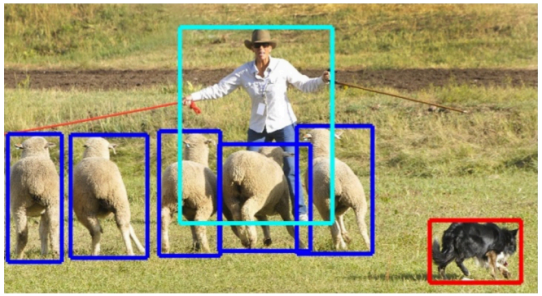
Accel: A Corrective Fusion Network for Efficient Semantic Segmentation on Video

Samvit Jain, Xin Wang**,** Joseph Gonzalez
RISE Lab, UC Berkeley

Semantic segmentation

Image classification



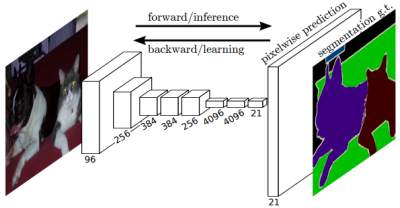
Object detection

Semantic segmentation

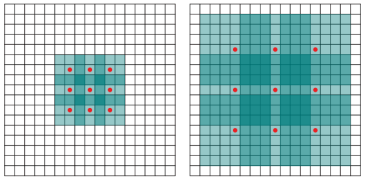
Evolution

Efficient Graph-Based Image Segmentation (2004)

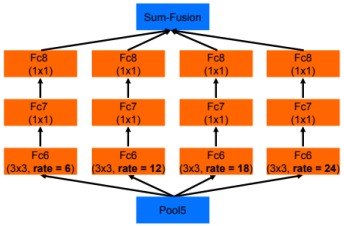
...



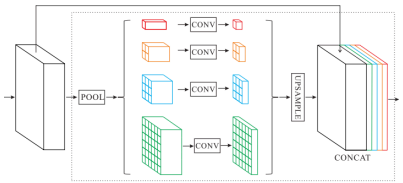
Fully Convolutional Networks for SS (2014)



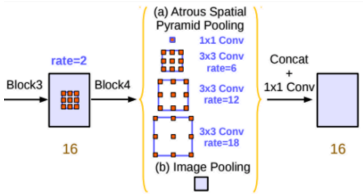
Multi-Scale Aggregation by Dilated Convolutions (2015)



DeepLab-v2 (2016)

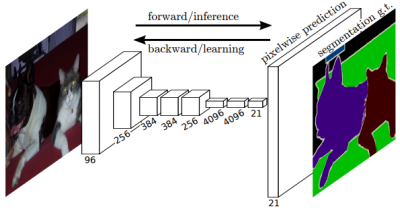


PSPNet (2017)

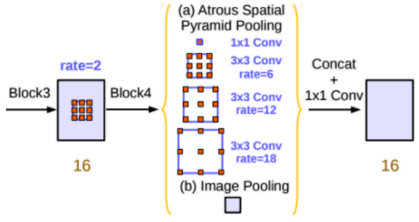


DeepLab-v3 (2017)

Evolution



Fully Convolutional Networks (2014)



DeepLab-v3 (2017)

Dataset	Pascal VOC 2012	
Accuracy (mIoU)	62.2	85.7
Inference Time	175 ms	750 ms

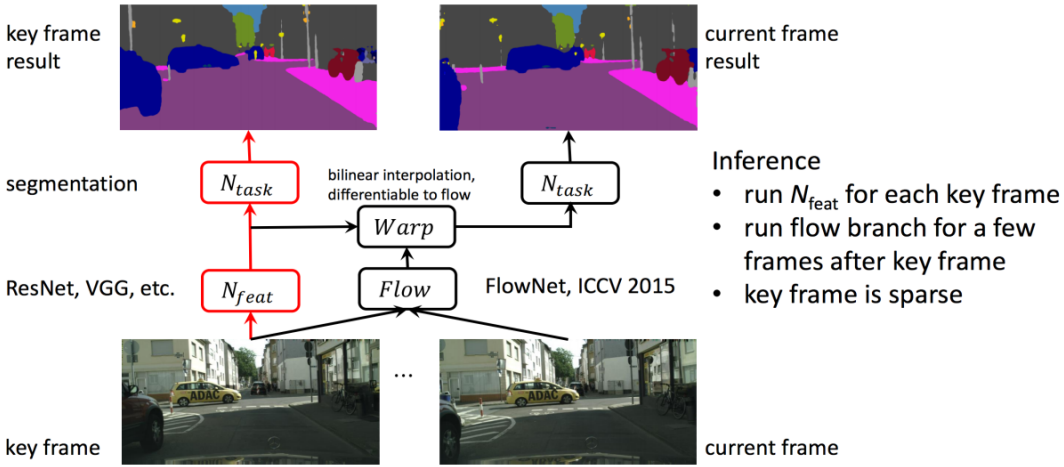
Motivation

- Image models don’t translate to video
  - High frame rates (e.g. 30 fps)
  - High resolution (e.g. full-HD, 1920 x 1080 p)
  - Scene complexity (e.g. ego motion, urban streets)

Cityscapes dataset: Frankfurt

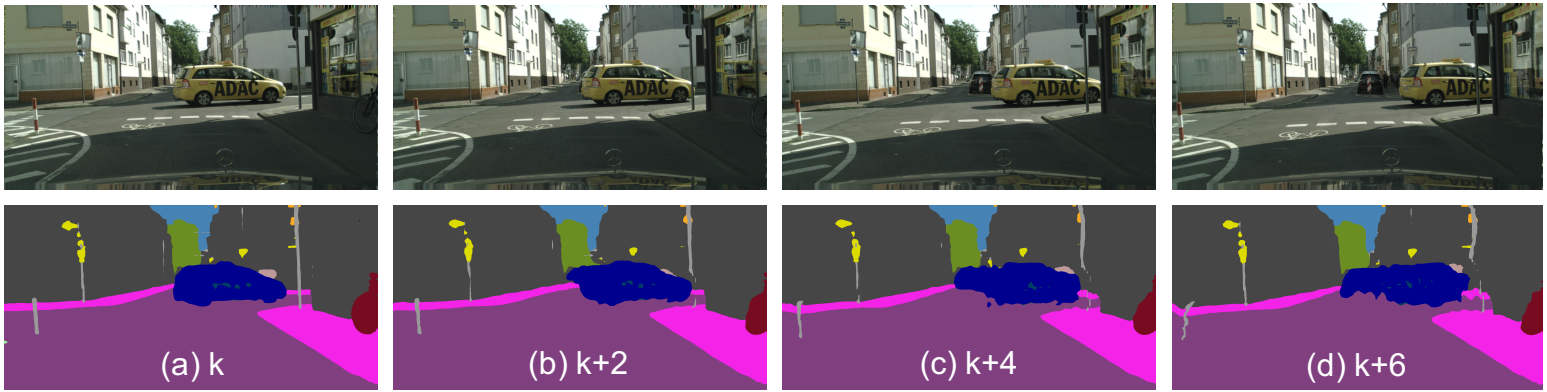
Deep Feature Flow

- Idea: run feature net on **keyframes**, warp features to **intermediate frames**

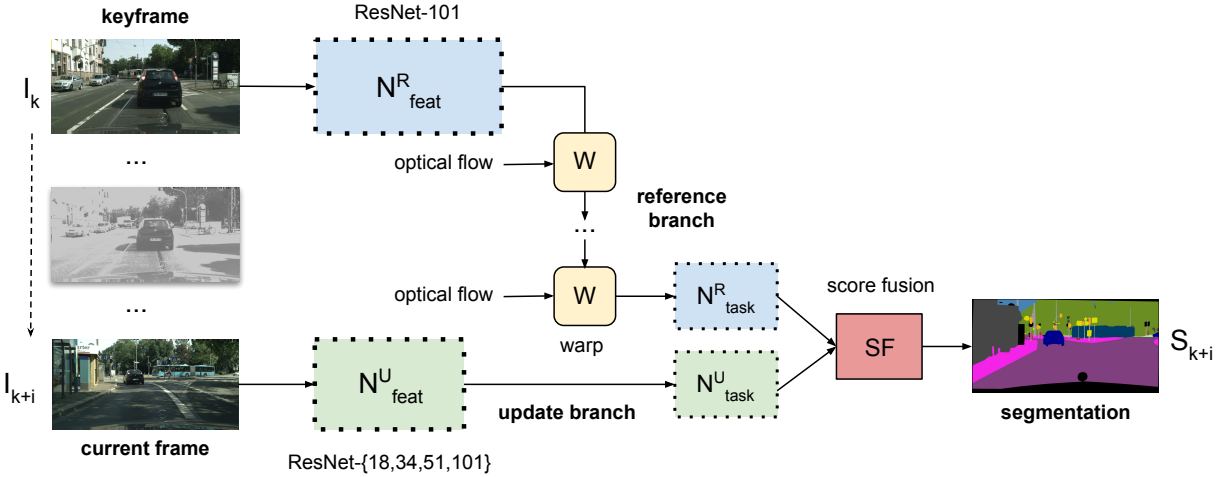


Problems

- Accuracy degradation
  - Warping with a flow field is a coarse operation
  - Non-translational temporal change (e.g. new objects, occlusions, lighting) ignored



Accel



Accel: a family of corrective, two-stream fusion networks combining:

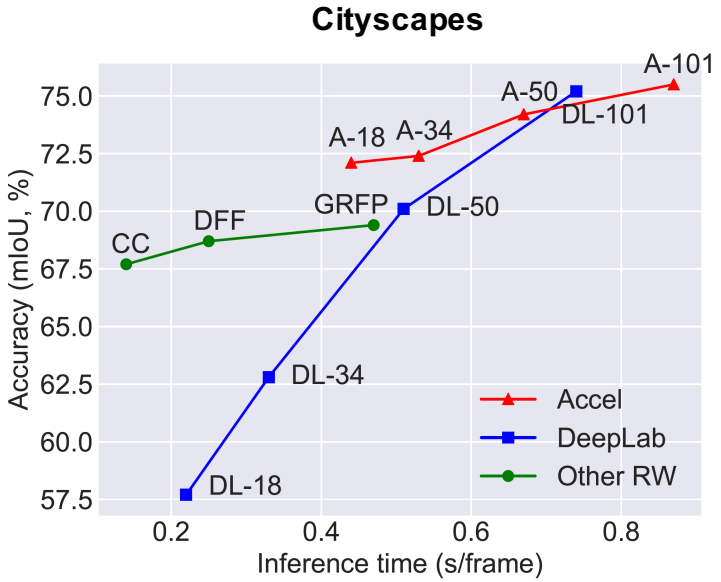
(1) N^R (**reference branch**) – optical flow-based keyframe feature warping

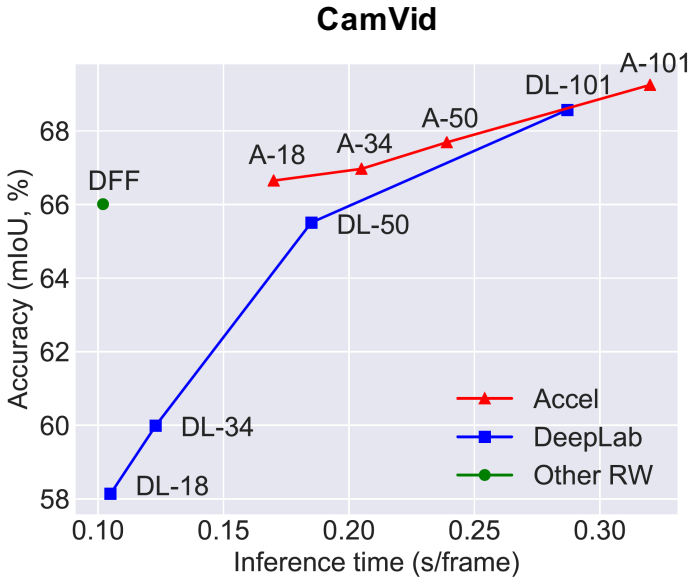
(2) N^U (**update branch**) – per-frame correction with residual segmentation network

Accel

N^R_{feat} (reference branch)	N^U_{feat} (update branch)	$N^R + N^U$ (full network)
ResNet-101	**ResNet-18**	**Accel-18**
ResNet-101	ResNet-34	Accel-34
ResNet-101	ResNet-51	Accel-51
ResNet-101	**ResNet-101**	**Accel-101**

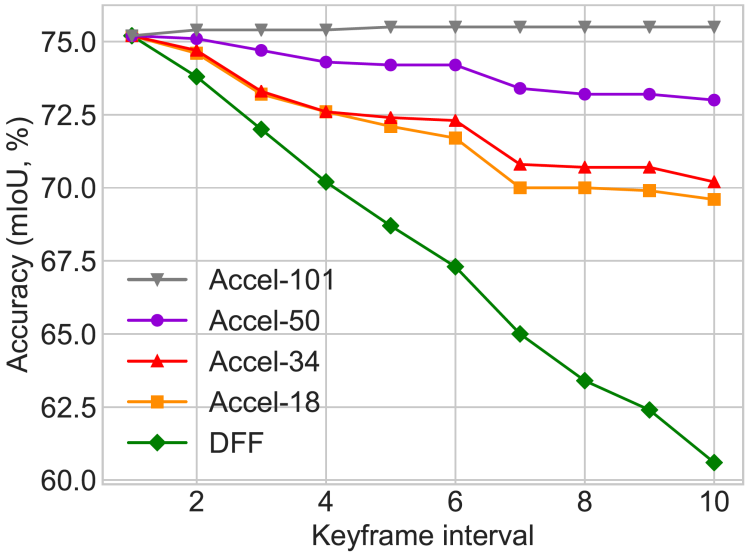
Results





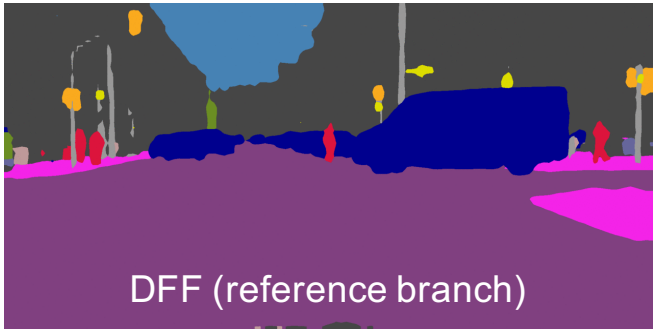
Accuracy (mIoU) vs. **inference time** (s/frame)

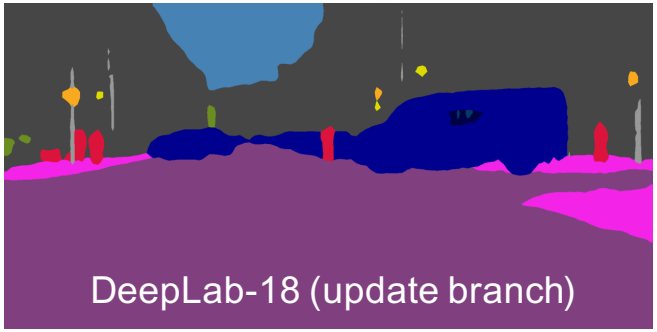
Results

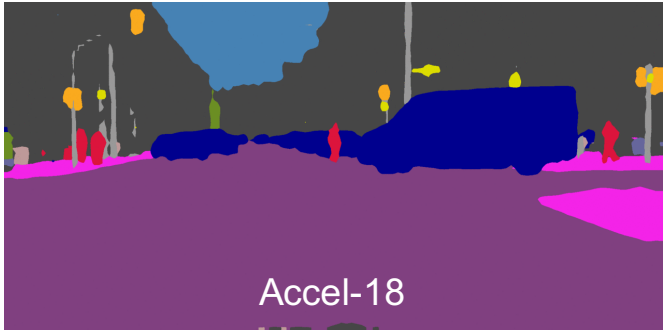


Accuracy (mIoU) vs. **keyframe interval**

Visualizations







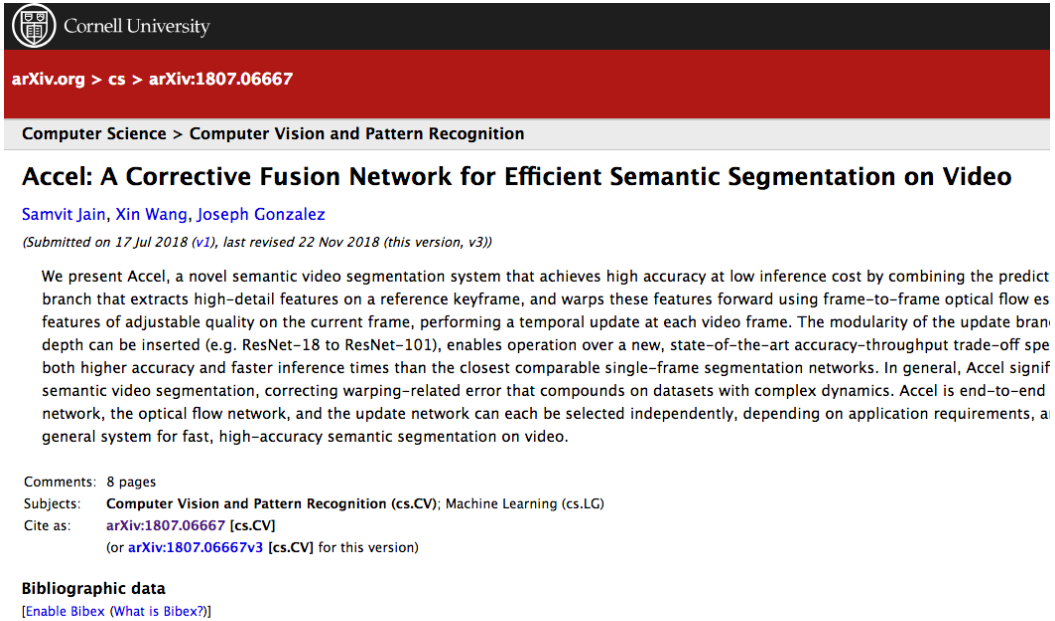
Thank you!

Accel: A Corrective Fusion Network for Efficient Semantic Segmentation on Video

S. Jain, X. Wang, J. Gonzalez

In: CVPR 2019 (oral)

https://arxiv.org/abs/1807.06667



Cornell University

arXiv.org > cs > arXiv:1807.06667

Computer Science > Computer Vision and Pattern Recognition

Accel: A Corrective Fusion Network for Efficient Semantic Segmentation on Video

Samvit Jain, Xin Wang, Joseph Gonzalez

(Submitted on 17 Jul 2018 (v1), last revised 22 Nov 2018 (this version, v3))

We present Accel, a novel semantic video segmentation system that achieves high accuracy at low inference cost by combining the predict branch that extracts high-detail features on a reference keyframe, and warps these features forward using frame-to-frame optical flow es features of adjustable quality on the current frame, performing a temporal update at each video frame. The modularity of the update bran depth can be inserted (e.g. ResNet-18 to ResNet-101), enables operation over a new, state-of-the-art accuracy-throughput trade-off spe both higher accuracy and faster inference times than the closest comparable single-frame segmentation networks. In general, Accel signif semantic video segmentation, correcting warping-related error that compounds on datasets with complex dynamics. Accel is end-to-end network, the optical flow network, and the update network can each be selected independently, depending on application requirements, a general system for fast, high-accuracy semantic segmentation on video.

Comments: 8 pages
Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Machine Learning (cs.LG)
Cite as: **arXiv:1807.06667 [cs.CV]**
(or **arXiv:1807.06667v3 [cs.CV]** for this version)

Bibliographic data

[Enable Bibex (What is Bibex?)]