Learning a Weakly-Supervised Video Actor-Action Segmentation Model with a Wise Selection

Jie Chen  Zhiheng Li  Jiebo Luo  Chenliang Xu
Department of Computer Science, University of Rochester
{jiechen, zhiheng.li, jiebo.luo, chenliang.xu}@rochester.edu

Abstract

We address weakly-supervised video actor-action segmentation (VAAS), which extends general video object segmentation (VOS) to additionally consider action labels of the actors. The most successful methods on VOS synthesize a pool of pseudo-annotations (PAs) and then refine them iteratively. However, they face challenges as to how to select from a massive amount of PAs high-quality ones, how to set an appropriate stop condition for weakly-supervised training, and how to initialize PAs pertaining to VAAS. To overcome these challenges, we propose a general Weakly-Supervised framework with a Wise Selection of training samples and model evaluation criterion (WS2). Instead of blindly trusting quality-inconsistent PAs, WS2 employs a learning-based selection to select effective PAs and a novel region integrity criterion as a stopping condition for weakly-supervised training. In addition, a 3D-Conv GCAM is devised to adapt to the VAAS task. Extensive experiments show that WS2 achieves state-of-the-art performance on both weakly-supervised VOS and VAAS tasks and is on par with the best fully-supervised method on VAAS.

1. Introduction

Video actor-action segmentation (VAAS) has recently received significant attention from the community [46, 45, 47, 14, 28, 13, 6]. Extended from general video object segmentation (VOS) which aims to segment out foreground objects, VAAS goes one step further by assigning an action label to the target actor. Spatial information within a single frame may be sufficient to infer the actors, but it alone can hardly distinguish the actions, e.g., running v.s. walking. VAAS requires spatiotemporal modeling of videos. A few existing works have addressed this problem using supervoxel-based CRF [45], two-stream branch [14, 13], Conv-LSTM integrated with 2D-/3D-FCN [28], 3D convolution involved Mask-RCNN [13], or under the guidance of a sentence instead of predefined actor-action pairs [6]. Although these fully-supervised models have shown promising results on the Actor-Action Dataset (A2D) [46], the scarcity of extensive pixel-level annotation prevents them from being applied to real-world applications.

We approach VAAS in the weakly-supervised setting where we only have access to video-level actor and action tags, such that model generalization is boosted by benefiting from abundant video data without fine-grained annotations. The only existing weakly-supervised method on A2D we are aware of is by Yan et al. [47]. Their method replaces the classifiers in [45] with ranking SVMs, but still uses CRF for the actual segmentation, which results in slow inference.

We consider weakly-supervised VOS, a more widely-studied problem. To fill the gap between full- and weak-supervision, a line of works first synthesize a pool...
of pseudo-annotations (PAs) and then refine them iteratively [5, 49, 21]. This synthesize-refine scheme is most related to our work but faces the following challenges:

Challenge 1: How to select from a massive amount of PAs high-quality ones? In general, PAs are determined by unsupervised object proposals [17, 39, 22], superpixel / supervoxel segmentations [1, 18], or saliency [32] inferred from low-level features. Hence, they can hardly handle challenging cases when there is background clutter, objects of multiple categories, or motion blur. The VOS performance is largely limited by the PA quality for models lacking a PA selection mechanism [53, 37], or simply relying on hand-crafted filtering rules [12, 19] that can barely generalize to broader cases. To tackle this challenge, we make a learning-based wise selection among massive PAs rather than blindly trusting the entire PA set. We will show that with only about 25%-35% of the full PAs, the selected PAs manage to provide more efficient and effective supervision to the segmentation network that outperforms the full-PA counterpart by 4.46% mean Intersection over Union (mIoU), a relative 22% improvement, on the test set (see Table 1). Note that there is another selection criterion in [20, 49, 50] with a focus on easy/hard samples, whereas ours is good/bad. They are quite different.

Challenge 2: How to select an appropriate stop condition for weakly-supervised training? In supervised training, it is safe to stop training upon the convergence of validation mIoU. However, it gets complicated when the obtained validation mIoU is no longer reliable when calculating against the PAs due to the complete absence of the real ground-truth. Fixing the number of training iterations is a simple yet brute solution [32, 37]. Instead, we propose a novel no-reference metric—region integrity criterion (RIC)—that does not blindly trust PAs and injects certain boundary constraints in the model evaluation. The convergence of RIC acts as the stop condition in training. Moreover, it turns out that the model with the highest RIC always produces better PAs of the next version than the model with the highest mIoU computed with PAs (see Table 2).

Challenge 3: How to initialize PAs when actions are considered along with actors? This is a question pertaining to VAAAS. Recent works in weakly-supervised image segmentation [44, 32, 19] and VOS [9] have shown that gradient-weighted class activation mapping (GCAM) [31] is capable of generating initial PAs from attention maps. However, GCAM is implemented with the network composed of 2D convolutions and trained on object labels; we denote this type of GCAM as 2D-Conv GCAM. Hence, it can only operate on video data frame-by-frame as on images. The spatiotemporal dynamics cannot be captured by 2D-Conv GCAM. Motivated by the success of 3D convolutions [3] in action recognition, we extend 2D-Conv GCAM to 3D-Conv GCAM to generate action-guided attention maps that are eventually converted to PAs with action labels.

In brief, we propose a general Weakly-Supervised framework with a Wise Selection of training samples and model evaluation criterion, instead of blindly trusting quality-inconsistent PAs. We thereby name it WS2, and Figure 1 depicts the framework. In Stage-1, attention maps are generated for each frame and subsequently refined to produce sharp-edged PAs for initializing a segmentation network. In Stage-2, we devise a simple but effective select-train-predict cycle to facilitate PA evolution. Upon a novel region integrity criterion, the performance of video segmentation is enhanced monotonically.

Customizing the above general-purposed WS2 to the specific VAAAS task is achieved by adding an action-guided attention map obtained by the proposed 3D-Conv GCAM, which has played a complementary role to its 2D counterpart in capturing the motion-discriminate part in the frames. For example, in the first two pairs of Figure 2, where a bird and an adult are walking, the 3D-Conv GCAM trained on action classification finds the area around legs most discriminative in identifying walking, regardless of the appearance difference between adult’s and bird’s legs, as the motion related to walking always resides on legs.

In summary, our contributions are as follows:

- We propose an effective two-stage framework WS2 for weakly-supervised VOS with a novel select-train-predict cycle and a new region integrity criterion to ensure a monotonic increase of segmentation capability.
- We customize WS2 with a 3D-Conv GCAM model to reliably locate the motion-discriminate parts in videos to generate PAs with action labels for the VAAAS task.
- Our model achieves state-of-the-art for weakly-supervised VOS on the YouTube-Object dataset [33, 11], as well as weakly-supervised VAAS on A2D, which is on par with the best fully-supervised model [13].

2. Related Work

In addition to the aforementioned weakly-supervised models of the synthesize-refine scheme for VOS or object detection [36, 51], we also summarize other non-refinement literature on weakly-supervised video object segmentation, as well as action localization.

VOS. Motion cue is a good source of knowledge. Using optical flow, Pathak et al. [26] group foreground pixels that move together into a single object, and set it as the segmentation mask to train a model. Similarly, the PAs in [49] are initialized from segmentation proposals using optical flow.
and fed into a self-paced fine-tuning network. Tokmakov et al. [37] propose to obtain the foreground appearance from motion segmentation and object appearance from the fully convolutional neural network trained with video labels only. The appearances are then combined to generate the final labels through graph-based inference. However, we try to avoid optical flow in our design due to its inability to handle large motion and brightness constancy constraint.

The use of non-parametric segmentation approaches is also common. For instance, mid-level super-pixels/voxels are extracted for weakly-supervised semantic segmentation [16] and human segmentation [21]. Tang et al. [34] enforce a Normalized Cut (NC) loss in weakly-supervised learning. Since it is accompanied by relatively strong supervision—scribbles, the model has already achieved a 85% full-supervised accuracy even without NC loss. Similarly, in [35], shallow regularizers, i.e., relaxations of MRF/CRF potentials, are integrated into the loss. Hence the model could do away with the explicit inference of PAs.

**Action localization.** Mettes et al. [23] introduce five cues, i.e., action proposal, object proposal, person detection, motion, and center bias, for action-aware PA generation and a correlation metric for automatically selecting and combining them. In contrast, our proposed 3D-Conv GCAM is much simpler by wrapping everything in a unified model.

### 3. WS$^2$ for Weakly-Supervised VOS

In this section, we illustrate how we design the two-stage WS$^2$ framework for weakly-supervised video object segmentation. The first stage provides the initial version of pixel-wise supervision on the full training set. The second stage continually improves PAs by iterations of select-train-predict cycles. In each cycle, a portion of more reliable PAs are selected to train a segmentation network, which, in turn, goes through an inference pass to predict a new version of PAs, and a new cycle starts all over again. The whole iteration stops when the highest RIC in each cycle is converged. The overall WS$^2$ approach is shown in Algorithm 1.

#### 3.1. Initial Pseudo-Annotation Generation

We first apply 2D-Conv GCAM [31] to the video frames to locate the most appearance-discriminate regions with a classification network trained on the object labels. Training frames are uniformly sampled over a video. The obtained attention map is subsequently converted to the binary mask $M_{init}$ using Otsu threshold [24], which produces the optimal threshold such that the intra-class variance is minimized.

Note that the attention maps calculated from the last convolutional layer are of low-resolution (typically of size 16x smaller than the input size using ResNet-50), the resultant $M_{init}$ is mostly a blob, which can hardly serve as qualified PAs to provide segmentation network with supervision that precisely pinpoints the borders of objects. This issue naturally suggests the use of simple linear iterative clustering (SLIC) [1], a fast low-level superpixel algorithm known for its ability to adhere to object boundaries well. We impose the $M_{init}$ on the SLIC segmentation map, thus treating $M_{init}$ as a selector of the superpixels $\{p_i\}$. The superpixel selection process is described in Algorithm 2.

The basic idea is to select superpixel $p_i$ with sufficient overlap with $M_{init}$ (line 3), meanwhile $p_i$ is not likely to be a background superpixel (line 4). Some overly-large foreground objects may be rejected by line 4, but there is a tradeoff between high recall and high precision. For $P_{A,v0}$, we aim to construct a more precise PA for the network to start with. Results in Figure 8 (I-R) show that our model manages to gradually delineate the entire body of large objects. Finally, the union of the selected superpixels constructs the refined mask $M_{refine}$.

Figure 3 shows how the superpixel selection process refines the initial blob-like mask: the false positive part (red) is removed, while the false negative part (green) is successfully retrieved. Such refinement imposes an effective

---

**Algorithm 1** WS$^2$ for weakly-supervised VOS

**Require:** weakly-labeled video frames $\{f_i\}$, trained classifier $\Phi$.

1: # Stage-1: Initial PA generation
2: for $f \in \{f_i\}$ do
3:  Generate attention map $S = \text{GCAM}(\Phi, f)$
4:  Generate initial mask $M_{init} = \text{Otsu}(S)$
5:  Generate refined mask $M_{refine} = \text{SLIC\_Refine}(M_{init})$
6:  $P_{A,v0} = \{M_{refine}\}$
7: # Stage-2: Iterative PA evolution
8: Set current version $i = 0$
9: do
10:  Select a subset of high-quality $P_{select}$ from $P_{A,v_i}$
11:  $RIC^\text{max} = 0$ $\triangleright$ The maximum $RIC$ achieved at $v_i$
12:  do
13:  Train model.v with $P_{select}$
14:  Evaluate model.v using $RIC$ $\triangleright$ for current epoch
15:  if $RIC > RIC^\text{max}$ then
16:  $RIC^\text{max} = RIC$
17:  while $RIC$ on the validation set not converge
18:  Produce new version $P_{A,v_i++}$ by model.v with $RIC^\text{max}$
19:  while $RIC^\text{max}$ on the validation set not converge
20: return model.v

---

**Algorithm 2** Mask refinement

**Require:** initial mask $M_{init}$, SLIC superpixels $\{p_i\}$, $\alpha, \beta$

1: $P_{select} = \emptyset$
2: for $p \in \{p_i\}$ do
3:  if $\text{IoU}(p, M_{init}) > \alpha$ then
4:  if $P_{p_{\text{area}}} = \frac{\text{Area}(p)}{\text{Area}(\text{frame})} < \beta$ then
5:     $P_{select} \cup p$
6: $M_{refine} = \bigcup P_{select}$
7: return $M_{refine}$
boundary constraint on PA, which is very critical to the dense (pixel-wise accurate) segmentation task.

### 3.2. Iterative PA Evolution

The quality of PA_v0 is inconsistent over the full training set, because some challenging cases can hardly be addressed in the initial PAs. To improve the overall quality of PAs, we design a select-train-predict mechanism. First, a subset of PAs is selected to train a segmentation network. Once the network is well-trained, it will make its predictions on the full training set as the new version of PA. The same select-train-predict procedure repeats iteratively until the RIC is converged.

**Selection criteria.** The PAs are recognized as of high quality if they either cover the entire object with a sharp boundary (strict criterion), or cover the most discriminate part of the object (relaxed criterion). Satisfying the relaxed criterion means that a classifier is easy to predict its type if only pseudo-annotated foreground part is visible. This leniency seems to risk taking inferior PAs to training samples as shown in Figure 4. However, these samples are still valuable, because they provide abundant training samples with precise localization. And, its inaccuracy can be remedied by the temporal consistency in video data, because by aggregating the information in the adjacent similar frames, the segmentation network could still learn to piece up the full body of the object despite of the noise in annotations. Taking the video clip in Figure 4 as an example, the missed arm in frame $t_5$ can be retrieved from the neighboring frame $t_4$.

To select the training samples using the above criteria, we employ two networks—a cut-and-paste patch discriminator and an object classifier, as shown in Figure 5. Inspired by [29], the samples qualified for the strict criterion will cover the whole object with a clear boundary. With such a mask, we can crop out the foreground object and paste it to another background region extracted from the same video, and the cut-and-paste patch still looks real to the binary discriminator. However, the samples matching the relaxed criterion are easily denied by the discriminator, so we add an object classifier to identify them. As long as the mask unveils a certain discriminative part of the object, it will send a strong signal to the object classifier and guide it to recognize its object category.

To prepare the inputs to these two networks, we first sample sets of foreground patches $\{p^i_{fg}\}$ and background patches $\{p^i_{bg}\}$ for each video. Foreground patches are squares enclosing the pseudo-annotated objects, and background patches are those not containing any pseudo-annotated objects (Background patches are mostly close to the frame boundary or from frames of scenic shots). Each foreground patch is coupled with a background patch of the same size. Note that they do not necessarily need to come from the same frame but need to be from the same video. It is particularly useful in close shots, where the foreground nearly occupies the full portion of the frame, or when there are multiple objects. In these cases, there is hardly enough space in the same frame for its paired background patch.

**Relation to Remez et al. [29].** Note that our framework is different from the image-based cut-and-paste model [29] in two ways. First, their weakly-supervised model is under the bounding-box level of supervision, whereas our task is of higher complexity with video-level labels. Second, they employ a GAN framework, where the generator tries to refine the input bounding-box to a tight mask under the guidance of GAN loss, whereas our model is an iterative evolution framework, in which the discriminator plays the role of a selector to pick high-quality PAs for training. There is no generator or adversarial loss involved in our framework, which eases the training process.
3.3. Region Integrity Criterion (RIC)

Without the supervision of real ground-truth, it is hard to evaluate the trained model properly. At the end of each select-train-predict cycle, if only mean Intersection over Union ($mIoU$) calculated using PAs is considered to evaluate the model on the validation set:

\[
mIoU_{PA} = mIoU(M_{\text{refine}}, \text{PAs}),
\]

where $M_{\text{refine}}$ is the refined network prediction, it may risk misleading the network to learn the noises in PAs as well. In the absence of ground-truth annotation for reference, we thereby introduce a new no-reference metric called Region Integrity Index ($\text{RII}$). This metric to some extent estimates how much the prediction has recovered the full body of the foreground objects from a low-level perspective. As shown in Figure 3, the initial masks can be refined by SLIC superpixels to fit the boundary of the object. If the difference of the masks before and after the refinement is minor, then it indicates that $M_{\text{init}}$ is already fairly precise. Therefore, we define $\text{RII}$ in a way that measures how close $M_{\text{init}}$ is to its refined version $M_{\text{refine}}$:

\[
\text{RII} = mIoU(M_{\text{init}}, M_{\text{refine}}),
\]

The trained models are thus evaluated with region integrity criterion ($\text{RIC}$) that combines $mIoU_{PA}$\(^1\) with $\text{RII}$:

\[
\text{RIC} = mIoU_{PA} + \alpha \ast \text{RII},
\]

where $\alpha = 0.5$ in our setting. Such design incorporates the boundary constraint in model evaluation that is necessary to avoid the blind trust in automated PAs.

Experiments show that at the turn of each evolution, new version PA generated by the highest RIC model is always superior to that by the highest $mIoU_{PA}$ model. Also, we stop the iterative PA evolution when the highest RIC for each version converges.

4. WS\(^2\) for Weakly-Supervised VAAS

In the typical VOS, each pixel is assigned an object label, whereas in VAAS it is an actor-action label. To adapt the VOS-oriented WS\(^2\) framework to VAAS, we add another branch in Stage-1 for the additional action label as shown in Figure 1. Hence, apart from the actor-guided attention map that is generated in the same way as in weakly-supervised VOS by a 2D-Conv GCAM, a 3D-Conv GCAM is proposed to generate the action-guided attention map. After binary thresholding, we take the union of the actor and action masks, $M_{\text{init}} = M_{\text{actor}} \cup M_{\text{action}}$, as the initial mask. Next, following the same steps in Section 3.1, we refine the blob-like mask $M_{\text{init}}$ with SLIC [1] to produce PA.v0.

\(^1\)To distinguish, $mIoU_{CT}$ is calculated based on the real ground-truth.

Figure 6. 3D-Conv GCAM. Take dog-running as an example, the 3D-Conv Network takes one video clip (consisting of $t$ frames) and the video-level action label, i.e., running, as inputs. During back-propagation, the gradients of all classes are set to zeros, except running to 1. In total, $t'$ action-guided attention maps corresponding to $t'$ frames uniformly sampled from the input clip are generated to estimate a sparse trajectory of the running dog.

To implement 3D-Conv GCAM for the given action label, we first obtain a well-trained action classification network denoted as 3D-Conv Model in Figure 6. Then, we conduct 3D-Conv GCAM to produce action-guided attention maps with the trained models.

Actor-action attention map generation. GCAM [31] is very popular in weakly-supervised learning [44, 9, 32, 19], as it can locate the most appearance-discriminate region purely using the classification network trained with the image-level label. We extend GCAM from 2D to 3D to produce the action-guided attention maps for VAAS.

As shown in Figure 6, the action attention map is calculated as the weighted average of the feature maps in the last convolutional layer. Specifically, for the target action class $c$, a one-hot vector $y^c$ is back-propagated to the feature maps $\{A_m\}$ of the last convolutional layer, the weight $w_m^c$ is the gradient with respect to the $m^{th}$ feature map $A_m$:

\[
\omega_m^c = \frac{1}{Z} \sum_i \sum_j \sum_k \frac{\partial y^c}{\partial A_m^k},
\]

where $Z$ is a normalization factor. Once the weights are obtained, the action-guided attention map $S^c_{\text{action}}$ can be calculated by:

\[
S^c_{\text{action}} = \text{ReLU}(\sum_m \omega_m^c A_m).
\]

Compared with 2D-Conv GCAM, each $A_m$ is a 3-dimensional feature map of size $(t', h', w')$ with an additional time dimension, thus the obtained $S^c_{\text{action}}$ is also of size $(t', h', w')$, which can be split into $t'$ attention maps $\{S^c_{\text{action}}\}^t'$. A non-trivial question is how to find the most critical $t'$ (out of $t$) input frames that stimulate the response in the action-guided attention maps. Our empirical findings suggest that a uniform sampling works the best.

Discriminate multiple instances. One benefit of using 3D-Conv GCAM as the initialization for weakly-supervised
VAAS is its ability to distinguish multiple instances. Unlike the instance definition in [53, 10], instances here may be the same-type actors doing different actions or vice versa. For some easier scenes that contain multiple different actors, we can set 1 to the interested actor type in the one-hot vector $y^c$, and localize it only with the actor-guided attention map. However, the actor-guided attention map cannot discriminate actors of the same actor type but performing different actions. As illustrated in Figure 7, showing a flock of birds on the beach, some are walking while others are flying. In this case, walking-guided and flying-guided action attention map will highlight different regions, which enables us to assign the action label to the corresponding actors.

In light of these observations, we further applied 3D-Conv GCAM to weakly-supervised spatial-temporal localization on AVA dataset in Section 5.4. It turns out that 3D-Conv GCAM shows great potential to focus on the object the person interacts with.

5. Experiments

In this section, we first present the quantitative and qualitative performance of the proposed VWS$^2$ on A2D for weakly-supervised VAAS and on YouTube-Object for weakly-supervised VOS. Then we apply the 3D-Conv GCAM to frame-level weakly-supervised action localization on a subset of the AVA dataset to demonstrate its appealing potential in person-object interaction detection.

5.1. Datasets

**A2D** [46] is an actor-action video segmentation dataset containing 3782 videos. Different from classic video object segmentation datasets [27, 2, 30, 11], A2D assigns actor-action to the mask, e.g., cat-eating. In total, 7 actors and 9 actions are involved. The dataset is quite challenging in terms of unconstrained video quality, action ambiguity, and multi-actor/action, etc. We split the 3036 training videos into two parts, 2748 for training and the rest for validation.

**YouTube-Object** [33, 11] consists of 5507 video shots that fall into 10 object classes. Among them, 126 video shots that have pixel-wise ground-truth annotation in every 10th frame [11] are used for testing, and the rest are for training following the same common setting in [43, 52, 38, 49]. AVA [7] is densely annotated with bounding boxes locating the performers with actions. Videos that fall in 10 classes with evident interactions and a balanced amount of training data are selected for weakly-supervised action localization. We denote it as AVA-10 hereafter.

5.2. Implementation Details

In general, weakly-supervised VOS and VAAS share the two-stage framework, except that the latter has an extra action recognition network in initial PA generation of Stage-1 to account for the action label.

**Initial PA generation.** For A2D, the 2D- and 3D-GCAM are implemented with ResNet-50 [8] pretrained on ImageNet [30] for actor classification, and inflated 3D ConvNet (I3D) [3] pretrained on Kinetics-400 [15] for action recognition. To finetune the two models on the A2D, 2794 videos with a single-actor label are used in training & validation set to train a ResNet-50, and 2639 videos with a single-action label are selected to train an I3D. Once they are well-trained—ResNet-50 achieves 87.74% accuracy on the single-actor test set, and I3D achieves 76.60% accuracy on the single-action test set—we apply the two classification networks to its respective GCAM settings for actor/action-guided attention map generation. Next, the binary attention masks are refined by SLIC with the thresholds set to $\alpha = 0.5$, $\beta = 0.4$. For YouTube-Object, we follow the similar procedure, except that only ResNet-50 is used for object classification and attention map generation.

**Iterative PA evolution.** To select a subset of high-quality PAs, a small network with five layers of Conv-LeakyRelu is constructed to discriminate original foreground patches from cut-and-paste patches. Note that the ResNet-50 trained in Stage-1 is directly used here in testing mode to predict the actor type for the cropped patches.

As for segmentation network, we choose DeepLab-v2 [4]. During training, the inputs to the network are patches of size $224 \times 224$ pixels randomly cropped from the frame. We use the “poly” learning rate policy as suggested by [4], with base learning rate set to $7 \times 10^{-4}$ and power to 0.9. We fix a minibatch size of 12 frames, momentum 0.9. In the testing, we output the full-size segmentation map for each frame. A simple action-alignment post-processing is used to unify the action label for the same actor, since frame-based segmentation network can hardly capture the temporal information throughout the

---

2The selected classes are fight/hit (a person), give/serve (an object) to (a person), ride, answer phone, smoke, eat, read, play musical instrument, drink, and write.
shows how the model’s prediction power

<table>
<thead>
<tr>
<th>Models</th>
<th>Settings</th>
<th>(mIoU_{GT}) (actor-action)</th>
<th>(mIoU_{PA})</th>
<th>(RIC)</th>
<th>(mIoU_{GT}) (a.-a./actor/action)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>full</td>
<td>24.62</td>
<td>20.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model-S</td>
<td>subset</td>
<td>27.65</td>
<td>24.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(WS^2)</td>
<td>subset</td>
<td>29.32</td>
<td>26.74</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Comparison of model variants with different settings in Stage-2. The settings specify whether the model is trained on PAs from the full training set or only the selected subset. And in each PA version upgrade, whether model with the highest validation \(mIoU_{PA}\) or \(RIC\) is selected to predict the next version of PA.

video, which may cause action-inconsistency in the same actor appearing in multiple frames. To tackle this issue, we take the poll of neighboring frames, and assign the action label with the maximum votes to the actor of interest. This procedure is similar to the effective temporal segment network [41], which belongs to the video-level action recognition, whereas ours is in instance-level.

Evaluation metrics. We use mean intersection over union (\(mIoU\)) averaged across all classes to evaluate the performance of the model. To compare with the weakly-supervised model [47] on A2D, we also adopt average per-class pixel accuracy (\(cls\_acc\)) and global pixel accuracy (\(glo\_acc\)) for evaluation.

5.3. Weakly-Supervised VAAS & VOS

We first investigate the effectiveness of the key components in iterative PA evolution on A2D. Then we compare our \(WS^2\) model with other state-of-the-art fully- and weakly-supervised methods on A2D and YouTube-Object. Results show that \(WS^2\) outperforms all video-level weakly-supervised models with the performance that is highly competitive even against the fully-supervised models.

5.3.1 Ablation Study

The iterative PA evolution is running in select-train-predict cycles, in which train is no more special than training a segmentation network as in the fully-supervised setting. The two key factors that influence how much PA can be improved iteration by iteration mainly reside in 1) the overall quality of the selected training samples compared with the original full set, and 2) the performance of the model chosen by \(RIC\) to predict the next version of PA, compared with that chosen by plain \(mIoU_{PA}\). To quantitatively evaluate their respective contribution to the final model, we conduct an ablation study with three model variants in Table 1.

The results show that, Model-S trained on the subset outperforms Baseline trained on the full PAs, because the selected training samples are of higher-quality. It is also verified in Table 2 with the training samples evaluated by the real ground-truth. The selected subsets always have higher \(mIoU_{GT}\) than the full set, which means the selected training samples tend to have more clear boundary and complete coverage of the full object. In comparison, there is more noise and inconsistency in the full set, which may confuse the model and impede it from converging. More importantly, models can be trained much more efficiently on the subset than the full set with 65%-75% less training frames.

Employing \(RIC\) rather than the plain \(mIoU_{PA}\) also helps us choose better models in each PA version upgrade. \(mIoU_{PA}\) calculated by noisy PAs is not guaranteed to assess the true performance of the model as in the fully-supervised setting. It is possible that models with high validation \(mIoU_{PA}\) may also produce noisy prediction that matches exactly the noise in PAs [49]. To overcome this problem, we propose \(RIC\) that considers both \(mIoU_{PA}\) and \(RIL\) (Eq. 3). \(RIL\) measures the shape change ratio in mask before and after the SLIC refinement. Since refinement drags the segmentation boundary closer to the real object’s boundary, if there is not much change on the original prediction after refinement (i.e., high \(RIL\)), then it is likely that the original prediction has already produced edge-preserving masks that approximate the ground-truth segmentation maps. Table 2 clearly exhibits that the-highest-\(RIC\) model performs better PAs of the next version than the-highest-\(mIoU_{PA}\) model.

5.3.2 Comparison with the State-of-the-Art Methods

A2D. We compare our weakly-supervised model with the state-of-the-art fully- and weakly-supervised models on the VAAS task. Table 3 indicates that our model evolves iteration by iteration, and eventually achieves about 72% performance of the best fully-supervised model [13], which is actually a two-stream method that makes use of optical flow for action recognition, whereas our model only takes RGB frames as input. To make a fair comparison with the only existing weakly-supervised model we know of on A2D, we report in Table 4 with the evaluation metric used in [47].

Figure 8 shows how the model’s prediction power
Table 3. Comparison with the state-of-the-art fully-supervised models on the A2D test set. The subscript denotes the performance percentage to the best fully-supervised model [13].

<table>
<thead>
<tr>
<th>Models</th>
<th>mIoU(_{GT}) (actor-action/actor/action)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPM+TSP [45]</td>
<td>19.9/53.9% / 33.4/50.3% / 32.0/69.1%</td>
</tr>
<tr>
<td>TSMT [14]+GBH</td>
<td>24.9/77.5% / 42.7/64.3% / 35.5/70.7%</td>
</tr>
<tr>
<td>TSMT [14]+SM</td>
<td>29.7/80.5% / 49.5/74.5% / 42.2/91.1%</td>
</tr>
<tr>
<td>DST-FCN [28]</td>
<td>33.4/90.5% / 47.4/71.4% / 45.9/99.1%</td>
</tr>
<tr>
<td>Gavrilyuk et al. [6]</td>
<td>34.8/94.3% / 53.3/80.9% / 49.4/106.7%</td>
</tr>
<tr>
<td>Ji et al. [13]</td>
<td>36.9/100% / 66.4/100% / 46.3/100%</td>
</tr>
</tbody>
</table>

Table 4. Comparison with the state-of-the-art weakly-supervised model on the A2D test set. cls\(_{acc}\) and glo\(_{acc}\) are shown in order of actor-action/actor/action.

<table>
<thead>
<tr>
<th>Models</th>
<th>cls(_{acc})</th>
<th>glo(_{acc})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yan et al. [47]</td>
<td>41.7 / - / -</td>
<td>81.7 / 83.1 / 83.8</td>
</tr>
<tr>
<td>Ours (model.v3)</td>
<td>43.06 / 49.16 / 35.12</td>
<td>87.10 / 91.30 / 87.44</td>
</tr>
</tbody>
</table>

Figure 8. Evolution of the model prediction on some tough test samples. Two samples are shown on each row (left→right): input frame, prediction by models from v0 to v3, ground-truth (GT). Although in the complete absence of pixel-wise annotation from the GT, our model can still handle challenge cases like occlusion (g-L, i-L, j-L), out of view (a-L, j-R), low illumination (c-L, b-R, e-R), small objects (d-R, h-R), blur (k-L), multitype-actors (f-L, k-R), fast motion (j-R), background clutter (a-R, g-R), etc.

evolves through versions. Especially for challenging cases like when only the adult’s upper body is observed in a-L, the model correctly predicts adult-crawling by seeing his arms erected on the floor. The case of a boy covering his head with a towel (g-L) really gives our model a hard time in the beginning, and it finally figures it out in model.v3. In other hard cases, such as background clutter, motion blur, low illumination, occlusion, out of view, and small scale, the model sometimes fails to output something reasonable in its early versions. Its ability gradually grows as the PA evolves, and finally it gets the prediction right. As for the less complicated cases, the model is able to catch the approximate location of the actors in its early versions, but the predicted masks may suffer from under-/over-segmentation, or wrong action label, which are self-corrected by later versions (see more examples in the supplementary material).

YouTube-Object. WS\(^2\) achieves promising segmentation results which outperform the previous video-level weakly-supervised methods as shown in Table 5. Qualitative results are given in the supplementary video.

5.4. Weakly-Supervised Action Localization

To further validate the ability of the proposed 3D-Conv GCAM in localizing motion-discriminate part in video, we apply it to weakly-supervised spatial-temporal action localization on AVA-10. We train an I3D [3] action classification network on AVA-10 with only frame-level supervision (without bounding-boxes). In the testing mode, I3D predicts an action label, with which 3D-Conv GCAM attends to the most relevant region. The visualization of the attention maps in the supplementary material indicates that 3D-Conv GCAM can accurately localize the object the person is interacting with, such as the cigar of a smoking person, or the hand of the person giving/serving something.

6. Conclusion

Given only video-level categorical labels, we tackle the weakly-supervised VOS and VAAS problems. A two-stage framework called WS\(^2\) is proposed to overcome common challenges faced by many synthesize-refine scheme-based methods that are most successful in weakly-supervised VOS. Our proposed select-train-predict cycle utilizes a different cut-and-paste model than [29] to effectively select high-quality PAs and is customized to handle videos. The new region integrity criterion (RIC) is proposed to better guide the convergence of training in the absence of ground-truth segmentation. Extensive experiments on A2D and YouTube-Object show that WS\(^2\) performs the best among weakly-supervised methods. Our proposed framework and techniques are general and can be used for other weakly-supervised video segmentation problems.

Acknowledgments. This work was supported in part by NSF 1741472, 1813709, and 1909912. The article solely reflects the opinions and conclusions of its authors but not the funding agents.
References


[20] Siyang Li, Xiangxin Zhu, Qin Huang, Hao Xu, and C-C Jay Kuo. Multiple instance curriculum learning for weakly supervised object detection. 2017. 2


