

# We don't Need Thousand Proposals: Single Shot Actor-Action Detection in Videos

Aayush J Rana

aayushjr@knights.ucf.edu

Yogesh S Rawat

yogesh@crcv.ucf.edu

Center for Research in Computer Vision  
University of Central Florida  
Orlando, FL, 32816

## Abstract

We propose SSA2D, a simple yet effective end-to-end deep network for actor-action detection in videos. The existing methods take a top-down approach based on region-proposals (RPN), where the action is estimated based on the detected proposals followed by post-processing such as non-maximal suppression. While effective in terms of performance, these methods pose limitations in scalability for dense video scenes with a high memory requirement for thousands of proposals. We propose to solve this problem from a different perspective where we don't need any proposals. SSA2D is a unified network, which performs pixel level joint actor-action detection in a single-shot, where every pixel of the detected actor is assigned an action label. SSA2D has two main advantages: 1) It is a fully convolutional network which does not require any proposals and post-processing making it memory as well as time efficient, 2) It is easily scalable to dense video scenes as its memory requirement is independent of the number of actors present in the scene. We evaluate the proposed method on the Actor-Action dataset (A2D) and Video Object Relation (VidOR) dataset, demonstrating its effectiveness in multiple actors and action detection in a video. SSA2D is 11x faster during inference with comparable (sometimes better) performance and fewer network parameters when compared with the prior works. Code available at <https://github.com/aayushjr/ssa2d>

## 1. Introduction

Actor-action detection in videos is a challenging problem where the goal is to detect all the actors in the video and determine which different actions they are performing. One natural solution to this problem is to perform object detection and identify all the actors and classify those detected actors for corresponding actions. Motivated by this,

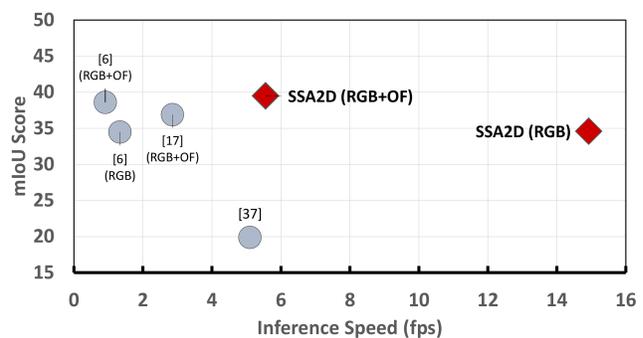


Figure 1. A comparative analysis of SSA2D with existing methods in terms of speed and performance. We observe that SSA2D is faster with comparable performance. The x-axis represents inference speed in frames per second and the y-axis represents the mean pixel-wise intersection over union score for joint actor-action detection. (RGB+OF - Using both RGB and optical flow).

existing methods utilize region proposal (RPN) [29] based approach [6, 17], where they first detect proposal for objects and use them for action detection.

Region proposal and ROI pooling based methods [29, 10] have been shown to be extremely successful for object detection in images. The approach usually involves detection of thousands of proposals followed by a post-processing cleanup using non-maximal suppression. In videos, we have an additional time dimension, and detecting proposals for every frame can be *memory intensive*. Therefore, the existing methods [6, 17] relying upon RPN approach can perform actor-action detection only on *single frame* at a time. Also, with multiple actors and multiple actions per actor in a scene, these approaches become complex and inefficient as thousands of region proposals per frame are required for training. Due to these limitations, such networks have to be trained in multiple stages [6], leading to an increase in the training and inference time.

Single shot detection methods [23, 33, 27, 28] overcome such region proposal limitations, but they don't translate well for video actor-action detection.

In this work, we present an alternative approach which is simple yet effective and overcomes these limitations. We propose SSA2D, a novel end-to-end deep network which does not require proposals and performs *single shot* actor-action detection in videos. SSA2D is an encoder-decoder based unified network, which utilizes spatio-temporal contextual information between objects and their surrounding pixels for joint detection of *multiple* objects and activities in *multiple* input video frames at once. In RPN based approach, *ROI-Pooling* [29] is used to extract object focused features for each proposal and therefore it leads to a higher memory consumption with increase in number of objects in the scene. In contrast, we propose *Single-Shot Attentive Masking (SSA-Masking)*, which utilize a spatio-temporal unified mask (*STU-Mask*) for all the objects to perform feature filtering at once. This unified masking makes the proposed approach much more efficient for dense video scenes.

A prior knowledge of actor presence provides crucial information regarding the performed action. A joint training of actor and action detection allows us to utilize actor-prior (*A-Prior*) to assist in action detection. We propose actor-prior infusion (*AP-Infusion*), which make use of spatio-temporal actor features as actor-priors. We demonstrate that *AP-Infusion* helps in improving performance of action detection as well as actor detection. Apart from this, SSA2D utilize atrous convolutions [3] and feature pyramid network [22], which we adapt for videos, to address the issue of *multiple scales* of actors present in a scene. SSA2D is jointly trained end-to-end for three objectives, actor detection, action detection, and STU-Mask, with the help of a combination of pixel-wise cross-entropy loss and a generalized dice-loss [32], which we extend for videos as Generalized 3D Dice Loss to balance the foreground and the background pixels in videos.

The proposed SSA2D model has several advantages over existing methods: 1) It requires fewer network parameters making it memory efficient, 2) It has faster inference and also requires less training time, 3) The memory requirement of SSA2D does not depend upon the number of actors present in the video, and therefore it does not have any scalability issues for dense scenes, and finally 4) SSA2D can perform spatio-temporal actor-action detection on multiple input video frames at once making it more efficient when compared with existing approaches which can perform detection only on single frame at a time. We make the following contributions in this work:

- We present SSA2D, a novel end-to-end network, to perform pixel-wise detection of actors and actions in videos. SSA2D is trained jointly to detect multiple actors and their actions simultaneously, while not relying

on region proposals or any external post-processing.

- We propose SSA-Masking, which utilize unified spatio-temporal mask for selective feature extraction of multiple actors at once. In contrast to ROI-Pooling, it is much more efficient for dense video scenes.
- We propose the use of actor-prior for spatio-temporal pixel-wise action detection with the help of spatio-temporal actor features, leading to an improved performance of both action as well as actor detection in joint learning.

We perform extensive experiments on A2D and VidOR dataset and demonstrate SSA2D to be efficient as well as effective for joint actor-action detection in videos. The proposed method is significantly efficient in inference ( $\sim 11x$  faster for RGB and  $\sim 6x$  faster for RGB+optical-flow) as well as in training (2x faster than state-of-the-art) with fewer network parameters and yet achieve performance comparable (sometimes even better) to the prior works.

## 2. Related work

### 2.1. Human action detection in videos

Action detection in videos require spatio-temporal localizations of actors in each frame which is then used for classification. Extending the image classification models [29, 27, 28, 23, 11], prior CNN based works detect humans in each frame and combine them temporally to form action tubes while classifying at clip level [41, 25, 40, 21], leveraging existing classification techniques from [1, 12, 16, 26, 36, 34, 31, 35]. [16] uses RPN based approach to detect humans in each frame and then forms action tubes by stitching them together, followed by Tube of Interest (TOI) pooling and action classification. [15] does TOI pooling based on foreground segmentation map from an encoder-decoder based network. [9] uses RPN along with transformer based attention map that detects and classifies actions. [7] uses a 3D capsule based CNN, where the authors apply routing-by-agreement algorithm to capture various action representations, leading to localize actions and classify them at the same time. Although prior works show great improvements on action detection in videos, they are limited by complex region proposal network coupled with region pooling or can only detect and classify single actor per video, creating a challenge to adapt it to denser real-life scenarios.

### 2.2. Actor and action detection in videos

Actor-action detection problem is related to identifying the actors and their corresponding actions in a given clip, where both semantic localization and classification is required. The authors in [38] proposed the A2D dataset,

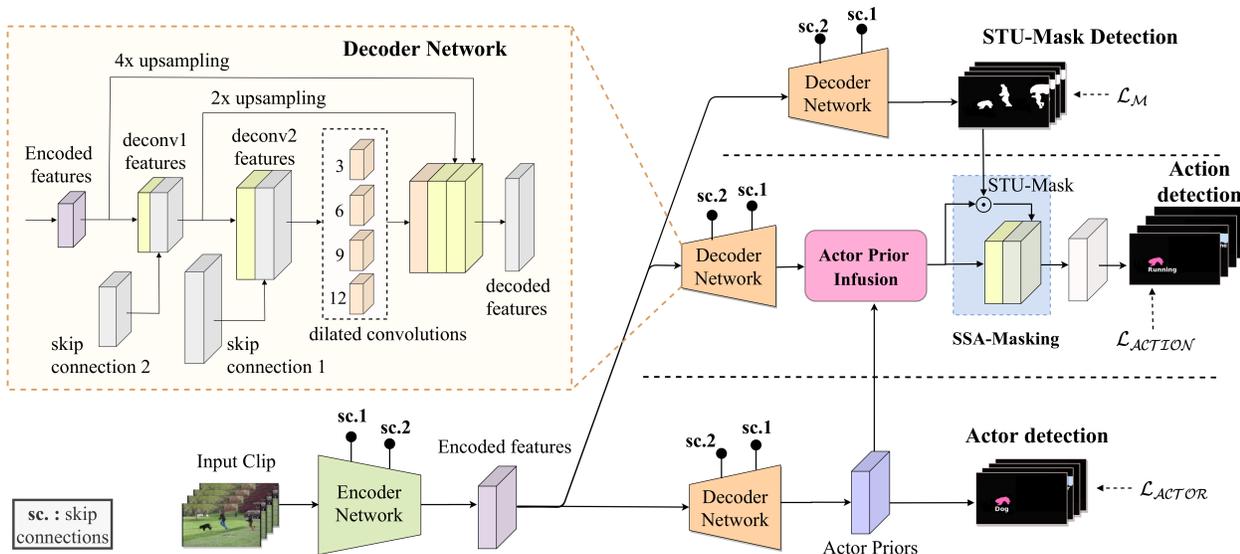


Figure 2. Overview of the proposed method for pixel-wise actor-action detection showing the overall SSA2D architecture. The 3D convolution based encoder network extracts features which are used in three separate branches for actor, action, and *STU-Mask* detection. The action detection branch use *AP-Infusion* to utilize actor-priors and *SSA-Masking* to focus on activity regions in the video. All three branches use a decoder network which has similar architecture, however the weights are not shared among these tasks.

a large scale benchmark dataset to study actor-action detection problem. An early approach of adaptive grouping of segments during inference improves segmentation in A2D [37]. [39] proposed weakly supervised method and train the model using only video-level tags. A two-stage model was proposed by [18], where objects are detected first and their bounding box are refined for segmentation outputs. [8, 24] use sentence priors to detect actor-actions in videos. Authors in [17] propose a joint end-to-end model which uses two-stream input (RGB + optical flow) to classify object regions and perform segmentation on them. Conceptually based on [13], they generate semantic features and use RPN to segment and classify actor-action pairs. [6] use similar approach and propose segmentation based region proposal and pooling to detect actor and action classes. They apply a region pooling based fully convolutional segmentation network for their actor segmentation, followed by 2D ResNet-101 [14] for action classification. Although prior works show great improvements on joint actor-action classification, they are limited by expensive region proposal and pooling which increases the approach’s complexity.

### 3. Proposed method

Given a video  $V \in \mathbb{R}^{T \times H \times W \times 3}$  as input, SSA2D jointly predicts actor detection  $ActorD \in \mathbb{R}^{T \times H \times W \times C_{actor}}$  and action detection  $ActionD \in \mathbb{R}^{T \times H \times W \times C_{action}}$ . Here,  $T$  is the number of frames in the input clip,  $H$  is the height and  $W$  is the width of the video frames,  $C_{actor}$  is total number of actor classes, and  $C_{action}$  is the total number of action categories. In addition to these two, SSA2D also predicts

a spatio-temporal mask  $STUMask \in \mathbb{R}^{T \times H \times W \times 2}$ , which is used for *SSA-Masking*. SSA2D consists of an encoder network  $E$ , and three separate branches for actor, action, and *STU-Mask* detection. Each task utilize a decoder network ( $D$ ), which has similar architecture for all the three branches. An overview of SSA2D is shown in Figure 2.

#### 3.1. Encoder network

Understanding and extracting relevant features both spatially and temporally is crucial in learning a video’s actor-action relations. We utilize a 3D convolution based encoder  $E$  that extracts actor and action related feature volume  $f_{enc}$  from a given input video clip  $V \in \mathbb{R}^{T \times H \times W \times 3}$ . The network takes a video clip as input with  $[T \times H \times W \times 3]$  with  $T$  frames at a resolution of  $H \times W$  and outputs a feature volume  $f_{enc} \in \mathbb{R}^{\frac{T}{4} \times \frac{H}{16} \times \frac{W}{16}}$  as video encodings. We use I3D [1] model as our encoder where we adapt the network by controlling the pooling strides (more details in supplementary). This encoder can use any state-of-the art 3D convolution network.

#### 3.2. Decoder network

The spatio-temporal features  $f_{enc}$  extracted using the encoder network  $E$  needs to be decoded into a larger fine-grained resolution for jointly detecting the actors and actions. The decoder network  $D$  takes  $f_{enc}$  as input and performs a series of 3D deconvolution [42] and upsampling operations to get the desired resolution for fine-grained pixel-wise detection. We upsample the encoded features to  $[\frac{T}{2} \times \frac{H}{4} \times \frac{W}{4}]$  as a final resolution to reduce parameters. We add skip connections from the encoder network to ev-

ery deconvolution layer to preserve the suppressed features during downsampling. Adapting multi-scale object feature learning techniques from images, we extend atrous convolutions [3, 4] and feature pyramid network [22, 20] to 3D architecture for videos. Atrous convolutions helps encode multi-scale contextual information around each pixel while feature pyramid helps preserve features of smaller objects.

The same decoder architecture  $D$  is used in all the three branches, however, the network parameters are not shared as these branches solve different tasks. The final output from all the branches is upsampled to match the resolution of the input video with the help of linear interpolation. A detailed architecture of  $D$  is shown in Figure 2 and more details are provided in the supplementary.

### 3.3. Actor detection

For pixel-wise actor detection, the actor detection branch utilizes encoded video features  $f_{enc}$  and learns pixel-wise  $C_{ap}$  actor prior ( $A\text{-prior} \in \mathbb{R}^{T \times H \times W \times C_{ap}}$ ) with the help of a decoder network  $D_{actor}$ . A final 3D convolution layer takes the learned  $A\text{-Prior}$  and predicts  $C_{Actor}$  channels for each pixel ( $C_{Actor}$  being the total number of actors present in the dataset). A *Softmax* activation is applied across actor channels for each pixel location as each pixel will correspond to only one of the actors. This gives us  $ActorD$  for pixel-wise actor detection in the input video. The scores in each channel corresponds to one of the actor class and indicates its presence in that spatio-temporal location.

### 3.4. Action detection

The action detection branch  $Action$  takes the spatio-temporal features  $f_{enc}$  from the encoder network  $E$  as input and uses the decoder architecture  $D_{action}$  from section 3.2 to learn action relevant feature maps  $f_a$  for action detection. As each actor’s interaction with surrounding objects is decisive in inferring its actions, the actor detection branch will have more meaningful features corresponding to each actor. However, it is also important to focus only on the spatio-temporal region where the action is occurring. To address these issues, we propose Actor Prior Infusion (*AP-Infusion*) and Single-Shot Attentive Masking (*SSA-Masking*), which allow the network to filter and learn meaningful interaction between the detected actors for action detection.

**Actor Prior Infusion (*AP-Infusion*)** The Actor Prior Infusion (*AP-Infusion*) provides additional information to the action detection network in form of latent actor representations. This is done by integrating  $A\text{-priors}$  with action related features, adding more actor focused contextual information and helps in action detection. As shown in Figure 2, the  $A\text{-priors}$   $f_{ap}$  are integrated with action features  $f_a$  from the decoder network in  $Action$  branch as  $f_{act} = Conv3D(\langle f_a, f_{ap} \rangle)$ , where  $Conv3D$  is 3D convolution

operation and  $\langle \rangle$  represented feature concatenation. We also experimented with feature addition and observed similar performance.

**Single-Shot Attentive Masking (*SSA-Masking*):** Instead of generating proposal boxes from external networks [29] or using all possible region boxes [27], we use single-shot attentive masking for feature filtering. A fine-grained spatio-temporal region is helpful to filter and improve the coarse actor-action detection results. To get this spatio-temporal mask, the features  $f_{enc}$  from the encoder network  $E$  are passed to a decoder network  $D_{STU-Mask}$  which predicts pixel-wise scores  $STU - Mask \in \mathbb{R}^{T \times H \times W \times 2}$  for each spatio-temporal location in the input video. Each pixel’s score in the  $STU-Mask$  indicates whether it is relevant to the action or not. The network learns to identify potential actor regions through the  $STU-Mask$ . This mask from the  $D_{STU-Mask}$  is used as spatio-temporal unified mask  $f_{mask} \in \mathbb{R}^{T \times H \times W \times 1}$  to filter the spatio-temporal features for action detection.

The action features  $f_{act}$  augmented with actor-priors are filtered using *SSA-Masking*. The augmented features  $f_{act}$  are integrated with the  $STU-Mask$  [ $f'_{act} = f_{act} \odot f_{mask}$ ] to get the filtered features  $f'_{act}$ . The filtered features  $f'_{act}$  are integrated back with the original action features  $f_{act}$  [ $f''_{act} = \langle f'_{act}, f_{act} \rangle$ ] to keep both action as well as contextual background features for an effective learning. With this masking, forward pass only learns detection of useful feature regions while backward pass has minimal gradient update for unrelated regions. Furthermore, the *SSA-Masking* can be done on the whole frame in a single-shot, removing the need for extracting multiple region proposal boxes and performing ROI/TOI pooling. The masking can be done within the network, making this an end-to-end architecture. During training, we use the ground truth  $STU-Mask$ . While testing, we extract the  $D_{STU-Mask}$  detection results and pass that as the  $STU-Mask$  within the network. Finally, the output feature from *SSA-Masking* is used to predict  $ActionD$  with  $C_{Action}$  channels using 3D convolution, where each channel corresponds to one action class. Since each pixel is evaluated individually, it can be formulated to have multi-labels and multi-class predictions.

### 3.5. Objective function

The proposed network is trained end-to-end with joint learning of three tasks: actor detection, action detection, and STU-Mask detection. Since we predict pixel-wise maps for each branch, we have to consider the large imbalance in active and non-active pixels, with fewer active pixels for sparse scenes. This imbalance is handled using ratio loss for the scene. In case of image segmentation, this can be computed as a ratio of foreground pixels to background using the Generalized Dice Loss [32]. We extended it to videos as

Generalized 3D Dice Loss with the following formulation:

$$\mathcal{L}_{\mathcal{DL}} = 1 - \frac{2 \sum_{c=1}^C \sum_{i=1}^N p_{ci} * \hat{p}_{ci}}{\sum_{c=1}^C (\sum_{i=1}^N p_{ci}^2 + \sum_{i=1}^N \hat{p}_{ci}^2 + \epsilon)} \quad (1)$$

where the dice coefficient score is computed per class  $C$  of given task,  $N$  is total number of pixels in segmentation map of a video clip, probability  $p_{ci} \in (0, 1)$  is the ground-truth segmentation map, and  $\hat{p}_{ci} \in (0, 1)$  is the network’s predicted segmentation map probability.

The actor detection loss is defined as the negative log-likelihood of the ground truth class and is computed as categorical cross-entropy per pixel. For  $C_{actor}$  set of actor classes, the actor detection head generates  $C_{actor}$  segmentation maps, where each pixel’s ground truth actor class is  $x$  and predicted actor class is  $\hat{x}$ . The loss is calculated for each pixel across all classes and then averaged over all pixels, which gives us the following loss formulation:

$$\mathcal{L}_{ACTOR} = [-\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{C_{actor}} (x_{i,j} \log(\hat{x}_{i,j}))] + \mathcal{L}_{\mathcal{DL}}. \quad (2)$$

Action detection is also defined similarly to actor detection, with  $C_{action}$  segmentation maps generated. For each pixel’s ground truth action class  $y$  an action class  $\hat{y}$  is predicted, and the loss is:

$$\mathcal{L}_{ACTION} = [-\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{C_{action}} (y_{i,j} \log(\hat{y}_{i,j}))] + \mathcal{L}_{\mathcal{DL}}. \quad (3)$$

We look at the STU-Mask detection task as a binary segmentation task, where all the actor pixels are considered as positive and all others as negative. The loss is computed using binary cross-entropy in combination with the dice loss:

$$\mathcal{L}_{\mathcal{M}} = [-\frac{1}{N} \sum_{i=1}^N p_i \log(\hat{p}_i) - (1 - p_i) \log(1 - \hat{p}_i)] + \mathcal{L}_{\mathcal{DL}}, \quad (4)$$

where  $\hat{p}_i$  is the prediction and  $p_i$  is the ground-truth. The total loss is a combination of these losses and is defined as:

$$\mathcal{L} = \mathcal{L}_{ACTOR} + \mathcal{L}_{ACTION} + \mathcal{L}_{\mathcal{M}}. \quad (5)$$

### 3.6. Implementation and training details

We implement the proposed method in Keras [5] with Tensorflow backend. The encoder block uses I3D [1] pre-trained on Kinetics-400. We input a video clip of temporal resolution (T) of 16 frames and spatial resolution 224 x 224. The final output of the encoder network is 4 x 14 x 14, which we then upsample to 8 x 112 x 112 for *STU-Mask* detection branch and 8 x 56 x 56 for actor and action detection branch. For the RGB + optical flow approach, we perform two stream implementation where two encoders are

used for each input type. The encoders share some of the final layers to reduce network size, and skip connections are passed from both streams. Since our network does not have any fully connected layers or an extra region proposal network, the network has fewer number of parameters and can be trained end-to-end in a single stage.

**Optimization** We use Adam optimizer [19] with an initial learning rate of 1e-4 and finetune at a rate of 1e-5. For our joint training task, we can fit an effective batch size of 14 clips per iteration. The model is trained for 5 epochs with initial learning rate and fine-tuned for another 6 epochs.

**Joint training** We train all three branches together with the loss weights assigned based on class distribution per task. For the A2D dataset, *STU-Mask* detection is given the weight of 0.3, while both actor and action detection task is given weights of 1.3 (based on per class pixel distribution).

**STU-Mask** We input the *STU-Mask* of size 4 x 56 x 56 for the action detection task, which helps to increase focus on the related pixels. For training, we use all actor regions from ground truth as the *STU-Mask*. During inference, we use the *STU-Mask* predicted by the *STU-Mask* branch.

## 4. Experiments

### 4.1. Datasets

**A2D dataset:** A2D [38] is the first video dataset with multiple actor classes and action classes in the same clip along with semantic labels. It provides pixel-level semantic labels of 3-5 frames for each video and is the only joint actor-action segmentation benchmark reported in prior works [38, 37, 18, 17]. The dataset consists of 3,782 YouTube videos, consisting of 7 actor classes performing one of the 9 action classes. A total of 43 actor-action pairs are valid and used for joint actor-action segmentation task. Both pixel level and bounding box annotation per actor-action pair are provided in this dataset.

**VidOR dataset:** We also evaluate our method on the VidOR dataset [30] which contains 10,000 videos with 80 object categories and 42 action predicates with bounding box annotations. Although it has more videos for training, the dataset is more challenging as it has a wide range of objects with a skewed distribution where 92% of objects are from only 30 categories. Each action is part of a triplet and consists of a subject and an object, with the subject performing the action. Thus, action detection using object and its surrounding context is more meaningful.

### 4.2. Metric

Following the evaluation protocols from [37] and [17], we measure average per-class accuracy (*ave*), global pixel

Input	Method	Actor			Action			Joint (A,A)			Time (ms) per frame
		glo	ave	mIoU	glo	ave	mIoU	glo	ave	mIoU	
RGB	GPM + TSP [37]	85.2	58.3	33.4	85.3	60.5	32.0	84.2	43.3	19.9	-
	GPM + GBH [37]	84.9	61.2	33.3	84.8	59.4	31.9	83.8	43.9	19.9	-
	Chen et al. [2]*	91.3	49.16	49.2	87.44	35.12	38.7	87.1	43.06	26.7	-
	Ji et al. [17]	93.7	79.5	66.5	86.3	60.4	36.8	87.8	46.2	29.4	-
	Dang et al. [6]	95.0	85.5	67.0	92.9	68.8	48.1	92.5	51.5	34.5	750
	SSA2D (Ours)	96.1	79.4	66.8	94.4	66.2	46.5	93.8	49.3	34.6	67
RGB + OF	TSMT + GBH [18]	85.8	72.9	42.7	84.6	61.4	35.5	83.9	48.0	24.9	-
	TSMT + SM [18]	90.6	73.7	49.5	89.3	60.5	42.2	88.7	47.5	29.7	-
	Gavrilyuk et al. [8] <sup>†</sup>	92.8	71.4	53.7	92.5	69.3	49.4	91.7	52.4	34.8	-
	Ji et al. [17]	94.5	79.1	66.4	92.6	62.9	46.3	92.5	51.4	36.9	350**
	Dang et al. [6]	95.3	86.0	68.1	93.4	70.7	51.1	93.0	56.4	38.6	1100
	SSA2D (Ours)	96.2	80.1	67.5	94.9	69.1	51.3	95.0	54.7	39.5	180

Table 1. Quantitative comparison of SSA2D on A2D dataset with prior approaches using RGB and RGB + optical flow as input, reporting average per-class accuracy (*ave*), global pixel accuracy (*glo*) and mean pixel Intersection-over-Union (*mIoU*) for each task. <sup>†</sup> Uses sentence priors. \*Uses weakly-supervised training. \*\* Is time adjusted for same hardware setting by correspondence with authors.

Dataset	Method	Actor			Action			Joint (A,A)		
		glo	ave	mIoU	glo	ave	mIoU	glo	ave	mIoU
A2D	<b>Full (RGB only)</b>	<b>96.2</b>	<b>80.1</b>	<b>67.5</b>	94.4	<b>66.2</b>	<b>46.5</b>	<b>93.8</b>	<b>49.3</b>	<b>34.6</b>
A2D	w/o Actor-Prior	96.1	79.1	65.7	93.9	61.5	40.9	93.4	46.3	32.1
A2D	w/o SSA-Masking	96.1	79.9	67.2	<b>94.6</b>	63.9	43.6	93.7	48.2	33.8
A2D	w/o atrous convolutions	92.4	76.4	62	94.1	62.3	41.6	92.8	45.2	31.7
A2D	w/o multi-scale	96.0	79.8	66.5	94.2	63.8	43.1	93.6	47.9	33.1
VidOR	<b>Full (RGB only)</b>	<b>72.2</b>	<b>7.6</b>	<b>5.1</b>	<b>66.8</b>	<b>33.2</b>	<b>7.9</b>	<b>41.7</b>	<b>15.7</b>	<b>2.1</b>
		(54.1)	(20.5)	(12.5)	(70.2)	(40.8)	(11.7)	(44.2)	(18.8)	(5.1)
VidOR	w/o Actor-Prior	71.1	7.1	4.1	61.8	28.3	5.9	37.7	12.1	1.1
VidOR	w/o SSA-Masking	71.8	7.1	4.3	65.4	31.7	7.1	39.2	15.2	1.8
VidOR	w/o atrous convolutions	71.5	6.4	3.4	63.1	30.8	6.2	38.3	14.4	1.3
VidOR	w/o multi-scale	71.2	6.1	3.1	61.7	29.5	6.0	37.8	14.1	1.1

Table 2. Ablation study of various components of SSA2D and their effect on actor-action detection on A2D and VidOR dataset. We report scores on average per-class accuracy (*ave*), global pixel accuracy (*glo*) and mean pixel Intersection-over-Union (*mIoU*). Values in bracket represent scores for the 20 most frequent classes.

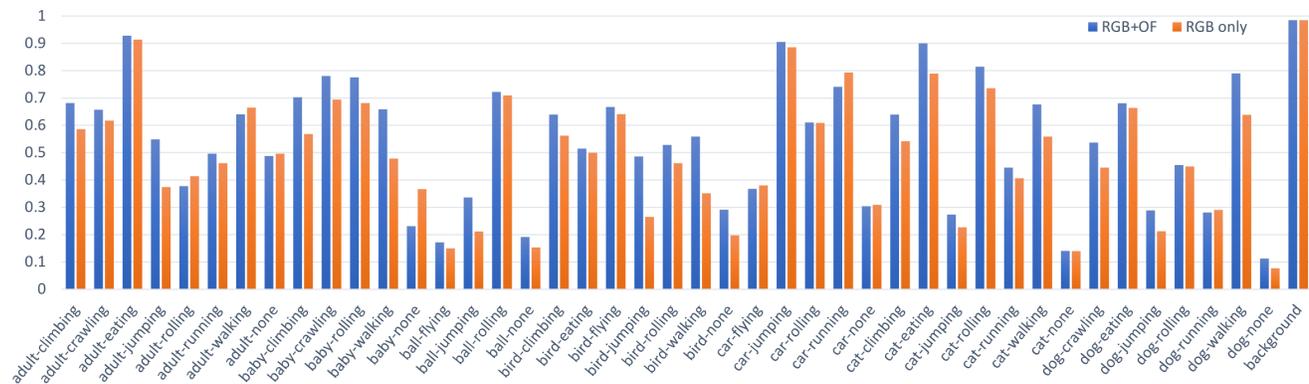


Figure 3. Per actor-action category average accuracy score for A2D. The orange bars show results for RGB modality and blue bar for combined RGB and Optical Flow. We observe that on average, most of the classes benefit from having extra optical flow information.

accuracy (*glo*) and mean pixel Intersection-over-Union (*mIoU*) as evaluation metrics. Accuracy is the percent of pixels with correct label prediction, where (*glo*) is computed over all pixels and (*ave*) is first computed per class

and then averaged. Since background covers a large area and most models are biased towards background, *mIoU* is the most representative metric for correct pixel prediction over all classes [18]. We report results for actor, action and

joint actor-action detection for (*ave*), (*gIoU*) and (*mIoU*) metrics for a fair comparison with existing methods.

### 4.3. Results

The performance of SSA2D on A2D is shown in Table 1. Using only RGB stream, SSA2D gives improved joint actor-action mIoU with significant reduction in inference time ( $\sim 11x$  faster). This demonstrates that the network’s joint training is able to learn action features based on actors while computing video level detection faster than previous methods. Moreover, using RGB+OF input we observe that the network gives improved mIoU scores on action and joint task as expected, demonstrating that our approach generalizes to different types on input modalities. We also analyze per class performance and the scores are shown for both of our RGB model and RGB+OF model in Figure 3. We observe that the proposed method can detect most classes accurately in RGB model and the scores are further increased with additional flow information.

We report the performance of proposed method on ViDOR dataset in Table 2. We evaluate the model on same evaluation metrics as used for the A2D dataset. Due to its long tail distribution, the dataset suffers from large data unbalance. As such, even when our network performs well on those classes, the average accuracy and mean IoU scores drop due to tail classes with fewer training samples.

**Qualitative evaluation:** Figure 4 and 5 show qualitative results for actor-action detection. We observe that the proposed method can predict reasonable detections for most of the cases. Figure 6 shows that the network predicts correctly even though ground truth annotation is missing labels. The network is able to generalize and learn effective actor-action features to predict the missing labels. The last column shows the network detecting a hard sample correctly. Even though the cat blends with the background, it is well segmented and detected as *cat-jumping* class. Using 3D convolution on videos where the object is better visible in other frames, detection improves in such challenging frames as features is evaluated together for the entire video.

### 4.4. Ablation Studies

We further validate the importance of different components proposed in our model through ablation experiments. Since our contribution is agnostic to input type, we evaluate all variations against the full RGB only model in Table 2.

**Actor Prior Infusion (AP-Infusion):** One of the key components in improving action detection in our model is the use of actor prior for inferring activities. The *A-prior* coming from actor detection branch provides contextual information regarding all actors around each pixel. It is reasonable to have an understanding of the actors involved in order to better judge the actions happening. While [11] shows that using pair-wise actor features helps improve ac-

tion classification in images, our *AP-Infusion* approach uses all of the involved actor’s features together because of the pixel-level detection. We train our model without using the *AP-Infusion* to evaluate its effectiveness. As seen in *Full* and *w/o Actor-prior* models of Table 2, *A-prior* provides a significant gain in the action detection task ( $\sim 6\%$  improvement in mIoU for A2D) and subsequently increases the scores for all other tasks. Since we perform a joint training, we also observe the decrease of scores for actor detection task when feedback from the *AP-Infusion* block is not present.

**SSA-Masking:** *SSA-Masking* is used in action detection task to filter and enhance focus on action regions for pixel-wise detection. This reduces the surplus background noise and helps in a faster convergence. Our motivation to use the *STU-Mask* is to provide emphasis on features related to actors while filtering out excess background data. In RPN based methods, ROI-Pooling play the role of feature filtering. However, pooling is performed for each proposal independently making it computationally expensive. We use a unified mask for all the actors in the scene for this filtering making SSA2D more efficient. *SSA-Masking* enables the network to focus more on the actor pixels while suppressing the background pixels, which leads to an improved network performance for action detection ( $\sim 3\%$  increase in mIoU for A2D) and also provides a faster network convergence ( $\sim 3x$ ).

### 4.5. Comparative analysis

Figure 1 shows a comparative view of our method along with [17, 37, 6] in terms of performance and speed. Compared to [6], our training **does not** use weights pre-trained on segmentation task and trains the decoders from scratch, while [6] uses pre-trained weights on segmentation tasks. We observe that our method performs significantly better compared to [17, 37] in all evaluation metrics as seen in Table 1. We see that despite fast inference time for [17], it under-performs and has a larger model. Furthermore, our quantitative scores are similar or slightly better than previous state of the art method [6] and has significantly better inference time ( $\sim 11x$ ). This large gap in inference time makes our approach better suited for actor-action detection in videos as compared to all prior works.

**Network parameters:** Another key aspect of the proposed method is the smaller network size (35M params for RGB and 55M params for RGB+OF) compared to [6, 17] (44M params for RGB and 88M params for RGB+OF). Compared to prior works, SSA2D has reduced network size which relates to the overall efficiency and performance speed. The memory-efficient reduced network also enables end-to-end training for all tasks simultaneously as compared to multi-stage training [6], which is time consuming.



Figure 4. Qualitative results of our method on A2D dataset. The first row shows the input key frame, second row shows the ground truth with annotation labels and third row shows our joint actor-action detection result with predicted labels.

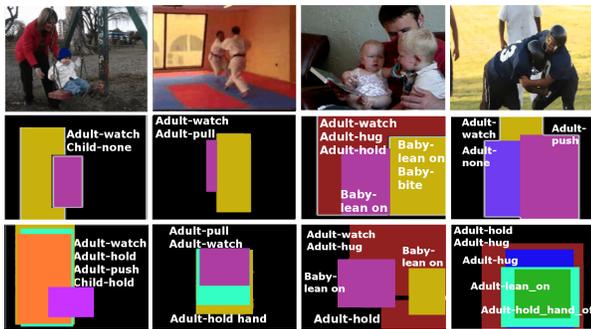


Figure 5. Qualitative results of our method on ViDOR dataset. The top, middle and bottom row represents input key frame, ground truth and our joint actor-action predictions with label respectively.



Figure 6. Qualitative analysis of some success cases where the network predicts better than the ground truth. The top, middle and bottom row represents input key frame, ground truth and our joint actor-action detection predictions with label respectively. The network correctly predicts labels for actions which are not annotated in the ground truth but present in the clip.

**Running time:** A crucial difference between SSA2D and prior works is that previous works rely on RPN as an auxiliary task during training to obtain actor regions for ROI pooling. Our method uses end-to-end pixel-wise detection and jointly trains actor-action tasks on pixel level while keeping the implementation efficient and effective. For a

fair evaluation, we evaluate the time taken to perform the evaluations on a single core of an Intel Xeon 2.3GHz CPU using a single NVidia Tesla K80 GPU [6]. During inference, our system takes *180 ms per frame* with the RGB + OF model, while it takes only *67 ms per frame* for single stream RGB model. [6] report computational time of *1100 ms per frame* for their full system with optical flow, with around *350 ms* being used for optical flow estimation.

## 5. Conclusions

We propose SSA2D, a simple yet effective approach for single-shot actor-action detection in videos. We demonstrate that actor-action detection in videos can be performed without relying on region proposal network where thousand of proposals are required making it in-efficient for dense video scenes. We evaluate the proposed approach on A2D and ViDOR datasets and achieve comparable (sometimes even better) performance when compared with prior works. The proposed model can be efficiently trained (2x faster) with a fast inference ( $\sim 11x$  faster for RGB and  $\sim 6x$  faster for RGB+optical-flow) with fewer network parameters when compared with best performing prior works.

**Acknowledgments** This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. D17PC00345. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

## References

- [1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [2] Jie Chen, Zhiheng Li, Jiebo Luo, and Chenliang Xu. Learning a weakly-supervised video actor-action segmentation model with a wise selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9901–9911, 2020.
- [3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.
- [5] François Chollet et al. Keras, 2015.
- [6] Kang Dang, Chunlun Zhou, Zhigang Tu, Michael Hoy, Justin Dauwels, and Junsong Yuan. Actor-action semantic segmentation with region masks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.
- [7] Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Video-capsulenet: A simplified network for action detection. In *Advances in Neural Information Processing Systems*, pages 7610–7619, 2018.
- [8] Kirill Gavriluk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966, 2018.
- [9] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.
- [10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [11] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.
- [12] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [15] Rui Hou, Chen Chen, and Mubarak Shah. An end-to-end 3d convolutional neural network for action detection and segmentation in videos. *arXiv preprint arXiv:1712.01111*, 2017.
- [16] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *IEEE International Conference on Computer Vision*, 2017.
- [17] Jingwei Ji, Shyamal Buch, Alvaro Soto, and Juan Carlos Niebles. End-to-end joint semantic segmentation of actors and actions in video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 702–717, 2018.
- [18] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Joint learning of object and action detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4163–4172, 2017.
- [19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [20] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. *arXiv preprint arXiv:1901.02446*, 2019.
- [21] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 303–318, 2018.
- [22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.
- [23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [24] Bruce McIntosh, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Visual-textual capsule routing for text-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9942–9951, 2020.
- [25] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream r-cnn for action detection. In *European conference on computer vision*, pages 744–759. Springer, 2016.
- [26] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.
- [27] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [28] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint*, 1612, 2016.
- [29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [30] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on Interna-*

- tional Conference on Multimedia Retrieval*, pages 279–287. ACM, 2019.
- [31] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
- [32] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.
- [33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 9627–9636, 2019.
- [34] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 4489–4497. IEEE, 2015.
- [35] Shruti Vyas, Yogesh S Rawat, and Mubarak Shah. Multi-view action recognition using cross-view video prediction. In *Proceedings of the European Conference on Computer Vision*, 2020.
- [36] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018.
- [37] Chenliang Xu and Jason J Corso. Actor-action semantic segmentation with grouping process models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3083–3092, 2016.
- [38] C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso. Can humans fly? Action understanding with multiple classes of actors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [39] Yan Yan, Chenliang Xu, Dawen Cai, and Jason J Corso. Weakly supervised actor-action segmentation via robust multi-task ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1298–1307, 2017.
- [40] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2019.
- [41] Zhenheng Yang, Jiyang Gao, and Ram Nevatia. Spatio-temporal action detection with cascade proposal and location anticipation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.
- [42] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Robert Fergus. Deconvolutional networks. In *proceedings of the IEEE International Conference on Computer Vision*, volume 10, pages 2528–2535, 2010.