

Towards Robust Video Object Segmentation with Adaptive Object Calibration

Xiaohao Xu*

Huazhong University of Science & Technology
xuh1102019@outlook.com

Xiang Ming

Microsoft Research Asia
xiangming@microsoft.com

Jinglu Wang

Microsoft Research Asia
jinglwa@microsoft.com

Yan Lu

Microsoft Research Asia
yanlu@microsoft.com

ABSTRACT

In the booming video era, video segmentation attracts increasing research attention in the multimedia community. Semi-supervised video object segmentation (VOS) aims at segmenting objects in all *target* frames of a video, given annotated object masks of *reference* frames. Most existing methods build pixel-wise reference-target correlations and then perform pixel-wise tracking to obtain target masks. Due to neglecting object-level cues, pixel-level approaches make the tracking vulnerable to perturbations, and even indiscriminate among similar objects. Towards robust VOS, the key insight is to calibrate the representation and mask of each specific object to be expressive and discriminative. Accordingly, we propose a new deep network, which can adaptively construct object representations and calibrate object masks to achieve stronger robustness. First, we construct the object representations by applying an adaptive object proxy (AOP) aggregation method, where the proxies represent arbitrary-shaped segments at multi-levels for reference. Then, prototype masks are initially generated from the reference-target correlations based on AOP. Afterwards, such proto-masks are further calibrated through network modulation, conditioning on the object proxy representations. We consolidate this conditional mask calibration process in a progressive manner, where the object representations and proto-masks evolve to be discriminative iteratively. Extensive experiments are conducted on the standard VOS benchmarks, YouTube-VOS-18/19 and DAVIS-17. Our model achieves the state-of-the-art performance among existing published works, and also exhibits superior robustness against perturbations.

CCS CONCEPTS

- Computing methodologies → Video segmentation.

KEYWORDS

video object segmentation, robustness, neural network

*The work was done when Xiaohao Xu was an intern at MSRA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '22, October 10–17, 2022, Lisbon, Portugal

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/XXXXXXX.XXXXXXX>

ACM Reference Format:

Xiaohao Xu, Jinglu Wang, Xiang Ming, and Yan Lu. 2022. Towards Robust Video Object Segmentation with Adaptive Object Calibration. In *Proceedings of ACM Multimedia '22 (MM '22)*. ACM, Lisbon, Portugal, 19 pages. <https://doi.org/XXXXXXX.XXXXXXX>

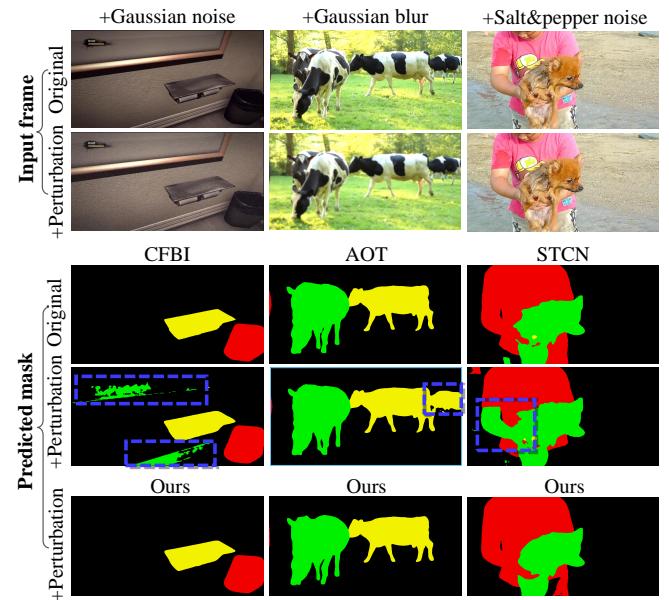


Figure 1: Existing advanced VOS models, including CFBI [63], AOT [64], and STCN [7], are fragile to natural perturbations, especially for the scene containing multiple objects. Our model with adaptive object calibration shows superior robustness against perturbations.

1 INTRODUCTION

Video Object Segmentation (VOS) is one of the most attractive video-related research problems in the multimedia area. This work focuses on the semi-supervised setting, which aims to segment one or multiple objects in a video sequence given annotated object masks of reference frames. In real-world videos, perturbations commonly exist due to signal noise, camera defocus, and fast motion. The model robustness becomes critical for real applications, especially for some safety-aware applications, such as autonomous driving.

However, existing VOS methods have seldom discussed the robustness against perturbations. According to our pilot study, the performance of current advanced models [7, 63, 64] degrade largely under simple image perturbations, as is shown in Fig. 1, especially for the scenes where multiple objects exist. Towards robust VOS, we consider two key factors that matter: 1) a robust object representation extracted from references for building correlations to the target frames; 2) a robust mask calibration process to produce pixel-wise classification conditioning on the referenced object representation.

For the representation of objects in VOS, most previous works [6, 16, 17, 28, 32, 39, 46, 52, 57, 63, 69?] directly employ pixel-level features. Concretely, pixel-level matching [17] is utilized to track pixels across frames. Object-related information is only implicitly encoded in the pixel-level feature in terms of limited receptive fields. Such pixel-based representation is often error-prone, leading to noisy results, especially in the cases that similar objects co-exist [28]. Meanwhile, following the fashion of object-tracking [56], some VOS works turn to object-level representations. For instance, objects are represented as box-bounded proposals [25, 51] generated with an off-the-shelf object detector and predictions are made by correlating object proposals of the current frame with historical templates. Despite these methods can eliminate noisy predictions to some extent, fine-grained correspondences are lost, leading to inaccurate results in details.

For the mask calibration process, some recent works [39?] introduce network modulation [11, 18, 40, 53] to conditionally decode the target object masks, which have already achieved successful results. Most methods [39?] consider individual objects and neglect interactions between different objects, which often fail in scenes with multiple objects. We only find a recent method, AOT [64], encoding object-aware interactions using transformer-based association, which demonstrates that such kind of interactions can contribute to a large performance gain. However, AOT could not discriminate similar objects well against perturbations (Fig. 1).

To achieve the robustness of VOS, we propose a deep network that adaptively calibrates the object representation and masks. First, we introduce an adaptive object proxy representation to extract object-specific features from the reference. The new representation is constructed from multiple granularities for building robust correlations between the reference and target frames afterward. Then, prototype masks are initially generated from the calculated correlations. After that, we progressively perform the object mask calibration to refine the masks, conditioning on the learned object proxies. The mask calibration is implemented with network modulation, which performs channel-wise reweighting according to the learned conditioning weights. Different from previous methods, we also update the learned conditioning weights to discriminate each object from other co-existing ones during the mask evolving process. Thus, the object representation and mask are calibrated in an interleaving manner.

Our contributions are summarized as follows.

- We are the first to conduct a comprehensive study of the robustness of VOS models against perturbations. Towards robust VOS, we rethink the problem from the perspective of object representation and mask calibration, and propose a framework with stronger perturbation robustness.

- We introduce an adaptive object proxy representation for referenced objects robustly, which reduces errors incurred by unstable pixel-level matching.
- We calibrate the object masks by updating object representation and masks in an interleaving manner progressively, achieving discrimination among co-existing objects.

Extensive experiments are conducted on the standard VOS benchmarks and our constructed pilot robustness benchmark. Our model not only achieves the state-of-the-art results on the standard YouTube-VOS-18/19 and DAVIS benchmarks among existing published methods, but also exhibits superior robustness under perturbations.

2 RELATED WORK

Video Object Segmentation. Early video object segmentation methods [3, 8, 23, 42, 61, 68] use online fine-tuning, calculate optical flow at high computational cost, or perform segmentation in a sequence-to-sequence manner. Recent online VOS methods [57] aim at achieving good performance while maintaining a real-time speed, which can be divided into propagation-based and matching-based models. For propagation-based models [3, 23, 42], the guidance of segmentation masks from past frames are introduced during the process of mask decoding. For matching-based models [5, 10, 17, 17, 26, 27, 30, 35, 49, 55, 63, 65, 67], an embedding space is learnt for target objects. Recently, STM-based networks [6, 16, 28, 32, 39, 46, 47, 52?] achieve impressive results with memory networks that memorize and read information from past frames. With adaptive object representation in multiple levels and object mask calibration for multi-object discrimination during mask decoding, our model can enhance model robustness besides better performance.

Model Robustness. As robustness is crucial for real-world applications with safety concerns, there is a growing trend [12, 14, 19, 21, 22, 36] to evaluate and enhance the model robustness against corruptions and perturbations. Previous study [58] suggests that adversarial perturbations on images may lead to noise in the features constructed by these networks, thus making the final prediction unstable. To reduce model fragility, many attempts have been proposed to boost the robustness of image-related tasks [54, 70]. However, there is no work to study the vulnerability of VOS models against perturbations. This work aims to fill this gap and proposes two components for the robustness enhancement of VOS models.

3 ADAPTIVE OBJECT CALIBRATION NETWORK FOR ROBUST VOS

We propose the adaptive object calibration network, which improves the robustness of VOS from the two key factors, *i.e.*, the object representation and mask calibration.

3.1 Overview.

Fig. 2 illustrates the overview of the proposed network. Given the target frame \mathbf{x}_t and reference frames $\{\mathbf{x}_r\}, r \in \mathcal{S}_{ref}$ with annotated object masks $\{\mathbf{y}_r\}$, the goal is to predict the segmentation mask for each object $i \in \{0, 1, \dots, N\}$ ($i=0$ indicates the background). Our basic setting uses the first and previous frame as references, namely, $\mathcal{S}_{ref} = \{1, t-1\}$. After extracting image features with the backbone, reference features are combined with downsampled reference masks to form basic object-specific embeddings $\{\mathbf{e}_r^i\}_{i=1}^N$. To robustly

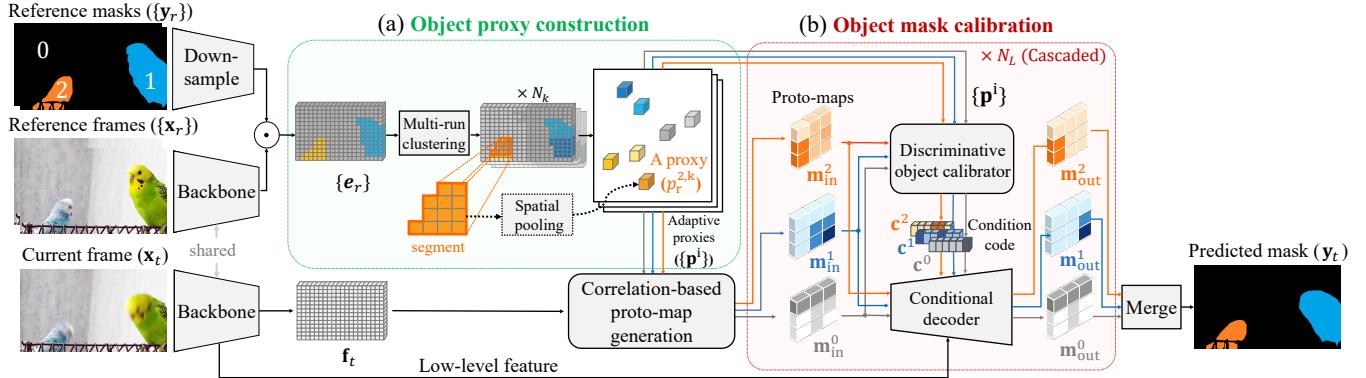


Figure 2: Overview of the proposed model. Given reference frames $\{x_r\}$ with annotated object masks $\{y_r\}$ and the target frame x_t , our goal is to predict the target mask y_t for all annotated objects ($\{0, 1, 2\}$ for this case). The object calibration network contains two stages, i.e., object proxy construction and object mask calibration. (a) We first construct the adaptive object-specific proxies p^i for each object i from the reference features and masks. The prototype maps $\{m^i\}$ are generated with the correlation between current frame feature and adaptive proxy set $\{p^i\}$. (b) The object masks are progressively calibrated from $\{m^i_{in}\}$ to $\{m^i_{out}\}$ with condition codes $\{c^i\}$. Meanwhile, the condition code for a specific object evolves to be discriminative among co-existing objects. The calibrated outputs of the last iteration are merged as the final mask of all objects y_t .

build correlations from the reference to the target frames, we introduce an adaptive proxy representation for reference object context. The proxies convey object-specific information at multiple levels, thus reducing feature matching noise. Then, initial proto-maps $\{m^i\}_{i=1}^N$ of objects are generated from the calculated correlations. Afterwards, the proto-maps are progressively calibrated with condition codes $\{c^i\}$ with N_L iterations. In the calibration process, each condition code c^i of object i evolves to be discriminative from other objects and background, and then serves as channel-wise weights to modulate proto-maps from m^i_{in} to m^i_{out} with the conditional decoder. The final mask y_t is obtained by merging the output of the last conditional decoder with an *argmax* operation.

3.2 Object Proxy Construction

The VOS problem is also known as mask tracking, and one of the principles is to build robust correlations from reference to target frames. The first essential step is to construct object-aware representations from the reference frames.

Object Proxy. We denote a representative embedding of each object as object proxy, e.g., $p^i \in \mathbb{R}^{1 \times 1 \times C_p}$ is a proxy of object i . Each pixel conveys an object proxy for later pixel-wise matching between target and reference frames.

All previous matching-based VOS methods [6, 10, 17, 28, 32, 35, 39, 46, 49, 63] employ **pixel-level** proxies, that is, each pixel conveys the feature of itself subject to the specific object mask (column 1 in Fig. 3 (b)). Let us consider the proxy map $p^i \in \mathbb{R}^{H \times W \times C_p}$ for an object i from reference frames:

$$p^i = [\dots; e_r^i; \dots], \quad r \in \mathcal{S}_{ref}, \quad e_r^i = f_r \odot \mathbb{1}_r^i, \quad (1)$$

where $[\dots; \cdot]$ denotes channel-wise concatenation, e_r^i denotes the basic object-specific embedding, f_r is the extracted feature map from the backbone, $\mathbb{1}_r^i$ is the downsampled object-specific binary mask from y_r , \odot denotes element-wise multiplication. Obviously, this representation is not robust against cases where similar objects

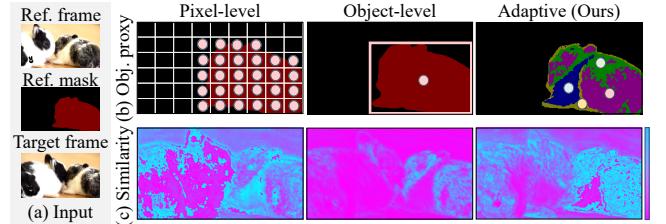


Figure 3: Object proxy representation. (a) Given the image and object mask of the reference frame, we aim to obtain robust object representation for correlation calculation between the target and reference frames. (b) Object proxies are denoted as dots. For our adaptive proxy representation, we observe pixel-level embeddings can be inherently categorized into semantic clusters. Note that we only show one run of clustering here for clear visualization, while multiple runs are performed in the implementation. (c) Correlation maps in which bluer means higher similarity. Pixel-level representation is vulnerable to noise, while object-level one loses details. The proposed adaptive object representation can solve such dilemma by representing an object as a set of semantically-similar proxies via clustering-based aggregation.

co-exist or object appearances change drastically across frames. Besides, the effective receptive field [34] of a single pixel-level embedding could not cover large objects or backgrounds, regardless of the global context. These issues make the embedding error-prone for object-aware correlation calculation.

To represent the proxies with object-specific cues, a straightforward practice is to use the global average pooled feature of the object. However, the **object-level** representation could lose details.

Adaptive Object Proxy (AOP). To address the problems with previous object representations, we propose an adaptive proxy representation, which is a learned combination of embeddings from multiple granularities. We embrace the observation that pixel-level embedding corresponding to semantics can be inherently grouped into meaningful clusters, thus we construct such proxies with clustering algorithms (we use K-Means [29] in implementation). We first cluster the pixel-level feature embeddings \mathbf{e}_r^i and the centroid $p_r^{i,k}$ of each cluster $c_k, k = 1, \dots, K$ serves as a part of the proxy at different level. Clustering is performed in multiple runs with different cluster numbers $K \in \mathcal{L}_{clu} = [K_1, \dots, K_{N_k}]$. Thus, the proxy can represent object information at different granularities. The proxy map $\mathbf{p}_r^{i,K}$ is produced by propagating each cluster centroid over the pixels belonging to its cluster.

We construct the final adaptive proxy representation \mathbf{p}^i of each object i as the concatenated embeddings of all the object-specific adaptive proxies, namely,

$$\mathbf{p}^i = [\dots; \mathbf{p}_r^{i,K}; \dots], \quad r \in \mathcal{S}_{ref}, K \in \mathcal{L}_{clu}. \quad (2)$$

Accordingly, representing the objects with proxies constructed with clusters from different granularities can improve the robustness against noises compared to pixel-level ones, and also preserve more semantic details compared to object-level ones. Fig. 3 illustrates proxy representation with different levels. The pink dots in Fig. 3 (b) represents proxies. With different proxies, the calculated correlations between reference and target frames are different. While correlations from pixel-level and object-level are either noisy or over-smoothing, adaptive proxies can generate more robust correlation maps.

3.3 Object Mask Calibration

After constructing object representation from reference frames, we initialize prototype object masks (proto-map) with the correlation-based proto-map generation block. The object mask calibration process is performed progressively with N_L iterations. In each iteration, the input proto-map \mathbf{m}_{in}^i of object i with the other $\{\mathbf{m}_{in}^j\}_{j \neq i}$ is first used to generate the condition code \mathbf{c}_i for object i . \mathbf{c}_i is calibrated to be discriminative from the other objects. Then, \mathbf{m}_{in}^i is further calibrated with network modulation conditioning on \mathbf{c}_i via the conditional decoder. The overall decoded outputs $\{\mathbf{m}_{out}^i\}_{i=1}^N$ are merged into the final predicted mask \mathbf{y}_t .

Preliminaries of Network Modulation. The mask calibration is implemented by network modulation, which has already been demonstrated to be effective in existing works [62, 63, 65]. Network modulation is an operation to re-weight responses in different channels of a feature map, following the mechanism used in SE-Net [15], taking the form:

$$\mathbf{z}_{out}^m = \mathbf{w}^m \mathbf{z}_{in}^m, \quad (3)$$

where $\mathbf{z}_{in}^m \in \mathbb{R}^{H \times W \times 1}$ and $\mathbf{w}^m \in \mathbb{R}$ are the m -th channel of feature map $\mathbf{z}_{in} \in \mathbb{R}^{H \times W \times C}$ and object-specific weights $\mathbf{w} \in \mathbb{R}^{1 \times 1 \times C}$ respectively. The modulation operation is defined as:

$$\mathbf{z}_{out} = \mathbf{w} \otimes \mathbf{z}_{in}, \quad (4)$$

where \otimes denotes channel-wise multiplication.

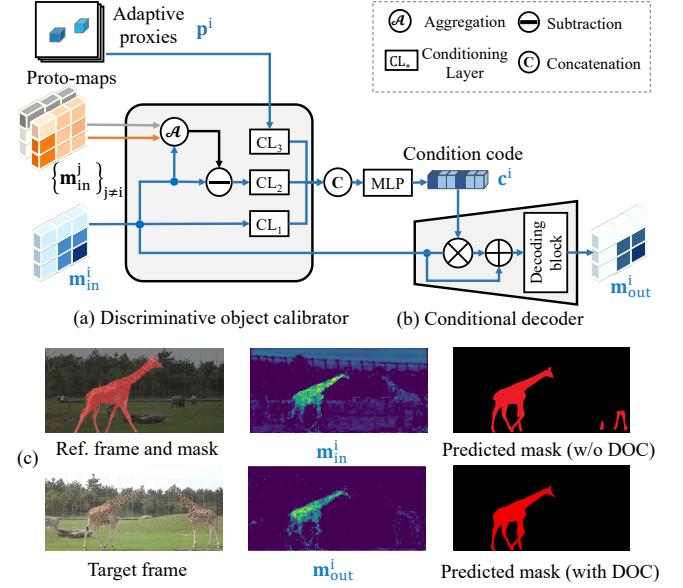


Figure 4: Object mask calibration. (a) Detailed architecture of the discriminative object calibrator (DOC). (b) The condition code c^i with multi-object discrepancy integration is used in the conditional decoder for proto-map calibration. (c) Our mask calibration with the proposed DOC can discriminate the target giraffe from other similar ones better.

The modulation weight \mathbf{w} , which we call condition code, is usually aggregated as a contextual representation. The condition code is used to calculate channel-wise modulating weight for each object in the mask calibration process. Previous methods modulate the features \mathbf{f}_t of the current frame with features \mathbf{f}_r in reference frames according to the cues (e.g., appearance, location) of the specific object, but object-wise interactions have hardly been considered, resulting in failure cases where visually similar objects co-exist. We propose a discriminative object mask calibration procedure for mask modulation where information of various objects in intermediate decoding layers is exchanged with each other, thus enhancing multi-object discrimination.

Discriminative Object Calibration (DOC). To calibrate object mask \mathbf{m}^i to be discriminative from the other $\{\mathbf{m}^j\}_{j \neq i}$, a prerequisite is to make the condition code \mathbf{c}^i to be discriminative. Intuitively, we enhance the discrimination of \mathbf{c}^i by suppressing cues existing in the other objects, namely, by first aggregating cues from all $\{\mathbf{m}^j\}_{j=1}^N$ and then screening out similar cues in \mathbf{c}^i . The detailed operations are illustrated in Fig. 4 (a). The condition code \mathbf{c}^i is calibrated by taking the form:

$$\mathbf{c}^i = MLP([CL_1(\mathbf{m}_{in}^i); CL_2(\mathcal{A}(\{\mathbf{m}_{in}^j\}_{j=1}^N) - \mathbf{m}_{in}^i); CL_3(\mathbf{p}^i)]), \quad (5)$$

where $\mathcal{A}(\cdot)$ is an order-invariant aggregation layer for multiple inputs with a channel-wise pooling operator to aggregate object cues. Specifically, we use a channel-wise max pooling with a 1×1 conv in the implementation. CL_* denotes a Conditioning Layer block, which encodes a feature map into a vector ($H \times W$ to 1×1 in the spatial dimension). Detailed implementation is introduced

in Section 3.4. In the discriminative object calibrator, we utilize three conditioning layer blocks CL_1 , CL_2 and CL_3 to aggregate object information from the target object \mathbf{m}^i , the other objects in the current frame $\mathcal{A}(\{\mathbf{m}^j\}_{j=1}^N) - \mathbf{m}^i$ and target object in reference frames \mathbf{p}^i respectively.

Our motivation for this procedure is to incorporate discrepancy between different objects into object representation, thus reducing ambiguities between objects.

Conditional Decoder. We then utilize the more discriminative condition code \mathbf{c}^i to calibrate the object mask \mathbf{m}^i through the conditional decoder θ_{dec} . We build θ_{dec} based on a modulation block adopted in [63, 66] with an additional residual block. As is shown in Fig. 4 (b), given a proto-map \mathbf{m}^i and the condition code \mathbf{c}^i , the mask calibration is given by:

$$\mathbf{m}_{out}^i = \theta_{dec}(\mathbf{m}_{in}^i + \mathbf{m}_{in}^i \otimes \mathbf{c}^i). \quad (6)$$

We cascade N_L combinations of discriminative object calibrators and conditional decoders to progressively refine and up-sample the proto-map \mathbf{m}^i for each object i . Besides, the proto-maps are concatenated with the low-level features of the current frame from the backbone in the second-to-last conditional decoder, introducing more fine-grained pixel-level image cues for mask calibration. As is shown in Fig. 4 (c), we can find the proto-map evolves to be more discriminative, i.e., cues of other (even similar) objects are substantially suppressed, thus producing a more accurate object mask for the target object.

3.4 Network Details

Correlation-based Proto-map Generation. Given the overall adaptive object proxy map $\mathbf{p}^i = [\dots; \mathbf{p}_r^{i,K}; \dots]$ for each object i as discussed in Sec.3.2, we first calculate the similarity maps between the queries of the current feature map \mathbf{f}_t and each element $\mathbf{p}_r^{i,K}$ separately. Then, the proto-map \mathbf{m}^i is generated by translating the concatenation of all the similarity maps and current feature map \mathbf{f}_t . Formally,

$$\mathbf{m}^i = \theta_{ens}([\varphi_s([\dots; Sim(\mathbf{f}_t, \mathbf{p}_r^{i,K}); \dots]), \mathbf{f}_t]), \quad r \in \mathcal{S}_{ref}, K \in \mathcal{L}_{clu} \quad (7)$$

where $\theta_{ens}(\cdot)$ consists of the first two stages of bottleneck blocks in the ensembler of [63], φ_s is a 1×1 convolution layer to project the similarity maps in the channel dimension, and $Sim(\cdot, \cdot)$ is a L2-norm-based similarity function.

Conditioning Layer. Conditioning Layer (CL) is a block in the discriminative object calibrator to calculate the condition code \mathbf{c} from a feature map \mathbf{z}_{in} , taking the form:

$$CL(\mathbf{z}_{in}) = MLP(GAP(\mathbf{z}_{in} \odot \pi_\beta(\varphi(\mathbf{z}_{in})))), \quad \pi_\beta(x) = \begin{cases} x & x \geq \beta \\ 0 & x < \beta \end{cases} \quad (8)$$

where $\varphi(\cdot)$ represents a 1×1 convolution layer with ReLU activation to project the input \mathbf{z}_{in} to a confidence map, GAP denotes global average pooling. We incorporate a confidence gate π_β in the conditioning layer to filter out unreliable cues in the input feature map, where β is chosen as a percentile value in $\varphi(\mathbf{z}_{in})$.

4 EXPERIMENT

To evaluate the performance and robustness of our model, we conduct experiments on both standard VOS benchmarks and our constructed perturbed benchmark.

4.1 Standard Benchmark

Datasets. We evaluate our method on two widely-used multi-object VOS benchmarks, i.e., YouTube-VOS [60] and DAVIS17 [44]. The unseen object categories make YouTube-VOS a good benchmark to measure the generalization ability of various methods. Besides, the comparison between our method and other methods on a single-object VOS benchmark DAVIS16 [43] will also be included.

Metrics. We adopt the evaluation metrics from DAVIS [43], i.e., the region accuracy \mathcal{J} and boundary accuracy \mathcal{F} . \mathcal{J} measures the intersection-over-union (IoU) between the predicted masks and the ground-truth masks, and \mathcal{F} measures the accuracy of masks on the boundaries via bipartite matching between the boundary pixels. For both metrics, we will report the performance on seen and unseen categories as \mathcal{J}_s , \mathcal{J}_u and \mathcal{F}_s , \mathcal{F}_u respectively.

4.2 Pilot Robustness Benchmark

Perturbed datasets. For a type of perturbation $\epsilon \sim E$, we can generate a perturbed dataset $\mathcal{D}_\epsilon = \epsilon(\mathcal{D})$. Concretely, we established a pilot validation benchmark, namely *YouTube-VOS-P*, to evaluate VOS robustness against image perturbations based on the clean YouTube-VOS-2019 [60] validation set, which is the largest multi-object VOS dataset. In *YouTube-VOS-P*, we apply 6 types of perturbations on clean data for perturbed dataset construction. The perturbation set is constructed with noises [9] and blurring [31], which widely exist in real-world videos. Specifically, perturbation types include *Gaussian blur with 7×7 and 9×9 kernel*, *salt and pepper noise with 1k or 5k points*, and *Gaussian noise with mean of 0 and standard deviation of 10 or 30*. All perturbations are implemented with OpenCV [2].

Robustness metrics. Following robustness evaluation metrics commonly used in other tasks [14, 20, 21, 24, 48], we put forward two metrics for the evaluation of robustness against perturbation for a VOS model as follows.

After-perturbation accuracy (Q_p). After-perturbation accuracy Q_p represents the averaged remaining overall performance after perturbation. Specifically, given all the perturbation operation $\epsilon \sim E$, the after-perturbation accuracy is defined as

$$Q_p = \frac{1}{|E|} \sum_{\epsilon \sim E} Q_\epsilon. \quad (9)$$

Perturbation robustness (\mathcal{R}_p). Given the performance Q_c on the original clean dataset and the after-perturbation accuracy (Q_p) on the perturbed dataset, we can approximate the overall robustness to perturbation \mathcal{R}_p for a VOS model with the average overall performance drop, which can be formulated as

$$\mathcal{R}_p = Q_c - Q_p. \quad (10)$$

Here, smaller \mathcal{R}_p indicates better robustness for a VOS model.

4.3 Implementation Details.

We use SGD [1] with momentum 0.9 as the optimizer and use cross-entropy loss following our baseline setting [63]. For all the experiments, we set the batch size to 8. For training on YouTube-VOS, we do not use any external data. The total training iterations is 400k. We set the learning rate as 0.02 for the first half (200k) and 0.01

Methods				YouTube-VOS 2018 Validation					YouTube-VOS 2019 Validation				
	AF	MF	EXD	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}_s	\mathcal{J}_u	\mathcal{F}_s	\mathcal{F}_u	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}_s	\mathcal{J}_u	\mathcal{F}_s	\mathcal{F}_u
PReM [33]				66.9	71.4	56.5	75.9	63.7	-	-	-	-	-
CFBI [63]				81.4	81.1	75.3	85.8	83.4	81.0	80.6	75.2	85.1	83.0
CFBI+ [65]				82.8	81.8	77.1	86.6	85.6	82.6	81.7	86.2	77.1	85.2
AOT-B [64]			✓	83.2	82.6	77.3	87.4	85.6	83.3	82.5	77.8	87.0	86.0
Ours-Base				83.6	82.6	78.3	87.2	86.3	83.7	82.3	79.0	86.6	86.9
STM [39]		✓	✓	79.4	79.7	72.8	84.2	80.9	-	-	-	-	-
LCM [16]		✓	✓	82.0	82.2	75.7	86.7	83.4	-	-	-	-	-
MiVOS+km [6]		✓	✓	82.6	81.1	77.7	85.6	86.2	82.8	81.6	77.7	85.8	85.9
DMN-AOA [26]		✓	✓	82.7	82.6	76.7	87.0	84.8	-	-	-	-	-
STCN [7]	✓	✓	✓	83.0	81.9	77.9	86.5	85.7	82.7	81.1	78.2	85.4	85.9
JOINT [35]		✓		83.1	81.5	78.7	85.9	86.5	82.8	80.8	79.0	84.8	86.6
AOT-L [64]		✓	✓	83.7	82.5	77.9	87.5	86.7	83.6	82.2	78.3	86.9	86.9
Ours-MF		✓		84.0	82.7	78.8	87.4	87.1	84.1	82.7	79.4	86.9	87.2
CFBI ^{MS} [63]				82.7	82.2	76.9	86.8	85.0	82.4	81.8	76.9	86.1	84.8
CFBI+ ^{MS} [65]				83.3	82.8	77.3	87.5	85.7	-	-	-	-	-
Ours-Base^{MS}				84.4	83.2	79.3	87.8	87.3	84.4	82.7	80.0	87.1	87.8

Table 1: Quantitative comparison on YouTube-VOS [60]. AF denotes using All-Frames (30FPS) videos instead of default (6FPS) videos. MF denotes multiple historical frames are leveraged as guidance for current frame, otherwise only using the first and the previous frame. EXD denotes using external (static image transformation) data for training. MS denotes using multi-scale and flip testing in evaluation.

Reference frames	1&(t - 1) frames							Multiple frames										
	Method	OnAVOS* [50]	RGMP* [38]	FEEL [49]	PReM* [33]	CFBI [63]	AOT-B [64]	Ours -Base	Ours ^{FR} -Base	STM [39]	SST* [10]	MiVOS [6]	RMN [59]	LCM [16]	JOINT [35]	HMMN [47]	STCN [7]	Ours -MF
Davis16 Valid	$\mathcal{J} \& \mathcal{F}$	85.0	81.8	81.7	86.8	89.4	89.9	90.7	91.2	89.3	-	91.0	88.8	90.7	-	90.8	91.6	91.6
	\mathcal{J}	85.7	81.5	81.1	84.9	88.3	88.8	87.1	88.0	88.7	-	89.7	88.9	89.9	-	89.6	90.8	88.5
	\mathcal{F}	84.2	82.0	82.2	88.6	90.5	90.9	94.2	94.4	89.9	-	92.1	88.7	91.4	93.9	92.0	92.5	94.7
Davis17 Valid	$\mathcal{J} \& \mathcal{F}$	65.4	66.7	71.5	77.8	81.9	82.1	83.1	84.0	81.8	82.5	83.3	83.5	83.5	83.5	84.7	85.4	83.8
	\mathcal{J}	61.6	64.8	69.1	73.9	79.1	79.4	80.5	81.0	79.2	79.9	80.6	81.0	80.5	80.8	81.9	82.2	81.7
	\mathcal{F}	69.1	68.6	74.0	81.8	84.6	84.8	85.7	86.9	84.3	85.1	85.1	86.0	86.5	86.2	87.5	88.6	85.9
Davis17 Test-dev	$\mathcal{J} \& \mathcal{F}$	52.8	52.9	57.8	71.6	74.8	75.5	76.5	77.5	72.3	-	76.5	75.0	78.1	-	78.6	76.1	79.3
	\mathcal{J}	49.9	51.3	55.1	67.5	71.1	71.8	72.4	73.6	69.3	-	72.7	71.9	74.4	-	74.7	72.7	74.7
	\mathcal{F}	55.7	54.4	60.4	75.7	78.5	79.1	80.6	81.3	75.2	-	80.2	78.1	81.8	-	82.5	79.6	83.9

Table 2: Quantitative comparisons on DAVIS16[43] and DAVIS17[44]. FR denotes full-resolution testing, otherwise methods are tested on 480p. * denotes training with DAVIS only, otherwise with both DAVIS and YouTube-VOS.

for the rest. We use the DeepLabv3+ [4] architecture with ResNet-101 [13] as the backbone of our model. For multi-scale inference, we apply a scale set of [1.0, 1.15, 1.3, 1.5] as previous works [63, 65]. For training on DAVIS [43], we finetune the model pre-trained on YouTube-VOS for 40k iterations with learning rate of 0.1 and we use both DAVIS and YouTube-VOS datasets with a sampling ratio of 2:1. An NVIDIA Linux workstation (GPU: 8× Tesla V100) is used for our experiments. Our codebase is built on PyTorch 1.8.0 [41] and reuses some components implemented in [63]. We set multi-run clustering list $\mathcal{L}_{clu} = [1, 16, |H \times W|]$. The parameter β used in the conditioning layer as discussed in Sec. 3.4 is 0.3 through grid search. The number of cascaded mask decoding blocks N_L is set to 6.

We set two model variants for a fair comparison with previous methods: (1) **Ours-Base** is the default setting with the first and previous frame as reference, i.e., $\mathcal{S}_{ref} = \{1, t - 1\}$ as mentioned in Section 3.1; (2) **Ours-MF** uses multiple historical frames following [39, 64], i.e., $\mathcal{S}_{ref} = \{1, 1 + \delta, 1 + 2\delta, 1 + 3\delta, \dots\}$, $\delta = 5$.

4.4 Result on Standard VOS Benchmark

Quantitative comparison. The comparison between our method and other state-of-the-art methods on YouTube-VOS 2018 and YouTube-VOS 2019 validation set is shown in Table 1, which shows our model outperforms all the previous SOTA methods even without using any external data. Our method stands out under various evaluation metrics, especially on unseen categories, which demonstrates the great generalization ability of our method. The comparison on DAVIS is provided in Table 2. Our model also achieves the best performance on the challenging DAVIS17 test-dev split.

Qualitative comparison. Fig. 5 shows the qualitative comparison between previous SOTA methods and our model (Ours-Base) under various hard cases, and our method performs well for all these cases. The accuracy changes of different methods over time on YouTube-VOS 2019 validation set are illustrated in Fig 6. Our method has the least performance decay, demonstrating better robustness to suppress error propagation.

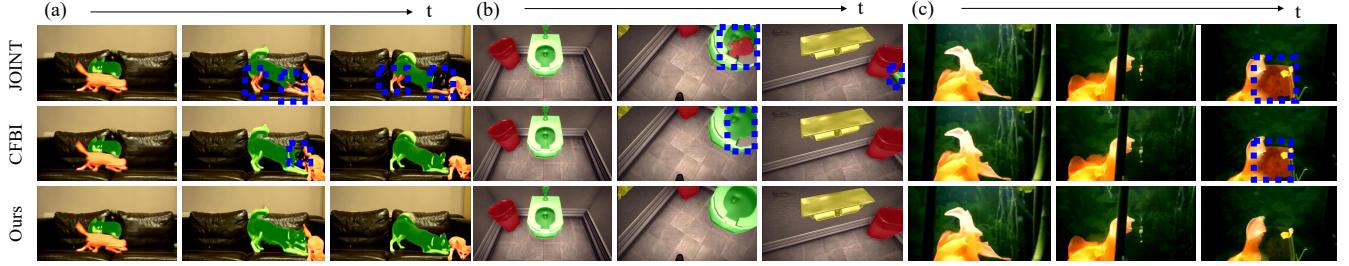


Figure 5: Qualitative comparison to competitive methods, JOINT and CFBI on YouTube-VOS 19 validation set. With the proposed adaptive proxy representation and object mask calibration for mask decoding, our model can tackle cases such as (a) object occlusion, (b) large camera rotation, and (c) fast motion better. Error regions are highlighted with blue bounding boxes.

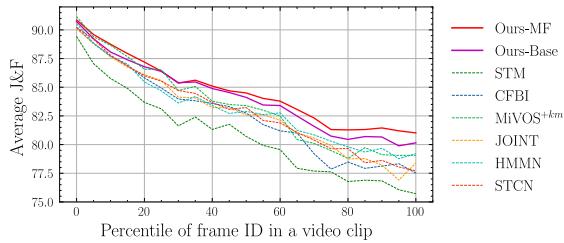


Figure 6: Comparison of average overall performance (\mathcal{J} & \mathcal{F}) on YouTube-VOS19 over time. 0(%) and 100(%) represents the beginning and the end of a video clip respectively. Performance on all the videos are normalized to the same length. Ours has the least performance decay.

Reference frames	1&(t - 1) frames			Multiple frames		
	Ours -Base	CFBI [63]	AOT-B [64]	Ours -MF	AOT-L [64]	STCN [7]
Clean accuracy $Q_c \uparrow$	83.7	81.0	83.3	84.1	83.6	82.7
+Gaussian noise ($\sigma = 10$)	83.1	80.5	82.7	83.4	82.8	80.8
+Gaussian noise ($\sigma = 30$)	80.8	76.6	77.0	81.1	77.6	78.6
+Salt&pepper noise (1k)	83.3	80.0	83.2	83.8	83.4	80.1
+Salt&pepper noise (5k)	82.2	79.1	82.0	82.5	81.4	78.0
+Gaussian blur (7 \times 7)	82.7	80.4	82.8	83.0	82.8	80.9
+Gaussian blur (9 \times 9)	82.0	79.9	81.7	82.4	82.2	79.9
After-perturbation accuracy $Q_p \uparrow$	82.3	79.4	81.6	82.7	81.7	79.7
Perturbation robustness $R_p \downarrow$	1.4	1.6	1.7	1.4	1.9	3.0

Table 3: Pilot study of perturbation robustness for VOS models on the perturbed dataset YouTube-VOS-P. We also list the official results on YouTube-VOS (Clean) for reference. Results are reported with \mathcal{J} & \mathcal{F} .

4.5 Analysis on Robustness to Perturbation.

In our pilot study on VOS model robustness, the proposed perturbed VOS validation dataset, *YouTube-VOS-P*, is used for robustness evaluation, as introduced in Sec.4.2. Note that all the models to be evaluated are trained with clean datasets. The results under random perturbations for AOT-B [64], AOT-L [64], and STCN [7] (without using external BL30K dataset[6]) are evaluated with their official model checkpoints while CFBI [63] is retrained by us and

the retrained result on the original clean dataset matches the performance reported in their paper. Experiments are averaged for 3 runs for validity. The results are summarized in Table 3 and we provide the insights and discussions as follows.

Are current VOS models robust to image perturbations? No. We notice a clear performance drop for all the VOS models investigated. Noticeably, these models are vulnerable to attack even when the input is injected with simple random Gaussian noise. Meanwhile, we can notice that the after-perturbation performance will drop largely as the severity of perturbations increases.

Does a model with higher performance on clean benchmarks guarantee better robustness against perturbations? Not sure. When making a comparison across various models, the correlation of the average overall performance (Q_c) on the original clean YouTube-VOS dataset and the after-perturbation accuracy (Q_p) is not clear. When making comparisons within a certain model, higher performance for one model also may not ensure stronger robustness because prediction errors and noises may be propagated during correlation calculation for VOS.

Can the proposed VOS framework help improve perturbation robustness? Yes! Our VOS framework with adaptive proxy aggregation and multi-object discrepancy discrimination mechanism stands out in the perturbed dataset setting, achieving higher after-perturbation accuracy Q_p (82.3 V.S. 79.4 in \mathcal{J} & \mathcal{F}) and better perturbation robustness R_p (1.4 V.S. 1.6 in \mathcal{J} & \mathcal{F}) than CFBI which is our baseline model. Though our baseline is strong enough to well handle some cases with multiple similar objects on the original clean datasets, the prediction of our model is more stable and consistent under perturbations, as is shown in Fig. 7.

4.6 Ablation Study

We make a thorough analysis of various components used in our method. All the ablation studies are based on our default setting (**Ours-Base**) as mentioned in Section 3.1 and conducted on the YouTube-VOS 2019 validation set.

Component effectiveness study. Table 4 demonstrates the effectiveness of the proposed adaptive object proxy (AOP) representation and discriminative object calibrator (DOC) for mask decoding. The first row shows the result of our baseline method [63] which adopts pixel-level correlation and uses modulation during mask decoding.

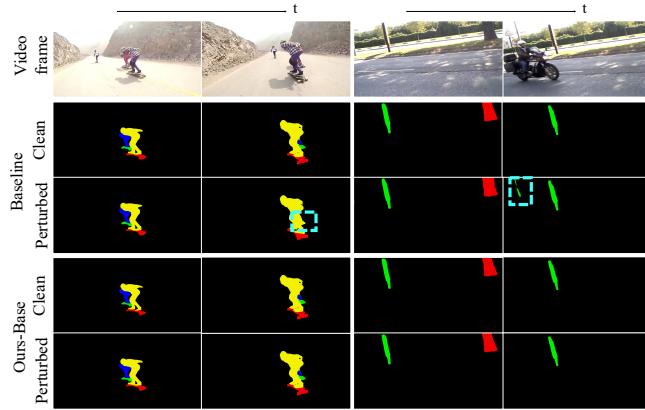


Figure 7: Qualitative comparison between our model (Ours-Base) and the baseline model [63] on robustness against perturbations. Two models performs well on original clean videos, but our model outperforms the baseline on perturbed videos with Gaussian Noise.

Model	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}_s	\mathcal{J}_u	\mathcal{F}_s	\mathcal{F}_u	FPS	Param
Baseline[63]	81.0	80.6	75.2	85.1	83.0	3.4	66.1
GSPR	82.4	81.6	77.4	85.8	85.0	3.3	66.1
AOP ($K=8$)	82.8	82.1	77.6	86.5	85.0	3.2	66.1
AOP ($K=16$, ours)	82.9	82.1	77.8	86.4	85.2	3.2	66.1
AOP ($K=32$)	82.8	82.2	77.6	86.4	85.1	3.2	66.1
AOP ($K=256$)	82.2	81.6	77.4	86.0	83.8	3.0	66.1
DOC ($\beta = 0$)	83.0	82.2	77.9	86.5	85.4	3.3	66.7
DOC ($\beta = 0.3$, ours)	83.4	82.1	78.4	86.5	86.3	3.3	66.7
Ours-Base	83.7	82.3	79.0	86.6	86.9	3.2	66.7

Table 4: Ablation study of the proposed adaptive object proxy (AOP) representation and discriminative object calibration (DOC). Here AOP and DOC denote models using AOP or DOC only. Grid-sampling-based proxy representation (GSPR) is for comparison with AOP. The inference time is reported in *multi-object FPS* as previous works [63, 64] and measured on a single V100 NVIDIA GPU with *batchsize* = 1. The number of model parameters is reported in MB.

Ablation on adaptive object proxy (AOP) representation. Compared to the commonly-used pixel-level object representation used in the baseline model, our adaptive proxy representation formed with help of clustering-based aggregation can significantly improve the overall performance, especially for unseen categories (+2.6% in \mathcal{J}_u and +2.2% in \mathcal{F}_u). Meanwhile, our adaptive proxy representation also performs better than simply aggregating the features grid-to-grid to form grid-sampled proxies (GSPR). Fig. 8 illustrates the segments constructed from K-Means clustering algorithm, as we can observe, each segment corresponds to a meaningful semantic part of an object, which can be aggregated to robust proxies for further correlation calculation across frames.

Ablation on discriminative object calibrator (DOC) for mask decoding. The multi-object discrimination mechanism help boost

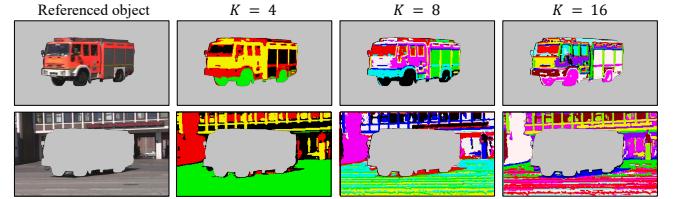


Figure 8: Visualization of clustering segments of different proxy number K for the construction of adaptive proxy representation. Semantically similar pixels in the foreground or background are grouped into semantic parts, such as wheels of the car and windows in the background.

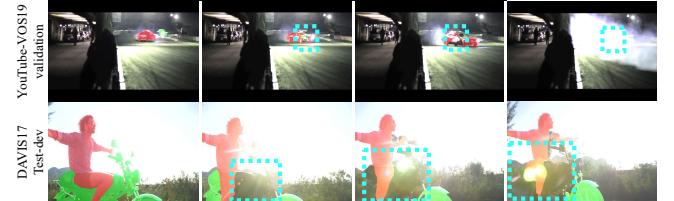


Figure 9: Limitation. We show two failure cases with strong natural perturbations, including the smoke (top) and the blur caused by a strong halo (bottom), of our model.

the performance from 82.9% to 83.7% given the model with adaptive proxy representation. Meanwhile, the confidence gate of the conditioning layer in DOC can also help improve the overall performance (+0.4% in $\mathcal{J} \& \mathcal{F}$) compared to the setting when disabling the confidence gate ($\beta=0$), which is owing to its de-noising mechanism to filter unreliable cues.

Complexity analysis. As is shown in the running time and model size ablation study in Table 4, compared to our baseline, our model achieves much better overall performance (83.7% V.S. 81.0% in $\mathcal{J} \& \mathcal{F}$) with a negligible cost of inference speed (-0.2 multi-object FPS) and model parameters (+0.6 MB).

5 CONCLUSION

Our pilot study on model robustness to perturbations for VOS reveals the fragility of current advanced methods. Towards robust VOS, we propose an end-to-end network with two tightly coupled modules to generate adaptive proxy representation and perform object mask calibration with multi-object discrimination considered. The two modules are demonstrated to help achieve a significant gain on standard VOS benchmarks and better robustness against perturbations compared to the baseline method.

Limitation and discussion. Though the performance of our adaptive proxy aggregation with non-parametric K-Means clustering is not sensitive to the number of clusters, we consider it is better to design a learning-based way to generate adaptive parts. Moreover, as perturbations in natural videos are more diverse, such as the smoke and the severe blur shown in Fig. 9, perturbation types can be extended to more diverse natural corruptions and adversarial attacks. For the robustness enhancement of VOS models, the data organization part can be further explored.

REFERENCES

- [1] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 177–186.
- [2] Gary Bradski and Adrian Kaehler. 2008. *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.".
- [3] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. 2017. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 221–230.
- [4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.
- [5] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. 2018. Blazingly fast video object segmentation with pixel-wise metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1189–1198.
- [6] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. 2021. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5559–5568.
- [7] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. 2021. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems* 34 (2021).
- [8] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. 2017. Segflow: Join learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*. 686–695.
- [9] Charles-Alban Deledalle, Loïc Denis, and Florence Tupin. 2012. How to compare noisy patches? Patch similarity beyond Gaussian noise. *International journal of computer vision* 99, 1 (2012), 86–102.
- [10] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. 2021. SSTVOS: Sparse spatiotemporal transformers for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5912–5921.
- [11] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2016. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629* (2016).
- [12] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. 2018. Generalisation in humans and deep neural networks. *Advances in neural information processing systems* 31 (2018).
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [14] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (2019).
- [15] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7132–7141.
- [16] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. 2021. Learning Position and Target Consistency for Memory-based Video Object Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4144–4154.
- [17] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. 2018. Videomatch: Matching based video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 54–70.
- [18] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.
- [19] Zhaoyang Jia, Han Fang, and Weiming Zhang. 2021. *MBRS: Enhancing Robustness of DNN-Based Watermarking by Mini-Batch of Real and Simulated JPEG Compression*. Association for Computing Machinery, New York, NY, USA, 41–49.
- [20] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is BERT Really Robust? Natural Language Attacks on Text Classification and Entailment. *arXiv preprint arXiv:1907.11932* (2019).
- [21] Christoph Kamann and Carsten Rother. 2020. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8828–8838.
- [22] Christoph Kamann and Carsten Rother. 2020;2021;. Benchmarking the Robustness of Semantic Segmentation Models with Respect to Common Corruptions. *International journal of computer vision* 129, 2 (2020;2021;), 462–483.
- [23] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. 2019. Lucid data dreaming for video object segmentation. *International Journal of Computer Vision* 127, 9 (2019), 1175–1197.
- [24] Alfred Laugros, Alice Caplier, and Matthieu Ospici. 2019. Are adversarial robustness and common perturbation robustness independent attributes? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–8.
- [25] Shuxian Liang, Xu Shen, Jianqiang Huang, and Xian-Sheng Hua. 2021. Video Object Segmentation with Dynamic Memory Networks and Adaptive Object Alignment. *IEEE*, 8045–8054.
- [26] Shuxian Liang, Xu Shen, Jianqiang Huang, and Xian-Sheng Hua. 2021. Video Object Segmentation with Dynamic Memory Networks and Adaptive Object Alignment. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 8045–8054. <https://doi.org/10.1109/ICCV48922.2021.00796>
- [27] Yongqing Liang, Navid Jafari, Xing Luo, Qin Chen, Yanpeng Cao, and Xin Li. 2020. WaterNet: An adaptive matching pipeline for segmenting water with volatile appearance. *Computational Visual Media* (2020), 1–14.
- [28] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. 2020. Video Object Segmentation with Adaptive Feature Bank and Uncertain-Region Refinement. *Advances in Neural Information Processing Systems* 33 (2020).
- [29] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. 2003. The global k-means clustering algorithm. *Pattern recognition* 36, 2 (2003), 451–461.
- [30] Huajia Lin, Xiaojuan Qi, and Jiaya Jia. 2019. Agss-vos: Attention guided single-shot video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3949–3957.
- [31] Yu-Qi Liu, Xin Du, Hui-Liang Shen, and Shu-Jie Chen. 2020. Estimating generalized gaussian blur kernels for out-of-focus image deblurring. *IEEE Transactions on circuits and systems for video technology* 31, 3 (2020), 829–843.
- [32] Xinkai Lu, Wenguang Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. 2020. Video object segmentation with episodic graph memory networks. *arXiv preprint arXiv:2007.07020* (2020).
- [33] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. 2018. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*. Springer, 565–580.
- [34] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. 2016. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Proceedings of the Advances on Neural Information Processing Systems*. 4905–4913.
- [35] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. 2021. Joint Inductive and Transductive Learning for Video Object Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9670–9679.
- [36] Jan H. Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. 2017. On Detecting Adversarial Perturbations. (2017).
- [37] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. 2021. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4133–4143.
- [38] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. 2018. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7376–7385.
- [39] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. 2019. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9226–9235.
- [40] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2337–2346.
- [41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* (2019).
- [42] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. 2017. Learning video object segmentation from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2663–2672.
- [43] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 724–732.
- [44] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017).
- [45] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017).
- [46] Hongje Seong, Junhyuk Hyun, and Euntai Kim. 2020. Kernelized Memory Network for Video Object Segmentation. In *European Conference on Computer Vision*. Springer, 629–645.
- [47] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. 2021. Hierarchical Memory Matching Network for Video Object Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12889–12898.
- [48] Florian Tramer and Dan Boneh. 2019. Adversarial training and robustness for multiple perturbations. *Advances in Neural Information Processing Systems* 32 (2019).
- [49] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. 2019. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9481–9490.
- [50] Paul Voigtlaender and Bastian Leibe. 2017. Online Adaptation of Convolutional Neural Networks for Video Object Segmentation. In *Proceedings of the British*

- Machine Vision Conference (BMVC)*. Article 116, 13 pages.
- [51] Paul Voigtlaender, Jonathon Luiten, and Bastian Leibe. 2019. Boltvos: Box-level tracking for video object segmentation. *arXiv preprint arXiv:1904.04552* (2019).
 - [52] Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai. 2021. SwiftNet: Real-time Video Object Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1296–1305.
 - [53] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 606–615.
 - [54] Zilei Wang, Jiashi Feng, and Shuicheng Yan. 2014. Collaborative Linear Coding for Robust Image Classification. *International journal of computer vision* 114, 2-3 (2014), 322–333.
 - [55] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. 2019. Ranet: Ranking attention network for fast video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3978–3987.
 - [56] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. 2019. Towards Real-Time Multi-Object Tracking. (2019).
 - [57] Peisong Wen, Ruolin Yang, Qianqian Xu, Chen Qian, Qingming Huang, Runmin Cong, and Jianlou Si. 2020. *DMVOS: Discriminative Matching for Real-Time Video Object Segmentation*. Association for Computing Machinery, New York, NY, USA, 2048–2056. <https://doi.org/10.1145/3394171.3414035>
 - [58] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. 2019. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 501–509.
 - [59] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. 2021. Efficient Regional Memory Networks for Video Object Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1286–1295.
 - [60] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. 2018. Youtub-evos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 585–601.
 - [61] Yu-Syuan Xu, Tsu-Jui Fu, Hsuan-Kung Yang, and Chun-Yi Lee. 2018. Dynamic video segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6556–6565.
 - [62] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K. Katsaggelos. 2018. Efficient Video Object Segmentation via Network Modulation. *CVPR* (2018).
 - [63] Zongxin Yang, Yunchao Wei, and Yi Yang. 2020. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*. Springer, 332–348.
 - [64] Zongxin Yang, Yunchao Wei, and Yi Yang. 2021. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems* 34 (2021).
 - [65] Zongxin Yang, Yunchao Wei, and Yi Yang. 2021. Collaborative Video Object Segmentation by Multi-Scale Foreground-Background Integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
 - [66] Zongxin Yang, Linchao Zhu, Yu Wu, and Yi Yang. 2020. Gated channel transformation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11794–11803.
 - [67] Xiaohui Zeng, Renjie Liao, Li Gu, Yuwen Xiong, Sanja Fidler, and Raquel Urtasun. 2019. Dmm-net: Differentiable mask-matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3929–3938.
 - [68] Bao Zhang, Handong Zhao, and Xiaochun Cao. 2012. Video object segmentation with shortest path. In *Proceedings of the 20th ACM international conference on Multimedia*. 801–804.
 - [69] Kaihua Zhang, Long Wang, Dong Liu, Bo Liu, Qingshan Liu, and Zhu Li. 2020. Dual temporal memory network for efficient video object segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1515–1523.
 - [70] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. 2020. Polysemy deciphering network for human-object interaction detection. In *European Conference on Computer Vision*. Springer, 69–85.

A MORE QUALITATIVE RESULTS

More qualitative comparisons to the state-of-the-art models [7, 39, 63] and ablation studies on our two main components, our adaptive object proxy representation and discriminative object calibration, are presented in the video.¹

A.1 Qualitative Results on Video Clips with An Extremely Large Number of Objects

In Fig. A, we use our model (Ours-Base) to propagate the masks in two cases where more than 40 objects co-exist. We leverage the video clips from a video scene parsing dataset, VSPW[37], and use the mask of the first frame as the reference for mask propagation. Even though we only segment each video frame with the guidance from the first frame and the previous frame, we can segment most objects in such a crowded scene well. Notice that we infer the video clip in $480p$ and the given mask in the first frame is not accurate enough especially in the boundary, so it is hard for our model to segment some extremely tiny objects very precisely.

A.2 Qualitative Results on Extremely Long Video Clips

Considering that the standard large-scale VOS benchmarks, including DAVIS [43] and YouTube-VOS [60], are tailored for evaluation on short-term video clips, we further evaluate our proposed model (Ours-Base) on two long-term video clips (*blueboy* and *dressage*) from [28]. These videos contain more than 1000 video frames and include more diverse appearance changes. Specifically, the video clip *blueboy* contains 2406 frames and the video clip *dressage* contains 3589 frames. For the mask propagation in these two video sequences, We use our model (Ours-Base) which is trained only on the YouTube-VOS [60] and only leverages the guidance from the reference (first) and the previous frame. Even if our model (Ours-Base) only maintains a constant memory bank of reference object proxies from the first frame and the previous frame, it can handle these cases with an extremely large number of video clips and large appearance changes well. The qualitative results of the mask predictions of our model are illustrated in Fig. B and Fig.C. We sample 20 frames here for illustration.

A.3 More Qualitative Results on Robustness to Perturbations

We show more qualitative cases to show the performance and the robustness of state-of-the-art models, including CFBI [63] (our baseline), AOT [64], and STCN [7], and our model (Ours-Base and Ours-MF) when the input video clips are under perturbations. Concretely, we show three cases from YouTube-VOS [60] 2019-version dataset with the perturbation of Gaussian noise, salt and pepper noise, and Gaussian blur in Fig. D, Fig. E, and Fig. F respectively. Our model shows stronger robustness to perturbations compared to other previous strong competitors in these challenging cases.

A.4 More Cases for Adaptive Object Proxy Representation

We provide more cases to illustrate the adaptive object proxy representation constructed with K -means clustering in Fig. G. Each segment in Fig. G represents a cluster to be aggregated as an adaptive object proxy. For the segments generated via clustering, we can find that regions of similar objects, the same object, or the spatially neighboring areas are clustered into the same group, which decomposes a complex region into several semantically-alike parts. Such an adaptive representation constructs a meaningful candidate pool for a query to do retrieval and calculate correlations for prototypical mask initialization. Meanwhile, due to the de-noising property of representation aggregation, our adaptive object proxies are more robust than the widely-used pixel-level proxies.

A.5 More Qualitative Results on DAVIS

Fig. H shows two hard cases of the qualitative comparison to previous state-of-the-art methods (STM [39], CFBI [63] and MiVOS [6]) on the challenging DAVIS17 [45] test-dev split. In the first case, although our model produces some minor prediction errors in the intermediate frames, errors are reduced in subsequent predictions. The second case shows that our model can handle cases with similar objects or occluded objects.

B MORE IMPLEMENTATION DETAILS

For data augmentations, we apply flipping, scaling, and balanced random-crop the same as [63]. Specifically, the balanced random crop is the input size for the model is 465×465 .

For the training strategy, following [38, 63, 65], the proposed network leverages the sequential training strategy where a clip of consecutive frames is sampled in each iteration. Concretely, a batch of video clips is sampled in each turn. For each video clip, we randomly sample a frame as the reference frame and a continuous $N + 1$ frames as the previous frame, and the current frame sequence with N frames. For the prediction of the reference frame, we use the ground-truth segmentation mask of the previous frame as the previous mask. For the prediction of the following frames, we use the latest prediction as the previous mask.

For the construction of object proxy aggregation, we leverage K -means clustering algorithm for implementation. We do not apply specific strategies for the initialization. The model is insensitive to the clustering algorithm and simple K-means clustering can already provide good performance. The deviation of our final model for different initialization of clustering is small (about 0.1% on YouTube-VOS[60] 2019-validation set).

¹The video demo is available at <https://youtu.be/3F6n7tcwWkA>

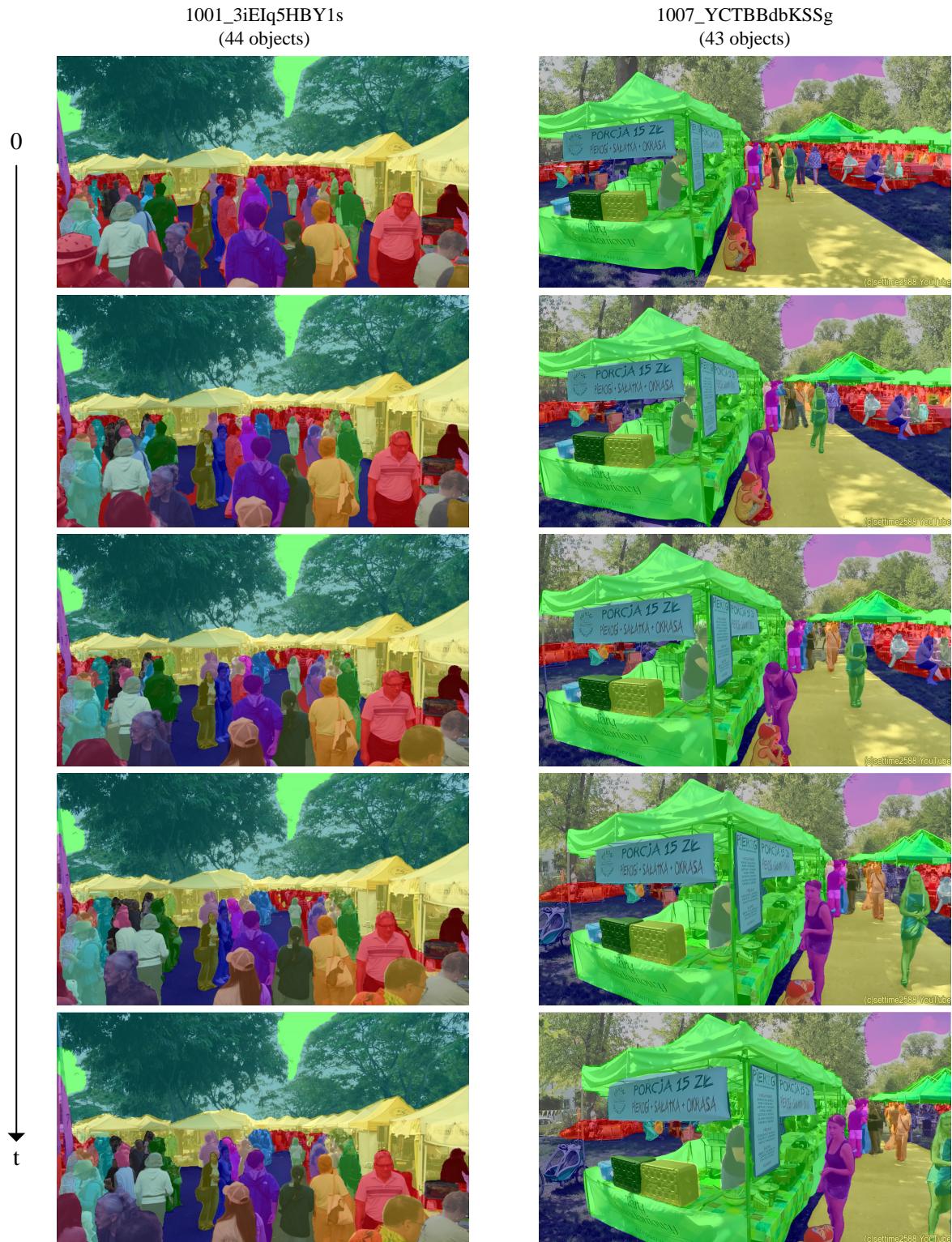


Figure A: We show two cases with an extremely large number of target objects to be segmented, including *1001_3iEIq5HBY1s* with 44 objects (left) and *1007_YCTBBdbKSSg* with 43 objects (right), from a video scene parsing dataset, VSPW [37]. Even if our model (Ours-Base) only leverages the guidance from the first and the previous frame, it can handle such cases with an extremely large number of objects well. Please zoom in on the figure to view it better.



Figure B: We show a long-video case (*dressage*) of *a man riding a horse* with 3589 frames from [28] and we sample 20 frames overtime here for illustration. The first frame with its mask (in the top left corner) is given as the reference for mask propagation, the propagated masks along the time are arranged from left to right and from top to bottom. Even if our model (Ours-Base) only leverages the guidance from the first and the previous frame, it can handle such cases with an extremely large number of video clips and large appearance changes well. Please zoom in on the figure to view it better.

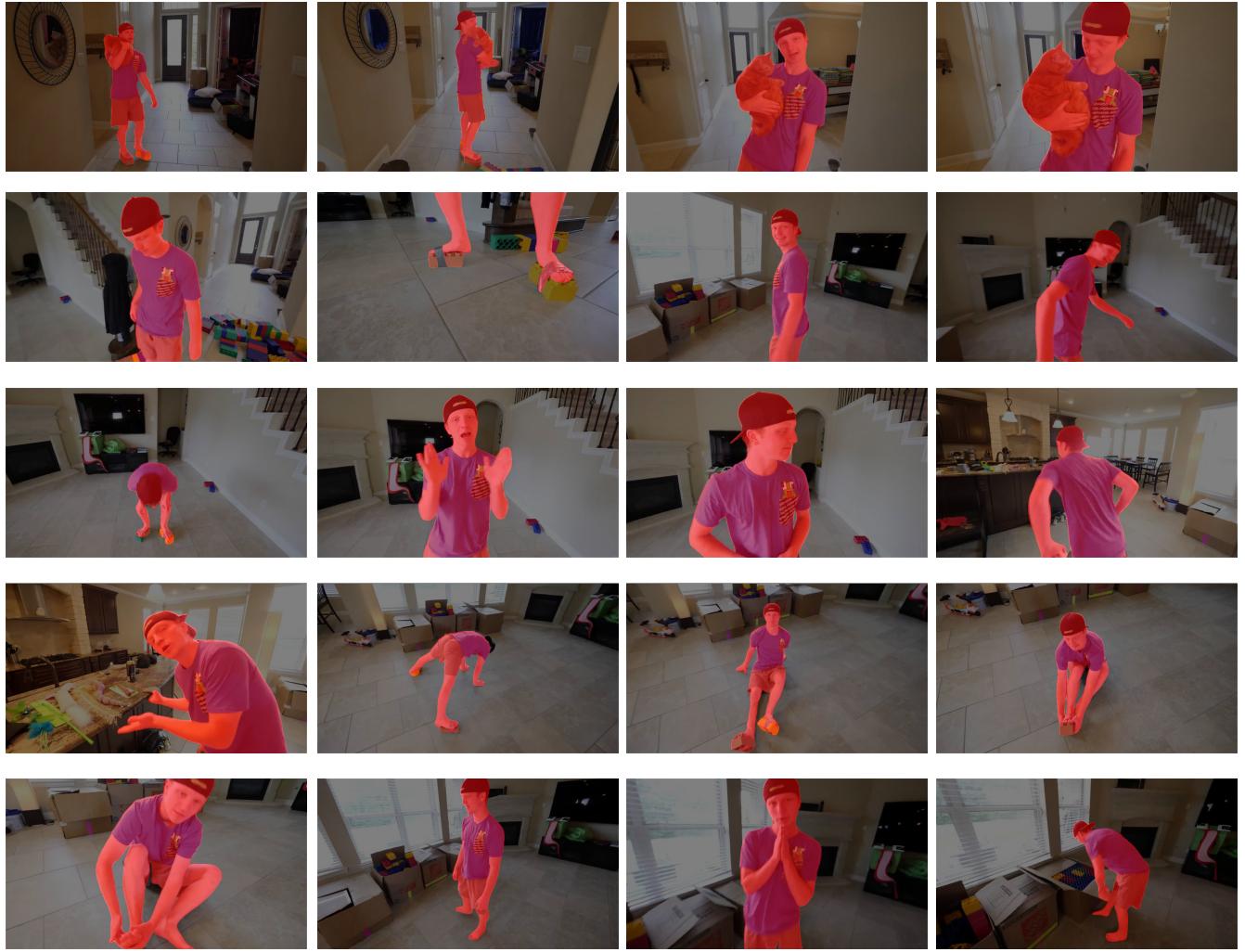


Figure C: We show a long-video case (*blueboy*) of a man in a blue shirt who does a lot of actions with 2406 frames from [28] and we sample 20 frames over time for illustration. The first frame with its mask (in the top left corner) is given as the reference for mask propagation, the propagated masks along the time are arranged from left to right and from top to bottom. Even if our model (Ours-Base) only leverages the guidance from the first and the previous frame, it can handle such cases with an extremely large number of video clips and large appearance changes well. Please zoom in on the figure to view it better.

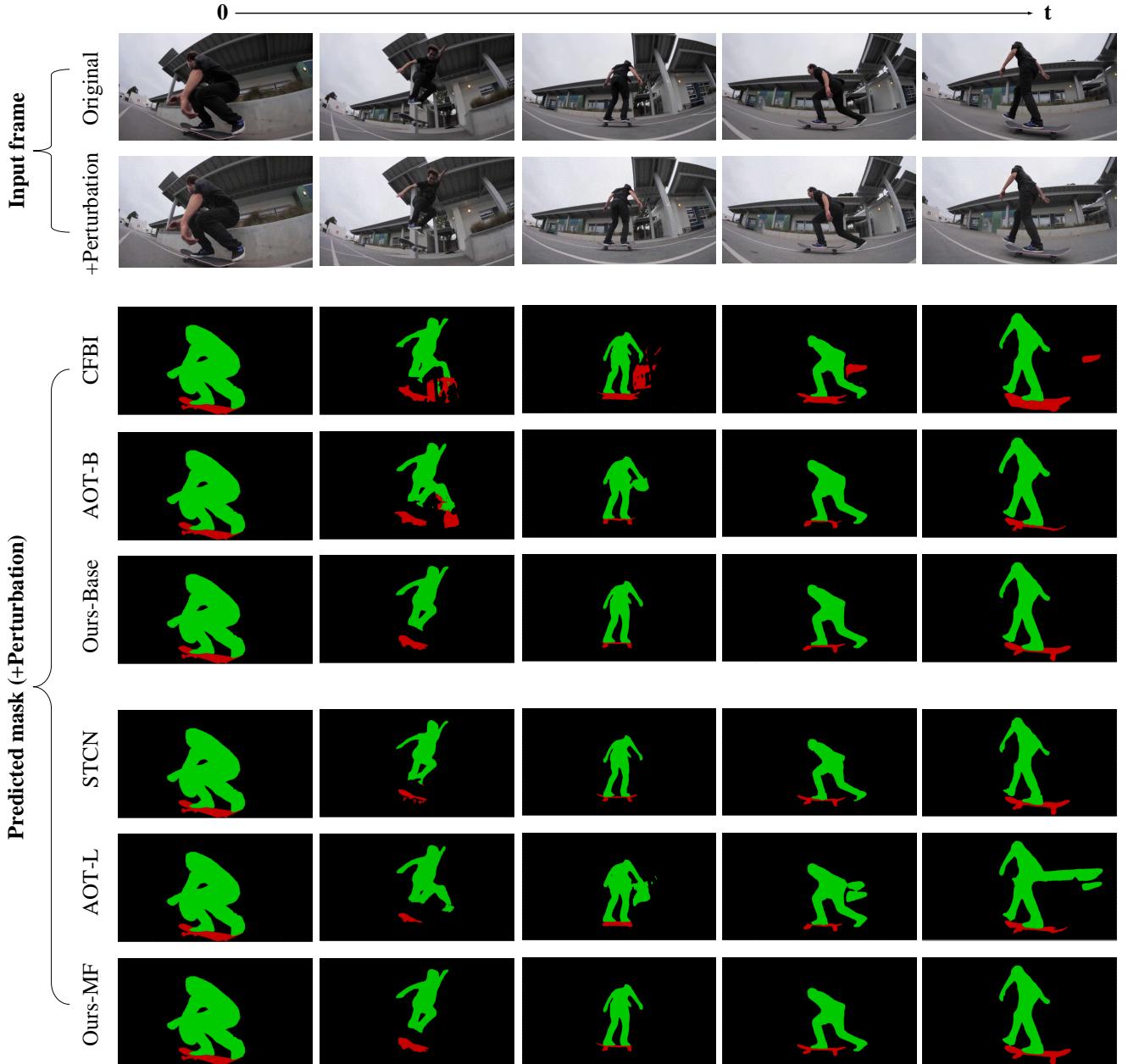


Figure D: A qualitative comparison between our model (Ours-Base, Ours-MF) and state-of-the-art models, including CFBI[63] (our baseline), AOT[64], and STCN[7], on robustness against perturbations. This video clip is challenging due to the fast motion of the man playing the skateboard. Here, Gaussian noise is injected into the video as the perturbation. Our baseline model, CFBI [63], can not handle such a case with fast motion and perturbations of Gaussian noise well. On the contrary, thanks to the proposed adaptive object proxy representation and discriminative object calibration, our model shows the best performance and robustness to perturbations among all the models.

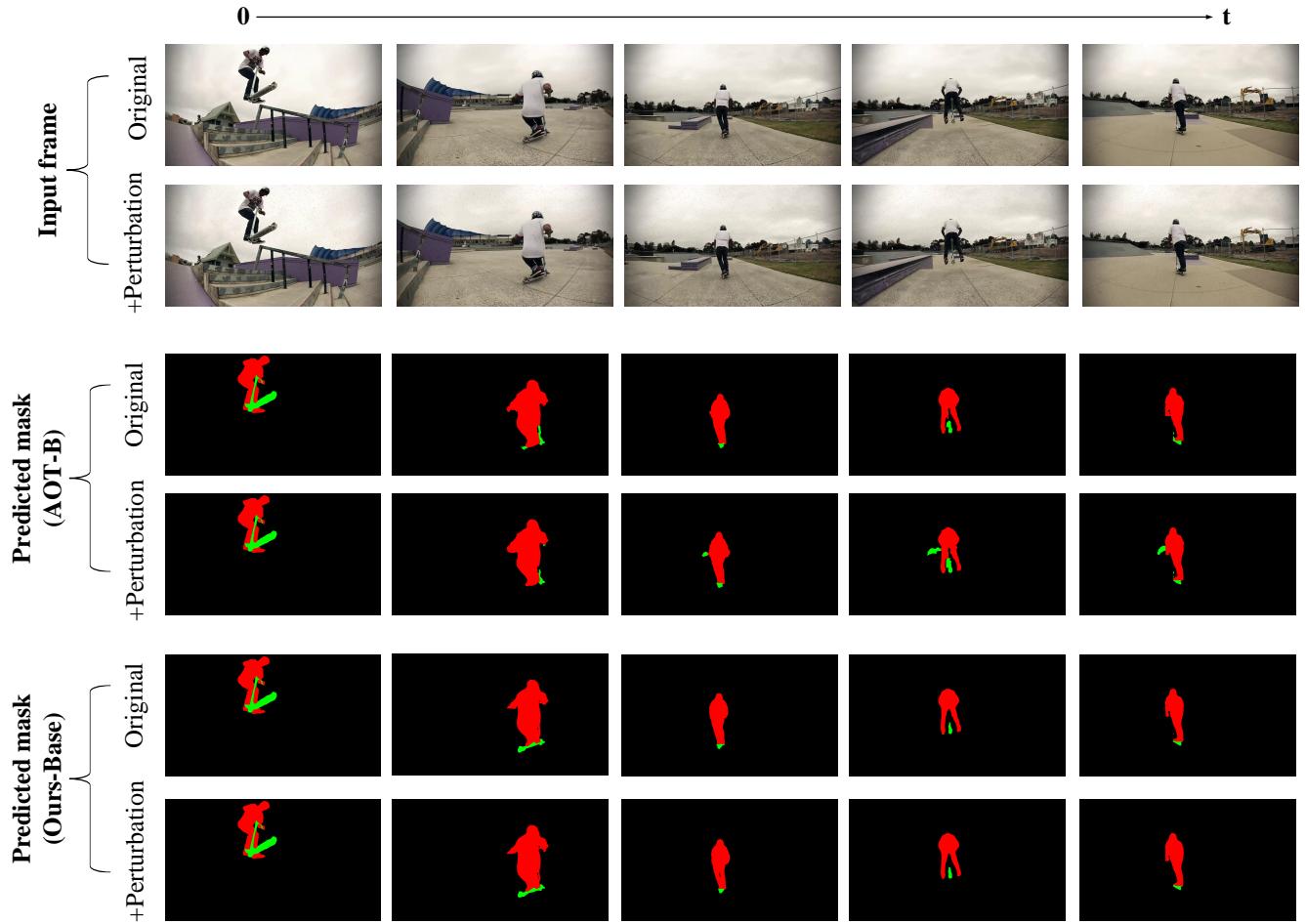


Figure E: A qualitative comparison between our model (Ours-Base) and a state-of-the-art model AOT-B [64] on robustness against perturbations of salt and pepper noise with 1k points. Both two models perform well on the original clean video clip. However, our model outperforms AOT-B on perturbed videos with noise injected.

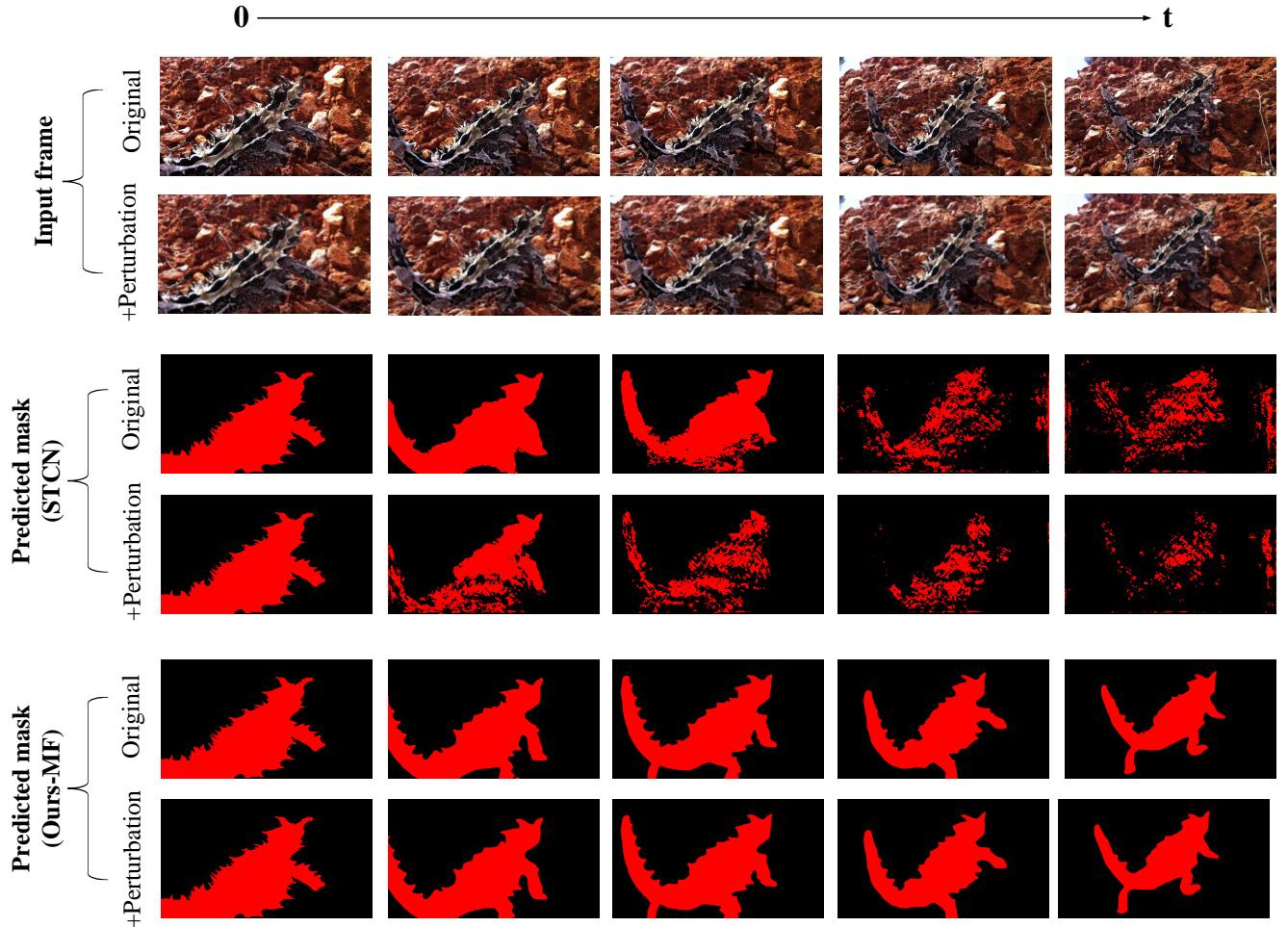


Figure F: A qualitative comparison between our model (Ours-MF) and a state-of-the-art model STCN [7] on robustness against perturbations of Gaussian blur with 9×9 kernel. In this video clip, the target object to be segmented is partially similar to the background in texture. For mask prediction given the original clean video clip, the performance of STCN will degrade and become more and more unstable over time while our model can tackle this case well. For the prediction when the video clip is under perturbation, the performance degradation of STCN goes more violent. Contrary, our model shows stronger robustness to perturbations and there is no drop in performance under perturbation in such a video case.

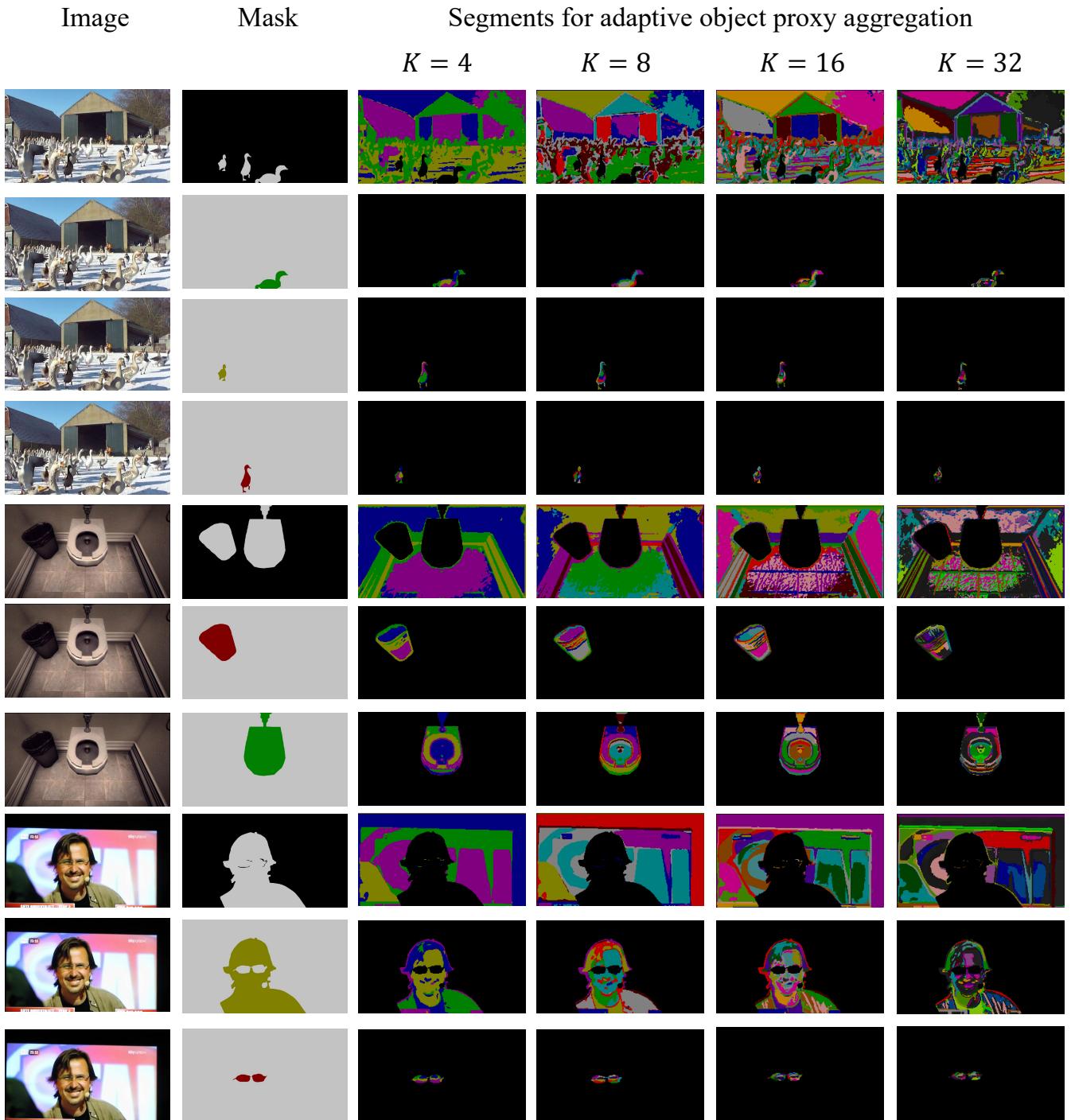


Figure G: Visualization of clustered K segments implemented with K -means clustering for adaptive object proxy aggregation. Each segment in this figure represents a cluster to be aggregated as an adaptive object proxy during adaptive object proxy aggregation. The adaptive object proxies are further used to construct correspondence with the current frame feature.

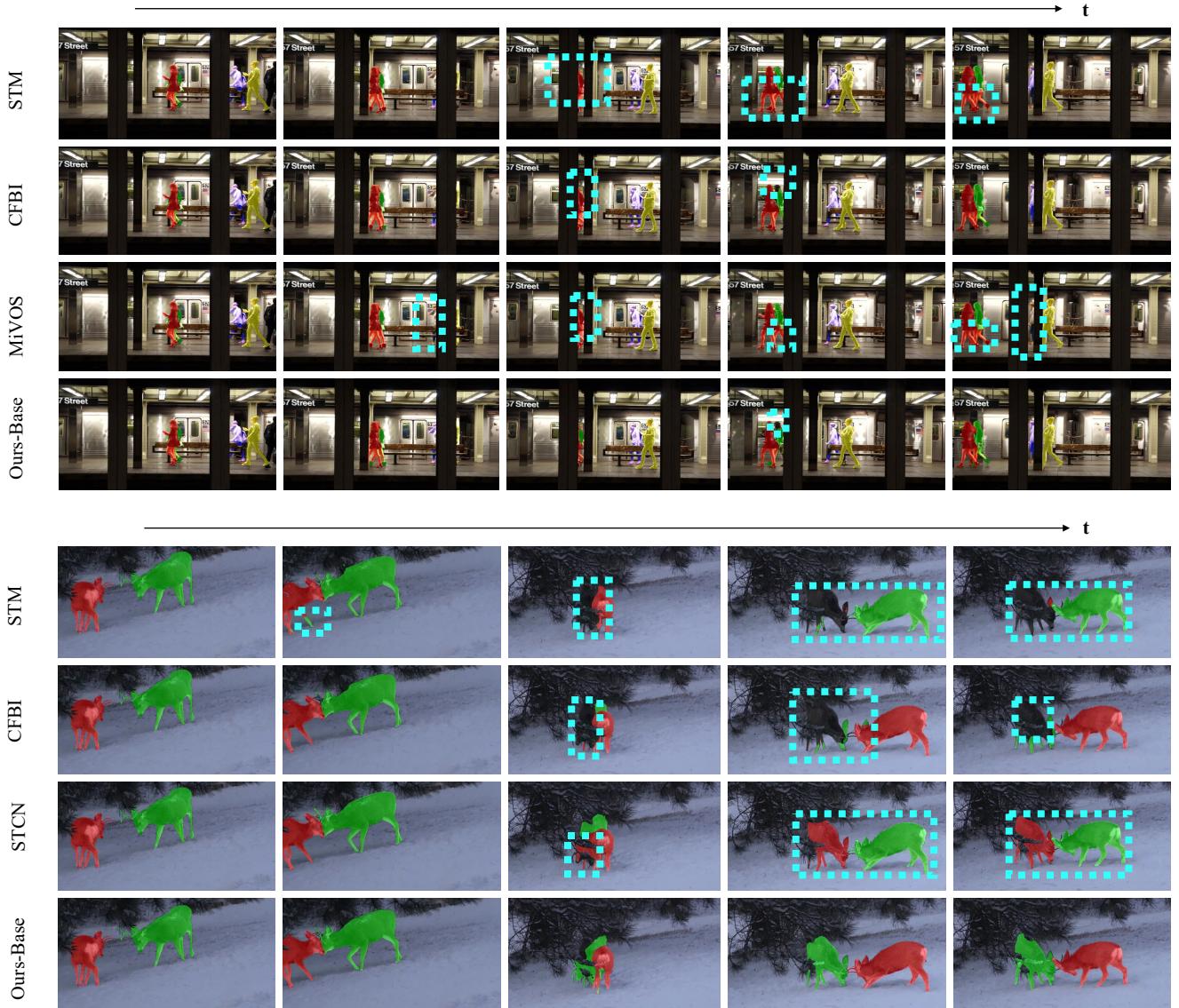


Figure H: Qualitative comparison to previous state-of-the-art methods, STCN [7], STM [39], CFBI [63] and MIVOS [6] on DAVIS17 [45] test-dev split. All of them are predicted with input resolution 480p. Our model (Ours-Base) shows stronger performance in these cases with multiple objects and object occlusions. Error regions are highlighted with light blue bounding boxes. Zoom in to view better.