Tracking and Segmentation of Transparent Objects

Alan Lukežič[1], **Matej Kristan**[1]

[1] *Faculty of computer and information science, University of Ljubljana, Večna pot 113, 1000 Ljubljana*
E-mail: {alan.lukezic, matej.kristan}@fri.uni-lj.si

Abstract

Transparent object tracking is a challenging, recently introduced, problem. Existing methods predict target location as a bounding box, which is often only a poor approximation of actual location. Segmentation mask is a more accurate prediction, but benchmarks for evaluating tracking and segmentation performance of transparent objects does not exist. In this paper we address this drawback by introducing a new dataset for tracking and segmentation of transparent objects. In particular we sparsely re-annotate the existing bounding box TOTB dataset with ground-truth segmentation masks. A comprehensive analysis demonstrates that existing segmentation methods perform surprisingly well on this task indicating good design generalization and potential for transparent object tracking tasks. In addition, we show that existing bounding box trackers can be easily transformed into segmentation trackers using modern mask refinement methods.

1 Introduction

Visual object tracking is a task of continuous target localization in a video given a single supervised example of target appearance at the beginning. It is a fundamental computer vision problem with a large potential of downstream applications including autonomous vehicles, robotics, surveillance and video editing systems. In the past, most of the research had been focused on tracking opaque objects [15, 8, 9], but recently, the TOTB benchmark [6] introduced a new sub-field of transparent object tracking.

Transparent object tracking evolved further with a new training dataset [11], developed for fine-tuning trackers on transparent objects. This allowed an elegant adaptations of existing deep trackers developed for tracking opaque objects without algorithmic modifications. Fine-tuning consistently boosted performance of existing opaque object trackers on the TOTB dataset, which was mainly a result of the improved prediction accuracy, while tracking robustness did not improve significantly. This was addressed with the distractor-aware transparent object tracker (DiTra) [12]. This tracker improved also tracking robustness by treating localization accuracy and localization robustness as two separated tasks.

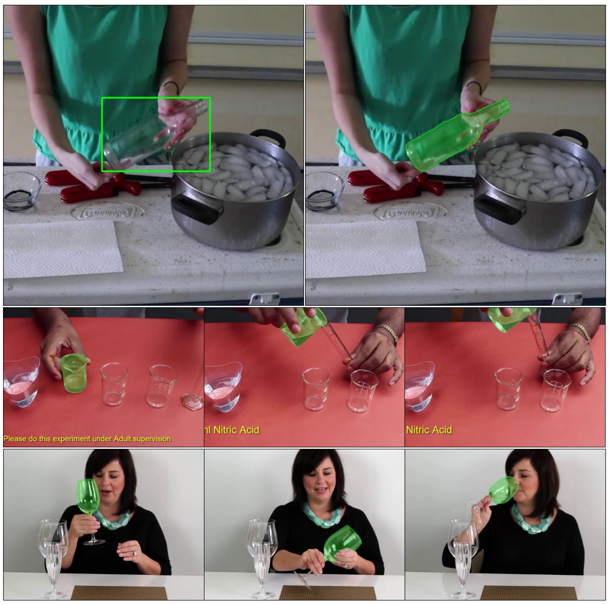

Figure 1: A bounding box is often a poor approximation of object location, while segmentation mask is much more accurate representation (first row). The second and third row show two video sequences from TOTB dataset with new ground-truth segmentation masks.

All existing transparent object trackers predict target position as an axis aligned bounding box. In some cases bounding box represents only a poor approximation of target location, for example: a rotated glass bottle in Figure 1. In such cases a large amount of pixels within a bounding box are actually background pixels and it would be an extremely challenging task for a robotic hand to grasp this bottle. A much more accurate representation would be a segmentation mask, but the task of tracking and segmentation of transparent objects does not exist yet. One reason for this is lack of evaluation datasets, which we address in this work.

In this paper we present a semi-automatic dataset annotation process and demonstrate it on annotating the TOTB dataset with segmentation masks. Since the TOTB dataset is already annotated with bounding boxes, they are used as approximate annotations and refined until high quality segmentation masks are obtained. The new dataset allows evaluation of tracking and segmentation performance on transparent objects. Such an analysis of existing tracking methods has been conducted on the new dataset and is presented in this paper along with several insights.

2 Dataset

A crucial requirement to evaluate tracking and segmentation performance of transparent objects is a dataset with annotated objects of interest with ground-truth segmentation masks in every sequence. An existing dataset for tracking transparent objects TOTB [6] consists of bounding box annotations on every video frame i.e., dense annotations. In this paper we extend the TOTB dataset with sparsely annotated segmentation masks. A detailed description of the dataset annotation is presented in the following.

TOTB dataset is composed of 225 video sequences, which is 85,631 frames in total. Manual annotation of such a large number of frames with high-quality segmentation masks is an expensive and a time-consuming task. To address this issue we developed a three-stage semi-automatic annotation process. In the **first stage** we selected every Δ frame and used a segmentation tool based on Segment Anything Model [7] (SAM), which automatically segments an object selected by an existing TOTB ground-truth bounding box. We empirically determined that $\Delta = 30$ is a good trade-off for reliable evaluation. Result of the first annotation stage was 2,944 segmentation masks, which were manually checked and refined in the **second stage**. In this stage 2,206 masks (75%) were corrected using positive or negative user clicks used as SAM prompts, or accepted without a correction. 674 masks (23%) were marked for a manual segmentation, while the remaining 64 masks (2%) were removed from annotation process due to the poor quality. The main reasons for poor quality were significant motion blur, or extreme transparency of an object, which prevented unambiguous manual annotation. In the **third stage** the 674 frames determined in the second stage were manually segmented using a simple annotation tool [14]. The main reasons which prevented automatic segmentation on these frames were poor visibility of object boundary or visual ambiguity due to the transparency. The annotation process ended with 2,880 high-quality segmentation masks.

3 Methods

We benchmark several state-of-the-art trackers, that predict target position as a bounding box, as well as a recent segmentation-based tracker Cutie [3]. A set of bounding box trackers include representatives of siamese deep trackers SiamBAN [2] and SiamRPN [10], deep correlation filters ATOM [4] and DiMP [5] and transformer-based trackers STARK [17] and TOMP [13]. Since the focus of this paper is on transparent object tracking, we include TrATOM [6], which is a variation of ATOM, developed specifically for transparent objects. Finally, DiTra [12] is included as a recent state-of-the-art transparent object tracker.

In addition to pure visual object trackers, we include two methods for computing segmentation masks based on approximate object locations. AlphaRefine [18] was developed for refining bounding box tracking outputs, while Segment Anything Model [7] (SAM) is a general object segmentator.

4 Experiments

In the experiments, all trackers are run using the one-pass evaluation protocol [16] under which the target is initialized in the first frame and then tracked until the end of the video without interventions. The new dataset with segmentation ground-truth (Section 2) is denoted by TOTB^S to make a distinction from the original bounding box TOTB dataset [6]. The primary performance measure is the area under the success rate curve (AUC), which approximates average overlap over the entire dataset [1]. To provide additional insights into tracking performance, we present two auxiliary measures: tracking accuracy and robustness [8]. Tracking robustness is defined as an average ratio of successfully tracked frames, while tracking accuracy is an average overlap on these frames. Following [8], a frame is considered as successfully tracked frame if the overlap of the prediction with the ground-truth is larger than $\theta = 0$.

Since the new ground-truth segmentation masks in TOTB^S are annotated sparsely (Section 2), i.e., every Δ frames, the sparse evaluation protocol is evaluated in Section 4.1. Then, the state-of-the-art trackers are compared on the TOTB^S dataset in Section 4.2. Two segmentation refinement methods are evaluated in Section 4.3, while the importance of the initialization quality is analyzed in Section 4.4.

4.1 Sparse evaluation

This experiment evaluates our hypothesis, which assumes that results obtained with sparse evaluation (using only every $\Delta = 30$ frame) does not differ significantly from results obtained with dense evaluation (using all video frames). Tracking performance evaluated on all frames is presented in Table 1 under the *Dense eval* column. In contrast, tracking performance computed on every Δ frames is shown in Table 1 under the *Sparse eval* column. Note that the original bounding box annotations provided in TOTB dataset [6] are used in both experiments, to rule out the impact of different annotation types.

These results clearly show that difference in tracking performance between the two evaluation setups is negligible. The largest relative change in tracking performance is observed at STARK tracker – approximately 1% AUC, while performance of all other methods is significantly below one percent. Changes of both secondary measures, tracking accuracy and robustness, are comparable to changes of the AUC. These results also confirm that annotating every $\Delta = 30$ frame of the TOTB dataset with a segmentation mask is a good approximation of tracking performance evaluated on every frame.

Table 1: Relative change (RC) of the AUC performance measure between dense and sparse evaluation. Tracking accuracy (Acc) and robustness (Rob) are secondary measures shown to provide additional insights in tracking performance. Note that original TOTB dataset with bounding box ground-truth is used in both evaluation protocols.

	Dense eval.			Sparse eval.			RC
Method	AUC	Acc.	Rob.	AUC	Acc.	Rob.	
Cutie	0.803	83.3	98.0	0.806	83.0	98.2	0.4%
DiTra	0.773	81.3	96.5	0.777	81.5	96.7	0.5%
TOMP	0.738	78.4	95.2	0.742	78.7	95.3	0.5%
STARK	0.738	81.0	91.8	0.746	81.4	92.4	1.1%
DiMP	0.699	74.4	94.2	0.700	74.5	94.2	0.1%
SiamBAN	0.680	73.2	90.2	0.679	72.5	92.4	0.1%
TrATOM	0.664	72.0	92.4	0.662	71.6	91.9	0.3%
SiamRPN	0.655	72.2	90.2	0.653	71.6	90.1	0.3%
ATOM	0.642	71.0	90.4	0.637	70.1	90.2	0.8%

Table 2: Segmentation performance on the TOTB^S dataset. Gray-coloured line represent tracker with segmentation outputs, while the remaining are bounding box trackers.

Method	AUC	Acc.	Rob.
Cutie	0.808	83.6	97.9
DiTra	0.583	60.7	96.0
TOMP	0.555	58.7	94.5
STARK	0.575	62.5	91.8
DiMP	0.552	58.4	93.4
SiamBAN	0.543	58.2	91.2
TrATOM	0.527	57.0	91.0
SiamRPN	0.537	58.7	89.2
ATOM	0.507	55.9	89.3

4.2 State-of-the-art comparison

Table 2 evaluates tracking and segmentation performance on the newly annotated segmentation-based TOTB^S dataset using sparse evaluation protocol. Eight trackers predicting bounding boxes and a single segmentation tracker are included in the evaluation. The top-performer Cutie outperforms the second-best DiTra by a large margin of 39% in AUC. A closer inspection reveals, that both trackers achieve a similar robustness (97.9% vs 96%), i.e., are able to successfully track the target for similar duration. A large difference is observable in accuracy, where Cutie outperforms DiTra by 38%. This is an expected outcome since Cutie predicts much more accurate segmentation masks, compared to bounding boxes predicted by DiTra.

Results in Table 2 clearly show that existing state-of-the-art trackers achieve outstanding robustness when tracking transparent objects. On the other hand, accuracy of bounding box trackers is limited, compared to segmentation-based trackers. Qualitative examples of tracking and segmentation with Cutie [3] are shown in Figure 4.2.

4.3 Segmentation refinement

Experiment in Section 4.2 demonstrates that bounding box trackers achieve excellent robustness with a significantly lower accuracy compared to segmentation trackers. We thus compare two approaches for refining tracking output: AlphaRefine (AR) [18] and the Segment Anything Model (SAM) [7] in Table 3. This experiment demonstrates how well the bounding box predictions can be transformed into segmentation masks.

Table 3: Comparison of two mask refinement methods AR and SAM on top of Cutie and DiTra on TOTB^S dataset. Gray-coloured lines represent segmentation outputs, while the remaining line is a bounding box tracker.

Method	AUC	Acc.	Rob.
Cutie+SAM	0.839	88.9	97.0
DiTra+SAM	0.829	89.1	95.6
Cutie	0.808	83.6	97.9
DiTra+AR	0.752	79.1	96.2
DiTra	0.583	60.7	96.0

Extending a state-of-the-art bounding box tracker DiTra with AlphaRefine (DiTra+AR) improves overall tracking performance by 29% AUC. Performance boost comes due to the more accurate predictions in form of segmentation masks, compared to the bounding boxes in original DiTra. Tracking accuracy increases by 30%, while robustness remains almost unchanged. Using SAM instead of AlphaRefine on top of DiTra output (DiTra+SAM) improves the results even further, i.e., overall tracking performance increases by 42%. Performance boost is again a consequence of improved accuracy by 47% compared to the original DiTra.

To demonstrate the strong capability of improving the tracking accuracy using SAM, we extend the segmentation based tracker Cutie with SAM refined segmentation masks (Cutie+SAM). In particular, we applied a min-max operation on Cutie segmentation masks to obtain bounding boxes, which were used as prompts in SAM to get the refined segmentation masks. Results in Table 3 demonstrate that overall tracking performance of Cutie is improved by approximately 4% AUC, mostly due to the 6% improvement in accuracy, while robustness drops minimally, by only 1%.

These results demonstrate that existing refinement methods (AlphaRefine and SAM) are capable of providing accurate segmentation masks on transparent objects and effectively extend bounding box trackers with segmentation outputs. Furthermore, SAM achieves excellent tracking accuracy – almost 90% and demonstrates capability of improving both types of tracking outputs: bounding boxes as well as segmentation masks.

4.4 Evaluation of initialization quality

Segmentation-base trackers e.g., Cutie [3] require a segmentation mask on the first video frame for initialization. In Table 4 we compare different initialization variations for Cutie. The upper performance limit is set by using a ground-truth, i.e., manually annotated mask.

The second-best performing approach estimates the mask by refining a ground-truth bounding box using SAM model [7], which reduces an overall tracking performance by 1.4% AUC. An alternative to SAM is AlphaRefine [18] (denoted by *AR* in Table 4). Compared to SAM, Al-

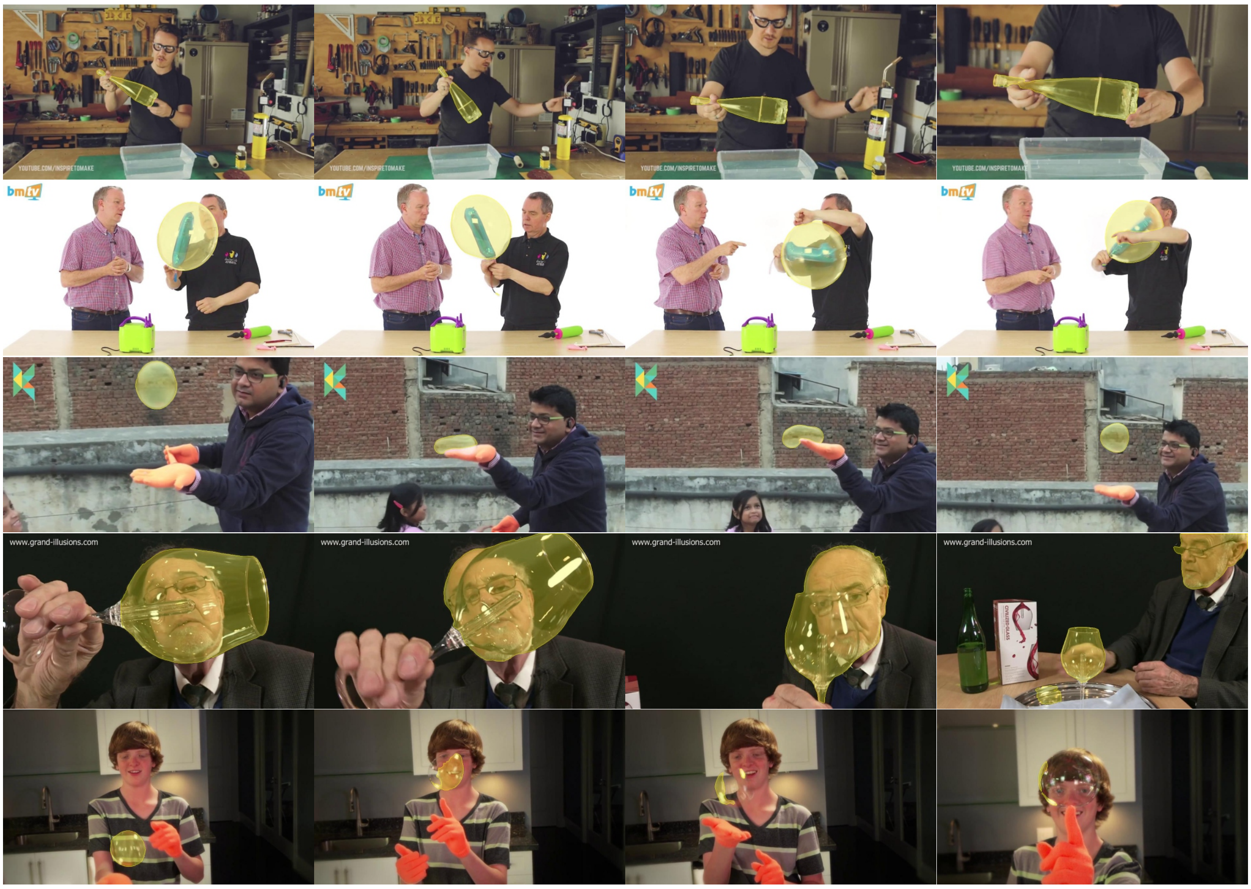

Figure 2: Qualitative examples of tracking and segmentation with Cutie [3] on TOTB dataset [6]. First three lines show successful tracking while failure cases are demonstrated in the last two lines.

phaRefine is much less resource demanding and faster deep model. An overall tracking performance when using initialization mask estimated by AlphaRefine drops by 3% compared to ground-truth initialization mask.

The simplest initialization method constructs an initialization mask by setting to one all values within the ground-truth bounding box (denoted by *Box* in Table 4) and does not require any additional computation cost. Such initialization reduces overall tracking performance obtained by a ground-truth mask by 6.3%.

When initialization mask is not available, or is too expensive to obtain, requirements of the downstream task are crucial to know, i.e., the application which uses the result of a tracking algorithm. In a case when prediction accuracy is very important and the computation resources are not critical, the SAM model is preferred to use, to obtain the initialization mask. In an opposite case, when segmentation accuracy is not so important and computation resources are limited, initialization from a bounding box mask is preferred. This experiment also demonstrates that quality of initialization mask does not impact tracking robustness – only a minor 0.1 or 0.2% performance drop has been noted.

5 Conclusion

A new dataset to evaluate tracking and segmentation performance on transparent objects was presented. An ex-

Table 4: Comparison of different methods to compute the initialization mask on the TOTB^S dataset.

Init	AUC	Accuracy	Robustness
GT mask	0.808	83.6	97.9
SAM	0.797 (↓ 1.4%)	82.3 (↓ 1.6%)	97.8 (↓ 0.1%)
AR	0.784 (↓ 3.0%)	81.2 (↓ 2.9%)	97.7 (↓ 0.2%)
Box	0.757 (↓ 6.3%)	78.2 (↓ 6.5%)	97.7 (↓ 0.2%)

isting TOTB dataset [6], annotated with bounding boxes was re-annotated with segmentation ground-truth. A semi-automatic protocol for sparse dataset annotation was presented and evaluated on transparent object tracking task. State-of-the-art bounding box and segmentation trackers were evaluated on the new dataset, which lead to several insights: (i) sparse evaluation does not significantly differ from dense evaluation, (ii) both types of trackers (bounding box and segmentation) achieve extremely high tracking robustness, while bounding box trackers lag behind segmentation trackers in accuracy, (iii) tracking accuracy of bounding box trackers could be easily improved using modern mask refinement methods (e.g., AlphaRefine or SAM) and (iv) quality of the initialization mask impacts only the accuracy of segmentation trackers, which demonstrates their robustness on initialization noise. We expect that results of this paper will allow a fair evaluation of transparent object tracking and segmentation methods, encouraging further development of this field.

Acknowledgement

This work was supported on part by the ARIS research projects and programs J2-2506, L2-3169, Z2-4459 and P2-0214, and by supercomputing network SLING (ARNES, EuroHPC Vega - IZUM).

References

[1] Cehovin, L., Leonardis, A., Kristan, M.: Visual object tracking performance measures revisited. IEEE Trans. Image Proc. **25**(3), 1261–1274 (2016)

[2] Chen, Z., Zhong, B., Li, G., Zhang, S., Ji, R.: Siamese box adaptive network for visual tracking. In: Comp. Vis. Patt. Recognition. pp. 6668–6677 (2020)

[3] Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: Comp. Vis. Patt. Recognition. pp. 3151–3161 (2024)

[4] Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: ATOM: Accurate tracking by overlap maximization. In: Comp. Vis. Patt. Recognition (2019)

[5] Danelljan, M., Bhat, G., Van Gool, L., Timofte, R.: Learning discriminative model prediction for tracking. In: Int. Conf. Computer Vision. pp. 6181–6190 (2019)

[6] Fan, H., Miththanthaya, H.A., Harshit, Rajan, S.R., Liu, X., Zou, Z., Lin, Y., Ling, H.: Transparent object tracking benchmark. In: Int. Conf. Computer Vision. pp. 10734–10743 (2021)

[7] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Int. Conf. Computer Vision. pp. 4015–4026 (2023)

[8] Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Cehovin, L.: A novel performance evaluation methodology for single-target trackers. IEEE Trans. Pattern Anal. Mach. Intell. (2016)

[9] Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kämäräinen, J.K., Danelljan, M., Zajc, L.Č., Lukežič, A., Drbohlav, O., et al.: The eighth visual object tracking VOT2020 challenge results. In: European Conference on Computer Vision Workshops (2020)

[10] Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: SiamRPN++: Evolution of siamese visual tracking with very deep networks. In: Comp. Vis. Patt. Recognition (2019)

[11] Lukezic, A., Trojer, Z., Matas, J., Kristan, M.: Trans2k: Unlocking the power of deep models for transparent object tracking. In: In Proceedings of the British Machine Vision Conference (BMVC) (2022)

[12] Lukezic, A., Trojer, Z., Matas, J., Kristan, M.: Trans2k: Unlocking the power of deep models for transparent object tracking. A New Dataset and a Distractor-Aware Architecture for Transparent Object Tracking **132**, 2729–2742 (2024)

[13] Mayer, C., Danelljan, M., Bhat, G., Paul, M., Paudel, D.P., Yu, F., Gool, L.V.: Transforming model prediction for tracking. In: Comp. Vis. Patt. Recognition (2022)

[14] Pelhan, J., Kristan, M., Lukežič, A., Matas, J., Zajc, L.Č.: Guided video object segmentation by tracking. Electrotechnical Review/Elektrotehniski Vestnik **90**(4) (2023)

[15] Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1834–1848 (2015)

[16] Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: Comp. Vis. Patt. Recognition. pp. 2411–2418 (2013)

[17] Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10448–10457 (October 2021)

[18] Yan, B., Zhang, X., Wang, D., Lu, H., Yang, X.: Alpha-refine: Boosting tracking performance by precise bounding box estimation. In: Comp. Vis. Patt. Recognition. pp. 5289–5298 (2021)