

Actor-Critic Instance Segmentation

Nikita Araslanov*

Constantin A. Rothkopf^{†,§}

Stefan Roth^{*,§}

*Dept. of Computer Science

[†]Institute of Psychology

[§]Centre for Cognitive Science

TU Darmstadt

Abstract

Most approaches to visual scene analysis have emphasised parallel processing of the image elements. However, one area in which the sequential nature of vision is apparent, is that of segmenting multiple, potentially similar and partially occluded objects in a scene. In this work, we revisit the recurrent formulation of this challenging problem in the context of reinforcement learning. Motivated by the limitations of the global max-matching assignment of the ground-truth segments to the recurrent states, we develop an actor-critic approach in which the actor recurrently predicts one instance mask at a time and utilises the gradient from a concurrently trained critic network. We formulate the state, action, and the reward such as to let the critic model long-term effects of the current prediction and incorporate this information into the gradient signal. Furthermore, to enable effective exploration in the inherently high-dimensional action space of instance masks, we learn a compact representation using a conditional variational auto-encoder. We show that our actor-critic model consistently provides accuracy benefits over the recurrent baseline on standard instance segmentation benchmarks.

1. Introduction

Methods for instance segmentation have for the most part relied on the idea of parallel processing of the image elements and features within images [13]. However, previous work [31, 32] suggests that instance segmentation can be formulated as a sequential visual task, akin to human vision, for which substantial evidence has revealed that many vision tasks beyond eye movements are solved sequentially [36]. While the segmentation accuracy of feed-forward pipelines hinges on a large number of object proposals, proposal-free recurrent models have a particular appeal for instance segmentation where the number of instances is unknown. Also, the temporal context can facilitate a certain order of prediction: segmenting “hard” instances can be improved by conditioning on the masks of “easy” instances segmented first (e.g., due to occlusions,

ambiguities in spatial context etc.; [20]).

A pivotal question of a recurrent formulation for instance segmentation is the assignment of the ground-truth segments to timesteps, since the order in which they have to be predicted is unknown. Previously this was addressed using the Kuhn-Munkres algorithm [18], computing the max-matching assignment. We provide some insight, however, that the final prediction ordering depends on the initial assignment. Furthermore, the loss for every timestep is not informative in terms of its effect on future predictions. Intuitively, considering the future loss for the predictions early on should improve the segmentation accuracy at the later timesteps. Although this can be achieved by unrolling the recurrent states for gradient backpropagation, such an approach quickly becomes infeasible for segmentation networks due to high memory demands.

In the past years, reinforcement learning (RL) has been showing promise in solving increasingly complex tasks [23, 26, 27]. However, relatively little work has explored applications of RL outside its conventional domain, which we attribute to two main factors: (1) computer vision problems often lack the notion of the environment, which provides the interacting agent with the reward feedback; (2) actions in the space of images are often prohibitively high-dimensional, leading to tough computational challenges.

Here, we use an actor-critic (AC) model [5] to make progress regarding both technical issues of a recurrent approach to instance segmentation. We use exploration noise to reduce the influence of the initial assignments on the segmentation ordering. Furthermore, we design a reward function that accounts for the future reward in the objective function for every timestep. Our model does not use bounding boxes – often criticised due to their coarse representation of the objects’ shape. Instead, we built on an encoder-decoder baseline that makes pixelwise predictions directly at the scale of the input image. To enable the use of RL for instance segmentation with its associated high-dimensional output space, we propose to learn a compact action-space representation through the latent variables of a conditional variational auto-encoder [16], which we integrate into a recurrent prediction pipeline.

Our experiments demonstrate that our actor-critic model improves the prediction quality over its baseline trained with the max-matching assignment loss, especially at the later timesteps, and performs well on standard instance segmentation benchmarks.

2. Related Work

Instance segmentation has received growing attention in the recent literature. One family of approaches focuses on learning explicit instance encodings [4, 9, 17, 24, 35], which are then clustered into individual instance masks using post-processing. Another common end-to-end approach is to first predict a bounding box for each instance using dynamic pooling and then to produce a mask of the dominant object within the box using a separate segmentation network [13, 22]. These methods are currently best-practice, which can be attributed to the maturity of deep network-based object detection pipelines. However, this strategy is ultimately limited by the detection performance, proposal set, and the need of additional processing to account for pixel-level context [2, 8].

Making the predictions sequentially points at an alternative line of work. Romera-Paredes & Torr [32] used a convolutional LSTM [39] with a spatial softmax, which works well for isotropic object shapes and moderate scale variation. At each timestep, the recurrent model of Ren & Zemel [31] predicts a box location and scale for one instance. However, the extent of the available context for subsequent segmentations is limited by the box. Some benefits of the temporal and spatial context have been also re-asserted on the task of object detection [7, 21] and, much earlier, on image generation [12] and recognition [19]. In contrast to these works, our method obviates the need for the intermediate bounding box representation and predicts masks directly at the image resolution.

We cast the problem as a sequential decision process, as is studied by reinforcement learning (RL; [34]). Using the actor-critic framework [5], we define the actor as the model that sequentially produces instance masks, whereas the critic learns to provide a score characterising the actor’s performance. Leveraging this score, the actor can be trained to improve the quality of its predictions. This is reminiscent of the more recent Generative Adversarial Networks (GANs; [11]), in which a generator relies on a discriminator to improve. In particular, our model is similar to Wasserstein GANs [1] in that the discriminator is trained on a regression-like loss, and to SeqGAN [40] in that the generator’s predictions are sequential.

One obstacle is the action dimensionality, since the sampling complexity required for exploration grows exponentially with the size of the actions. A naive action representation for dense pixelwise predictions would lead to an action space of dimension in the order of $O(2^{H \times W})$ for im-



Figure 1. *Illustration of the max-matching assignment for instance segmentation.* Consider an image with ground-truth instances a and b , and a recurrent model making predictions 1 and 2. In the constructed bipartite graph, each edge is assigned a weight corresponding to the IoU of the prediction with the connected ground truth. From the set of possible assignments, depicted by the orange and grey edges, max-matching finds the one that maximizes the sum of the IoUs. The loss is then computed independently for each timestep w.r.t. to this assignment to the ground truth.

ages with resolution $H \times W$. This is significantly higher than the action spaces of standard problems studied by reinforcement learning (usually, between 1 and 20), or even its applications to natural language processing [3, 30]. To address this, we suggest learning a compact representation using variational auto-encoders [16] to enable the crucial reduction of the problem from a high-dimensional discrete to a lower-dimensional continuous action space.

3. Motivation

As discussed above, we follow previous work in modelling instance segmentation as a sequential decision problem [31, 32], yielding one instance per timestep t .

To motivate our work, we revisit the standard practice of using the Kuhn-Munkres algorithm [18] to assign the ground-truth instances to the predictions of a recurrent model. Let θ parametrise the model and $U_\theta \in \mathbb{R}^{n \times n}$ denote a matrix of elements u_{ij} measuring the score of the i^{th} prediction w.r.t. the j^{th} ground truth (e.g., the IoU). The Kuhn-Munkres algorithm finds a permutation matrix as a solution to the max-matching problem

$$\arg \max_{P \in \mathcal{P}} \text{tr}(U_\theta P), \quad (1)$$

where \mathcal{P} is the set of n -dimensional permutation matrices, i.e. such that for all $P \in \mathcal{P}$, we have $\sum_j P_{ij} = 1$, $\sum_i P_{ij} = 1$, $P_{ij} \in \{0, 1\}$. Given a differentiable loss function $l_\theta(i, j)$ (e.g., the binary cross-entropy), the model parameters θ are then updated to minimise $\sum_{i,j, P_{ij}=1} l_\theta(i, j)$.

Consider a simple case of two ground-truth segments, a and b , illustrated in Fig. 1. Without loss of generality, assume that the initial (random) model parameters yield $u_{1a} + u_{2b} < u_{1b} + u_{2a}$, i.e. the sum of scores for segmenting instance a first and b second is lower than in the opposite order. This implies that max-matching will perform a gradient update step maximising the second sum, i.e. $u_{1b} + u_{2a}$, but not the first. As a consequence, for the updated parameters, the score for the ordering $b \rightarrow a$ is likely to dominate also

at the later iterations of training.¹

Previous work [20, 37] suggests that sequential models are not invariant to the order of predictions, including object segments (*c.f.* supplemental material). The implication from the example above is that $\sup u_{1a} + u_{2b} \neq \sup u_{1b} + u_{2a}$ (the sup is w.r.t. θ). One conceivable remedy to alleviate the effect of the initial assignment is to introduce noise ϵ to the score matrix U (*e.g.*, *i. i. d.* Gaussian), such that Eq. (1) becomes

$$\arg \max_{P \in \mathcal{P}} \text{tr}((U_\theta + \epsilon)P). \quad (2)$$

However, the noise in the loss function will not account for the inherent causality of the temporal context in recurrent models: perturbation of one prediction affects the consecutive ones.

In this work, we consider a more principled approach to encourage exploration of different orderings. We inject *exploration noise* at the level of individual predictions made by an *actor* network, while a jointly trained *critic* network keeps track of the long-term outcome of the early predictions. This allows to include in the gradient to the actor not only the immediate loss, but also the contribution of the current prediction to the future loss. We achieve this by reformulating the instance segmentation problem in the RL framework, which we briefly introduce next.

4. Notation and Definitions

In the following, we define the key concepts of Markov decision processes (MDPs) in the context of instance segmentation. For a more general introduction, we refer to [34].

We consider finite-horizon MDPs defined by the tuple (S, A, T, r) , where the state space S , the action space A , the state transition $T : S \times A \rightarrow S$, and the reward $r : S \times A \rightarrow \mathbb{R}$ are defined as follows.

State. The state $s_t \in S$ of the recurrent system is a tuple of the input image (and its task-specific representations) and an aggregated mask, *i.e.* $s_t = (I, M_t)$. The mask M_t simply accumulates previous instance predictions, which encourages the model to focus on yet unassigned pixels. Including I enables access to the original input at every timestep.

Action. To limit the dimensionality of the action space, we define the action $a_t \in A$ in terms of a compact mask representation. To achieve this, we pre-train a conditional variational auto-encoder (cVAE; [16]) to reproduce segmentation masks. As a result, the action $a_t \in A = \mathbb{R}^l$ is a continuous latent representation of a binary mask and has dimensionality $l \ll H \cdot W$, while the decoder $\mathcal{D} : \mathbb{R}^l \rightarrow \mathbb{R}^{H \times W}$ “expands” the latent code to a full-resolution mask.

¹Note that a formal proof is likely non-trivial due to the stochastic nature of training.

State transition. As implied by the state and action definitions above, the *state transition*

$$T((I, M_t), a_t) = (I, \max(M_t, \mathcal{D}(a_t))) \quad (3)$$

uses a pixelwise max of the previous mask and the decoded action, *i.e.* integrating the currently predicted instance mask into the previously accumulated predictions.

Reward. We design the *reward function* to measure the progress of the state transition towards optimising a certain segmentation criterion. The building block of the reward is the *state potential* [28], which we base on the max-matching assignment of the current predictions to the ground-truth segments, *i.e.*

$$\phi_t := \max_{k \in \mathcal{P}(N)} \sum_{i=1}^t \mathcal{F}(\mathcal{S}_i, \mathcal{T}_{k_i}), \quad (4)$$

where \mathcal{T}_i and $\mathcal{S}_{1 \leq i \leq t}$ are the N ground-truth masks and t predicted masks; $\mathcal{P}(N)$ is a collection of all permutations of the set $\{1, 2, \dots, N\}$. $\mathcal{F}(\cdot, \cdot)$ denotes a distance between the prediction and a ground-truth mask and can be chosen with regard to the performance metric used by the specific benchmark (*e.g.*, IoU, Dice, etc.). We elaborate on these choices in the experimental section.

The state potential in Eq. (4) allows us to define the reward as the difference between the potentials of subsequent states

$$r_t := \phi(s_{t+1}) - \phi(s_t). \quad (5)$$

Note that since the $(t+1)^{\text{st}}$ prediction might re-order the optimal assignment (computed with the Kuhn-Munkres algorithm), our definition of the reward is less restrictive w.r.t. the prediction order compared to previous work [31, 32], which enforces a certain assignment to compute the gradient. Instead, our immediate reward allows to reason about the relative improvement of one set of predictions over another.

5. Actor-Critic Approach

5.1. Overview

The core block of the actor model, shown in Fig. 2, is a conditional variational auto-encoder (cVAE; [16]). The encoder computes a compact vector of latent variables, encoding a full-resolution instance mask. The decoder recovers such a mask from the latent code. Using the transition function defined by Eq. (3), the latest prediction updates the state, and the procedure repeats until termination.

The actor relies on two types of context with complementary properties. As discussed above, the mask M_t is a component of the state s_t , which accumulates the masks produced in the previous steps. It provides permutation-invariant temporal context of high-resolution cues, encouraging the network to focus on yet unlabelled pixels. The

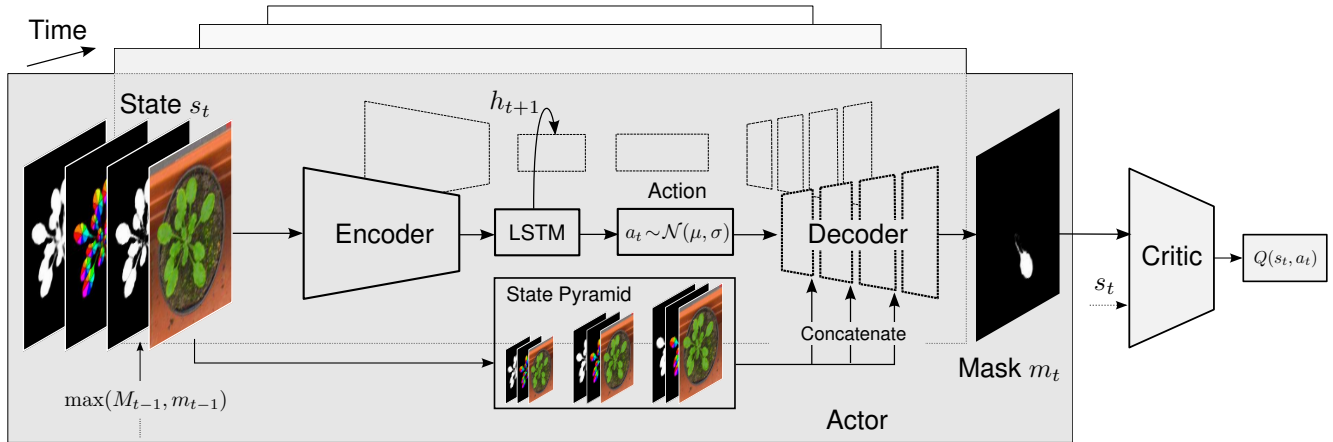


Figure 2. *Actor-critic model for instance segmentation.* The model relies on two types of context: a spatial permutation-invariant state s_t accumulates the masks, whereas the hidden LSTM state h_t models a temporal context sensitive to the prediction ordering. The State Pyramid propagates the high-res information at multiple scales to the decoder to compensate the loss of resolution at the bottleneck section.

hidden state of actor h_t is implemented by the LSTM [14] at the bottleneck section and is unknown to the critic. In contrast to state s_t , the representation of the hidden state is learned and can be sensitive to the prediction ordering due to the non-commutative updates of the LSTM state. The hidden state, therefore, contributes to the temporal context and is shown to be particularly helpful for counting in the ablation study.

We train our model in two stages as described next.

5.2. Pre-training

We pre-train the actor cVAE to reconstruct the mask of the target segment. The input to the network consists of the image and the binary mask of a randomly chosen ground-truth instance. To account for the loss of high-resolution information at the latent level, the decoder is conditioned on the input image and auxiliary channels of instance-relevant representations supplied at multiple scales, which we term *State Pyramid*. One channel contains the foreground prediction, while the other 8 channels encode the instance angle quantisation of [35], thereby binning the pixels of the object segment into quantised angles relative to the object’s centroid. These features assist in instance detection and disambiguation, since a neighbourhood of orthogonal quantisation vectors indicates occlusion boundaries and object centroids. Following Ren & Zemel [31], we predict the angles with a pre-processing network [25] trained in a standalone fashion.

The auto-encoder uses the binary cross-entropy (BCE) as the reconstruction loss. For the latent representation, we use a Gaussian prior with zero mean and unit variance. The corresponding loss function is taken as the Kullback-Leibler divergence [16].

5.3. Training

During training, we learn a new encoder to sequentially predict segmentation masks. In addition to the image, the encoder also receives the auxiliary channels used during pre-training. In contrast, however, the encoder is encouraged to learn instance-sensitive features, since the decoder expects the latent code of ground-truth masks.

Algorithm 1 provides an outline of the training procedure. The actor is trained jointly with the critic from a buffer of experience accumulated in the *episode execution* step. In the *policy evaluation*, the critic is updated to minimise the error of approximating the expected reward, while in the *policy iteration* the actor receives the gradient from the critic to maximise the Q-value.

Episode execution. For an image with N instances, we define the episode as the sequence of N predictions. The algorithm randomly selects a mini-batch of images without replacement and provides it as inputs to the actor. Using the reparametrisation trick [16], the actor samples an action corresponding to the next prediction of the instance mask. The results of the predictions and corresponding rewards are saved in a buffer. At the end of each episode, the target Q-value can be computed for each timestep t as a sum of immediate rewards.

Policy evaluation. The critic network parametrised by ϕ , maintains an estimate of the Q-value defined as a function of the state and action guided by the policy μ :

$$Q_\phi(s_t, a_t) = \mathbb{E}_{a_j \sim \mu(s_j), j > t} \left[\sum_{i=t}^N \gamma^{i-t} r_i(s_i, a_i) \right]. \quad (6)$$

Note the finite sum in the expectation due to the finite number of instances in the image. The critic’s loss $\mathcal{L}_{\text{Critic}, t}$ for

timestep t is defined as the squared \mathcal{L}_2 -distance with a discounted sum of rewards

$$\mathcal{L}_{\text{Critic},t} = \left\| Q_\phi(s_t, a_t) - \sum_{i=t}^N \gamma^{i-t} r_i \right\|_2^2, \quad (7)$$

where $\gamma \in (0, 1)$ is a *discount factor* that controls the *time horizon*, i.e. the degree to which future rewards should be accounted for. The hyperparameter γ allows to trade off the time horizon for the difficulty of the reward approximation: as $\gamma \rightarrow 1$, the time horizon extends to all states, but the critic has to approximate a distant future reward based only on the current state and action. We update the parameters of the critic to minimise Eq. (7) using the samples of the state-actions and rewards in the buffer, and set $\gamma = 0.9$ throughout our experiments.

Policy iteration. The actor samples an action $a_t \in A$ from a distribution provided by the current policy $\mu_\theta : S \rightarrow A$, parametrised by θ and observes a reward r_t computed by Eq. (5). Given the initial state s_1 , the actor’s goal is to find the policy maximising the expected total reward, $\theta^* = \arg \max_\theta \mathbb{E}_{a_j \sim \mu_\theta(s_j)} [\sum_{i=1}^N \gamma^i r_i(s_i, a_i)]$, approximated by the critic. To achieve this, the state $s_t = (I, M_t)$ and the actor’s mask prediction m_t are passed to the critic, which produces the corresponding Q-value. The gradient maximising the Q-value is computed via backpropagation and returned to the actor for its parameter update.

We found that fixing the decoder during training led to faster convergence. Since the critic only approximates the true loss, its gradient is biased, which in practice can break the assumption we maintain during training – that an optimal mask can be reconstructed from a normally-distributed latent space. We fix the decoder and maintain the KL-divergence loss \mathcal{L}_{KL} while sampling new actions, thus encouraging exploration of the action space. In our ablation study, we verify that such exploration improves the segmentation quality. Note that we do not pre-define the layout of the actions, but only maintain the Gaussian prior.

To further improve the stability of the joint actor-critic training, we use a *warm-up phase* for the critic: episode execution and update of the critic happen without updating the actor for a number of epochs. This gives the critic the opportunity to adapt to the current action and state space of the actor. We could confirm in our experiments that pre-training the decoder was crucial; omitting this step resulted in near-zero rewards from which it proved to be difficult to train the critic even with the warm-up phase.

Termination. We connect the hidden state h_t and the last layer preceding it (via a skip connection) to a single unit predicting “1” to continue prediction, and “0” to indicate the terminal state. Using the ground-truth number of instances, we train this unit with the BCE loss.

Algorithm 1: Actor-critic training

```

Initialise actor  $\mu_\theta(s)$  from pre-training and critic  $Q_\phi(s, a)$ 
for epoch = 1, NumEpochs do
    foreach minibatch do
        // accumulate buffer for replay
        buffer  $\leftarrow []$ 
        foreach (Image,  $\{\mathcal{T}\}_{1,\dots,N}$ ) in minibatch do
            Initialise mask  $M_1 \leftarrow \text{Empty}$ 
            Initialise state  $s_1 \leftarrow (\text{Image}, M_1)$ 
            episode  $\leftarrow []$ 
            for  $t = 1, N$  do
                Sample action  $a_t \sim \mu_\theta(s_t)$ 
                Obtain next state  $s_{t+1} = T(s_t, a_t)$  with Eq. (3)
                Add  $(s_t, a_t, s_{t+1})$  to episode
            end
            Compute rewards for episode with Eq. (5)
            Add episode with rewards to buffer
        end
        // Batch-update critic from buffer
        foreach  $(s_t, a_t, r_t, s_{t+1})$  in buffer do
             $\phi \leftarrow \phi - \alpha_{\text{critic}} \nabla_\phi (Q_\phi(s_t, a_t) - \sum_{i=t}^N \gamma^{i-t} r_i)^2$ 
        end
        // Batch-update actor using critic
        Initialise states  $s_1$  from buffer
        for  $t = 1, N$  do
            Sample action  $a_t \sim \mu_\theta(s_t)$ 
             $\theta \leftarrow \theta + \alpha_{\text{act}} \nabla_{a_t} Q_\phi(s_t, a_t) \nabla_\theta \mu_\theta(s_t) - \beta_{\text{act}} \nabla_\theta \mathcal{L}_{\text{KL}}$ 
            Obtain next state  $s_{t+1} = T(s_t, a_t)$  using Eq. (3)
        end
    end
end

```

Inference. We recurrently run the actor network until the termination prediction. To obtain the masks, we discard the deviation part and only take the mean component of the action predicted by the encoder and pass it through the decoder. We do not use the critic network at inference time.

Implementation details.² We use a simple architecture similar to [32] for both the critic and the actor networks trained with Adam [15] until the training loss on validation data does not improve (*c.f.* supplemental material).

5.4. Discussion

In the actor-critic model the critic plays the role of modelling the subsequent rewards for states $s_{i>t}$ given state s_t . Hence, if the critic’s approximation is exact, the backpropagation through time (BPTT; [38]) until the first state is not needed: to train the actor, we need to compute the gradient w.r.t. the future rewards already predicted by the critic. The implication of this property is that memory-demanding networks, such as those for dense prediction, can be effectively trained with truncated BPTT and the critic, even in case of long sequences. Moreover, using the critic’s approximation allows the reward be a non-differentiable, or even a discontinuous function tailored specifically to the task.

²Code is available at <https://github.com/visinf/acis/>.

6. Experiments

In our experiments, we first quantitatively verify the importance of the different components in our model and investigate the sources of the accuracy benefits of the actor-critic over the baseline. Then, we use two standard datasets of natural images for the challenging task of instance segmentation, and compare to the state of the art.

6.1. Ablation study

We design a set of experiments to investigate the effect of various aspects of the model using the A1 benchmark of the *Computer Vision Problems in Plant Phenotyping* (CVPPP) dataset [33]. It comprises a collection of 128 images of plants taken from a top view with leaf annotation as ground-truth instance masks. We downsized the original 128 images in the training set by a factor of two and used a centre crop of size 224×224 for training. For the purpose of the ablation study, we randomly select 103 images from the CVPPP A1 benchmark for training and report the results on the remaining 25 images.

To compute the reward for our actor-critic model (Eq. 4), we use the Dice score computed as $\mathcal{F}(\mathcal{S}, \mathcal{T}) = \frac{2 \sum_i \mathcal{S}_i \mathcal{T}_i}{\sum_i \mathcal{S}_i + \sum_i \mathcal{T}_i}$. The dimensionality of the latent action space is fixed to 16.

In the first part of the experiment, we look into how different terms in the loss influence the segmentation quality, measured in Symmetric Best Dice (SBD), and the absolute Difference in Counting (|DiC|). Specifically, we train five models: *BL* is an actor-only recurrent model trained with BPTT through all states. We use the BCE loss and Dice-based max-matching as a heuristic for assigning the ground truth to predictions, similar to [31, 32]. *BL-Trunc* is similar to *BL*, but is trained with a truncated, one-step BPTT. We train our actor-critic model *AC-Dice* with the gradient from the critic approximating the Dice score. *AC-Dice-NoKL* is similar to the *AC-Dice* model, *i.e.* the actor is trained jointly with the critic, but we remove the KL-divergence term, which encourages exploration, from the loss of the actor. Lastly, we verify the benefit of the State Pyramid, the multi-res spatial information provided to the decoder, by comparing to a baseline without it (*AC-Dice-NoSP*).

The side-by-side comparison of these models summarised in Table 1 reveals that *AC-Dice* exhibits a superior accuracy compared to the baselines, both in terms of Dice and counting. Using the KL-divergence term in the loss improves the actor, which shows the value of action exploration in a consistent action space. We also observed that training *AC-Dice-NoKL* would sometimes diverge and require a restart with the critic warm-up. Furthermore, the State Pyramid aids the decoder, as removing it leads to a significant drop in mask quality. Surprisingly, *BL-Trunc* is only slightly worse than *BL*, which however has by far

Model	SBD \uparrow	DiC \downarrow
BL	80.0	1.08
BL-Trunc	79.4	1.32
AC-Dice	80.5	0.88
AC-Dice-NoKL	75.4	1.36
AC-Dice-NoSP	61.3	1.52

Table 1. *Evaluation on CVPPP val.* We compare our baseline with fully-unrolled (*BL*) and truncated BPTT (*BL-Trunc*) to the actor-critic with Dice-based reward, with (*AC-Dice*) and without (*AC-Dice-NoKL*) exploration, and without the State Pyramid (*AC-Dice-NoSP*).

Model	LSTM + Mask		Mask only		LSTM only	
	Dice* \uparrow	DiC \downarrow	Dice* \uparrow	DiC \downarrow	Dice* \uparrow	DiC \downarrow
BL	78.6	1.04	76.6	4.36	6.5	3.96
BL-Trunc	77.9	1.72	77.5	6.24	6.0	4.8
AC-Dice	78.4	0.88	78.5	1.92	5.8	4.36

* computed by max-matching and ground-truth stopping

Table 2. *Contribution of recurrent states to mask quality* measured by Dice and absolute Difference in Counting |DiC| on CVPPP val.

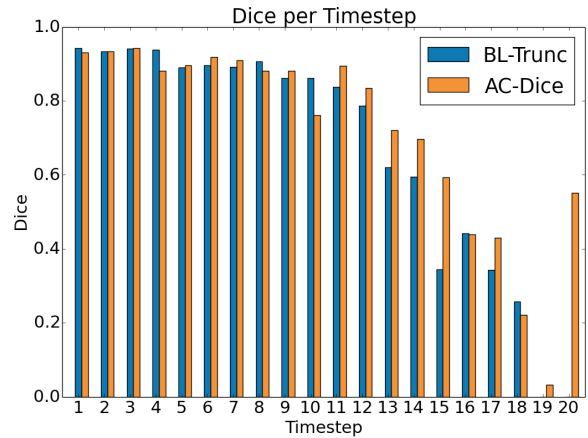


Figure 3. *Dice score of our actor-critic model (AC-Dice) vs. our baseline with truncated BPTT (BL-Trunc) on CVPPP val, averaged for each timestep.* We observe the advantage of our actor-critic model at later timesteps, which is an expected benefit of including the estimate of the expected reward in the loss at the earlier timesteps. Note that few images contain 20 instances, hence a large variance for this timestep.

higher memory demands than both *AC-Dice* and *BL-Trunc* in the setting of long sequences and high resolutions.

To further investigate the accuracy gains of the actor-critic model, we report the average Dice score w.r.t. the corresponding timestep of the prediction in Fig. 3. The histogram confirms our intuition that incorporating the future reward into the loss function for every timestep, as modelled by the critic, should improve the segmentation quality at later stages of prediction: the Dice score of the actor-

Model	SBD \uparrow	DiC \downarrow
RIS [32]	66.6	1.1
MSU [33]	66.7	2.3
Nottingham [33]	68.3	3.8
IPK [29]	74.4	2.6
DLoss [9]	84.2	1.0
E2E [31]	84.9	0.8
Ours (AC-Dice)	79.1	1.12

Table 3. *Segmentation quality of our actor-critic model on CVPPP test with Dice-based reward (AC-Dice) in terms of Symmetric Best Dice (SBD) and absolute Difference in Counting (|DiC|).*

critic model tends to be tangibly higher especially at the later timesteps. Note that the contribution of this benefit to the average score across the dataset is moderated by not all images in the dataset having many instances.

In the next part of the experiment, we are interested in the reliance of the model on the recurrent state. Recall that our model maintains the mask accumulating the previous predictions as well as the hidden LSTM state. We alternately “block” either of the states by providing a zero tensor at every timestep. We consider only the first n predictions to compute the Dice score, where n is the number of ground-truth masks. We stop the iterations if no termination was predicted after 21 timesteps, since the largest number of instances in our validation set is 20. The results in Table 2 show that the LSTM plays an important role for counting (or, termination prediction), while having almost no effect on the mask quality. The networks have learned a sequential prediction strategy given only the binary mask of previously predicted pixels. Note that in contrast to the baseline models, actor-critic training reduced the dependence on the LSTM state for counting (AC-Dice), which suggests that the actor makes a better use of the state mask to make the next prediction.

6.2. Instance segmentation

We compare our method to other approaches on two standard instance segmentation benchmarks, each containing a rich variety of small segments as well as occlusions.

CVPPP dataset. For the CVPPP dataset used in our ablation study, this time we evaluate on the official 33 test images and train only our actor-critic model (AC-Dice) on the total 128 images in the training set.

The results on the test set in Table 3 show that our method is on par with the state of the art in terms of counting while maintaining competitive segmentation accuracy. From a qualitative analysis, see examples in Fig. 4a, we observe that the order of prediction follows a consistent, interpretable pattern: large leaves are segmented first, whereas small and occluded leaves are segmented later. This follows our intuition for an optimal processing sequence: “easy”,

Model	MWCov \uparrow	MUCov \uparrow	AvgFP \downarrow	AvgFN \downarrow
DepthOrder [42]	70.9	52.2	0.597	0.736
DenseCRF [41]	74.1	55.2	0.417	0.833
AngleFCN+D [35]	79.7	75.8	0.201	0.159
E2E [31]	80.0	66.9	0.764	0.201
Ours (BL-Trunc)	72.2	50.7	0.393	0.432
Ours (AC-IoU)	75.6	57.3	0.338	0.309

(a) KITTI test set

Model	MWCov \uparrow	MUCov \uparrow	AvgFP \downarrow	AvgFN \downarrow
E2E (Iter-1)	64.1	54.8	0.200	0.375
E2E (Iter-3)	71.3	63.4	0.417	0.308
E2E (Iter-5)	75.1	64.6	0.375	0.283
Ours (BL-Trunc)	70.4	55.8	0.313	0.339
Ours (AC-IoU)	71.9	59.5	0.262	0.253

(b) KITTI validation set

Table 4. *Segmentation quality on KITTI.* We evaluate our baseline with truncated BPTT (BL-Trunc) and the actor-critic with IoU-based reward (AC-IoU) in terms of mean weighted (MWCov) and unweighted (MUCov) coverage, average false positive (AvgFP), and false negative (AvgFN) rates.

more salient instances should be predicted first to alleviate consecutive predictions of the “harder” ones. We also note, however, that the masks miss some fine details, such as the stalk of the leaves, which limits the benefits of the context for occluded instances. We believe this stems from the limited capacity of the critic network to approximate a rather complex reward function.

KITTI benchmark. We use the instance-level annotation of cars in the KITTI dataset [10] to test the scalability of our method to traffic scenes. We used the same data split as in previous work [31, 35], which provides 3712 images for training, 144 images for validation, and 120 images for testing. While the validation and test sets have high-quality annotations [6, 41], the ground-truth masks in the training set are largely ($> 95\%$) coarse or incomplete [6]. Hence, good generalisation from the training data would indicate that the algorithm can cope well with inaccurate ground-truth annotation.

The evaluation criteria for this dataset are: the mean weighted coverage loss (MWCov), the mean unweighted coverage loss (MUCov), the average false positive rate (AvgFP), and the average false negative rate (AvgFN). MUCov is the maximum IoU of the ground truth with a predicted mask, averaged over all ground-truth segments in the image. MWCov additionally weighs the IoUs by the area of the ground-truth mask. AvgFP is the fraction of mask predictions that do not overlap with the ground-truth segments. Conversely, AvgFN measures the fraction of the ground-truth segments that do not overlap with the predictions.

We use an IoU-based score function to compute the rewards, i.e. $\mathcal{F}(\mathcal{S}, \mathcal{T}) = \frac{\sum_i \mathcal{S}_i \mathcal{T}_i}{\sum_i \mathcal{S}_i + \sum_i \mathcal{T}_i - \sum_i \mathcal{S}_i \mathcal{T}_i}$. To show the

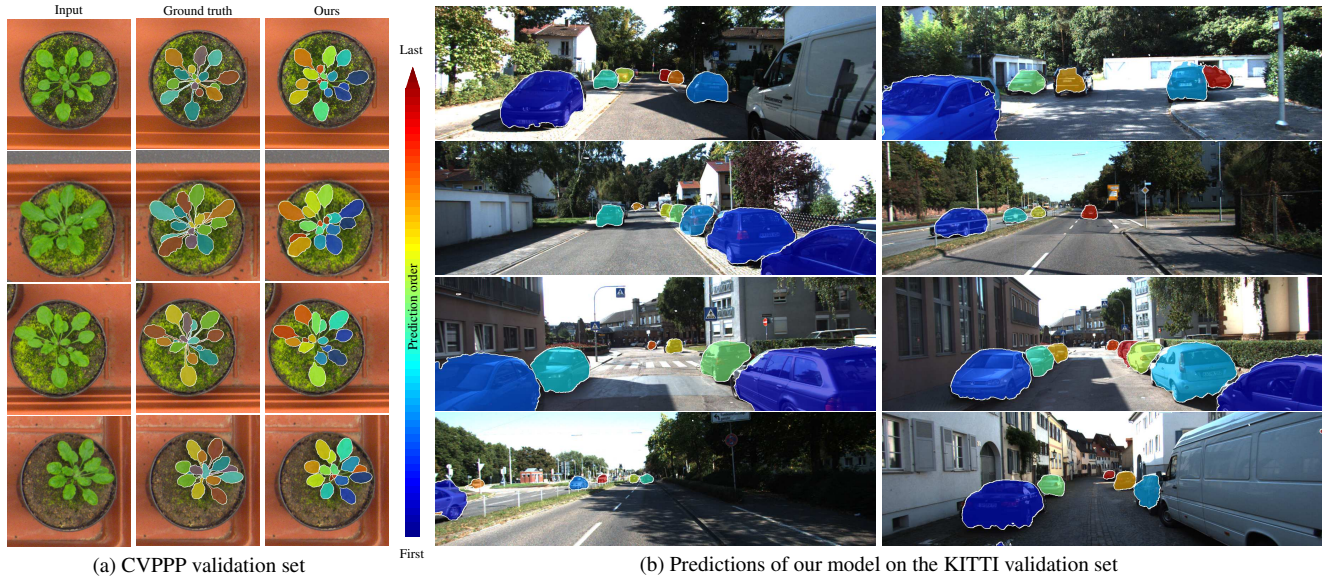


Figure 4. Predictions from our AC model on the CVPPP (a) and KITTI datasets (b). The colourmap (middle) encodes the prediction order and ranges from blue (first prediction) to red (last prediction). Note how the prediction order follows a consistent pattern: large unoccluded segments tend to be segmented first, whereas small and occluded segments are usually predicted last.

benefits of our Actor-Critic model (*AC-IoU*) for structured prediction at higher resolutions, we also train and report results for a baseline, the actor-only model trained with one step BPTT (*BL-Trunc*). Considering the increased variability of the dataset compared to CVPPP, we used 64 latent dimensions for the action space.

The results on the test are shown in Table 4a. Given the relatively small size of the test set, we also report the results on the validation set in Table 4b, and use the available results from the equivalent evaluation of a state-of-the-art method [31] for reference.

The results indicate that our method scales well to larger resolutions and action spaces and shows competitive accuracy despite not using bounding box representations. Similar to our results on CVPPP, our model does not quite reach the accuracy of a recurrent model using bounding boxes [31] and a non-recurrent pipeline. We believe the segmentation accuracy is currently limited by the degree of the reward approximation by the critic and the representational power of the network architecture used by the actor model. As can be seen in some examples in Fig. 4b, without post-processing the masks are not always well aligned with the object and occlusion boundaries. However, we note that the prediction order also follows a consistent, interpretable pattern: nearby instances are segmented first, while far-away instances are segmented last. Without hard-coding such constraints, the network appears to have learned a strategy that agrees with human intuition to segment larger, close-by objects first and exploits the resulting context to make predictions in the order of increasing difficulty.

7. Conclusions

In the current study, we formalised the task of instance segmentation in the framework of reinforcement learning. Our proposed actor-critic model utilises exploration noise to alleviate the initialisation bias on the prediction ordering. Considering the high dimensionality of pixel-level actions, we enabled exploration in the action space by learning a low-dimensional representation through a conditional variational auto-encoder. Furthermore, the critic approximates a reward signal that also accounts for future predictions at any given timestep. In our experiments, it attained competitive results on established instance segmentation benchmarks and showed improved segmentation performance at the later timesteps. Our model predicts instance masks directly at the full resolution of the input image and does not require intermediate bounding box predictions, which stands in contrast to proposal-based architectures [13] or models delivering only a preliminary representation for further post-processing, *e.g.* [9, 35].

These encouraging results suggest that actor-critic models have potentially a wider application spectrum, since the critic network was able to learn a rather complex loss function to a fair degree of approximation. In future work, we aim to improve our baseline model of the actor network, which currently limits the attainable accuracy.

Acknowledgements. The authors would like to thank Stephan R. Richter for helpful discussions.

References

- [1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017. 2
- [2] Anurag Arnab and Philip H. S. Torr. Bottom-up instance segmentation using deep higher-order CRFs. In *BMVC*, 2016. 2
- [3] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. In *ICLR*, 2017. 2
- [4] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, pages 2858–2866, 2017. 2
- [5] Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983. 1, 2
- [6] Liang-Chieh Chen, Sanja Fidler, Alan L. Yuille, and Raquel Urtasun. Beat the MTurkers: Automatic image labeling from weak 3D supervision. In *CVPR*, pages 3198–3205, 2014. 7
- [7] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In *ICCV*, pages 4106–4116, 2017. 2
- [8] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, pages 3150–3158, 2016. 2
- [9] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. In *CVPR Workshop on Deep Learning for Robotic Vision*, 2017. 2, 7, 8
- [10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI Vision Benchmark Suite. In *CVPR*, pages 3354–3361, 2012. 7
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*2014*, pages 2672–2680. 2
- [12] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. DRAW: A recurrent neural network for image generation. In *ICML*, pages 1462–1471, 2015. 2
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017. 1, 2, 8
- [14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4
- [15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5
- [16] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1, 2, 3, 4
- [17] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. InstanceCut: From edges to instances with MultiCut. In *CVPR*, pages 7322–7331, 2017. 2
- [18] Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics*, 2(1):83–98, 1955. 1, 2
- [19] Hugo Larochelle and Geoffrey E. Hinton. Learning to combine foveal glimpses with a third-order Boltzmann machine. In *NIPS*2010*, pages 1243–1251. 2
- [20] Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, and Xiaoou Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *CVPR*, pages 6459–6468, 2017. 1, 3
- [21] Yao Li, Guosheng Lin, Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Sequential person recognition in photo albums with a recurrent network. In *CVPR*, pages 5660–5668, 2017. 2
- [22] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, pages 4438–4446, 2017. 2
- [23] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *ICLR*, 2016. 1
- [24] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. SGN: Sequential grouping networks for instance segmentation. In *ICCV*, pages 3516–3524, 2017. 2
- [25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 4
- [26] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*, 2013. 1
- [27] Volodymyr Mnih, Adrià Puigdomènèch Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, pages 1928–1937, 2016. 1
- [28] Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, pages 278–287, 1999. 3
- [29] Jean-Michel Pape and Christian Klukas. 3-D histogram-based segmentation and leaf detection for rosette plants. In *ECCV Workshops*, pages 61–74, 2014. 7
- [30] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016. 2
- [31] Mengye Ren and Richard S. Zemel. End-to-end instance segmentation and counting with recurrent attention. In *CVPR*, pages 293–301, 2017. 1, 2, 3, 4, 6, 7, 8
- [32] Bernardino Romera-Paredes and Philip H. S. Torr. Recurrent instance segmentation. In *ECCV*, volume 6, pages 312–329, 2016. 1, 2, 3, 5, 6, 7
- [33] Hanno Schar, Massimo Minervini, Andrew P. French, Christian Klukas, David M. Kramer, Xiaoming Liu, Imanol Luengo, Jean-Michel Pape, Gerrit Polder, Danijela Vukadinovic, Xi Yin, and Sotirios A. Tsafaris. Leaf segmentation in plant phenotyping: A collation study. *Mach. Vis. Appl.*, 27(4):585–606, 2016. 6, 7
- [34] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An Introduction*. Adaptive computation and machine learning. MIT Press, 1998. 2, 3

- [35] Jonas Uhrig, Marius Cordts, Uwe Franke, and Thomas Brox. Pixel-level encoding and depth layering for instance-level semantic labeling. In *GCPR*, pages 14–25, 2016. [2](#), [4](#), [7](#), [8](#)
- [36] Shimon Ullman. Visual routines. In *Readings in Computer Vision*, pages 298–328. Elsevier, 1987. [1](#)
- [37] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. In *ICLR*, 2016. [3](#)
- [38] Paul J. Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1:339–356, 1988. [5](#)
- [39] Shi Xingjian, Zhouong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NIPS*2015*, pages 802–810. [2](#)
- [40] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seq-GAN: Sequence generative adversarial nets with policy gradient. In *AAAI*, pages 2852–2858, 2017. [2](#)
- [41] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected MRFs. In *CVPR*, pages 669–677, 2016. [7](#)
- [42] Ziyu Zhang, Alexander G. Schwing, Sanja Fidler, and Raquel Urtasun. Monocular object instance segmentation and depth ordering with CNNs. In *ICCV*, pages 2614–2622, 2015. [7](#)