# Objects do not disappear: Video object detection by single-frame object location anticipation

Thanks to all reviewers. We are excited they find our work:

Novel **R1***: interesting and could be inspiring*, **R2***: presents a novel video object detection approach*, **R3***: a novel method to improve the computation efficiency, trajectory loss is new and effective.*

Powerful **R1***: improvements on ImageNet VID is significant*, **R2***: achieves great computational efficiency, competitive results on widely used benchmarks, has great practical potential*, **R3***: effective, computationally efficient.*

Elegant **R1***: a clear and transparent acknowledgement of the limitations*, **R2***: straightforward to use and free of unnecessary complexity*, **R3***: good and easy to follow.*

**About the mentioned detector/backbone experiments:**

**R1***: The proposed method seems to be detector agnostic, ..., consider doing it with more advanced detectors*;

**R2***: Table 4 is missing some recent VOD methods ... using a stronger backbone than ResNet-101*:

We added ImageNet VID results below, showing on-par results with PTSEFormer with the Deformable DETR detector and R101 backbone, and our method improves over all state-of-the-art with a SwinB backbone.

| Methods | Base Detector | Backbone | mAP (%) | Runtime (FPS) |
|---|---|---|---|---|
| TransVOD | Deformable DETR | R101 | 81.9 | 32.3 |
| PTSEFormer | Deformable DETR | R101 | 88.1 | - |
| TransVOD | Deformable DETR | SwinB | 90.1 | 14.9 |
| Ours | Faster-RCNN | R101 | 87.2 | 39.6 |
| Ours | Deformable DETR | R101 | 87.9 | 36.4 |
| Ours | Deformable DETR | SwinB | 91.3 | 18.1 |

**About the mentioned additional datasets:**

**R1***: method only applicable to very simple datasets*:

Yes, ImageNet VID (≈90% mAP) can be considered simple. Yet, EPIC KITCHENS-55 (≈45% mAP), and YouTube-BoundingBoxes (≈60% mAP) are definitely not simple, and our method is still applicable.

**R1***: running on MOT datasets and Waymo Open Dataset*:

Given the limited rebuttal time, we could only run a small subset of Waymo. We are the first to do VOD on this dataset, and thus we can only compare to a static detector. We will add the full dataset results to the paper.

| Methods | AP/L1 (%) | AP/L2 (%) |
|---|---|---|
| Faster-RCNN | 55.66 | 49.63 |
| Ours $L_{\text{traj}}^{(\text{sa})}$ | 58.72 | 51.56 |

Unfortunately, we cannot do MOT datasets as they have only a few (eg 4) training videos, which is ok for tracking, but not sufficient for VOD training, which might explain why there are also no VOD baselines for MOT.

**About predicting the future from a static image:**

**R1***: predicting future locations from a single input image is ill-posed.*;

**R2***: Why using a single static frame is sufficient for predicting accurate object trajectories*:

**R3***: [17] has already validated that a model can predict motion from objects' appearance in a static frame*:

Yes, it is ill-posed. However, conditioning on a data-distribution still allows predicting a likely future. Others also use a single frame to predict future appearance [30, 43, 49, 61], actions [1, 18, 46, 55], and motion [17, 49, 62]. Our method differs, by predicting box trajectories in video.

**R1***: Ideally we should have a way to model the uncertainty. it'll make more sense to predict a list of future trajectories*

Definitely. Our method does this now implicitly, by predicting multiple future object trajectories – one for each proposal box, where multiple boxes correspond to the same object. We clarified the text. Future work by explicitly encoding uncertainty sounds exciting!

**About MovingDigits:**

**R1***: what motions are applied to mnist*:

Each class has its own, unique, linear motion. We will release the dataset and clarify the paper.

**R3***: MovingDigits ... once the model learns to recognize the digit, it can predict the trajectory without necessarily relying on cues from the object's appearance*:

No, to recognize the digit, it first needs to use appearance cues; then it can predict the corresponding trajectory.

**R3***: could have tried using the MovingMNIST dataset*:

Unfortunately, it has no detection bounding boxes nor the desired motion-appearance correlations. Thus, we had to create our own. We added this motivation to the paper.

**Other remarks:**

**R2***: how does Eq. (2) handle occlusion cases...Are the coordinates cropped if they go out of images?*:

In Eq. (2), we ignore the coordinates calculation if the ground truth coordinates are not valid. So, we may over-predict when the objects disappear or are occluded along the predicted trajectory. We clarified this.

**R2***: Section 3.1, association of trajectories to ground truth*:

Each trajectory inherits the class of its keyframe detection guaranteeing that the highest IoU box always belongs to the same predicted object class. We clarified the text.

**R3***: Section 4.1 is not clear to me. ... motion prediction is not used in the pipeline for other experiments*:

We apologise for the confusion, motion prediction is used in all experiments. We clarified the text and figures.