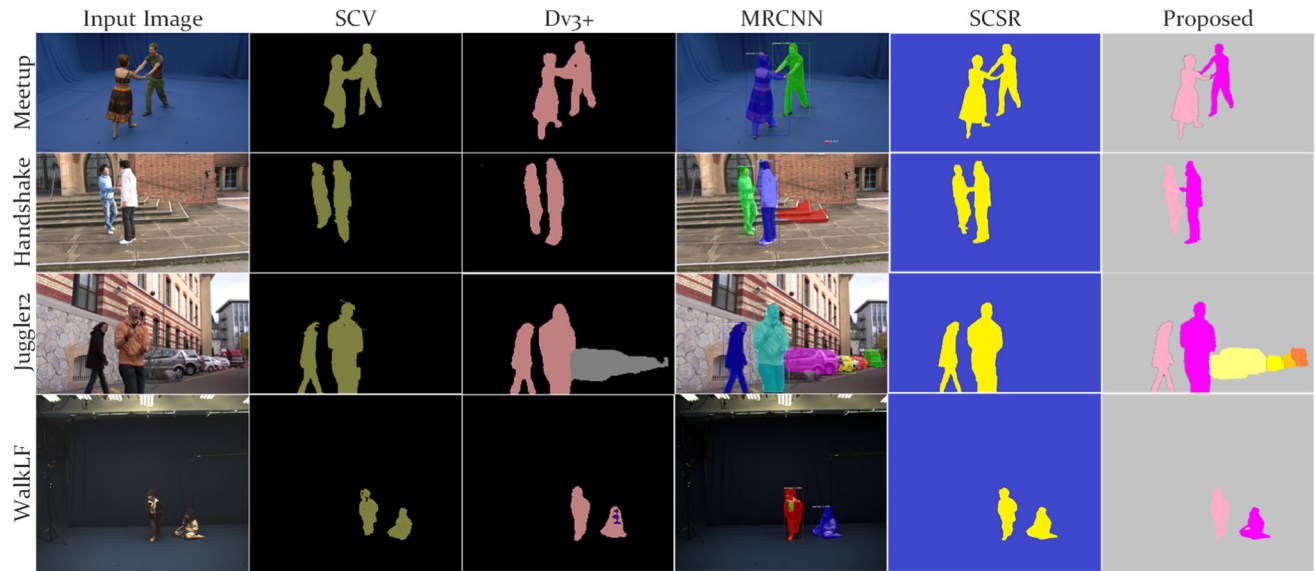


Fig. 9 Semantic segmentation comparison against state-of-the-art methods. In the proposed method shades of pink depicts instances of humans and shades of yellow depict instances of cars (Color figure online)

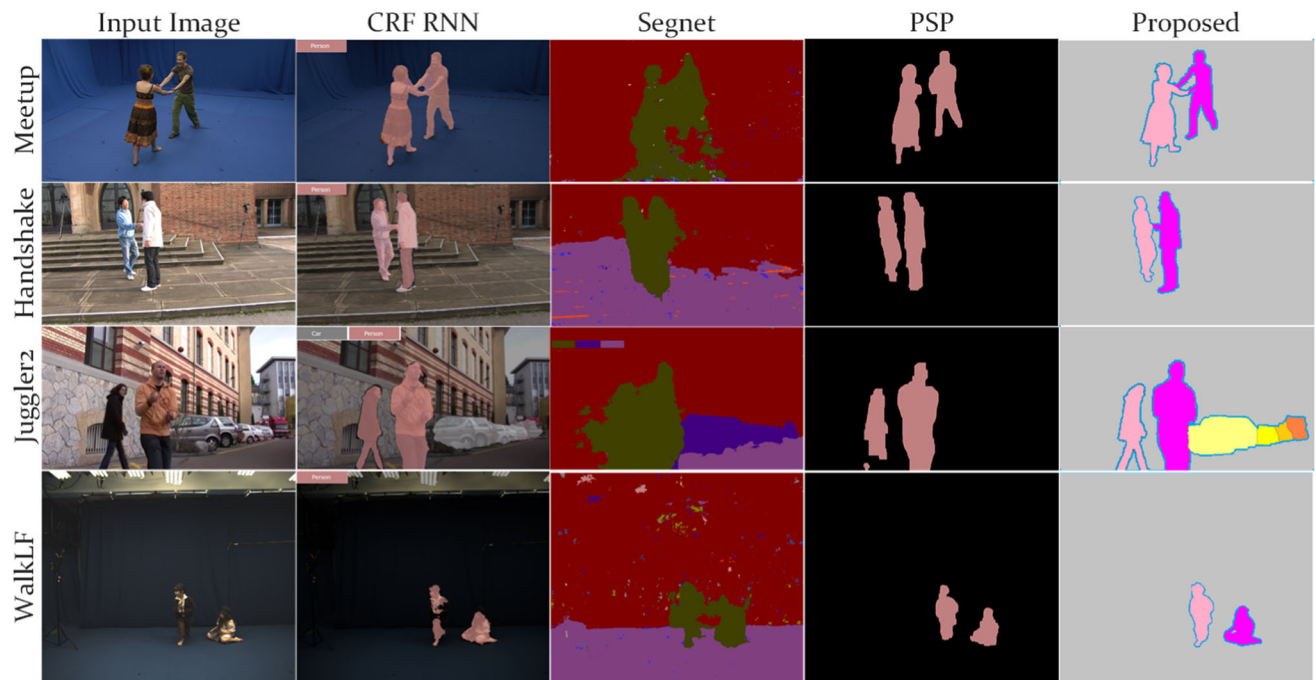


Fig. 10 Semantic segmentation comparison against state-of-the-art methods. In the proposed method shades of pink depicts instances of humans and shades of yellow depict instances of cars (Color figure online)

Table 9 Table caption

Methods	Handshake	Handstand	Rachel	Juggler1	Juggler2	Magician	Dance	Meetup	Human3.6	MagicianLF	WalkLF
PRSM Xu et al. (2017)	1.80	2.15	1.54	1.65	1.79	1.96	1.87	2.11	2.34	1.87	1.52
Deepflow Weinzaepfel et al. (2013)	1.15	1.48	1.01	1.08	1.16	1.27	1.21	1.37	1.52	1.05	0.81
DCFlow Vogel et al. (2015)	0.90	1.17	0.97	0.87	0.93	1.03	0.96	1.12	1.21	0.83	0.79
4DMatch Mustafa et al. (2016)	0.79	0.98	0.75	0.69	0.87	0.81	0.77	0.87	0.94	0.80	0.77
$P_{P,S}$	0.75	1.01	0.85	0.78	0.91	0.93	0.86	0.99	1.07	0.81	0.78
P_P	0.71	0.93	0.80	0.73	0.84	0.87	0.78	0.92	0.99	0.76	0.73
P_S	0.64	0.77	0.63	0.61	0.65	0.72	0.65	0.76	0.81	0.64	0.61
Proposed	0.51	0.61	0.48	0.49	0.52	0.58	0.55	0.63	0.68	0.53	0.44

Bold values indicate best performing method



Fig. 11 Ground-truth semantic segmentation comparison against state-of-the-art methods JSR and SCSR

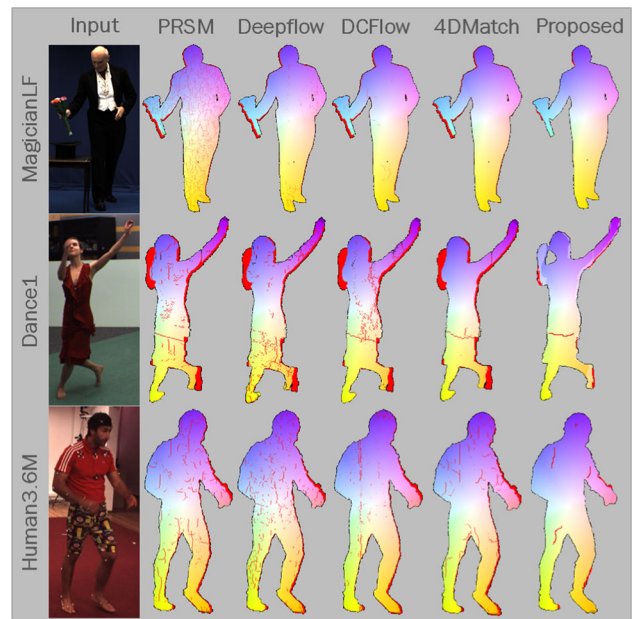


Fig. 12 Temporal coherence evaluation against existing methods

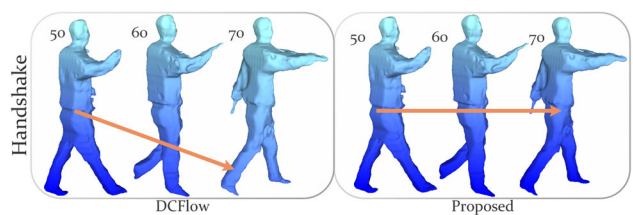


Fig. 13 4D alignment evaluation against DCFlow Xu et al. (2017)

and semantic constraints play an equal role in reconstruction and the motion term contributes less to this task. *Semantics* Pose constraints contribute more to the performance than the motion term. *Flow* Pose constraints contribute more than semantics to the performance. To sum up, the motion term gives proposed 4D flow, however it contributes least to the overall performance, followed by semantic, pose then depth. Also, the terms in Equation 2 are chosen based on a series of experiments to achieve optimal performance.

Table 10 Temporal coherence evaluation for Meetup dataset against existing methods

Methods	Frame-to-frame		Keyframe-to-frame	
	Mean	S.D.	Mean	S.D.
Proposed	3.604	1.653	4.181	2.317
4DMatch	5.896	2.513	8.344	5.006
DCflow	6.085	3.314	16.673	8.55
Deepflow	7.525	4.179	18.115	9.052
PRSM	8.794	4.908	20.876	11.493

Bold values indicate best performing method
S.D. is the standard deviation

Table 11 Silhouette overlap error for multi-view datasets for evaluation of long-term temporal coherence

Datasets	Proposed without key-frame detection	Proposed
Handshake	0.60	0.51
Handstand	0.71	0.61
Juggler1	0.57	0.49
Juggler2	0.59	0.52
Magician	0.67	0.58
Meetup	0.72	0.63
Human3.6	0.78	0.68
WalkLF	0.51	0.44

Bold values indicate best performing method

The method requires pose as prior for human reconstruction, however Table 7, 8 and 9 demonstrate results without pose constraints (P_P). The performance is reduced but the method still performs better than other state-of-the-art approaches.

4.5 Ablation Study Without Key-Frame Detection

The higher the number of key-frames the better the quality of alignment. However if no key-frames are detected for a sequence, it will degrade the performance of 4D long-term scene flow. To evaluate the effect of key-frame detection we evaluate the performance of 4D scene flow for proposed joint optimization with and without key-frames in Table 11. The results show an $\approx 15\%$ improvement in scene flow with key-frame detection.

4.6 Computation Time Comparison

Computation times for the proposed approach vs other methods that perform joint estimation are presented in Table 12. The proposed approach to reconstruct temporally coherent 4D models is comparable in computation time to per-frame multiple view reconstruction and gives a $\sim 50\%$ reduction in computation cost compared to previous joint segmentation and reconstruction approaches using a known background. This efficiency is achieved through improved per-frame initialisation based on temporal propagation and the introduction of the geodesic star constraint in joint optimisation.

4.7 Results on Sports Data in the Wild

Qualitative results using the proposed 4D scene understanding framework on three sports sequences Football, Wembley and Soccer with multiple people are shown in Figs. 14, 15 and 16. Football sequence is a synthetic dataset, Wembley and Soccer are real datasets. Properties of the sports datasets are listed in Table 4.

The proposed method obtains robust semantic instance segmentation on Football and Wembley sequences demonstrating the generalizing capability of the method for in the wild datasets, illustrated in Fig. 14. 4D Reconstruction obtained using the proposed 4D scene understanding framework is shown in Fig. 15. Reconstruction is shown from different viewpoints demonstrating the applicability of automatic scene understanding framework for in the wild data. Proposed long term temporal coherence is shown in Fig. 16. Each player is uniquely coloured and the colours are propagated using proposed motion estimate from the 4D scene understanding framework. In spite of the low resolution ($\approx 12 \times 30$ px) of each player the colours are reliably propagated across the sequence.

For comparative evaluation on sports datasets, majority of the state-of-the-art methods are unable to obtain a reliable segmentation, mesh and flow for the players. Hence we have evaluated on the selected methods that work, and comparative evaluation of segmentation, reconstruction and motion estimation is shown in Table 13. The results demonstrate



Fig. 14 Semantic instance segmentation on two sports sequences with multiple people

Table 12 Comparison of computational efficiency for a few dynamic sequences against state-of-the-art methods (time in seconds)

Dataset	PRSM Xu et al. (2017)	JSR Guillemaut and Hilton (2010)	SCSR Mustafa and Hilton (2017)	Proposed
Magician	342 s	608 s	362 s	353 s
Rachel	397 s	582 s	379 s	362 s
Handstand	348 s	566 s	353 s	325 s
Juggler2	413 s	621 s	405 s	421 s
MagicianLF	659 s	1227 s	622 s	611 s

Bold values indicate best performing method

that proposed method achieves significant improvement over existing methods for all tasks.

4.8 Complexity Analysis on Synthetic Multi-View Data

To evaluate how proposed method works with crowded scenes we use a multi-view synthetic dataset with multiple people in the scene ranging from 2–10 people at each time instant (Caliskan et al. 2020). Qualitative results are shown in Fig. 17.

We also perform complexity analysis for the proposed method through a quantitative evaluation on the synthetic dataset. The number of people are increased in the scene

and reconstruction and segmentation accuracy is calculated for different number of people in the scene as seen in Table 14. The reconstruction and segmentation accuracy slightly decreases with the increase in the number of people in the scene, due to increased occlusion and clutter in the scene.

4.9 Limitations

Gross errors in initial semantic instance segmentation and 3D pose estimation lead to degradation in the quality of results (e.g. the cars in Juggler2 - Fig. 9).

Small errors in semantic segmentation, initial coarse reconstruction and 3D pose are handled gracefully by the

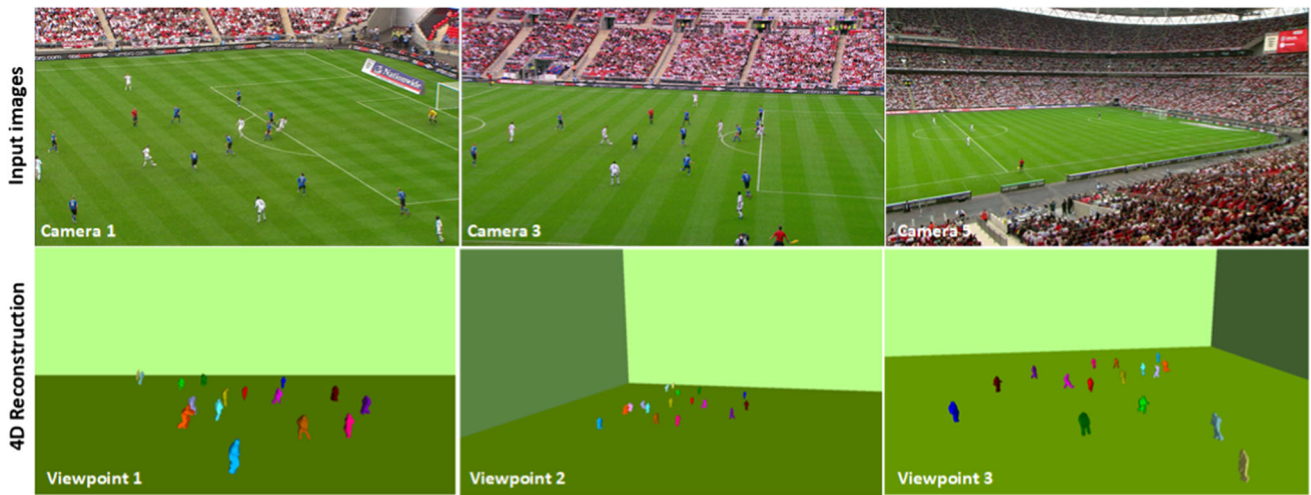


Fig. 15 4D Reconstruction demonstrated on Wembley dataset. Top row shows different cameras for Frame 105 and bottom rows show reconstruction from 3 different randomly picked viewpoints

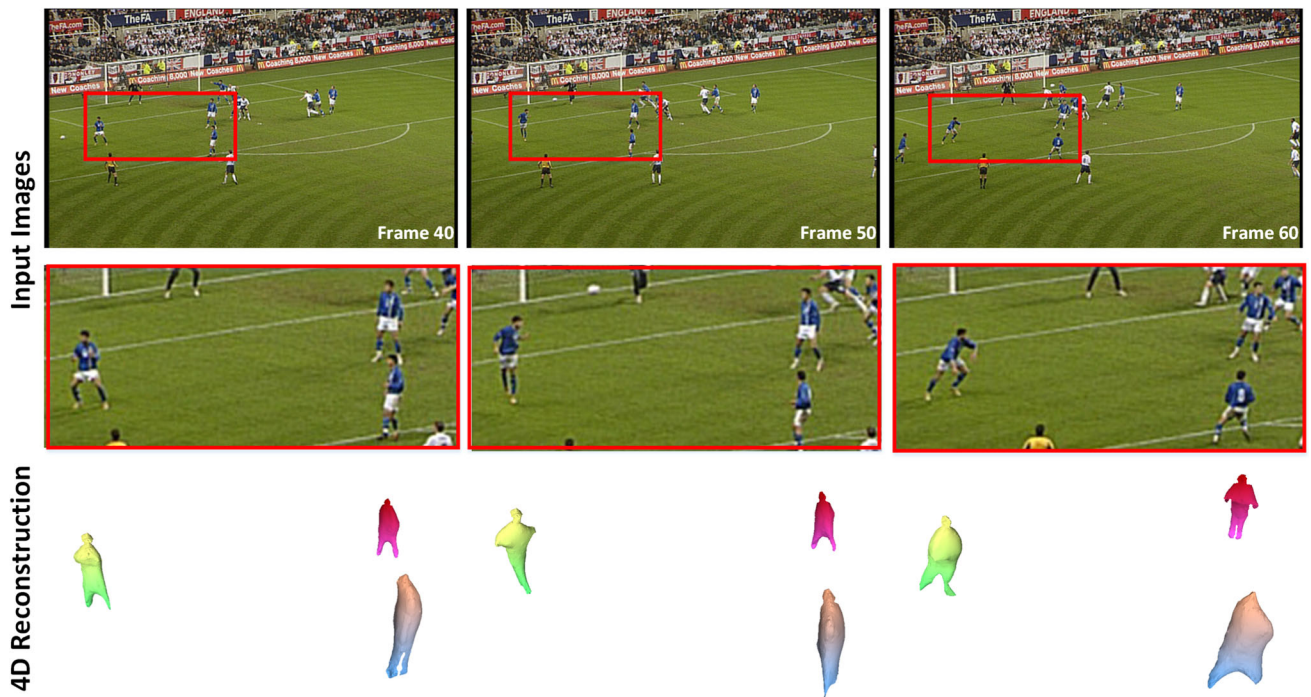


Fig. 16 4D Reconstruction demonstrated on Soccer dataset for frame 40, 50 and 60. Unique colours are assigned to the reconstruction of each player in the cropped image for frame 1 and the colours are reliably propagated across the sequence

proposed method. This is evident from Fig. 9 where the initial semantic segmentation MRCNN has small errors and the final result of the proposed method improves the semantic segmentation. However, large errors over successive frames in semantic segmentation propagate through the optimization, which reduces the quality of the final segmentation and reconstruction as seen in Fig. 16. The final reconstructions look incomplete because of large errors in initial semantic segmentation.

Also errors in the initial coarse reconstruction, for example an incomplete initial coarse reconstruction with missing human limb will lead to incomplete final reconstruction. Gross errors in 3D pose estimation for crowded scenes may lead to errors in the final reconstruction. Failure or errors in key-frame detection only slightly degrades the quality of the flow estimate from the proposed approach. Although 3D

Table 13 Quantitative comparison of segmentation, reconstruction and motion estimation obtained from proposed 4D scene understanding framework on sports datasets

Segmentation results			
Dataset	Football	Wembley	Soccer
Dv3+ Chen and Zhu (2018)	49.9	46.5	39.8
MRCNN He et al. (2017)	50.4	46.7	40.2
SCSR Mustafa and Hilton (2017)	58.1	51.6	43.4
Proposed	71.7	68.5	65.3
Reconstruction results			
Dataset	Football	Wembley	Soccer
SCSR Mustafa and Hilton (2017)	1.19	1.36	1.62
Proposed	0.70	0.79	0.82
Motion estimation results			
Dataset	Football	Wembley	Soccer
Deepflow Weinzaepfel et al. (2013)	1.98	2.21	2.43
4DMatch Mustafa et al. (2016)	1.66	1.89	2.05
Proposed	0.91	1.02	1.15

Bold values indicate best performing method



Fig. 17 Reconstruction Results on the Multi-view Synthetic Dataset

human pose helps in robust 4D reconstruction of interacting people in dynamic scenes, current 3D pose estimation is unreliable for highly crowded environments resulting in degradation of the proposed approach.

Table 14 Quantitative comparison of segmentation and reconstruction accuracy on multi-view synthetic multiple people dataset

	Reconstruction results	Segmentation results
2 people	0.45	91.7
4 people	0.45	91.5
6 people	0.47	91.1
8 people	0.48	90.8
10 people	0.51	90.4

For reconstruction evaluation, projection error across views is shown and for segmentation evaluation Intersection-of-union metric is shown

5 Conclusions

This paper introduced the first automatic method for 4D dynamic scene understanding of multiple interacting people from multi-view video that does not need any ground-truth data for training or manual intervention. A novel joint flow, reconstruction and semantic instance segmentation framework is introduced exploiting 2D/3D human-pose, motion, semantic, shape and appearance information in space and time. Ablation study on the joint optimisation demonstrates the effectiveness of the proposed scene understanding framework for general scenes with multiple interacting people. The semantic, motion and depth information per view is fused spatially across views for 4D semantically and temporally coherent scene understanding. A fully automatic system is presented for multiple view semantic instance segmentation, 4D reconstruction and motion estimation from moving broadcast cameras to allow full 4D scene understanding of in the wild sports data such as soccer. Extensive evalua-

tion against state-of-the-art methods on a variety of complex indoor and outdoor datasets with large non-rigid deformations demonstrates a significant improvement in the accuracy in semantic segmentation, reconstruction, motion estimation and 4D alignment.

Acknowledgements This research was supported by the Royal Academy of Engineering Research Fellowship RF-201718-17177 and the EPSRC Platform Grant on Audio-Visual Media Research EP/P022529.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- 4d repository, <http://4drepository.inrialpes.fr/>. In: Institut national de recherche en informatique et en automatique (INRIA) Rhone Alpes.
- Multiview video repository, <http://cvssp.org/data/cvssp3d/>. In: Centre for Vision Speech and Signal Processing, University of Surrey, UK.
- Kundu, A., Yin, X., Fathi, A., Ross, D., Brewington, B., Funkhouser, T., & Pantofaru, C. (2020). Virtual multi-view fusion for 3d semantic segmentation. In: *ECCV*.
- Gilbert, A., Trumble, M., Hilton, A. & Collomosse, J. (2020) Semantic estimation of 3d body shape and pose using minimal cameras. In: *BMVC*.
- Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*.
- Ballan, L., Brostow, G. J., Puwein, J., & Pollefeys, M. (2010). *Unstructured video-based rendering: Interactive exploration of casually captured videos*. Graph: ACM Trans.
- Basha, T., Moses, Y., Kiryati, N. (2010). Multi-view scene flow estimation: A view centered variational approach. In: *CVPR*, pp. 1506–1513.
- Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *TPAMI*, 26(11), 1124–1137.
- Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *TPAMI*, 23(11), 1222–1239.
- Cai, Y., Huang, L., Wang, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., Shen, X., Liu, D., Thalman, N.M. (2020). Learning progressive joint propagation for human motion prediction. In: A. Vedaldi, H. Bischof, T. Brox, J.M. Frahm (eds.) *Computer Vision – ECCV 2020*, pp. 226–242.
- Caliskan, A., Mustafa, A., Imre, E., Hilton, A. (2020). Multi-view consistency loss for improved single-image 3d reconstruction of clothed people. In: *Asian Conference on Computer Vision (ACCV)*.
- Cao, Z., Simon, T., Wei, S.E., Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In: *CVPR*.
- Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., Yan, Y. (2020). Blend-mask: Top-down meets bottom-up for instance segmentation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L. (2016). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR arXiv:1606.00915*
- Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation.
- Chen, P.Y., Liu, A.H., Liu, Y.C., Wang, Y. (2019). Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In: *CVPR*.
- Chiu, W.C., Fritz, M. (2013). Multi-class video co-segmentation with a generative multi-video model. In: *CVPR*.
- Dai, A., Nießner, M. (2018). 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In: *ECCV*.
- Djelouah, A., Franco, J.S., Boyer, E., Perez, P., Drettakis, G. (2016). Cotemporal Multi-View Video Segmentation. In: *3DV*.
- Dosovitskiy, A., Fischery, M., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Smagt, P., Cremers, D., Brox, T. (2015). FlowNet: Learning optical flow with convolutional networks. In: *ICCV*.
- Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S.R., Kowdle, A., Escolano, S.O., Rhemann, C., Kim, D., Taylor, J., Kohli, P., Tankovich, V., Izadi, S. (2016). Fusion4d: Real-time performance capture of challenging scenes. *ACM Trans. Graph.* 35(4).
- Eigen, D., Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: *ICCV*.
- Engelmann, F., Stückler, J., Leibe, B. (2016). Joint object pose estimation and shape reconstruction in urban street scenes using 3D shape priors. In: *GCLR*.
- Evangelidis, G. D., & Psarakis, E. Z. (2008). Parametric image alignment using enhanced correlation coefficient maximization. *TPAMI*, 30(10), 1858–1865.
- Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A. (2012). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>
- Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (2013). Learning hierarchical features for scene labeling. *TPAMI*, 35(8), 1915–1929.
- Floros, G., Leibe, B. (2012). Joint 2d-3d temporally consistent semantic segmentation of street scenes. In: *CVPR*, pp. 2823–2830.
- Godard, C., Mac Aodha, O., Brostow, G.J. (2017). Unsupervised monocular depth estimation with left-right consistency. In: *CVPR*.
- Guerry, J., Boulch, A., Saux, B.L., Moras, J., Plyer, A., Filliat, D. (2017). Snapnet-r: Consistent 3d multi-view semantic labeling for robotics. In: *ICCVW*.
- Guillemaut, J. Y., & Hilton, A. (2010). Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. *IJCV*, 93, 73–100.
- Gupta, S., Girshick, R.B., Arbelaez, P., Malik, J. (2014). Learning rich features from RGB-D images for object detection and segmentation, pp. 345–360.
- Hane, C., Zach, C., Cohen, A., Pollefeys, M. (2016). Dense semantic 3d reconstruction. *TPAMI* p. 1.
- Hariharan, B., Arbeláez, P.A., Girshick, R.B., Malik, J. (2015). Hypercolumns for object segmentation and fine-grained localization. In: *CVPR*, pp. 447–456.
- Hasler, N., Rosenhahn, B., Thormahlen, T., Wand, M., Gall, J., Seidel, H.P. (2009). Markerless motion capture with unsynchronized moving cameras. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 224–231. <https://doi.org/10.1109/CVPR.2009.5206859>.

- He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask R-CNN. In: *ICCV*.
- Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M. J. (2017). Towards accurate marker-less human shape and pose estimation over time. In: *3DV*.
- Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7), 1325–1339.
- Kazhdan, M., Bolitho, M., Hoppe, H. (2006). Poisson surface reconstruction. In: *Eurographics Symposium on Geometry Processing*, pp. 61–70
- Kendall, A., Gal, Y., Cipolla, R. (2017). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. CoRR [arXiv:1705.07115](https://arxiv.org/abs/1705.07115).
- Kendall, A., Gal, Y., Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: *CVPR*.
- Kim, H., Sarim, M., Takai, T., Yves Guillemaut, J., Hilton, A. (2012). Outdoor dynamic 3-D scene reconstruction. *T-CSVT*, 22(11), 1611–1622.
- Klodt, M., Vedaldi, A. (2018). Supervising the new with the old: learning sfm from sfm. In: *ECCV*.
- Kundu, A., Li, Y., Dellaert, F., Li, F., Reh, J.M. (2014). Joint semantic segmentation and 3d reconstruction from monocular video. In: *ECCV*, vol. 8694, pp. 703–718.
- Kundu, A., Vineet, V., Koltun, V. (2016). Feature space optimization for semantic video segmentation. In: *CVPR*, pp. 3168–3175.
- Lai, H., Tsai, Y., Chiu, W. (2019). Bridging stereo matching and optical flow via spatiotemporal correspondence. In: *CVPR*.
- Langguth, F., Sunkavalli, K., Hadap, S., Goesele, M. (2016). Shading-aware multi-view stereo. In: *ECCV*.
- Larsen, E.S., Mordohai, P., Pollefeys, M., Fuchs, H. (2007). Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In: *ICCV*, pp. 1–8.
- Li, X., You, A., Zhu, Z., Zhao, H., Yang, M., Yang, K., Tong, Y. (2020). Semantic flow for fast and accurate scene parsing. In: *ECCV*.
- Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L. (2014). Microsoft COCO: common objects in context. CoRR [arXiv:1405.0312](https://arxiv.org/abs/1405.0312).
- Luo, B., Li, H., Song, T., Huang, C. (2015). Object segmentation from long video sequences. In: *ACM Multimedia*, pp. 1187–1190.
- Menze, M., Heipke, C., Geiger, A. (2015). Discrete optimization for optical flow. In: *German Conference on Pattern Recognition (GCPR)*, vol. 9358, (pp. 16–28). Springer International Publishing.
- Mostajabi, M., Yadollahpour, P., Shakhnarovich, G. (2015). Feedforward semantic segmentation with zoom-out features. In: *CVPR*, pp. 3376–3385.
- Mustafa, A., Hilton, A. (2017). Semantically coherent co-segmentation and reconstruction of dynamic scenes. In: *CVPR*.
- Mustafa, A., Kim, H., Guillemaut, J., Hilton, A. (2016). Temporally coherent 4d reconstruction of complex dynamic scenes. In: *CVPR*.
- Mustafa, A., Kim, H., Hilton, A. (2016). 4d match trees for non-rigid surface alignment. In: *ECCV*.
- Mustafa, A., Kim, H., & Hilton, A. (2019). Msfd: Multi-scale segmentation-based feature detection for wide-baseline scene reconstruction. *IEEE Transactions on Image Processing*, 28, 1118–1132.
- Mustafa, A., Russell, C., Hilton, A. (2019). U4d: Unsupervised 4d dynamic scene understanding. In: *ICCV*.
- Mustafa, A., Volino, M., Guillemaut, J., Hilton, A. (2017). 4d temporally coherent light-field video. In: *3DV*.
- Newcombe, R.A., Fox, D., Seitz, S.M. (2015). Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR* pp. 343–352.
- Ranjan, A., Jampani, V., Kim, K., Sun, D., Wulff, J., Black, M.J. (2019). Adversarial collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: *CVPR*.
- Ranjan, A., Romero, J., Black, M.J. (2018). Learning human optical flow. In: *BMVC*.
- Rodriguez, A.L., Mikolajczyk, K. (2020). Desc: Domain adaptation for depth estimation via semantic consistency. In: *BMVC*.
- Rossi, M., Gheche, M.E., Kuhn, A., Frossard, P. (2020). Joint graph-based depth refinement and normal estimation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Roussos, A., Russell, C., Garg, R., Agapito, L. (2012). Dense multibody motion estimation and reconstruction from a handheld camera. In: *ISMAR*.
- Rusu, R.B. (2009). Semantic 3d object maps for everyday manipulation in human living environments. Ph.D. thesis, Computer Science department, Technische Universitaet Muenchen, Germany.
- Bi, S., Xu, Z., Sunkavalli, K., Hasan, M., Hold-Geoffroy, Y., Kriegman, D., & Ramamoorthi, R. (2020). Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In: *ECCV*.
- Schönberger, J.L., Frahm, J.M. (2016). Structure-from-motion revisited. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M. (2016). Pixelwise view selection for unstructured multi-view stereo. In: *European Conference on Computer Vision (ECCV)*.
- Sevilla-Lara, L., Sun, D., Jampani, V., Black, M.J. (2016). Optical flow with semantic segmentation and localized layers. In: *CVPR*, pp. 3889–3898.
- Shelhamer, E., Long, J., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In: *CVPR*.
- Siam, M., Gamal, M., Abdel-Razek, M., Yogamani, S., Jägersand, M. (2018). Rtseg: Real-time semantic segmentation comparative study. In: *ICIP*.
- Sorkine, O., Alexa, M. (2007). As-rigid-as-possible surface modeling. In: *SGP*, pp. 109–116.
- Szeliski, R. (1999). A multi-view approach to motion and stereo. In: *CVPR*.
- Taniai, T., Matsushita, Y., Sato, Y., & Naemura, T. (2018). Continuous 3D label stereo matching using local expansion moves. *TPAMI*, 40(11), 2725–2739. <https://doi.org/10.1109/TPAMI.2017.2766072>.
- Tao, M.W., Bai, J., Kohli, P., Paris, S. (2012). Simpleflow: A non-iterative, sublinear optical flow algorithm. *Computer Graphics Forum (Eurographics 2012)*, 31(2).
- Tome, D., Russell, C., Agapito, L. (2017). Lifting from the deep: Convolutional 3d pose estimation from a single image. In: *CVPR*.
- Tomè, D., Toso, M., Agapito, L., Russell, C. (2018). Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. In: *3DV*.
- Trager, M., Hebert, M., Ponce, J. (2019). Coordinate-free carlsson-weinshall duality and relative multi-view geometry. In: *CVPR*.
- Tsai, Y.H., Zhong, G., Yang, M.-H., e.B., Matas, J., Sebe, N., Welling, M. (2016). Semantic co-segmentation in videos. In: *ECCV*, pp. 760–775.
- Ulusoy, A.O., Black, M.J., Geiger, A. (2017). Semantic multi-view stereo: Jointly estimating objects and voxels. In: *CVPR*.
- Vineet, V., Miksik, O., Lidegaard, M., Nießner, M., Golodetz, S., Prisacariu, V.A., Kähler, O., Murray, D.W., Izadi, S., Perez, P., Torr, P.H.S. (2015). Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In: *ICRA*.
- Vlasic, D., Baran, I., Matusik, W., Popović, J. (2008). Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.*, 27(3).
- Vogel, C., Schindler, K., Roth, S. (2015). 3d scene flow estimation with a piecewise rigid scene model pp. 1–28.

- Wang, L., Zhang, J., Wang, O., Lin, Z., Lu, H. (2020). Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., & Cremers, D. (2011). Stereoscopic scene flow computation for 3d motion understanding. *IJCV*, 95(1), 29–51.
- Wei Zeng, S.K., Gevers, T. (2020). Pano2scene: 3d indoor semantic scene reconstruction from a single indoor panorama image. In: *BMVC*.
- Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C. (2013). Deep-flow: Large displacement optical flow with deep matching. In: *ICCV*, pp. 1385–1392.
- Xia, F., Wang, P., Chen, X., Yuille, A.L. (2017). Joint multi-person pose estimation and semantic part segmentation. In: *CVPR*.
- Xie, J., Kiefel, M., Sun, M.T., Geiger, A. (2016). Semantic instance annotation of street scenes by 3d to 2d label transfer. In: *CVPR*.
- Xu, J., Ranftl, R., Koltun, V. (2017). Accurate optical flow via direct cost volume processing. In: *CVPR*.
- Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L. (2018). Mvsnet: Depth inference for unstructured multi-view stereo. In: *ECCV*.
- Zanfir, A., Sminchisescu, C. (2015). Large displacement 3d scene flow with occlusion reasoning. In: *ICCV*.
- Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Pyramid scene parsing network. In: *CVPR*.
- Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S. (2015). Conditional random fields as recurrent neural networks. In: *ICCV*.
- Zhong, Y., Ji, P., Wang, J., Dai, Y., Li, H. (2019). Unsupervised deep epipolar flow for stationary or dynamic scenes. In: *CVPR*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.