International Journal of Computer Vision (2024) 132:2729–2742
https://doi.org/10.1007/s11263-024-02010-0

A New Dataset and a Distractor-Aware Architecture for Transparent Object Tracking

Alan Lukežič[1] · Žiga Trojer[1] · Jiří Matas[2] · Matej Kristan[1]

Received: 31 March 2023 / Accepted: 16 January 2024 / Published online: 16 February 2024

Abstract
Performance of modern trackers degrades substantially on transparent objects compared to opaque objects. This is largely due to two distinct reasons. Transparent objects are unique in that their appearance is directly affected by the background. Furthermore, transparent object scenes often contain many visually similar objects (distractors), which often lead to tracking failure. However, development of modern tracking architectures requires large training sets, which do not exist in transparent object tracking. We present two contributions addressing the aforementioned issues. We propose 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. Standard trackers trained on this dataset consistently improve by up to 16%. Our second contribution is a new distractor-aware transparent object tracker (DiTra) that treats localization accuracy and target identification as separate tasks and implements them by a novel architecture. DiTra sets a new state-of-the-art in transparent object tracking and generalizes well to opaque objects.

Keywords Visual object tracking · Transparent object tracking · Distractors

1 Introduction

Visual object tracking is a fundamental computer vision problem with far-reaching applications in human-computer interaction, surveillance, autonomous robotics, and video editing, among others. The significant progress observed over the past decade can be attributed to the emergence of challenging evaluation datasets (Wu et al., 2015; Kristan et al., 2016; Huang et al., 2019; Fan et al., 2019) and diverse training sets (Muller et al., 2018; Russakovsky et al., 2015; Huang et al., 2019), which have facilitated the end-to-end learning of modern deep tracking architectures.

While numerous benchmarks have addressed opaque objects, tracking of transparent objects has received comparatively little attention. These objects are unique due to their reflective nature and dependence on background texture, which reduces the effectiveness of deep features trained for opaque objects, as shown in Fig. 1.

Transparent objects, e.g., cups and glasses are common in households. Thus a household robot or ambient intelligence systems for scene and activity understanding will rely on accurate tracking and localization of such objects. Furthermore, transparent object localization is already crucial in modern automated end-of-line quality control systems such as in the glassmaking industry.

The recent transparent object tracking benchmark (TOTB), demonstrated that trackers based on deep learning outperform traditional (non-deep learning) methods on transparent objects, even when trained on opaque objects. Furthermore, the benchmark revealed that the performance of the state-of-the-art trackers designed for opaque objects drops when applied to transparent objects. However, these results were obtained without re-training the trackers on representative training sets, which raises the question of whether the

Communicated by Guang Yang.

✉ Alan Lukežič
alan.lukezic@fri.uni-lj.si

Žiga Trojer
ziga.trojer20@gmail.com

Jiří Matas
matas@fel.cvut.cz

Matej Kristan
matej.kristan@fri.uni-lj.si

[1] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia

[2] Center for Machine Perception, Czech Technical University in Prague, Karlovo namesti 13, 12135 Prague, Czech Republic

Fig. 1 A tracker STARK* trained on opaque objects fails on transparent objects, while its performance remarkably improves after training on the proposed Trans2k dataset (first and second row). Both versions, however, fail in presence of visual distractors (third and fourth row), while the proposed DiTra comfortably tracks due to the new distractor-aware visual model

observed performance drop is a consequence of a domain shift, rather than an inherent property of the problem. Consequently, there is an urgent need for a high-quality transparent object training video dataset to address this question and potentially unlock the power of deep learning trackers.

In general, the construction of training datasets presents numerous challenges. First, the dataset must be large, diverse, and focused on visual attributes and challenging scenarios specific to transparent objects, that are not covered in opaque tracking datasets. Second, the targets should be annotated accurately. Given these challenges, various sequence selection and annotation protocols have been developed for related problems (Huang et al., 2019; Kristan et al., 2013; Fan et al., 2021; Kristan et al., 2014). In 6DoF estimation (Hodan et al., 2020a, 2018) and scene parsing (Du et al., 2018; Zhang et al., 2019; Qi et al., 2019), image rendering has been employed to circumvent these issues. While the realism of rendered opaque objects may be limited, transparent objects are unique in that non-textured transparent materials can be rendered faithfully by modern renderers (Denninger et al., 2020). This allows generating highly realistic sequences with precisely specified visual attributes and pixel-level ground truth, free of subjective annotation bias. We exploit this property and introduce the first transparent object tracking training dataset, Trans2k, which is our first contribution.[1]

An intriguing aspect of videos featuring transparent objects is the frequent presence of multiple visually similar transparent objects, or distractors (Fig. 1). For instance, kitchen or laboratory scenes often contain several glasses and bottles, while crowded scenes are commonplace in industrial manufacturing lines producing identical object types (i.e., distractors). Efficient distractor handling mechanisms are thus essential to achieve robust tracking of transparent objects.

Our second contribution is a new distractor-aware transparent object tracker (DiTra), which addresses the situations when multiple objects, visually similar to the target (distractors), are present in the scene. DiTra treats target localization accuracy (i.e., precise bounding box estimation) and localization robustness (i.e., selecting the correct target among similar objects) as distinct problems (Fig. 1). A common backbone encodes the image, while separate branches are utilized to extract features specialized for localization accuracy and robustness. These features are then fused into target-specific localization features and regressed into a bounding box. The proposed tracker is able to track an arbitrary transparent object, regardless of the object category.

In summary, our contributions include: (i) Trans2k, the first training dataset for transparent object tracking that unlocks the power of deep trainable trackers and allows for training bounding box or segmentation trackers, and (ii) the accuracy/robustness split architecture with the distractor-aware block for computing robust localization features.

A variety of trackers representing major modern deep learning approaches is evaluated on TOTB (Fan et al., 2021). After re-training on Trans2k, a consistent performance boost (up to 16%) is observed across all architectures. The proposed DiTra outperforms all re-trained trackers, setting a new state-of-the-art on the TOTB benchmark Fan et al. (2021), making it a strong baseline for this task.

We initially presented the Trans2k dataset in a conference paper Lukezic et al. (2022). Here we further explore its performance contributions to existing state-of-the-art trackers. We also propose a new tracker DiTra specialized for transparent objects, and demonstrate substantial benefits from training on Trans2k.

2 Related Work

2.1 Object Tracking

Deep trackers excel across various benchmarks (Kristan et al., 2020, 2021; Huang et al., 2019; Fan et al., 2019; Wu et al.,

[1] The data that support the findings of this study are openly available in Github repository at https://github.com/trojerz/Trans2k.

2015; Fan et al., 2021) compared to their hand-crafted counterparts. Initially, pre-trained general backbones were used for feature extraction, primarily by the discriminative correlation filter (DCF) trackers (Danelljan et al., 2016, 2017; Bhat et al., 2018; Danelljan et al., 2015; Liu, 2021), which learned a discriminative localization models online during tracking. Later, backbone end-to-end training techniques that maximize DCF localization were proposed (Valmadre et al., 2017). Most recently, the DCF optimization has been introduced as part of the deep network. Milestone representatives were proposed in (Danelljan et al., 2019a, b; Bhat et al., 2020), 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 (Bertinetto et al., 2016) trained AlexNet-based network (Krizhevsky et al., 2012) 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 (Li et al., 2018, 2019) and recently an anchor-free extension has been proposed (Chen et al., 2020) with improved localization performance. Drawing on advances in object detection (Carion et al., 2020), transformer-based trackers have recently emerged (Yan et al., 2021; Chen et al., 2021; Wang et al., 2021; Mayer et al., 2022). These are the current state-of-the-art, and computationally efficient with remarkable real-time performance (Kristan et al., 2021).

2.2 Benchmarks

The developments in visual object tracking have been facilitated by introduction of benchmarks. The first widely-used benchmark (Wu et al., 2015, 2013) 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 (Kristan et al., 2016, 2013, 2014). Further improvements were made in the subsequent yearly challenges, e.g., (Kristan et al., 2020, 2021). With advent of deep learning, tracking training sets have emerged. (Muller et al., 2018) constructed a huge training set from public video repository and applied a semi-automatic annotation. Recently, (Huang et al., 2019) presented ten thousand annotated video sequences, divided into a large training and a smaller evaluation set. Concurrently, a long-term tracking benchmark (Fan et al., 2019) 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 (Fan et al., 2021) has been proposed. However, training datasets for transparent object tracking have not been proposed.

2.3 Use of Synthesis

Rendering has been previously considered in computer vision to avoid costly manual dataset acquisition. In (Krahenbuhl, 2018; Richter et al., 2017), 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 (Ros et al., 2016), was shown to substantially improve the trained deep models for semantic segmentation. A similar dataset (Wrenninge & Unger, 2018) was proposed for training and evaluation of scene parsing networks. A fine-grained vegetation and terrain dataset (Metzger et al., 2021) was recently proposed for training drivable surfaces and natural obstacles detection networks in outdoor scenes. (Saleh et al., 2018) 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 (Hodan et al., 2020b). The major 6DoF object detection challenge (Hodaň et al., 2020) thus provides a combination of real and synthetic images for training as well as evaluation. Synthesis has been used in the UAV123 tracking benchmark (Mueller et al., 2016) in which eight of the sequences are rendered by a game engine. A rendering approach was used in (Cehovin Zajc et al., 2017) to parameterize camera motion for fine-grained tracker performance analysis. However, using synthetic data for training in visual tracking remains unexplored.

2.4 Transparent Object Sensing

Highlighting the difference from opaque counterparts, transparent objects have been explored in computer vision in various tasks. Recognition of transparent objects was studied in (Fritz et al., 2009; Maeno et al., 2013), while 3D shape estimation and reconstruction of transparent objects on RGB-D images was proposed in (Klank et al., 2011; Sajjan et al., 2020). Segmentation of transparent objects has been studied in (Xu et al., 2015; Kalra et al., 2020), while a benchmark was proposed in (Xie et al., 2020). All these works consider single-image tasks and little attention has been dedicated to videos. In fact, a transparent object tracking benchmark (Fan et al., 2021) 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 (Denninger et al.,

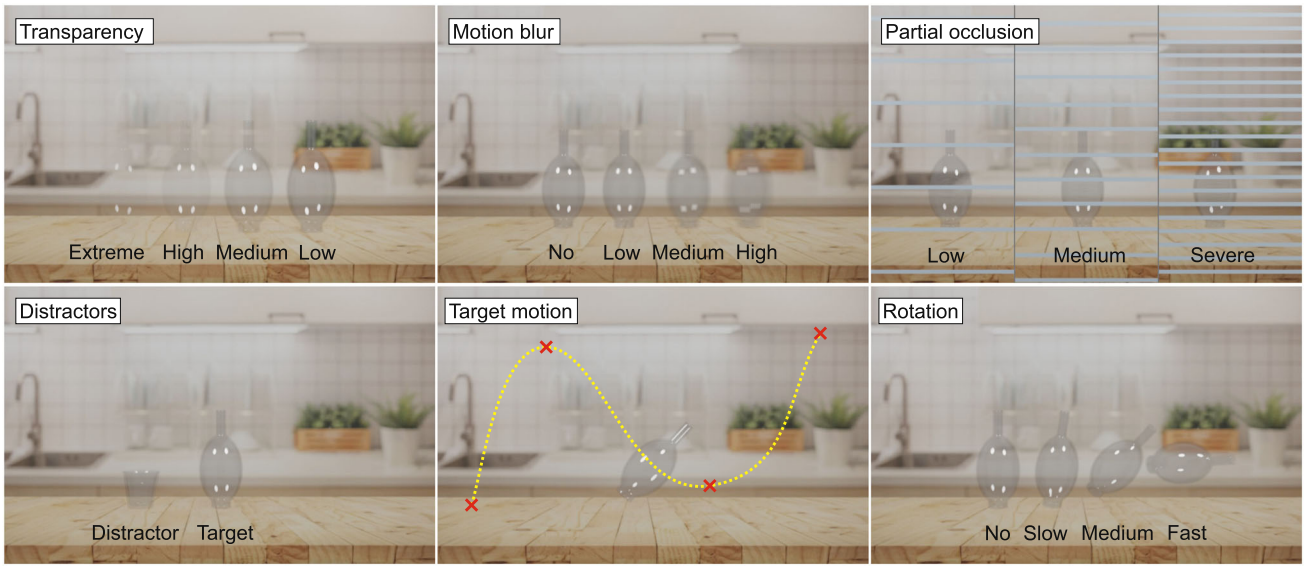

Fig. 2 Trans2k attribute levels for "Transparency", "Motion blur", "Partial occlusion", "Distractor" (binary), "Target motion" (four control points) and "Rotation"

2020). In our approach, we first identify and parameterize the sequence attributes specific to transparent objects in Sect. 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 presented in Sect. 3.2.

3.1 Parametrization of Sequence Attributes

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

3.1.1 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 (Huang et al., 2019) training set and use them as backgrounds over which the transparent object is rendered. Sampled background sequences are highly diverse, including indoor and outdoor environments, and scenes from sports, nature, marine and traffic, to name just a few.

3.1.2 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, which are visualized in Fig. 3.

3.1.3 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 2D points. The motion dynamics is not critical in training, since deep models are typically trained on pairs of image patches cropped at target position. Thus a constant velocity is applied.

3.1.4 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 of a different type to keep the appearance-based localization learning task feasible.

3.1.5 Transparency

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

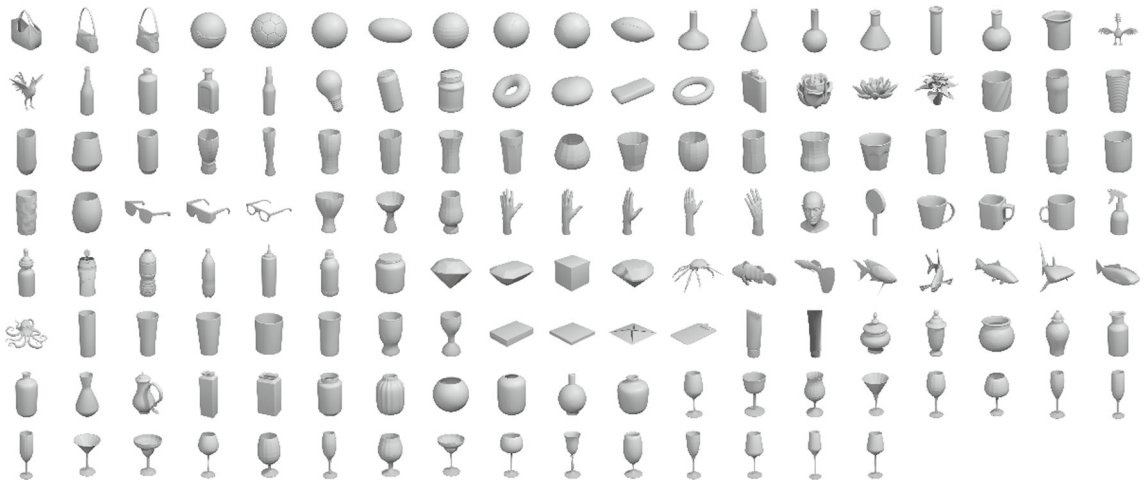

Fig. 3 A diverse set of object instances used in rendering Trans2k sequences

3.1.6 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.

3.1.7 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.

3.1.8 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 Trans2k Dataset Generation

Our preliminary study (Lukezic et al., 2022) reveals that most of the attribute levels described in Sect. 3.1 result in performance reduction and are thus kept as relevant in our final dataset. Two attribute levels including the lowest transparency level and zero rotation were identified as already well addressed by the opaque object training sets and are thus omitted from the dataset for better use of its capacity.

The following parameters are thus applied when rendering Trans2k. The GoT10k (Huang et al., 2019) 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.

Fig. 4 Targets in Trans2k are annotated by axis-aligned bounding boxes (first row) or by segmentation masks (second and third rows). The dataset also contains annotated distractors (third row)

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, e.g., see (Kristan et al., 2020). The ground truths for distractors are generated as well. Trans2k is thus the first dataset with per-frame distractor annotation to facilitate development of distractor-aware methods. Some qualitative examples of the generated Trans2k sequences are shown in Fig. 4.

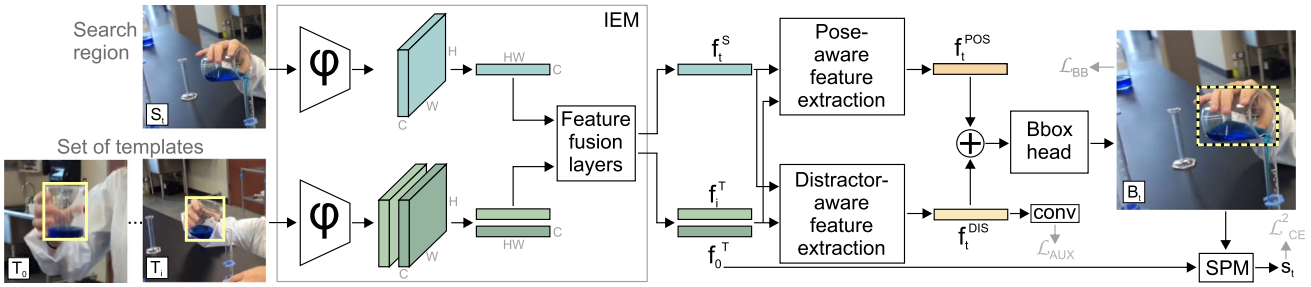

Fig. 5 Overview of the proposed DiTra architecture. Features are first extracted from the search region and from a set of templates by Image encoding module (IEM). These features are then processed by two parallel branches generating pose-aware and distractor-aware features ($\mathbf{f}_t^{POS}$ and $\mathbf{f}_t^{DIS}$). Both features are summed together and processed by a bounding box prediction head to predict the target bounding box $\mathbf{B}_t$. Localization confidence score s_t is estimated using the score prediction module (SPM)

4 A Distractor-Aware Tracker

We now introduce our second contribution – a distractor-aware transparent object tracker DiTra (Fig. 5). Given N_T target templates $\mathbf{T} \in \{\mathbf{T}_i\}_{i=1:N_\mathrm{T}}$ and their bounding boxes $\mathbf{B} \in \{\mathbf{B}_i\}_{i=1:N_\mathrm{T}}$, DiTra localizes the target in the search region $\mathbf{S}_t$ at the current time-step t by predicting a target bounding box $\mathbf{B}_t$.

The templates and the search region are of the same spatial size (i.e., $H_{im} \times W_{im} \times 3$) and are first encoded by the Image encoding module (Sect. 4.1). Next, the search region features are transformed into two sets of features by separate computational branches. The distractor-aware branch (Sect. 4.2) extracts features specialized only for discriminating between the target and similar objects. In parallel the pose-aware branch (Sect. 4.3) extracts features tuned for maximally accurate pose estimation.

The two types of extracted features are then fused using a Target localization head (Sect. 4.4) and regressed into the final estimated bounding box. Finally, a target presence confidence score is computed and the target template set is updated (Sect. 4.5). The following subsections detail each of the aforementioned computational blocks.

4.1 Image Encoding Module

The Image encoding module, IEM, (Fig. 5) first encodes the RGB templates and the search region by passing each through a backbone network $\varphi(\cdot)$, e.g., ResNet (He et al., 2016) and reduces the dimensionality of the output by a 1×1 convolution to C channels. This is followed by L transformer-based *Feature fusion layers*, adopted from Chen et al. (2021). The templates and the search region are thus mapped into template and search region features, i.e., $\mathbf{T} \rightarrow \mathbf{F}^T \in \{\mathbf{f}_i^T\}_{i=1:N_\mathrm{T}}$ and $\mathbf{S}_t \rightarrow \mathbf{f}_t^S$, each the size of $HW \times C$. The search region is then further encoded separately by two feature extraction branches described in the following two subsections.

4.2 Distractor-Aware Feature Extraction

The task of this branch is to extract features focusing on discriminating the target from similar objects in its vicinity. We exploit the fact that the distractors will appear in the larger neighborhood of the localized targets in the previous frames. Assuming frequent template updating, distractors can be captured by constructing sufficiently large templates. In our tracker we thus set the template size equal to the search region.

The distractor-aware feature extraction branch is implemented as a single multi-head attention block (Vaswani et al., 2017). The search region features $\mathbf{f}_i^S$ are used as queries, the template features $\mathbf{f}_i^T$ as keys, while the values are obtained by summing the template features and the template encodings $\mathbf{f}_i^E$. The latter enables the attention mechanism to distinguish the target from the potential distractors. The output of the multi-head attention block is followed by two linear transformations and ReLu activations to produce the final distractor-aware features $\mathbf{f}_t^{DIS} \in \mathbb{R}^{HW \times C}$.

The template encodings are computed as follows. For each template a two-channel binary mask $\hat{\mathbf{m}}_i \in \mathbb{R}^{HW \times 2}$ is constructed. The values in the first channel are set to one within the target bonding box and to zero elsewhere, while the second channel is the inverse of the first. The mask is then linearly transformed into the template encoding $\mathbf{f}_i^E$.

4.3 Pose-Aware Feature Extraction

The features specializing on discriminating the target from the distractors are extracted by the distractor-aware branch. This allows exploiting the entire capacity of the pose-aware branch solely for bounding box estimation, without compromising discrimination between similar objects, handled already by the distractor-aware branch.

To facilitate bounding box prediction learning, the templates are cropped to contain only the target appearance and not the potential distractors in their vicinity. Crop-

ping thus converts the backbone template features $\mathbf{f}_i^T$ into $\mathbf{f}_i^{*T} \in \mathbb{R}^{hw \times C}$. The features are then processed by a single multi-head attention block (Vaswani et al., 2017), where the search region $\mathbf{f}_t^S$ is used as a query and templates $\mathbf{f}_i^{*T}$ are used as keys and values. This is followed by two linear transformations and ReLu activation to produce the final target pose-aware features $\mathbf{f}_t^{POS} \in \mathbb{R}^{HW \times C}$.

4.4 Target Localization

The target location is predicted by considering both positional and discriminative features, which are summed into localization features, i.e., $\mathbf{f}_t^{LOC} = \mathbf{f}_t^{POS} + \mathbf{f}_t^{DIS}$, and a convolutional bounding box head (Yan et al., 2021) is applied to predict a single bounding box $\mathbf{B}_t \in \{x_t, y_t, w_t, h_t\}$, where x_t, y_t are coordinates of the top-left corner and w_t, h_t are width and height, respectively.

4.5 Updating the Template Set

Target localization requires multiple templates to represent the target appearance, which can significantly change during tracking. Thus the set of templates $\mathbf{T}$ is dynamically updated by adding a new template in the set every 10 frames. When the number of the templates exceeds N_{T}, the oldest one is removed, while keeping the initial template extracted in the first frame always within the set.

Since the distractors that lead to tracking failures are likely positioned in the vicinity of the target in the previous frame, the set is additionally updated by a template extracted at the previous frame. This template, however, is only used in the distractor-aware feature computation and not in the pose-aware feature extraction.

Tracking quality highly depends on ensuring targets are well localized in the templates. Updating the template set when the target is poorly localized, (i.e., during occlusion, momentary drift or target absence) can lead to tracking failure. DiTra thus estimates the confidence score $s_t \in (0, 1)$ after the target localization step and updates the template set only if the confidence is high enough, i.e., $s_t > 0.5$.

The confidence score s_t is estimated by the score prediction module (SPM) (Cui et al., 2022) as follows. The localized target appearance is encoded by a learnable token attended to the search region features $\mathbf{f}_t^S$ extracted from the estimated bounding box $\mathbf{B}_t$. Next, the token is attended to the features from the initial target template $\mathbf{f}_0^{*T}$ and regressed into s_t by a MLP with a sigmoid function. The reader is referred to Cui et al. (2022) for additional details.

4.6 Training Details

DiTra is trained in two phases. The first phase is dedicated to training robust and accurate target localization, while the second phase is dedicated to training the target presence prediction module SPM (Sect. 4.5).

4.6.1 Phase 1

The whole network (except the SPM) is trained for target localization by optimizing the following localization loss:

$$\mathcal{L}_{BB} = \lambda_{GIOU}\mathcal{L}_{GIOU}(\mathbf{B}_t, \mathbf{B}_{GT}) + \lambda_{L1}\mathcal{L}_1(\mathbf{B}_t, \mathbf{B}_{GT}), \quad (1)$$

where $\mathbf{B}_t$ and $\mathbf{B}_{GT}$ are predicted and ground-truth bounding boxes, respectively, $\mathcal{L}_{GIOU}$ represents generalized IoU loss (Rezatofighi et al., 2019) and $\mathcal{L}_1$ is the ℓ_1 loss. The losses are weighted by $\lambda_{GIOU} = 2$ and $\lambda_{L1} = 5$.

To guide the network towards learning distractor-aware features, we add an auxiliary loss $\mathcal{L}_{AUX}$ to the output of the distractor-aware feature extraction block. A 1×1 convolution, denoted as $\phi(\cdot)$ is first used to map the distractor-aware features $\mathbf{f}_t^{DIS}$ to a single channel segmentation mask. The auxiliary loss is then computed as a standard cross-entropy loss $\mathcal{L}_{CE}(\cdot)$, i.e.,

$$\mathcal{L}_{AUX} = \mathcal{L}_{CE}(\phi(\mathbf{f}_t^{DIS}), \mathbf{m}_t^{GT}), \quad (2)$$

where $\mathbf{m}_t^{GT}$ is obtained by setting pixels within the ground truth bounding box to one and zero outside. The first-stage final training loss is then composed of the individual losses, i.e., $\mathcal{L} = \mathcal{L}_{BB} + \mathcal{L}_{AUX}$. Combination of the two losses $\mathcal{L}_{BB}$ and $\mathcal{L}_{AUX}$ guides the network pose- and distractor-aware feature extraction branches to focus on their individual tasks.

4.6.2 Phase 2

Only the score prediction module (SPM) is trained in this phase by minimizing the cross-entropy loss

$$\mathcal{L}_{CE}^2 = y_t log(s_t) + (1 - y_t) log(1 - s_t), \quad (3)$$

where s_t is the predicted target presence confidence score and $y_t \in \{0, 1\}$ is the ground-truth label of the t-th training sample, i.e., whether the search region contains the target or not. Note that the superscript 2 in $\mathcal{L}_{CE}^2$ denotes training in the second phase.

5 Validation of Trans2k

This section reports empiric validation of the proposed Trans2k training dataset. A set of trackers (Sect. 5.1) is trained on Trans2k and evaluated against their versions trained on opaque object training sets (Sect. 5.2).

5.1 Trackers and Training Setup

State-of-the-art learning-based trackers that cover the major trends in modern architecture designs are selected: (i) two siamese trackers SiamRPN++ (Li et al., 2019) and SiamBAN (Chen et al., 2020), (ii) two deep correlation filter trackers ATOM (Danelljan et al., 2019a) and DiMP (Danelljan et al., 2019b), (iii) the recent state-of-the-art transparent object tracker TransATOM (Fan et al., 2021), and (iv) a transfomer-based STARK (Yan et al., 2021). These trackers localize the target by a bounding box. To account for the recent trend in localization by per-pixel segmentation (Kristan et al., 2020), we include (v) the recent state-of-the-art segmentation-based tracker D3S (Lukežič et al. 2020).

For training on Trans2k, 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 10,000 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 object training datasets GoT10k (Huang et al., 2019), LaSoT (Fan et al., 2019) and TrackingNet (Muller et al., 2018) into a single dataset, abbreviated as an *opaque object training dataset* (OTD). A training batch is then constructed by sampling from Trans2k and OTD with 5:3 ratio.

5.2 Results

The contribution of the new training dataset Trans2k is validated by measuring tracking performance on the recent transparent object tracking benchmark TOTB (Fan et al., 2021). Following the training regime described in Sect. 5.1 the selected trackers were re-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 the AUC and center error measures on success and precision plots. For more information on the evaluation protocol, please refer to Wu et al. (2015) and Fan et al. (2021).

The results are shown in Fig. 6. 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 (Fan et al., 2021), a transparent object tracking extension of ATOM (Danelljan et al., 2019a), 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.

Since Trans2k provides segmentation ground truths in addition to bounding boxes, it boosts the segmentation-based tracker D3S (Lukežič et al. 2020) as well. The version trained with Trans2k gains a remarkable 6% in performance.

Consistent with the observation on opaque object tracking benchmarks, the transformer-based tracker STARK achieves the best performance among existing trackers. Note that even without training with Trans2k, these tracker surpasses all (non-transformer) trackers. When trained with Trans2k, additional performance boost of 2.5% is observed.

6 Validation of DiTra

This section provides an experimental evaluation of the proposed distractor-aware transparent object tracker (DiTra). Implementation details are given in Sect. 6.1, while DiTra is evaluated on transparent and opaque object tracking in Sects. 6.2 and 6.3, respectively. Ablation study is reported in Sect. 6.4.

6.1 Implementation Details

6.1.1 Tracking Implementation Details

Features extracted from the fourth layer of the ResNet-50 (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) for object classification are used in DiTra Image encoding module. The backbone features are extracted from the image region resized to $H_{im} = W_{im} = 320$ pixels, while the spatial dimensions of the features are reduced 16-times, i.e., $H = W = 20$. The channel dimension C is set to 256.

$N_{\mathrm{T}} = 6$ templates are used in tracking. All attention blocks in DiTra contain 8 heads and the standard sine-based positional embeddings (Vaswani et al., 2017; Carion et al., 2020; Yan et al., 2021) are used on queries and keys. Tracking performance of the proposed tracker is not sensitive to the exact values of the parameters, thus we use the same values in all experiments.

6.1.2 Training Implementation Details

As described in Sect. 4.6, the training process is divided into two phases. In the *phase 1*, a search region is randomly sampled from a random training sequence and two templates are sampled within 200 frames from the same sequence. In the *phase 2* (i.e., training the score prediction module – SPM),

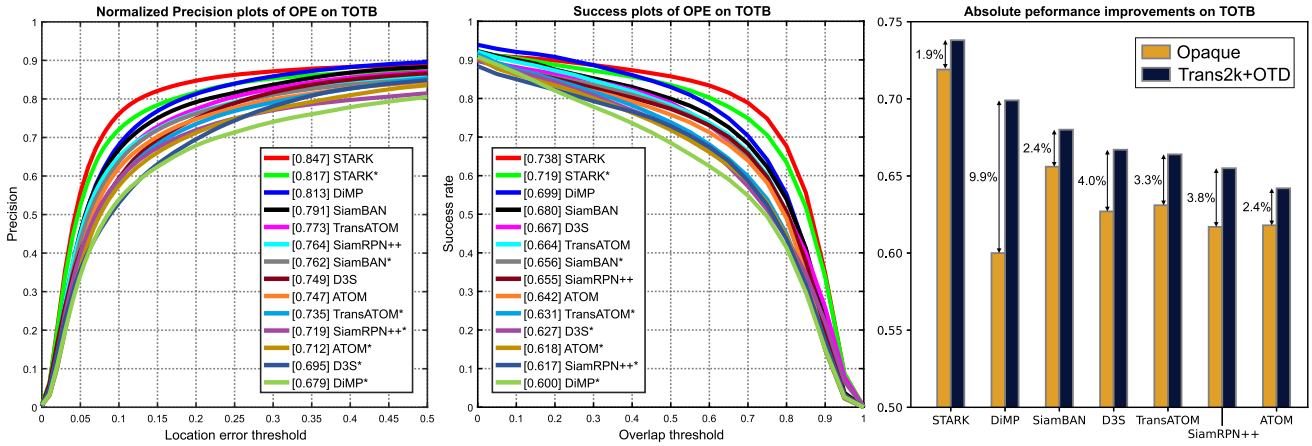

Fig. 6 Trackers evaluated on TOTB dataset shown in precision and success plots. Trackers trained on opaque datasets only are denoted by a star (*). The right graph shows absolute improvements in tracking performance measured by the AUC measure after training with the proposed Trans2k

the positive and negative training samples are sampled with equal probability. A positive sample is constructed by sampling a template and search region from the same sequence, while the negative sample is constructed by sampling them from different sequences.

In the phase 1, DiTra is trained for 300 epochs using ADAM optimizer (Kingma & Ba, 2015) with learning rate set to 10^{-4} decreasing by factor 10 after 250 epochs. Training takes approximately 4 days on two NVidia V100 with batch size 32 per-gpu. In phase 2, DiTra is trained for 40 epochs using ADAM optimizer (Kingma & Ba, 2015) with learning rate set to 10^{-4} decreasing by factor 10 after 30 epochs. Training takes approximately 8 h on two NVidia V100 with batch size 64 per-gpu.

6.2 Transparent Object Tracking

Transparent object tracking performance is evaluated on the recent TOTB benchmark (Fan et al., 2021). The following state-of-the-art trackers are considered for comparison: two siamese trackers SiamRPN++ (Li et al., 2019) and SiamBAN (Chen et al., 2020), three deep correlation filter trackers ATOM (Danelljan et al., 2019a), DiMP (Danelljan et al., 2019b) and KYS (Bhat et al., 2020), the recent state-of-the-art transparent object tracker TransATOM (Fan et al., 2021), two transfomer-based trackers STARK (Yan et al., 2021) and TOMP (Mayer et al., 2022) and a segmentation-based tracker D3S (Lukežič et al. 2020). The trackers are evaluated by the standard one-pass evaluation protocol (OPE) that quantifies the performance by the AUC score (Fan et al., 2021).

Results reported in Fig. 7 show that the proposed DiTra outperforms all trackers and sets new state-of-the-art on TOTB. In particular, it outperforms the second-best STARK and TOMP for approximately 5% in AUC. These results

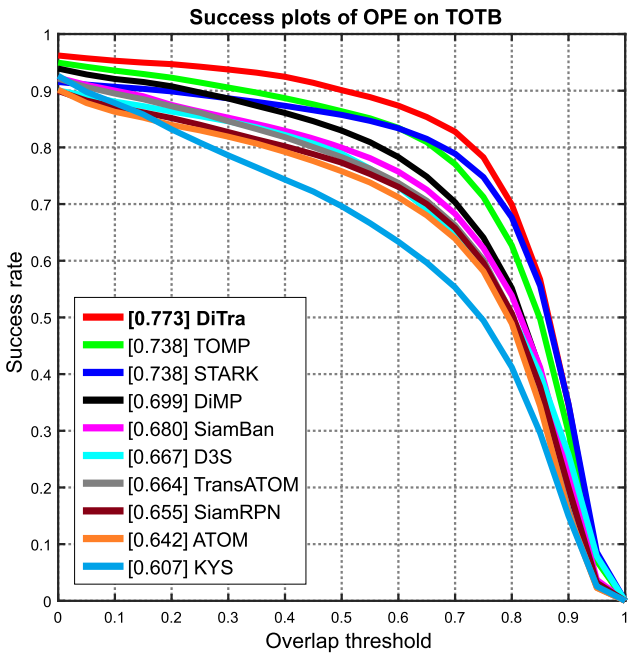

Fig. 7 Performance on transparent object tracking benchmark TOTB (Fan et al., 2021)

show that the disctractor-aware mechanism in the proposed DiTra successfully handles distractors in challenging scenarios and represents a strong baseline for the further research in tracking transparent objects. A qualitative comparison of DiTra with state-of-the-art trackers evaluated on TOTB is shown in Fig. 8.

6.3 Opaque Object Tracking

For evaluation completeness, DiTra is evaluated on opaque object tracking problems as well. The trackers are first evaluated on the challenging VOT2020 dataset (Kristan et al.,

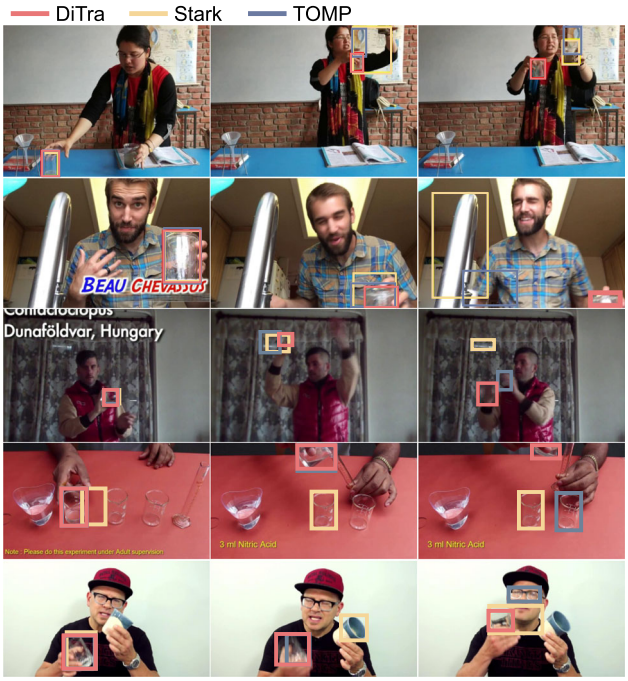

Fig. 8 Qualitative comparison of DiTra, Stark and TOMP on TOTB (Fan et al., 2021)

Table 1 Performance on the VOT2020 Kristan et al. (2020) opaque object tracking benchmark

Tracker	EAO	Accuracy	Robustness
DiTra	0.314	0.447	0.821
STARK	0.308	0.478	0.799
TOMP	0.297	0.453	0.789
DiMP	0.274	0.457	0.734
ATOM	0.271	0.462	0.734
SiamRPN++	0.255	0.424	0.730

Table 2 Performance on opaque tracking datasets GoT10k (Huang et al., 2019) and LaSoT (Fan et al., 2019)

Tracker	GoT10k	LaSoT
DiTra	76.1	66.0
MixFormer-1k	73.2	67.9
STARK	68.0	66.4
TOMP	67.0	67.6
KYS	63.6	55.4
DiMP	61.1	56.9
ATOM	55.6	51.5
SiamRPN++	51.7	49.6

2020), which is part of the annual VOT challenges (Kristan et al., 2016). Trackers are run on each sequence multiple times from different pre-defined starting points and let to track until the end of the sequence. Tracking performance is measured by two complementary measures: (i) accuracy, computed as the average overlap and (ii) robustness, which counts how often tracker fails to localize the target. Both, accuracy and robustness are combined in the primary measure, called expected average overlap (EAO).

DiTra achieves top results among the compared trackers, in particular it outperforms the second-best STARK by 2% EAO. Results show that DiTra particularly excells in robustness. This is due to the discriminative formulation, which allows to resolve challenging situations with multiple distractors (Table 1).

Next, we evaluate DiTra on GoT10k (Huang et al., 2019) test dataset, which is a large-scale high-diversity tracking dataset. It consists of approximately 10 thousand training sequences, while a set of 180 sequences is used for evaluating tracking performance. A tracker is initialized at the beginning and let to track to the end of the sequence. Tracking performance is measured by the area under the success-rate curve (AUC). Results in Table 2 show that DiTra outperforms the compared recent state-of-the-art tracker (Cui et al., 2022) by nearly 4%. This result demonstrates that DiTra achieves state-of-the-art results on opaque object tracking and also generalizes well across different short-term datasets.

Despite being a short-term tracker, we evaluate DiTra on the long-term tracking dataset LaSoT (Fan et al., 2019). In this dataset, the targets often disappear from the image, which emphasizes *long-term capabilities* of a tracker. The dataset contains a total of 1400 sequences with 70 object categories, where 280 sequences are used for evaluation and others are used for training. A tracker is initialized at the beginning and let to track to the end of the sequence. Tracking performance is measured by the area under the success-rate curve (AUC).

Results in Table 2 show that DiTra performs comparable to the top-performing MixFormer Cui et al. (2022), TOMP (Mayer et al., 2022) and STARK (Yan et al., 2021). Based on the results obtained on VOT, GoT10k and LaSoT datasets, we conclude that DiTra excels both in tracking of transparent objects as well as opaque objects, indicating the generality of the proposed distractor-aware formulation.

6.4 Ablation Study

Ablation study on the TOTB benchmark (Fan et al., 2021) is conducted for further insights. The following variations of DiTra are analyzed: (i) DiTra without fine-tuning on transparent objects (DiTra$^{\overline{TRS}}$), i.e., trained on opaque objects only; (ii) DiTra without the distractor-aware feature extraction branch (DiTra$^{\overline{DIS}}$); (iii) DiTra without the pose-aware feature extraction branch (DiTra$^{\overline{POS}}$); and (iv) DiTra without the most recent template in the distractor-aware feature

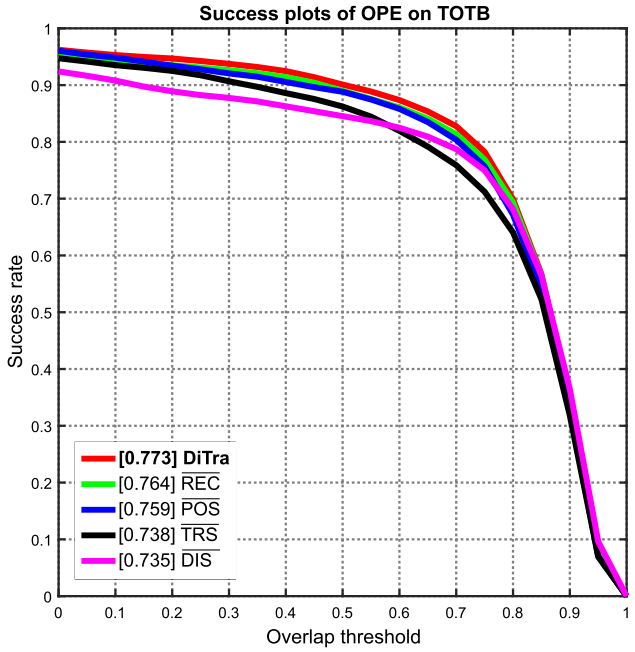

Fig. 9 Ablation study on TOTB. Removing: the most recent template ($\overline{REC}$), the pose-aware feature ($\overline{POS}$) and the distractor-aware features ($\overline{DIS}$). The ($\overline{TRS}$) denotes a version of DiTra without fine-tuning on transparent objects

extraction (DiTra$^{\overline{REC}}$). Note that the variants (ii) and (iii) are trained using the same training setup as the original DiTra, while the variant (iv) does not require re-training.

Results of the ablation study are presented in Fig. 9. Omitting the fine-tuning step on transparent objects (DiTra$^{\overline{TRS}}$) reduces the tracking performance by 4.5%. This confirms the contribution of the Trans2k dataset and shows the importance of including transparent objects in training the process.

Removing the distractor-aware feature extraction branch (DiTra$^{\overline{DIS}}$) causes a 5% performance drop, while removal of the pose-aware feature extraction branch (DiTra$^{\overline{POS}}$) reduces the performance by 2%. These results show the importance of splitting the tracking task into two separate branches. However, the distractor-aware features are more important for good tracking performance than the pose-aware features, since they prevent irreversible tracking failures.

Finally, removing the most recent template in the distractor-aware feature extraction (DiTra$^{\overline{REC}}$) reduces the tracking performance by approximately 1%. This demonstrates that the most recent template is not essential, but helps when target appearance is changing quickly in presence of multiple distractors.

Qualitative comparison is given in Fig. 10. Removing the distractor-aware feature extraction branch (DiTra$^{\overline{DIS}}$) reduces the tracking capability especially when multiple similar objects (distractors) are present in the same scene (third row). A version without the pose-aware feature extraction

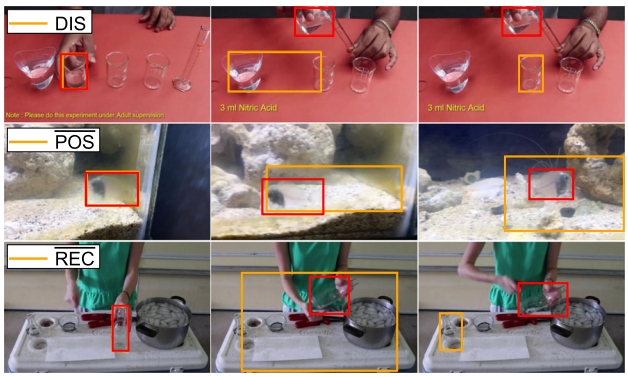

Fig. 10 The proposed DiTra (red bounding box) is compared to the versions: without distractor-aware features ($\overline{DIS}$), without pose-aware features ($\overline{POS}$) and without the most recent template ($\overline{REC}$). Please see text in Sect. 6.4 for discussion

branch (DiTra$^{\overline{POS}}$) fails to accurately localize the target in challenging scenarios (fourth row). Removal of the most recent template in the distractor-aware feature extraction branch (DiTra$^{\overline{REC}}$) causes tracking failure when the target appearance changes significantly in presence of distractors (fifth row).

To provide additional insights of the proposed tracker, we visualize the pose-aware and distractor-aware feature extraction attention maps in Fig. 11. Attention operation of the pose-aware feature extraction focuses on object shape, highlighting shapes of individual (multiple) objects, not only the target. On the other hand, attention of the distractor-aware feature extraction successfully suppresses distractors and only provides the object center. Combination of both, the pose-aware and distractor-aware feature extraction results in an accurate and robust tracker.

6.5 Failure Cases

An analysis of failure cases of the proposed tracker is presented in Fig. 12. We observe two major reasons causing DiTra to fail. First one is an extreme level of transparency, which results in a poorly visible target. Two examples showing such objects are shown in Fig. 12a and b. In these examples DiTra tends to focus on the background, which is visible through the target, instead of tracking it. Note that such examples are extremely difficult for humans as well. We believe that these failures could be addressed by specializing the feature extractor and forcing it to focus on the really fine visual details, specific for such objects.

Another group of failures are situations where occlusion appears together with distractors, shown on Fig. 12c. When the target gets occluded, the tracker localizes the object, which is visually the most similar to the target. If the target re-appears in the position outside of the search region, the tracker is not able to localize it and keeps tracking the

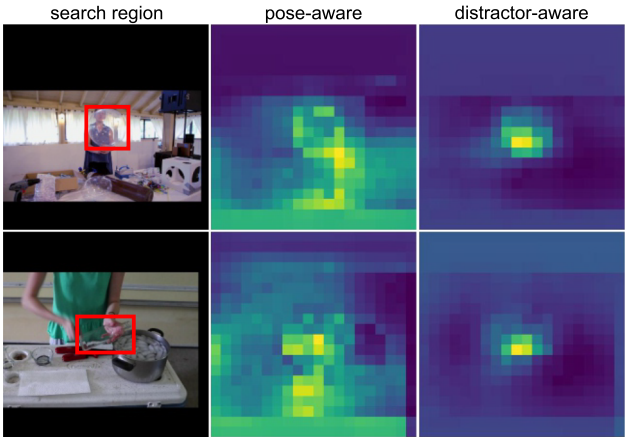

Fig. 11 Visualization of the attention maps of the pose-aware and distractor-aware feature extraction for the corresponding search regions

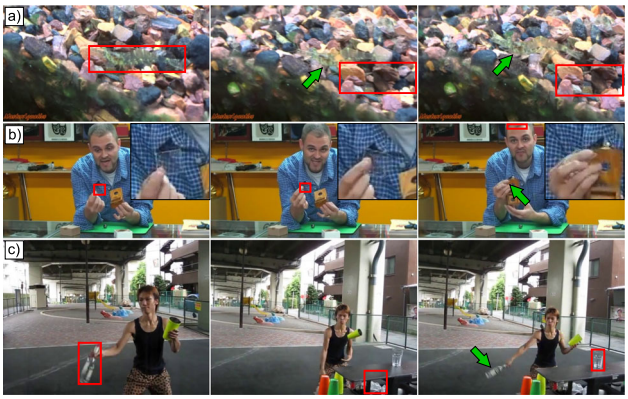

Fig. 12 Failure cases of DiTra. Two most frequent reasons for a failure are extreme transparency of the target (**a**) and (**b**) and combination of occlusion and distractors (**c**)

wrong target (distractor). A possible solution to such failures would be to incorporate long-term tracking components, e.g., image-wide re-detection mechanism or motion priors.

7 Conclusion

Two contributions to transparent object tracking were presented. The first contribution is the first transparent object tracking training dataset Trans2k, which exploits the fact that transparent objects can be sufficiently realistically rendered by modern renderers. Trans2k was validated on the recent transparent object tracking benchmark TOTB (Fan et al., 2021). 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.

The second contribution is a new distractor-aware transparent object tracker (DiTra). DiTra addresses tracking in presence of multiple visually similar objects (distractors), which are common in transparent object tracking scenes. The proposed tracker achieves state-of-the-art performance on the transparent object tracking task and is competitive in opaque object tracking. Trans2k, its rendering engine and DiTra will be publicly released.

While an excellent test set Fan et al. (2021) was recently introduced for transparent object tracking, the second 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. We envision that the Trans2k generation engine will allow innovative 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.

Based on the excellent generalizaton to opaque object tracking, we hope that the proposed distractor-aware formulation in DiTra will ignite exploration of similar modules dedicated for opaque object tracking, thus leading to further advancements in both tracking sub-domains.

Acknowledgements This work was supported by Slovenian research agency program P2-0214 and projects Z2-4459 and J2-2506.

References

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. S. (2016). Fully-convolutional siamese networks for object tracking. In *European conference on computer vision workshops* (pp. 850–865).

Bhat, G., Danelljan, M., Gool, L.V., & Timofte, R. (2020). Know your surroundings: Exploiting scene information for object tracking. In *Proceedings of the European conference on computer vision*.

Bhat, G., Johnander, J., Danelljan, M., Shahbaz Khan, F., & Felsberg, M. (2018). Unveiling the power of deep tracking. In *Proceedings of the European conference on computer vision* (pp. 493–509).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Proceedings of the European conference on computer vision* (pp 213–229).

Cehovin Zajc, L., Lukezic, A., Leonardis, A., & Kristan, M. (2017). Beyond standard benchmarks: Parameterizing performance evaluation in visual object tracking. In *International conference on computer vision*.

Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., & Lu, H. (2021). Transformer tracking. In *Computer vision and pattern recognition* (pp. 8126–8135).

Chen, Z., Zhong, B., Li, G., Zhang, S., & Ji, R. (2020). Siamese box adaptive network for visual tracking. In *Computer vision and pattern recognition* (pp. 6668–6677).

Cui, Y., Jiang, C., Wang, L., & Wu, G. (2022). Mixformer: End-to-end tracking with iterative mixed attention. In *Computer vision and pattern recognition* (pp. 13,608–13,618).

Danelljan, M., Bhat, G., Khan, F. S., & Felsberg, M. (2019a). ATOM: Accurate tracking by overlap maximization. In *Computer vision and pattern recognition*.

Danelljan, M., Bhat, G., Shahbaz Khan, F., & Felsberg, M. (2017). Eco: Efficient convolution operators for tracking. In *Computer vision and pattern recognition* (pp. 6638–6646).

Danelljan, M., Bhat, G., Van Gool, L., & Timofte, R. (2019b). Learning discriminative model prediction for tracking. In *International conference on computer vision* (pp. 6181–6190).

Danelljan, M., Häger G, Khan, F. S., & Felsberg, M. (2015). Convolutional features for correlation filter based visual tracking. In *2015 IEEE international conference on computer vision workshop (ICCVW)* (pp. 621–629).

Danelljan, M., Robinson, A., Khan, F. S., & Felsberg, M. (2016). Beyond correlation filters: learning continuous convolution operators for visual tracking. In *Proceedings of the European conference on computer vision*, (pp. 472–488).

Denninger, M., Sundermeyer, M., Winkelbauer, D., Olefir, D., Hodan, T., Zidan, Y., Elbadrawy, M., Knauer, M., Katam, H., & Lodhi, A. (2020). Blenderproc: Reducing the reality gap with photorealistic rendering. In *International conference on robotics: Science and systems RSS 2020*.

Du, Y., Liu, Z., Basevi, H., Leonardis, A., Freeman, B., Tenenbaum, J., & Wu, J. (2018). Learning to exploit stability for 3d scene parsing. In *Advances in neural information processing systems*.

Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., & Ling, H. (2019). Lasot: A high-quality benchmark for large-scale single object tracking. In *Computer vision and pattern recognition*.

Fan, H., Miththanthaya, H. A., Harshit, Rajan, S. R., Liu, X., Zou, Z., Lin, Y., & Ling, H. (2021). Transparent object tracking benchmark. In International conference on computer vision (pp. 10,734–10,743).

Fritz, M., Bradski, G., Karayev, S., Darrell, T., & Black, M. (2009). An additive latent feature model for transparent object recognition. *Advances in neural information processing systems* 22.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Computer vision and pattern recognition* (pp. 770–778).

Hodan, T., Barath, D., & Matas, J. (2020a). Epos: Estimating 6d pose of objects with symmetries. In *Computer vision and pattern recognition*.

Hodan, T., Barath, D., & Matas, J. (2020b). Epos: Estimating 6d pose of objects with symmetries. In *Computer vision and pattern recognition*, (pp. 11,703–11,712).

Hodan, T., Michel, F., Brachmann, E., Kehl, W., GlentBuch, A., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., Sahin, C., Manhardt, F., Tombari, F., Kim, T.-K., Matas, J., & Rother, C. (2018). Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision*.

Hodaň, T., Sundermeyer, M., Drost, B., Labbé, Y., Brachmann, E., Michel, F., Rother, C., & Matas, J. (2020). Bop challenge 2020 on 6d object localization. In *Proceedings of the European conference on computer vision* (pp. 577–594).

Huang, L., Zhao, X., & Huang, K. (2019). GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kalra, A., Taamazyan, V., Rao, S. K., Venkataraman, K., Raskar, R., & Kadambi, A. (2020). Deep polarization cues for transparent object segmentation. In *Computer vision and pattern recognition* (pp. 8602–8611).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations (ICLR)*.

Klank, U., Carton, D., & Beetz, M. (2011). Transparent object detection and reconstruction on a mobile platform. In *International conference on robotics and automation*. IEEE (pp. 5971–5978).

Krahenbuhl, P. (2018). Free supervision from video games. In *Computer vision and pattern recognition* (pp. 2955–2964).

Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kämäräinen, J.-K., Danelljan, M., Zajc, L., Lukežič, A., Drbohlav, O. et al. (2020). The eighth visual object tracking VOT2020 challenge results. In *European conference on computer vision workshops*.

Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kämäräinen, J.-K., Chang, H. J., Danelljan, M., Cehovin, L., Lukežič, A., Drbohlav, O., Käpylä, J., Häger, G., Yan, S., Yang, J., Zhang, Z., & Fernández, G. (2021). The ninth visual object tracking vot2021 challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV) Workshops* (pp. 2711–2738).

Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., & Cehovin, L. (2016). A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Čehovin, L., Nebehay, G., Vojir, T., Fernandez, G., et al. (2014). The visual object tracking vot2014 challenge results. In *Proceedings of the European conference on computer vision* (pp. 191–217).

Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Čehovin, L., Nebehay, G., Fernandez, G., Vojir, T., et al. (2013). The visual object tracking vot2013 challenge results. In *Visual object tracking challenge VOT2013, In conjunction with ICCV2013* (pp. 98–111).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., & Yan, J. (2019). SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *Computer vision and pattern recognition*.

Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018). High performance visual tracking with siamese region proposal network. In *Computer vision and pattern recognition*.

Liu, X. (2021). Deep correlation filters for robust visual tracking. In *2021 IEEE international conference on multimedia and expo (ICME)* (pp. 1–6).

Lukežič, A., Matas, J., & Kristan, M. (2020). D3S—A discriminative single shot segmentation tracker. In *Computer vision and pattern recognition* (pp. 7133–7142).

Lukezic, A., Trojer, Z., Matas, J., & Kristan, M. (2022). Trans2k: Unlocking the power of deep models for transparent object tracking. In *Proceedings of the British machine vision conference (BMVC)*.

Maeno, K., Nagahara, H., Shimada, A., & Taniguchi, R. (2013). Light field distortion feature for transparent object recognition. In *Computer vision and pattern recognition* (pp. 2786–2793).

Mayer, C., Danelljan, M., Bhat, G., Paul, M., Pani Paudel, D., Yu, F., & Van Gool, L. (2022). Transforming model prediction for tracking. In *Computer vision and pattern recognition*.

Metzger, K. A., Mortimer, P., & Wuensche, H. J. (2021). A fine-grained dataset and its efficient semantic segmentation for unstructured driving scenarios. In *Proceedings of the international conference on pattern recognition* (pp. 7892–7899).

Mueller, M., Smith, N., & Ghanem, B. (2016). A benchmark and simulator for uav tracking. In *Proceedings of the European conference on computer vision* (pp. 445–461).

Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., & Ghanem, B. (2018). TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision*.

Qi, M., Wang, Y., Qin, J., & Li, A. (2019). Ke-gan: Knowledge embedded generative adversarial networks for semi-supervised scene parsing. In *Computer vision and pattern recognition*.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Computer vision and pattern recognition* (pp. 658–666).

Richter, S. R., Hayder, Z., & Koltun, V. (2017). Playing for benchmarks. In *International conference on computer vision* (pp. 2213–2222).

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Computer vision and pattern recognition* (pp. 3234–3243).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.

Sajjan, S., Moore, M., Pan, M., Nagaraja, G., Lee, J., Zeng, A., & Song, S. (2020). Clear grasp: 3d shape estimation of transparent objects for manipulation. In *International conference on robotics and automation* (pp. 3634–3642). IEEE.

Saleh, F. S., Aliakbarian, M. S., Salzmann, M., Petersson, L., & Alvarez, J. M. (2018). Effective use of synthetic data for urban scene semantic segmentation. In *Proceedings of the European conference on computer vision* (pp. 84–100).

Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., & Torr, P. H. S. (2017). End-to-end representation learning for correlation filter based tracking. In Computer vision and pattern recognition (pp. 2805–2813).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*.

Wang, N., Zhou, W., Wang, J., & Li, H. (2021). Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Computer vision and pattern recognition* (pp. 1571–1580).

Wrenninge, M., & Unger, J. (2018). Synscapes: A photorealistic synthetic dataset for street scene parsing. arXiv preprint arXiv:1810.08705

Wu, Y., Lim, J., & Yang, M. H. (2013). Online object tracking: A benchmark. In *Computer Vision and Pattern Recognition* (pp. 2411–2418).

Wu, Y., Lim, J., & Yang, M. H. (2015). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(9), 1834–1848.

Xie, E., Wang, W., Wang, W., Ding, M., Shen, C., & Luo, P. (2020). Segmenting transparent objects in the wild. In *Proceedings of the European conference on computer vision* (pp. 696–711).

Xu, Y., Nagahara, H., Shimada, A., & Taniguchi, R. (2015). Transcut: Transparent object segmentation from a light-field image. In: *International conference on computer vision* (pp. 3442–3450).

Yan, B., Peng, H., Fu, J., Wang, D., & Lu, H. (2021). Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 10,448–10,457).

Zhang, J., Chen, Z., Huang, J., Lin, L., & Zhang, D. (2019). Few-shot structured domain adaptation for virtual-to-real scene parsing. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV) workshops*.