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

Brent A. Griffin Jason J. Corso
University of Michigan
{griffb,jjcorso}@umich.edu

arXiv:1903.11779v1 [cs.CV] 28 Mar 2019

Abstract

Semi-supervised video object segmentation has made significant progress on real and challenging videos in recent years. The current paradigm for segmentation methods and benchmark datasets is to segment objects in video provided a single annotation in the first frame. However, we find that segmentation performance across the entire video varies dramatically when selecting an alternative frame for annotation. This paper address the problem of learning to suggest the single best frame across the video for user annotation—this is, in fact, never the first frame of video. We achieve this by introducing BubbleNets, a novel deep sorting network that learns to select frames using a performance-based loss function that enables the conversion of expansive amounts of training examples from already existing datasets. Using BubbleNets, we are able to achieve an 11% relative improvement in segmentation performance on the DAVIS benchmark without any changes to the underlying method of segmentation.

1. Introduction

Video object segmentation (VOS), the dense separation of objects in video from background, remains a hotly studied area of video understanding. Motivated by the high cost of densely-annotated user segmentations in video [5, 38], our community is developing many new VOS methods that are regularly evaluated on the benchmark datasets supporting VOS research [22, 31, 33, 37, 45]. Compared to unsupervised VOS [12, 21, 29, 44], semi-supervised VOS, the problem of segmenting objects in video given a single user-annotated frame, has seen rampant advances, even within

Figure 1. The current paradigm for video object segmentation is to segment an object annotated in the first frame of video (yellow, left). However, selecting a different frame for annotation changes performance across the entire video [for better (green) *or* worse (red)]. To best use an annotator's time, our deep sorting framework suggests a frame that will improve segmentation performance.

is critical that we improve performance of semi-supervised VOS methods by providing the best single annotation frame possible. However, we are not aware of any work that seeks to learn which frame to annotate for VOS.

To that end, this paper addresses the problem of selecting a single video frame for annotation that will lead to greater performance. Starting from an untouched video, we select an annotation frame using our deep bubble sorting framework, which makes relative performance predictions between pairs of frames using our custom network, BubbleNets. BubbleNets iteratively compares and swaps adjacent video frames until the frame with the greatest predicted performance is ranked highest, at which point, it is selected for the user to annotate and use for VOS. To train Bub-

BubbleNets 🛁

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

Griffin, B. A., & Corso, J. J. (2019). BubbleNets: Learning to Select the Guidance Frame in Video Object Segmentation by Deep Sorting Frames. Retrieved from http://arxiv.org/abs/1903.11779

Computer Vision Lab, Hanyang University
Paper review 30 Sep 2019
Jihun Kim

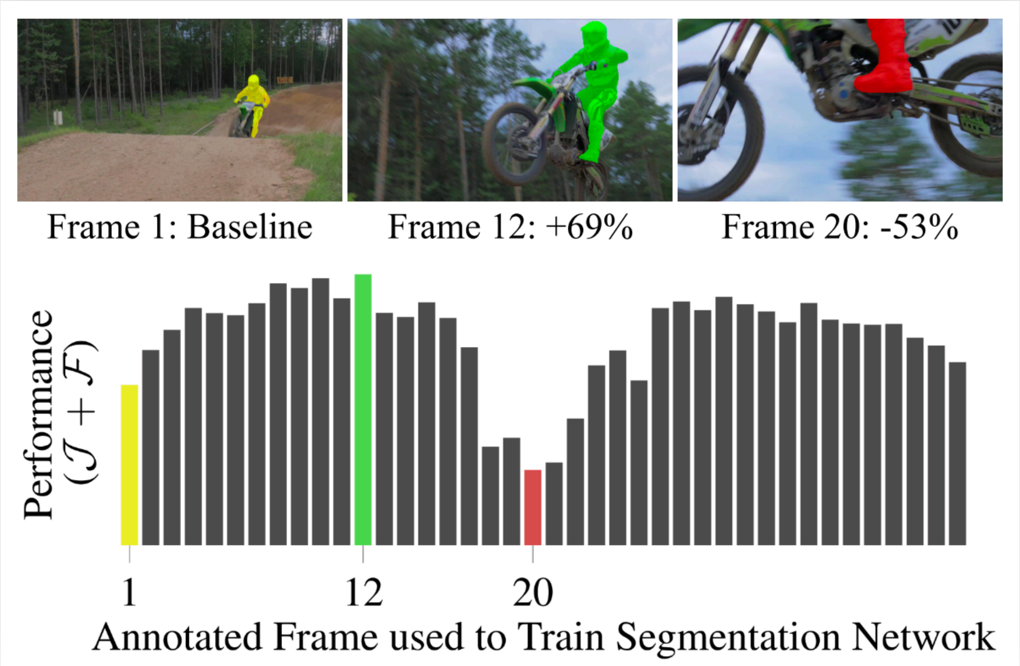
Introduction

Introduction 👀

Video object segmentation (VOS), the dense separation of objects in video from background, remains a hotly studied area of video understanding.

Using different frames for annotation changes performance dramatically.

This paper addresses the problem of selecting a single video frame for annotation that will lead to greater performance.



Selecting a different frame for annotation changes performance across the entire video.

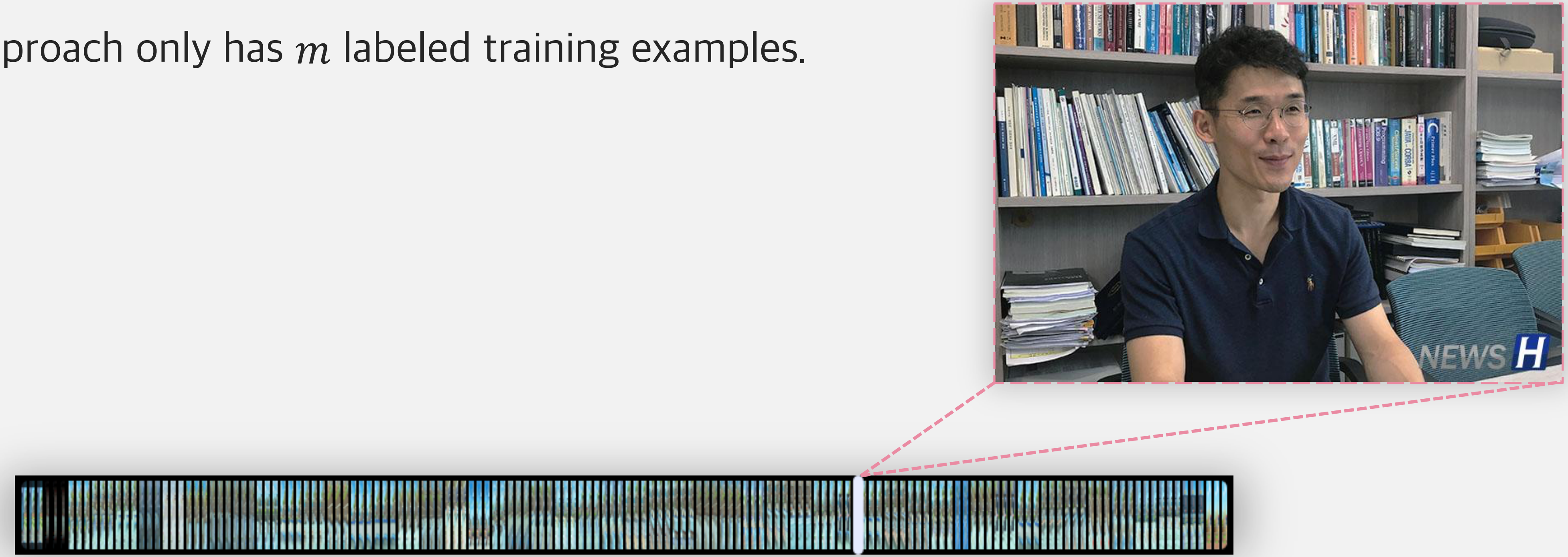
Methodology

Predicting Relative Performance 🎯

Assume we are given a set of m training videos wherein
each video has n frames with labels corresponding to some performance metric, $y \in \mathbb{R}$.

One way to accomplish this task is to use the entire video as input to a network
and output the frame index with the greatest predicted performance.

However, this approach only has m labeled training examples.

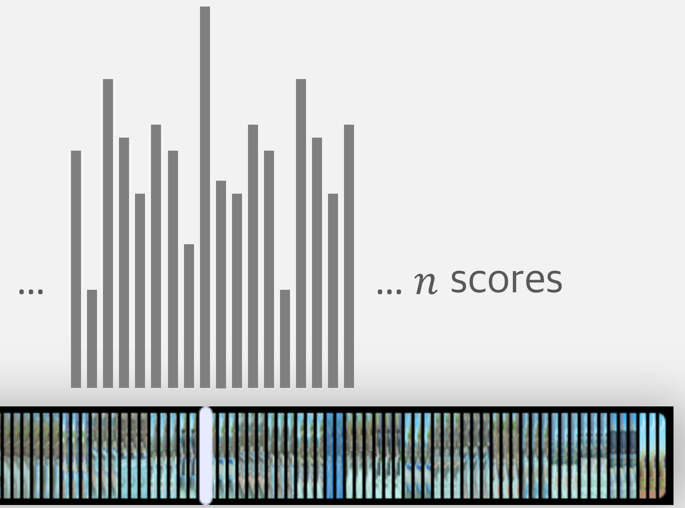


Predicting Relative Performance 🎯

Assume we are given a set of m training videos wherein
each video has n frames with labels corresponding to some performance metric, $y \in \mathbb{R}$.

A second way to formulate this problem is to use individual frames as input
to a network and output the predicted performance of each frame.

Using this formulation, the frame with the maximum predicted performance can be selected
from each video and there are $m \times n$ labeled training examples.

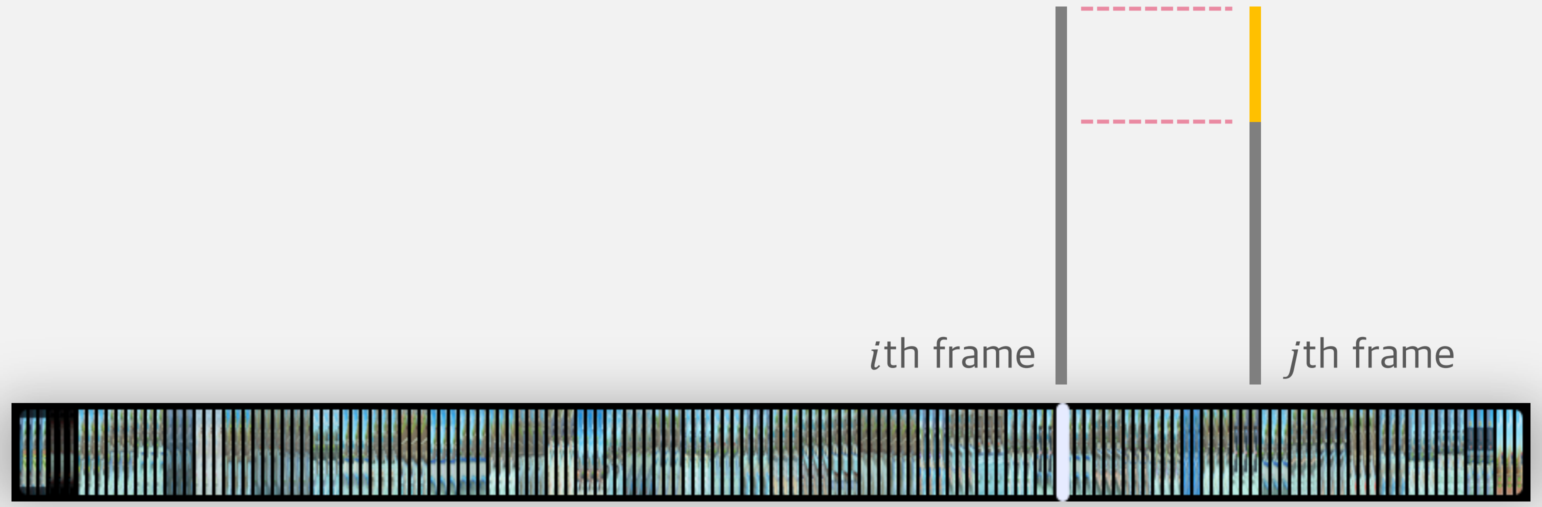


Predicting Relative Performance 🎯

Assume we are given a set of m training videos wherein
each video has n frames with labels corresponding to some performance metric, $y \in \mathbb{R}$.

BN predicts the relative difference in performance of two frames being compared.

This difference may seem trivial, but it effectively increases the number of labels and training examples from $m \times n$ to $\mathrm{m} \times \binom{n}{2} \approx \frac{mn^2}{2}$.



Predicting Relative Performance 🎯

Assume we are given a set of m training videos wherein
each video has n frames with labels corresponding to some performance metric, $y \in \mathbb{R}$.

×To further increase the number of unique training examples and increase BN's accuracy,
we use k random video reference frames as an additional network input.

When predicting the relative performance between two frames,
additional consideration can be given to the frame that better represents the reference frames.

Loss Function 📉

Finally, we define our performance loss function as:

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

where W are the trainable parameters of BN,
y_i is the performance label associated with the ith video frame,
x_i is the image and normalized frame index associated with the ith video frame,
$X_{ref.}$ is the set of k reference images and frame indices,
and f is the predicted relative performance.

For later use, denote the normalized frame index for the ith frame of an n-frame video as

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

Including I as an input enables BN to also consider temporal proximity of frames.

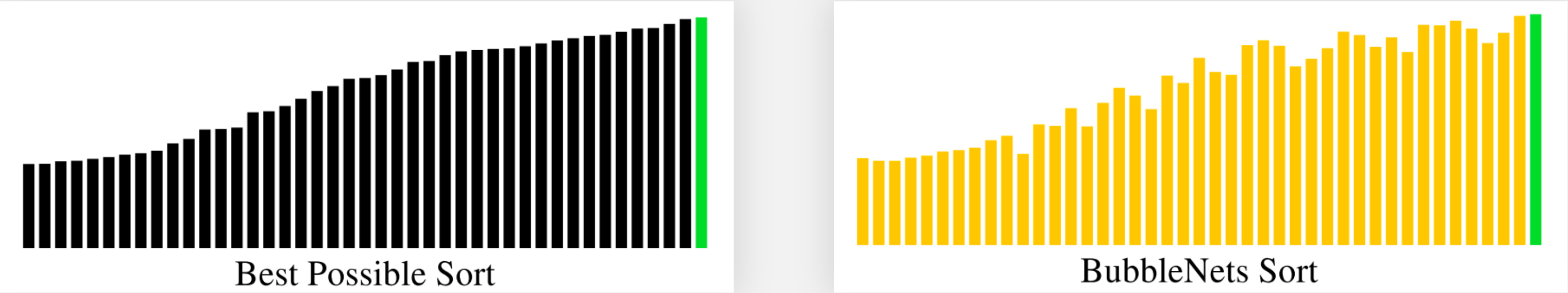
Deep Bubble Sorting 🛁

Same as bubble sort, except using BN as comparison function.

Normally, bubble sort is deterministic and only needs one pass through a list to promote the greatest element to the top.

BN uses k random video reference frames as input for each prediction, and using a different set of reference frames can change that prediction.

Thus, a BN comparison for the same two frames can change.



BubbleNets Prediction Sort of Motorbike Video.

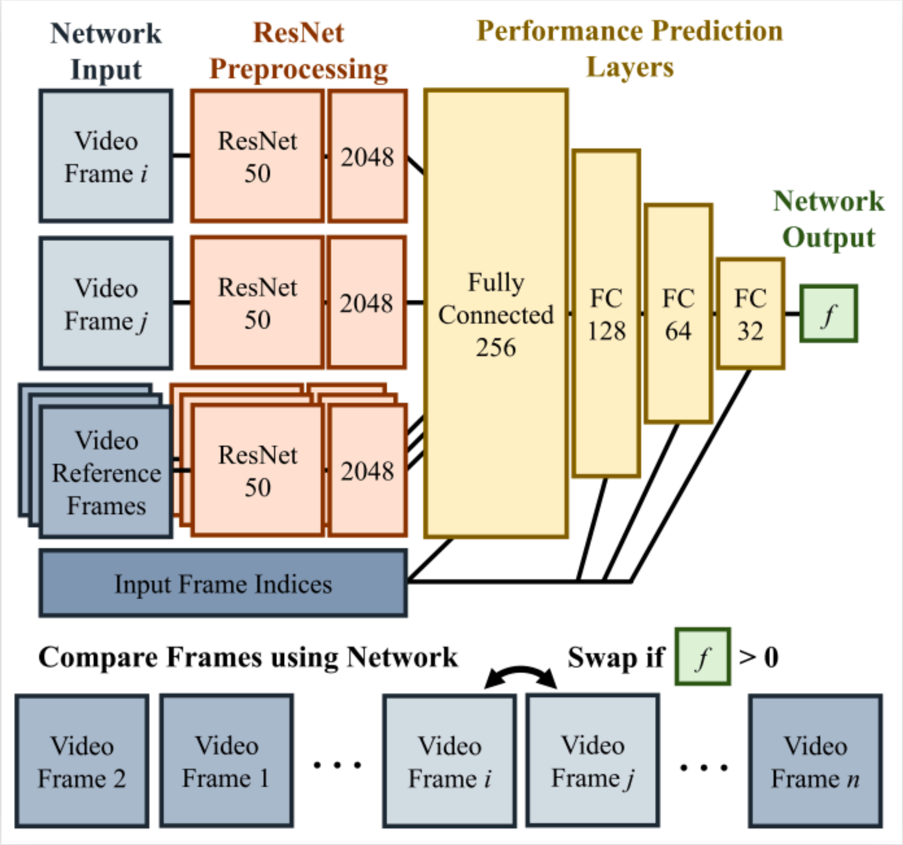
BubbleNets Architecture ⚙️

The input has two comparison images, three reference images, and normalized indices for all five frames.

All performance prediction layers include the normalized frame indices as input and use a Leaky ReLU activation function;
The later three prediction layers have 20% dropout for all inputs during training.

After the performance prediction layers,
BN architecture ends with one last fully connected neuron that is the output relative performance prediction:

$$f(x_i, x_j, X_{\text{ref.}}, \mathbf{W}) \in \mathbb{R}$$



BubbleNets Framework.

Experiment

Configurations 🔧

To test the efficacy of new concepts and establish best practices, we implement five BN configurations for VOS.

The first configuration (BN_0) uses the standard BN architecture.
The second and third configurations are similar to BN_0 but use
No Input Frame Indices (BN_{NIFI}) or No Reference Frames (BN_{NRF}).
The fourth and fifth configurations are similar to BN_0 but use loss functions modified
that predict Single-frame Performance (BN_{LSP}) or bias toward middle Frame selection (BN_{LF}).

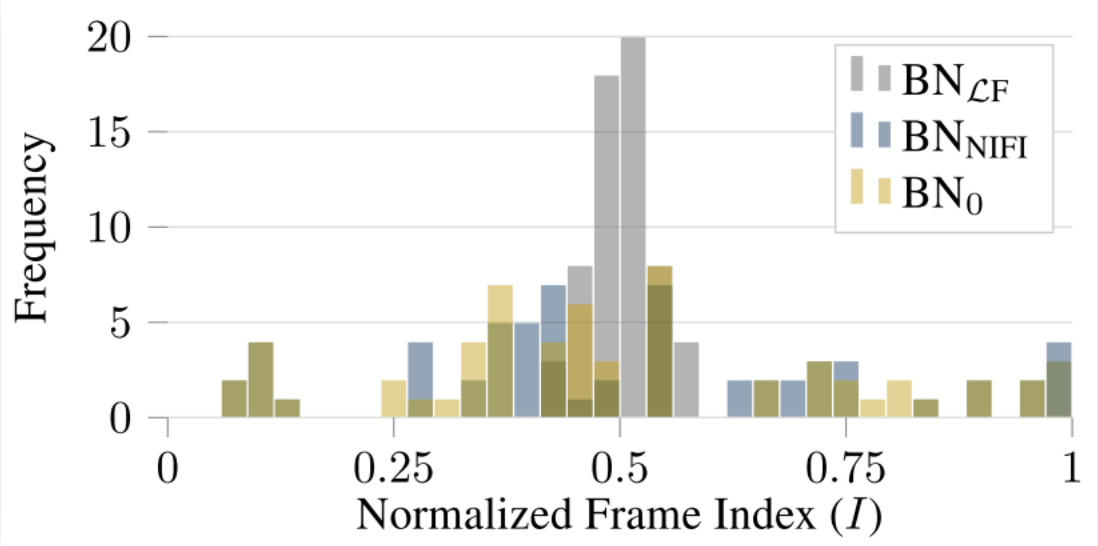
Table 2. **BubbleNets Configurations.**

Config. ID	Input Frame Index	Input Ref. Frame	Loss Funct.	Total Training Iterations	Total Training Time	DAVIS '17 Val. Mean $\mathcal{J}$	DAVIS '17 Val. Mean $\mathcal{F}$
BN_0	Yes	Yes	$\mathcal{L}$ (1)	3,125	5m 11s	**59.7**	**65.5**
BN_{NIFI}	**No**	Yes	$\mathcal{L}$ (1)	2,500	3m 52s	58.7	65.0
$BN_{\mathcal{L}F}$	Yes	Yes	$\mathcal{L}_F$ (5)	8,125	15m 30s	57.8	63.8
BN_{NRF}	Yes	**No**	$\mathcal{L}$ (1)	3,125	2m 20s	55.4	62.3
$BN_{\mathcal{L}SP}$	Yes	Yes	$\mathcal{L}_{SP}$ (4)	1,875	2m 32s	55.1	62.3

DAVIS Validation 🧪

Middle frame selection has the best performance of all simple strategies.

All BN configurations outperform the simple selection strategies, and BN0 performs best of all BN configurations.



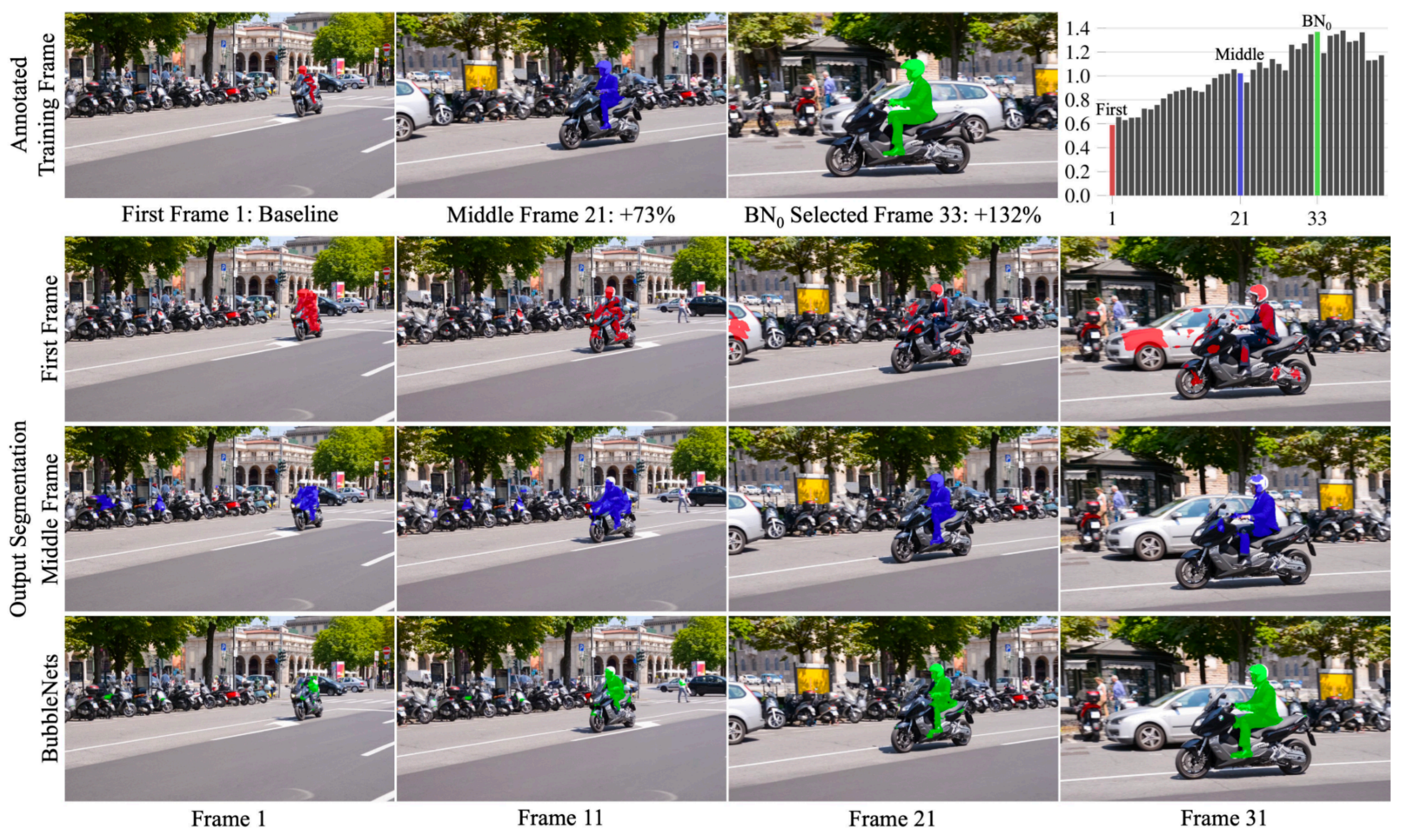
Frame-Selection Locations in Video.

Table 4. **Dataset Annotated Frame Selection Results.**

Annotation Frame Selection	Segmentation Performance ($\mathcal{J}+\mathcal{F}$) Mean	Med.	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$_{\text{NIFI}}$	**123.8**	**129.9**	8.7–194.2	**0.35**
BN$_{\mathcal{L}\text{F}}$	**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
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$_{\text{NIFI}}$	**157.3**	**165.7**	**72.6–194.5**	**0.18**
BN$_{\mathcal{L}\text{F}}$	**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

Complete annotated frame selection results for the DAVIS 2016 and 2017 validation sets.

DAVIS Validation 🧪



Qualitative Comparison on DAVIS 2017 Validation Set:

Segmentations from different annotated frame selection strategies.


arXiv:1903.11779v1 [cs.CV] 28 Mar 2019

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

Brent A. Griffin Jason J. Corso
University of Michigan
{griffb,jjcorso}@umich.edu

Abstract

Semi-supervised video object segmentation has made significant progress on real and challenging videos in recent years. The current paradigm for segmentation methods and benchmark datasets is to segment objects in video provided a single annotation in the first frame. However, we find that segmentation performance across the entire video varies dramatically when selecting an alternative frame for annotation. This paper address the problem of learning to suggest the single best frame across the video for user annotation—this is, in fact, never the first frame of video. We achieve this by introducing BubbleNets, a novel deep sorting network that learns to select frames using a performance-based loss function that enables the conversion of expansive amounts of training examples from already existing datasets. Using BubbleNets, we are able to achieve an 11% relative improvement in segmentation performance on the DAVIS benchmark without any changes to the underlying method of segmentation.

1. Introduction

Video object segmentation (VOS), the dense separation of objects in video from background, remains a hotly studied area of video understanding. Motivated by the high cost of densely-annotated user segmentations in video [5, 38], our community is developing many new VOS methods that are regularly evaluated on the benchmark datasets supporting VOS research [22, 31, 33, 37, 45]. Compared to unsupervised VOS [12, 21, 29, 44], semi-supervised VOS, the problem of segmenting objects in video given a single user-annotated frame, has seen rampant advances, even within

Frame 1: Baseline Frame 12: +69% Frame 20: -53%

Performance ($\mathcal{J}+\mathcal{F}$)

1 12 20

Annotated Frame used to Train Segmentation Network

Figure 1. The current paradigm for video object segmentation is to segment an object annotated in the first frame of video (yellow, left). However, selecting a different frame for annotation changes performance across the entire video [for better (green) *or* worse (red)]. To best use an annotator's time, our deep sorting framework suggests a frame that will improve segmentation performance.

is critical that we improve performance of semi-supervised VOS methods by providing the best single annotation frame possible. However, we are not aware of any work that seeks to learn which frame to annotate for VOS.

To that end, this paper addresses the problem of selecting a single video frame for annotation that will lead to greater performance. Starting from an untouched video, we select an annotation frame using our deep bubble sorting framework, which makes relative performance predictions between pairs of frames using our custom network, BubbleNets. BubbleNets iteratively compares and swaps adjacent video frames until the frame with the greatest predicted performance is ranked highest, at which point, it is selected for the user to annotate and use for VOS. To train Bub-

BubbleNets 🛁

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

Griffin, B. A., & Corso, J. J. (2019). BubbleNets: Learning to Select the Guidance Frame in Video Object Segmentation by Deep Sorting Frames. Retrieved from http://arxiv.org/abs/1903.11779

Computer Vision Lab, Hanyang University
Paper review 30 Sep 2019
Jihun Kim