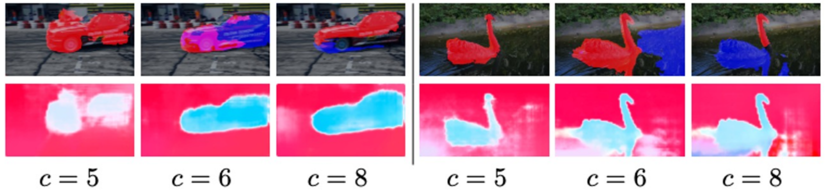
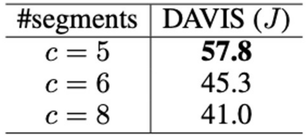
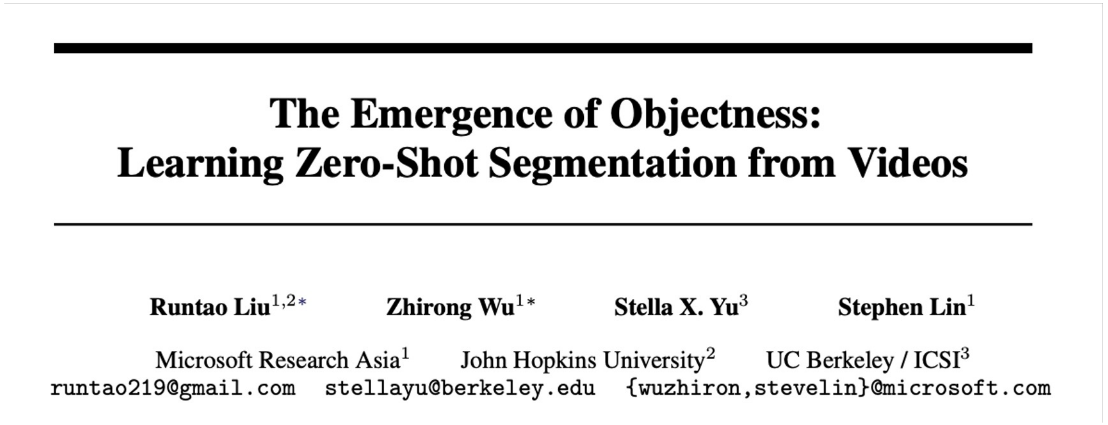




#segments	DAVIS (J)
$c = 5$	**57.8**
$c = 6$	45.3
$c = 8$	41.0



The Emergence of Objectness:
Learning Zero-Shot Segmentation from Videos

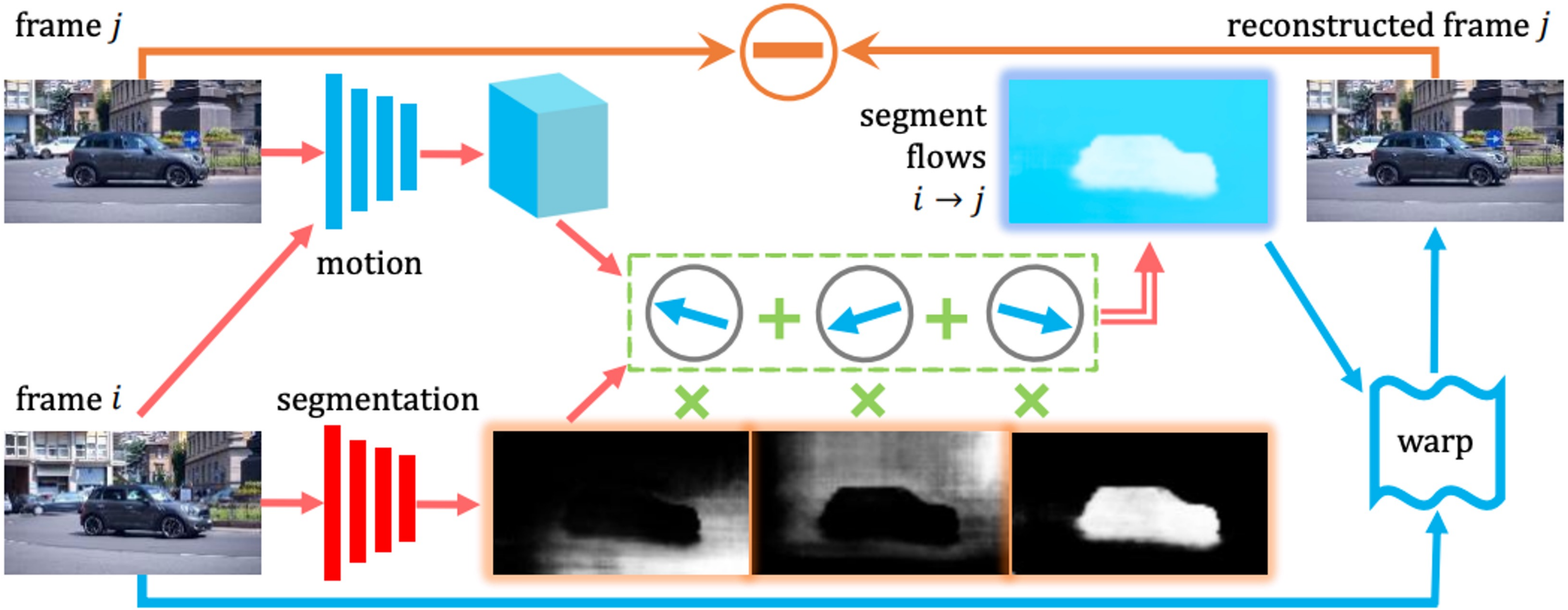
Runtao Liu[1,2*] **Zhirong Wu**[1*] **Stella X. Yu**[3] **Stephen Lin**[1]

Microsoft Research Asia[1] John Hopkins University[2] UC Berkeley / ICSI[3]
runtao219@gmail.com stellayu@berkeley.edu {wuzhiron,stevelin}@microsoft.com

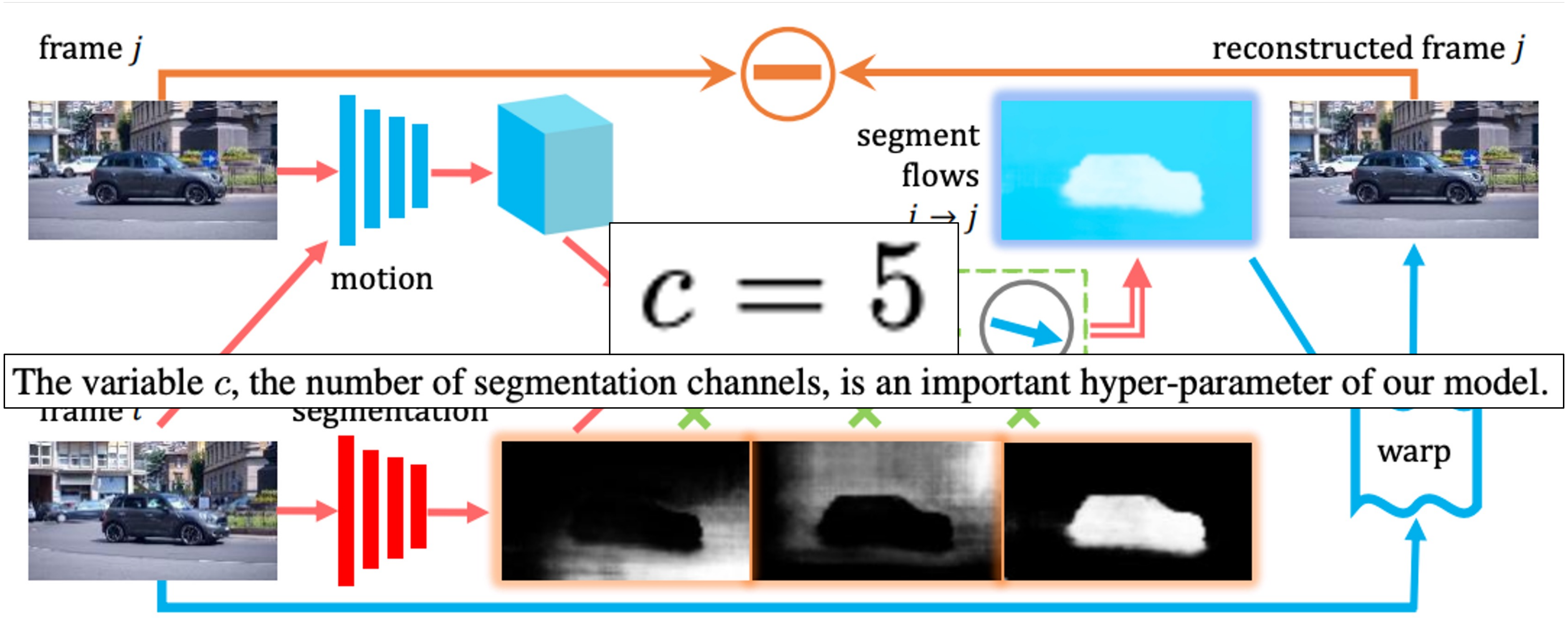
Investigating the Emergence of Objectness

Oliver Wang

Appearance Motion Decomposition (AMD)

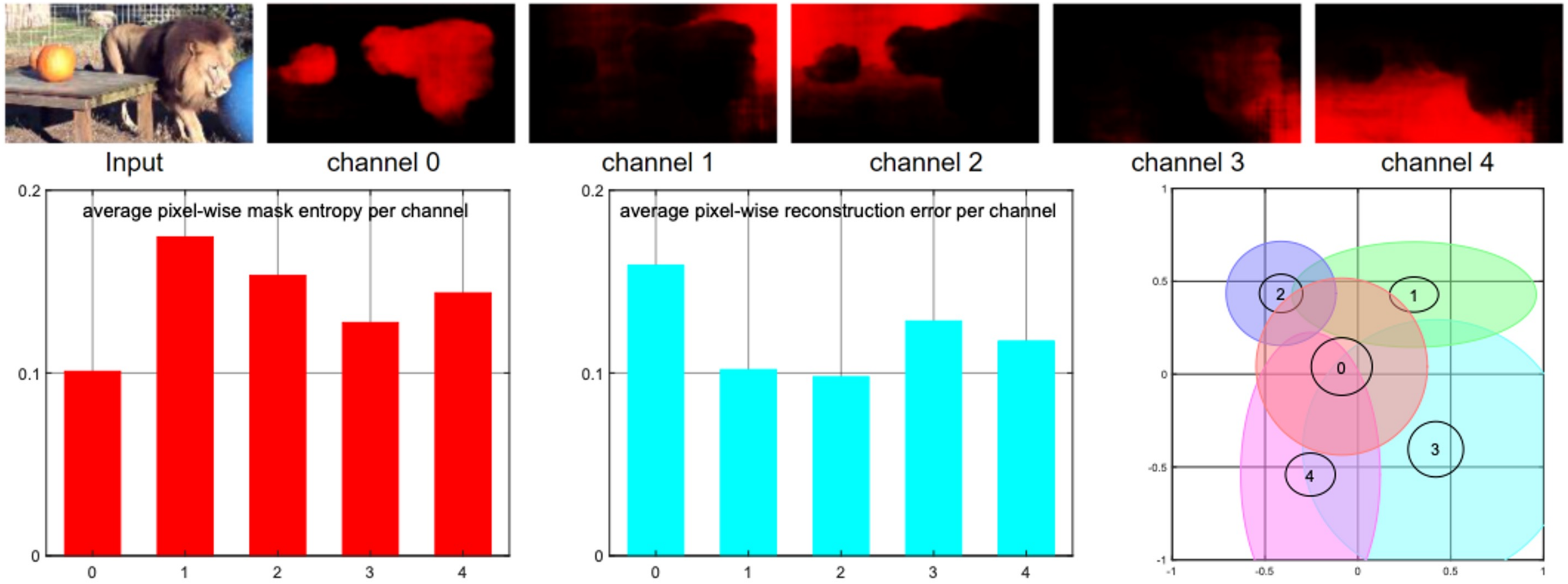


Appearance Motion Decomposition (AMD)



$$c = 5$$

The variable c, the number of segmentation channels, is an important hyper-parameter of our model.



a) segmentation uncertainty **b)** reconstruction error **c)** mask center and spread

Figure 3: **Top)** Our model shows the surprising emergence of objectness in a particular channel. Note that the index of the channel (channel 0 here) could be random in different training runs, but there is always a channel concentrated with objects from all the training videos. **Bottom)** Channel-wise statistics over 17,500 sample training frames of our segmentation network reveal that our objectness channel has **a)** the least segmentation uncertainty (measured by the entropy of S_m), **b)** the largest reconstruction training error (measured by SSIM), and **c)** mostly central locations (the average of the mean and standard deviation of mask centers marked by the channel number and the small black circle) and relatively focused areas (the half of average mask spread shown as the color-shaded disk).

Fore and Back Features

- Is this related to the data itself?
- Are foreground and background distinct in feature space?

- Pass Youtube-VOS dataset through Imagenet pretrained model (separate from AMD)
- Collect pixels in feature space from foreground and background, using annotation mask

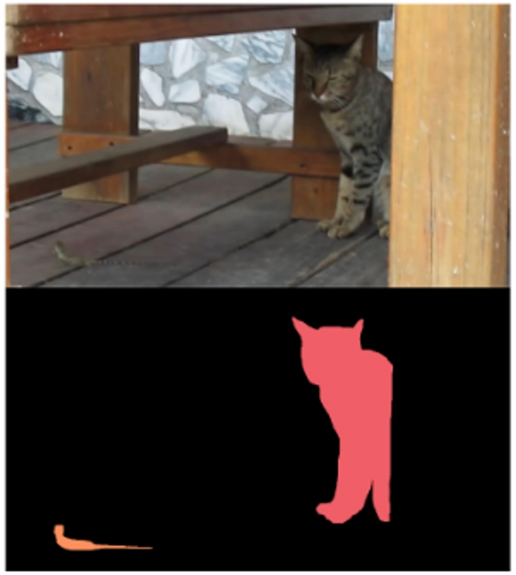
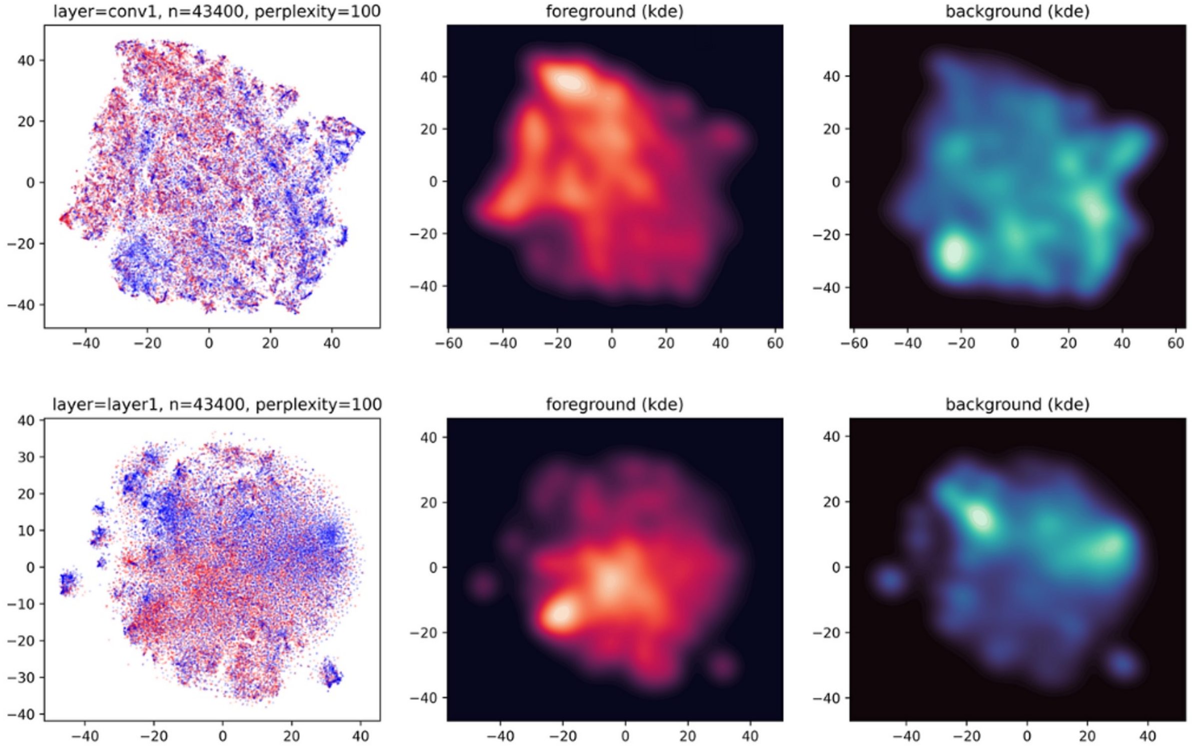
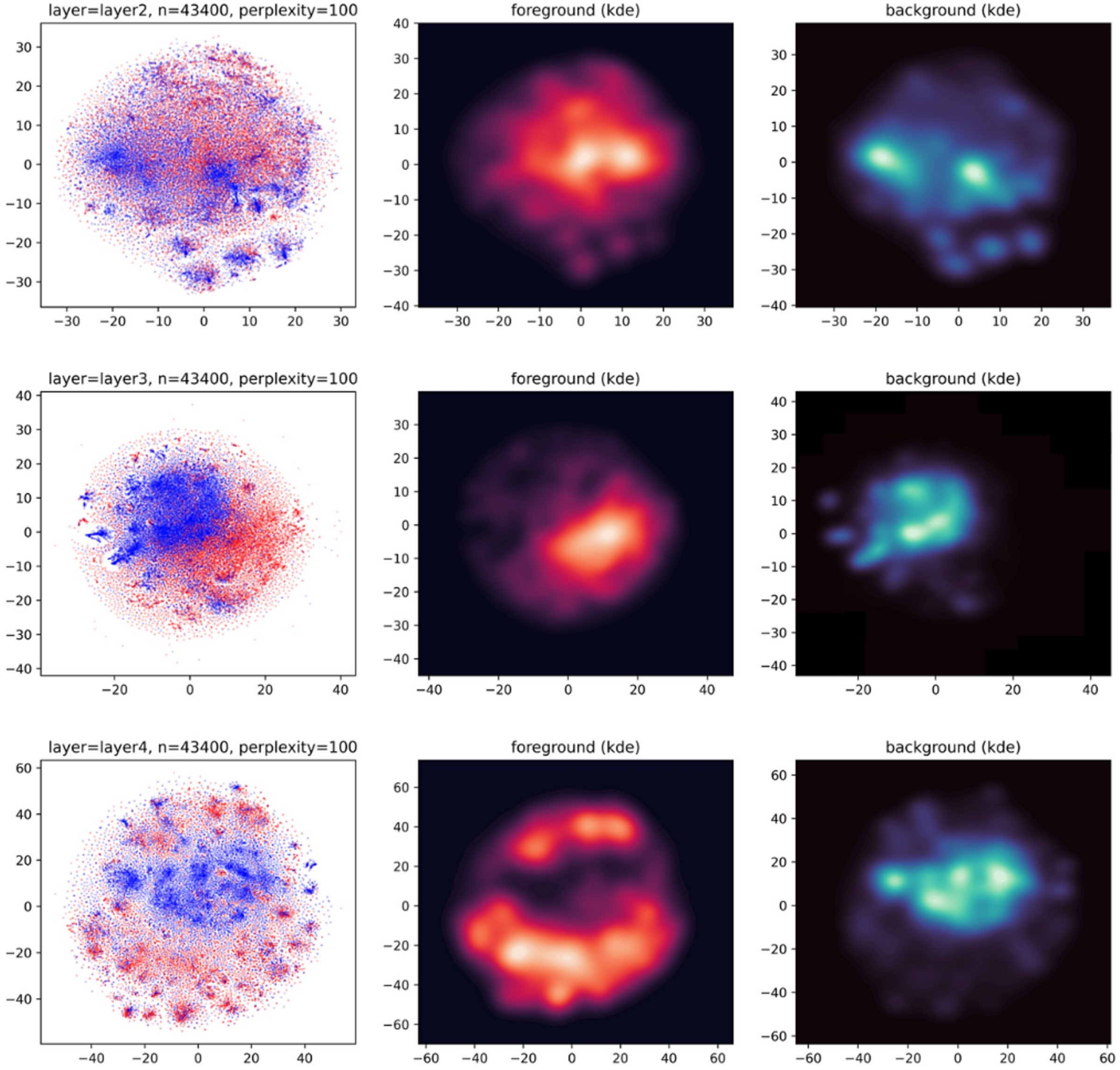


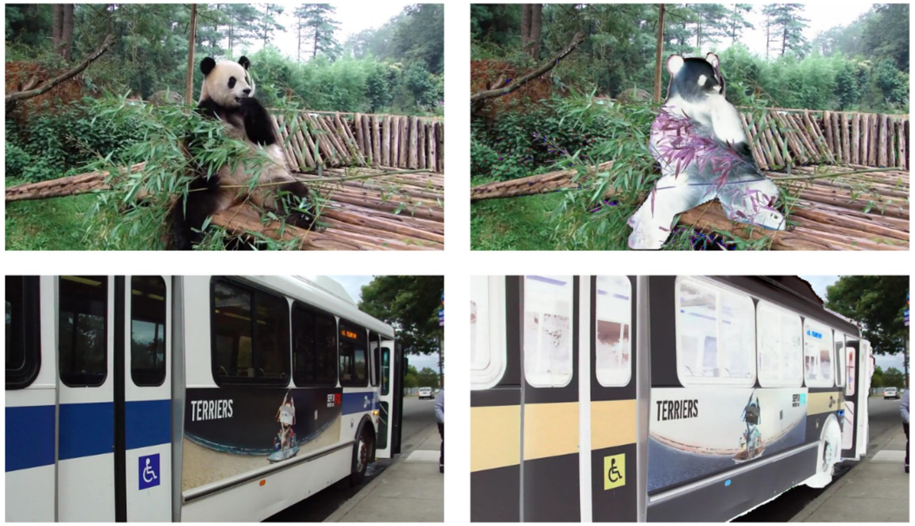
Figure 2. Example foreground mask from the Youtube-VOS dataset.


layer=conv1, n=43400, perplexity=100
foreground (kde)
background (kde)
layer=layer1, n=43400, perplexity=100
foreground (kde)
background (kde)


layer=layer2, n=43400, perplexity=100
foreground (kde)
background (kde)
layer=layer3, n=43400, perplexity=100
foreground (kde)
background (kde)
layer=layer4, n=43400, perplexity=100
foreground (kde)
background (kde)

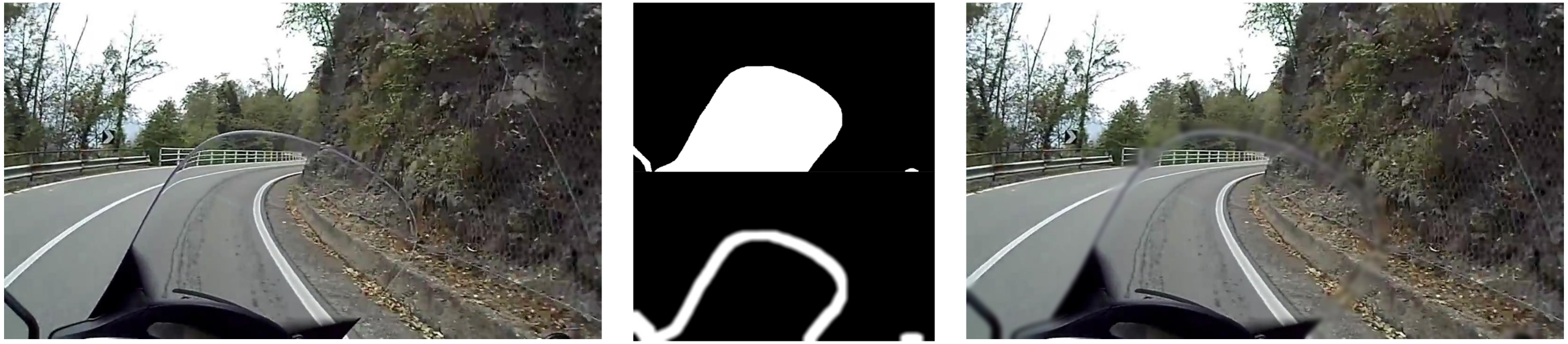
Foreground Color Inversion

- Since segmentation is appearance based, what happens if we change the appearance of the main objects?
  - Would expect the segmentation network to be strongly disrupted
  - AMD model trained only with resize, crop, flip transformations, no color
- Out of distribution data

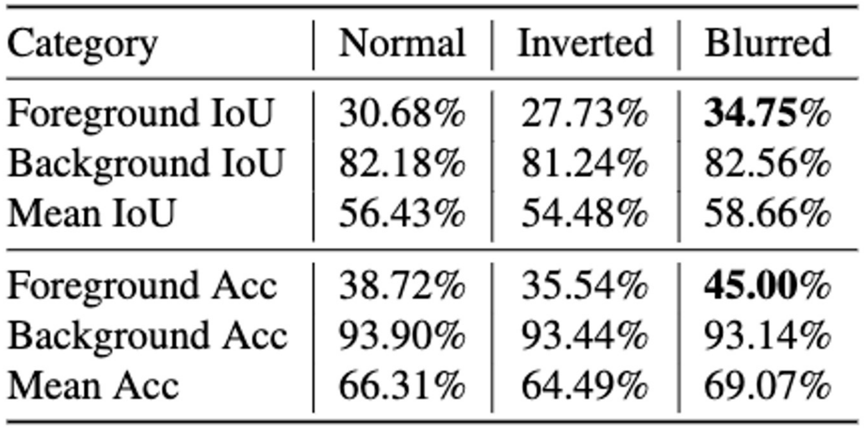


Mask Edge Blur

- Does the appearance of motion lead to better segmentations, since the model is trained to predict segment flows?
  - Human added motion cues
  - AMD’s training objective is reconstruction through 1) predicting the correct segments to move (segmentation), and 2) how they will move (optical flow)



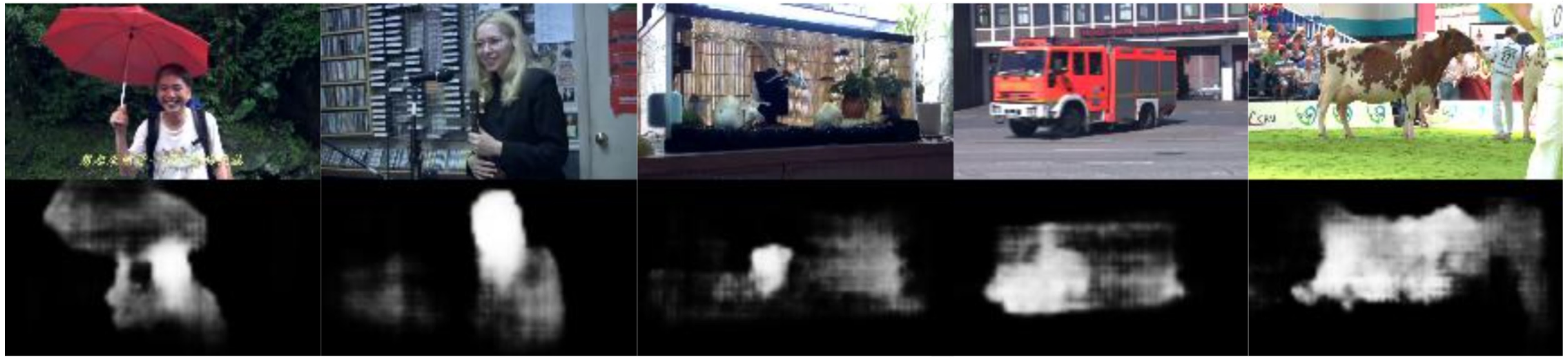
Results



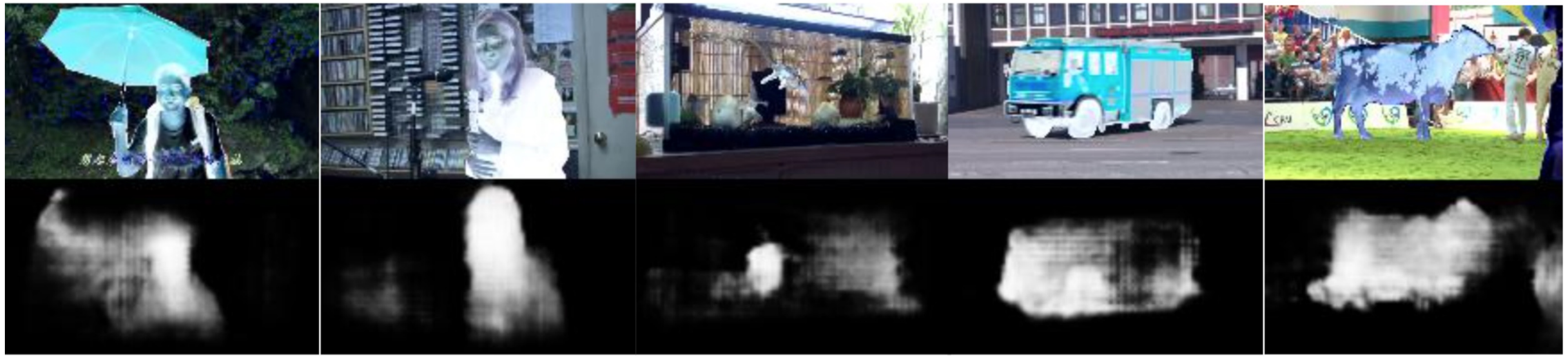
Category	Normal	Inverted	Blurred
Foreground IoU	30.68%	27.73%	**34.75%**
Background IoU	82.18%	81.24%	82.56%
Mean IoU	56.43%	54.48%	58.66%
Foreground Acc	38.72%	35.54%	**45.00%**
Background Acc	93.90%	93.44%	93.14%
Mean Acc	66.31%	64.49%	69.07%

Table 1. Results: Intersection over Union scores and Intersection over Ground Truth (Accuracy) scores for the unmodified Youtube-VOS validation set, inverted foreground colors set, and the blurred edges set. Inference on the pretrained AMD model. Note that the edge blurred set achieves 4% higher foreground IoU compared to the original dataset.

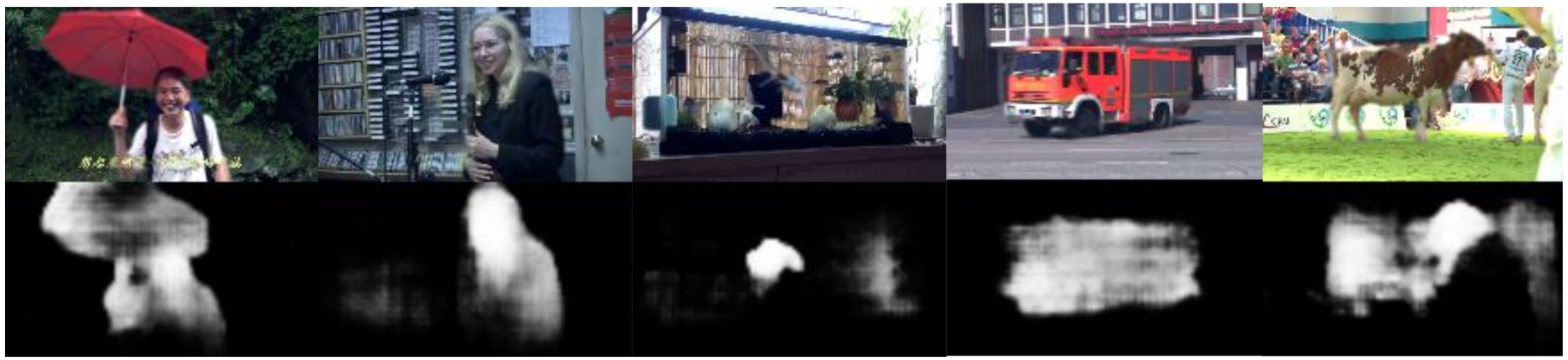
Unmodified



Inverted Foreground



Blurred Edges



Questions?