DAVIS Challenge 2016 (Unguided)

EECS 542 Final Project

Densely Annotated VIdeo Segmentation (DAVIS)

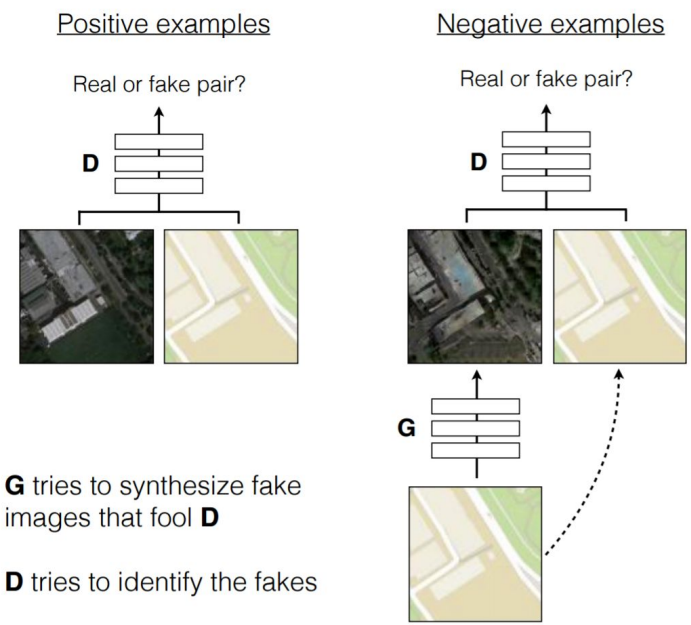
- Background Clutter
- Deformation
- Motion Blur
- Fast Motion
- Low Resolution
- Occlusion
- Out-of-view
- Scale-Variation

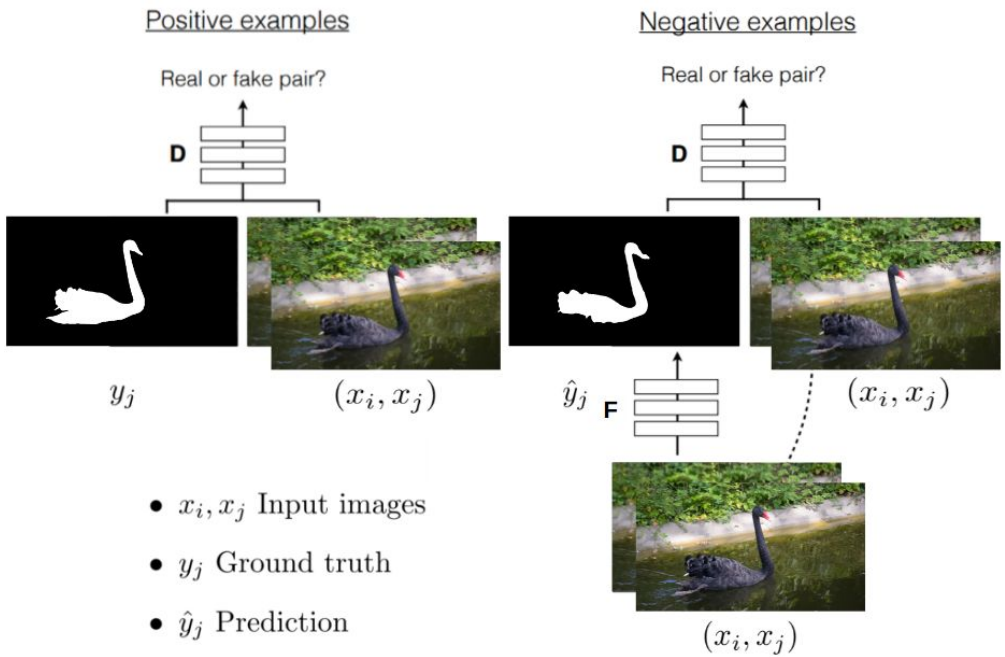
Approaches

Group 1:
Adversarial Dual-Frame FCN

Ming-Yuan Yu S.R.Manikandasriram Kai Jia Yunwen Zhou

Network Architecture





[1] *Image-to-Image Translation with Conditional Adversarial Networks*, Isola, Phillip and Zhu, Jun-Yan and Zhou, Tinghui and Efros, Alexei A

Dual-Frame FCN

$x_i = \text{the } i^{th} \text{ frame}$

$x_j = \text{the } j^{th} \text{ frame}$

$\hat{y}_j = F(x_i, x_j) = \text{prediction}$

$y_j = \text{ground truth}$

function $F(x_i, x_j)$

$\{\text{pool}_{1i}, \ldots, \text{pool}_{5i}\} \leftarrow \text{VGG16}(x_i)$

$\{\text{pool}_{1j}, \ldots, \text{pool}_{5j}\} \leftarrow \text{VGG16}(x_j)$

$e_k = \text{dropout}([\text{pool}_{ki}; \text{pool}_{kj}]) \; \forall k \in \{1, \ldots, 5\}$

$d_5 = \text{upsample}(e_5)$

$d_k = \text{upsample}([d_{k+1}; e_k]) \; \forall k \in \{4, \ldots, 1\}$

$\hat{y}_j = \text{sigmoid}(d_1)$

return $\hat{y}_j$

end function

Losses

$$x_i = \text{the } i^{th} \text{ frame}$$
$$x_j = \text{the } j^{th} \text{ frame}$$
$$\hat{y}_j = F(x_i, x_j) = \text{prediction}$$
$$y_j = \text{ground truth}$$

Cross entropy loss:
$$\mathcal{L}_{cls}(x_i, x_j, y_j) = -\frac{1}{hw}\sum_{k}^{hw} y_{jk}\log(\hat{y}_{jk})$$

Adversarial loss:
$$\mathcal{L}_{adv}(x_i, x_j, y_j) = \log(D(x_i, x_j, y_j)) + \log(1 - D(x_i, x_j, \hat{y}_j))$$

Total loss:
$$\mathcal{L}(x_i, x_j, y_j) = \mathcal{L}_{adv}(x_i, x_j, y_j) + \lambda\mathcal{L}_{cls}(x_i, x_j, y_j)$$

Training details

- Initialized F with pre-trained weights from ImageNet
- ADAM optimizers with fixed learning rate 10^{-5} for 40k iterations

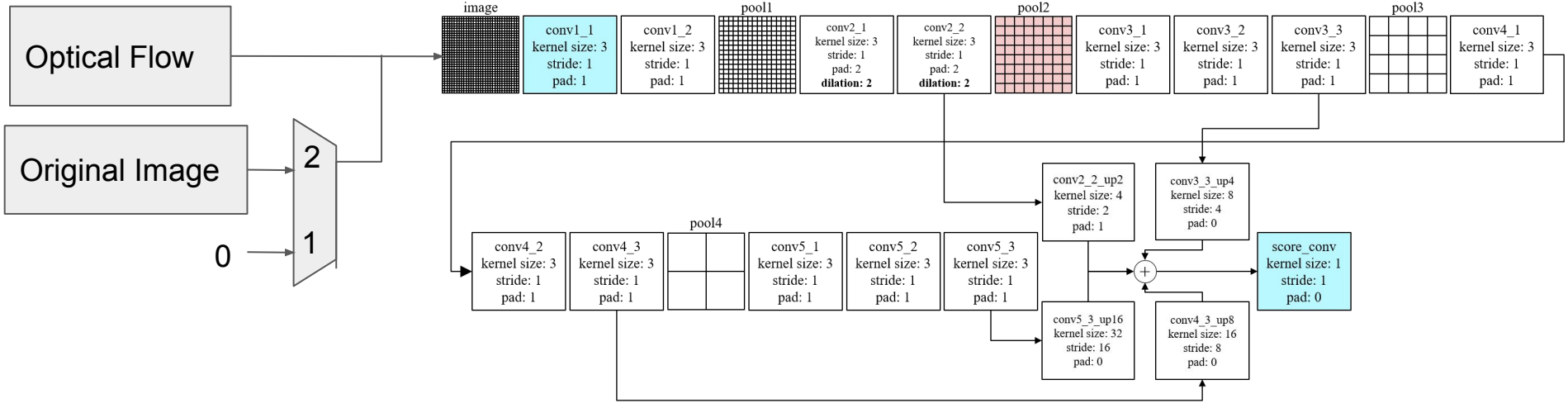
- Tried 4 choices for datasets
  - Case 1: DAVIS 2016
  - Case 2: DAVIS 2016 + reversed pairs
  - Case 3: DAVIS 2017
  - Case 4: DAVIS 2017 + reversed pairs

Group 15:
Optical Flow and Object Segmentation Network

Xiaolin Chen, Ci-Jyun Liang, Lichao Xu, and Fu-Chun Yeh

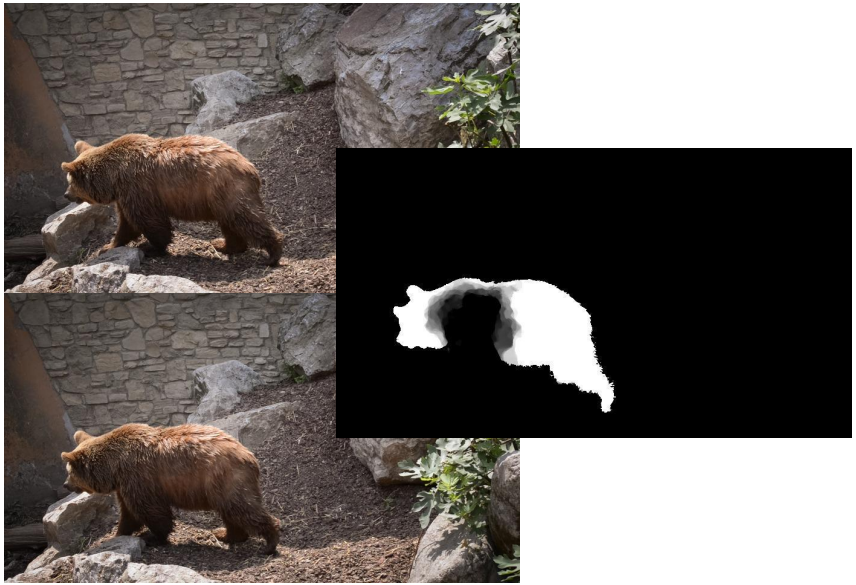
Approach

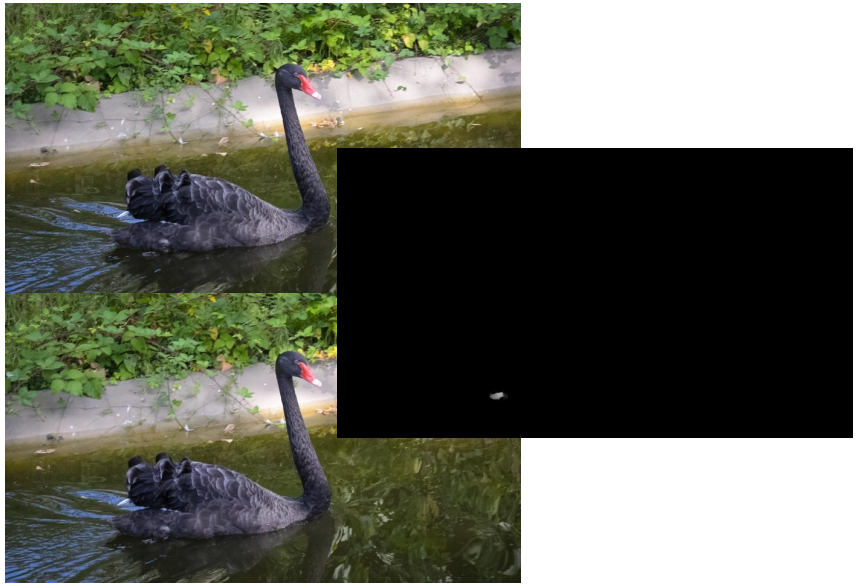
1. Optical Flow Image + Object Segmentation Network
2. Original RGB Image + Optical Flow Image + Object Segmentation Network
3. Tried to use dilated convolution layers to get more global information



Optical Flow

- Applied an optical flow estimation method (Classic+NL) to calculate neighboring two frames' flow.
- The optical flow can get most parts of the object but also fail sometimes.



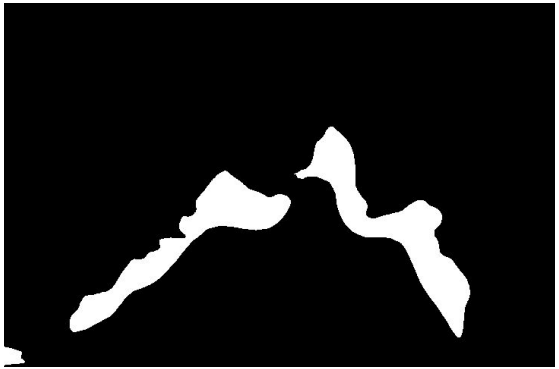


Training

- Using pre-trained model from OSVOS paper (stage 2: separate foreground and background).
- SGD optimizer with learning rate 1E-5 for 100k iterations
- Trained on DAVIS 2016 dataset

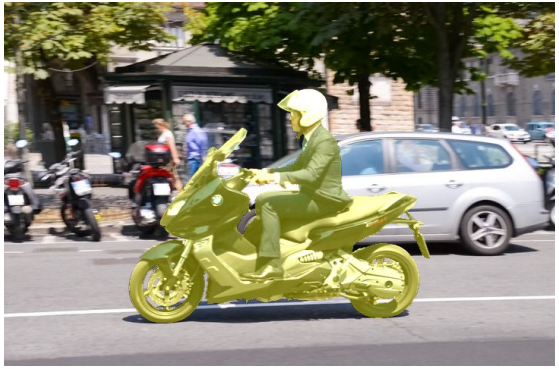
Comparison

Optical flow



Optical flow + OS Network

RGB + Optical flow + OS Network



Ground truth

Group 23:
Two-stream Mask CNN

Yuhao Tang, Haoyu Yu, Hong Moon, Cheng Ouyang

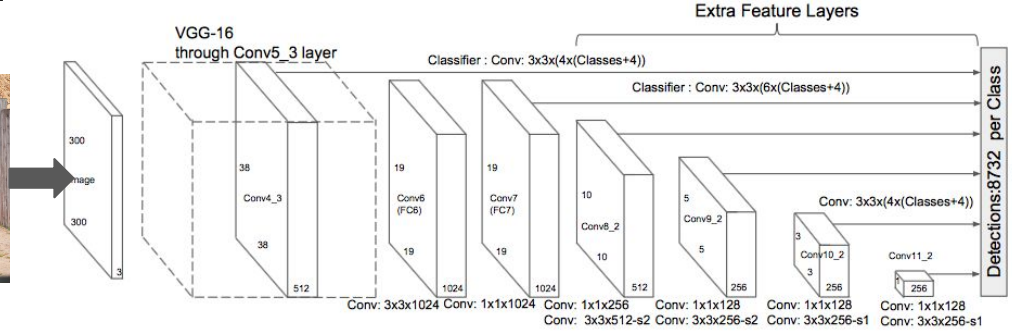
Approach

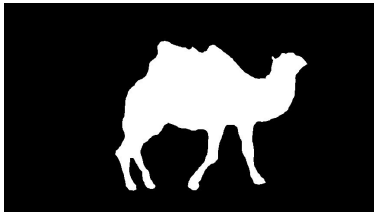
1. Using Single Shot MultiBox Detector(SSD) to generate bounding box and object proposals for each of the frames.

2. Inspecting the whole clip and decide which object classes to track. In each of the frames, simply choose the largest bounding box in the selected classes as our proposed “attention” regions.

3. Use FCN to segment object according to the given regions.

Approach





FCN 8s

[480, 854, 4]

Training Detail

FCN8s training:

We compute ground truth bounding boxes from ground truth segmentations and randomly then resize and translate them. These randomly generated bounding boxes serve as training data together with corresponding RGB images.

We use DAVIS, MSRA10K_Imgs_GT, ECSSD as our training dataset, and trained FCN on K80m GPU about 44 hrs with mini-batchsize 10. Adam Optimizer with 10^-4 (1-10000 itrs), 10^-5(10000-20000 itrs),10^-6(20000-30000 itrs)

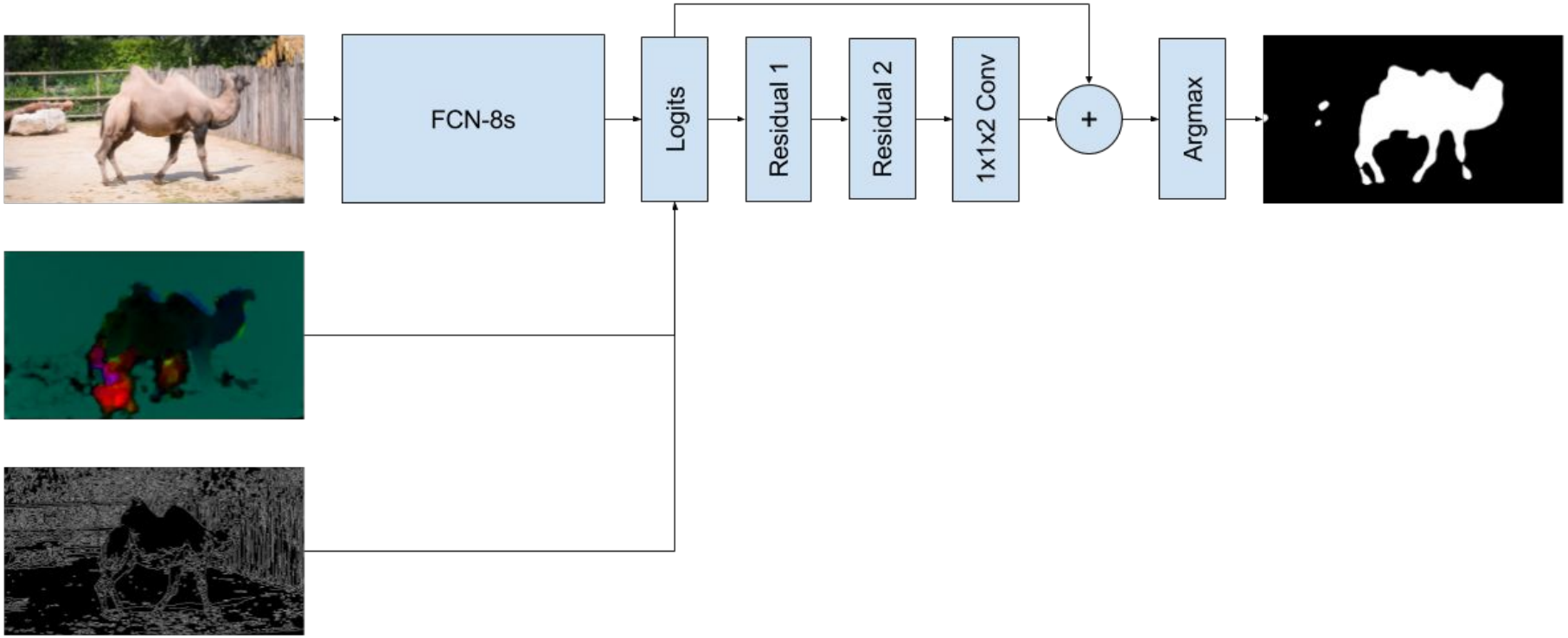
SSD:

We only trained FCN8s, and directly use pretrained SSD model !

Group 17: Foreground/Background FCN with Optical Flow Correction

Derrick Dominic Joshua Mangelson Sudhanva Sreesha Jorge Vilchis

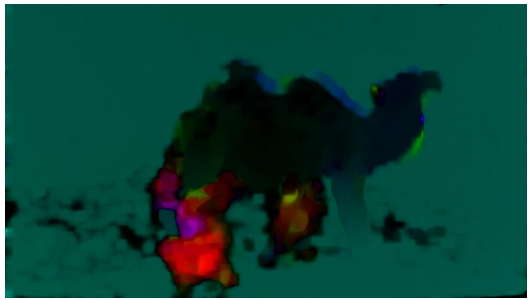
Architecture



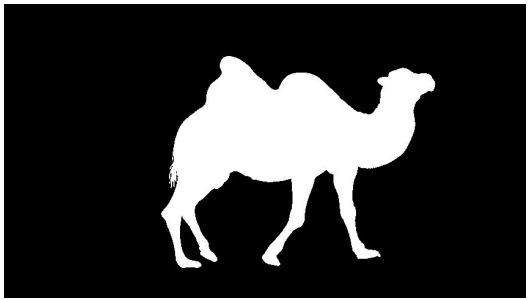
Optical Flow Correction

Previous Frame

Current Frame



Optical Flow

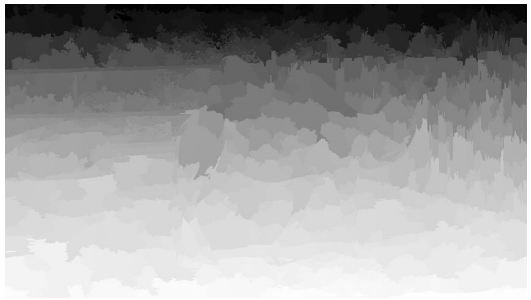


Ground Truth

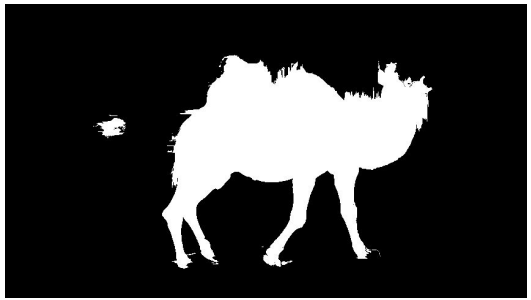
Superpixels

RGB

RGB+Flow

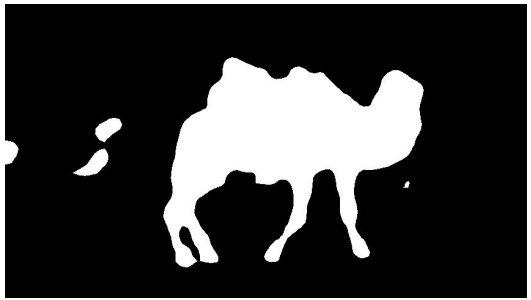


Quickshift



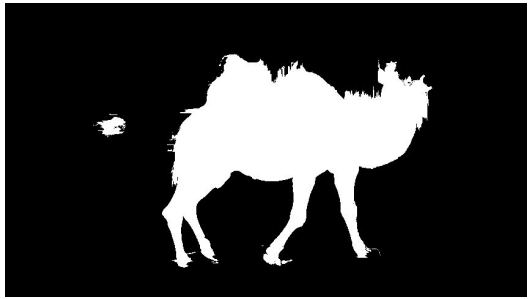
Refined Mask

Boundary Refinement

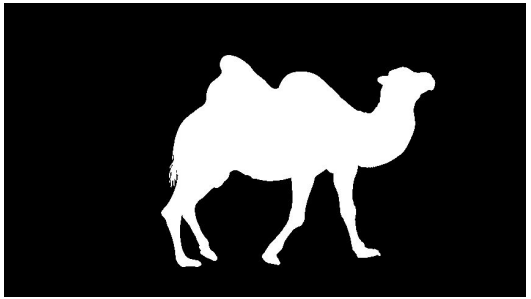


RGB Pre-Training

RGB+Flow



RGB+Flow+Superpixels



Ground Truth

Training

- Pretrain FG/BG on DAVIS to get good initialization
  - Used the Momentum Optimizer with 0.9 momentum
  - 1e-4 learning rate for learning the weights
  - 2e-4 learning rate for learning the biases
  - Ran it for 60 epochs with a batch size of 4
  - Loss saturated after 20 epochs
- Continue training full network including optical flow tail on DAVIS
  - Used the Momentum Optimizer with 0.9 momentum
  - 1e-4 learning rate for learning the weights
  - 2e-4 learning rate for learning the biases
  - Ran it for 11 epochs with a batch size of 4

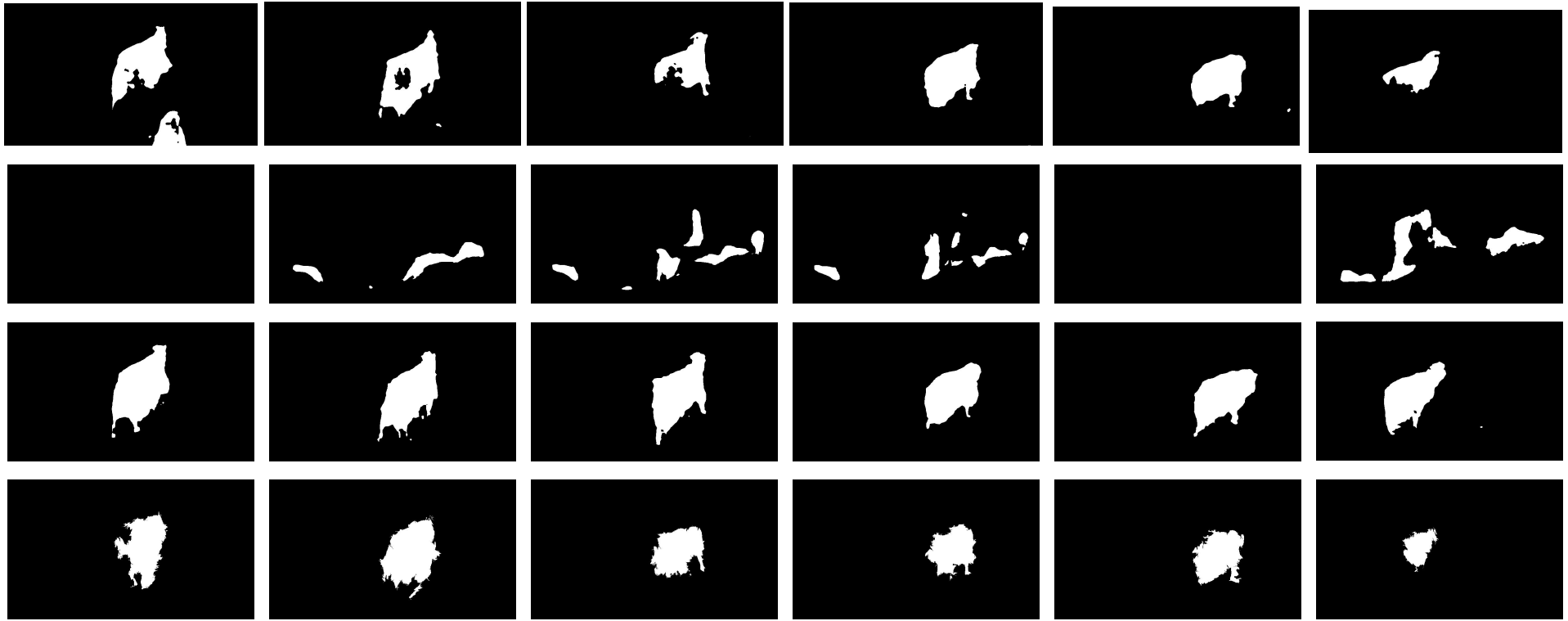
Results

Table of Quantitative Results (2016 Validation)

	Group 1	Group 15	Group 23	Group 17	FST	SAL	KEY	MSG	TRC	CVOS	NLC
J Mean	55.5	30.9	**69.2**	55.2	**55.8**	39.3	49.8	53.3	47.3	48.2	55.1
J Recall	60.2	28.1	**82.4**	60.6	**64.9**	30.0	59.1	61.6	49.3	54.0	55.8
J Decay	**-1.8**	5.3	2.2	-	**0.0**	6.9	14.1	2.4	8.3	10.5	12.6
F Mean	54.6	26.2	**66.7**	49.7	51.1	34.4	42.7	50.8	44.1	44.7	**52.3**
F Recall	60.7	15.5	**78.1**	50.5	51.6	15.4	37.5	**60.0**	43.6	52.6	51.9
F Decay	**-2.0**	5.5	2.0	-	**2.9**	4.3	10.6	5.1	12.9	11.7	11.4

Goat (Ground Truth)

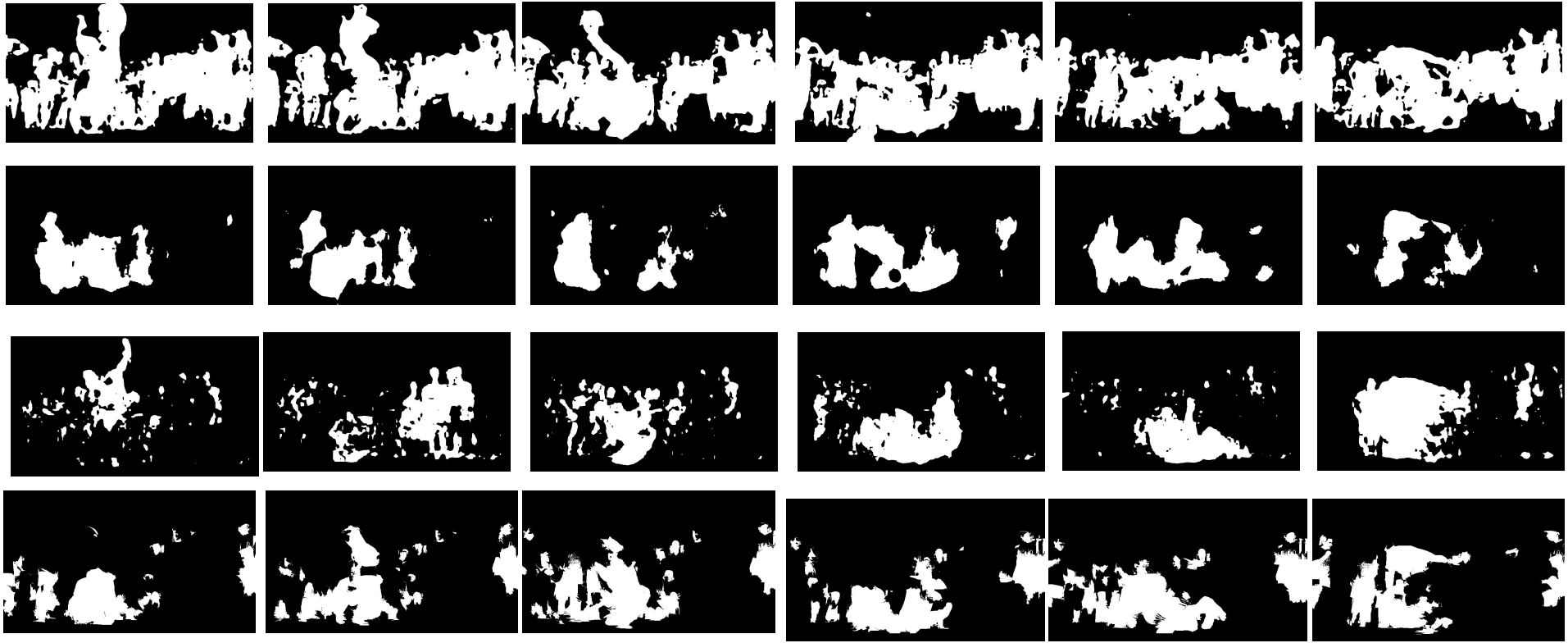
Qualitative Segmentation Examples - Goat



From top to bottom, groups 1, 15, 23, 17.

Breakdance (Ground Truth)

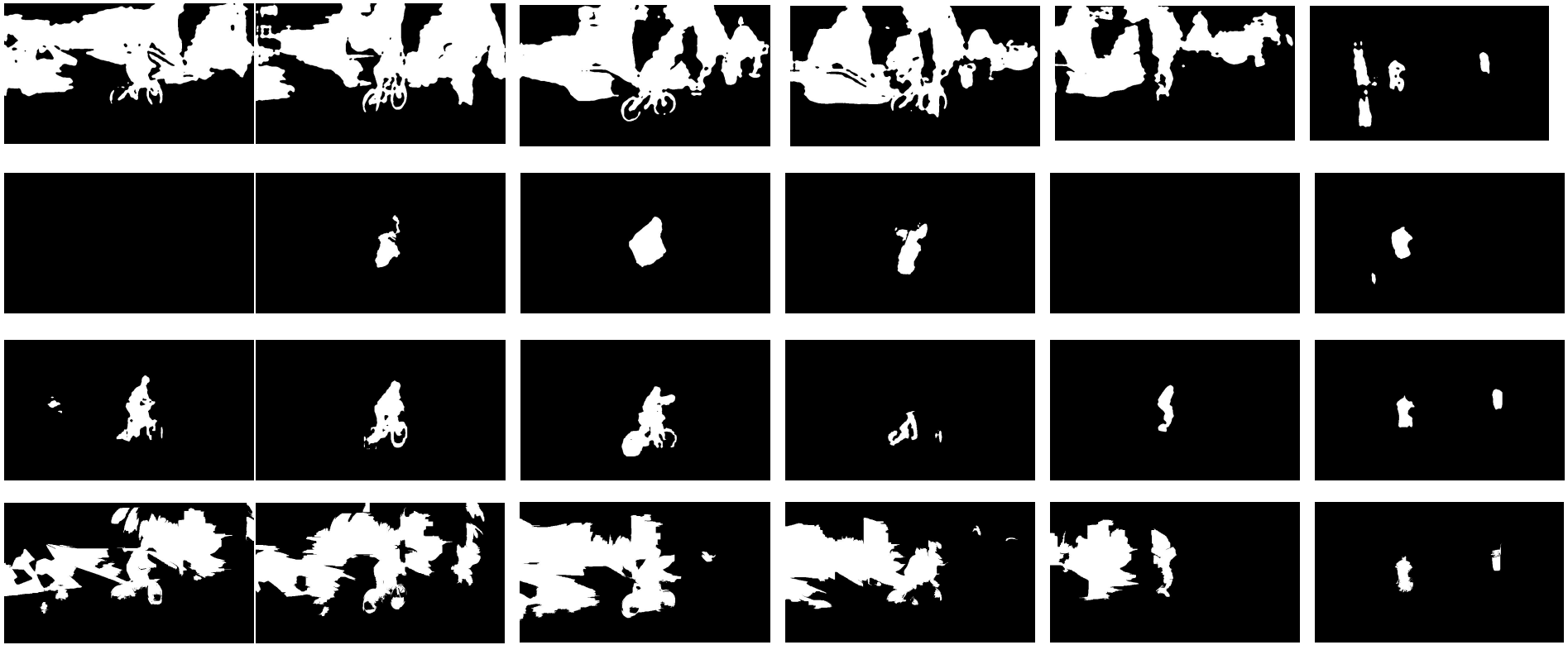
Qualitative Segmentation Examples - Breakdance



From top to bottom, groups 1, 15, 23, 17.

BMX-Trees (Ground Truth)

Qualitative Segmentation Examples - BMX Trees



From top to bottom, groups 1, 15, 23, 17.