OSVOS
First-frame fine-tuning

• Goal: Learn the appearance of the object to track

• Main contribution: separate training steps
  – Pre-training for ‘objectness’.
  – First-frame adaptation to specific object-of-interest using fine-tuning.
One-shot VOS

1. **Pre-trained**
   - Base Network: Pre-trained on ImageNet
   - Results on frame N of test sequence

2. **Training**
   - Parent Network: Trained on DAVIS training set
   - Learns how to do video segmentation

3. **Finetuning**
   - Test Network: Fine-tuned on frame 1 of test sequence
   - Learns which object to segment

Edges and basic image features
One-shot VOS

• One-shot: we see the first frame ground truth

• Finetuning step: this is used to technically *overfit* to the test sequence first frame. Overfitting is therefor used to learn the appearance of the foreground object (and the background!)

• Test time: each frame is processed independently ➔ no temporal information
Frame-based segmentation

• **PRO**: it recovers well from occlusions (unlike mask propagation or optical flow-based methods)

• **CON**: it is temporally inconsistent
Experiments: highly dynamic scenes
Experiments: accuracy vs annotations

Another annotation where the 2nd camel is background

Another annotation

Two camels!

Mask is refined
Finetuning time

Object flow

102ms – One forward pass (parent network)

11.8 pp.
Observations

• OSVOS does not have an object of object shape.

• It is a pure appearance-based method, if the foreground (or the background) appearance changes too much, the method fails.
Introducing Semantics

First frame
He was occluded in the first frame, therefore the network never learned he was background.
But wait....

- We have already seen models that have an idea of object shape.

- Instance segmentation methods!
OSVOS-S: Semantic propagation

Semantic prior branch that gives us proposals to select from

Semantic Prior

Semantic Instance Segmentation → Semantic Selection & Propagation

Instance Proposals → Top Matching Instances

Semantic Prior branch that gives us proposals to select from

Prior: semantics stay coherent throughout the sequence

Input Image → CNN → Appearance Model

First-Round Foreground Estimation → Foreground Estimation

Conditional Classifier → Result

K.-K. Maninis et al. “Video object segmentation without temporal information”. TPAMI 2018
OSVOS-S: Semantic propagation

Instance Segmentation Proposals

Ground Truth

Semantic Selection

Selected Instances: Person and Motorbike

First-Round Foreground Estimation

Semantic Propagation

Frame 0

Frame 18  Frame 24  Frame 30  Frame 36

K.-K. Maninis et al. “Video object segmentation without temporal information”. TPAMI 2018
Drifting problem

• If the object greatly changes its appearance (e.g., though pose or camera changes), then the model is not powerful anymore

• But this change was gradual....
Drifting problem

- If the object greatly changes its appearance (e.g., though pose or camera changes), then the model is not powerful anymore.

Why not gradually update the model?
OnAVOS: Online Adaptation

- Online adaptation: adapt model to appearance changes every frame – not just the first frame.
- Iteratively fine-tune the model on previous prediction every frame.

- **CON**: Extremely slow.

OnAVOS: Online Adaptation

Mask Refinement

• Assumption: an object, i.e., a mask, does not move a lot from frame to frame.
• We can often start with an approximate mask (either from previous frame or from coarse estimate).
• We can then use a refinement network to accurately refine the mask estimate.
• This can take advantage of crop-and-zoom to do segmentation at a higher resolution.
A. Khoreva et al. "Learning Video Object Segmentation from Static Images" CVPR 2017

Why the name?
• Training inputs can be simulated!
  – Like displacements to train the regressor of Faster-RCNN
  – Very similar in spirit to Tracktor

(a) Annotated image  (b) Example training masks
Worth reading


• X. Li et al. „Video object segmentation with re-identification“ CVPRW 2017. → use reidentification techniques to recover from occlusions
Proposal-based approaches
Proposal Generation

Until now:
- Input is the whole image
- Proposals are put on top just to refine

Now:
Input are proposals
Goal is to “link” them (much like we did in tracking-by-detection)

- Instance Segmentation Networks (E.g. Mask-RCNN) give object instance segmentation proposals.
- One can approach video object segmentation as taking these proposals in each frame and then linking them over time using a merging algorithm.
PReMVOS

• An approach that combines all of the previous VOS principles and gives state-of-the-art results.

• Combines the following principles:
  – First-frame fine-tuning
  – Mask Refinement
  – Optical Flow Mask Propagation
  – Data Augmentation
  – Object Appearance Re-Identification
  – Proposal Generation

PReMVOS: Overview

- **Proposal generation**
  - Category-agnostic Mask R-CNN proposals

- **Refinement**
  - Fully-convolutional segmentation network trained to refine the segmentation given a proposal bounding box
PReMVOS: Overview

- **Merging**
  - Greedy decision process, chooses proposal(s) with best score
  - Optional proposal expansion through Optical Flow propagation
  - Proposal score as combination of
    - **Objectness** score
    - **Mask propagation** IoU score (Optical Flow warping)
    - **ReID** score
    - **Object-Object interaction** scores
PReMVOS: results

• Very complex but a winner

• DAVIS Challenge 2018 Winner

• Youtube-VOS Challenge 2018 Winner
Lessons Learned

• Challenge 1: How to generate proposals?
  – Deep-learning based region proposal generators are fit for the task
  – Experimented with SharpMask and Mask R-CNN

• Challenge 2: How to track region proposals?
  – Region overlap works as a consistency measure
  – Optical flow based propagation really helps
  – ReID score also helpful

• Open issues
  – PReMVOS has no notion of 3D objects moving through 3D space.
  – Track initialization / termination logic needed for real tracking.
  – How to obtain the initial segmentation?
Pixel-wise retrieval

- Re-Identification networks based on bounding-box region proposals work really well.
- This idea can be extended to a Re-Identification embedding for every pixel.
- This pixel-wise Re-ID embedding vectors can then be used to directly extract a mask by taking the pixel with an embedding similar to the first frame embeddings.
Pixel-wise retrieval

- The user input can be in any form and we do not need to retrain the model for each sequence, nor finetune.

We are dealing with video

• Which is a sequence of images....

• And we have not talked about....

• Recurrent Neural Networks!
Temporal LSTM

- One-shot video object segmentation
- If we have multiple objects, each of them is predicted independently
R-VOS: temporal and spatial LSTM


C. Ventura et al. „RVOS: end-to-end recurrent network for video object segmentation“. CVPR 2019
R-VOS: temporal and spatial LSTM

- Instance generation and temporal coherence are both trained end-to-end

- Image just needs to be processed once (unlike ConvLSTM example before)
Overview of the methods

- Video Object Segmentation (VOS)
  - OSVOS: First-frame fine-tuning (appearance model)
  - OSVOS-S: + semantic guidance through proposals (shape)
  - OnAVOS: Online Adaptation (stronger appearance model)
  - MaskTrack: Mask Refinement
  - Lucid: clever data augmentation
  - ReID-VOS: Object Appearance Re-Identification
  - PReMVOS: putting it all together
  - Seq2seq and RVOS: recurrent architectures
Evaluation and metrics
Metrics for VOS

- **Region similarity**: Jaccard index (IoU) of ground truth mask and predicted mask.
- **Contour Accuracy**: measures the precision and recall of the boundary pixels. This is put together in the F-measure.

\[
\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}
\]

\[
F = \frac{2 \times \text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}
\]
Metrics for VOS

- **Temporal stability**: measures the evolution of object shapes, i.e., how stable the boundaries are in time.
  - Estimate the deformation of the mask from $t$ to $t+1$
  - If the transformation is smooth and precise, the result is considered stable.
  - A bad results is a jittery mask evolution
  - Note: this measure has been dropped due to its instability during occlusions.
Metrics for VOS

• You can use error measure statistics

• Region similarity: Jaccard index (IoU) of ground truth mask and predicted mask.
  – Mean: average for the dataset
  – Decay: quantifies the performance loss (or gain) over time. ➔ This is currently used to judge temporal stability
  – Recall: fraction of sequences scoring higher than a threshold
Tracking and Segmentation
VOS -> MOTS

• Video Object Segmentation (VOS) is limited by:
  – First frame mask given (in the supervised case)
  – Short video clips with objects present in almost all frames
  – Objects in a video are (mostly) of different categories
  – Few objects to track (max around 7 per video)

• Multi-Object Tracking and Segmentation (MOTS)
  – Scenarios with a large number of objects (20-40), mostly of the same category (e.g., pedestrians)
  – Long sequences
  – No first frame annotation provided, one has to deal with appearing and disappearing objects.
MOTS dataset

- Segmentations coming to MOTChallenge pedestrian tracking dataset

P. Voigtlaender et al. „MOTS: Multi-Object Tracking and Segmentation“. CVPR 2019
Next lectures

• Last lecture on January 31\textsuperscript{st}

• Taking the leap towards 3D detection, tracking and segmentation

• February 7\textsuperscript{th}: your presentations – best methods of the challenge will be selected to present!
Disclaimer

- This lecture was done borrowing material from:
  - Prof. Xavier Giró, Technical University of Catalonia (UPC)
  - Jonathon Luiten, RWTH Aachen