One-shot Learning-based Animal Video Segmentation

Tengfei Xue, Yongliang Qiao, He Kong, Daobilige Su, Shirui Pan, Khalid Rafique, Salah Sukkarieh

Abstract—**Deep learning-based video segmentation methods can offer good performance after being trained on the large-scale pixel labeled datasets. However, pixel-wise manual labeling of animal images is challenging and time consuming due to irregular contours and motion blur. To achieve desirable trade-offs between accuracy and speed, a novel one-shot learning-based approach is proposed to segment animal video with only one labeled frame. The proposed approach consists of three main modules: (1) Guidance Frame Selection (GFS) utilizes "BubbleNet" to choose one frame for manual labeling, which can leverage the fine-tuning effects of the only labeled frame; (2) Xception-based Fully Convolutional Network (XFCN) localizes dense prediction using depthwise separable convolutions based on one single labeled frame; (3) Post-processing (POST) is used to remove outliers and sharpen object contours, which consists of two sub-modules—Test Time Augmentation (TTA) and Conditional Random Field (CRF). Extensive experiments have been conducted on the DAVIS 2016 animal dataset. Our proposed video segmentation approach achieved mean intersection-over-union score of 89.5% on the DAVIS 2016 animal dataset with less run-time, and outperformed the state-of-art methods (OSVOS and OSMN). The proposed one-shot learning-based approach achieves real-time and automatic segmentation of animals with only one labeled video frame. This can be potentially used further as a baseline for intelligent perception-based monitoring of animals and other domain specific applications. The source code, datasets, and pre-trained weights for this work are publicly available at https://github.com/tengfeixue-victor/One-Shot-Animal-Video-Segmentation.**

Index Terms—**One-shot learning, Video segmentation, Deep learning, Convolutional neural network, Animal monitoring**

I. Introduction

Video segmentation aims to separate objects from the background in all frames of a given video, and is of great significance for pixel-level precision demanded applications such as autonomous robots, unmanned vehicles, environmental monitoring and surveillance, and so on [1, 2, 3, 4]. Some well-known semantic image segmentation models, including Fully Convolutional Network (FCN) [5], DeepLab [6], PSPNet [7], etc., usually incorporate extra layers (e.g. spatial pyramid pooling, multi-scale dilated convolution, multi-scale input paths, and conditional random field) for boosting accuracy. However, these models rely on large-scale pixel-level labelled datasets (e.g. PASCAL [8]) to train models and obtain desirable segmentation results. Therefore, there is an increasing demand for high segmentation accuracy with minimum data labeling costs. Recently, methods such as one-shot learning [9, 10] have been applied to video segmentation. One-shot learning methods can generalize the neural network to new tasks with one annotated sample by utilizing prior knowledge such as data, model, and algorithm [11, 12]. Wang et al. [13] unified object tracking and video segmentation to estimate the bounding boxes and segmentation masks for the rest of the frames with only the initial bounding box. OSVOS [14] implemented the skip connection network architecture to learn the appearance feature of the first frame and look for the matching appearance on the follow-ups. Yang et al. [15] established two modules which utilize the information in the one labeled frame and the previous frame's spatial information to adjust the segmentation results.

However, most of the above video segmentation paradigms label the first frame of video and then automatically segment object from the remaining frames. Actually, for video segmentation, the best single labeling frame is not always the first one [16]. In addition, methods like [14] achieved object segmentation based on complex networks, which is not feasible for real-time video segmentation tasks due to long computing time.

Moreover, despite the wide applicability of existing approaches, there are little research focusing on animal video segmentation. Animals are widespread in nature and the analysis of their shape and motion is important in many fields and industries [17, 18]. For example, segmenting animals from video and tracking their motion is a prerequisite for body condition scoring and behaviour analysis in precision livestock farming [19]. However, animals present unique challenges, compared to humans or other moving objects such as vehicles. First, the shape variation and motion blur is usually larger than humans or other moving objects [19]. Second, the amount of available data for training is usually limited, particularly for endangered animals, in contrast to public datasets related to humans or autonomous vehicles. Thus the lack of training data will inevitably lead to the performance degradation of the state of art deep learning methods when implemented on animals [20]. Last but not the least, animals such as cows and sheep often live in outdoor field environments where their appearance is camouflaged, making automatic segmentation

Tengfei Xue and Yongliang Qiao contributed equally to this work and share the first authorship. Corresponding author: Yongliang Qiao. Email: yongliang.qiao@sydney.edu.au

Tengfei Xue, Yongliang Qiao, He Kong, Khalid Rafique and Salah Sukkarieh are with Australian Centre for Field Robotics (ACFR), Faculty of Engineering, The University of Sydney, NSW 2006, Australia (e-mail: txue4133@uni.sydney.edu.au; yongliang.qiao@sydney.edu.au; he.kong@sydney.edu.au; khalid.rafique@sydney.edu.au; salah.sukkarieh@sydney.edu.au).

Daobilige Su is with College of Engineering, China Agricultural University, 100083, Beijing, China (e-mail: sudao2020@outlook.com).

Shirui Pan is with Department of Data Science and AI, Faculty of Information Technology, Monash University, Clayton, 3800, VIC, Australia (e-mail: shirui.pan@monash.edu).

even more challenging.

Motivated by the above observations, in this paper, we propose a novel one-shot learning-based real-time animal video segmentation approach in order to achieve high segmentation accuracy with one labeled image in complex background environments. As demonstrated in Fig. 1, the proposed approach consists of three main modules: BubbleNets-based guidance frame selection, XFCN, and post-processing. More specifically, the BubbleNet-based guidance frame selection module uses the deep bubble sorting framework to choose single best frame across the video for manual labeling, and then this labeled frame is used for the follow-up model fine-tuning; XFCN is deployed to accurately localize dense prediction through one-shot learning based on CNN feature fusion; the post-processing module comprises of two sub-modules, namely, test time augmentation and conditional random field, which are helpful to remove outliers and sharpen the animal contours.

In our work, we have introduced a lightweight and one-shot learning-based deep learning architecture for animal video segmentation. The main contributions of this work are: (1) The proposed lightweight and power-efficient network architecture, XFCN, encodes and decodes the images features for video segmentation, which significantly improved computation efficiency and segmentation accuracy. In our XFCN encoding part, the proposed 20 layers' Xception-lite uses depthwise separable convolutions to extract CNN features with fewer parameters. Although it looks similar to Xception, the proposed Xception-lite significantly increases representational efficiency and reduces over-fitting. In the XFCN decoding part, five different level feature maps were upsampled to the same size of the frame and concatenated into the final feature map for segmentation. Here dilated convolutions were implemented with different dilated rates to increase the receptive field for video segmentation. (2) An effective GFS was utilized to select one guidance frame from the video, which leverages the fine-tuning effects of the only labeled frame. In addition, in order to further improve the segmentation performance, the POST module (consists of Test Time Augmentation (TTA) and Conditional Random Field (CRF)) was used to reduce some noises and edge blurring. (3) Systemic experiments were conducted on DAVIS 2016 animal [21]. The proposed approach achieved 89.5% mIoU and 93.2% $\mathcal{F}$(Mean) with a running speed of 1.16s per frame, which outperformed the state-of-art methods (OSVOS, OSVOS + GFS, and OSMN). We also studied the effects of pre-training on segmentation accuracy. Experimental results show that pre-training can moderately enhance segmentation accuracy.

The remainder of this paper is organized as follows. Section II illustrates the our proposed video segmentation approach. Section III presents the experimental setup, including dataset, training details and evaluation methods. Evaluations of the proposed method on the DAVIS 2016 animal dataset and discussions of the performance are presented in Section IV. Section V contains ablation study. Finally, conclusions and areas for future research are given in Section VI.

II. Methodology

A. Overview of proposed approach

As illustrated in Fig. 1, the proposed approach has three main modules: Guidance Frame Selection (GFS), Xception-based Fully Convolutional Network (XFCN), post-processing (POST). In our proposed approach, a given video with one of its labeled frame is used to fine-tune a pre-trained network for the video segmentation. In addition, in order to remove the outliers from segmented results and refine animal contours, post-processing (it has two sub-modules: Test Time Augmentation and Conditional Random Field) is implemented to further improve the segmentation performance.

Denote the video sequence F with n frames as $F = [f_1, f_2, \cdots, f_n]$. In our proposed one-shot learning-based approach, GFS utilizes 'BubbleNet' [16] to select one guidance frame f_{guide} from one video sequence F (see Section II.B). The selected one frame will be labeled manually to get the ground

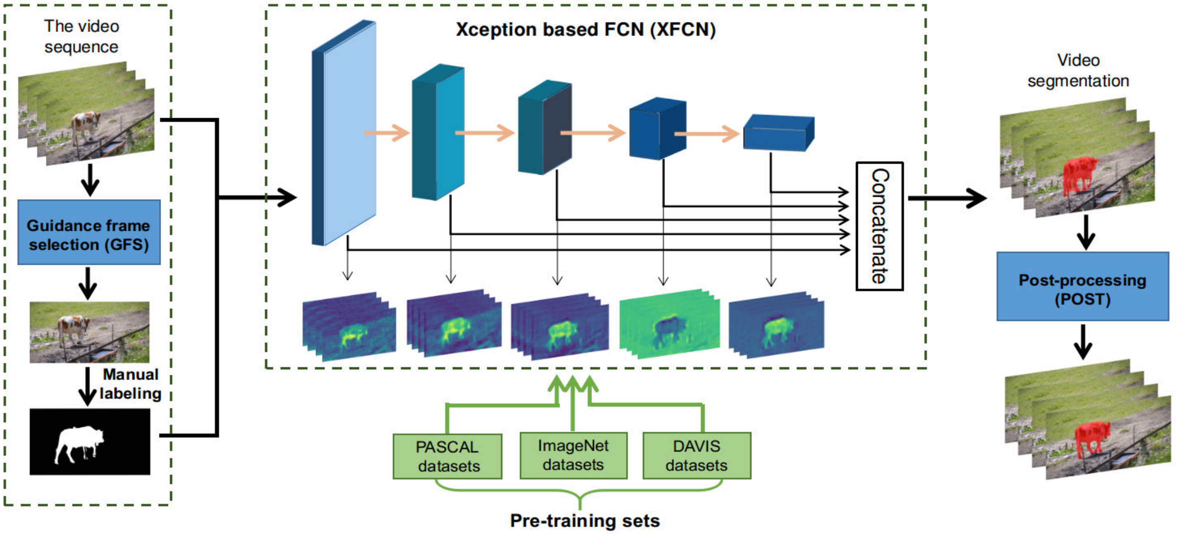


Fig. 1: The overall flowchart for one-shot learning based animal video segmentation.

truth mask s_{gt}. The only labeled frame s_{gt} with video sequence F will be fed into XFCN for network fine-tuning. Here XFCN is pre-trained offline with large dataset $D_{pretrain}$ before the fine-tuning process. After XFCN is fine-tunned with the only label frame f_{guide}, model M is obtained and segmentation results are denoted as $Y = [y_1,\ y_2,\ \cdots,\ y_n]$. Considering the fact that the segmentation results Y may have some outliers and blurring edges, TTA and CRF post-processing are implemented to obtain the final results $Y_f = [y_{f1},\ y_{f2},\ \cdots,\ y_{fn}]$ (Section II.D). The details of each step are discussed below.

B. Guidance frame selection (GFS)

In one-shot learning-based video segmentation approaches, the model needs to be fine-tuned with one labeled frame in the test video sequence. The state of the art video segmentation models and benchmark datasets always use a single labeling in the first frame to segment objects in the video sequence. However, as it has been illustrated in [16], the first frame is usually not the optimum selection for the video sequence. Therefore, in our work, the BubbleNet [16] was utilized to select one guidance frame from the video, then the selected frame was labeled for fine-tuning to leverage the model effects.

BubbleNet is an unsupervised deep sorting network based on loss functions without the need of labels. It compares the performance of using each frame as a guide, and swaps frames until the corresponding frame giving the best-predicted result is selected [16]. Then the selected frame was manually labeled for the subsequent model fine-tuning. By this, BubbleNet can effectively improve the performance of one-shot learning-based video segmentation.

The process of how BubbleNet chooses the guidance frame is illustrated in Algorithm 1. Our later experiments will confirm that the fine-tuning process with the selected labeled frame can boost the segmentation accuracy effectively.

Algorithm 1 BubbleNet guidance frame selection

```
Inputs:
  $F_{ref} = [f_1,\ f_2,\ \cdots,\ f_n]$ {The video sequence};
  $K = [1,\ 2,\ \cdots,\ i,\ \cdots,\ n]$ {Frame indexes of the video sequence};
Outputs:
  $f_{guide}$ {The guidance frame of the video sequence};
Algorithm:
  for $i \leftarrow 1$ to $N$ do
    $f_i,\ f_{i+1},\ F_{ref},\ K \leftarrow$ relative loss computation;
    $p \leftarrow$ The predicted relative loss by neural network;
    if $p > 0$ then
      swap $f_i,\ f_{i+1}$;
    end if
  end for
  return $f_{guide} \leftarrow$ The selected frame in the video sequence after the iteration;
```

C. XFCN architecture

Although OSVOS [14], with its VGG backbone and the skip-connection architecture, showed good feature representation for segmentation, it can be further improved in terms of computation efficiency and segmentation accuracy. Aiming at finding an effective network that can output accurate segmentation results with less computation time, and inspired by Xception65 backbone [22] of DeeplabV3+ [23], in this paper, we propose XFCN to encode and decode the extracted CNN features for video segmentation. The overall architecture of the network can be seen in Fig. 1.

In DeeplabV3+, the original Xception-65 backbone has 65 layers to extract visual features, which is time-consuming during the training or fine-tuning phrases. In our work, after experimenting on different layer-length backbone, a 20 layers' Xception-lite was proposed to extract features in XFCN, which incorporates residual connections and separable convolutions can achieve best trade-off between the accuracy and speed.

As demonstrated in Fig. 2, the proposed Xception-lite comprises five stages, the first three stages mainly focus on spatial feature extraction (e.g. edge, texture, shape) whilst the last two stages pay more attention to the semantic information extraction. Between different stages, MaxPooling operation is used to downsample the feature maps.

Generally, the texture information with high spatial resolution are represented in those front layers. With the layers going deeper, the semantic and abstract information are extracted. Leveraging features from different levels is significantly beneficial to video segmentation. According to our experimental comparison, five outputs (feature maps) from shallow layers to deep layers were selected and concatenated together for video segmentation.

As these five outputs have different sizes, they were firstly upsampled to the same size as the video frame and then concatenated to a feature map X_{final} with multiple dimensional information:

$$X_{final} = U\{X_5\}_{\uparrow 4} \mathbin{+\!\!+} U\{X_8\}_{\uparrow 8} \mathbin{+\!\!+} U\{X_{11}\}_{\uparrow 16} \mathbin{+\!\!+} U\{X_{17}\}_{\uparrow 16} \mathbin{+\!\!+} U\{X_{20}\}_{\uparrow 32} \tag{1}$$

where $X_5,\cdots,X_{20}$ are feature maps from five stages and their subscripts indicate the corresponding layer numbers; $U\{X\}_{\uparrow 4}$, $\cdots$, $U\{X\}_{\uparrow 32}$ are the upsampling operation with corresponding rate (i.e. 4, 8, 16, 32); $\mathbin{+\!\!+}$ means concatenation. Subsequently, multiple dimensional of X_{final} will be linearly fused by convolution with $1{\times}1$ kernel to form a one-dimensional probability map as the final result.

D. Post-processing

Although the proposed one-shot learning-based approach can segment the animals from videos, there are some noises and edge blurring problems. In order to further improve the segmentation performance, two popular post-processing methods, namely, TTA and CRF were used. TTA relies on augmenting test datasets, then performs the prediction both on the original and on the augmented versions of the image, followed by merging the predictions [24]. CRF is a discriminative statistical modelling method that is used when the class labels for different inputs are not independent [25], which is also a useful post-processing tool to improve the performance of segmentation.

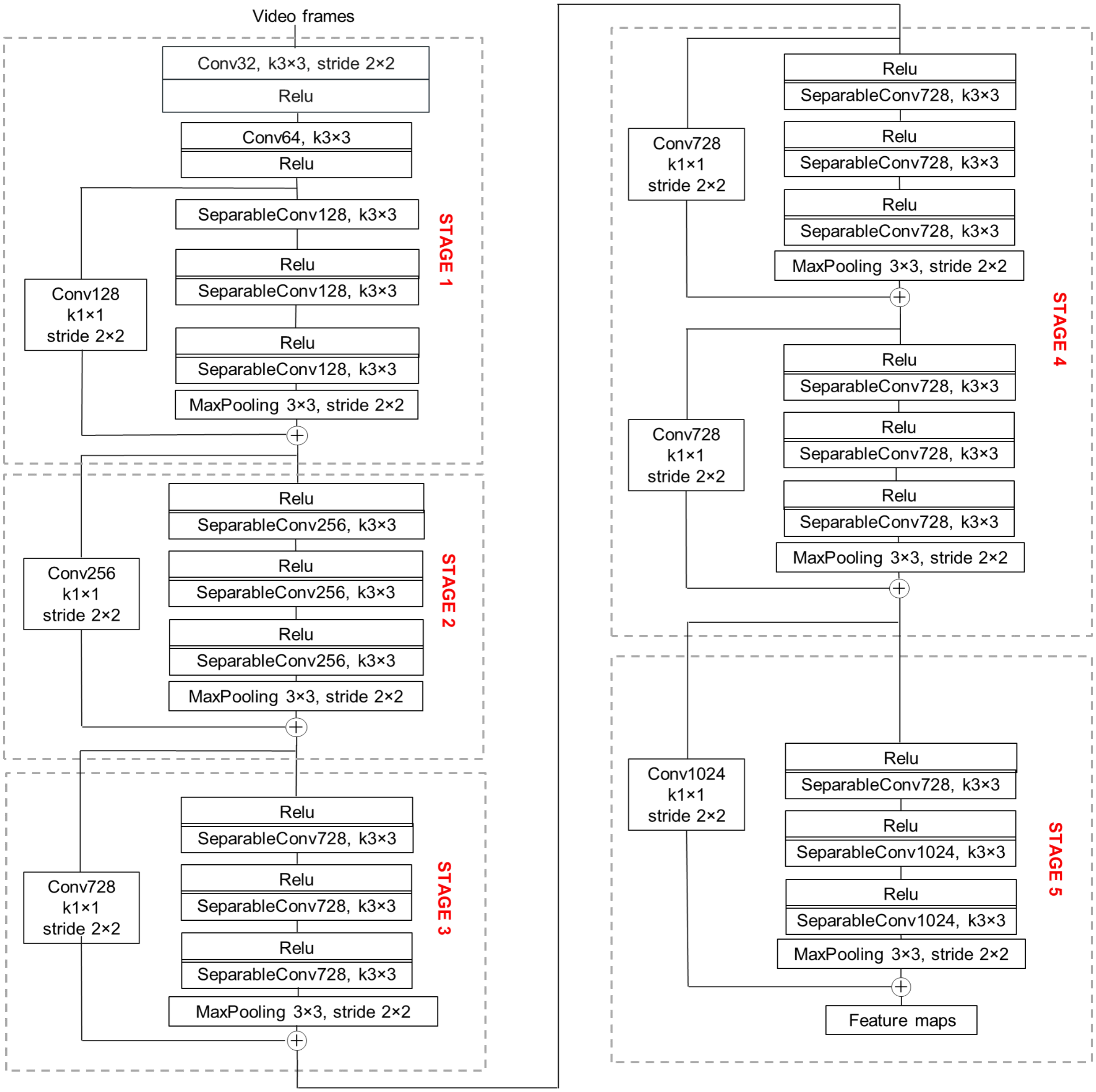


Fig. 2: Architecture of the proposed Xception-lite.

For the TTA method, flip data augmentation was implemented on testing videos, and segmentation was then performed on the original and augmented data. The final segmentation results were obtained from averaging all predictions of augmented videos, which can remove some outliers and slightly improve the segmentation performance.

In terms of CRF, fully connected CRF was used to refine the coarse output based on the label at each location itself, and their neighboring positions' labels and locations. The CRF is characterized by Gibbs distribution of a form:

$$P(X = x|f) = \frac{1}{Z(f)} e^{-E(x|f)} \tag{2}$$

where x is the segmentation results for one frame; f is the original frame; *E(x)* is the energy function, which is composed of unary potential and pairwise potential; $Z(f)$ is the partition function which is just the sum of all $e^{-E(x|f)}$. The energy function is calculated as the following equation:

$$E(X|f) = \sum_i \psi_u(x_i) + \sum_{i,j} \psi_p(x_i, y_j) \tag{3}$$

where $\sum_i \psi_u(x_i)$ and $\sum_{i,j} \psi_p(x_i, y_j)$ are unary and pairwise potential, respectively; x_i, y_j are segmentation results for pixels of the frame f.

E. One-shot learning-based animal video segmentation

In our proposed one-shot learning-based animal video segmentation, the selected guidance frame is manually labeled and fed into the XFCN for fine-tuning. The XFCN is pre-trained on ImageNet [26], extended PASCAL VOC 2012 [8], and DAVIS 2016 [21] training datasets before fine-tuning. After fine-tuning, CNN features were extracted from XFCN. Then five different layers' features were fused for animal video segmentation. In the final step, the obtained segmentation results from XFCN were post-processed by using TTA and CRF to further improve the segmentation accuracy.

Considering that there are usually a larger number of background pixels than that of foreground (i.e. animal) pixels in animal videos, in our model training, a class-balanced loss function [27] was used to solve the problem of imbalance between the foreground and background. The loss function equation was listed as below:

$$\mathcal{L}_{mod} = -\beta \sum_{j \in Y_+} \log P\left(y_j = 1 \mid f\right) - (1-\beta) \sum_{j \in Y_-} \log P\left(y_j = 0 \mid f\right) \tag{4}$$

where f is the input frame, $y_i \in \{0, 1\}$, Y_+ and Y_- represent pixels with positive and negative labels respectively; Y represents all pixels and $Y = Y_+ + Y_-$; $\beta = \frac{|Y_-|}{|Y|}$ is the key weighting factor for balancing the pixels.

III. Experimental setup

A. The used dataset

To validate the proposed one-shot learning-based animal video segmentation approach, experiments were conducted on DAVIS 2016 animal dataset. The DAVIS 2016 animal dataset was constructed by selecting the animal-related videos from public-opened DAVIS 2016 validation dataset [21], which includes seven different animal videos (i.e. blackswan, camel, cows, dog, goat, horse and libby). Each video has different frame numbers varying from 49 to 104 with the resolution 854×480 (Table I). The segmentation for these animal videos is challenging considering that the complex background, occlusions, motion blur, and shadow influence.

TABLE I: Description of the dataset used in the experiments

Animal video	No. frames	Resolution	Description
Blackswan	50	854×480	Dynamic background Complicated shape
Camel	90		Confusing background Complicated shape
Cow	104		Complex background Occlusions
Dog	60		Blurry video Confusing background
Goat	90		Complicated shape Confusing background
Horse	50		Fast motion Occlusions
Libby	49		Blurry background Strong occlusions

B. Network training and fine-tuning details

In our work, all experiments were conducted on a computer equipped with NVIDIA RTX 2080Ti GPU and Ryzen 5 3600 CPU@3.6 GHz. The proposed one-shot learning-based approach was firstly pre-trained on several open datasets, then fine-tuned with the labeled guidance frame for the video segmentation. According to the datasets used, the pre-training process can be classified as base training and objectness training.

- Base training: XFCN was trained on the PASCAL VOC 2012 dataset [8] with 632 images and an extended dataset with 11,208 training images [28]. For the base training on these two datasets, flipping and zooming in data augmentation were used. The used optimization algorithm was Stochastic Gradient Descent (SGD) with learning rate of 1e-6, and 25000 iterations occurred at this stage. After base training on these two large image segmentation datasets, the network has ability to segment foreground objects from the background.
- Objectness training: Although the network can segment objects from the background after base training, it still has some noise and blurry contour in the segmented images. Therefore, the XFCN is further pre-trained using DAVIS 2016 training dataset for pixel objectness. Noticeably, animal types in the test videos are not included in the training videos. Due to the relatively small size of used dataset (30 videos), a data augmentation (i.e. random flipping, zooming in, cropping, brightness and contrast change) was implemented to avoid over-fitting. For the objectness training, SGD optimization algorithm with momentum 0.9 was adopted, and the learning rate was gradually decreased from 10^{-6} to 2.5×10^{-7}. The whole objectness training process had 20000 iterations.

After pre-training, the proposed network was fine-tuned using the guidance frame (manually labeled) from testing videos. Here, data augmentation–random flipping, zooming, cropping, brightness and contrast change were implemented. The learning rate was set to 10^{-7}. As the fine-tuning time has a large influence on efficiency, our proposed XFCN uses light-weight architecture and separable convolution to accelerate the fine-tuning process. In addition, different fine-tuning iterations were also investigated. According to our experiments, 200 iterations can achieve the best trade-off between speed and accuracy for DAVIS 2016 animal dataset.

C. Metrics

In our experiments, three popular metrics from DAVIS video segmentation competition, namely, region similarity regarding intersection over union $\mathcal{J}$, contour accuracy $\mathcal{F}$ and temporal instability of the masks $\mathcal{T}$ were used in the measures of our experimental results.

Given Y is an output segmentation and G is the corresponding ground-truth mask, $\mathcal{J}$ is defined as $\mathcal{J} = \frac{Y \cap G}{|Y \cup G|}$, which measures the how well the pixels of ground truth and prediction match. The contour accuracy $\mathcal{F}$ contains contour-based precision P_c and recall R_c, which is define as $\mathcal{F} = \frac{2P_cR_c}{P_c+R_c}$. The temporal instability $\mathcal{T}$ is used to estimate the deformation between frames. The occlusions and very strong deformations will lead to high temporal instability.

In our experiments, the values of mean and recall were calculated for both $\mathcal{J}$ and $\mathcal{F}$. Here, mean is the average results of all frames in the video; recall is to calculate the average results only for frames with a high score over a threshold (0.5 in our experiments).

IV. Experimental results

A. *Animal video segmentation results*

Our proposed approach was firstly compared with the state-of-art OSVOS and OSMN methods on DAVIS 2016 animal dataset [21]. Here, the evaluation measures of OSVOS and OSMN were directly computed from their published segmentation video results [14, 15]. For OSVOS + GFS, we integrated the same GFS module as our method to OSVOS. As illustrated in Table II, the achieved $\mathcal{J}$ (Mean) of our proposed approach was 89.5%, which is 0.8%, 2.2%, and 10.5% higher than that of the OSVOS + GFS, OSVOS and OSMN respectively; the $\mathcal{F}$ (Mean) of our method is 93.2%, which is 3.2% higher than OSVOS + GFS, 1.7% over the OSVOS, and 12.0% better than the OSMN. In addition, our one-shot learning-based video segmentation achieves 100% recall of $\mathcal{J}$ and $\mathcal{F}$ with the lower $\mathcal{T}$ (8.3%), which outperforms the OSVOS + GFS, OSVOS and OSMN. Compared to OSVOS + GFS, OSVOS and OSMN, our approach achieves better overall performance in the animal video segmentation task. As the run-time of CRF module accounts for the large proportion in the total computing time, our proposed approach can work in two modes: the fast mode can be applied in real-time by removing the CRF from the POST module; the slow mode (complete model) can achieve the highest accuracy with more fine-tuning iterations, which is suitable for the scenario that the object to be segmented is known beforehand.

In terms of time efficiency, according to Table II, in fast mode (Ours-CRF), our approach achieves an $\mathcal{J}(Mean)$ of 88.7% and an $\mathcal{F}(Mean)$ 92.6% with a speed of 0.57 s/frame, which outperforms OSVOS (87.3% $\mathcal{J}$ and 91.5% $\mathcal{F}$ with roughly 9 seconds per frame). With shorter running time, it also obtains 2.6% higher $\mathcal{F}(Mean)$ and lower $\mathcal{T}$ (7.4%) than OSVOS + GFS. In terms of OSMN, although it is slightly faster than our model, the $\mathcal{J}(Mean)$ and $\mathcal{F}(Mean)$ of OSMN are 9.7% and 11.4% below our approach.

Fig. 3 shows the qualitative segmentation samples of our proposed approach on libby, goat and cows sequences. It can be seen that animals are segmented with high accuracy, even when they are walking or heavily occluded by trees. Additionally, Fig. 4 demonstrates the comparison of different methods' qualitative results on four sequences in DAVIS 2016 animal dataset. Our approach has fewer outliers, false predictions and sharper contours than OSVOS, OSMN in terms of the occlusion, fast movement and confusion colors in the scene.

TABLE II: Comparison of video segmentation results (%)

Methods / Metrics	OSMN	OSVOS	OSVOS+GFS	Ours	Ours-CRF (Fast mode)
$\mathcal{J}$(Mean)↑	79.0	87.3	88.7	**89.5**	88.7
$\mathcal{J}$(Recall)↑	90.7	99.1	99.0	**100.0**	100.0
$\mathcal{F}$(Mean)↑	81.2	91.5	90.0	**93.2**	92.6
$\mathcal{F}$(Recall)↑	90.2	100.0	96.3	**100.0**	100.0
$\mathcal{T}$↓	9.3	13.8	8.5	8.3	**7.4**
Time/frame (s)	**0.14**	9.00	1.17	1.16	0.57

* Ours is the complete model (slow mode).
** Ours-CRF means the approach without CRF module (fast mode).
*** In OSVOS+GFS, the used post processing is CRF. Note that since the code of OSVOS with its original post-processing module is not publicly available, here we use CRF instead for comparison purposes.

V. Ablation study

Our approach contains several vital modules, namely, GFS, XFCN, and POST, which contribute to improving the video segmentation results. For analyzing and quantifying the significance and effects of each module in our approach, comparison experiments were conducted by removing some modules, and those ablated versions were applied to the DAVIS 2016 animal dataset [21].

A. *The influence of GFS and POST*

We firstly investigated the influence of GFS and POST for video segmentation performance. Table III illustrates the segmentation results with and without GFS and POST modules.

It can be seen that an $\mathcal{J}(Mean)$ of 87.7% and an $\mathcal{F}(Mean)$ of 91.1% are achieved without GFS, which is 1.8% and 2.1% lower than that of complete model respectively. When

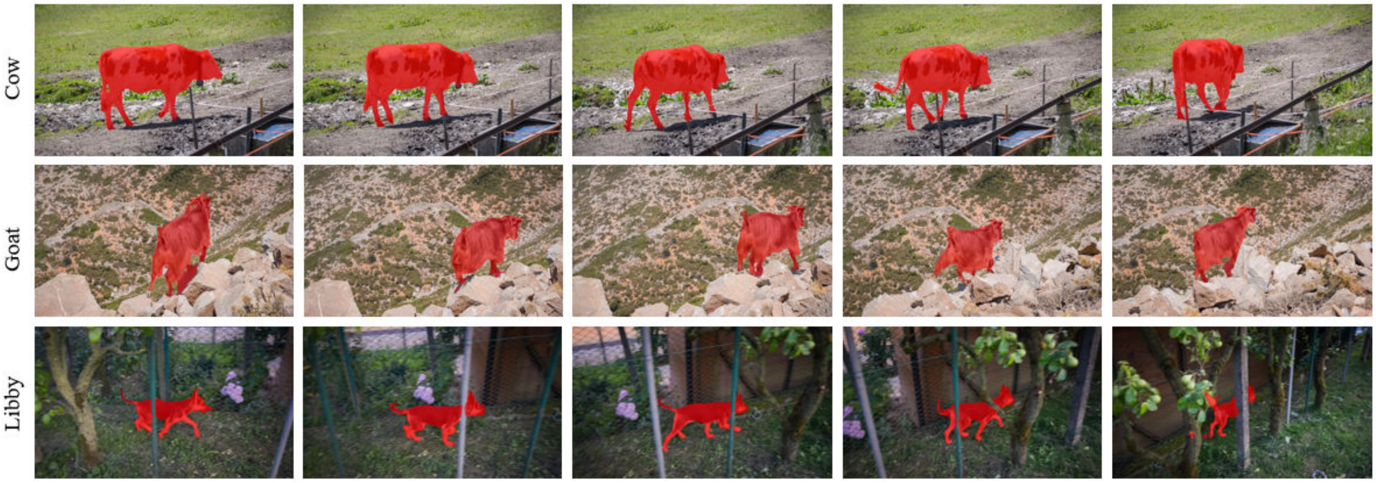


Fig. 3: Qualiative segmentation samples of the proposed one-shot learning based video segmentation.

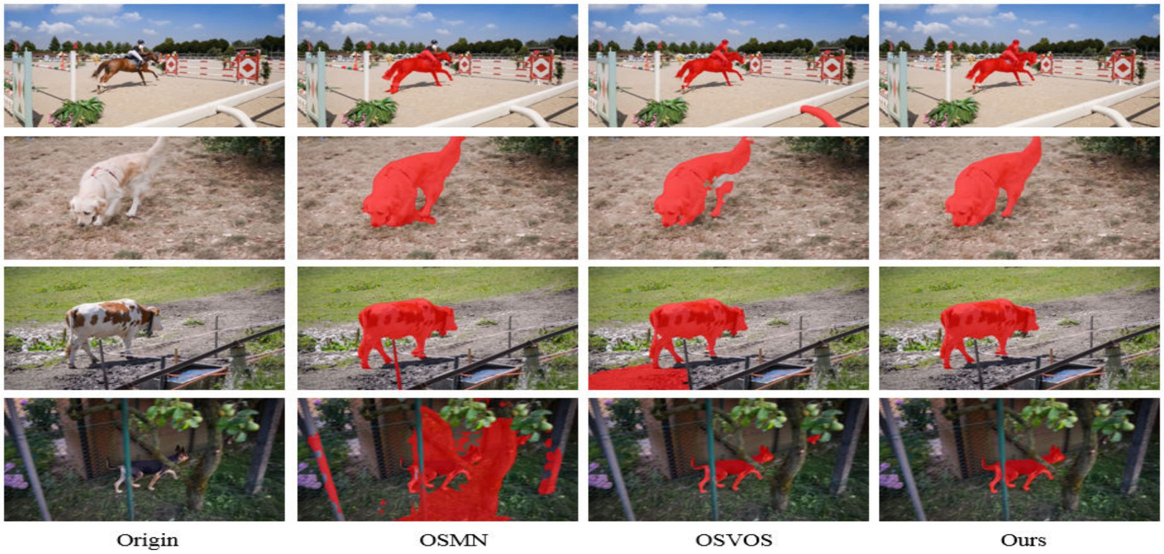


Fig. 4: Comparison of segmentation results among OSVOS, OSMN and our proposed approach.

the POST module is removed, the achieved $\mathcal{J}(Mean)$ and $\mathcal{F}(Mean)$ are 88.0% and 92.0%, which is 1.5% and 1.2% lower than that of the complete model, respectively. It demonstrates that GFS plays a more important role in our video segmentation approach than POST. GFS and POST modules contribute 4.6% accuracy on $\mathcal{J}(Mean)$. In addition, the lowest time instability occur at the complete model. Therefore, the segmentation stability in video sequences has been enhanced with these two extended modules as well.

TABLE III: Comparison with pipelines without some sub-blocks (%)

Methods Metrics	Ours	-GFS	-POST	-GFS-POST
$\mathcal{J}$(Mean)↑	**89.5**	87.7	88.0	84.9
$\mathcal{F}$(Mean)↑	**93.2**	91.1	92.0	87.6
$\mathcal{T}$ ↓	**8.3**	9.7	8.5	9.6

⋇ '-GFS' means the proposed approach without GFS module; '-POST' means the proposed approach without POST module; '-GFS-POST' means the proposed approach without GFS and POST modules.

B. The influence of pre-training and fine-tuning

In order to further analyze the influence of pre-training and fine-tuning, the performance of our approach without base training (BT), objectness-training (OT) and fine-tuning (FT) were investigated. Note that GFS and POST were implemented for all cases in Table IV. As shown in Table IV, for DAVIS 2016 animal dataset, BT, OT and FT improve the performance 2.0%, 4.5% and 19.7% of $\mathcal{J}(Mean)$ and 2.1%, 5.4% and 21.0% of $\mathcal{F}(Mean)$. Obviously, FT plays the most significant role in improving the model's overall performance, since FT informs the model what the specific object is. The worst time stability also occurs at the pipeline without FT and the complete model has the lowest $\mathcal{T}$ value (the best time stability).

TABLE IV: Comparison with pipelines without one of training process (%)

Methods Metrics	Ours	-BT	-OT	-FT
$\mathcal{J}$(Mean)↑	**89.5**	87.5	85.0	69.8
$\mathcal{F}$(Mean)↑	**93.2**	91.1	87.8	72.2
$\mathcal{T}$↓	**8.3**	8.7	9.6	25.9

⋇ '-BT' means the proposed approach without base training; '-OT' means the proposed approach without objectness training; '-FT' means the proposed approach without fine-tuning.

VI. Conclusions and Future Work

An accurate and real-time animal video segmentation method is crucial to automatic behavior and health monitoring of animals. For segmenting animal in the video with only one labeled frame, an effective one-shot learning-based video segmentation approach was proposed. Our approach selected one guidance frame for manual labeling to leverage the effects of the fine-tuning process on test videos. Xception-based FCN was used to extract features and generate segmentation results. Then the post-processing module improved the segmentation results by removing noise and sharpening the contour. According to the experiments on DAVIS 2016 animal, our proposed approach achieved 89.5% $\mathcal{J}(Mean)$ with a relatively fast speed, and outperformed the state-of-art methods OSVOS (87.3% $\mathcal{J}(Mean)$) and OSMN (79.0% $\mathcal{J}(Mean)$). Extensive ablation studies have been performed to validate the role of several modules in enhancing the performance of the whole approach. Overall, our approach achieved a good trade-off between speed and accuracy for one-shot animal video segmentation, and can be adapted to applications with different timing requirements. For future work, the spatial-temporal information in video sequences will be exploited to potentially improve the segmentation performance and computational efficiency.

References

[1] C.-C. Wong, Y. Gan, and C.-M. Vong, "Efficient outdoor video semantic segmentation using feedback-based fully convolution neural network," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5128–5136, 2019.

[2] J. Zhang, J. Chen, Z. Wang, S. Chen, and J. Zhang, "Detection and segmentation of unlearned objects in unknown environment," *IEEE Transactions on Industrial Informatics*, 2020.

[3] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, "Youtube-vos: Sequence-to-sequence video object segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 585–601.

[4] S. Eiffert, N. Wallace, H. Kong, N. Pirmarzdashti, and S. Sukkarieh, "Resource and response aware path planning for long-term autonomy of ground robots in agriculture," *Field Robotics*, Accepted and to appear, 2021 (also available at arXiv:2105.10690).

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine iIntelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[8] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[9] Z. Abderrahmane, G. Ganesh, A. Crosnier, and A. Cherubini, "A deep learning framework for tactile recognition of known as well as novel objects," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 423–432, 2019.

[10] W. Zhang, Q. J. Wu, Y. Yang, T. Akilan, and H. Zhang, "A width-growth model with subnetwork nodes and refinement structure for representation learning and image classification," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 1562–1572, 2020.

[11] M. Zhang, H. Li, S. Pan, T. Liu, and S. Su, "One-shot neural architecture search via novelty driven sampling," in *Proceedings of the Twenty-Ninth International Joint Conference on Artiicial Intelligence Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization*, 2020, pp. 3188–3194.

[12] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3855–3865, 2020.

[13] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern rRecognition*, 2019.

[14] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 221–230.

[15] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos, "Efficient video object segmentation via network modulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6499–6507.

[16] B. A. Griffin and J. J. Corso, "Bubblenets: Learning to select the guidance frame in video object segmentation by deep sorting frames," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8906–8915.

[17] S. Zuffi, A. Kanazawa, and M. J. Black, "Lions and tigers and bears: Capturing non-rigid, 3d, articulated shape from images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3955–3963.

[18] Y. Zhao, L. Yang, S. Zheng, and B. Xiong, "Advances in the development and applications of intelligent equipment and feeding technology for livestock production," *Smart Agriculture*, vol. 1, no. 1, pp. 20–31, 2019.

[19] Y. Qiao, M. Truman, and S. Sukkarieh, "Cattle segmentation and contour extraction based on mask r-cnn for precision livestock farming," *Computers and Electronics in Agriculture*, vol. 165, p. 104958, 2019.

[20] Y. Yin, D. Xu, X. Wang, and L. Zhang, "Agunet: Annotation-guided u-net for fast one-shot video object segmentation," *Pattern Recognition*, vol. 110, p. 107580, 2021.

[21] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 724–732.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[24] N. Moshkov, B. Mathe, A. Kertesz-Farkas, R. Hollandi, and P. Horvath, "Test-time augmentation for deep learning-based cell segmentation on microscopy images," *Scientific reports*, vol. 10, no. 1, pp. 1–7, 2020.

[25] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Ad-*

vances in neural information processing systems, 2011, pp. 109–117.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[27] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

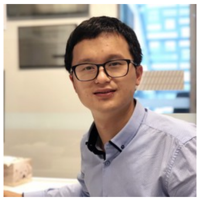
[28] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 991–998.

Tengfei Xue received his B.Eng. in Mechatronic Engineering from Beijing Jiaotong University, China and University of Wollongong, Australia, in 2019, and M.Eng. in Automation and Manufacturing Systems from University of Sydney, Australia in 2021. He is currently a Ph.D. candidate in the School of Computer Science, University of Sydney, Australia. His research interests include deep learning, pattern recognition, and computer vision.

Yongliang Qiao received B.S.in Electrical Engineering and Automation and M.S. degree in Agricultural Electrification and Automation from Northwest A&F University, Yangling, China, respectively. He received Ph.D. degree in computer science from the University of Technology of Belfort-MontbÃl'liard, France. Then he work as a research associate at the Australian Centre for Field Robotics (ACFR), the University of Sydney, Australia. Since 2021, he has been on the youth editorial board of the journal of Smart Agriculture. His research interests include computer vision, agricultural robots, deep learning, multi-sensor fusion, and pattern recognition.

He Kong received the Bachelor degree in Electrical Engineering from China University of Mining and Technology and Master degree in Control Science and Engineering from Harbin Institute of Technology (Centre for Control Theory and Guidance Technology), China. He then undertook doctoral studies at the Centre for Complex Dynamic Systems and Control, the University of Newcastle, Australia, and received the PhD degree in Electrical Engineering. He is currently a research fellow at the Sydney Institute for Robotics and Intelligent Systems as well as the Australian Centre for Field Robotics, the University of Sydney, Australia. His research interests centre on the intersections of systems and control, field robotics, and machine learning. He is particularly interested in optimization-based estimation/control, robot path planning, active perception, robot audition, and their applications in agriculture, environmental monitoring, etc. Since 2021, he has been on the youth editorial board of the journal of Smart Agriculture, sponsored by Agricultural Information Institute of Chinese Academy of Agricultural Sciences.

Daobilige Su received his B. Eng. in Mechatronic Engineering from Zhejiang University, China in 2010, M. Eng. in Automation and Robotics from Warsaw University of Technology, Poland and M. Eng. in Automation from University of Genova, Italy through European Master on Advanced Robotics (EMARO) program in 2012, and Ph. D. in robotics at Centre for Autonomous System (CAS), University of Technology Sydney (UTS), Australia in 2017. He was a post-doctoral research associate at Australian Centre for Field Robotics (ACFR), The University of Sydney from 2017 to 2020. He is currently an Associate Professor at College of Engineering, China Agricultural University, China. His current research areas include field robotics, SLAM, computer vision, robot audition and machine learning.



Shirui Pan is an ARC Future Fellow and a Senior Lecturer with the Faculty of Information Technology, Monash University, Australia. Shirui received his Ph.D degree in computer science from UTS, Australia. To date, Dr Pan has published over 100 research papers in top-tier journals and conferences, including TPAMI, TKDE, NeurIPS, and KDD. He is a (senior) program committee member and an invited reviewer for many top conferences and journals including NeurIPS, ICLR, KDD, ICML, TPAMI, and TKDE. Dr Pan's research interests include data mining and machine learning, specialized in graph mining and network analysis.

Khalid Rafique is a program manager with the Australian Centre for Filed Robotics (ACFR), The University of Sydney. He received Bachelor's degree in Avionics from National University of Sciences and Technology (NUST) and Master's in Interdisciplinary/Systems Engineering from Purdue University. His career history extends to defence, academia, and commercial industries, and includes successfully building, coaching, and managing cross-functional, geographically dispersed project teams on large and complex technology projects. His career interests are driven by both passion & dedication and focus on technologies those can bring better future to humanity and environment.

Salah Sukkarieh received a Bachelor degree in mechanical (mechatronics) engineering and the Ph.D. degree from the University of Sydney, Australia. He is currently Professor of Robotics and Intelligent Systems and Associate Dean (Industry and Innovation) of Faculty of Engineering at the University of Sydney. Salah is also the CEO of Agerris, a new AgTech startup company from the Australian Centre for Field Robotics (ACFR), developing autonomous robotic solutions to improve agricultural productivity and environmental sustainability. He was the Director of Research and Innovation at the ACFR from 2007 to 2018, where he led the strategic research and industry engagement program. He is an international expert in the research, development and commercialisation of field robotic systems and has led a number of robotics and intelligent systems R&D projects in logistics, commercial aviation, aerospace, education, environment monitoring, agriculture and mining. Salah was awarded the NSW Science and Engineering Award for Excellence in Engineering and Information and Communications Technologies in 2014, and the 2017 CSIRO Eureka Prize for Leadership in Innovation and Science, and the 2019 NSW Australian of the Year nominee. He is a Fellow of the Australian Academy of Technological Sciences and Engineering, and has served/is serving on the editorial board for Field Robotics, Autonomous Robots, amongst others.