BubbleNets:
Learning to Select the Guidance Frame in Video Object Segmentation by Deep Sorting Frames

Roger Lee, Yu-Si Hsu, Yun-Hsin Kuo

Video Object Segmentation (VOS)

A task to learn where the objects are in the video in pixel level

– Unsupervised Video Object Segmentation

– Semi-supervised Video Object Segmentation

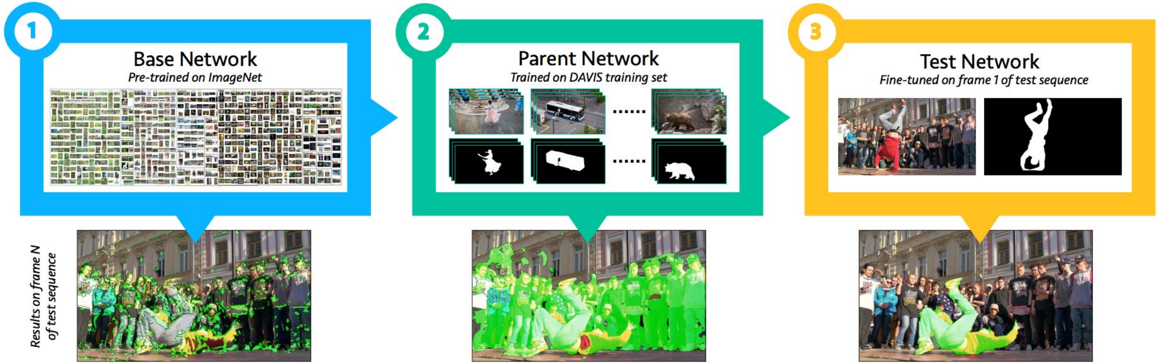
– Supervised Video Object Segmentation

Semi-Supervised Video Object Segmentation

Takes only one frame with annotation from the video, and segment the objects for the rest of the video.

One Shot Video Object Segmentation (OSVOS)

- OSVOS is a semi-supervised video object segmentation model
  - Pre-train a model to classify each pixel to be either foreground or background
  - Fine-tune the model with the first frame (ground truth annotation)
  - Output the object segmentation for the rest of frames
- Uses no temporal information
  - i.e. the order of frames fed into the network does not influence the outcome



The Order Doesn't Matter!

- Therefore we are allowed to pick any frame being used to fine-tune the model
- OSVOS fine-tuned the model with the first frame
  - Is first frame the best choice to fine-tune the model?

Contributions

- The very first paper that discusses about frame selection for semi-supervised VOS.
  - Motivated by the high cost of densely-annotated user segmentations
- Demonstrates BubbleNets to improve VOS performance

BubbleNets

- Segmentation performance **varies** when selecting an alternative frame.
- Select one single frame for user annotation, from an **untouched** video.
- Improve the performance of semi-supervised VOS. OSVOS, specifically.

BubbleNets: I/O selection

To increase training examples.
Because labeled video data are expensive.

m training videos. n labeled frames per video.

Training examples	$\boldsymbol{m}$	$\boldsymbol{m}{*}n$	$m \times \binom{n}{2} \approx \frac{mn^2}{2}$
Input	entire video	individual frame	two compared frames
Output	$\arg\max_i \ y_i$	predicted performance	predicted **relative** performance

BubbleNets: Input

k reference frames as additional input. Provide some video context.

INPUT: 2 compared frames (i, j) + k reference frames

$k = 3$

Compromise between video-wide awareness and network complexity

Training examples: $m \times \binom{n}{2} \approx \frac{mn^2}{2} \quad \Rightarrow \quad m \times \binom{n}{k+2} \approx \frac{mn^{(k+2)}}{k+2}$

BubbleNets: Output & Loss function

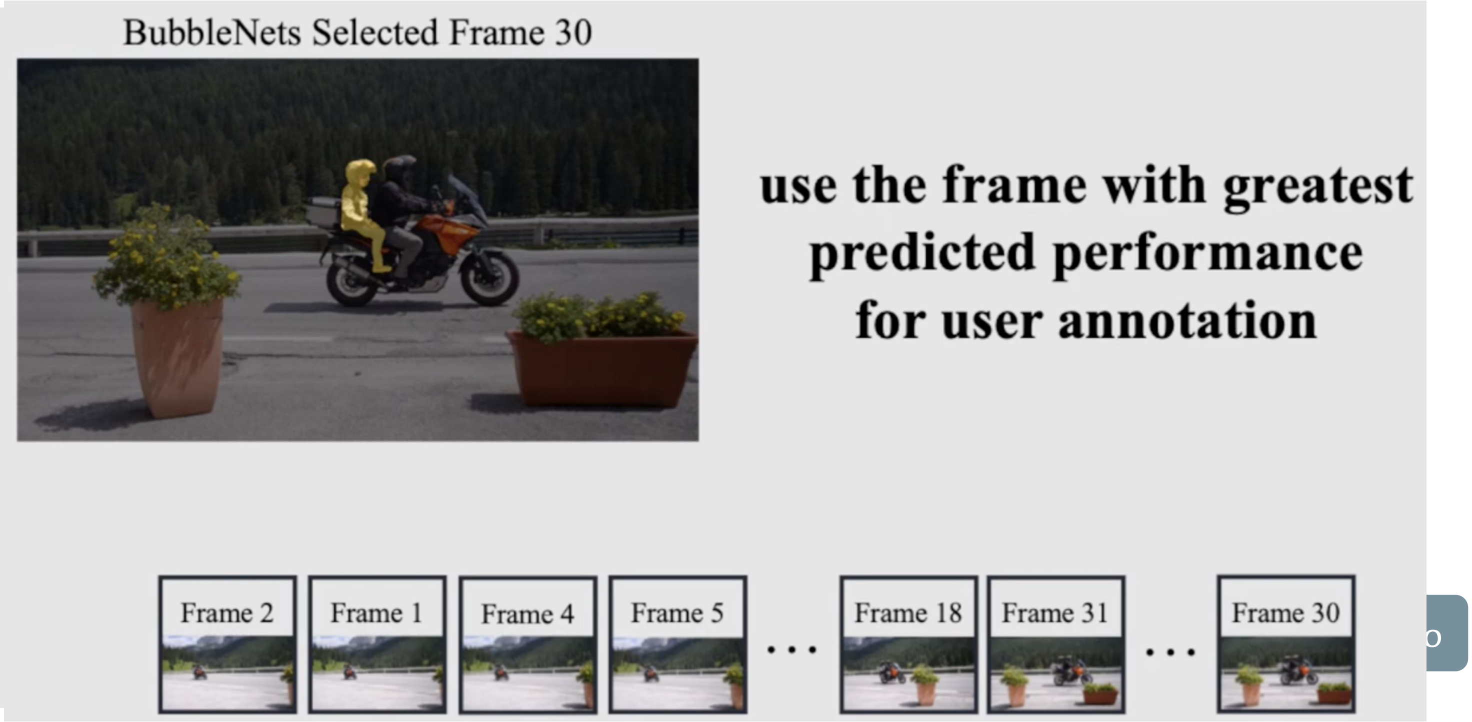
set of reference frames

OUTPUT: predicted **relative** performance, $f(x_i, x_j, X_{\text{ref.}}, \mathbf{W})$

2 compared frames

$$\mathcal{L}(\mathbf{W}) := |(y_i - y_j) - f(x_i, x_j, X_{\text{ref.}}, \mathbf{W})|$$

BubbleNets: Basic sorting framework



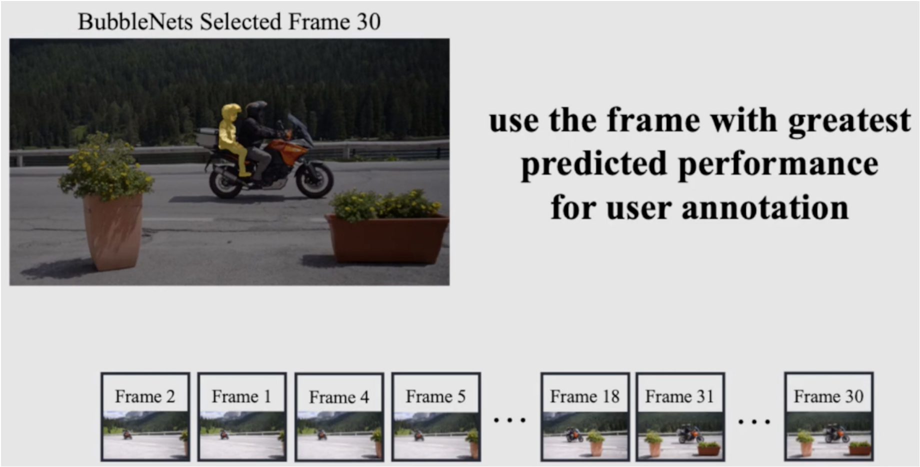
Source: https://youtu.be/XBEMuFVC2lg

BubbleNets: Basic sorting framework

Seems to be deterministic, and only one pass is needed.

But BubbleNets is **stochastic**.
Different set of reference frames results in different relative performance.

How to increase BubbleNets' consistency?



BubbleNets: Consistency

(1) Bubble sort feature: redundancy, thus sub-optimal

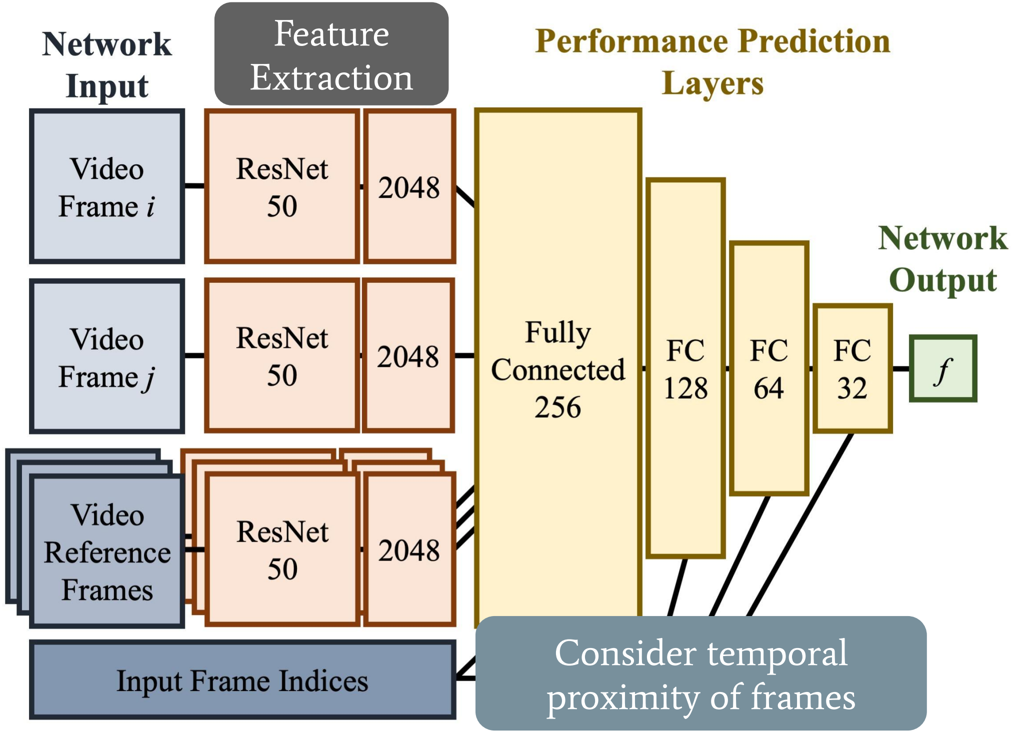
→ *n*-forward passes for an *n*-frame video.

Effective given BN's stochastic nature

(2) Batch each prediction over multiple sets of reference frames.

→ Summation over entire batch. Reduce variability.

Increasing batch size:
[1] more easily to hit local minimum
[2] longer execution time



- Nomarlized Frame Indices:

$$I_i = \frac{i}{n}.$$

- Leaky ReLU as activation function
- 20% dropout in last three layers
- Output is a **scalar**

BubbleNets: Implementation

- Training dataset: DAVIS 2017
  - Complete set of fully-annotated video frames
  - i.e., every frame has ground truth
- Performance indicator: Region similarity + Contour accuracy
  - Video-wide mean performance by selecting frame *i* for annotation (ground truth performance)

Region similarity
a.k.a. Jaccard Idx, IoU

Contour accuracy
a.k.a. F_1 score

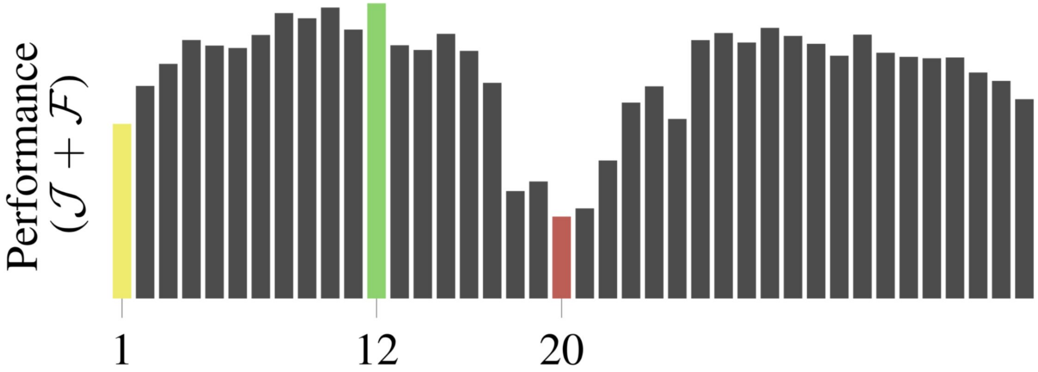
the bigger the better

$$y_i := \frac{1}{n} \sum_{j=1}^{n} \mathcal{J}_j + \mathcal{F}_j$$

n frames in each video

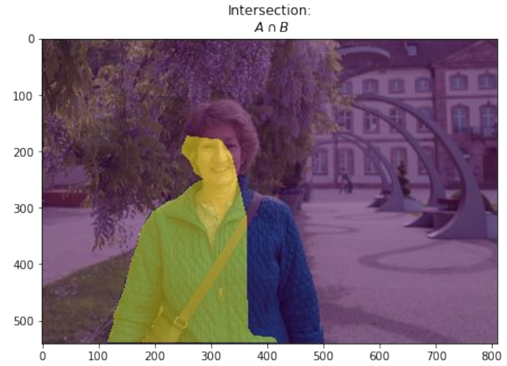
BubbleNets: Training labels

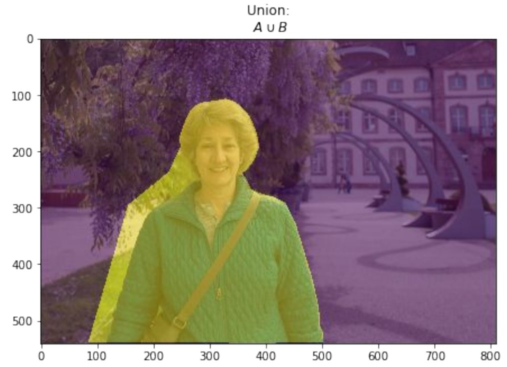
- Pre-calculation for each frame's ground truth performance
  - We will know the best single frame, as well as the worst one.
  - Simle frame selection strategies (First, Middle, Last)
  - Decrease training time



Region Similarity J (IoU, Jaccard Idx)

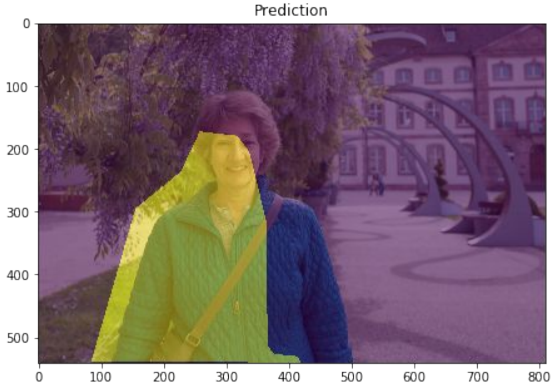
$$\mathcal{J} = \frac{M \cap G}{M \cup G}$$





M : foreground mask (prediction)

G : ground truth



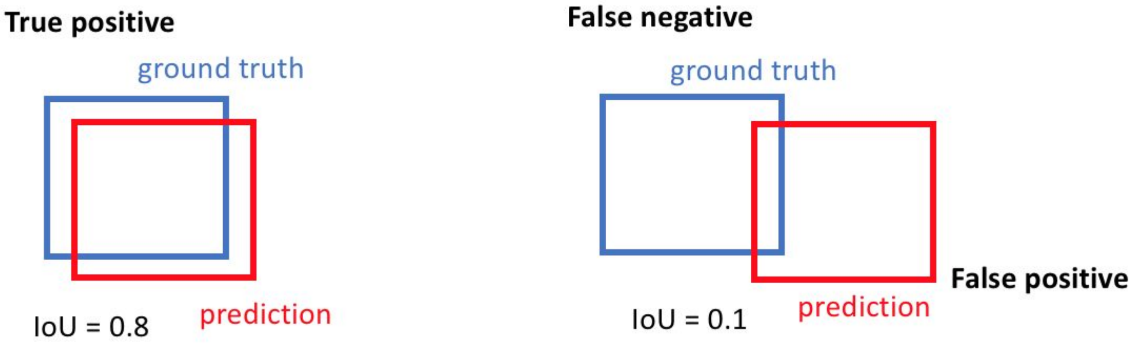
Source

Contour Accuracy F (F_1 score)

$$F_1 = \left(\frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}\right) = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Example
Threshold: 0.5

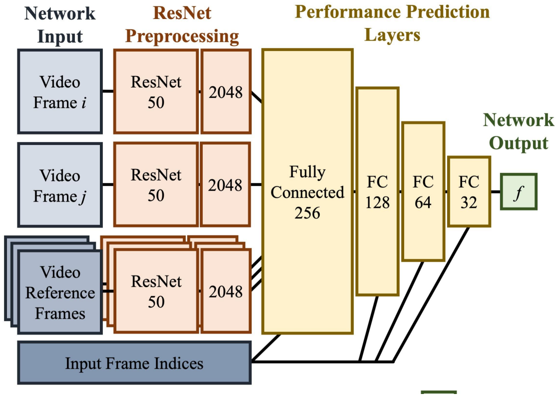


Source

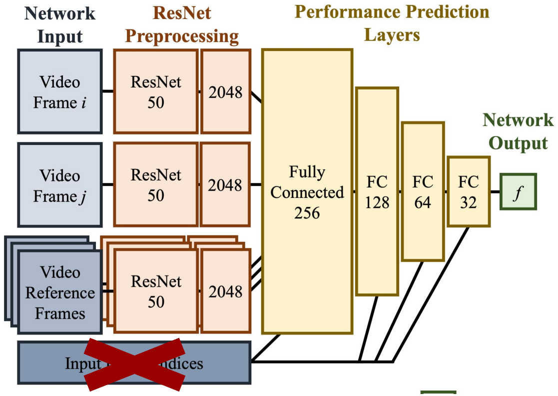
BubbleNets: Configurations

5 configurations are implemented to test efficacy.

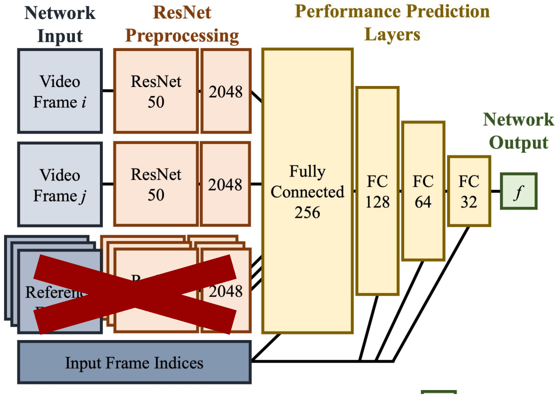
BN_0 : Standard BubbleNets



BN_{NIFI} : No Input Frame Indices



BN_{NRF} : No Reference Frames



Focus on Input Selection

BubbleNets: Configurations

BN_0 : Standard BubbleNets

$$\mathcal{L}(\mathbf{W}) := |(y_i - y_j) - f(x_i, x_j, X_{\text{ref.}}, \mathbf{W})|$$

BN_{LSP} : Single Frame Preference

$$\mathcal{L}_{\text{SP}}(\mathbf{W}) := |y_i - f(x_i, X_{\text{ref.}}, \mathbf{W})|$$

BN_{LF} : Bias toward Middle Frame

$$\mathcal{L}_{\text{F}}(\mathbf{W}) := |(y_i - y_j) - (d_i - d_j) - f(x_i, x_j, X_{\text{ref.}}, \mathbf{W})|$$

$$d_i = \lambda |I_i - I_{\text{MF}}| \qquad I_i = \frac{i}{n}.$$

$\lambda = 0.5$, $I_{MF} = 0.5$

Focus on Loss function Design

Experimental setup

- Primary Dataset: DAVIS 2017
- Best and worst possible frame selection - upper and lower bound
- Simple frame selection: First, Middle, Last, Random
- Determine Batch size:

Table 3. **Ablation Study on DAVIS 2017 Val. Set:** Study of BN input batch size for bubble sort comparisons and end performance.

Batch Size	Performance ($\mathcal{J}+\mathcal{F}$) BN_0	$\mathrm{BN}_{\mathrm{NIFI}}$	$\mathrm{BN}_{\mathcal{L}\mathrm{F}}$	Mean Video Sort Time
1	124.1	122.9	120.5	3.88 s
3	125.2	122.0	121.6	4.83 s
5	**125.2**	**123.8**	**121.7**	5.32 s
10	125.2	122.0	120.3	6.52 s
20	123.6	123.4	120.7	9.34 s

DAVIS dataset annotated frame selecting results

Annotation Frame Selection	Segmentation Performance ($\mathcal{J}+\mathcal{F}$) Mean	Median	Range	Coef. of Variation
DAVIS 2017 Val.				
Best	141.2	143.2	14.9–194.9	0.26
BN_0	**125.2**	**128.9**	**7.6–194.2**	**0.34**
BN_{NIFI}	**123.8**	**129.9**	8.7–194.2	**0.35**
$BN_{\mathcal{L}F}$	**121.7**	**128.0**	**7.6–194.3**	**0.38**
Middle	119.2	124.0	7.6–193.6	0.41
Random	116.5	119.7	1.6–193.2	0.38
First	113.3	117.2	3.5–192.5	0.39
Last	104.7	110.3	4.4–190.1	0.42
Worst	86.3	88.2	1.6–188.9	0.56

Annotation Frame Selection	Segmentation Performance ($\mathcal{J}+\mathcal{F}$) Mean	Median	Range	Coef. of Variation
DAVIS 2016 Val.				
Best	171.2	176.3	130.6–194.9	0.11
BN_0	**159.8**	**168.5**	**72.6–194.5**	**0.18**
BN_{NIFI}	**157.3**	**165.7**	**72.6–194.5**	**0.18**
$BN_{\mathcal{L}F}$	**155.6**	**170.5**	**72.6–193.8**	**0.21**
Middle	155.2	169.5	77.1–193.8	0.21
First	152.8	153.4	115.2–191.7	0.15
Random	147.5	157.3	83.1–194.5	0.25
Last	147.5	153.0	72.0–189.6	0.23
Worst	127.7	141.3	68.3–188.9	0.31

- BN0's use of normalized frame indices is more benificial

DAVIS dataset annotated frame selecting results

Annotation Frame Selection	Segmentation Performance ($\mathcal{J}+\mathcal{F}$)			
	Mean	Median	Range	Coef. of Variation
DAVIS 2017 Val.				
Best	141.2	143.2	14.9–194.9	0.26
BN_0	**125.2**	**128.9**	**7.6–194.2**	**0.34**
$\mathrm{BN}_{\mathrm{NIFI}}$	**123.8**	**129.9**	8.7–194.2	**0.35**
$\mathrm{BN}_{\mathcal{L}\mathrm{F}}$	**121.7**	**128.0**	**7.6–194.3**	**0.38**
Middle	119.2	124.0	7.6–193.6	0.41
Random	116.5	119.7	1.6–193.2	0.38
First	113.3	117.2	3.5–192.5	0.39
Last	104.7	110.3	4.4–190.1	0.42
Worst	86.3	88.2	1.6–188.9	0.56

Annotation Frame Selection	Segmentation Performance ($\mathcal{J}+\mathcal{F}$)			
	Mean	Median	Range	Coef. of Variation
DAVIS 2016 Val.				
Best	171.2	176.3	130.6–194.9	0.11
BN_0	**159.8**	**168.5**	**72.6–194.5**	**0.18**
$\mathrm{BN}_{\mathrm{NIFI}}$	**157.3**	**165.7**	**72.6–194.5**	**0.18**
$\mathrm{BN}_{\mathcal{L}\mathrm{F}}$	**155.6**	**170.5**	**72.6–193.8**	**0.21**
Middle	155.2	169.5	77.1–193.8	0.21
First	152.8	153.4	115.2–191.7	0.15
Random	147.5	157.3	83.1–194.5	0.25
Last	147.5	153.0	72.0–189.6	0.23
Worst	127.7	141.3	68.3–188.9	0.31

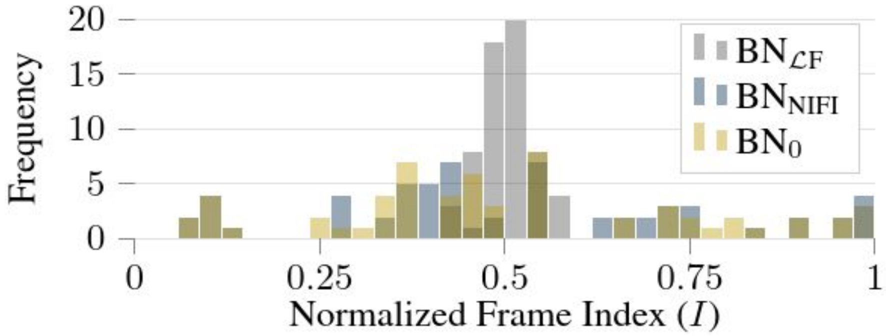
- Middle frame selection has the best performance of all simple frame selections

DAVIS dataset annotated frame selecting results

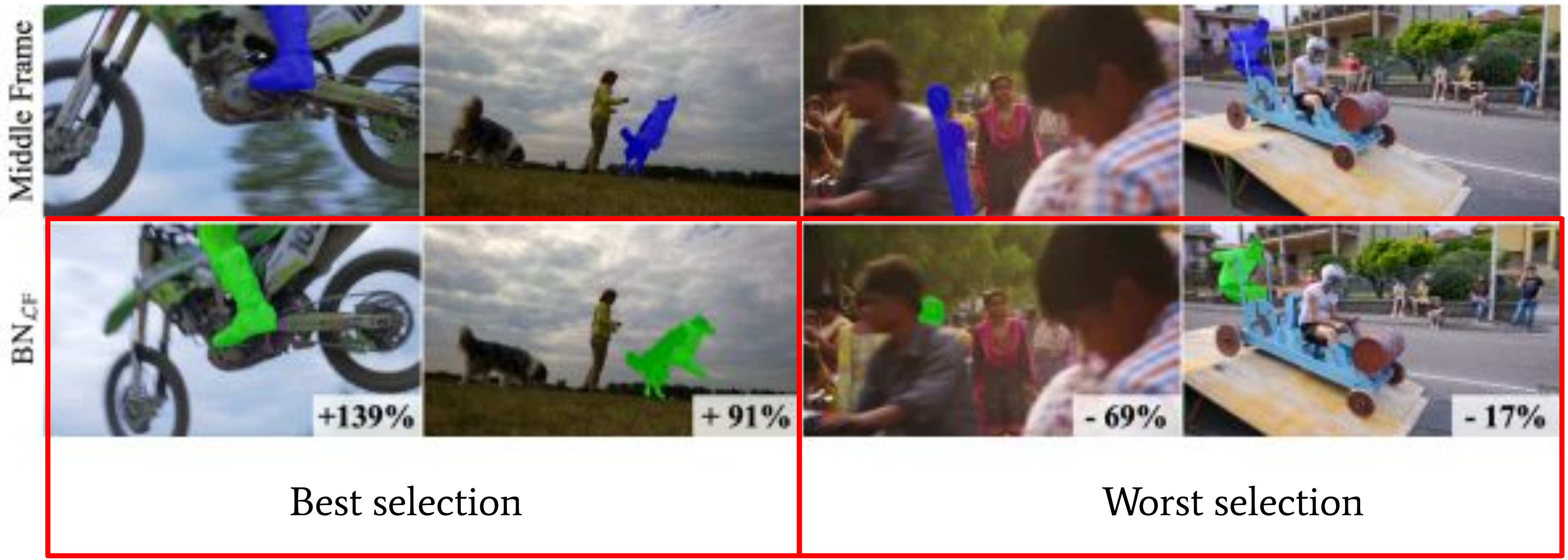
Annotation Frame Selection	Segmentation Performance ($\mathcal{J}+\mathcal{F}$) Mean	Median	Range	Coef. of Variation
DAVIS 2017 Val.				
Best	141.2	143.2	14.9–194.9	0.26
BN_0	**125.2**	**128.9**	**7.6–194.2**	**0.34**
BN_{NIFI}	**123.8**	**129.9**	8.7–194.2	**0.35**
$BN_{\mathcal{L}F}$	**121.7**	**128.0**	**7.6–194.3**	**0.38**
Middle	119.2	124.0	7.6–193.6	0.41
Random	116.5	119.7	1.6–193.2	0.38
First	113.3	117.2	3.5–192.5	0.39
Last	104.7	110.3	4.4–190.1	0.42
Worst	86.3	88.2	1.6–188.9	0.56

Annotation Frame Selection	Segmentation Performance ($\mathcal{J}+\mathcal{F}$) Mean	Median	Range	Coef. of Variation
DAVIS 2016 Val.				
Best	171.2	176.3	130.6–194.9	0.11
BN_0	**159.8**	**168.5**	**72.6–194.5**	**0.18**
BN_{NIFI}	**157.3**	**165.7**	**72.6–194.5**	**0.18**
$BN_{\mathcal{L}F}$	**155.6**	**170.5**	**72.6–193.8**	**0.21**
Middle	155.2	169.5	77.1–193.8	0.21
First	152.8	153.4	115.2–191.7	0.15
Random	147.5	157.3	83.1–194.5	0.25
Last	147.5	153.0	72.0–189.6	0.23
Worst	127.7	141.3	68.3–188.9	0.31

- BN_{LF} performs better than Middle selection
- BN_{LF} biases selections toward the middle of each video



- $BN_{\mathcal{L}F}$ and Middle frame comparison on DAVIS 2017 Val. set

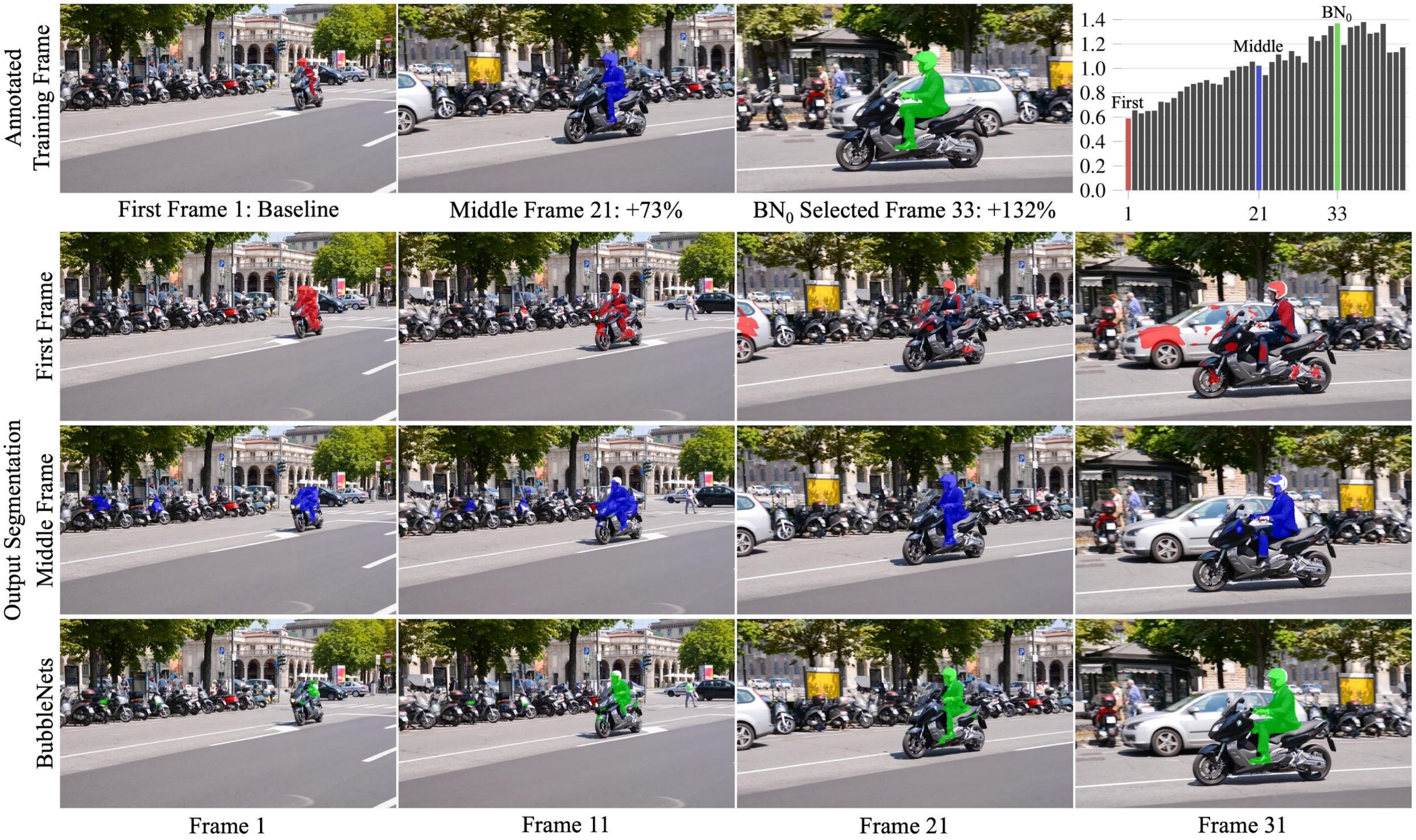


DAVIS dataset annotated frame selecting results

Annotation Frame Selection	Segmentation Performance ($\mathcal{J}+\mathcal{F}$) Mean	Median	Range	Coef. of Variation
DAVIS 2017 Val.				
Best	141.2	143.2	14.9–194.9	0.26
BN_0	**125.2**	**128.9**	**7.6–194.2**	**0.34**
BN_{NIFI}	**123.8**	**129.9**	8.7–194.2	**0.35**
$BN_{\mathcal{L}F}$	**121.7**	**128.0**	**7.6–194.3**	**0.38**
Middle	119.2	124.0	7.6–193.6	0.41
Random	116.5	119.7	1.6–193.2	0.38
First	113.3	117.2	3.5–192.5	0.39
Last	104.7	110.3	4.4–190.1	0.42
Worst	86.3	88.2	1.6–188.9	0.56

Annotation Frame Selection	Segmentation Performance ($\mathcal{J}+\mathcal{F}$) Mean	Median	Range	Coef. of Variation
DAVIS 2016 Val.				
Best	171.2	176.3	130.6–194.9	0.11
BN_0	**159.8**	**168.5**	**72.6–194.5**	**0.18**
BN_{NIFI}	**157.3**	**165.7**	**72.6–194.5**	**0.18**
$BN_{\mathcal{L}F}$	**155.6**	**170.5**	**72.6–193.8**	**0.21**
Middle	155.2	169.5	77.1–193.8	0.21
First	152.8	153.4	115.2–191.7	0.15
Random	147.5	157.3	83.1–194.5	0.25
Last	147.5	153.0	72.0–189.6	0.23
Worst	127.7	141.3	68.3–188.9	0.31

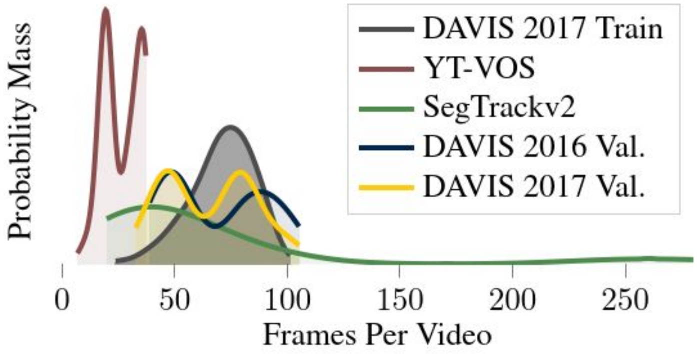
- BN0 has better performance than all simple frame selections



Annotated Training Frame
First Frame 1: Baseline
Middle Frame 21: +73%
BN_0 Selected Frame 33: +132%
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
First
Middle
BN_0
1
21
33
Output Segmentation
First Frame
Middle Frame
BubbleNets
Frame 1
Frame 11
Frame 21
Frame 31

If we have limited number of annotated frames

Number of	DAVIS 2017 Train	DAVIS 2017 Val.	DAVIS '16 Val.	SegTrack v2	YT-VOS (1st 1,000)
Objects	144	61	20	24	1,000
Videos	60	30	20	14	607
Annotated Frames	4,209	1,999	1,376	1,066	16,715
Object Annotations	10,238	3,984	1,376	1,515	26,742
Annotated Frames Per Video					
Mean	70.2	66.6	68.8	76.1	27.5
Median	71	67.5	67.5	**39**	**30**
Range	25–100	34–104	40–104	21–279	8–36
Coef. of Variation	0.22	0.31	0.32	1.03	0.29



If we have limited number of annotated frames

- $BN_{\mathcal{LF}}$ outperforms all other simple selections strategies

Table 5. **Results on Datasets with Limited Frames Per Video.**

Annotation Frame Selection	Segmentation Performance ($\mathcal{J}+\mathcal{F}$)			
	Mean	Median	Range	Coef. of Variation
SegTrackv2				
$BN_{\mathcal{LF}}$	**134.7**	**145.9**	**14.3–184.6**	**0.32**
Middle	134.5	143.5	14.3–182.8	0.32
BN_{NIFI}	134.3	144.2	33.9–178.5	0.30
BN_0	130.6	127.3	50.0–183.2	0.30
Last	123.6	130.4	14.3–178.4	0.36
First	122.3	122.5	45.8–181.7	0.31
YT-VOS (1st 1,000)				
$BN_{\mathcal{LF}}$	**115.5**	**126.6**	**0.0–197.3**	**0.46**
Middle	115.0	124.2	0.0–196.2	0.46
BN_{NIFI}	111.8	121.0	0.0–196.3	0.47
BN_0	110.4	121.5	0.0–194.1	0.49
First	107.3	114.0	0.0–196.3	0.49
Last	101.2	108.1	0.0–195.4	0.56

If we have limited number of frames

- Additional experiment:
  Analyze 10 longest and shortest videos from DAVIS 2017 validation set

Videos from DAVIS 2017 Val.	Number of Frames	Relative Mean ($\mathcal{J}+\mathcal{F}$)		
		BN_0	$\mathrm{BN}_{\mathrm{NIFI}}$	$\mathrm{BN}_{\mathcal{L}\mathrm{F}}$
10 Longest	81–104	**+ 11.8%**	**+ 10.9%**	+ 4.0%
All	34–104	**+ 10.5%**	**+ 9.3%**	+ 7.4%
10 Shortest	**34–43**	+ 4.9%	+ 5.0%	+ 3.3%

Cross evaluation results for different segmentation methods

Segmentation Method	Frame Selection and DAVIS $\mathcal{J}$ & $\mathcal{F}$ Mean				
	First	Middle	$BN_{\mathcal{L}F}$	BN_{NIFI}	BN_0
OSVOS	56.6	59.6	60.8	61.9	**62.6**
OnAVOS	63.9	68.4	68.5	68.4	**69.2**

Strengths and Weaknesses

Strengths :

- Further improves the performance of OSVOS using fine-tune frame selection network.
- Applies loss function to learn from fewer initial frame labels, as labeling video data are expensive.

Weaknesses:

- The reason of BubbleNets performs better on longer video is not very reliable to us.

Open Research Questions

- There are some other VOS models that uses multiple frames as training examples. Can BubbleNets possibly be applied on those models and select the frames that can improve the performance the most?

Reference Link

- https://www.youtube.com/watch?v=XBEMuFVC2Ig&fbclid=IwAR0fM9tzUGwruiqzg_epQeVzoGWNPKY6BpMSuYXYydahUXNWrozgHgg9RiE
- https://towardsdatascience.com/semantic-segmentation-with-deep-learning-a-guide-and-code-e52fc8958823
- https://techburst.io/video-object-segmentation-the-basics-758e77321914