IEEE International Conference in Computer Vision & Pattern Recognition CVPR, 2017

Semantically Coherent Co-segmentation and Reconstruction of Dynamic Scenes

Armin Mustafa Adrian Hilton
CVSSP, University of Surrey, United Kingdom
a.mustafa@surrey.ac.uk

Abstract

In this paper we propose a framework for spatially and temporally coherent semantic co-segmentation and reconstruction of complex dynamic scenes from multiple static or moving cameras. Semantic co-segmentation exploits 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. We demonstrate that semantic coherence results in improved segmentation and reconstruction for complex scenes. A joint formulation is proposed for semantically coherent object-based co-segmentation and reconstruction of scenes by enforcing consistent semantic labelling between views and over time. Semantic tracklets are introduced to enforce temporal coherence in semantic labelling and reconstruction between widely spaced instances of dynamic objects. Tracklets of dynamic objects enable unsupervised learning of appearance and shape priors that are exploited in joint segmentation and reconstruction. Evaluation on challenging indoor and outdoor sequences with hand-held moving cameras shows improved accuracy in segmentation, temporally coherent semantic labelling and 3D reconstruction of dynamic scenes.

1. Introduction

Advances in visual scene understanding using deep learning, with convolutional neural network architectures and large annotated image collections [56, 10, 40], have achieved excellent performance in per-pixel labelling of semantic categories in complex real-world scenes from images. Due to the inherent ambiguity in visual segmentation and classification from a single camera view the output may include errors in pixel labelling and object boundary segmentation resulting in a lack of temporal coherence in semantic labelling. Likewise independent classification for different views of the same scene may result in inconsistent per-pixel semantic labelling for the same object.

This paper introduces a framework for semantically coherent per-pixel segmentation and reconstruction of dynamic scenes. The approach enforces semantic coherence both spatially across different views of the scene and temporally across different observations of the same object. Semantic tracklets are introduced to associate semantic labels between different observations of a dynamic object with similar shape and appearance over time. This enables improved temporal coherence in semantic labelling and co-segmentation for monocular video. Joint semantic co-segmentation and reconstruction across multiple views of dynamic objects enforces spatial coherence in semantic labelling resulting in improved performance over previous approaches which did not exploit semantic information.

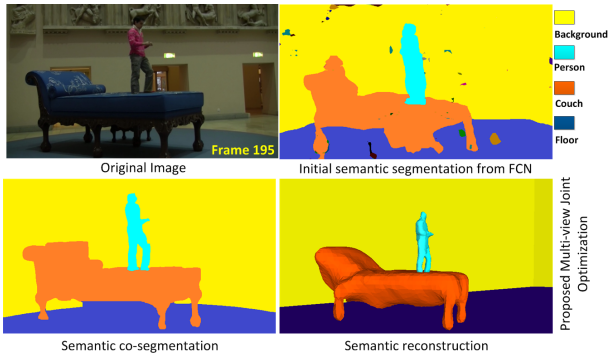

Figure 1. Example of input image from Magician dataset [3] and standard image classification from fully convolution network (FCN) [10] on the top. Bottom: Proposed framework resulting in an accurately labeled segmentation and 3D reconstruction.

Previous research has demonstrated the advantages of joint segmentation and reconstruction across multiple views [21, 24, 23, 34, 14, 32], co-segmentation of multiple view images [11, 31, 13, 12] and temporal coherence in reconstruction [20, 18, 36, 42]. Our contribution is the introduction of a framework for joint semantic co-segmentation and reconstruction of complex dynamic scenes to obtain semantically coherent per-view 2D object segmentation and 3D scene reconstruction from wide-baseline camera views. Semantic coherence refers to spatial and temporal coherence of semantic labels across the sequence. To the best of our knowledge, this is the first method addressing the problem of temporally coherent semantic co-segmentation and reconstruction for dynamic scenes.

Figure 1 shows an example of semantically coherent co-segmentation and reconstruction for the publicly available Magician dataset [3] captured with 5 hand-held unsynchronised moving cameras. An initial semantic class labelling is obtained independently for each view using fully convolutional networks (FCN) at each frame [10]. Joint semantic co-segmentation and reconstruction (bottom-row) results in significant improvement in both 2D segmentation and reconstruction. Contributions include:

- Joint semantic co-segmentation and reconstruction of dynamic objects in complex scenes
- Semantic tracklets for temporally coherent semantic labelling of video across wide-timeframes
- Improved 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 [41, 22]. 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 [17, 25]. The per-pixel prediction leads to segmentations with fuzzy boundaries and spatially disjoint regions. Another group of methods pioneered by [38] predict segmentations from the raw pixels. Methods were introduced to improve the spatial coherence of the semantic segmentation using conditional random fields (CRF) [33, 57, 9].

Co-segmentation: Co-segmentation was first introduced by [49] for simultaneous binary segmentation of object parts in an image pair. This was extended to co-segmentation of multiple images [5]. Multi-view co-segmentation in space and time was introduced in [13]. 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 [28, 40], segments [31], motion [49] and foreground propagation [20]. Recently, co-segmentation methods were introduced to segment common objects in a collection of videos for a single object [19] or multiple objects [11, 54].

2.2. Joint segmentation and reconstruction

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 [21, 42] by including the multiple view photo-consistency. This concept was extended to semantic segmentation and reconstruction to obtain additional information from the scene [24, 56]. Methods were introduced to utilize appearance-based pixel categories and stereo cues in a joint framework for street scenes from a monocular camera [34, 55, 18]. These methods used CRF to perform simultaneous dense reconstruction and segmentation of street scenes captured from a moving camera. A method to estimate the pose and 3D shape of rigid objects on street scenes was proposed [14]. Compact shape manifolds within an object class were used for joint object segmentation, pose and shape estimation. However these methods cannot be directly applied to multi-view wide-baseline scenes. A method for joint estimation of 3D scene geometry and semantic segmentation using multiple images was proposed for static scenes [23]. Dense semantic reconstruction of rigid objects was proposed by [4]. However, these methods are limited to static scenes and rigid objects.

This paper introduces joint semantic 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 and appearance information in space and time in a single optimization. Evaluation demonstrates improved accuracy and completeness of both segmentation and reconstruction for complex dynamic scenes.

3. Semantic Segmentation & Reconstruction

The proposed framework for semantic coherence, illustrated in Figure 2, comprises the following stages:

Initial Semantic Segmentation: Initial semantic labels are estimated for each pixel in the image per-view using fully convolutional networks (FCNs) [10].

Initial Semantic Reconstruction: 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: To enforce 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, and shape. Semantic trackets provide a prior for the joint space-time semantic co-segmentation and reconstruction to enforce temporal coherence.

Semantic Co-segmentation and Reconstruction: The initial semantic segmentation and reconstruction is refined per-view for each dynamic object through joint optimisation of 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 [29].

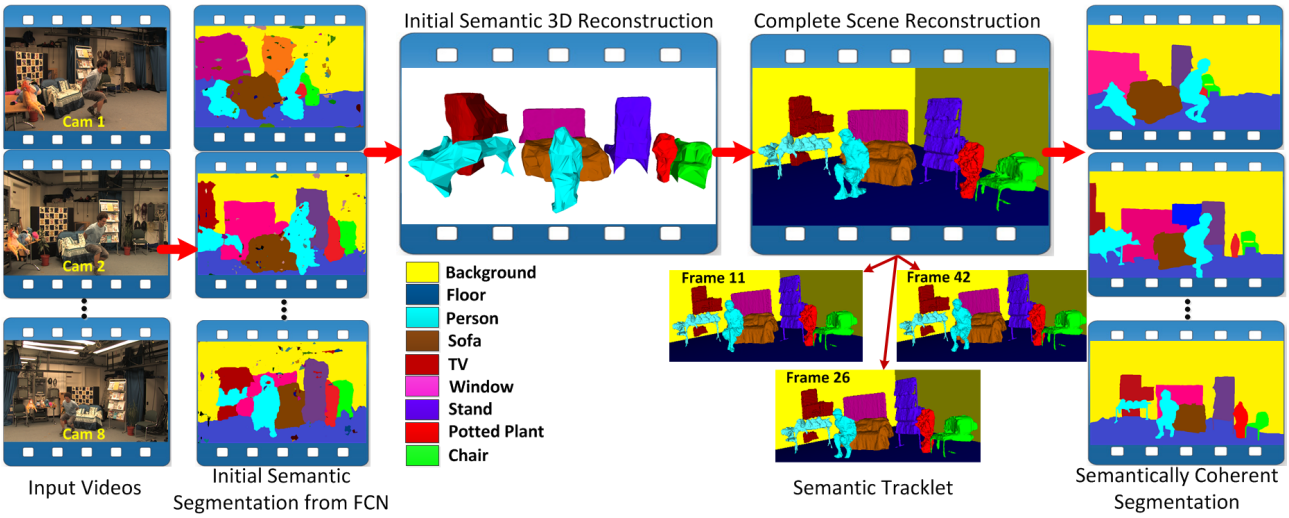

Figure 2. Semantically coherent co-segmentation and reconstruction framework.

The process is repeated for the entire sequence to obtain semantically coherent dense co-segmentation and reconstruction for the complete scene. The following sections include a detailed explanation of the proposed approach and highlight the novel contributions of this work.

3.1. Initial Segmentation & Reconstruction

Initial Semantic Segmentation: The state-of-the-art in semantic segmentation is currently represented by fully convolutional networks (FCNs). To predict semantic unary potentials we employ the DeepLab model, which is a fully convolutional adaptation of the VGG network [10]. 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[37] dataset with 81 classes and is refined on PASCAL VOC12 [16] dataset. FCNs use large receptive fields and many pooling layers, both of which cause blurring and low spatial resolution in the deep layers. As a result FCNs produce segmentations with poorly localized object boundaries as illustrated in Figure 3(b).

Initial Semantic Reconstruction: Sparse feature-based reconstruction of the scene is performed using SFD features [44] and SIFT descriptor[39] with the constraint that each 3D feature should be visible in 3 or more camera views for robustness [26]. The resulting point-cloud is clustered in 3D [50]. 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 p to enclose the object: $p = \frac{1}{N_h} * \sum_{c=1}^{N_h} \frac{B_s^c - B_r^c}{B_s^i}$, 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 priors.

3.2. Semantic Tracklets

In the case of general dynamic scenes with non-rigid objects, independent per-frame segmentation and reconstruction leads to incoherent results, for example failure to 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 [6, 46]. Previous work [54] has shown that semantic tracklets improve segmentation for single view video. To achieve robust temporally coherent reconstruction *semantic tracklets* are introduced linking instances of dynamic objects across wide-timeframes. This provides a prior to constrain co-segmentation and reconstruction. Semantic tracklets for a dynamic object are defined as a set of frames which have similar semantic labels, appearance and 2D shape as illustrated in Figure 4. Tracklets are used for long-term learning of semantic labels, appearance and shape information for per-view joint semantic co-segmentation and reconstruction of each object. This improves the semantic coherence in reconstruction and segmentation results as shown in Figure 5 and 12. Dynamic objects are identified in the scene using motion information from sparse temporal SIFT feature correspondences. The semantic, 2D shape and

Figure 3. The improvement of semantic segmentation using the proposed framework for Odzemok dataset.

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 is evaluated as follows:

Semantic Similarity: The semantic region associated with the object at each frame is identified using sparse wide-timeframe SIFT feature matches. An affine warp [15] 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: $L^c_{i,j} = \frac{z^c_{i,j}}{y^c_{i,j}}$

Appearance Similarity: The appearance metric $M^c_{i,j}$ between frame i and j for the semantic region segmentation in view c 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 $R^c_{i,j}$ [43]: $M^c_{i,j} = \frac{Q^c_{i,j}}{R^c_{i,j}}$

Shape Similarity: The shape 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 [15]. The 2D shape similarity metric $I^c_{i,j}$ is defined as the ratio of the intersection of the aligned segmentation $h^c_{i,j}$ to the union of the area $A^c_{i,j}$: $I^c_{i,j} = \frac{h^c_{i,j}}{A^c_{i,j}}$

Similarity metric: The metrics defined above are used to calculate the similarity between frames as follows:

$$S_{i,j} = \frac{1}{3N_S} \sum_{c=1}^{N_T} (M^c_{i,j} + I^c_{i,j} + L^c_{i,j}) \qquad (1)$$

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 $t_r \in [0, i-1]$.

3.3. Single-view Semantic Segmentation

Temporally coherent semantic segmentation can be optimised independently for a single-view video using the semantic tracklets without a requirement for multiple views. This is extended to spatially and temporally coherent joint co-segmentation and reconstruction from multiple view video in section 3.4. The goal of single-view semantic segmentation is to assign a semantic label from a set of semantic classes obtained as an initialization from FCN (section 3.1), $\mathscr{L} = \{l_1, ..., l_{|\mathscr{L}|}\}$, to each pixel p for the initial semantic segmentation region $\mathscr{S}$ of each object (Section 3.1), where $|\mathscr{L}|$ is the total number of classes in the network. This is achieved by optimization of a cost function:

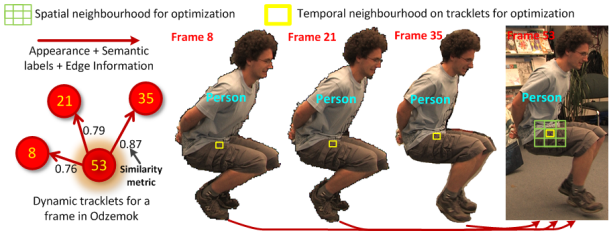

Figure 4. Example of dynamic tracklet generation (similar frames) for a dynamic object at current frame 53 based on appearance, shape and semantic information. The spatial and temporal neighbourhood are shown at the top in green and yellow respectively for the optimization.

$$E_{single}(l) = \lambda_{sem} E_{sem}(l) + \lambda_a E_a(l) + \lambda_c E_c(l) \qquad (2)$$

where individual cost terms enforce spatial and temporal coherence for dynamic objects in semantic labels E_{sem}, appearance E_a, and region boundary contrast E_c. Optimization is performed using α-expansion across spatial and temporal neighbourhoods as shown in Figure 4 by iterating through the set of labels in $\mathscr{L}$ [8].

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. For single view optimization the tracklets are estimated using the metric: $s^c_{i,j} = \frac{1}{3}(M^c_{i,j} + I^c_{i,j} + L^c_{i,j})$ derived from Eq. 1. In the color similarity metric $M^c_{i,j}$, $Q^c_{i,j}$ is replaced with correspondences obtained using the single view wide-timeframe matching approach by [45]. Optical flow is used to compute a dense flow field on the tracklets, initialized from the sparse temporal SIFT feature correspondences. EpicFlow [47] 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 $\psi_T = \{(p, q) \mid q = p + d_{i,j}\}$; where j 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: This cost is computed based on the prob-

ability of the class labels at each pixel for the initial FCN semantic segmentation [10]. Unlike previous approaches to achieve semantic coherence we enforce spatial and temporal consistency using tracklets across the neighbourhoods:

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

where $P_{sem}(I_p|l_p = l_i)$ denotes the probability of the layer l_i at pixel p in the classification image obtained from FCN.

Contrast cost: The contrast cost [10] 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:

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

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

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: This cost is computed using the negative log likelihood [7] 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:

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

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

An example of single-view semantic segmentation is shown in Figure 3(c). Enforcing temporal coherence with semantic tracklets for a single monocular video reduces noise in per-pixel labels. Errors in object segmentation remain due to the low spatial resolution of the FCN semantic boundaries and visual ambiguity in single view segmentation. In the following section we introduce multi-view joint semantic co-segmentation and reconstruction which combines information across multiple views to refine the segmentation as illustrated in Figure 3(d).

3.4. Multi-view Joint Semantic Co-segmentation and Reconstruction

Single view semantic segmentation is extended to multiple views to obtain semantically coherent co-segmentation and reconstruction. Co-segmentation is achieved by propagating the semantic labels across views and over time using tracklets in the framework. The initial semantic reconstruction obtained in Section 3.1 is refined for each dynamic object per-view. An accurate depth value is jointly assigned for each pixel p from a set of depth values $\mathscr{D} = \left\{d_1,...,d_{|\mathscr{D}|-1},\mathscr{U}\right\}$ along with a semantic label from the set $\mathscr{L}$ for the region $\mathscr{R}$ for each object, where d_i is obtained by sampling the optical ray from the camera and $\mathscr{U}$ is an unknown depth value to handle occlusions. Formulation of a cost function for semantically coherent depth 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).

The cost function enforces spatial and temporal constraints on the semantic, appearance and shape. Temporal semantic coherence is enforced using tracklets based on dynamic object similarity $S_{i,j}$ Eq.1. Joint optimisation of multiple view co-segmentation and reconstruction minimises:

$$\begin{aligned} E(l,d) &= E_{single}(l) + E_{multi}(l,d) \quad (3)\\ E_{multi}(l,d) &= \lambda_d E_d(d) + \lambda_{sm}E_{sm}(l,d) + \lambda_s E_s(l,d)\end{aligned}$$

where, d is the depth at each pixel and l is the semantic label. This is solved subject to a geodesic star-convexity constraint on the semantic labels l [42]:

$$\min_{\substack{(l,d)\\ s.t.\\ l\epsilon S^\star(\mathscr{C})}} E(l,d) \Leftrightarrow \min_{(l,d)} E(l,d) + E^\star(l|x,\mathscr{C}) \quad (4)$$

where $S^\star(\mathscr{C})$ is the set of all shapes which are geodesic star-convex wrt the features in $\mathscr{C} = \{c_1,...,c_n\}$ 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 $\mathscr{L}\times\mathscr{D}$ [8] and a solution is obtained using graph-cuts [7].

Semantic Cost: This term enforces multi-view consistency on the semantic labels of each pixel p. Inconsistent labels across views are penalised to ensure semantic coherence.

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

$e_{sm}(p,d_p,l_p) = \sum_{c=1}^{N_K} z(p,r,l_p)$, 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.

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

where l_r is the semantic label at pixel r in view c.

Matching cost: The photo-consistency matching cost across views is defined as:

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

where $e_d(p,d_p) = \sum_{i\in\mathscr{O}_k} m(p,r)$, if $d_p \neq \mathscr{U}$ else $M_{\mathscr{U}}$. $m(p,r)$ is inspired from [27]. $M_{\mathscr{U}}$ is the fixed cost of labelling a pixel unknown and r is as defined above. $\mathscr{O}_k$ is the set of k most photo-consistent pairs with reference camera.

Smoothness cost: The surface smoothness cost introduced in [42] is extended to spatial and temporal neighbourhoods:

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

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

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 of d_{max}^s to allow large movement in the object between tracklet frames.

The importance of the proposed semantically coherent optimization exploiting the information from semantic labels and tracklets for single and multiple views is shown in the Figure 5. Comparison is presented against optimization with/without semantic label and temporal tracklet information for single and multiple views. The proposed approach consistently performs better giving a more accurate segmentation. The final proposed multiple view co-segmentation and reconstruction using both semantic labels and tracklets gives a significantly improved segmentation.

4. Results and Evaluation

The proposed single-view approach (section 3.3) is evaluated on datasets previously used for 2D video co-segmentation (MOVICS [11] and ObMiC [19]) for comparison with state-of-the-art methods. Joint semantic co-segmentation and reconstruction (section 3.4) is evaluated on a variety of publically available multi-view indoor and outdoor dynamic scene datasets: DogJump[1], HumanEva[53], Odzemok [2], Handshake[30], Breakdance [58], Magician and Juggler [3].

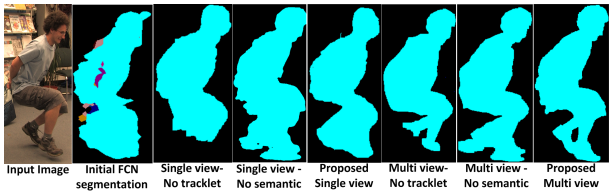

Figure 5. Comparison of segmentation of the proposed single and multi view optimization against optimization with no semantic and no tracklet information respectively for Odzemok dataset.

Figure 6. Comparison of semantic segmentation for 2D video segmentation datasets against MVC [11] and ObMiC [19].

Figure 7. Comparison of segmentation on dynamic datasets from [30] and [13] against MVVS [13].

4.1. Single-view segmentation evaluation

Single-view segmentation is evaluated against state-of-the-art semantic (MVC) [11] and non-semantic (ObMiC) [19] video co-segmentation methods. Qualitative comparison against ObMiC [19] and MVC [11] on four single view video co-segmentation datasets (Giraffe,Tiger,Person,Dog) are shown in Figure 6 and quantitative evaluation against ground-truth, is shown in the Table 1. Results indicate that the proposed approach achieves state-of-the-art performance for single view segmentation due to the introduction of semantic tracklets to enforce temporal coherence.

4.2. Multi-view evaluation

Segmentation Evaluation: Mutli-view co-segmentation is evaluated against a variety of state-of-the-art methods: ***(a) Non-Semantic methods:*** Multi-view segmenta-

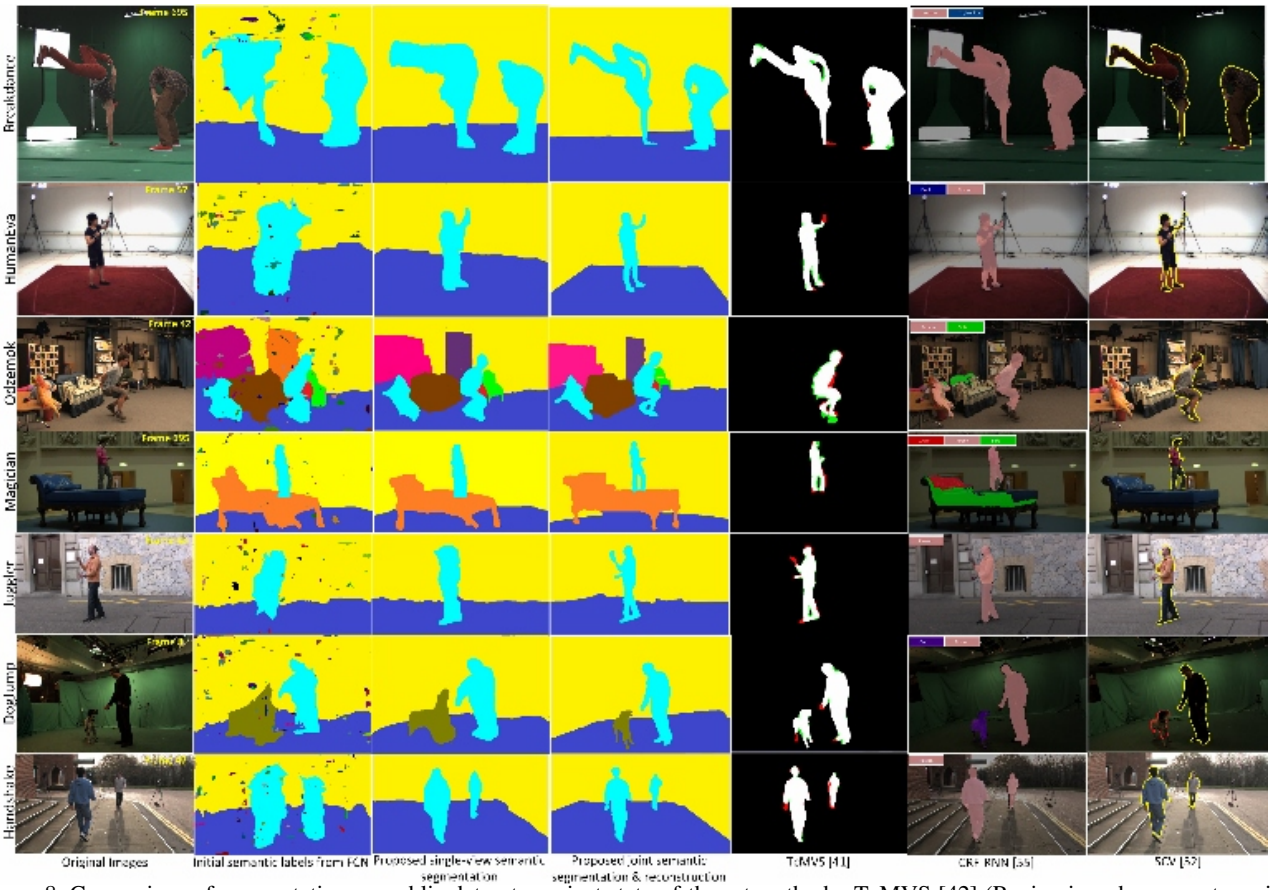

Figure 8. Comparison of segmentation on public datasets against state-of-the-art methods: TcMVS [42] (Region in red represents region missing from ground-truth and green represents region not present in ground-truth), CRF-RNN [57] and SCV [54].

Datasets	Multi-view segmentation							2D video segmentation			
Methods	Bdance	HEva	Oz	Mag	Juggler	Jump	HShake	Giraffe	Tiger	Person	Dog
MVC [11]	36.5	42.1	38.2	34.8	39.7	41.6	44.8	59.6	47.0	59.8	48.7
ObMiC [19]	39.4	49.6	45.5	41.4	44.0	45.9	48.1	66.2	71.0	54.3	74.0
CRF-RNN [57]	61.0	71.4	41.0	53.3	70.8	52.3	64.6	69.7	68.1	63.0	**77.1**
SCV [54]	48.9	51.0	53.3	61.0	56.6	60.2	49.5	59.0	**70.9**	61.2	76.6
TcMVS [42]	89.1	94.0	91.8	91.2	93.3	89.4	86.5	65.2	64.5	59.7	73.2
	Multi-view joint co-segmentation & reconstruction							Single-view segmentation			
Proposed	**93.2**	**95.6**	**94.5**	**93.0**	**94.7**	**92.6**	**91.5**	**72.5**	68.9	**66.4**	75.8

Table 1. Segmentation result comparisons for all datasets against state-of-the-art methods using the ***Intersection-over-Union*** metric. Representation of datasets: Bdance(Breakdance), HEva(HumanEva), Oz(Odzemok), Mag(Magician), HShake(Handshake) and Jump(Dogjump).

tion (MVVS) [13], Joint segmentation and reconstruction (TcMVS) [42], and ***(b) Semantic methods:*** Semantic co-segmentation in videos (SCV) [54] and Conditional random field as recurrent neural networks (CRF-RNN) [57]. Single view methods MVC[11] and ObMiC[19] are also applied independently on each view for comparison. Comparison against MVVS [13] is shown in Figure 7 and evaluation against TcMVS [42], SCV [54] and CRF-RNN [57] are shown in Figure 8 for dynamic datasets. Quantitative evaluation against state-of-the-art methods is measured by Intersection-over-Union with ground-truth, shown in the Table 1. Ground-truth is available online for most of the datasets and obtained by manual labelling for other datasets. The proposed semantically coherent joint multi-view co-segmentation and reconstruction achieves the best segmentation performance against ground-truth for all datasets tested. Results presented in Figure 8 indicate that the proposed approach accurately segments fine detail such as hands and feet where other approaches are unreliable.

Reconstruction Evaluation: The reconstruction results obtained from the proposed approach are compared against state-of-the-art approaches in joint segmentation and reconstruction (TcMVS [42]) and multi-view stereo (Colmap [51], MVE [52], SMVS [35]). MVE, SMVS and Colmap

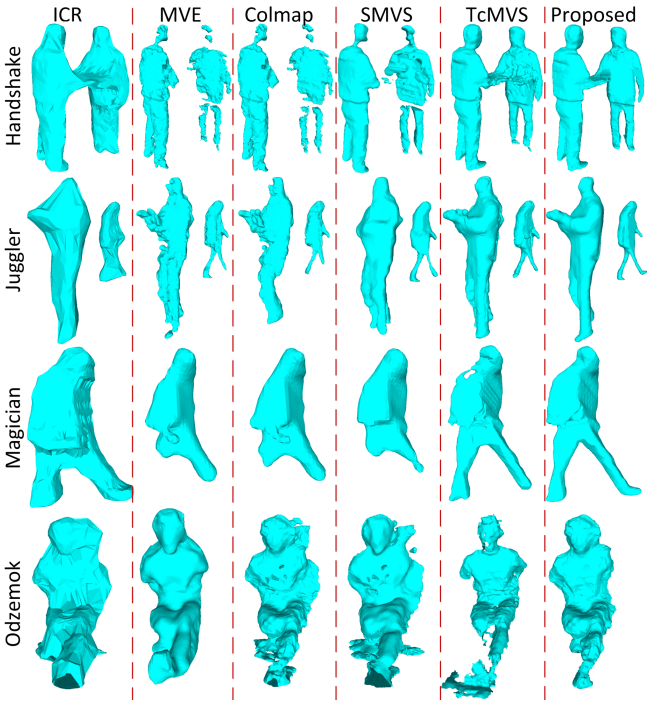

Figure 9. Comparison of reconstruction of dynamic objects against Colmap [51], MVE [52], SMVS [35] and TcMVS [42]) (Same semantic labels are assigned to all methods for fair comparison).

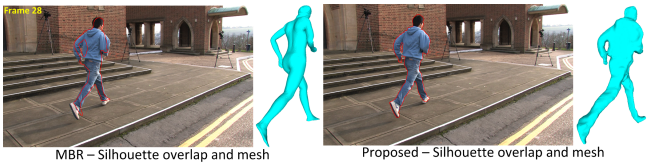

Figure 10. Comparison of reconstruction against MBR [48] from 4 views of Falling down [30] dataset.

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 (section 3.1) for fair comparison. Comparison of reconstructions Figure 9 demonstrates that the proposed method gives consistently more complete and accurate models. Figure 10 presents a comparison to a statistical model-based approach MBR [48] 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 clothing resulting in inaccurate silhouette overlap. Comparison of full scene reconstruction against MVE and SMVS is shown in Figure 11 showing improved completeness and accuracy. 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 Figure 12.

Limitations: The proposed approach is dependent on an initial semantic labelling of the scene for each view obtained using FCN. 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 Figure 2. Whilst enforcing semantic coherence is demonstrated to improve both segmentation and reconstruction for a wide-variety of scenes visual ambiguity in appearance and occlusion may degrade performance.

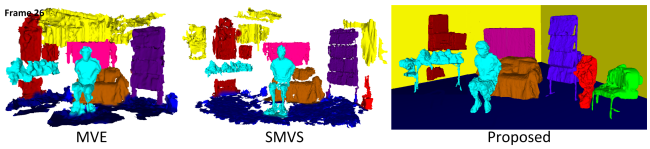

Figure 11. Comparison of full scene reconstruction against SMVS [51] and MVE [52] (Same semantic labels are assigned to all the approaches for fair comparison).

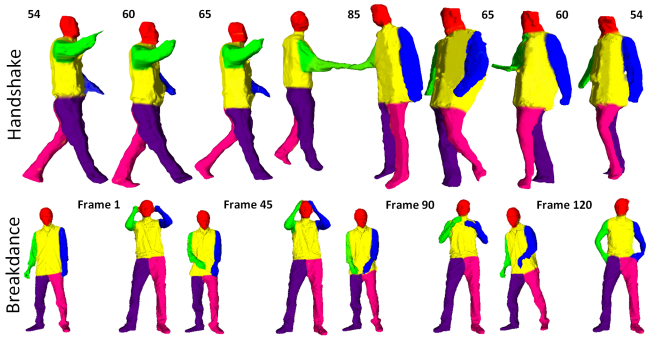

Figure 12. Semantic coherence results using proposed approach on two datasets. Color-coding: head is red, left-arm is blue, right-arm is green, left-leg is pink and right-leg is violet

5. Conclusion

This paper proposes a novel approach to joint semantically coherent multi-view co-segmentation and reconstruction of complex dynamic scenes. Temporal semantic coherence is enforced by semantic tracklets identifying similar frames using the semantic label, appearance and shape. Tracklets are used for long-term learning to constrain co-segmentation optimization on complex dynamic scenes. Joint optimization simultaneously improves the 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 both segmentation and reconstruction of general dynamic indoor and outdoor scenes captured with multiple hand-held cameras.

Acknowledgments: This research was supported by the InnovateUK grant for Live Action Lightfields for Immersive Virtual Reality Experiences (ALIVE) project (grant 102686). We would like to thank Helge Rhodin and Abdelaziz Djelouah for providing their data.

References

[1] 4d repository, http://4drepository.inrialpes.fr/. In *Institut national de recherche en informatique et en automatique (INRIA) Rhone Alpes*.

[2] Multiview video repository, http://cvssp.org/data/cvssp3d/. In *Centre for Vision Speech and Signal Processing, University of Surrey, UK*.

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

[4] Y. Bao, M. chandraker, Y. Lin, and S. Savarese. Dense object reconstruction using semantic priors. In *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[5] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[6] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner, and M. Gross. High-quality passive facial performance capture using anchor frames. *ACM Transaction in Graphics*, 30(4):75:1–75:10, 2011.

[7] Y. Boykov and V. Kolmogorov. 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, 2004.

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

[9] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014.

[10] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[11] W.-C. Chiu and M. Fritz. Multi-class video co-segmentation with a generative multi-video model. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

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

[13] A. Djelouah, J.-S. Franco, E. Boyer, P. Pérez, and G. Drettakis. Cotemporal Multi-View Video Segmentation. In *International Conference on 3D Vision (3DV)*, 2016.

[14] F. Engelmann, J. Stückler, and B. Leibe. 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)*, 2016.

[15] G. D. Evangelidis and E. Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(10):1858–1865, 2008.

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[17] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(8):1915–1929, 2013.

[18] G. Floros and B. Leibe. Joint 2d-3d temporally consistent semantic segmentation of street scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2823–2830, 2012.

[19] H. Fu, D. Xu, B. Zhang, and S. Lin. Object-based multiple foreground video co-segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[20] B. Goldluecke and M. Magnor. Space-time isosurface evolution for temporally coherent 3d reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 350–355, 2004.

[21] J. Y. Guillemaut and A. Hilton. Joint Multi-Layer Segmentation and Reconstruction for Free-Viewpoint Video Applications. *International Journal of Computer Vision (IJCV)*, 93(1):73–100, 2010.

[22] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. *Learning Rich Features from RGB-D Images for Object Detection and Segmentation*, pages 345–360. 2014.

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

[24] C. Hane, C. Zach, A. Cohen, and M. Pollefeys. Dense semantic 3d reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, page 1, 2016.

[25] B. Hariharan, P. A. Arbelez, R. B. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 447–456, 2015.

[26] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2003.

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

[28] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[29] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Eurographics Symposium on Geometry Processing*, pages 61–70, 2006.

[30] H. Kim, J. Guillemaut, T. Takai, M. Sarim, and A. Hilton. Outdoor Dynamic 3-D Scene Reconstruction. *IEEE transactions on Circuits and Systems for Video Technology (TCSVT)*, 22(11):1611–1622, 2012.

[31] K. Kolev, T. Brox, and D. Cremers. 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, 2012.

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

[33] A. Kundu, V. Vineet, and V. Koltun. Feature space optimization for semantic video segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3168–3175, 2016.

[34] L. Ladický, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H. S. Torr. Joint optimization for object class segmentation and dense stereo reconstruction. *International Journal of Computer Vision (IJCV)*, 100(2):122–133, 2012.

[35] F. Langguth, K. Sunkavalli, S. Hadap, and M. Goesele. Shading-aware multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[36] E. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[37] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[38] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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

[40] B. Luo, H. Li, T. Song, and C. Huang. Object segmentation from long video sequences. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1187–1190, 2015.

[41] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3376–3385, 2015.

[42] A. Mustafa, H. Kim, J.-Y. Guillemaut, and A. Hilton. Temporally coherent 4d reconstruction of complex dynamic scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[43] A. Mustafa, H. Kim, and A. Hilton. 4d match trees for non-rigid surface alignment. In *European Conference on Computer Vision (ECCV)*, 2016.

[44] A. Mustafa, H. Kim, E. Imre, and A. Hilton. Segmentation based features for wide-baseline multi-view reconstruction. In *International Conference on 3D Vision (3DV)*, 2015.

[45] G. Nebehay and R. Pflugfelder. Clustering of Static-Adaptive correspondences for deformable object tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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

[47] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. *CoRR*, abs/1501.02565, 2015.

[48] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *European Conference on Computer Vision (ECCV)*, pages 509–526, 2016.

[49] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 993–1000, 2006.

[50] R. B. Rusu. *Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments*. PhD thesis, Computer Science department, Technische Universitaet Muenchen, Germany, 2009.

[51] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[52] B. Semerjian. A new variational framework for multiview surface reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 719–734, 2014.

[53] L. Sigal, A. Balan, and M. J. Black. 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, 2010.

[54] Y.-H. Tsai, G.Zhong, and M.-H. Yang. Semantic co-segmentation in videos. In *European Conference on Computer Vision (ECCV)*, pages 760–775, 2016.

[55] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Perez, and P. H. S. Torr. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[56] J. Xie, M. Kiefel, M.-T. Sun, and A. Geiger. Semantic instance annotation of street scenes by 3d to 2d label transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[57] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

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