Trans2k: Unlocking the Power of Deep Models for Transparent Object Tracking

Alan Lukežič*[1]
alan.lukezic@fri.uni-lj.si
Žiga Trojer*[1]
ziga.trojer20@gmail.com
Jiří Matas[2]
matas@fel.cvut.cz
Matej Kristan[1]
matej.kristan@fri.uni-lj.si

[1] Faculty of Computer and Information Science
University of Ljubljana
Ljubljana, Slovenia

[2] Center for Machine Perception,
Czech Technical University in Prague,
Prague, Czech Republic

Abstract

Visual object tracking has focused predominantly on opaque objects, while transparent object tracking received very little attention. Motivated by the uniqueness of transparent objects in that their appearance is directly affected by the background, the first dedicated evaluation dataset has emerged recently. We contribute to this effort by proposing the first transparent object tracking *training dataset* Trans2k that consists of over 2k sequences with 104,343 images overall, annotated by bounding boxes and segmentation masks. Noting that transparent objects can be realistically rendered by modern renderers, we quantify domain-specific attributes and render the dataset containing visual attributes and tracking situations not covered in the existing object training datasets. We observe a consistent performance boost (up to 16%) across a diverse set of modern tracking architectures when trained using Trans2k, and show insights not previously possible due to the lack of appropriate training sets. The dataset and the rendering engine will be publicly released to unlock the power of modern learning-based trackers and foster new designs in transparent object tracking.

1 Introduction

Visual object tracking is a fundamental computer vision problem that emerges in a broad range of downstream applications such as human-computer interaction, surveillance, autonomous robotics and video editing, to name a few. The substantial advances observed in the last decade have been primarily driven by emergence of challenging evaluation datasets [14, 21, 26, 48] and diverse training sets [21, 38, 42] that enabled end-to-end learning of modern deep tracking architectures. While most benchmarks addressed opaque objects, very little attention has been dedicated to tracking of transparent objects. These are unique in that they are often reflective and their appearance is affected by the background texture, thus reducing the reliability of the deep features trained for opaque objects.

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

* The authors contributed equally.

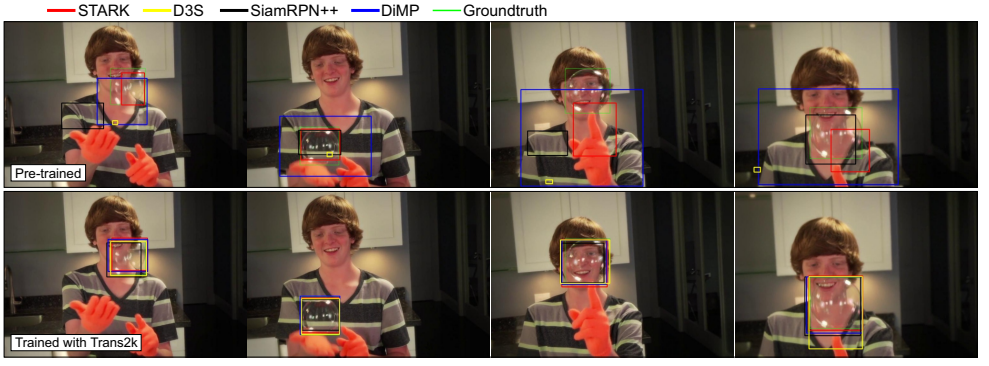

Figure 1: Trackers trained on opaque object training sets fail due to specifics of transparent object appearance dynamics (upper row). After training with the proposed Trans2k, their performance remarkably improves (bottom row).

Recently, the TOTB benchmark [15] was proposed to facilitate research in transparent object tracking. The benchmark results show that classical trackers underperform on transparent objects and that, contrary to opaque object tracking and to many other vision problems, shallow backbones outperform the deep ones. However, it is crucial to note that the results were obtained without re-training the state-of-the-art trackers on representative training sets, which opens the question whether these observations are not just a consequence of the domain shift rather than an inherent property of shallow and deep modern learning-based tracking architectures. There is thus a pressing need for a high quality transparent object training video dataset to answer this question and to potentially unlock the power of deep learning trackers, as well as to facilitate in-depth analysis and foster further research.

Construction of the training dataset presents many challenges. First, the training set should be large, diverse, and focus on visual attributes and challenging situations specific for transparent objects, which are not already covered in the opaque tracking datasets. Second, the targets should be accurately annotated. Presented with these challenges, various sequence selection and annotation protocols have emerged [15, 21, 25, 27]. In related fields like 6DoF estimation [17, 18] and scene parsing [13, 39, 53], image rendering has been applied to avoid the aforementioned issues. However, the realism of rendered general objects remains limited, reducing the training potential. We note that transparent objects are unique in that, contrary to their opaque counterparts, non-textured transparent materials may be faithfully rendered by modern renderers [17]. Thus *highly realistic* sequences with precisely specified visual attributes and pixel-level ground truth free of subjective annotation bias can be generated.

We propose the first transparent object tracking training dataset Trans2k. To maximise its utility, a protocol is designed that identifies challenging visual attributes and tracking situations not covered in existing datasets. The identified attribute ranges are then used in rendering over 2k sequences with 104,343 images overall.

A set of trackers representing the major modern deep learning approaches is evaluated on [15]. We report a consistent performance boost (up to 16%) across all architectures when trained with Trans2k. Contrary to [15], we show that deep backbones outperform shallow ones on transparent object tracking, which is consistent with observations in opaque tracking. We see transformers as the most promising approach and identify the visual attributes that future architectural designs should address to make significant progress in performance.

In summary, our contributions are: (i) Trans2k, the first training dataset for transparent object tracking that unlocks the power of deep trainable trackers and allows training bounding box or segmentation trackers, (ii) a complementary analysis on [15] with new findings indicating future research directions. The dataset and the sequence generation engine will be made publicly available. The paper reports two surprising observations: first, that transparent object tracking results are comparable to opaque object tracking for state-of-the-art trackers trained with Trans2k, and second, that training with Trans2k leads to substantial performance boost on transparent objects at minimal reduction on opaque objects. The dataset, rendering engine and instructions how to use it are available here: https://github.com/trojerz/Trans2k

2 Related Work

Object tracking. Deep trackers excel across various benchmarks [14, 15, 21, 28, 35, 48] compared to their hand-crafted counterparts. Initially, pre-trained general backbones were used for feature extraction, primarily by the discriminative correlation filter (DCF) trackers [2, 7, 8, 9, 32], which learned a discriminative localization models online during tracking. Later, backbone end-to-end training techniques that maximize DCF localization were proposed [45]. Most recently, the DCF optimization has been introduced as part of the deep network. Milestone representatives were proposed in [10, 11], which also proposed a post-processing network for bounding box refinement that accounted for target aspect changes. In parallel, siamese trackers have been explored and grown into a major tracker design branch. The seminal work [1] trained AlexNet-based network [29] such that localization accuracy is maximized simply by correlation between a template and search region in feature space. These trackers afford fast processing since no training is required during tracking. Siamese trackers were extended by anchor-based region proposal networks [30, 31] and recently an anchor-free extension has been proposed [6] with improved localization performance. Drawing on advances in object detection [3], transformer-based trackers have recently emerged [5, 46, 52]. These are the current state-of-the-art, and computationally efficient with remarkable real-time performance [28].

Benchmarks. The developments in visual object tracking have been facilitated by introduction of benchmarks. The first widely-used benchmark [48, 49] proposed a dataset and evaluation protocol that allowed standardised comparsion of trackers. Later, the VOT initiative explored dataset construction as well as performance evaluation protocols for efficient in-depth analysis [25, 26, 27]. Further improvements were made in the subsequent yearly challenges, e.g., [28, 35]. With advent of deep learning, tracking training sets have emerged. [38] constructed a huge training set from public video repository and applied a semi-automatic annotation. Recently, [21] presented ten thousand annotated video sequences, divided into a large training and a smaller evaluation set. Concurrently, a long-term tracking benchmark [14] with fifteen pre-defined categories, containing training and test set was proposed. All these benchmarks focus on opaque objects, while recently as transparent object tracking evaluation dataset [15] has been proposed. However, training datasets for transparent object tracking have not been proposed.

Use of synthesis. Rendering has been previously considered in computer vision to avoid costly manual dataset acquisition. In [24, 40], synthetic data was generated by a video game engine, which provided an unlimited amount of annotated training data for various computer vision tasks. A rendered dataset of urban scenes, Synthia [41], was shown to substantially

improve the trained deep models for semantic segmentation. A similar dataset [47] was proposed for training and evaluation of scene parsing networks. A fine-grained vegetation and terrain dataset [36] was recently proposed for training drivable surfaces and natural obstacles detection networks in outdoor scenes. [44] showed that foreground and background should be treated differently when training segmentation on synthetic images. The benefits of using mixed real and synthetic 6DoF training data has been recently shown in [19]. The major 6DoF object detection challenge [20] thus provides a combination of real and synthetic images for training as well as evaluation. Synthesis has been used in the UAV123 tracking benchmark [37] in which eight of the sequences are rendered by a game engine. A rendering approach was used in [4] to parameterize camera motion for fine-grained tracker performance analysis. However, using synthetic data for training in visual tracking remains unexplored.

Transparent objects. Highlighting the difference from opaque counterparts, transparent objects have been explored in computer vision in various tasks. Recognition of transparent objects was studied in [16, 34], while 3D shape estimation and reconstruction of transparent objects on RGB-D images was proposed in [23, 43]. Segmentation of transparent objects has been studied in [22, 51], while a benchmark was proposed in [50]. All these works consider single-image tasks and little attention has been dedicated to videos. In fact, a transparent object tracking benchmark [15] has been proposed only recently and reported a performance gap between transparent and opaque object tracking. However, due to the lack of a dedicated training dataset, the gap source remains unclear.

3 Trans2k dataset

Transparent objects, which are often reflective and glass-like, can be rendered with a high level of realism by the modern photo-realistic rendering engines [12]. In our approach, we first identify and parameterize the sequence attributes specific to transparent objects (Section 3.1). A BlenderProc-based sequence generator is implemented that enables parameterized sequence rendering. Attribute levels useful for learning are identified empirically and the final training dataset is generated (Section 3.2).

3.1 Parametrization of sequence attributes

The dataset should reflect the diversity of visual attributes typical for transparent object tracking scenes for efficient learning. After carefully examining various videos of transparent and opaque objects, the following attributes were identified (Figure 2).

Scene background. Since background affects the transparent object appearance, a high background diversity is required in training. We ensure this by randomly sampling videos from GoT10k [21] training set and use them as backgrounds over which the transparent object is rendered.

Object types. 3D models of 25 object types from open source online repositories are selected with several instances of the same type. The set was chosen such to cover a range of nontrivial as well as smooth shapes, with some objects rendered with empty and some with full volume. This amounts to 148 object instances.

Target motion. To increase the object-background appearance diversity, the objects are moving in the videos. The motion trajectory is generated by a cubic Hermite spline spanned by four uniformly sampled points. The motion dynamics is not critical in training, since

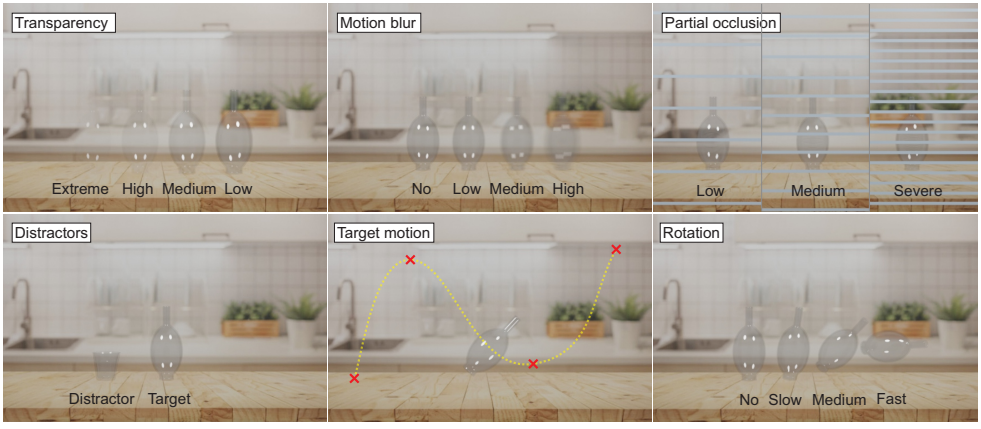

Figure 2: Trans2k attribute levels for "Transparency", "Motion blur", "Partial occlusion", "Distractor" (binary), "Target motion" (four control points) and "Rotation".

deep models are typically trained on pairs of image patches cropped at target position. Thus a constant velocity is applied.

Distractors. In realistic environments, the target may be surrounded by other visually similar transparent objects (e.g., glasses on a table), which act as distractors. We thus render an additional transparent object following the target object. The distractor object is from a different type to keep the appearance-based localization learning task feasible.

Transparency. The level of transparency crucially affects the target appearance. We thus identify four levels ranging from clearly visible to nearly invisible.

Motion blur. Fast motions, depending on the aperture speed, result in various levels of blurring. We identify four levels of blur intensity, ranging from no blurring to extreme blurriness.

Partial occlusion. Objects are commonly occluded by other objects in practical situations (e.g., handling of the target). We thus simulate partial occlusions by rendering coloured stripe pattern moving across the video frame. The stripe width is fixed, while the occlusion intensity is simulated by the number of stripes (0, 7, 11, 20) per image, i.e., from zero to severe occlusion.

Rotation. To present realistic object appearance change, the object rotates in 3D in addition to position change. The rotation dynamics is specified by the angular velocity along each axis, which is kept constant throughout the sequence. We identify four rotation speed levels, (0, 1.3, 5.4, 10.6) degrees per frame, thus ranging from no rotation to fast rotation.

3.2 Attribute selection and dataset generation

To maximize the dataset application utility, the sequences should be complementary to existing datasets from tracking perspective and should focus on attributes that the learning-based trackers cannot already learn from opaque object tracking training sets. An empirical study was designed to determine which intensity levels of the attributes (i) transparency, (ii) partial occlusion, (iii) rotation and (iv) motion blur should be considered in the final dataset. The intensity levels are visualized in Figure 2.

Seven state-of-the-art deep learning trackers pretrained on opaque object tracking datasets

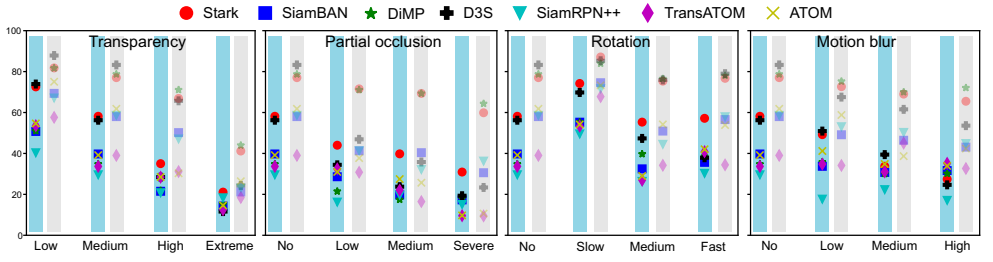

Figure 3: Average IoU of trackers reflect the difficulty level of individual attribute intensity. The blue shaded columns show performance of trackers pre-trained on opaque datasets, while the gray shaded column shows performance after training with Trans2k.

(see Section 4 for details) that cover the major current trends in tracking were selected. The difficulty level of individual attribute is quantified as the overall performance of these trackers on test sequences rendered with that attribute. Specifically, for each attribute level, five sequences were sampled from Got10K [21] training set and used as backgrounds in the rendered sequences. The same background sequences with same object, traveling over the same trajectory are used with all attributes to ensure consistent evaluation. This resulted in 80 test sequences (4 attributes $\times$ 4 levels $\times$ 5 variations).

The results are shown in Figure 3. We observe that most of the attribute levels result in performance reduction and are thus kept as relevant in our final dataset, except from two at which the trackers score quite high. The lowest transparency level and zero rotation appear to be well addressed by the opaque object training sets, thus we decide not to include them in our dataset for better use of its capacity. The following parameters are thus applied when rendering Trans2k. The GoT10k training set sequences are sampled at random and at most once. All object types are sampled with equal probability. The transparency levels (excluding the lowest level) are sampled with equal probability. Blur presence in a sequence is sampled with 0.15 probability, with blur levels sampled uniformly. Occlusion presence is sampled with 0.2 probability, while occlusion levels are sampled uniformly. Rotation level is uniformly sampled. The resulting training dataset Trans2k thus contains 2,039 challenging sequences and 104,343 frames in total.

Since the sequences are rendered, the ground truth can be exactly computed. We provide the ground truth in two standard forms, the widely accepted target enclosing axis-aligned bounding-box and the segmentation mask to cater to the emerging segmentation trackers [35]. The ground truths for distractors are generated as well. Trans2k is thus the first dataset with per-frame distractor annotation to facilitate development of future learning-based methods that could exploit this.

4 Experiments

4.1 Selected trackers and training setup

We selected state-of-the-art learning-based trackers that cover the major trends in modern architecture designs for validating Trans2k: (i) two siamese trackers SiamRPN++ [31] and SiamBAN [6], (ii) two deep correlation filter trackers ATOM [10] and DiMP [11], (iii) the recent state-of-the-art transparent object tracker TransATOM [35], and (iv) a transfomer-

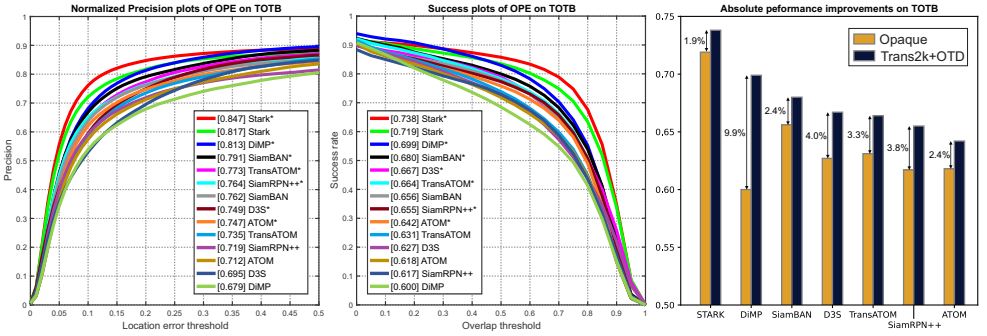

Figure 4: Trackers evaluated on TOTB dataset shown in precision and success plots. Trackers trained with Trans2k are denoted by a star (*). The right graph shows absolute improvements in tracking performance measured by the AUC measure after training with the proposed Trans2k.

based tracker STARK [52]. These trackers localize the target by a bounding box. To account for the recent trend in localization by per-pixel segmentation [35], we include (v) the recent state-of-the-art segmentation-based tracker D3S [33].

During training, the trackers were initialized by the pre-trained weights provided by their authors, while all the training details were the same as in the original implementations. The trackers were trained for 50 epochs with 10000 training samples per epoch. Since Trans2k was designed as a complementary dataset covering situations not present in existing datasets, the training considers samples from Trans2k as well as opaque object sequences. In particular, we merged the opaque training datasets GOT10k [21], LaSoT [14] and TrackingNet [38] into a single dataset, abbreviated as opaque object training dataset (OTD). A training batch is then constructed by sampling from Trans2k and OTD with 5:3 ratio.

4.2 Validation of Trans2k

We first validated the contribution of Trans2k by measuring performance of trackers on the recent transparent object tracking benchmark TOTB [15]. Following the regime described in Section 4.1 the selection of seven state-of-the-art trackers was trained using Trans2k. Their performance was then compared to their original performance, i.e., when trained only with opaque object tracking sequences. Thus any change in performance is contributed only by the training dataset. The trackers were evaluated by the standard one-pass evaluation protocol (OPE) that quantifies the performance by AUC and center error measures on success and precision plots. For more information on the protocol, please refer to [15, 48].

The results are shown in Figure 4. The performance of all trackers substantially improved when trained using Trans2k. The performance gains are at a level usually expected for a clear methodological improvement. Recently, TransATOM [15], a transparent object tracking extension of ATOM [10], was proposed, which outperformed ATOM by 2.1%. Without any methodological modification and only training with Trans2k, ATOM *outperforms* this extension by 1.7%. Nevertheless, TransATOM gains 3.3% when trained with Trans2k. The largest performance boost is achieved by DiMP, which improves by over 16% and scores as the second-best among all the tested trackers. Consistent with the observation on opaque object tracking benchmarks, the transformer-based tracker STARK achieves the best per-

formance. Note that even without training with Trans2k, STARK surpasses all trackers, but when trained with Trans2k, an additional healthy 2.5% performance boost is observed. Since Trans2k provides segmentation ground truths in addition to bounding boxes, it boosts the segmentation-based tracker D3S [33] as well. The version trained with Trans2k gains a remarkable 6% in performance.

4.3 Re-evaluating the significance of backbone depth

The recent benchmark [15] reported a remarkable case that, specific to transparent object tracking, shallow backbones outperform deep ones, which conflicts common observations in opaque object tracking. Since [15] could only analyze the performance with using opaque tracking training sets, we re-evaluate this claim but in the context of using a transparent object tracking training set. We select three deep discriminative correlation filters DiMP [10], ATOM [10] and TransATOM [15] and study their performance with a shallow (ResNet18) and a deep (ResNet50) backbone when trained with and without transparent objects.

Results in Table 1 reveal that ATOM and TransATOM with shallow backbones indeed outperform their deep backbone counterparts. In contrast, DiMP with deep backbone substantially outperforms the shallow backbone counterpart. This apparent discrepancy comes from the different designs of ATOM and DiMP. While ATOM uses a pre-trained backbone and allows training of only post-processing steps, DiMP trains the backbone as well. The ATOM's apparent preference of shallow backbones comes from the fact that shallow backbones generalize better to transparent objects when trained only with opaque training examples. To verify this, we replace ATOM's and TransATOM's backbone by those trained by DiMP. Both trackers improve their performance with deep backbones trained on transparent objects compared to shallow ones. Interestingly, their deep backbone variants reach performance near DiMP's and their *performance difference becomes negligible* – thus properly trained vanilla ATOM should be preferred to its more complex extension TransATOM. The experiments thus reveal, that deep backbones in fact lead to substantial improvements over shallow counterparts, if trained on the transparent object dataset.

Table 1: Tracking performance (AUC) of three trackers using different backbones. *Opaque* indicates training with only OTD, + *Trans2k* to using the transparent dataset as well. Pre-trained ResNet18 and ResNet50 backbones are denoted by R18 and R50, respectively, while their versions trained by DiMP are denoted by D18 and D50.

	DiMP		ATOM				TransATOM			
	D18	D50	R18	R50	D18	D50	R18	R50	D18	D50
Opaque	0.552	0.600	0.618	0.608	0.551	0.588	0.631	0.608	0.582	0.603
+ Trans2k	0.613	0.699	0.642	0.648	0.629	0.695	0.664	0.664	0.647	0.697

4.4 How does Trans2k affect opaque object tracking?

To quantify how much the trackers trained with Trans2k lose in generalization to opaque objects, we evaluate the trackers on the GOT10k [21] validation dataset. Table 2 shows results for trackers trained only with OTD and with added Trans2k (as described in Section 4.1). The tracking performance on opaque objects slightly drops, but still remains high.

This result suggests that, while substantial boosts are observed in transparent object tracking (Figure 4) with the use of Trans2k, the generalization to opaque objects is not lost.

Table 2: Tracking performance (AUC) on the opaque tracking dataset GoT-10k val. *Opaque* – training with only OTD, + *Trans2k* – using the transparent dataset as well.

	STARK	DiMP	SiamBAN	D3S	TransATOM	SiamRPN	ATOM
Opaque	0.777	0.706	0.679	0.676	0.662	0.656	0.650
+ Trans2k	0.752	0.696	0.676	0.663	0.650	0.656	0.650

4.5 The role of using opaque objects in training

To further study the impact of the training set content from perspective of the presence of opaque and transparent objects, the training sets were varied. We selected two well-known state-of-the-art trackers that performed well in our previous experiments, yet could be trained sufficiently fast. The deep discriminative correlation filter DiMP [1] and the siamese tracker SiamBAN [6] were selected. The original versions trained by the authors were evaluated on TOTB [15] along with the versions re-trained using the following variations of the training set: (i) only Trans2k without OTD, (ii) only OTD, (iii) Trans2k+OTD. In experiments (i), (ii) and (iii) the tracker networks are initialized by their pre-trained models provided by the authors. Thus an additonal experiment (iv) is performed where the trackers were trained from scratch using the dataset from (iii). The trained trackers were evaluated on TOTB [15].

Results in Table 3 show that using only Trans2k reduces the tracking performance compared to training on opaque objects training datasets. A closer look revealed that the trackers trained only with Trans2k tend to focus on transparent objects in general rather than localizing the target, which was reflected in tracker often jumping to nearby transparent objects when multiple such objects were close to the target. Training with OTD when initialized with original tracker parameters does not bring improvements in general (DiMP performance drops slightly, while that of SiamBAN increases a bit). However, when using transparent *as well as* opaque objects in the training set, the performance improves substantially. When training from scratch, the performance of both trackers drops compared to the version initialized with pre-trained networks. This suggests pre-training is beneficial for both trackers, but particularly for SiamBAN as it requires learning more parameters than DiMP.

Table 3: Comparison of different training setups for SiamBAN and DiMP. Performance of pre-trained trackers is indicated by *orig.*, while *scr.* indicates training from scratch.

	orig.	Trans2k	OTD	OTD+Trans2k	scr.
DiMP	0.600	0.554	0.584	0.699	0.667
SiamBAN	0.656	0.650	0.658	0.680	0.649

5 Conclusion

The first transparent object tracking training dataset Trans2k is proposed. The fact that transparent objects can be sufficiently realistically rendered by modern renderers is exploited. Us-

ing a specialized protocol, we identified visual attributes not covered well in existing datasets and rendered a dataset with over 2k training sequences containing transparent objects.

Trans2k was validated on the recent transparent object tracking benchmark TOTB [48]. Training with Trans2k improves performance at levels usually observed in fundamental methodological advancements in tracking algorithms. This behavior is observed over a wide range of tracking methodologies. Analysis shows that significant performance gains in transparent object tracking come at a minor performance loss in opaque object tracking, which indicates to excellent generalization of modern trackers. In contrast to the findings in [15], experiments show that trackers benefit from training deeper backbones on transparent objects. Additional experiments showed the benefits of using transparent as well as opaque objects in the training dataset. Overall, the best performance was observed with transformers.

While the field of transparent object tracking has recently obtained an excellent test set [15], the main ingredient crucial for advancements, i.e., a curated training set was missing. Trans2k fills this void and will enable future development of new learnable modules specifically addressing the challenges in transparent object tracking, thus fully unlocking the power of modern deep learning trackers on this scientifically interesting domain. Our sequence generator engine will be released along with Trans2k. We envision that the engine will allow inovative learning modes in which the sequences with specific challenges can be generated on demand to specialize the trackers to niche tasks or to improve their overall performance. In addition, the rendering engine could be used to generate training data for 6-DoF video pose estimation, thus benefiting research beyond 2D transparent object tracking.

Acknowledgements: This work was supported by Slovenian research agency program P2-0214 and projects Z2-4459, J2-2506 and J2-9433. J. Matas was supported by the Technology Agency of the Czech Republic, project No. SS05010008.

References

[1] Luca Bertinetto, Jack Valmadre, João F Henriques, Andrea Vedaldi, and Philip H S Torr. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision Workshops*, pages 850–865, 2016.

[2] Goutam Bhat, Joakim Johnander, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Unveiling the power of deep tracking. In *Proc. European Conf. Computer Vision*, pages 493–509, 2018.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. European Conf. Computer Vision*, pages 213–229, 2020.

[4] Luka Cehovin Zajc, Alan Lukezic, Ales Leonardis, and Matej Kristan. Beyond standard benchmarks: Parameterizing performance evaluation in visual object tracking. In *Int. Conf. Computer Vision*, 2017.

[5] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Comp. Vis. Patt. Recognition*, pages 8126–8135, 2021.

[6] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Comp. Vis. Patt. Recognition*, pages 6668–6677, 2020.

[7] Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Convolutional features for correlation filter based visual tracking. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 621–629, 2015.

[8] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: learning continuous convolution operators for visual tracking. In *Proc. European Conf. Computer Vision*, pages 472–488, 2016.

[9] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *Comp. Vis. Patt. Recognition*, pages 6638–6646, 2017.

[10] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate tracking by overlap maximization. In *Comp. Vis. Patt. Recognition*, 2019.

[11] Martin Danelljan, Goutam Bhat, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Int. Conf. Computer Vision*, pages 6181–6190, 2019.

[12] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Dmitry Olefir, Tomas Hodan, Youssef Zidan, Mohamad Elbadrawy, Markus Knauer, Harinandan Katam, and Ahsan Lodhi. Blenderproc: Reducing the reality gap with photorealistic rendering. In *International Conference on Robotics: Sciene and Systems, RSS 2020*, 2020.

[13] Yilun Du, Zhijian Liu, Hector Basevi, Ales Leonardis, Bill Freeman, Josh Tenenbaum, and Jiajun Wu. Learning to exploit stability for 3d scene parsing. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[14] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Comp. Vis. Patt. Recognition*, June 2019.

[15] Heng Fan, Halady Akhilesha Miththanthaya, Harshit, Siranjiv Ramana Rajan, Xiaoqiong Liu, Zhilin Zou, Yuewei Lin, and Haibin Ling. Transparent object tracking benchmark. In *Int. Conf. Computer Vision*, pages 10734–10743, 2021.

[16] Mario Fritz, Gary Bradski, Sergey Karayev, Trevor Darrell, and Michael Black. An additive latent feature model for transparent object recognition. *Advances in Neural Information Processing Systems*, 22, 2009.

[17] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiri Matas, and Carsten Rother. Bop: Benchmark for 6d object pose estimation. In *Proc. European Conf. Computer Vision*, September 2018.

[18] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *Comp. Vis. Patt. Recognition*, 2020.

[19] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *Comp. Vis. Patt. Recognition*, pages 11703–11712, 2020.

[20] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *Proc. European Conf. Computer Vision*, pages 577–594, 2020.

[21] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[22] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Comp. Vis. Patt. Recognition*, pages 8602–8611, 2020.

[23] Ulrich Klank, Daniel Carton, and Michael Beetz. Transparent object detection and reconstruction on a mobile platform. In *Int. Conf. Robotics and Automation*, pages 5971–5978. IEEE, 2011.

[24] Philipp Krahenbuhl. Free supervision from video games. In *Comp. Vis. Patt. Recognition*, pages 2955–2964, 2018.

[25] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Čehovin, G. Nebehay, G. Fernandez, and T. et al. Vojir. The visual object tracking vot2013 challenge results. In *Vis. Obj. Track. Challenge VOT2013, In conjunction with ICCV2013*, pages 98–111, 2013.

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

[27] Matej Kristan, Roman Pflugfelder, Aleš Leonardis, Jiri Matas, Luka Čehovin, Georg Nebehay, Tomas Vojir, and Gustavo et al. Fernandez. The visual object tracking vot2014 challenge results. In *Proc. European Conf. Computer Vision*, pages 191–217, 2014.

[28] Matej Kristan, Jiří Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Cehovin, Alan Lukežič, Ondrej Drbohlav, Jani Käpylä, Gustav Häger, Song Yan, Jinyu Yang, Zhongqun Zhang, and Gustavo Fernández. The ninth visual object tracking vot2021 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2711–2738, 2021.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[30] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Comp. Vis. Patt. Recognition*, 2018.

[31] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *Comp. Vis. Patt. Recognition*, 2019.

[32] X. Liu. Deep correlation filters for robust visual tracking. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021.

[33] Alan Lukežič, Jiří Matas, and Matej Kristan. D3S – a discriminative single shot segmentation tracker. In *Comp. Vis. Patt. Recognition*, pages 7133–7142, 2020.

[34] Kazuki Maeno, Hajime Nagahara, Atsushi Shimada, and Rin-ichiro Taniguchi. Light field distortion feature for transparent object recognition. In *Comp. Vis. Patt. Recognition*, pages 2786–2793, 2013.

[35] Matej Kristan, et al. The eighth visual object tracking VOT2020 challenge results. In *European Conference on Computer Vision Workshops*, 2020.

[36] Kai A. Metzger, Peter Mortimer, and Hans-Joachim Wuensche. A fine-grained dataset and its efficient semantic segmentation for unstructured driving scenarios. In *Proc. Int. Conf. Pattern Recognition*, pages 7892–7899, 2021.

[37] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *Proc. European Conf. Computer Vision*, pages 445–461, 2016.

[38] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In *Proc. European Conf. Computer Vision*, September 2018.

[39] Mengshi Qi, Yunhong Wang, Jie Qin, and Annan Li. Ke-gan: Knowledge embedded generative adversarial networks for semi-supervised scene parsing. In *Comp. Vis. Patt. Recognition*, 2019.

[40] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Int. Conf. Computer Vision*, pages 2213–2222, 2017.

[41] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Comp. Vis. Patt. Recognition*, pages 3234–3243, 2016.

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252, 2015.

[43] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *Int. Conf. Robotics and Automation*, pages 3634–3642. IEEE, 2020.

[44] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M Alvarez. Effective use of synthetic data for urban scene semantic segmentation. In *Proc. European Conf. Computer Vision*, pages 84–100, 2018.

[45] Jack Valmadre, Luca Bertinetto, Joao Henriques, Andrea Vedaldi, and Philip H. S. Torr. End-to-end representation learning for correlation filter based tracking. In *Comp. Vis. Patt. Recognition*, pages 2805–2813, 2017.

[46] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Comp. Vis. Patt. Recognition*, pages 1571–1580, 2021.

[47] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018.

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

[49] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Comp. Vis. Patt. Recognition*, pages 2411– 2418, 2013.

[50] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *Proc. European Conf. Computer Vision*, pages 696–711. Springer, 2020.

[51] Yichao Xu, Hajime Nagahara, Atsushi Shimada, and Rin-ichiro Taniguchi. Transcut: Transparent object segmentation from a light-field image. In *Int. Conf. Computer Vision*, pages 3442–3450, 2015.

[52] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10448–10457, October 2021.

[53] Junyi Zhang, Ziliang Chen, Junying Huang, Liang Lin, and Dongyu Zhang. Few-shot structured domain adaptation for virtual-to-real scene parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.