Video Propagation Networks

Varun Jampani[1], Raghudeep Gadde[1,2], Peter V. Gehler[1,2,3]

[1]MPI for Intelligent Systems, Tübingen; [2]Bernstein Center for Computational Neuroscience, Tübingen; [3]University of Würzberg, Würzberg

VPNs are end-to-end trainable, can propagate any type of video content, handle large number of input frames and have favorable runtime.

Code: `http://varunjampani.github.io/vpn`

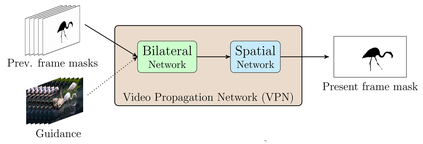
1 Video Information Propagation

Aim: To devise a general neural network approach for structured information propagation across video frames.

Useful for many video applications:
- Video object segmentation, Semantic video segmentation, Video color propagation etc.

Challenges:
- To account for *scene and camera motion*.
- To process *large number of video frames*.

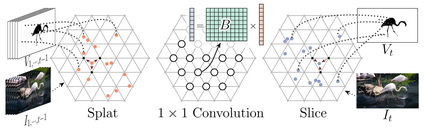
Approach: A neural network with *video adaptive receptive fields* and *some tricks* for computational efficiency.



Overview of Video Propagation Network with Bilateral and Spatial Networks

2 Video Propagation Networks with Learnable Bilateral Filters

Bilateral Convolution Layer (BCL) [1, 2] for forward propagation across time:
1. **Splat** previous frame results onto 6 dimensional permutohedral lattice constructed with *(x,y,r,g,b,t)* features.
2. **Convolve** the high-dimensional signal with learnable filters [1].
3. **Slice** the resulting signal onto the *present frame*.

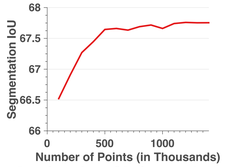


Bilateral Convolution Layer (BCL) for Forward Information Propagation

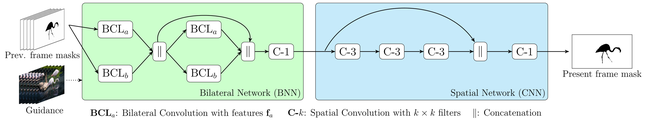
Compared to spatial convolutions:
- Efficient sparse high-dimensional filtering with hash table.
- Image adaptive receptive fields.
- Input and output points can be different.

Some tricks for efficient video propagation:
- Splat only a *random set of points* or *superpixels* from previous frames.
- Use 1x1 convolutions in high-dimensional space as building neighborhood matrix is time-consuming.

Advantages of VPN:
- *General applicability*
- *Online propagation*
- *End-to-end trainable*
- *Long range and image-adaptive filtering*
- *Runtime*



Random Set of Points vs. IoU



Schematic of Video Propagation Network Architecture

3 Application: Object Segmentation

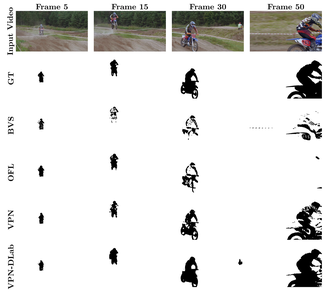
Task: Given an object mask for the first video frame, predict the corresponding object masks for the entire video.

State-of-the-art results on DAVIS dataset [3] at the time of submission with *favorable runtime* despite VPN *not* using any strong deep appearance model.

VPN can be easily integrated into other deep architectures. We show a boost in performance when augmenting VPN with DeepLab [4] model (*VPN-DeepLab*).

Results on DAVIS [3] Dataset

	IoU↑	$\mathcal{F}$ ↑	$\mathcal{T}$ ↓	*Time*(s)
BNN-Identity	67.0	67.1	36.3	0.21
VPN-Stage1	70.1	68.4	30.1	0.48
VPN-Stage2	71.3	68.9	30.2	0.75
With pre-trained appearance models				
DeepLab	57.0	49.9	47.8	0.15
VPN-DeepLab	**75.0**	**72.4**	29.5	0.63
OFL[Tsai et al.'16]	71.1	67.9	22.1	>60
BVS[Märki et al.'16]	66.5	65.6	31.6	0.37
NLC[Faktor et al.'14]	64.1	59.3	35.6	20
FCP[Perazzi et al.'15]	63.1	54.6	28.5	12
JMP[Fan et al.'15]	60.7	58.6	**13.2**	12
HVS[Grundmann et al.'10]	59.6	57.6	29.7	5
SEA[Ramakanth et al.'14]	55.6	53.3	13.7	6



Sample Result of Video Object Segmentation

4 Application: Semantic Video Segmentation

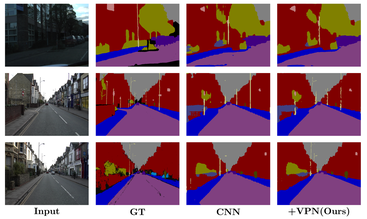
Task: Refine a semantic video segmentation result obtained by a CNN run on each frame independently.

VPN works *better yet faster* compared to video CRF techniques such as FSO-CRF [6].

Using optical flow to normalize the bilateral features (VPN-Flow) results in IoU improvements.

Results on CamVid [5] Test Dataset

	IoU	*Time*(s)
CNN [Yu et al.'16]	65.3	0.38
+ FSO-CRF [6]	66.1	>10
+ BNN-Identity	65.3	0.31
+ BNN-Identity-Flow	65.5	0.33
+ VPN (Ours)	66.5	0.35
+ VPN-Flow (Ours)	**66.7**	0.37
CNN [Richter et al.'16]	68.9	0.30
+ VPN-Flow (Ours)	**69.5**	0.38



Sample Results of Semantic Video Segmentation

5 Application: Video Color Propagation

Task: Propagate color from first frame to the remaining video frames in a given grayscale video.

VPN works *faster and better* compared to traditional optimization based techniques [7].

PSNR Results on DAVIS [3] Dataset

	PSNR	*Runtime*(s)
BNN-Identity	27.89	0.29
VPN	**28.15**	0.90
Levin et al.'04	27.11	19

Sample Result of Video Color Propagation

6 Conclusion

VPN is the first neural network approach for video information propagation that is both *computationally efficient* and is *useful* for a *wide range* of video propagation tasks.

References:
1. Jampani, V. et al. Learning sparse high-Dimensional Filters: Image filtering, dense CRFs and bilateral neural networks . *In CVPR*, 2016.
2. Adams, A., Baek, J., & Davis, M. A. Fast high-dimensional filtering using the permutohedral lattice. *In Computer Graphics Forum 29(2)*, 2010.
3. Perazzi, F. et al. A benchmark dataset and evaluation methodology for video object segmentation, *In CVPR*, 2016.
4. Liang Chieh, C. et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *In ICLR*, 2015.
5. Brostow, G. et al. Semantic object classes in video: A high-definition ground truth database, *In Pattern Recog. Letters*, 2009.
6. Kundu, A. et al. Feature space optimization for semantic video segmentation, *In CVPR*, 2016.
7. Levin, A. et al. Colorization using optimization, *In ToG*, 2004.

{varun.jampani, raghudeep.gadde, peter.gehler}@tuebingen.mpg.de