International Journal of Computer Vision (2020) 128:319–335
https://doi.org/10.1007/s11263-019-01241-w

Semantically Coherent 4D Scene Flow of Dynamic Scenes

Armin Mustafa[1] · Adrian Hilton[1]

Received: 25 January 2019 / Accepted: 19 September 2019 / Published online: 3 October 2019

Abstract

Simultaneous semantically coherent object-based long-term 4D scene flow estimation, co-segmentation and reconstruction is proposed exploiting the coherence in semantic class labels both spatially, between views at a single time instant, and temporally, between widely spaced time instants of dynamic objects with similar shape and appearance. In this paper we propose a framework for spatially and temporally coherent semantic 4D scene flow of general dynamic scenes from multiple view videos captured with a network of static or moving cameras. Semantic coherence results in improved 4D scene flow estimation, segmentation and reconstruction for complex dynamic scenes. Semantic tracklets are introduced to robustly initialize the scene flow in the joint estimation and enforce temporal coherence in 4D flow, semantic labelling and reconstruction between widely spaced instances of dynamic objects. Tracklets of dynamic objects enable unsupervised learning of long-term flow, appearance and shape priors that are exploited in semantically coherent 4D scene flow estimation, co-segmentation and reconstruction. Comprehensive performance evaluation against state-of-the-art techniques on challenging indoor and outdoor sequences with hand-held moving cameras shows improved accuracy in 4D scene flow, segmentation, temporally coherent semantic labelling, and reconstruction of dynamic scenes.

Keywords Semantic 4D flow · Reconstruction · Segmentation

1 Introduction

Advances in visual scene understanding using deep learning, with convolutional neural network architectures and large annotated image collections (Chen et al. 2016, 2018; Xie et al. 2016; Luo et al. 2015), have achieved excellent performance in per-pixel labelling of semantic categories in real-world scenes from images.

These advances in semantic segmentation have been exploited to improve scene flow estimation between pairs of frames for dynamic scenes (Behl et al. 2017). However semantic segmentation from a single view suffers from errors due to the inherent visual ambiguity which leads to errors in flow estimation at object boundaries and for regions of uniform appearance. Errors may also be introduced in scene flow estimated between pairs of frames due to large non-rigid motions and self-occlusions for dynamic sequences. In the case of multiple views, independent classification for different views and different time instants of the same scene may result in inconsistent per-pixel flow and semantic labelling for the same object.

This paper introduces a framework for semantically coherent long-term 4D scene flow (aligning entire dynamic sequence of > 150 frames), co-segmentation and reconstruction of dynamic scenes, as shown in Fig. 1 for the publicly available Juggler dataset (Ballan et al. 2010) captured with 6 hand-held unsynchronised moving cameras. Joint semantic co-segmentation(top-row), flow estimation, and 4D reconstruction (bottom-row) results in significant improvement in per-view 2D segmentation, 4D scene flow and reconstruction. The approach enforces semantic coherence both spatially across different views of the scene and temporally across different observations of the same object for robust long-term 4D flow estimation. Semantic tracklets are introduced to identify similar frames in time across a sequence, exploiting semantic, motion, shape and appearance information between different observations of a dynamic object over time. This gives improved temporal coherence enabling long-term flow estimation along-with consistent semantic

Communicated by Karteek Alahari.

✉ Armin Mustafa
a.mustafa@surrey.ac.uk

Adrian Hilton
a.hilton@surrey.ac.uk

[1] CVSSP, University of Surrey, Guildford, UK

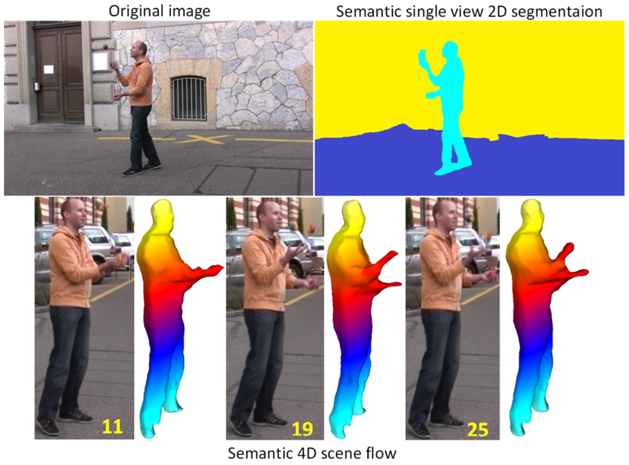


Fig. 1 Example of input image from Juggler dataset (Ballan et al. 2010) and proposed framework resulting in an accurately labeled segmentation, 4D reconstruction and scene flow (represented by color mask propagation in dynamic object of the scene) (Color figure online)

co-segmentation of long sequences across multiple views. Joint semantic scene flow, co-segmentation, and reconstruction enforces spatio-temporal semantic coherence in flow estimation resulting in improved performance over previous approaches which did not exploit semantic and depth information in space and time.

Previous research has demonstrated the advantages of joint semantic segmentation and flow estimation (Tokmakov et al. 2019; Behl et al. 2017; Sevilla-Lara et al. 2016a; Zhu et al. 2018), joint segmentation and reconstruction across multiple views (Yang et al. 2018; Hane et al. 2013, 2016; Engelmann et al. 2016; Kundu et al. 2014), co-segmentation of multiple view images (Khoreva et al. 2019; Chiu and Fritz 2013; Kolev et al. 2012; Djelouah et al. 2015, 2016) and temporal coherence in reconstruction (Li et al. 2018; Goldluecke and Magnor 2004; Floros and Leibe 2012; Larsen et al. 2007; Mustafa et al. 2016a). Our contribution is the introduction of a framework for joint semantically coherent 4D scene flow with co-segmentation and reconstruction of complex dynamic scenes to obtain semantically coherent per-view long-term scene flow, 2D object segmentation and 4D scene reconstruction from wide-baseline camera views. Our approach to long-term scene flow, co-segmentation and 4D dynamic shape reconstruction leverages recent advances in single-view semantic segmentation and semantic flow estimation.

The input to the framework is multi-view videos. Per-view initial semantic segmentation is obtained using Mask RCNN (He et al. 2017) and FCN (Chen et al. 2018), this could in principle use any semantic video segmentation approach. The initial semantic segmentation is combined with sparse reconstruction to obtain initial semantic reconstruction. A joint semantic flow, co-segmentation and reconstruction optimization is proposed to refine this initial segmentation and reconstruction. Semantic coherence is enforced using semantic tracklets, which link frames to enforce temporal coherence between widely spaced timeframes. Semantic coherence refers to spatial and temporal coherence of semantic labels across the sequence. The per-view semantic flow and reconstruction is combined across views for entire dynamic sequence to obtain semantically coherent long-term dense 4D scene flow, co-segmentation and reconstruction.

The primary contribution is semantically coherent scene flow, semantic co-segmentation and 4D reconstruction across multiple views. An initial version of this work was published in CVPR (Mustafa and Hilton 2017) where we proposed a method for semantic segmentation and reconstruction of dynamic scenes. The contributions of this paper over our previous work are as follows: (a) Semantically coherent long-term 4D scene flow estimation for dynamic scenes in addition to semantic segmentation and reconstruction; (b) Refined methodology enabling joint semantically coherent scene flow, co-segmentation and reconstruction by adding motion optimization in the energy defined in Eq. 6. The resulting 2D flow is projected to the 3D reconstruction to obtain the final 4D scene flow; (c) Refined methodology to estimate semantic tracklets by adding motion constrain in Eq. 1 and (d) Comprehensive performance evaluation of flow, segmentation, and reconstruction on challenging datasets. To the best of our knowledge, this is the first method addressing the problem of semantically and temporally coherent long-term 4D scene flow; semantic co-segmentation and reconstruction for dynamic scenes. The contributions of the paper include:

- A method to estimate scene flow, 4D mesh and 2D semantic video segmentation for natural dynamic scenes from multi-view videos.
- Joint semantic scene flow, co-segmentation, and reconstruction of dynamic objects in complex scenes exploiting spatial and temporal coherence.
- Semantic tracklets for long-term 4D reconstruction by enforcing spatial and temporal coherence in semantic labelling for improved scene flow of video across wide-timeframes.
- Improved flow, segmentation, and reconstruction of dynamic scenes from multiple moving cameras

2 Related Work

2.1 Semantic Segmentation

Various methods have been proposed in the literature for semantic segmentation of images. In the first category the image is initially segmented followed by a per-segment object category classification (Mostajabi et al. 2015; Gupta et al. 2014). However, errors in segmentation propagate to the semantic labelling. Several papers address these issues by proposing deep per-pixel CNN features followed by classification of each pixel in the image (Farabet et al. 2013; Hariharan et al. 2015). The per-pixel prediction leads to segmentations with fuzzy boundaries and spatially disjoint regions. Another group of methods pioneered by Long et al. (2015) and He et al. (2017) predict segmentations from the raw pixels. Methods were introduced to improve the spatial coherence of the semantic segmentation using conditional random fields (CRF) (Kundu et al. 2016; Zheng et al. 2015; Chen et al. 2014). End-to-end methods were proposed for semantic segmentation to overcome the limitations of methods using CRF (Chen et al. 2018; Zhang et al. 2018), improving the performance significantly.

Co-segmentation: This was first introduced by Rother et al. (2006) for simultaneous binary segmentation of object parts in an image pair and extended to simultaneous segmentation of multiple images (Batra et al. 2010). Multi-view co-segmentation in space and time was introduced in Djelouah et al. (2016). A common foreground is obtained from multiple views using the information from appearance and motion cues. Semantic co-segmentation methods from a single video use spatio-temporal object proposals (Joulin et al. 2012; Luo et al. 2015), segments (Kolev et al. 2012), motion (Rother et al. 2006) and foreground propagation (Goldluecke and Magnor 2004). Recently, co-segmentation methods were introduced to segment common objects in a collection of videos for a single object (Maninis et al. 2018; Fu et al. 2014) or multiple objects (Tokmakov et al. 2019; Chiu and Fritz 2013; Zhong and Yang 2016). A CNN method for both single and multiple object segmentation was introduced in Khoreva et al. (2019), exploiting an intuitive training strategy from less data.

2.2 Semantic Flow Estimation

Methods have been proposed to exploit semantic information to improve monocular flow or motion estimate per frame (Li et al. 2018; Behl et al. 2017; Sevilla-Lara et al. 2016a; Zhu et al. 2018; Tsai et al. 2016). Semantic 2D detections were exploited to improve the tracking for autonomous driving in Li et al. (2018). Advantages of segmentations, bounding boxes and object coordinates to flow estimation were reviewed in Behl et al. (2017) for the case of dynamic road scenes. Sevilla-Lara et al. (2016a) exploit the advances in static semantic segmentation to segment the image into objects of different types followed by modelling motion for each object depending on the type of object. However for non-rigid dynamic objects such as people defining a unique motion model for the entire object is not effective. A method to exploit flow information for video segmentation was proposed in Tsai et al. (2016), reporting improvement in video segmentation exploiting flow information. However all of these methods either work for street scenes or static scenes and do not exploit any stereo or multiple view information.

2.3 Joint Estimation

General multi-view image segmentation methods use appearance and contrast information which may not be sufficient in the case of complex real world scenes. To improve the results joint optimisation of segmentation with 3D reconstruction has been proposed (Mustafa et al. 2016a) by including the multiple view photo-consistency. This concept was extended to semantic segmentation and reconstruction to obtain additional information from the scene (Jiao et al. 2018; Hane et al. 2016; Xie et al. 2016). Methods were introduced to utilize appearance-based pixel categories and stereo cues in a joint framework for street scenes from a monocular camera (Vineet et al. 2015; Floros and Leibe 2012). These methods used CRF to perform simultaneous dense reconstruction and segmentation of street scenes captured from a moving camera. A method to estimate pose and shape of people was proposed in Zanfir et al. (2018) and another method to estimate the pose and 3D shape of rigid objects on street scenes was proposed (Engelmann et al. 2016). An unsupervised method to jointly learn depth and flow using cross-task consistency was proposed for monocular video (Zou et al. 2018). Another method jointly estimates dense depth, optical flow and camera pose (Yin and Shi 2018). Recently a method was proposed for joint unsupervised learning of depth, camera, motion, optical flow and motion segmentation (Ranjan et al. 2018). However these methods cannot be directly applied to multi-view wide-baseline scenes. A method for joint estimation of 3D geometry and pose was proposed for rigid objects (Tulsiani et al. 2018). Dense semantic reconstruction of rigid objects was proposed by Bao et al. (2013). Joint semantic segmentation and reconstruction using multiple images was proposed for static scenes (Hane et al. 2013). However, these methods are limited to static scenes and rigid objects.

Joint motion and reconstruction or segmentation (Roussos et al. 2012; Sevilla-Lara et al. 2016b) methods were proposed for dynamic scenes. Techniques have been introduced to align dense meshes using correspondence information between consecutive frames (Zanfir and Sminchisescu 2015; Mustafa et al. 2016b) or extracting the scene flow by estimating the pairwise surface or volume correspondence between reconstructions at successive frames (Wedel et al. 2011; Basha et al. 2010). State-of-the-art joint estimation methods give per frame reconstruction and semantic segmentation of the scenes (Chen et al. 2019; Kendall et al. 2017) exploiting a multi-task learning framework. However these methods do not align meshes for the entire sequence, give seman-

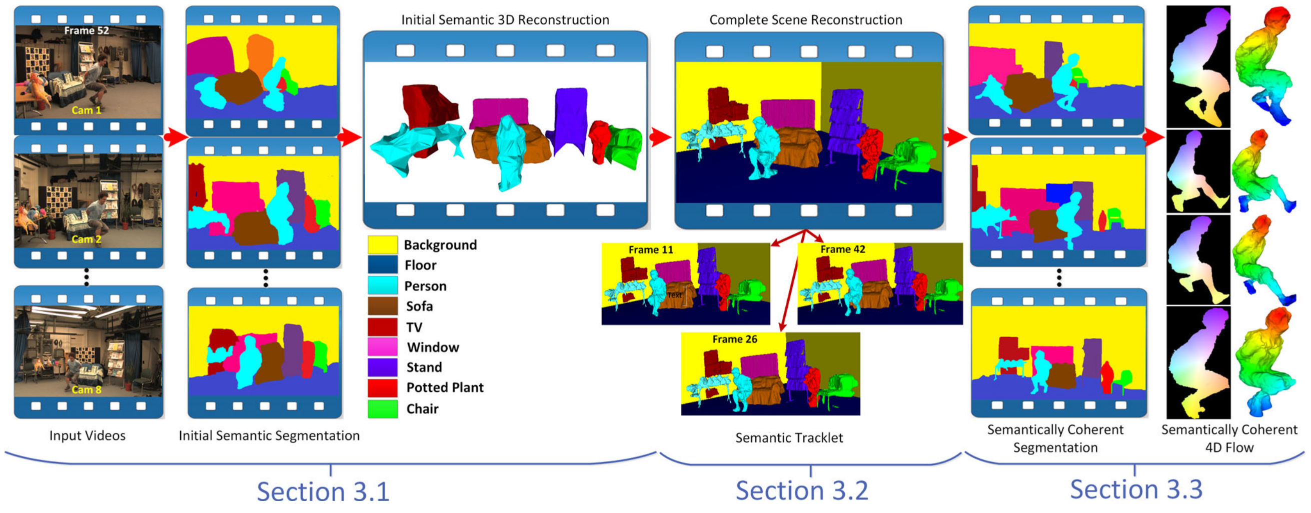


Fig. 2 Semantically coherent co-segmentation, reconstruction and flow estimation framework

tically coherent segmentation, or work for wide-baseline scenes. Our previous work (Mustafa and Hilton 2017) gives per-frame semantic segmentation and reconstruction of dynamic scenes, leading to unaligned meshes for dynamic sequence. The proposed method estimates 4D scene flow along with reconstruction and semantic co-segmentation, aligning meshes for entire dynamic sequence giving long-term semantic 4D scene flow.

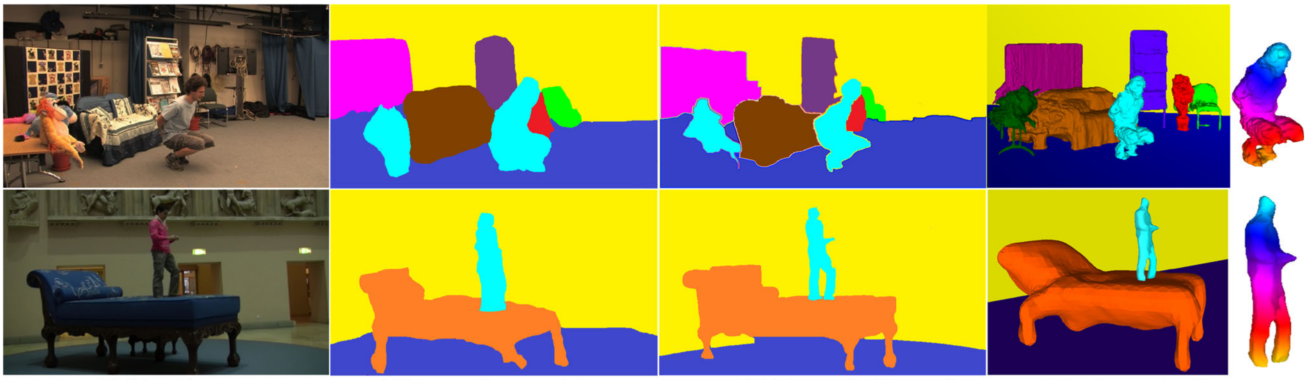
This paper introduces joint semantic flow, co-segmentation, and reconstruction enforcing coherence in both the spatial and temporal domains for scenes, with rigid and non-rigid dynamic objects, captured with multiple wide-baseline moving cameras. A key contribution of our work is that we combine semantics, shape, motion and appearance information in space and time in a single optimization to generate results automatically. The per-view motion, depth and semantic segmentation is combined across views and time for entire dynamic sequence to obtain 4D semantic flow. Evaluation demonstrates improved accuracy and completeness of flow, segmentation and reconstruction for complex dynamic scenes.

3 Semantic 4D Scene Flow and Segmentation

Overview:
This section gives an overview of the proposed framework for semantic temporal coherence, illustrated in Fig. 2. It comprises of following stages:

- **Input:** Multi-view videos are input to the system.
- **Initial Semantic Segmentation—Sect. 3.1:**
  Initial semantic labels are estimated for each pixel in the image per-view using state-of-the-art semantic segmentation (He et al. 2017; Chen et al. 2018).
- **Initial Semantic Reconstruction—Sect. 3.1:**
  Semantic information for each view is combined with sparse 3D feature correspondence between views to obtain an initial semantic 3D reconstruction. This initial reconstruction combines semantic information across views but results in inconsistency due to inaccuracies in the initial per-view segmentation.
- **Semantic Tracklets—Sect. 3.2:**
  To enforce long-term semantic coherence temporally we propose *semantic tracklets* that identify a set of similar frames for each dynamic object. Similarity between any pair of frames is estimated from the per-view semantic labels, appearance, shape and motion information.
  Semantic trackets provide a prior for the joint space-time semantic co-segmentation and reconstruction to enforce temporal coherence.
- **Joint Semantic Flow, Co-segmentation and Reconstruction—Sect. 3.3:** The initial semantic segmentation and reconstruction is refined per-view for each dynamic object through joint optimisation of flow, segmentation, and shape across multiple views and over time using the semantic tracklets. Per-view information is merged into a single 3D model using Poisson surface reconstruction (Kazhdan et al. 2006).
- **Semantic 4D Scene Flow and Segmentation—Sect. 3.3:**
  The process is repeated for the entire sequence and is combined across views and in time to obtain semantically coherent long-term dense 4D scene flow, co-segmentation, and reconstruction for the complete scene.

The following sections include a detailed explanation of the proposed approach and highlight the novel contributions.



(a) Original Image Initial semantic segmentation **(c)** Joint semantic 4D scene flow, segmentation & reconstruction

Fig. 3 The improvement of semantic segmentation using the proposed framework for Odzemok and Mgician datasets

3.1 Initial Segmentation and Reconstruction

Initial Semantic Segmentation: Mask RCNN is used for initial semantic segmentation because it is the state-of-the-art object detector that computes per instance masks and per instance class labels. This adopts a two-stage procedure to predict semantic segmentation of images. The object masks from Mask RCNN (He et al. 2017) are combined with background segmentation (Chen et al. 2018) to obtain dense semantic segmentation mask. For each frame in the sequence we perform deep semantic segmentation which estimates the probabilities of various classes at each pixel in the image. The network is trained on MS-COCO (Lin et al. 2014) dataset with 81 classes and is refined on PASCAL VOC12 (Everingham et al. 2012) dataset. In spite of being the state-of-the-art method the masks output still do not accurately align with the object boundaries as illustrated in Fig. 3b.

Initial Semantic Reconstruction: Sparse feature-based reconstruction of the scene is performed using SFD features (Mustafa et al. 2019) and SIFT descriptor (Lowe 2004) with the constraint that each 3D feature should be visible in 3 or more camera views for robustness (Hartley and Zisserman 2003). The resulting point-cloud is clustered in 3D (Rusu 2009). Clusters are formed between points with the same class labels across multiple views such that each cluster represents a semantically consistent object. Insufficient 3D features may occur on parts of an object due to lack of texture or visual ambiguity. To avoid incomplete reconstruction the sparse 3D object clusters are combined with the initial semantic segmentation to obtain the initial semantic reconstruction. A mesh is obtained for sparse 3D point clusters by triangulation to obtain an initial coarse reconstruction for each object. The initial coarse reconstruction is back-projected in each view onto the initial semantic segmentation. If the back-projected mask is smaller than its respective semantic region in 2 or more views then the initial coarse reconstruction is dilated in volume(3D) by v to enclose the object to match the segmentation boundaries in each view: $v = \frac{1}{N_h} * \sum_{c=1}^{N_h} \frac{B_s^c - B_r^c}{B_s^i}$, where N_h is the number of views with smaller back-projected mask, B_s^i is the area of the semantic segmentation and B_r^i is the area of the back-projected mask of the initial coarse reconstruction. This automatically initializes the reconstruction of each object in the scene without any strong initial prior.

3.2 Semantic Tracklets

In the case of general dynamic scenes with non-rigid objects, independent per-frame scene flow estimation, segmentation and reconstruction leads to incoherent results, for example failure to predict flow and reconstruct thin structures such as limbs and poorly localized object boundaries. Sequential methods for frame-to-frame temporal coherence are prone to errors due to drift and rapid motion (Beeler et al. 2011; Prada et al. 2016). Previous work Zhong and Yang (2016) introduced semantic tracklets for object segmentation in single view video based on co-segmentation across video collections. In this paper to achieve long-term scene flow, semantic co-segmentation and robust temporally coherent 4D reconstruction by introducing *semantic tracklets* which link instances of dynamic objects across wide-timeframes. This provides a prior to constrain long-term flow, co-segmentation and reconstruction. In our work semantic tracklets are defined for multiple views of the same dynamic scene to ensure temporal and spatial coherence in semantic 4D flow and 2D labelling, whereas in Zhong and Yang (2016) tracklets segment objects in a single video and relate them to similar object instances in multiple videos.

Semantic tracklets for a dynamic object are defined as a set of frames which have similar motion across 3 or more views, semantic labels, appearance and 2D shape as illustrated in Fig. 4. Tracklets are used for long-term learning of flow, semantic labels, appearance and shape information for per-view joint semantic 4D scene flow, co-segmentation and

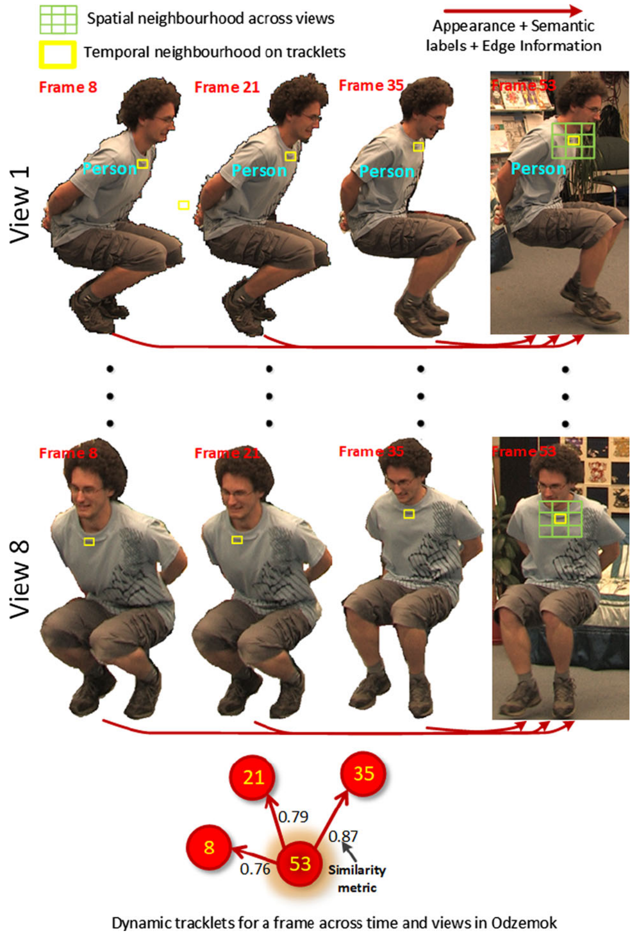

Spatial neighbourhood across views
Temporal neighbourhood on tracklets
Appearance + Semantic labels + Edge Information
Frame 8
Frame 21
Frame 35
Frame 53
View 1
Person
Person
Person
Person
View 8
Frame 8
Frame 21
Frame 35
Frame 53
21
35
8
53
0.79
0.87
0.76
Similarity metric
Dynamic tracklets for a frame across time and views in Odzemok

Fig. 4 Example of dynamic tracklet generation (similar frames) for a dynamic object at current frame 53 based on appearance, shape and semantic information. The spatial and temporal neighbourhood are shown at the top in green and yellow respectively for the optimization (Color figure online)

reconstruction of each object. This improves the temporal and semantic coherence in flow, reconstruction and segmentation results as shown in Fig. 5.

Dynamic objects are identified in the scene using motion information from sparse temporal SFD feature correspondences with SIFT descriptors. The semantic, 2D shape, motion and appearance similarity of the dynamic object is evaluated for each frame against all previous frames to identify the set of similar frames which form a tracklet. Similarity metric is defined as follows:

$$S_{i,j} = \frac{1}{4N_v} \sum_{c=1}^{N_v} (C^c_{i,j} + M^c_{i,j} + J^c_{i,j} + L^c_{i,j}) \tag{1}$$

where $C()$ is the measure of appearance similarity, $M()$ is the measure of motion similarity, $J()$ is the measure of shape similarity and $L()$ is the measure of semantic similarity. N_v is the number of views at each frame. These similarities are combined across time and views and all frames with similarity > 0.75 are selected as N_S similar frames to form a semantic tracklet T_i for each dynamic object at the i^{th} frame, $T_i = \{t_r\}_{r=1}^{N_S}$, where $t_r \in [0, i-1]$. An example of the frame-to-frame similarities is illustrated in Fig. 6 for Juggler sequence, depicting the differences in various measures and the overall similarity matrix.

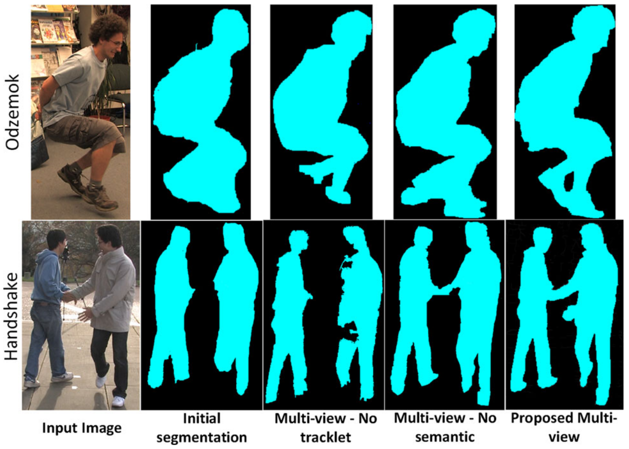


Fig. 5 Comparison of segmentation of the proposed multi view optimization against optimization with no semantic and no tracklet information respectively for Handshake and Odzemok datasets

Semantic Similarity: The semantic region associated with the object at each frame is identified using sparse wide-timeframe SFD feature matches combined with SIFT descriptor. An affine warp (Evangelidis and Psarakis 2008) based on the feature correspondence and region boundary is employed to transfer the semantic region segmentation to the current frame. The semantic similarity metric $L^c_{i,j}$ is defined as the ratio of the number of pixels with the same class label $z^c_{i,j}$ to the total number or pixels in the segmented region $y^c_{i,j}$ at frame i and j for view c:

$$L^c_{i,j} = \frac{z^c_{i,j}}{y^c_{i,j}} \tag{2}$$

Appearance Similarity: The appearance metric $C^c_{i,j}$ between frame i and j for the semantic region segmentation in view c corresponding to a dynamic object is based on the ratio of the number of temporal feature correspondences which are consistent across three or more views $q^c_{i,j}$ to the total number of feature correspondence in the segmented region $u^c_{i,j}$ (Mustafa et al. 2016b):

$$C^c_{i,j} = \frac{q^c_{i,j}}{u^c_{i,j}} \tag{3}$$

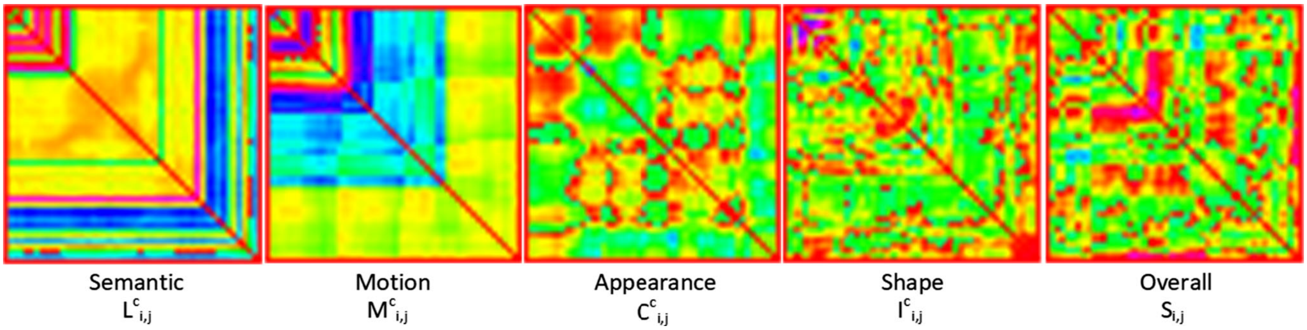


Fig. 6 Similarity matrix for each component in semantic tracklet estimation, along with all the components combined matrix

Motion Similarity: The motion metric $M^c_{i,j}$ between frame i and j for the semantic region segmentation in view c corresponding to a dynamic object is based on the average motion for the object across three or more views $s^c_{i,j}$ to the maximum motion between frames for entire sequence max (Mustafa et al. 2016b):

$$M^c_{i,j} = \frac{s^c_{i,j}}{max} \tag{4}$$

Shape Similarity: This metric gives a measure of the 2D region shape similarity between pairs of frames for each dynamic object. Semantic region segmentations are aligned using an affine warp (Evangelidis and Psarakis 2008). This is defined as the ratio of the intersection of the aligned segmentation $h^c_{i,j}$ to the union of the area $a^c_{i,j}$:

$$J^c_{i,j} = \frac{h^c_{i,j}}{a^c_{i,j}} \tag{5}$$

Importance of Semantic Tracklet: Semantic tracklets provide both temporal and multi-view priors for semantic 4D long-term flow estimation and co-segmentation. This is the importance of semantics in obtaining improved scene flow, segmentation and 4D reconstruction. Comparison is presented for optimization with/without semantic label and temporal tracklet information for multiple views in Fig. 5. Semantic tracklets result in significant improvement in scene flow estimation, reconstruction and multi-view video segmentation in comparison to state-of-the-art methods, as demonstrated in Sect. 4. The importance of the proposed semantically coherent optimization exploiting the information from semantic labels and tracklets for proposed multi-view joint optimization is shown in the Fig. 5. The proposed approach consistently performs better giving a more accurate flow and segmentation. The final proposed multiple view 4D flow, co-segmentation and reconstruction method using both semantic labels and tracklets gives significantly improved and more robust 4D flow and segmentation.

3.3 Joint Semantic Scene Flow, Co-segmentation and Reconstruction

The goal of multi-view joint semantic flow estimation, co-segmentation and reconstruction is to refine the initial semantic reconstruction obtained in Sect. 3.1 for each dynamic object for the region $\mathscr{R}$ per-view by optimizing the following variables: (a) Translation for each pixel location $p = (x_p, y_p)$ in image I, $m_p = (\delta x_p, \delta y_p)$ in time from a predefined set of flow vectors $\mathscr{M}$; (b) A semantic label from a set of semantic classes obtained as an initialization (Sect. 3.1), $\mathscr{L} = \left\{l_1, \ldots, l_{|\mathscr{L}|}\right\}$, to each pixel p for the initial semantic segmentation region $\mathscr{S}$ of each object, where $|\mathscr{L}|$ is the total number of classes in the network; and (c) An accurate depth value is jointly assigned for each pixel p from a set of depth values $\mathscr{D} = \left\{d_1, \ldots, d_{|\mathscr{D}|-1}, \mathscr{U}\right\}$, where d_i is obtained by sampling the optical ray from the camera and $\mathscr{U}$ is an unknown depth value to handle occlusions.

Long-term 4D flow and co-segmentation is achieved by propagating the semantic labels across views and over time using tracklets in the framework. Formulation of a cost function for semantically coherent depth and motion estimation and co-segmentation is based on the following principles:

- Local spatio-temporal coherence: Spatially and temporally neighbouring pixels are likely have the same semantic labels if they have similar appearance.
- Multi-view coherence: The surface is photo-consistent and semantically consistent across multiple views.
- Depth variation: The depth at spatially neighbouring pixels within an object varies smoothly for most of the surface (except internal depth discontinuities).
- Long-term temporal coherence: The semantic labels on each object remain consistent across a long time-frames in a sequence.

The cost function enforces spatial and temporal constraints on the semantic, appearance, motion and shape. Temporal semantic coherence is enforced using tracklets based on dynamic object similarity $S_{i,j}$ Eq. 1. An example of multi-view semantic scene flow, segmentation and reconstruction is shown in Fig. 3c. Enforcing temporal coherence with semantic tracklets for a multi-view video reduces noise in per-pixel labels. Errors in object segmentation remain due to the low spatial resolution of the initial semantic boundaries and visual ambiguity is addressed by combining information across multiple views. Joint optimisation of multiple view scene flow, co-segmentation and reconstruction minimises:

$$
\begin{aligned}
E(l,d,m) &= \lambda_d E_d(d) + \lambda_a E_a(l) + \lambda_c E_c(l) \\
&\quad + \lambda_{sm} E_{sm}(l,d) + \lambda_s E_s(l,d) + \lambda_m E_m(l,m)
\end{aligned} \tag{6}
$$

where, d is the depth at each pixel, m is the motion and l is the semantic label. $E_d()$ is the matching/depth cost, $E_a()$ is the appearance/color cost, $E_c()$ is the constrast cost, $E_{sm}()$ is the semantic labelling cost, $E_s()$ is the smoothness cost, and $E_m()$ is the motion/flow cost. Individual cost terms enforce spatial and temporal coherence for dynamic objects in semantic labels, appearance, region boundary contrast and motion cost. This is solved subject to a geodesic star-convexity constraint on the semantic labels l (Mustafa et al. 2016a):

$$
\underset{(l,d,m)}{min}\ \underset{l\in S^\star(\mathscr{C})}{E(l,d,m)} \Leftrightarrow \min_{(l,d,m)} E(l,d,m) + E^\star(l|x,\mathscr{C}) \tag{7}
$$

where $S^\star(\mathscr{C})$ is the set of all shapes which are geodesic star-convex wrt the features in $\mathscr{C} = \{c_1,\ldots,c_n\}$ within the initial semantic segmentation $\mathscr{R}$. $E^\star(l|x,\mathscr{C})$ is the geodesic star-convexity constraint enforced on the semantic labels l. α-expansion is used to iterate through the set of labels in $\mathscr{L} \times \mathscr{D} \times \mathscr{M}$ (Boykov et al. 2001) and a solution is obtained using graph-cuts (Boykov and Kolmogorov 2004) across spatial and temporal neighbourhoods as shown in Fig. 4. The initially reconstructed surface $\mathscr{R}$ is updated by minimizing the Energy in Eq. 6, by estimating the depth, segmentation and motion at each pixel within the projection of region $\mathscr{R}$ in each view.

Spatial neighbourhood: The spatial neighbourhood is defined as pairs of spatially close pixels in the image domain. A standard 8-connected spatial neighbourhood is used denoted by ψ_S; the set of pixel pairs (p,q) such that p and q belong to the same frame and are spatially connected.

Temporal neighbourhood: The temporal neighbourhood is defined based on the set of tracklets T_i generated for any frame i. Optical flow is used to compute a dense flow field on the tracklets, initialized from the sparse temporal SIFT feature correspondences. EpicFlow (Revaud et al. 2015) is used to preserve large displacements as the tracklets are distributed widely in time, and forward-backward flow consistency is enforced. Optical flow vectors define the temporal neighbourhood $\psi_T = \left\{(p,q) \mid q = p + d_{i,j}\right\}$; where j is the number of a frame in tracklet $T_i = \{j = t_r\}$, and $d_{i,j}$ is the displacement vector from image i to j.

Semantic cost $E_{sm}(l,d)$: This term enforces multi-view consistency on the semantic labels of each pixel p. Inconsistent labels across views are penalised to ensure semantic coherence. This cost is computed based on the probability of the class labels at each pixel for the initial semantic segmentation (Chen et al. 2016). Unlike previous approaches to achieve semantic coherence we enforce spatial and temporal consistency using tracklets across the neighbourhoods. The term is defined as:

$$
E_{sm}(l,d) = \sum_{p\in\psi_S} e_{sm}(p,d_p,l_p)
$$

$e_{sm}(p,d_p,l_p) = \sum_{c=1}^{N_K} z(p,r,l_p)$, if $d_p \neq \mathscr{U}$ else a fixed cost $S_{\mathscr{U}}$ is assigned. A $3D$ point $P(p,d_p)$ is assumed along the optical ray passing through pixel p located at a distance d_p from the reference camera. The projection of hypothesized point $P(p,d_p)$ in view c is defined by $r = \phi_c(P)$. N_K is the total number of views in which point $P(p,d_p)$ is visible.

$$
z(p,r,l_p) = \begin{cases} -log\, P_{sem}(I_p|l_p) & \text{if } l_p = l_r \\ -log\left(1 - P_{sem}(I_p|l_p)\right) & \text{if } l_p \neq l_r \end{cases}
$$

where l_r is the semantic label at pixel r in view c and $P_{sem}(I_p|l_p = l_i)$ denotes the probability of the semantic label l_i at pixel p in the classification image obtained from initial semantic segmentation.

Contrast cost $E_c(l)$: The contrast cost (Chen et al. 2016) is modified to introduce spatial and temporal semantic coherence and ensure that for dynamic objects the region boundaries have high contrast. Semantic region boundaries are propagated using the tracklets as a prior for the optimization:

$$
\begin{aligned}
E_c(l) = &\sum_{p,q\in\psi_T} e_c(p,q,l_p,l_q,\sigma_\alpha^t,\vartheta_{pq}^t,\sigma_\beta^t) \\
&+ \sum_{p,q\in\psi_S} e_c(p,q,l_p,l_q,\sigma_\alpha^s,\vartheta_{pq}^s,\sigma_\beta^s) \\
&e_c(p,q,l_p,l_q,\sigma_\alpha,\vartheta_{pq},\sigma_\beta) = \mu\left(l_p,l_q\right) \\
&\times \left(\lambda_{ca} exp^{-\left(\frac{\|B(p)-B(q)\|^2}{2(\sigma_\alpha)^2(\vartheta_{pq})^2}\right)} + \lambda_{cl} exp^{-\left(\frac{\|L(p)-L(q)\|^2}{2(\sigma_\gamma)^2}\right)}\right)
\end{aligned}
$$

where $\mu\left(l_p,l_q\right) = 1$ if $(l_p \neq l_q)$ else 0 and ϑ_{pq} is the Euclidean distance between pixel p and q. The first Gaussian kernel is a bilateral kernel which depends on RGB color ($B()$ is bilateral filtered image) and pixel positions, and the second kernel only depends on pixel positions L. The parameters σ_α, σ_β and σ_γ control the scale of the Gaussian kernels.

The first kernel forces pixels with similar color and position to have similar labels, while the second kernel only considers semantic spatial proximity when enforcing smoothness. The value of $\sigma_\alpha = \left\langle \frac{\|B(p)-B(p)\|^2}{\vartheta_{pq}^2} \right\rangle$, with the operator $\langle\rangle$ denoting the mean computed across the neighbourhoods ψ_S and ψ_T for spatial and temporally coherent contrast respectively.

Appearance cost $E_a(l)$: This cost is computed using the negative log likelihood (Boykov and Kolmogorov 2004) of the color models learned from the foreground object and background. In this work the foreground models are learnt from the sparse features of the dynamic object in the current frame and foreground regions from tracklets to improve the consistency of the results. Static background models are learnt from the sparse features outside the initial semantic segmentation of the dynamic object in the current frame and the region outside the semantic segmentation in the tracklets. Appearance cost is defined as:

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

where $P(I_p|l_p = l_i)$ is the probability of pixel p in the reference image belonging to label l_i. Color models use GMMs with 10 components each for foreground or background.

Matching cost $E_d(d)$: The photo-consistency matching cost across views is defined as:

$$E_d(d) = \sum_{p \in \psi_S} e_d(p, d_p)$$

where $e_d(p, d_p) = \sum_{i \in \mathcal{O}_k} m(p, r)$, if $d_p \neq \mathcal{U}$ else $M_{\mathcal{U}}$. $m(p, r)$ is inspired from Hu and Mordohai (2012). $M_{\mathcal{U}}$ is the fixed cost of labelling a pixel unknown. r denotes the projection of the hypothesised point P in an auxiliary camera where P is a 3D point along the optical ray passing through pixel p located at a distance d_p from the reference camera. $\mathcal{O}_k$ is the set of k most photo-consistent pairs with a reference camera across views. $\mathcal{O}_k$ are identified using the highest number or feature matches spatially across frames.

Motion cost $E_m(l, m)$: This adds the brightness consistency assumption to the cost function generalized for spatial and temporal neighbourhood, defined as:

$$E_m(l, m) = \sum_{p \in \psi_T} \lambda_l E_l(p, m_p, l_p) + \lambda_c E_c(p, m_p, l_p)$$

$$E_l(p, m_p, l_p) = \sum_{i=1}^{N_v} \left\|(I_i(p, t) - I_i(p + m_p, t + 1))\right\|^2$$
$$\text{if at t and t+1 } l_p = l_{p+m_p} \text{ else } 0$$

$$E_c(p, m_p, l_p) = \sum_{p \in \psi_T} \sum_{i=2}^{N_v} \left\|(I_1(p, t) - I_i(p + m_p, t))\right\|^2$$
$$\text{if at t } l_p = l_{p+m_p} \text{ else } 0$$

$E_l()$ penalizes deviation from the brightness constancy assumption in time for a single view. Term $E_c()$ penalizes deviation from the brightness constancy assumption between the reference view and each of the other views at other time instants. Here N_v is the number of views at each time frame and $I_i(p, t)$ is the intensity at a given pixel p at time instant t in view i, ψ_S and ψ_T are the spatial and temporal neighbourhood.

This term denotes that the flow vector m_p is located within a window from a sparse constraint at p and it forces the flow to approximate the sparse 2D temporal correspondences.

Smoothness cost $E_s(l, d)$: The surface smoothness cost introduced in Mustafa et al. (2016a) is extended to spatial and temporal neighbourhoods:

$$E_s(l, d) = \lambda_s^t \sum_{p,q \in \psi_T} e_s(l_p, d_p, l_q, d_q, d_{max}^t)$$
$$+ \lambda_s^S \sum_{p,q \in \psi_S} e_s(l_p, d_p, l_q, d_q, d_{max}^s)$$
$$e_s(l_p, d_p, l_q, d_q, d_{max})$$
$$= \begin{cases} min(\left|d_p - d_q\right|, d_{max}), & \text{if } l_p = l_q \text{ and } d_p, d_q \neq \mathcal{U} \\ 0, & \text{if } l_p = l_q \text{ and } d_p, d_q = \mathcal{U} \\ d_{max}, & \text{otherwise} \end{cases}$$

d_{max} is introduced to avoid over-penalising large discontinuities. d_{max}^s ensures spatial smoothness and d_{max}^t ensures smoothness over time between the temporal neighbourhood of the tracklets and is set to twice d_{max}^s to allow large movement in the object between tracklet frames.

Semantically and Temporally Coherent Reconstruction The estimated dense flow for each view is projected to the 3D visible surface to establish dense 3D correspondence (scene flow) between frames and between semantic tracklets T_i to obtain 4D semantically and temporally coherent dynamic scene reconstruction, as illustrated in Fig. 1. Temporal correspondence is first obtained for the view with maximum visibility of 3D points. To increase surface coverage correspondences are added in order of visibility of 3D points for different views. Dense temporal correspondence is propagated to new surface regions as they appear using the dense flow estimated from joint refinement. Temporal coherence is also estimated between semantic tracklets to overcome the limitations of sequential correspondence propagation by correcting any errors introduced in semantically and temporally coherent reconstruction. As a result along with segmentation and reconstruction of dynamic scenes, we have temporal and semantic per-pixel correspondence information in both

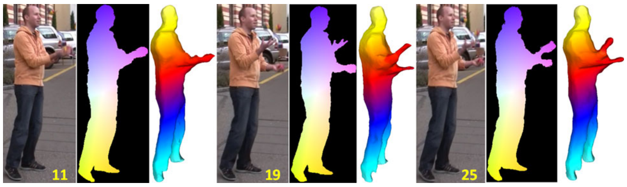


Fig. 7 3D temporal alignment between frames for Juggler dataset

2D and 3D, as shown for Juggler dataset in Fig. 7. The 2D per-view depth maps are combined using Poisson surface reconstruction (Kazhdan et al. 2006), which leads to loss in the details in mesh of the object compared to the semantic segmentation.

4 Results and Evaluation

Joint semantic co-segmentation, reconstruction and scene flow estimation (Sect. 3.3) is evaluated on a variety of publically available multi-view indoor and outdoor dynamic scene datasets, details in Table 1.

4.1 4D Flow Evaluation

We evaluate semantic and temporal coherence obtained using the proposed 4D semantic flow algorithm on all of the datasets. Stable long-term 4D correspondence propagation is illustrated using color coded results. First frame of the sequence is color coded and the colors are propagated between frames using the 2D-3D motion information obtained from the joint refinement explained in Sect. 3.3. Results of the proposed 4D temporal and semantic alignment, illustrated in Fig. 8 shows that the colour of the points remains consistent between frames. The proposed approach is qualitatively shown to propagate the correspondences reliably for complex dynamic scenes with large non-rigid motion.

For comparative evaluation we use:(a) state-of-the-art dense flow algorithm Deepflow (Weinzaepfel et al. 2013); (b) a recent algorithm for alignment of partial surfaces (4DMatch) (Mustafa et al. 2016b) and (d) Simple flow (Tao et al. 2012). Qualitative results against 4DMatch, Deepflow and Simpleflow shown in Fig. 9 indicate that the propagated colour map does not remain consistent across the sequence for large motion as compared to the proposed method (red regions indicate correspondence failure).

For quantitative evaluation we compare the silhouette overlap error (SOE). Dense correspondence over time is used to create propagated mask for each image. The propagated mask is overlapped with the silhouette of the projected surface reconstruction at each frame to evaluate the accuracy of the dense propagation. The error is defined as:

$$SOE = \frac{1}{MN} \sum_{i=1}^{N} \sum_{c=1}^{M} \frac{\text{Area of intersection}}{\text{Area of back-projected mask}} \tag{8}$$

Evaluation against the different techniques is shown in Table 2 for all datasets. As observed the silhouette overlap error is lowest for the proposed approach showing relatively high accuracy.

We evaluate the temporal coherence across the Magician sequence, by evaluating the variation in appearance for each scene point between frames and between semantic tracklets for state-of-the-art methods, defined as: $\sqrt{\frac{\Delta r^2+\Delta g^2+\Delta b^2}{3}}$, where Δ is the difference operator. Evaluation shown in Table 3 against state-of-the-art methods demonstrates the stability of long-term temporal tracking for proposed joint semantic scene flow, co-segmentation and reconstruction.

4.2 Segmentation Evaluation

Mutli-view co-segmentation is evaluated against a variety of state-of-the-art methods:
(a) Non-Semantic methods: Multi-view segmentation (MVVS) (Djelouah et al. 2016), Joint segmentation and reconstruction (TcMVS) (Mustafa et al. 2016a), and
(b) Semantic methods: Semantic co-segmentation in videos (SCV) (Zhong and Yang 2016), Mask RCNN (He et al. 2017)

Table 1 The characteristic properties of datasets used for evaluation

Dataset	Resolution	No. of views	Baseline
Odzemok (http://cvssp.org/data/cvssp3d/)	1920 × 1080	8 (2 moving)	25°
Handshake (Kim et al. 2012)	1920 × 1080	8 (all static)	20°
DogJump (http://4drepository.inrialpes.fr/)	1624 × 1080	16 (all static)	35°
Breakdance (Zitnick et al. 2004)	2048 × 2048	20 (all static)	15°
Magician (Ballan et al. 2010)	960 × 544	5 (all moving)	40°−55°
Juggler (Ballan et al. 2010)	960 × 544	6 (all moving)	25°−30°
HumanEva (Sigal et al. 2010)	640 × 480	3 (all static)	25°−30°
Human3.6 (Ionescu et al. 2014)	1000 × 1000	4 (all static)	25°−30°
Lightfield (Mustafa et al. 2017)	2048 × 2048	20 (all static)	5°

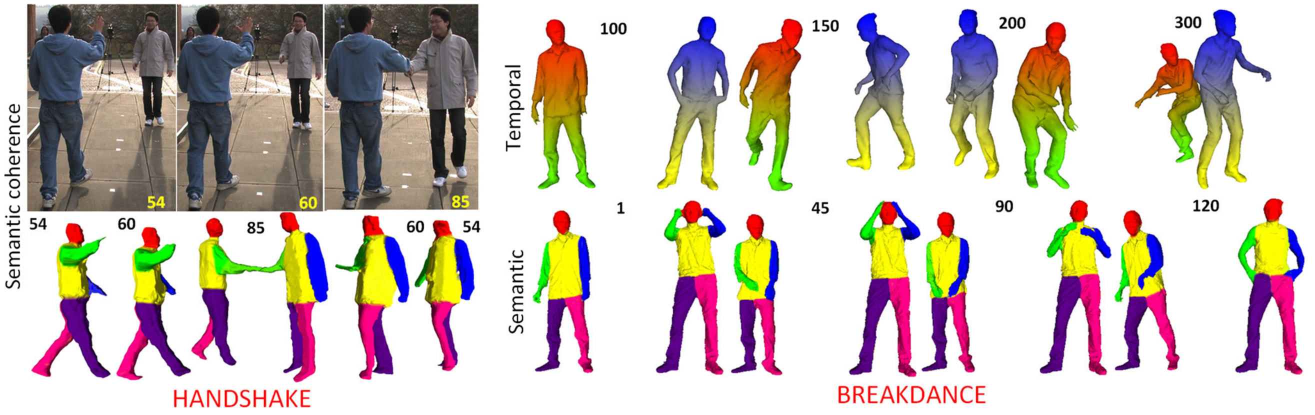


Fig. 8 Temporal and semantic coherence results using proposed approach on Handshake, Lightfield and Breakdance datasets. Color-coding for temporal coherence: Unique gradient colors are assigned to first frame of the sequence for each object. Color-coding for semantic coherence: head is red, left-arm is blue, right-arm is green, left-leg is pink and right-leg is violet. Colors are propagated using proposed 4D scene flow (Color figure online)

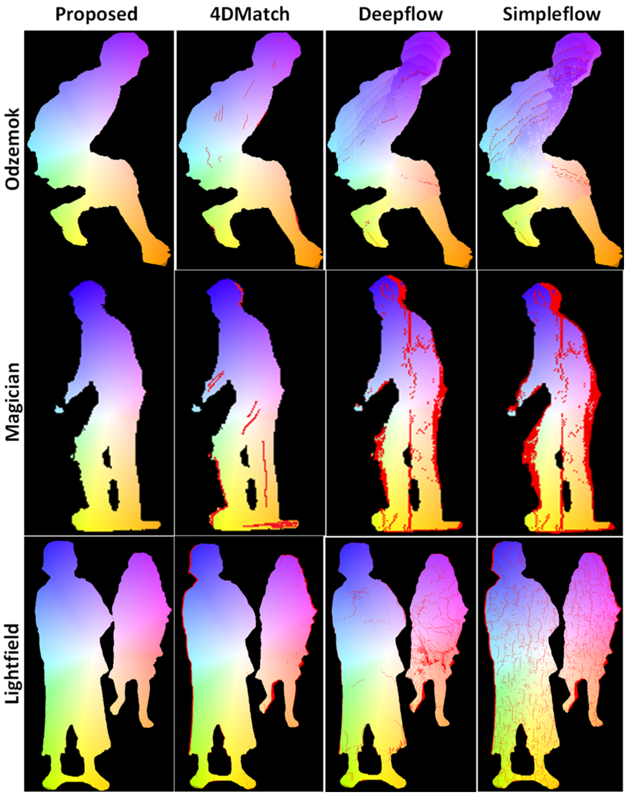


Fig. 9 Dense flow comparison results on different dynamic sequences

Table 2 Silhouette overlap error for multi-view datasets for flow evaluation, where SF is Simpleflow

Datasets	Proposed	4DMatch	Deepflow	SF
Breakdance	**0.60**	0.81	1.14	1.35
Odzemok	**0.66**	0.95	1.38	1.82
Magician	**0.78**	0.98	1.67	1.97
Juggler	**0.72**	0.92	1.53	2.11
HumanEva	**0.63**	0.83	1.46	1.89
Handshake	**0.81**	0.99	1.55	2.31
Dogjump	**0.69**	0.86	1.34	1.83
Human3.6M	**0.71**	0.95	1.66	1.97
Lightfield	**0.49**	0.71	1.13	1.83

Bold values represent best performance

Table 3 Temporal coherence evaluation for Magician dataset against existing methods

Methods	Frame-to-frame		Tracklets	
	Mean	SD	Mean	SD
Proposed	**4.67**	2.61	**5.63**	2.48
4DMatch	6.42	3.27	7.88	3.94
Deepflow	7.33	4.40	17.54	8.47
Simpleflow	8.21	4.60	23.30	9.55

Bold values represent best performance
SD standard deviation

and Conditional random field as recurrent neural networks (CRF-RNN) (Zheng et al. 2015).

Proposed segmentation is also evaluated against single-view segmentation methods MVC (Chiu and Fritz 2013) and ObMiC (Fu et al. 2014). These are applied independently on each view for comparison. Comparison against MVVS (Djelouah et al. 2016) is shown in Fig. 10 and evaluation against TcMVS (Mustafa et al. 2016a), SCV (Zhong and Yang 2016) and CRF-RNN (Zheng et al. 2015) are shown in Fig. 12 for dynamic datasets. Ground-truth segmentation comparison with TcMVS (Mustafa et al. 2016a) is shown in Fig. 11. Quantitative evaluation against state-of-the-art methods is measured by Intersection-over-Union with ground-truth, shown in the Table 4. Ground-truth is available online for most of the datasets and obtained by

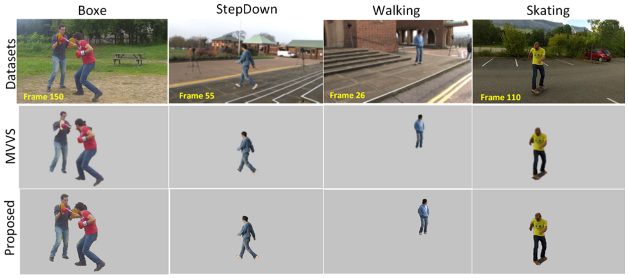


Fig. 10 Comparison of segmentation on dynamic datasets from Kim et al. (2012) and Djelouah et al. (2016) against MVVS (Djelouah et al. 2016)

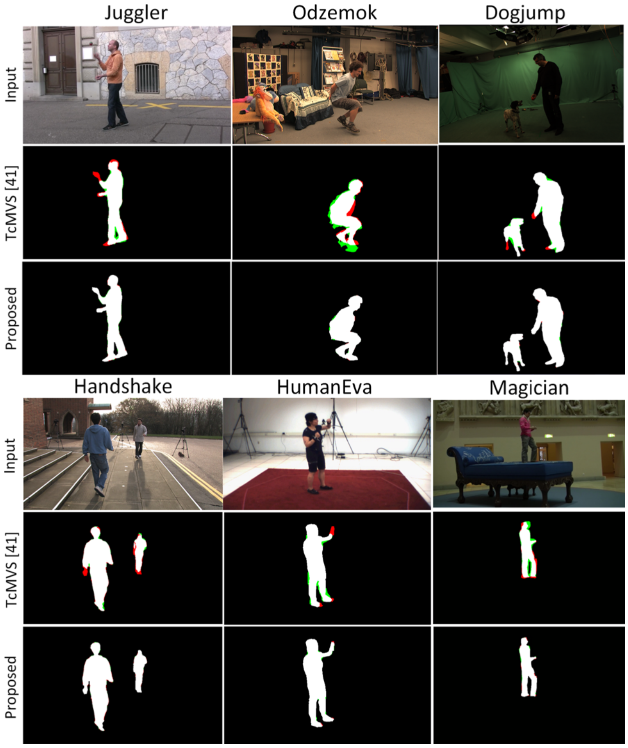


Fig. 11 Ground-truth segmentation comparison with TcMVS (Mustafa et al. 2016a) on multi-view datasets

manual labelling for other datasets. The proposed semantically coherent joint multi-view 4D flow, co-segmentation and reconstruction achieves the best segmentation performance against ground-truth for all datasets tested. Results presented in Fig. 12 indicate that the proposed approach accurately segments fine detail such as hands and feet where other approaches are unreliable.

Comparison with Tsai et al. (2016): Comparison to Zhong and Yang (2016) is shown in Fig. 12 and Table 4. The results show that the proposed approach achieves a significant improvement for multi-view video segmentation compared to co-segmentation approach using tracklets (Zhong and Yang 2016) (average 45% improvement in intersection-over-union of the segmentation vs. ground-truth).

Table 4 Segmentation comparison against state-of-the-art methods using the *intersection-over-union* metric

Methods	Breakdance	HumanEva	Odzemok	Magician	Juggler	Dogjump	Handshake	Human3.6	Lightfield
MVC (Chiu and Fritz 2013)	36.5	42.1	38.2	34.8	39.7	41.6	44.8	35.0	43.5
ObMiC (Fu et al. 2014)	39.4	49.6	45.5	41.4	44.0	45.9	48.1	40.9	47.9
CRF-RNN (Zheng et al. 2015)	61.0	71.4	41.0	53.3	70.8	52.3	64.6	40.8	66.8
Mask RCNN (He et al. 2017)	67.9	74.8	69.7	67.1	70.5	62.2	59.8	57.5	67.2
SCV (Zhong and Yang 2016)	48.9	51.0	53.3	61.0	56.6	60.2	49.5	47.5	51.9
TcMVS (Mustafa et al. 2016a)	89.1	94.0	91.8	91.2	93.3	89.4	86.5	63.8	75.2
Proposed	**94.9**	**96.7**	**95.6**	**94.8**	**95.8**	**93.1**	**92.2**	**78.4**	**88.7**

Bold values represent best performance

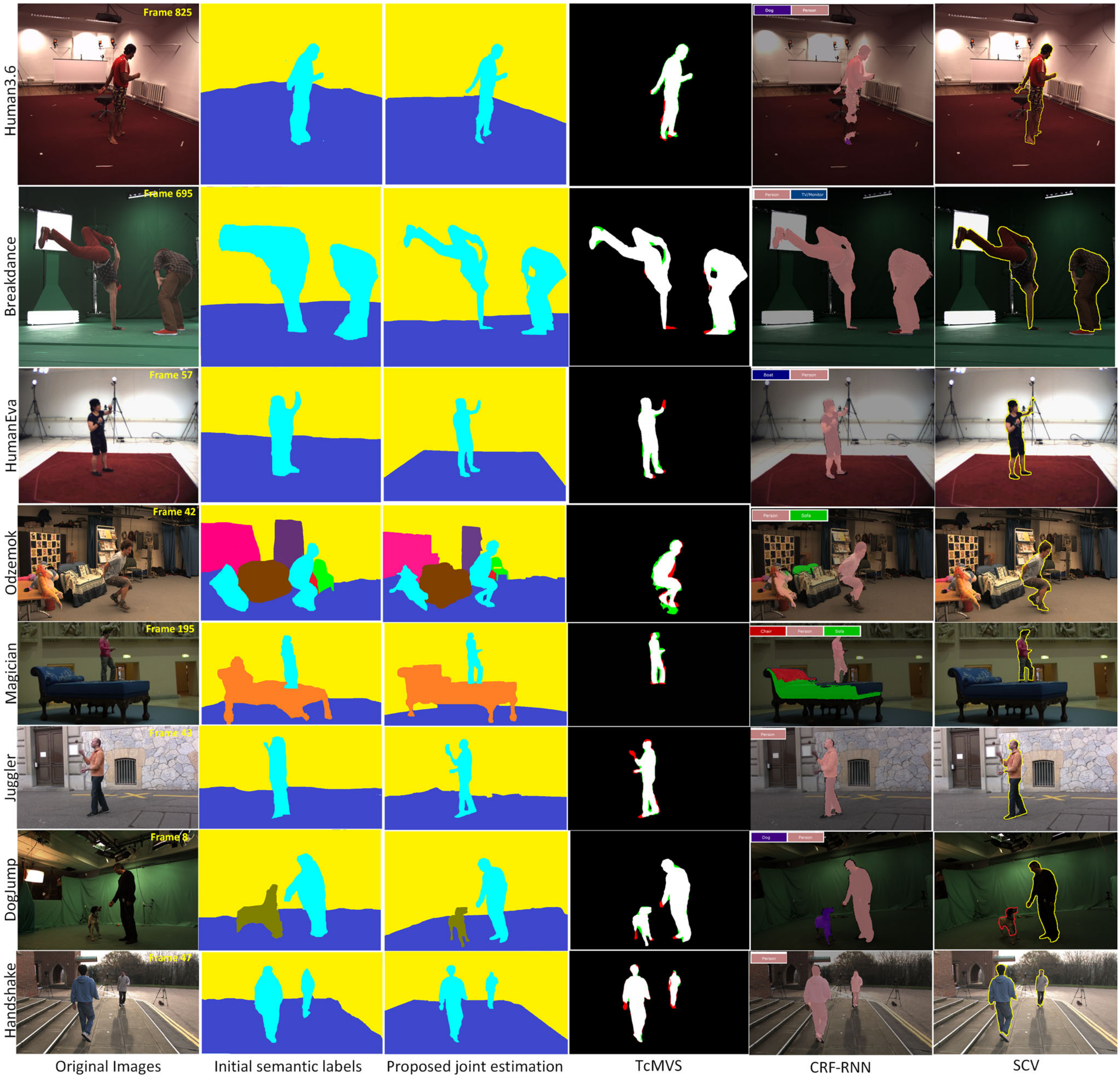


Fig. 12 Comparison of segmentation on public datasets against state-of-the-art methods: TcMVS (Mustafa et al. 2016a) (region in red represents region missing from ground-truth and green represents region not present in ground-truth), CRF-RNN (Zheng et al. 2015) and SCV (Zhong and Yang 2016) (Color figure online)

4.3 Reconstruction Evaluation

The reconstruction results obtained from the proposed approach are compared against state-of-the-art approaches in joint segmentation and reconstruction (TcMVS Mustafa et al. 2016a) and multi-view stereo (Colmap Schönberger et al. 2016, MVE Semerjian 2014, SMVS Langguth et al. 2016). MVE, SMVS and Colmap are state-of-the-art multi-view stereo techniques which do not refine the segmentation. All the methods are initialized with the same initial semantic reconstruction (Sect. 3.1) for fair comparison. Comparison of reconstructions Fig. 13 demonstrates that the proposed method gives consistently more complete and accurate models. Figure 14 presents a comparison to a statistical model-based approach MBR (Rhodin et al. 2016) which reconstructs a single human body shape from the whole sequence together with pose at each frame. This provides a good estimate of the underlying body shape but does not take into account cloth-

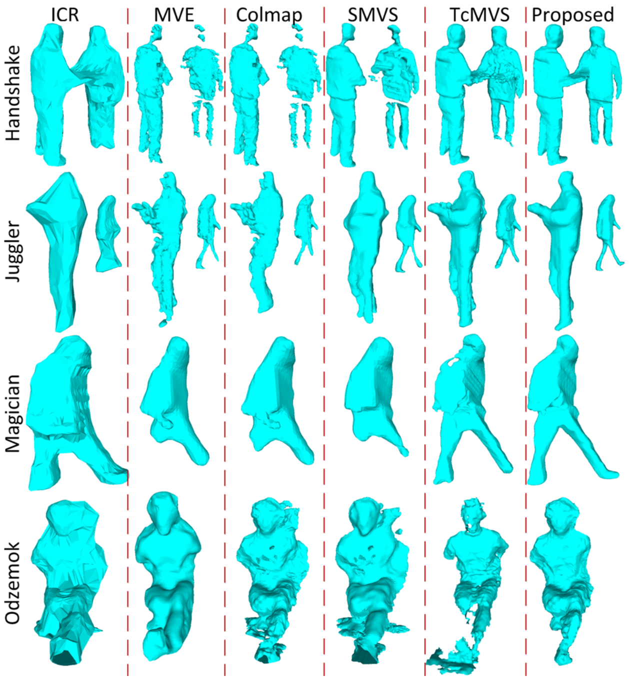


Fig. 13 Comparison of reconstruction of dynamic objects against Colmap (Schönberger et al. 2016), MVE (Semerjian 2014), SMVS (Langguth et al. 2016) and TcMVS (Mustafa et al. 2016a) (same semantic labels are assigned to all methods for fair comparison)

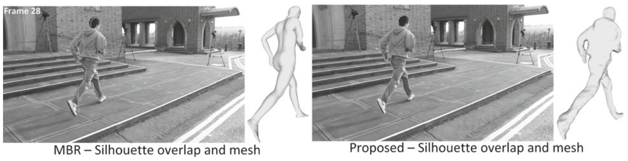

Frame 28
MBR – Silhouette overlap and mesh
Proposed – Silhouette overlap and mesh

Fig. 14 Comparison of reconstruction against MBR (Rhodin et al. 2016) from 4 views of falling down (Kim et al. 2012) dataset

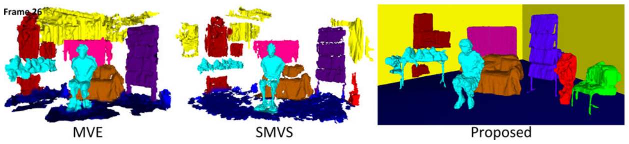


Fig. 15 Comparison of full scene reconstruction against SMVS (Schönberger et al. 2016) and MVE (Semerjian 2014) (same semantic labels are assigned to all the approaches for fair comparison)

Table 5 Parameters for all datasets

	E()					
	λ_d	λ_{sm}	λ_s	λ_m	λ_a	λ_c
Odzemok	1.0	0.7	0.2	0.4	0.5	5
Breakdance	1.0	0.7	0.2	0.4	0.5	5
DogJump	1.0	0.7	0.2	0.4	0.5	5
HumanEva	1.0	0.7	0.2	0.4	0.5	5
Lightfield	1.0	0.7	0.2	0.4	0.5	5
Juggler	1.0	0.9	0.1	0.6	0.6	7.5
Handshake	1.0	0.9	0.1	0.6	0.6	7.5
Magician	1.0	0.9	0.1	0.6	0.6	7.5
Human3.6	1.0	0.9	0.1	0.6	0.6	7.5

ing resulting in inaccurate silhouette overlap. Comparison of full scene reconstruction against MVE and SMVS is shown in Fig. 15 showing improved completeness and accuracy.

Joint semantic 4D scene flow, co-segmentation and reconstruction results in a 3D model for which every surface point has consistent surface labelling across all views and over time. To illustrate the semantic wide-timeframe coherence achieved using the proposed approach unique colors are assigned to human body parts in one frame and the colors are propagated using the estimated temporal coherence. The color in different parts of the object remains consistent over time as shown in Fig. 8.

Parameters: Results are insensitive to parameter setting for all indoor and outdoor scenes. Table 5 shows the parameters used, with constant contrast cost $\lambda_{ca} = \lambda_{cl} = 0.5$ and smoothness cost $\lambda_s^S = 0.4, \lambda_s^T = 0.6$.

Limitations: The proposed approach is dependent on an initial semantic labelling of the scene for each view obtained using Mask-RCNN. Gross errors or mislabeling may be propagated resulting in incorrect semantic reconstruction, such as the soft-toys labelled as people on the left hand side of the Odzemok dataset Fig. 2. Whilst enforcing semantic coherence is demonstrated to improve scene flow, segmentation and reconstruction for a wide-variety of scenes visual ambiguity in appearance and occlusion may degrade performance.

5 Conclusion

This paper proposes a novel approach to joint semantic 4D scene flow, multi-view co-segmentation and reconstruction of complex dynamic scenes. Temporal and semantic coherence is enforced over long-time frames by semantic tracklets identifying similar frames using the semantic label, appearance, shape and motion information. Tracklets are used for long-term learning to constrain flow per-frame and co-segmentation optimization on general dynamic scenes. Joint optimization simultaneously improves the scene flow, semantic segmentation and reconstruction of the scene by enforcing semantic coherence both spatially across views and temporal across widely-spaced similar frames. Comparative evaluation demonstrates that enforcing semantic coherence achieves significant improvement in scene flow and segmentation of general dynamic indoor and outdoor scenes captured with multiple hand-held cameras. Introduction of space-time semantic coherence in the proposed framework achieves better reconstruction and flow estimation against state-of-the-art methods.

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. We would like to thank Helge Rhodin and Abdelaziz Djelouah for providing their data.

References

4d repository. In *Institut national de recherche en informatique et en automatique (INRIA) Rhone Alpes*. http://4drepository.inrialpes.fr/.

Ballan, L., Brostow, G. J., Puwein, J., & Pollefeys, M. (2010). Unstructured video-based rendering: Interactive exploration of casually captured videos. *ACM Transactions on Graphics*, *29*(4), 1–11.

Bao, Y., chandraker, M., Lin, Y., & Savarese, S. (2013). Dense object reconstruction using semantic priors. In *The IEEE international conference on computer vision and pattern recognition (CVPR)*.

Basha, T., Moses, Y., Kiryati, N. (2010). Multi-view scene flow estimation: A view centered variational approach. In *The IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1506–1513).

Batra, D., Kowdle, A., Parikh, D., Luo, J., & Chen, T. (2010). icoseg: Interactive co-segmentation with intelligent scribble guidance. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., et al. (2011). High-quality passive facial performance capture using anchor frames. *ACM Transaction in Graphics*, *30*(4), 75:1–75:10.

Behl, A., Jafari, O. H., Mustikovela, S. K., Alhaija, H. A., Rother, C., & Geiger, A. (2017). Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *Proceedings IEEE international conference on computer vision (ICCV)*. IEEE.

Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *26*(11), 1124–1137.

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *23*(11), 1222–1239.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions in Pattern Analysis and Machine Intelligence (PAMI)*, *40*(4), 834–848.

Chen, L., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. CoRR arXiv:1412.7062.

Chen, L. C., 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, P.-Y., Liu, A. H., Wang, Y. C. F. (2019). Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Chiu, W. C., & Fritz, M. (2013). Multi-class video co-segmentation with a generative multi-video model. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Djelouah, A., Franco, J. S., Boyer, E., Le Clerc, F., & Perez, P. (2015). Sparse multi-view consistency for object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *37*(9), 1890–1903.

Djelouah, A., Franco, J. S., Boyer, E., Pérez, P., & Drettakis, G. (2016). Cotemporal multi-view video segmentation. In *International conference on 3D vision (3DV)*.

Engelmann, F., Stückler, J., & Leibe, B.(2016). Joint object pose estimation and shape reconstruction in urban street scenes using 3D shape priors. In *Proceedings of the German Conference on Pattern Recognition (GCPR)*.

Evangelidis, G. D., & Psarakis, E. Z. (2008). Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (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 (VOC2012) results*. Retrieved September 5, 2017 from 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *35*(8), 1915–1929.

Floros, G., & Leibe, B. (2012). Joint 2d-3d temporally consistent semantic segmentation of street scenes. In *The IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2823–2830).

Fu, H., Xu, D., Zhang, B., & Lin, S. (2014). Object-based multiple foreground video co-segmentation. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Goldluecke, B., & Magnor, M. (2004). Space–time isosurface evolution for temporally coherent 3d reconstruction. In *The IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 350–355).

Gupta, S., Girshick, R., Arbeláez, 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. (2013). Joint 3d scene reconstruction and class segmentation. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Hane, C., Zach, C., Cohen, A., & Pollefeys, M. (2016). Dense semantic 3d reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *39*, 1730–1743.

Hariharan, B., Arbeláez, P. A., Girshick, R. B., & Malik, J. (2015). Hypercolumns for object segmentation and fine-grained localization. In *The IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 447–456).

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision* (2nd ed.). Cambridge: Cambridge University Press.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. B. (2017). Mask R-CNN. CoRR arXiv:1703.06870.

Hu, X., & Mordohai, P. (2012). A quantitative evaluation of confidence measures for stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *34*(8), 2121–2133.

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1325–1339.

Jiao, J., Cao, Y., Song, Y., & Lau, R. (2018). Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *The European conference on computer vision (ECCV)*.

Joulin, A., Bach, F., & Ponce, J. (2012). Multi-class cosegmentation. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

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.

Khoreva, A., Benenson, R., Ilg, E., Brox, T., & Schiele, B. (2019). Lucid data dreaming for video object segmentation. *International Journal of Computer Vision (IJCV)*, *127*, 1175–1197.

Kim, H., Guillemaut, J., Takai, T., Sarim, M., & Hilton, A. (2012). Outdoor dynamic 3-D scene reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT)*, *22*(11), 1611–1622.

Kolev, K., Brox, T., & Cremers, D. (2012). Fast joint estimation of silhouettes and dense 3d geometry from multiple images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *34*(3), 493–505.

Kundu, A., Li, Y., Dellaert, F., Li, F., & Rehg, J. M. (2014). Joint semantic segmentation and 3d reconstruction from monocular video. *European Conference on Computer Vision (ECCV)*, *8694*, 703–718.

Kundu, A., Vineet, V., & Koltun, V. (2016). Feature space optimization for semantic video segmentation. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3168–3175).

Langguth, F., Sunkavalli, K., Hadap, S., & Goesele, M. (2016). Shading-aware multi-view stereo. In *European conference on computer vision (ECCV)*.

Larsen, E., Mordohai, P., Pollefeys, M., & Fuchs, H. (2007). Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *The IEEE international conference on computer vision (ICCV)* (pp. 1–8).

Li, P., Qin, T., & Shen, S. (2018). Stereo vision-based semantic 3d object and ego-motion tracking for autonomous driving. In *The European conference on computer vision (ECCV)*.

Lin, T. Y., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., et al. (2014). Microsoft COCO: Common objects in context. CoRR arXiv:1405.0312.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, *60*(2), 91–110.

Luo, B., Li, H., Song, T., & Huang, C. (2015). Object segmentation from long video sequences. In *Proceedings of the 23rd ACM international conference on multimedia* (pp. 1187–1190).

Maninis, K. K., Caelles, S., Pont-Tuset, J., & Van Gool, L. (2018). Deep extreme cut: From extreme points to object segmentation. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Mostajabi, M., Yadollahpour, P., & Shakhnarovich, G. (2015). Feedforward semantic segmentation with zoom-out features. In *The IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3376–3385).

Multiview video repository. In *Centre for vision speech and signal processing, University of Surrey, UK*. http://cvssp.org/data/cvssp3d/.

Mustafa, A., & Hilton, A. (2017). Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *CVPR*.

Mustafa, A., Kim, H., Guillemaut, J. Y., & Hilton, A. (2016). Temporally coherent 4d reconstruction of complex dynamic scenes. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Mustafa, A., Kim, H., & Hilton, A. (2016). 4d match trees for non-rigid surface alignment. In *European conference on computer vision (ECCV)*.

Mustafa, A., Kim, H., & Hilton, A. (2019). Msfd: Multi-scale segmentation-based feature detection for wide-baseline scene reconstruction. *IEEE TIP*, *28*, 1118–1132.

Mustafa, A., Volino, M., Guillemaut, J. Y., & Hilton, A. (2017). 4d temporally coherent light-field video. In *3DV*.

Prada, F., Kazhdan, M., Chuang, M., Collet, A., & Hoppe, H. (2016). Motion graphs for unstructured textured meshes. *ACM Transaction in Graphics*, *35*(4), 108:1–108:14.

Ranjan, A., Jampani, V., Kim, K., Sun, D., Wulff, J., & Black, M. J. (2018). Adversarial collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Revaud, J., Weinzaepfel, P., Harchaoui, Z., & Schmid, C. (2015). Epicflow: Edge-preserving interpolation of correspondences for optical flow. CoRR arXiv:1501.02565.

Rhodin, H., Robertini, N., Casas, D., Richardt, C., Seidel, H. P., & Theobalt, C. (2016). General automatic human shape and motion capture using volumetric contour cues. In *European conference on computer vision (ECCV)* (pp. 509–526).

Rother, C., Minka, T., Blake, A., & Kolmogorov, V. (2006). Cosegmentation of image pairs by histogram matching—Incorporating a global constraint into mrfs. In *The IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 993–1000).

Roussos, A., Russell, C., Garg, R., & Agapito, L. (2012). Dense multibody motion estimation and reconstruction from a handheld camera. In *The IEEE international symposium on mixed and augmented reality (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

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)*.

Semerjian, B. (2014). A new variational framework for multiview surface reconstruction. In *European conference on computer vision (ECCV)* (pp. 719–734).

Sevilla-Lara, L., Sun, D., Jampani, V., & Black, M. J. (2016a). Optical flow with semantic segmentation and localized layers. CoRR arXiv:1603.03911.

Sevilla-Lara, L., Sun, D., Jampani, V., & Black, M. J. (2016b). Optical flow with semantic segmentation and localized layers. In *The IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3889–3898).

Sigal, L., Balan, A., & Black, M. J. (2010). Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, *87*(1–2), 4–27.

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):345–353.

Tokmakov, P., Schmid, C., & Alahari, K. (2019). Learning to segment moving objects. *International Journal of Computer Vision (IJCV)*, *127*(3), 282–301.

Tsai, Y. H., Yang, M. H., & Black, M. J. (2016). Video segmentation via object flow. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Tsai, Y. H., Zhong, G., & Yang, M. H. (2016). Semantic co-segmentation in videos. In *European conference on computer vision (ECCV)* (pp. 760–775).

Tulsiani, S., Efros, A. A., & Malik, J. (2018). Multi-view consistency as supervisory signal for learning shape and pose prediction. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Vineet, V., Miksik, O., Lidegaard, M., Nießner, M., Golodetz, S., Prisacariu, V. A., et al. (2015). Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In *IEEE international conference on robotics and automation (ICRA)*.

Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., & Cremers, D. (2011). Stereoscopic scene flow computation for 3d motion understanding. *International Journal of Computer Vision (IJCV)*, *95*(1), 29–51.

Weinzaepfel, P., Revaud, J., Harchaoui, Z., & Schmid, C. (2013). Deepflow: Large displacement optical flow with deep matching. In *The IEEE international conference on computer vision (ICCV)* (pp. 1385–1392).

Xie, J., Kiefel, M., Sun, M. T., & Geiger, A. (2016). Semantic instance annotation of street scenes by 3d to 2d label transfer. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Yang, G., Zhao, H., Shi, J., Deng, Z., & Jia, J. (2018). Segstereo: Exploiting semantic information for disparity estimation. In *The European conference on computer vision (ECCV)*.

Yin, Z., & Shi, J. (2018). Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*.

Zanfir, A., Marinoiu, E., & Sminchisescu, C. (2018). Monocular 3d pose and shape estimation of multiple people in natural scenes—The importance of multiple scene constraints. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Zanfir, A., & Sminchisescu, C. (2015). Large displacement 3d scene flow with occlusion reasoning. In *The IEEE international conference on computer vision (ICCV)*.

Zhang, Z., Zhang, X., Peng, C., Xue, X., & Sun, J. (2018). Exfuse: Enhancing feature fusion for semantic segmentation. In *The European conference on computer vision (ECCV)*.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., et al. (2015). Conditional random fields as recurrent neural networks. In *The IEEE international conference on computer vision (ICCV)*.

Zhu, X., Xiong, Y., Dai, J., Yuan, L., & Wei, Y. (2017). Deep feature flow for video recognition. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4141–4150).

Zitnick, C. L., Kang, S. B., Uyttendaele, M., Winder, S., & Szeliski, R. (2004). High-quality video view interpolation using a layered representation. *ACM Transaction on Graphics*, *23*(3), 600–608.

Zou, Y., Luo, Z., & Huang, J. B. (2018). Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *European conference on computer vision*.