D^{2}Conv3D: Dynamic Dilated Convolutions for Object Segmentation in Videos – Supplementary Material

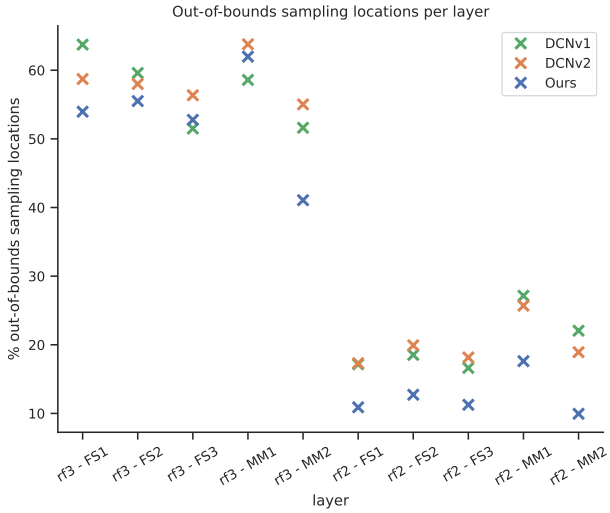

Figure 1: Percentage of out-of-bounds sampling locations, per layer. Measured during inference on DAVIS'16.

1. Modulation Map Visualization

In Fig. 2, we visualize a full volume of predicted modulation values for each of the convolutional layers in the refinement modules when D^{2}Conv3D is applied to [9]. It is visible that every channel reacts to different parts of the foreground or background. Kernel points that potentially sample from neighbouring frames receive higher modulation values on the object boundaries or on the background. Kernel points that sample the current frame, however, have low modulation values in the background and larger modulation values on the object.

2. Out-of-bounds Sampling Behaviour

As mentioned in Sec. 3 of the main paper, we perform a detailed comparison of the percentage of sampling locations that are sampled outside the input feature volume, per convolutional layer, in Fig. 1. It can be observed that D^{2}Conv3D predicts fewer sampling locations beyond the input features than DCNv1 or DCNv2 in most of the cases.

Model	#Params (M)	Time (s/frame)
AGNN [15]	82.3	2.96
CosNet [7]	81.2	0.45
STEm-Seg [1]	50.1	1.42
ADNet [21]	79.3	2.94
MatNet [23]	142.7	0.75*
DFNet [22]	64.7	0.28
3DC-Seg [9]	74.2	0.16
RTNet [11]	277.2	0.29†
Revised Baseline	74.2	0.2
Ours	77.1	0.22
Ours (dense)	77.1	1.07

Table 1: Runtimes during inference on DAVIS'16. Measured on an Nvidia GTX-1080Ti. †Not including time for CRF post-processing. * runtime reported on an Nvidia RTX-2080Ti.

3. Runtime

Although deformable convolutions are not as heavily optimized as regular convolutions, the impact on the runtime is small because we use them only on low-resolution feature maps. Detailed runtimes can be found in Tab. 1

4. Comparison with State-of-the-art

DAVIS 2019: Table. 2 reports the results of the state-of-the-art methods on DAVIS'19 unsupervised validation set. The methods that are grayed out do not use 3D convolutions and hence D^{2}Conv3D cannot be plugged-in to them for a direct comparison. UnOVOST [24] performs the best among all the methods with a $\mathcal{J}\&\mathcal{F}$ score of 67.0%, but it uses multiple 2D networks along with heuristic-based post-processing and hence D^{2}Conv3D cannot be used here as a drop-in replacement to further push its performance. In fact, STEm-Seg [1] is the only method that uses 3D convolutions to incorporate temporal context, and as seen in Table. 2, D^{2}Conv3D improves its performance from 63.4 to 64.6 $\mathcal{J}\&\mathcal{F}$.

YouTube-VIS: We provide an overview of current methods for video instance segmentation on YoutubeVIS[18] in

DAVIS 2019 Unsupervised			
Method	$\mathcal{J}\&\mathcal{F}$ Mean	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean
KIS* [4]	59.9	-	-
UnOVOST* [24]	**67.0**	**67.0**	**68.4**
RVOS [12]	41.2	36.8	45.7
AGNN [15]	61.1	58.9	63.2
STEm-Seg [1]	63.4	60.3	66.5
STEm-Seg +D^{2}Conv3D	**64.6**	**60.8**	**68.5**

Table 2: Results on the validation set of DAVIS'19 unsupervised VOS.

Method	mAP	AP50	AP75	AR1	AR10
FEELVOS[13]	26.9	42.0	29.7	29.9	33.4
IoUTracker+ [18]	23.6	39.2	25.5	26.2	30.9
OSMN [19]	27.5	45.1	29.1	28.6	33.1
DeppSORT [17]	26.1	42.9	26.1	27.8	31.3
MaskTrack R-CNN [18]	30.3	51.1	32.6	31.0	35.5
SeqTracker [18]	27.5	45.7	28.7	29.7	32.5
SipMask [3]	32.5	53.0	33.3	33.5	38.9
CSipMask [10]	35.1	55.6	38.1	35.8	41.7
CMaskTrack R-CNN [10]	32.1	52.8	34.9	33.2	37.9
CompFeat [5]	35.3	56.0	38.6	33.1	40.3
VisTR (Res50) [16]	36.2	59.8	36.9	37.2	42.4
VisTR (Res101) [16]	40.1	**64.0**	45.0	38.3	44.9
MaskProp [2]	**46.6**	–	**51.2**	**44.0**	**52.6**
STEm-Seg [1]	30.6	50.7	33.5	31.6	37.1
STEm-Seg + D^{2}Conv3D	**32.3**	**51.3**	**34.7**	**32.2**	**38.1**

Table 3: Performance comparison on the validation set of YoutubeVIS 2019 [18]. Baseline is STEm-Seg [1] with a ResNet50 backbone.

Method	mAP	AP50	AP75	AR1	AR10
CSipMask [10]	14.3	29.9	12.5	**9.6**	19.3
CMaskTrack R-CNN [10]	15.4	33.9	13.1	9.3	20.0
CrossVIS [20]	**18.1**	**35.5**	**16.9**	–	–
STEm-Seg [1]	14.3	31.5	12.4	10.2	20.7
STEm-Seg + D^{2}Conv3D	**15.2**	**33.8**	**13.7**	**10.6**	**22.2**

Table 4: Performance comparison on the validation set of OVIS [10].

Tab. 3. Again, methods in gray do not use 3D convolutions. The best performing method, MaskProp [2], achieves an impressive score of 46.6 mAP. It extends Mask R-CNN [6] with a mask propagation branch branch; there are no 3D convolutions which we can replace with D^{2}Conv3D in order to boost performance further. STEm-Seg [1] is the only method relying on 3D convolutions. Replacing regular convolutions with D^{2}Conv3D in the decoder increases performance from 30.6 mAP to 32.3 mAP. Despite a weaker ResNet50 backbone, STEm-Seg + D^{2}Conv3D is still competitive to many current architectures.

KITTI-MOTS: Recently, HOTA [8] has been proposed as a metric for tracking and segmentation. We provide HOTA scores for our models in Tab. 5, and compare our performance with Track R-CNN [14]. Our STEm-Seg baseline performs overall better than Track R-CNN; Track R-CNN provides a better detection accuracy (DetA in Tab. 5), while STEm-Seg achieves a better association accuracy. Both methods perform comparable in terms of localization accuracy.

Method	Car				Pedestrian			
	HOTA	DetA	AssA	LocA	HOTA	DetA	AssA	LocA
Track R-CNN [14]	72.3	**77.4**	67.8	88.3	42.1	**54.9**	32.7	78.6
STEm-Seg [1]	73.1	68.6	**78.2**	88.7	47.9	48.8	47.2	79.6
STEm-Seg + DCNv1	73.3	70.4	76.7	88.8	45.5	46.6	44.8	78.5
STEm-Seg + DCNv2	72.7	70.0	75.9	88.7	47.7	47.8	48.1	78.9
STEm-Seg + D^2Conv3D	**74.1**	70.5	**78.2**	**89.4**	**50.1**	50.3	**50.3**	**80.0**

Table 5: HOTA score on the validation set of KITTI MOTS. Baseline is STEm-Seg [1] with a ResNet50 backbone.

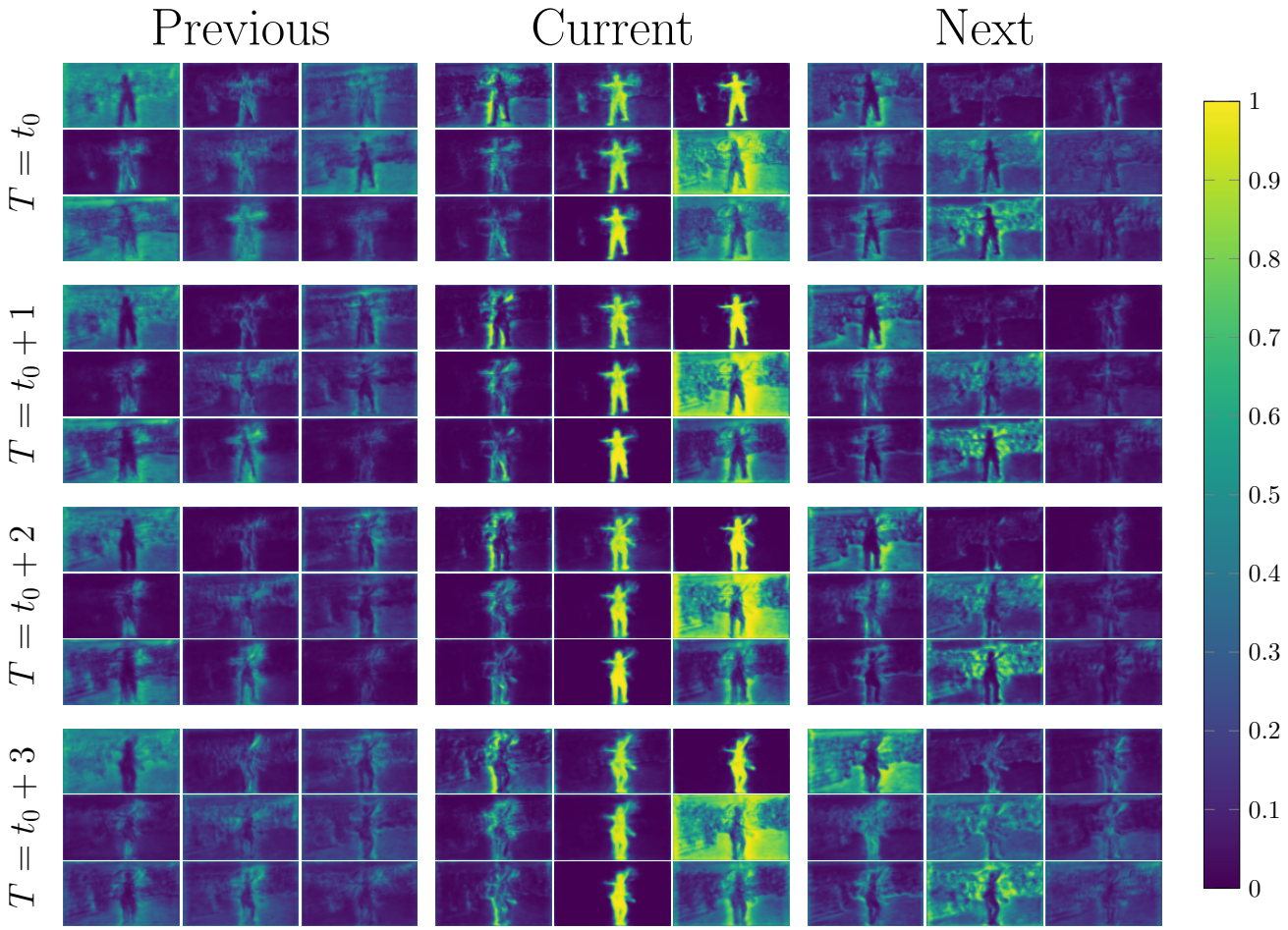

Figure 2: Modulation values predicted during inference on the *dance-twirl* sequence in DAVIS'16. Recall that for a $3 \times 3 \times 3$ convolution, the modulation map $\mathbf{M} \in \mathbb{R}^{T \times H \times W \times K}$ has $K = 27$ channels for each pixel in the input feature map. Here we visualize these 27 channels by splitting them into a row of 3 image blocks, with each block having size 3×3. Consider the row of image blocks for $T = t_0$: here the image block under "Previous" corresponds to the modulation values predicted for those kernel weights which will be applied to the video features in the previous timestep ($T = t_0 - 1$). Likewise, "Curent" and "Next" show the modulation values for the kernel weights which will be applied to the video features from the current ($T = t_0$) and next ($T = t_0 + 1$) timesteps, respectively. The modulation map $\mathbf{M}$ is shown here for a total of 4 time-steps ($t_0, ..., t_0 + 3$); thus, there are four sets of image blocks along the vertical dimension.

References

[1] Ali Athar, Sabarinath Mahadevan, Aljoša Ošep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020.

[2] Gedas Bertasius and L. Torresani. Classifying, segmenting, and tracking object instances in video with mask propaga-

tion. In *CVPR*, 2020.

[3] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. In *ECCV*, 2020.

[4] Donghyeon Cho, Sungeun Hong, Sungil Kang, and Jiwon Kim. Key instance selection for unsupervised video object segmentation. *CVPR Workshops*, 2019.

[5] Yang Fu, Linjie Yang, Ding Liu, Thomas S Huang, and Humphrey Shi. Compfeat: Comprehensive feature aggregation for video instance segmentation. *AAAI*, 2021.

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

[7] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, 2019.

[8] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *IJCV*, 2021.

[9] Sabarinath Mahadevan, Ali Athar, Aljoša Ošep, Sebastian Hennen, Laura Leal-Taixé, and Bastian Leibe. Making a case for 3d convolutions for object segmentation in videos. *BMVC*, 2020.

[10] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation. *arXiv preprint arXiv:2102.01558*, 2021.

[11] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He. Reciprocal transformations for unsupervised video object segmentation. In *CVPR*, 2021.

[12] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. Rvos: End-to-end recurrent network for video object segmentation. In *CVPR*, 2019.

[13] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019.

[14] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *CVPR*, 2019.

[15] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, 2019.

[16] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021.

[17] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017.

[18] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019.

[19] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K. Katsaggelos. Efficient video object segmentation via network modulation. In *CVPR*, 2018.

[20] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. *arXiv preprint arXiv:2104.05970*, 2021.

[21] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip H. S. Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, 2019.

[22] Mingmin Zhen, Shiwei Li, Lei Zhou, Jiaxiang Shang, Haoan Feng, Tian Fang, and Long Quan. Learning discriminative feature with crf for unsupervised video object segmentation. In *ECCV*, 2020.

[23] Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *TIP*, 2020.

[24] I. E. Zulfikar, J. Luiten, and B. Leibe. UnOVOST: Unsupervised Offline Video Object Segmentation and Tracking for the 2019 Unsupervised DAVIS Challenge. *CVPR Workshops*, 2019.