Video Object Segmentation using Deep Learning

Zack While

Advised by: Dr. Chen Chen, Rui Hou, and Dr. Mubarak Shah

July 14, 2017

Youngstown State University



UCF | Center for Research in Computer Vision

UNIVERSITY OF CENTRAL FLORIDA



Table of Contents

1. Problem Overview

2. Related Work

3. Our Approach

4. Results

5. Future Work

Problem Overview

Problem Description

- **Objective:** recognizing the predominant object in a video clip and distinguishing it per-pixel from the background in each frame.
- **Example:**

Figure 1: Input Video

Figure 2: Foreground Mask

Problem Description

- **Objective:** recognizing the predominant object in a video clip and distinguishing it per-pixel from the background in each frame.
- **Example:**

Figure 1: Input Video



Figure 2: Foreground Mask



Further Background

- **Applications:**
  - Video surveillance
  - Machine vision
  - Content-based video retrieval
- **Challenges:**
  - Spatial *and* temporal information
  - Camera motion
  - Main object changing appearance
  - Main object obstruction
- **Project goals:**
  - Training an end-to-end 3D CNN to perform video object segmentation.
  - Comparing results to the current state of the art on the DAVIS 2016 dataset.
  - Fine-tuning the model to generate results that are competitive with the current state of the art.

Further Background

- **Applications:**
  - Video surveillance
  - Machine vision
  - Content-based video retrieval
- **Challenges:**
  - Spatial *and* temporal information
  - Camera motion
  - Main object changing appearance
  - Main object obstruction
- **Project goals:**
  - Training an end-to-end 3D CNN to perform video object segmentation.
  - Comparing results to the current state of the art on the DAVIS 2016 dataset.
  - Fine-tuning the model to generate results that are competitive with the current state of the art.

Further Background

- **Applications:**
  - Video surveillance
  - Machine vision
  - Content-based video retrieval
- **Challenges:**
  - Spatial *and* temporal information
  - Camera motion
  - Main object changing appearance
  - Main object obstruction
- **Project goals:**
  - Training an end-to-end 3D CNN to perform video object segmentation.
  - Comparing results to the current state of the art on the DAVIS 2016 dataset.
  - Fine-tuning the model to generate results that are competitive with the current state of the art.

Further Background

- **Applications:**
  - Video surveillance
  - Machine vision
  - Content-based video retrieval
- **Challenges:**
  - Spatial *and* temporal information
  - Camera motion
  - Main object changing appearance
  - Main object obstruction
- **Project goals:**
  - Training an end-to-end 3D CNN to perform video object segmentation.
  - Comparing results to the current state of the art on the DAVIS 2016 dataset.
  - Fine-tuning the model to generate results that are competitive with the current state of the art.

Further Background

- **Applications:**
  - Video surveillance
  - Machine vision
  - Content-based video retrieval
- **Challenges:**
  - Spatial *and* temporal information
  - Camera motion
  - Main object changing appearance
  - Main object obstruction
- **Project goals:**
  - Training an end-to-end 3D CNN to perform video object segmentation.
  - Comparing results to the current state of the art on the DAVIS 2016 dataset.
  - Fine-tuning the model to generate results that are competitive with the current state of the art.

Further Background

- **Applications:**
  - Video surveillance
  - Machine vision
  - Content-based video retrieval
- **Challenges:**
  - Spatial *and* temporal information
  - Camera motion
  - Main object changing appearance
  - Main object obstruction
- **Project goals:**
  - Training an end-to-end 3D CNN to perform video object segmentation.
  - Comparing results to the current state of the art on the DAVIS 2016 dataset.
  - Fine-tuning the model to generate results that are competitive with the current state of the art.

Related Work

Literature Review

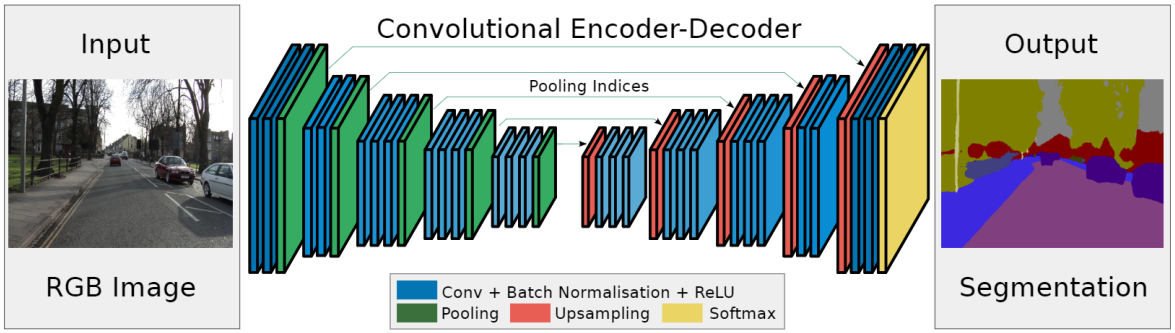
- **SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation** [1]

- **Learning Spatiotemporal Features with 3D Convolutional Networks** [4]

- **FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos** [2]

Literature Review

- **SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation** [1]



- **Learning Spatiotemporal Features with 3D Convolutional Networks** [4]

- **FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos** [2]

Literature Review

- SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation [1]

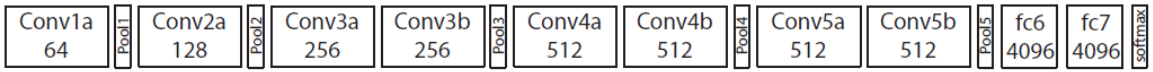
- **Learning Spatiotemporal Features with 3D Convolutional Networks** [4]

- FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos [2]

Literature Review

- SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation [1]

- **Learning Spatiotemporal Features with 3D Convolutional Networks** [4]



- FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos [2]

Literature Review

- SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation [1]

- Learning Spatiotemporal Features with 3D Convolutional Networks [4]

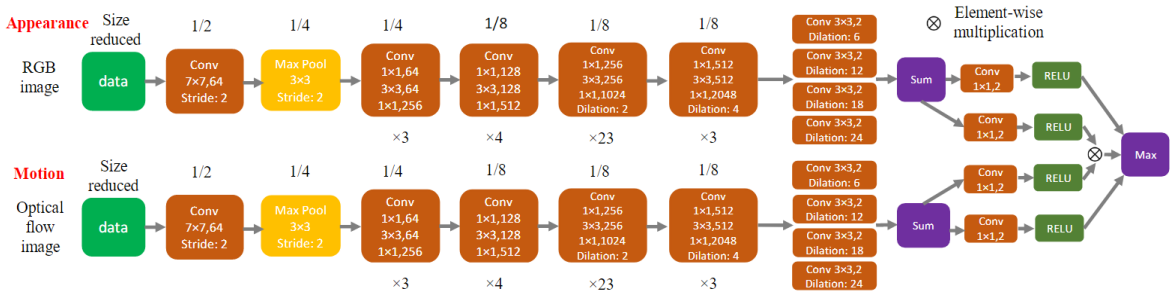
- **FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos** [2]

Literature Review

- SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation [1]

- Learning Spatiotemporal Features with 3D Convolutional Networks [4]

- **FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos** [2]



Our Approach

General Information

- **Dataset:** DAVIS 2016 [3]
  - 50 total HD video sequences with pixel-wise ground truth per frame
  - Split with 30 for training, 20 for evaluation
  - Using 480p versions of frames
- **Data Augmentation:**
  - Horizontal flip
  - Reversing the video sequence
- **Post-Processing:** morphological transformations
  1. Removing objects of 5500 connected pixels or less
  2. Filling in holes of 20,000 connected pixels or less

General Information

- **Dataset:** DAVIS 2016 [3]
  - 50 total HD video sequences with pixel-wise ground truth per frame
  - Split with 30 for training, 20 for evaluation
  - Using 480p versions of frames

- **Data Augmentation:**
  - Horizontal flip
  - Reversing the video sequence

- **Post-Processing:** morphological transformations
  1. Removing objects of 5500 connected pixels or less
  2. Filling in holes of 20,000 connected pixels or less

General Information

- **Dataset:** DAVIS 2016 [3]
  - 50 total HD video sequences with pixel-wise ground truth per frame
  - Split with 30 for training, 20 for evaluation
  - Using 480p versions of frames

- **Data Augmentation:**
  - Horizontal flip
  - Reversing the video sequence

- **Post-Processing:** morphological transformations
  1. Removing objects of 5500 connected pixels or less
  2. Filling in holes of 20,000 connected pixels or less

General Information

- **Dataset:** DAVIS 2016 [3]
  - 50 total HD video sequences with pixel-wise ground truth per frame
  - Split with 30 for training, 20 for evaluation
  - Using 480p versions of frames

- **Data Augmentation:**
  - Horizontal flip
  - Reversing the video sequence

- **Post-Processing:** morphological transformations
  1. Removing objects of 5500 connected pixels or less
  2. Filling in holes of 20,000 connected pixels or less

General Information

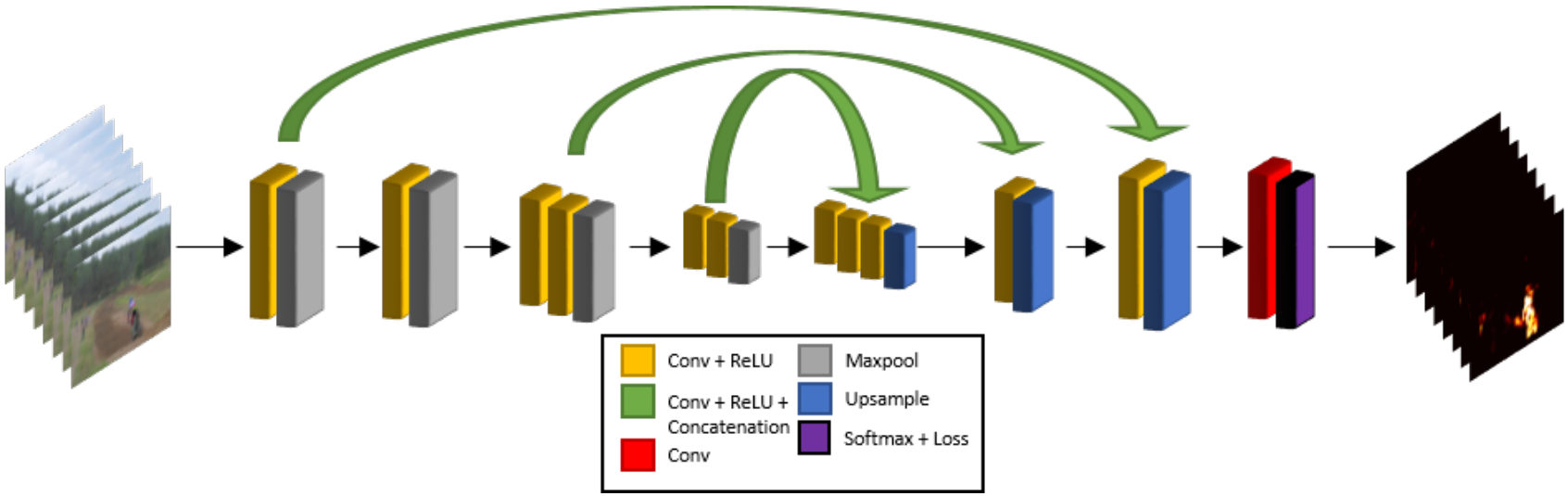
- **Dataset:** DAVIS 2016 [3]
  - 50 total HD video sequences with pixel-wise ground truth per frame
  - Split with 30 for training, 20 for evaluation
  - Using 480p versions of frames
- **Data Augmentation:**
  - Horizontal flip
  - Reversing the video sequence
- **Post-Processing:** morphological transformations
  1. Removing objects of 5500 connected pixels or less
  2. Filling in holes of 20,000 connected pixels or less

General Information

- **Dataset:** DAVIS 2016 [3]
  - 50 total HD video sequences with pixel-wise ground truth per frame
  - Split with 30 for training, 20 for evaluation
  - Using 480p versions of frames
- **Data Augmentation:**
  - Horizontal flip
  - Reversing the video sequence
- **Post-Processing:** morphological transformations
  1. Removing objects of 5500 connected pixels or less
  2. Filling in holes of 20,000 connected pixels or less

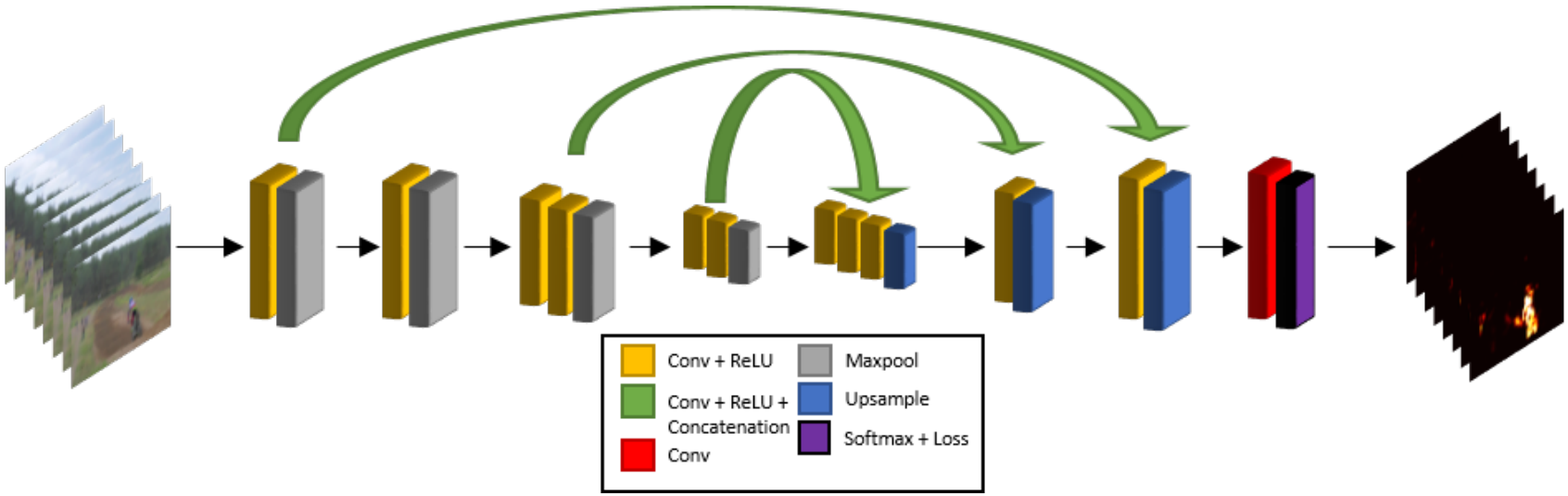
3D CNN Architecture

- **Input:** 8 frames, immediately reduced to 240×320 (bilinear interpolation)
- **Output:** 8 mask heatmaps, immediately resized to original dimensions (bicubic interpolation)
- Heatmap pixel values $\in [0, 1]$
- Each pixel's value represents likelihood that it is the foreground object
- Heatmap thresholded to create video object mask.



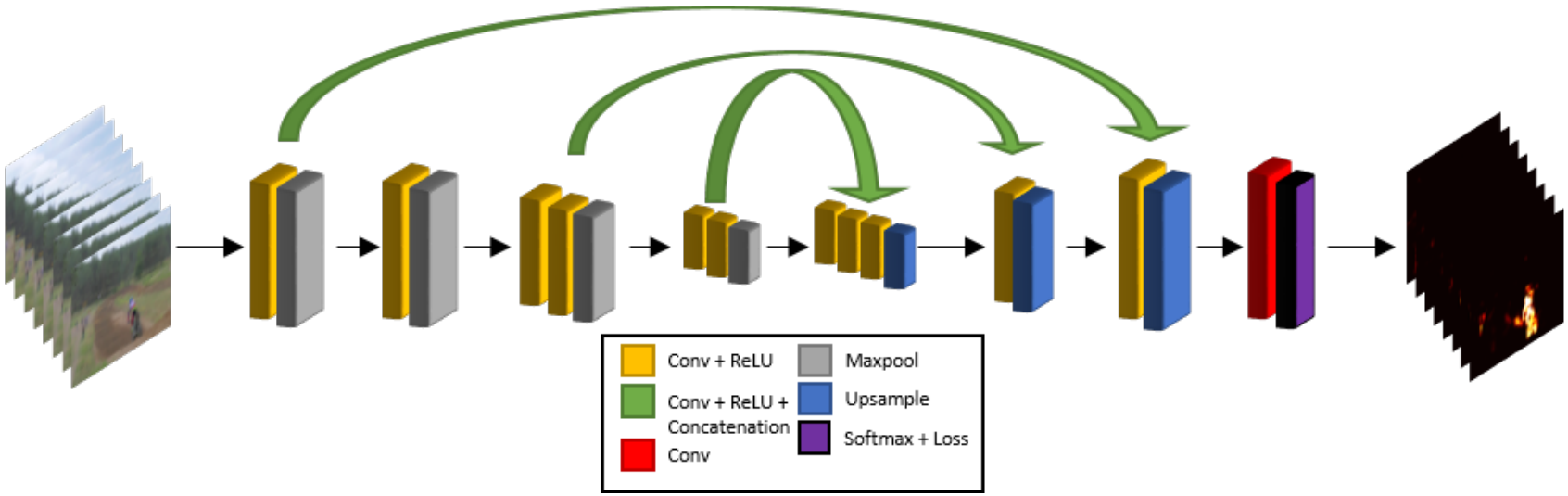
3D CNN Architecture

- **Input:** 8 frames, immediately reduced to 240×320 (bilinear interpolation)
- **Output:** 8 mask heatmaps, immediately resized to original dimensions (bicubic interpolation)
- Heatmap pixel values $\in [0, 1]$
- Each pixel's value represents likelihood that it is the foreground object
- Heatmap thresholded to create video object mask.



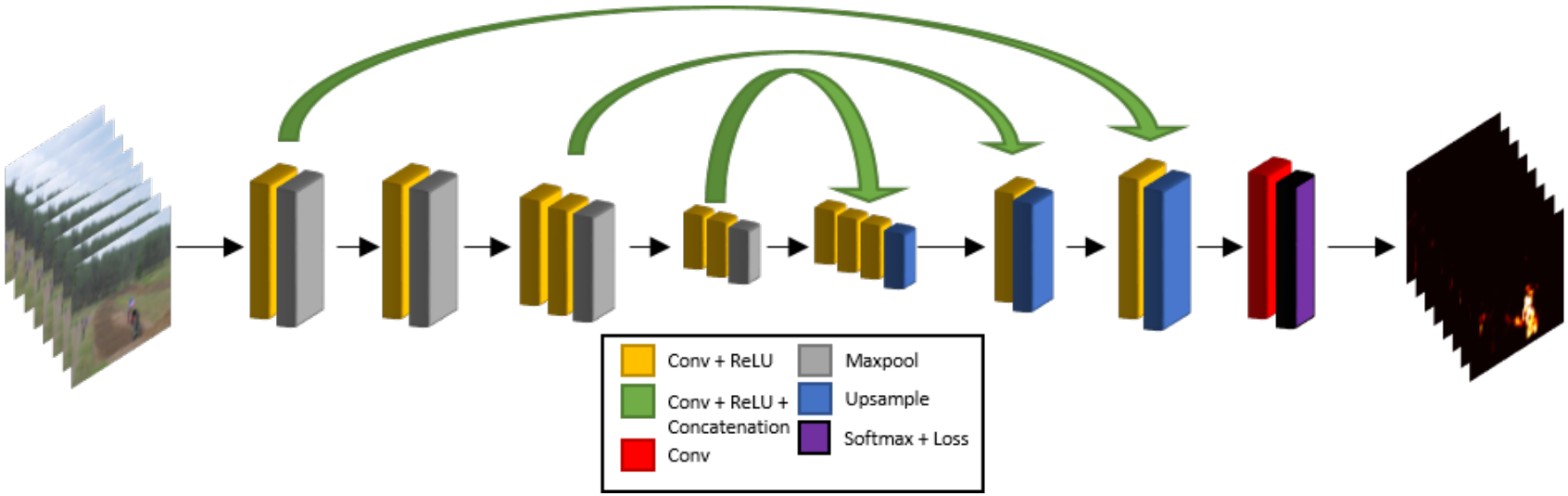
3D CNN Architecture

- **Input:** 8 frames, immediately reduced to 240×320 (bilinear interpolation)
- **Output:** 8 mask heatmaps, immediately resized to original dimensions (bicubic interpolation)
- Heatmap pixel values $\in [0, 1]$
- Each pixel's value represents likelihood that it is the foreground object
- Heatmap thresholded to create video object mask.



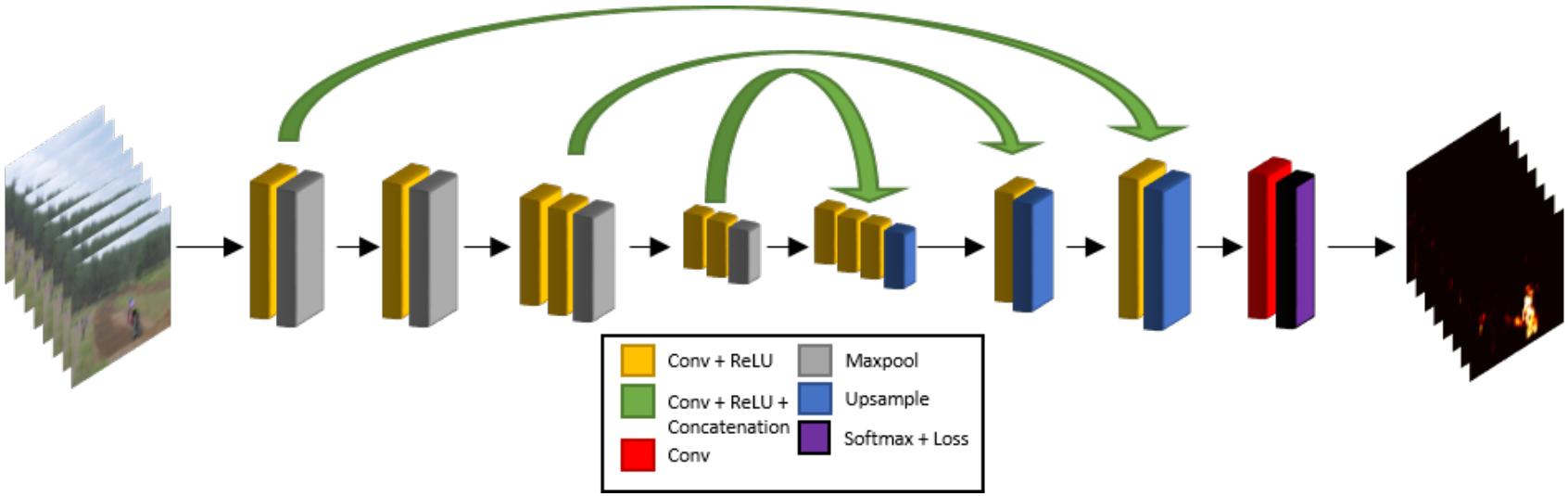
3D CNN Architecture

- **Input:** 8 frames, immediately reduced to 240×320 (bilinear interpolation)
- **Output:** 8 mask heatmaps, immediately resized to original dimensions (bicubic interpolation)
- Heatmap pixel values $\in [0, 1]$
- Each pixel's value represents likelihood that it is the foreground object
- Heatmap thresholded to create video object mask.



3D CNN Architecture

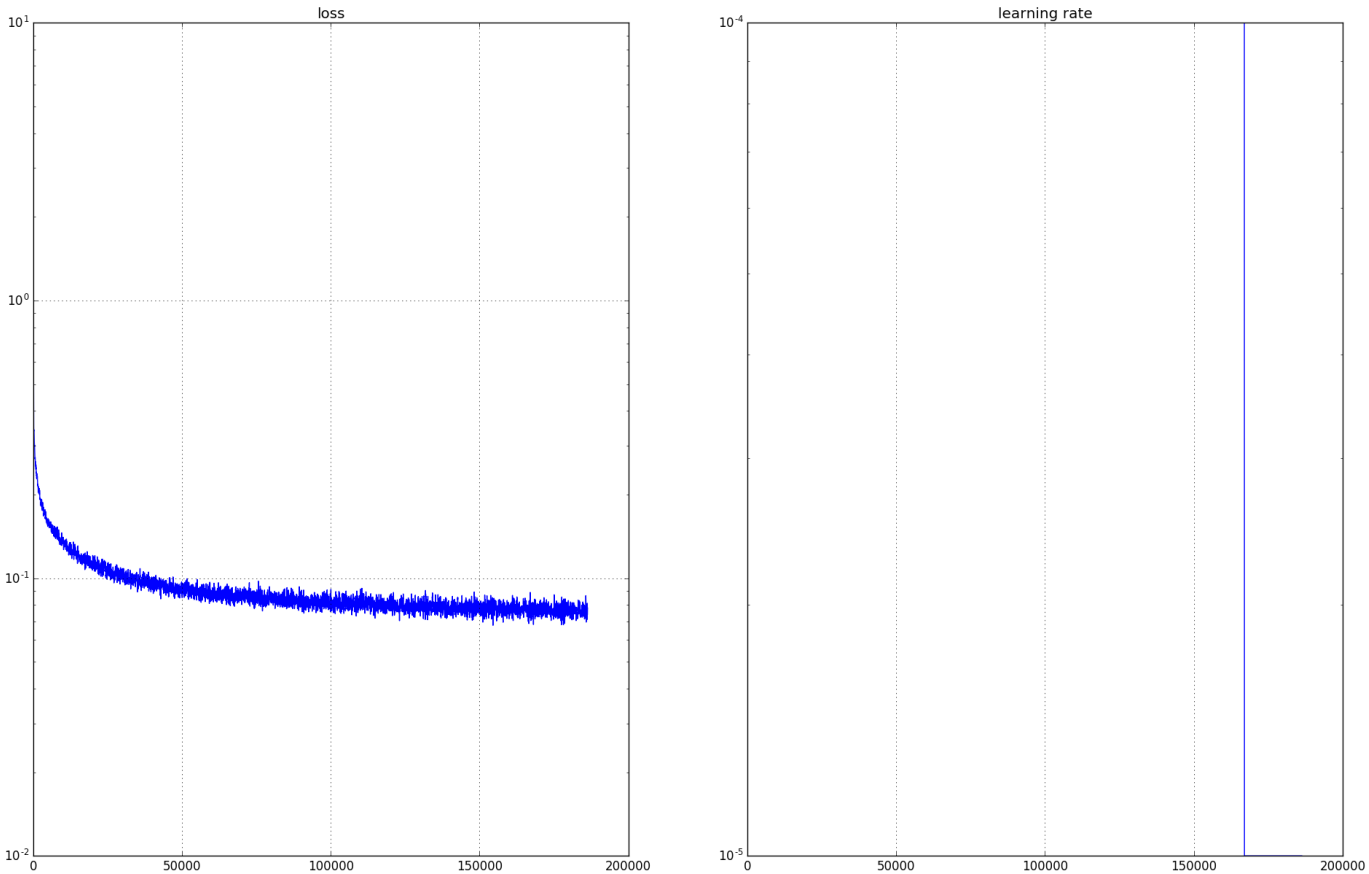
- **Input:** 8 frames, immediately reduced to 240×320 (bilinear interpolation)
- **Output:** 8 mask heatmaps, immediately resized to original dimensions (bicubic interpolation)
- Heatmap pixel values $\in [0, 1]$
- Each pixel's value represents likelihood that it is the foreground object
- Heatmap thresholded to create video object mask.



Results

Training Loss Curve

Figure 3: Training Loss Curve after 189,000 Iterations



- Calculate the Intersection over Union (IoU), also known as the *Jaccard Index* [3] to quantify the quality of a predicted mask
- For a ground truth mask T and predicted mask P, IoU is defined as $\mathcal{J} = \frac{|T \cap P|}{|T \cup P|}$ [3].
- Based on comparing pixels between the two masks, and the goal is to get as close to 1.0 (100%) as possible.

Table 1: DAVIS 2016 mIoU Comparison

Method	FST	CUT	NLC	MSG	KEY	UCF-C3D	CVOS	TRC	SAL
mIoU	0.558	0.552	0.551	0.533	0.498	0.486	0.482	0.473	0.393

- Calculate the Intersection over Union (IoU), also known as the *Jaccard Index* [3] to quantify the quality of a predicted mask
- For a ground truth mask T and predicted mask P, IoU is defined as $\mathcal{J} = \frac{|T \cap P|}{|T \cup P|}$ [3].
- Based on comparing pixels between the two masks, and the goal is to get as close to 1.0 (100%) as possible.

Table 1: DAVIS 2016 mIoU Comparison

Method	FST	CUT	NLC	MSG	KEY	UCF-C3D	CVOS	TRC	SAL
mIoU	0.558	0.552	0.551	0.533	0.498	0.486	0.482	0.473	0.393

- Calculate the Intersection over Union (IoU), also known as the *Jaccard Index* [3] to quantify the quality of a predicted mask
- For a ground truth mask T and predicted mask P, IoU is defined as $\mathcal{J} = \frac{|T \cap P|}{|T \cup P|}$ [3].
- Based on comparing pixels between the two masks, and the goal is to get as close to 1.0 (100%) as possible.

Table 1: DAVIS 2016 mIoU Comparison

Method	FST	CUT	NLC	MSG	KEY	UCF-C3D	CVOS	TRC	SAL
mIoU	0.558	0.552	0.551	0.533	0.498	0.486	0.482	0.473	0.393

- Calculate the Intersection over Union (IoU), also known as the *Jaccard Index* [3] to quantify the quality of a predicted mask
- For a ground truth mask T and predicted mask P, IoU is defined as $\mathcal{J} = \frac{|T \cap P|}{|T \cup P|}$ [3].
- Based on comparing pixels between the two masks, and the goal is to get as close to 1.0 (100%) as possible.

Table 1: DAVIS 2016 mIoU Comparison

Method	FST	CUT	NLC	MSG	KEY	UCF-C3D	CVOS	TRC	SAL
mIoU	0.558	0.552	0.551	0.533	0.498	0.486	0.482	0.473	0.393

- Calculate the Intersection over Union (IoU), also known as the *Jaccard Index* [3] to quantify the quality of a predicted mask
- For a ground truth mask T and predicted mask P, IoU is defined as $\mathcal{J} = \frac{|T \cap P|}{|T \cup P|}$ [3].
- Based on comparing pixels between the two masks, and the goal is to get as close to 1.0 (100%) as possible.

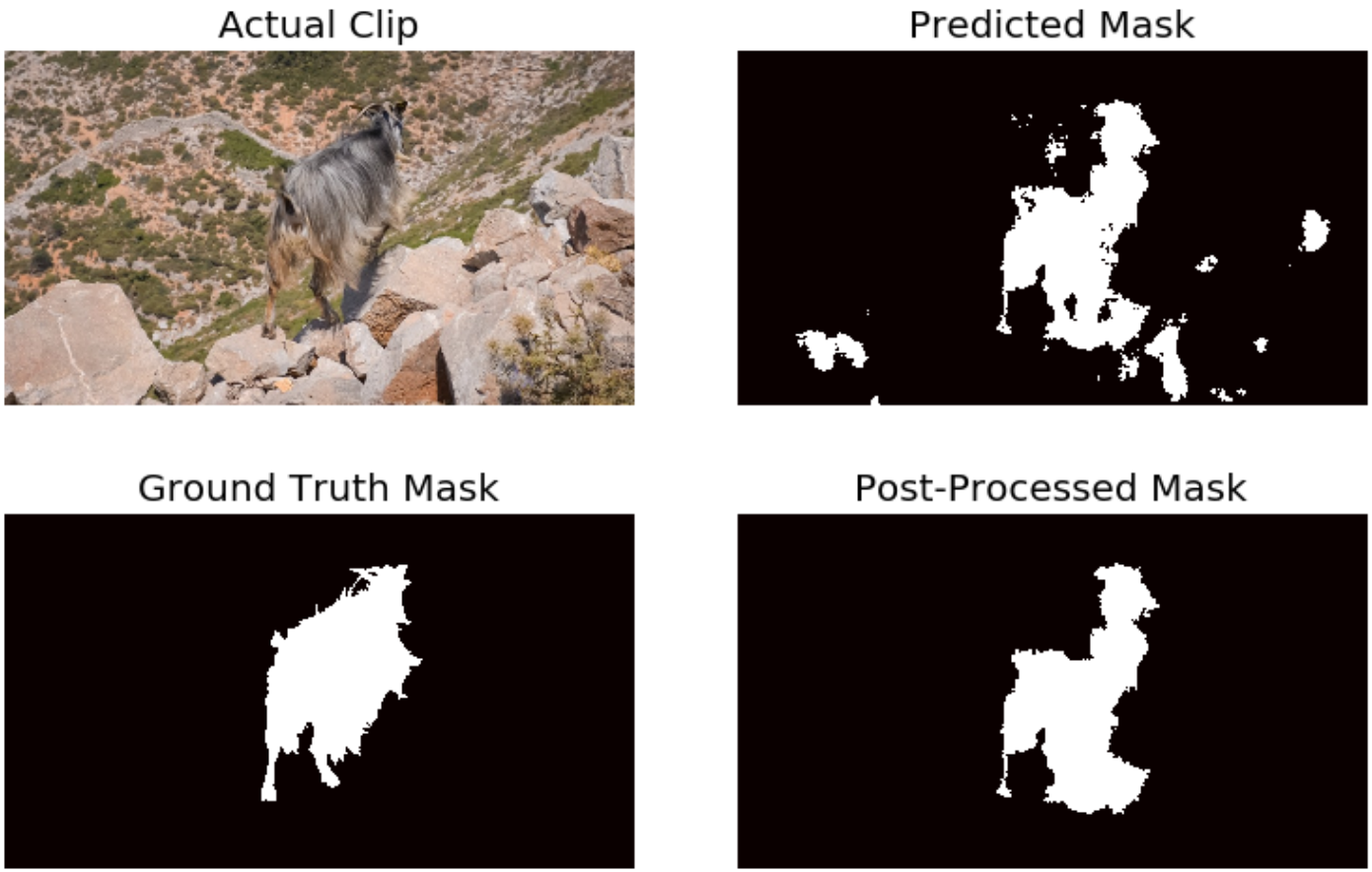
Table 1: DAVIS 2016 mIoU Comparison

Method	FST	CUT	NLC	MSG	KEY	UCF-C3D	CVOS	TRC	SAL
mIoU	0.558	0.552	0.551	0.533	0.498	0.486	0.482	0.473	0.393

Results (1)

Figure 4: GOAT

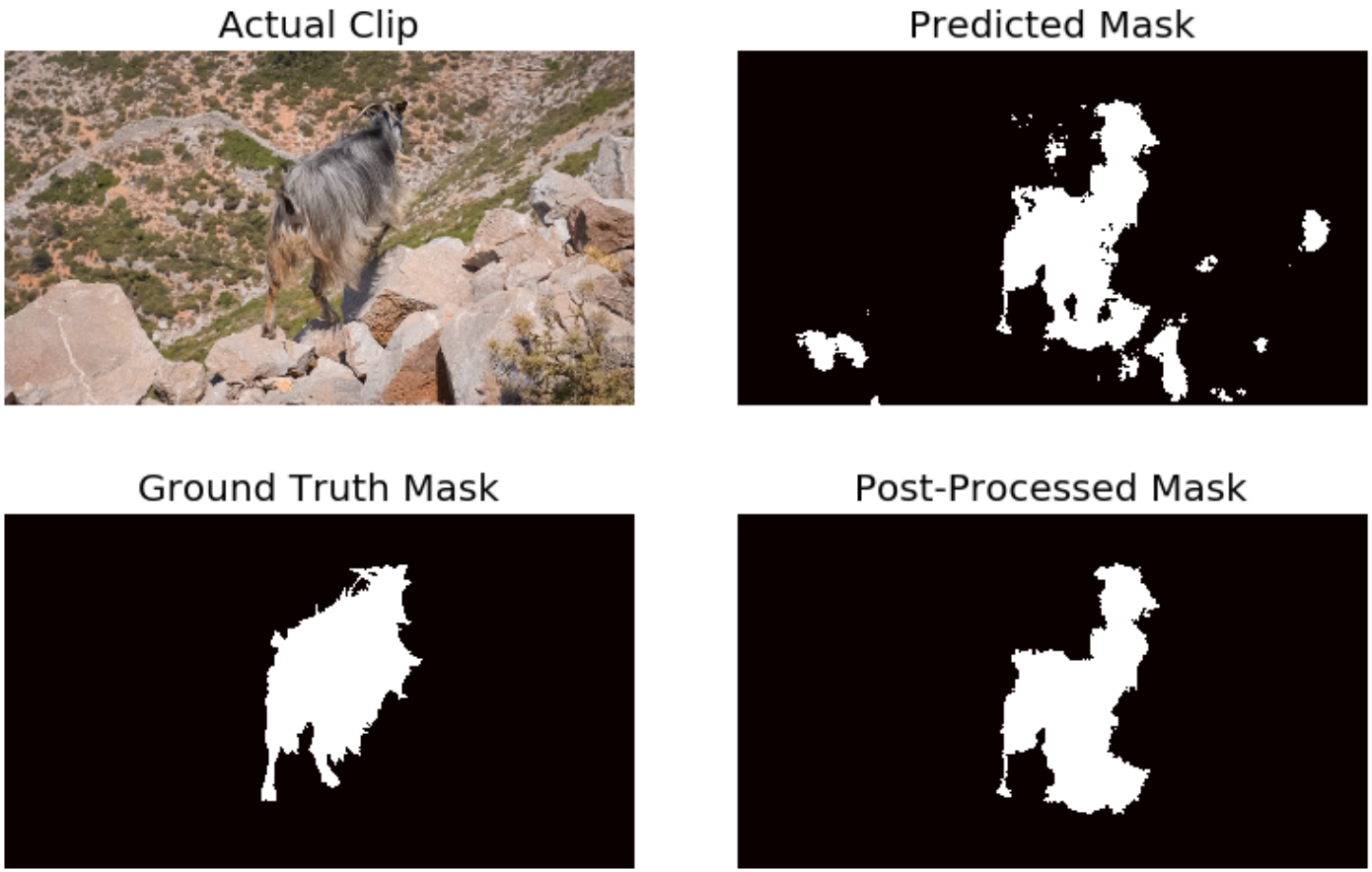
IoU: 0.592888658697



Results (1)

Figure 4: GOAT

IoU: 0.592888658697



Results (2)

Figure 5: MOTOCROSS-JUMP

IoU: 0.598848867416

Actual Clip



Predicted Mask



Ground Truth Mask



Post-Processed Mask



Results (2)

Figure 5: MOTOCROSS-JUMP

IoU: 0.598848867416

Actual Clip



Predicted Mask



Ground Truth Mask



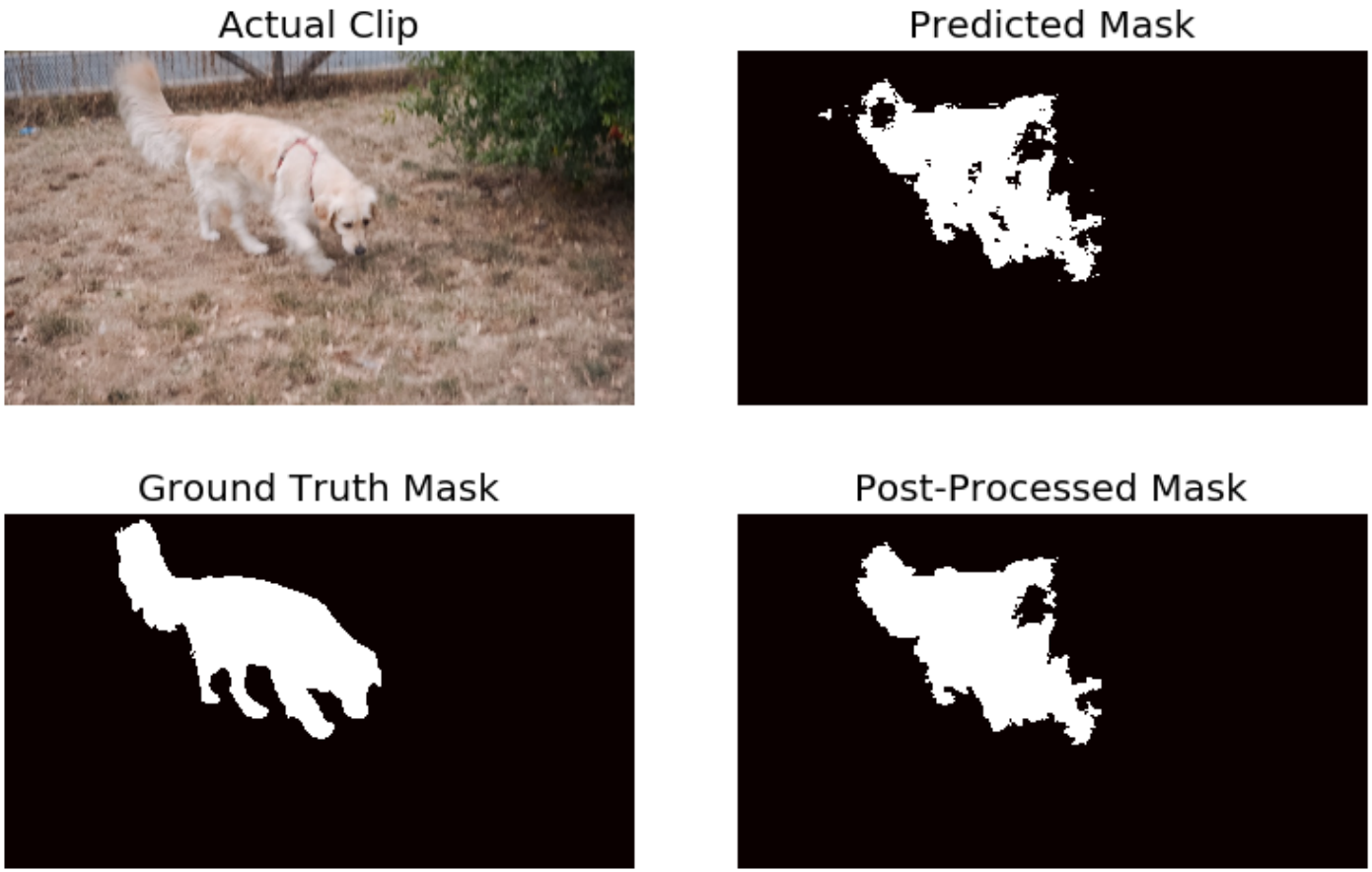
Post-Processed Mask



Results (3)

Figure 6: DOG

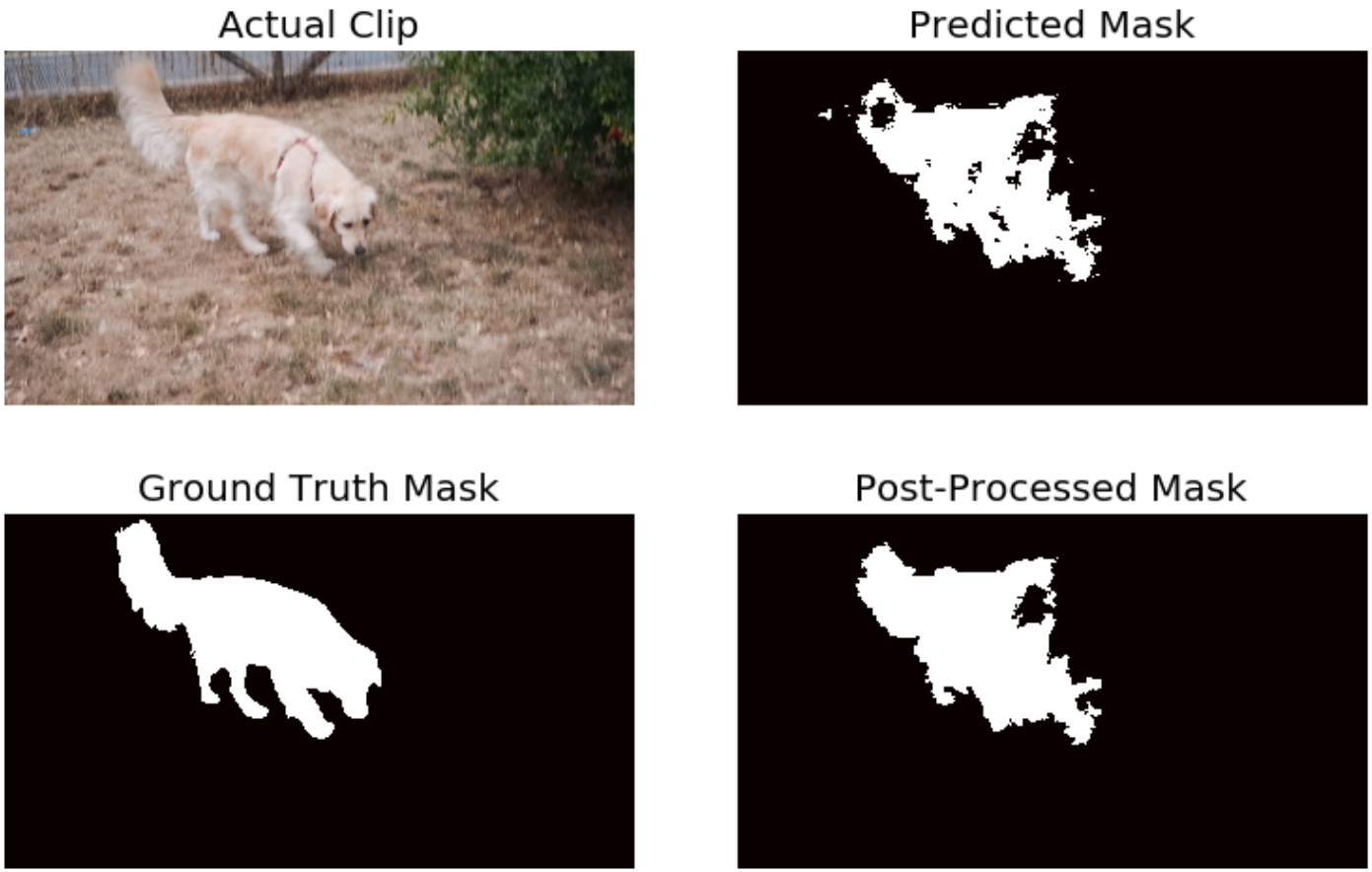
IoU: 0.6454608781



Results (3)

Figure 6: DOG

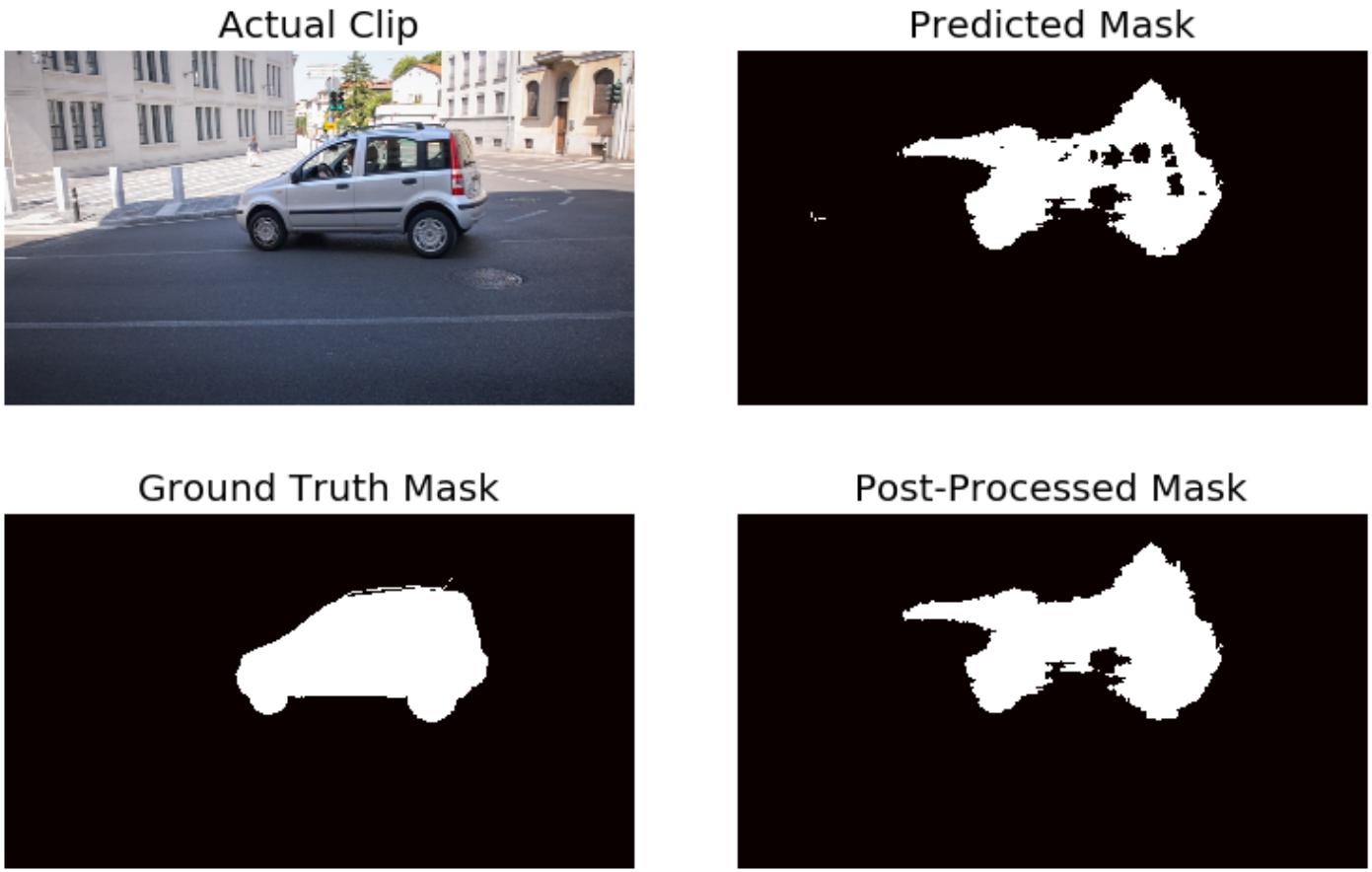
IoU: 0.6454608781



Results (4)

Figure 7: CAR-SHADOW

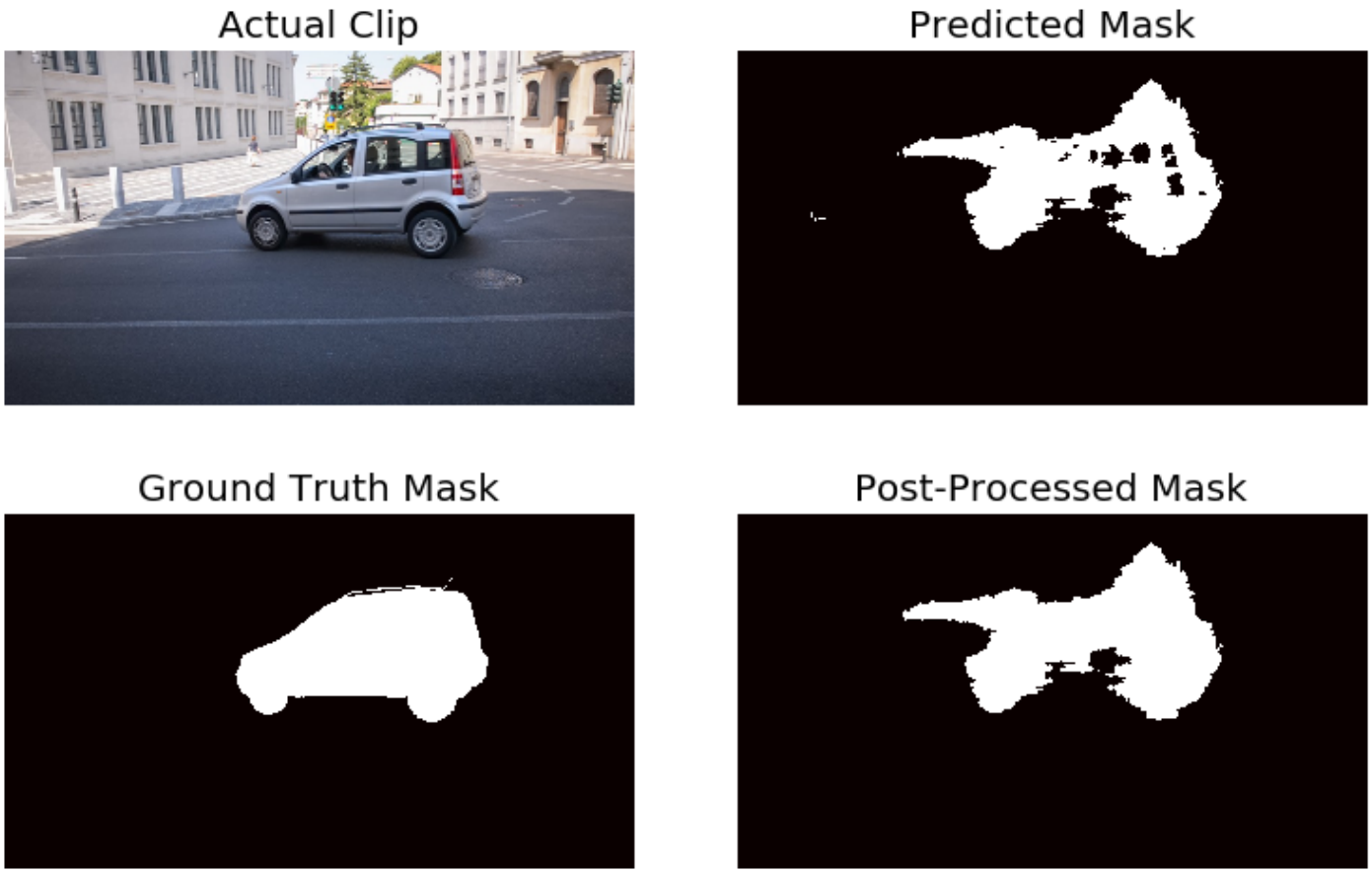
IoU: 0.782939815521



Results (4)

Figure 7: CAR-SHADOW

IoU: 0.782939815521



Future Work

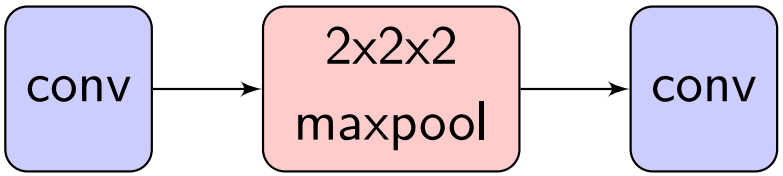
Near Future

- Implementing a variant of the PSPnet module [5] for pooling.
- Current version:

- Proposed change:

Near Future

- Implementing a variant of the PSPnet module [5] for pooling.
- Current version:

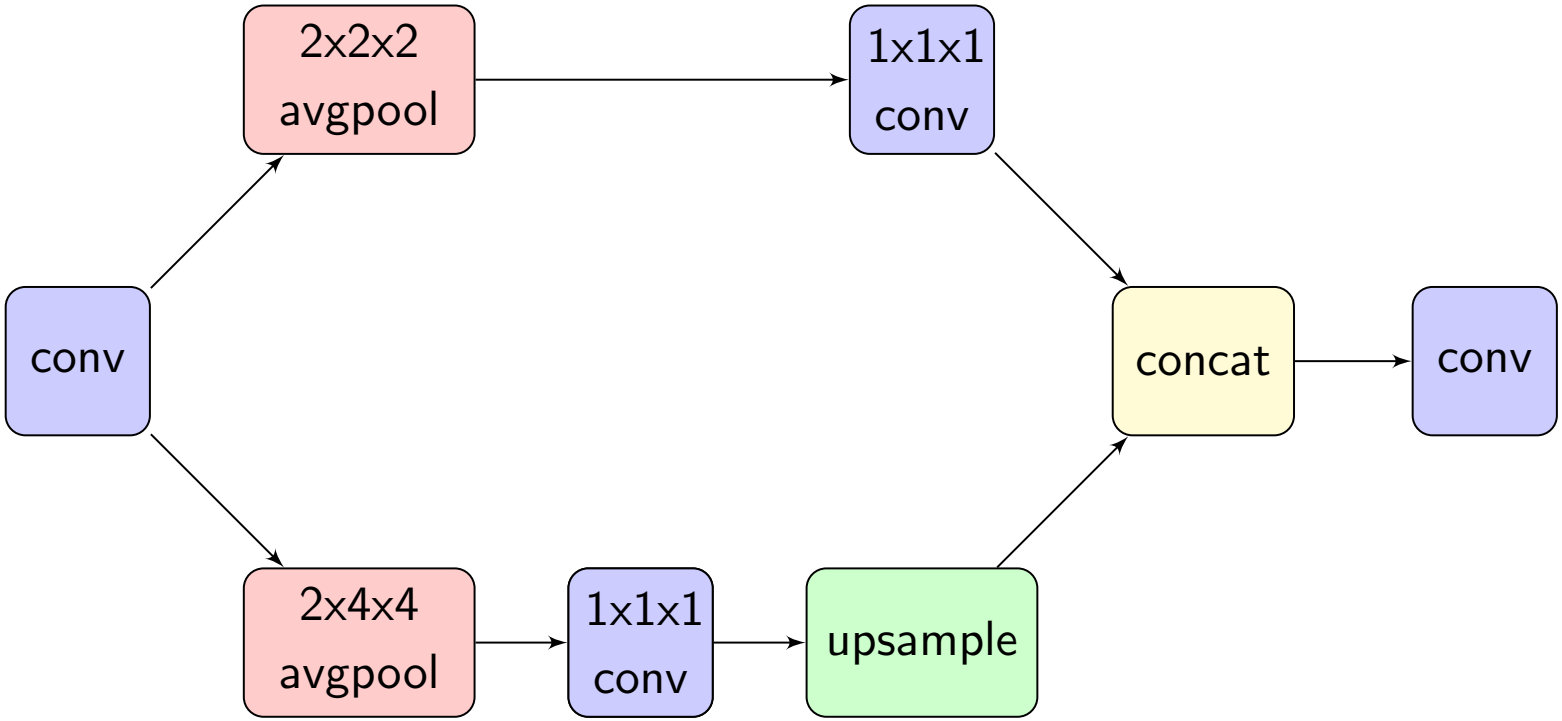


- Proposed change:

Near Future

- Implementing a variant of the PSPnet module [5] for pooling.
- Current version:

- Proposed change:



Distant Future

- Evaluating on two more benchmark datasets:
  1. SegTrack v2
  2. YouTube-Objects Masks Subset

- Changing foreground/background weights in the loss function to reflect the ratio of foreground/background pixels per clip.

Distant Future

- Evaluating on two more benchmark datasets:
  1. SegTrack v2
  2. YouTube-Objects Masks Subset

- Changing foreground/background weights in the loss function to reflect the ratio of foreground/background pixels per clip.

Distant Future

- Evaluating on two more benchmark datasets:
  1. SegTrack v2
  2. YouTube-Objects Masks Subset

- Changing foreground/background weights in the loss function to reflect the ratio of foreground/background pixels per clip.

References I


V. Badrinarayanan, A. Kendall, and R. Cipolla.
Segnet: A deep convolutional encoder-decoder architecture for image segmentation.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.


S. Jain, B. Xiong, and K. Grauman.
Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos.
CVPR, 2017.


F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung.
A benchmark dataset and evaluation methodology for video object segmentation.
In *Computer Vision and Pattern Recognition*, 2016.

References II


D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri.
Learning spatiotemporal features with 3d convolutional networks.
In *ICCV*, 2015.


H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia.
Pyramid scene parsing network.
In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Questions?