U4D: Unsupervised 4D Dynamic Scene Understanding - Supplementary material

1. Appendix A: Pairwise terms - $E_{pair}(l, d, m)$

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

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

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

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

Contrast term: This term is defined as:

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

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

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

2. Appendix B: Key-frame detection

As explained in the paper, key-frame detection is used to improve the long term temporal coherence in the proposed joint semantic instance segmentation and 4D reconstruction. The 3D meshes are aligned for frames in between two key-frames K_i and K_{i+1} and between key-frames N_K to obtain full 4D scene reconstruction for the sequence.

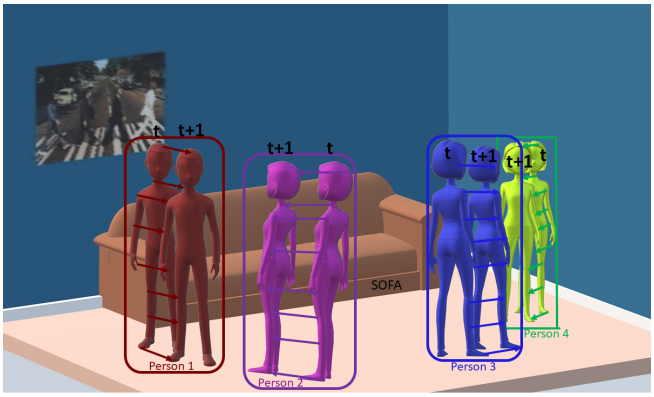

Figure 1. Illustration of 4D scene understanding which includes reconstruction, semantic instance segmentation and tracking in time.

	Semantic	Segment	Instance	3D	Motion
SCV [11]	✓	✓	×	×	✓
SCSR [7]	✓	✓	×	✓	×
JSR [4]	×	✓	×	✓	×
Dv3+ [2]	✓	✓	×	×	×
MRCNN [5]	✓	✓	✓	×	×
PSP [15]	✓	✓	×	×	×
CRF RNN [16]	✓	✓	×	×	×
Segnet [1]	✓	✓	×	×	×
RTSeg [9]	✓	✓	×	×	×
PRSM [12]	×	×	×	✓	✓
LocalStereo [10]	×	×	×	✓	×
SMVS [6]	×	×	×	✓	×
DCflow [14]	×	×	×	×	✓
Deepflow [13]	×	×	×	×	✓
4DMatch [8]	×	×	×	×	✓
Proposed	✓	✓	✓	✓	✓

Table 1. Illustration of tasks performed by state-of-the-art methods compared to the proposed method.

2.1. Key-frame similarity metric

Key-frame detection exploits sparse correspondence ($M^c_{i,j}$), pose ($P^c_{i,j}$), shape ($I^c_{i,j}$), semantic ($I^c_{i,j}$) and distance ($D^c_{i,j}$) information across views N_v between frame i and j for each object in view c. All the metrics are defined in detail below for the key-frame metric:

$$KS_{i,j} = 1 - \frac{1}{5N_v}\sum_{c=1}^{N_v}(M^c_{i,j} + L^c_{i,j} + D^c_{i,j} + P^c_{i,j} + I^c_{i,j})$$

2.1.1 Sparse correspondence Metric ($M^c_{i,j}$)

This measures appearance similarity between frames for each object, defined as the ratio of the number of sparse temporal correspondences Q to the total number of features R:

$$M^c_{i,j} = \frac{2Q^c_{i,j}}{R^c_i + R^c_j}$$

2.1.2 3D Pose Metric ($P^c_{i,j}$)

This metric measures the distance between the regularised pose:

$$P^c_{i,j} = \frac{\|P_i - P_j\|_F}{P^c_{max}}$$

where $j > i$ and P^c_{max} is the maximum change of pose between frames for view c. This term ensures that the distance between poses between key-frames is limited.

2.1.3 Semantic Metric ($L^c_{i,j}$)

An affine warp [3] is used to align semantic regions to measure semantic similarity between two frames. The metric is defined as the ratio of the number of pixels with similar class label $z^c_{i,j}$ to the pixels in the segmented region $y^c_{i,j}$:

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

2.1.4 Distance Metric ($D^c_{i,j}$)

This metric measures the distance between frames:

$$D^c_{i,j} = \frac{j - i}{D^c_{max}}$$

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

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

Table 2. Silhouette overlap error for multi-view datasets for evaluation of long-term temporal coherence, where P_K = Proposed method without key-frame detection.

2.1.5 Shape Metric ($I^c_{i,j}$)

It is defined as the ratio of the intersection of the aligned segmentation [3] (h) to the union of the area (a):

$$I^c_{i,j} = \frac{h^c_{i,j}}{a^c_{i,j}}$$

It gives shape overlap between frames.

2.2. Ablation study without key-frame detection

The higher the number of key-frames the better the quality of alignment. However if no key-frames are detected for a sequence, it will degrade the performance of 4D long-term scene flow. To evaluate the effect of key-frame detection we evaluate the performance of 4D scene flow for proposed joint optimization with and without key-frames in Table 2. The results show an $\approx 15\%$ improvement in scene flow with key-frame detection.

3. Appendix C: Results and Evaluations

This section contains additional results along with the original paper submission on 4D dynamic scene understanding (illustrated in Figure 1). We have added more qualitative and quantitative results on the datasets and state-of-the-art methods listed in the original manuscript. The results of the proposed method are compared with 15 methods listed in Table 1. None of these state-of-the-art methods exploit human pose information to refine the results.

3.1. Segmentation Comparison

Semantic segmentation comparison results against CRF RNN [16], Segnet [1], PSP [15] are shown in Figure 2 on four datasets. Ground-truth segmentation comparison is shown in Figure 3 against JSR [4] and SCSR [7]. The red and green regions highlight the error, green regions are present in segmentation but not ground-truth and red regions are present in ground-truth but not the segmentation.

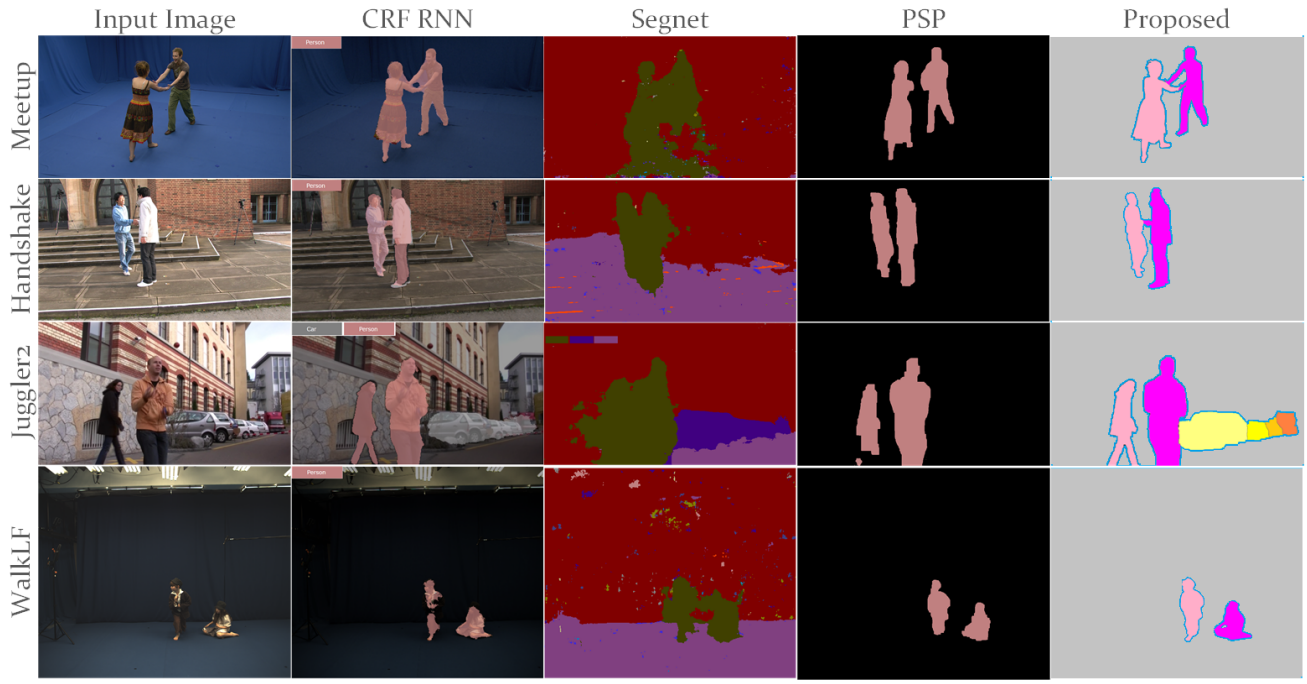

Figure 2. Semantic segmentation comparison against state-of-the-art methods. In the proposed method shades of pink depicts instances of humans and shades of yellow depict instances of cars.

Figure 3. Ground-truth semantic segmentation comparison against state-of-the-art methods JSR and SCSR.

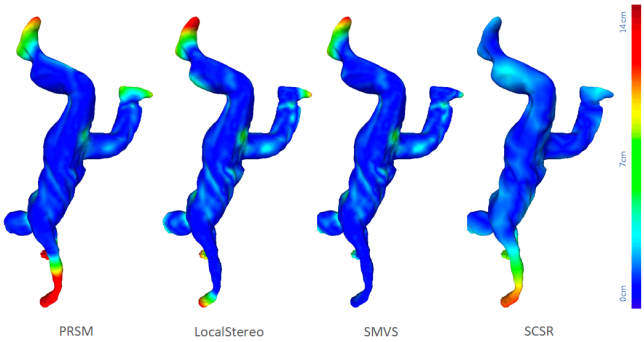

Figure 4. Comparison of reconstruction obtained using state-of-the-art methods against proposed method.

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

Table 3. Temporal coherence evaluation for Meetup dataset against existing methods: S.D. is the standard deviation

3.2. Reconstruction evaluation

In addition to the qualitative reconstruction shown in the paper, quantitative evaluation of the surface obtained using state-of-the-art methods is shown in Figure 4. The reconstructions shown in paper in Figure 7 for Handstand are compared against the proposed method and the errors are color coded, with red showing the maximum error.

3.3. Motion evaluation

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

Dataset	PRSM [14]	JSR [4]	SCSR [7]	Proposed
Magician	**342 s**	608 s	362 s	353 s
Rachel	397 s	582 s	379 s	**362 s**
Handstand	348 s	566 s	353 s	**325 s**
Juggler2	**413 s**	621 s	405 s	421 s
MagicianLF	659 s	1227 s	622 s	**611 s**

Table 4. Comparison of computational efficiency for a few dynamic sequences against state-of-the-art methods (time in seconds)

3.4. Computation time comparison

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

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 2017. 1, 2

[2] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs/1802.02611, 2018. 1

[3] G. D. Evangelidis and E. Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *TPAMI*, 30(10):1858–1865, 2008. 2

[4] J.-Y. Guillemaut and A. Hilton. Joint Multi-Layer Segmentation and Reconstruction for Free-Viewpoint Video Applications. *IJCV*, 93:73–100, 2010. 1, 2, 4

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017. 1

[6] F. Langguth, K. Sunkavalli, S. Hadap, and M. Goesele. Shading-aware multi-view stereo. In *ECCV*, 2016. 1

[7] A. Mustafa and A. Hilton. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *CVPR*, 2017. 1, 2, 4

[8] A. Mustafa, H. Kim, and A. Hilton. 4d match trees for non-rigid surface alignment. In *ECCV*, 2016. 1

[9] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, and M. Jägersand. Rtseg: Real-time semantic segmentation comparative study. In *ICIP*, 2018. 1

[10] T. Taniai, Y. Matsushita, Y. Sato, and T. Naemura. Continuous 3D Label Stereo Matching using Local Expansion Moves. *TPAMI*, 40(11):2725–2739, 2018. 1

[11] Y.-H. Tsai, G. Zhong, e. B. Yang, Ming-Hsuan", J. Matas, N. Sebe, and M. Welling. Semantic co-segmentation in videos. In *ECCV*, pages 760–775, 2016. 1

[12] C. Vogel, K. Schindler, and S. Roth. 3d scene flow estimation with a piecewise rigid scene model. pages 1–28, 2015. 1

[13] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *ICCV*, pages 1385–1392, 2013. 1

[14] J. Xu, R. Ranftl, and V. Koltun. Accurate Optical Flow via Direct Cost Volume Processing. In *CVPR*, 2017. 1, 4

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 2

[16] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 1, 2