LEARNING TO TRACK FROM PRIVILEGED TARGET APPEARANCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Target templates define what a visual tracker searches for, yet the templates available at inference trade off *localization certainty* with *appearance freshness*: the initial ground-truth template is exact but becomes stale, whereas recent templates better reflect the current appearance but are cropped from uncertain predictions. We quantify this bottleneck with a non-deployable oracle that supplies an exact current-frame target crop, improving AUC on LaSOT by 15.2 percentage points. This gap reveals a training-only opportunity: frame-level ground truths provide exact current- and future-frame target crops, although such crops are unavailable at deployment. We introduce Privileged Appearance Transfer for Tracking (**PATT**), a teacher-student training framework that transfers these privileged appearances to a deployable tracker through multi-level representation prediction. The privileged teacher observes exact target crops from past, current, and future frames, whereas the student receives only past-frame templates and learns to predict the teacher's search representations. To avoid transferring unreliable teacher signals, PATT weights this transfer by the teacher's relative localization advantage over the student and its absolute localization accuracy. After training, the teacher, latent predictor, reliability weights, and privileged crops are removed, leaving standard student-only inference. Across seven benchmarks at two model scales, PATT achieves consistent gains under both long- and short-term tracking protocols. Code and models will be made publicly available.

1 INTRODUCTION

Tracking is fundamentally a problem of maintaining object correspondence through time. Humans can preserve object identity across motion even in the presence of visually identical distractors (Pylyshyn & Storm, 1988), while object-file theory explains this ability as the integration of successive visual states into temporary object representations (Kahneman et al., 1992). Visual object tracking poses a computational analogue: given a target in the first frame, localize the same object as its position and appearance evolve (Lucas & Kanade, 1981; Wu et al., 2013). Modern trackers typically encode target identity through image templates, which specify *what* to track and guide localization in each new search region. This template-search paradigm underlies Siamese trackers (Bertinetto et al., 2016; Li et al., 2018), Transformer trackers (Chen et al., 2021; Yan et al., 2021; Cui et al., 2022; Ye et al., 2022), autoregressive trackers (Chen et al., 2023; Wei et al., 2023), and unified tracking models (Hong et al., 2024; Chen et al., 2025). Despite their architectural diversity, these methods share a common dependence on whether the available templates remain discriminative for the target's current state.

This dependence is complicated by the fact that the target representation needed for correspondence changes over time. The initial template is extracted from the given ground-truth bounding box and is therefore *exact* and *localization-certain*, but its appearance may become obsolete under deformation, rotation, illumination change, scale variation, or occlusion. A template acquired from a recent frame can better reflect the target's current appearance, yet its crop must be derived from the tracker's own predicted box. When this prediction contains substantial background or drifts toward a distractor, template updates can reinforce the resulting error. We refer to this limitation as the **inference-time template acquisition bottleneck**: at deployment, no reliably available template jointly provides localization certainty and appearance freshness.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

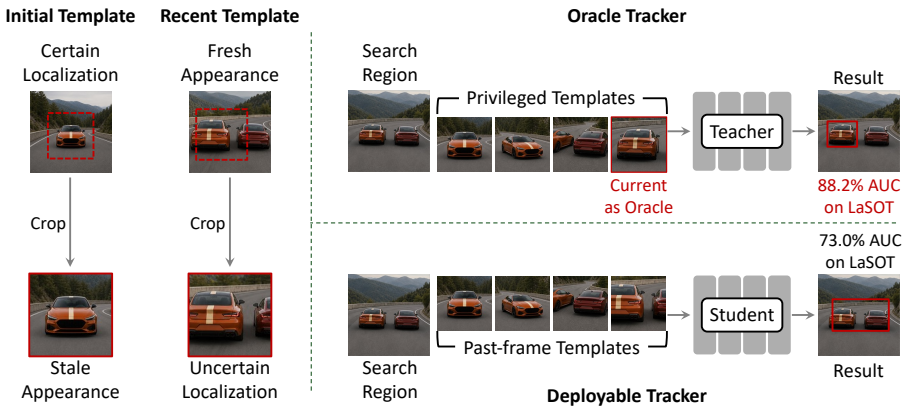


Figure 1: **Oracle study on LaSOT.** Standard trackers face an inference-time template-acquisition bottleneck: the initial template is accurately localized but becomes stale, whereas recent templates are fresher but cropped from uncertain predictions. A non-deployable oracle removes this bottleneck by using the ground-truth bounding box to supply an exact current-frame target crop. On LaSOT, AUC increases from 73.0% to 88.2%, a substantial gain of 15.2 percentage points.

Existing trackers mitigate this bottleneck by extracting additional target evidence from the inference history. Online discriminative learning (Bhat et al., 2019; Mayer et al., 2022), confidence-based template updates (Yan et al., 2021; Cui et al., 2022), temporal token propagation (Zheng et al., 2024; Feng et al., 2025; Huang et al., 2026; Chen et al., 2026), and autoregressive state modeling (Chen et al., 2023; Wei et al., 2023; Bai et al., 2024) all seek to preserve reliable target information as a video unfolds. However, they cannot remove the underlying information constraint: apart from the initial template, every newly acquired target crop is derived from a model prediction and may therefore inherit its localization errors. Without ground-truth supervision, an inference-time update rule cannot verify whether a new crop precisely depicts the target.

Training exposes precisely the information missing at inference. Frame-level ground-truth bounding boxes provide exact target crops not only before the search frame, but also from the current and future frames. To quantify the value of this information, we evaluate a *non-deployable* oracle that gives the tracker an exact current-frame target crop. As shown in Figure 1, this intervention raises the area under the success curve (AUC) on LaSOT (Fan et al., 2019) from 73.0% to 88.2%, a substantial gain of 15.2 percentage points. This oracle gap suggests that exact, up-to-date target appearances remain a major performance-limiting factor. Yet standard tracking objectives use the current ground-truth bounding box primarily for localization supervision and past target crops as causal templates; they barely exploit exact current and future target appearances as training-only inputs.

Motivated by this oracle gap, we introduce **P**rivileged **A**ppearance **T**ransfer for **T**racking (**PATT**), a teacher-student training framework that turns training-only target appearances into supervision for a deployable tracker. The student uses four past-frame target templates, whereas the privileged teacher replaces two of them with exact target crops from the current and a future frame. Apart from this difference, the two branches share the same architecture, backbone capacity, and four-slot template budget, making privileged target appearances the teacher's principal information advantage.

Rather than imitate the teacher's final box predictions, for which ground-truth bounding boxes already provide direct supervision, PATT transfers privileged information through multi-level search representations. A latent predictor maps the student's multi-level search representations to their counterparts produced by the privileged teacher, which is maintained as an exponential moving average (EMA) of the student to provide stable targets. However, even with privileged inputs, some teacher predictions may still be *unreliable*. PATT therefore weights the transfer using two complementary criteria: the teacher's relative localization advantage over the student and its absolute localization accuracy. At deployment, the teacher, latent predictor, reliability weights, and privileged crops are discarded, leaving the standard deployable tracker unchanged.

Experiments across seven benchmarks demonstrate that PATT achieves consistent gains across Base and Large model scales, long- and short-term tracking protocols, and datasets outside the training distribution. In particular, PATT-L256 achieves 76.7% AUC on LaSOT, 56.7% AUC on $\text{LaSOT}_{\text{ext}}$ (Fan et al., 2021), and 64.8% AUC on TNL2K (Wang et al., 2021b). These results show

that privileged target appearances can improve the representations of a deployable tracker without additional inference components.

In summary, our contributions are as follows:

- We identify the inference-time template acquisition bottleneck in visual tracking and quantify its impact with a non-deployable oracle that improves LaSOT AUC by 15.2 percentage points.
- We introduce PATT, a teacher-student framework that transfers privileged target appearances to a deployable tracker through multi-level representation prediction and reliability weighting.
- We demonstrate consistent improvements across seven benchmarks, two model scales, and multiple evaluation regimes, showing the effectiveness of training-only target appearances.

2 Related Work

In this section, we position PATT relative to two lines of work: template-based visual tracking and learning with privileged information. The former exposes the causal constraint at deployment, whereas the latter frames the use of training-only target crops as privileged supervision for a deployable tracker.

2.1 Template-based Visual Tracking

Template conditioning is a central paradigm in visual tracking. Siamese trackers instantiate this idea by comparing a target template with a search region, as in SiamFC (Bertinetto et al., 2016), and by adding region proposals, as in SiamRPN (Li et al., 2018). Transformer trackers further strengthen template-search interaction through explicit feature fusion, as in TransT (Chen et al., 2021), STARK (Yan et al., 2021), and MixFormer (Cui et al., 2022), or through joint token processing in a unified encoder, as in OSTrack (Ye et al., 2022). Recent work extends this paradigm with larger pretrained backbones (Lin et al., 2024), autoregressive target prediction (Chen et al., 2023; Wei et al., 2023; Bai et al., 2024), and architectures that unify tracking with other forms of target specification (Hong et al., 2024; Chen et al., 2025).

A complementary line of research concerns how target evidence can be maintained after initialization. Online discriminative methods such as ATOM (Danelljan et al., 2019), DiMP (Bhat et al., 2019), and ToMP (Mayer et al., 2022) update target models using newly collected positive and negative samples. Template-update methods such as STARK (Yan et al., 2021) and MixFormer (Cui et al., 2022) refresh stored templates from confident predictions, while KeepTrack (Mayer et al., 2021) associates candidate states across frames to reduce ambiguity. Other approaches, including ODTrack (Zheng et al., 2024), autoregressive trackers (Chen et al., 2023; Wei et al., 2023), and RELO (Chen et al., 2026) condition localization on previous predictions or latent target states. These mechanisms improve temporal adaptation, but they remain constrained by causal information available at deployment: except for the initial target crop, newly acquired target evidence is derived from model predictions and is therefore susceptible to localization error. PATT instead studies whether a deployable tracker can benefit from current- and future-frame target crops extracted from ground-truth bounding boxes during training, without changing its inference-time inputs or computation.

2.2 Learning with Privileged Information

Learning with privileged information formalizes settings in which auxiliary variables are available during training but absent at deployment (Vapnik & Vashist, 2009). The central challenge is to transfer this training-only information to a model that relies only on deployable inputs. Knowledge distillation provides a natural mechanism for such transfer: a student learns from targets produced by a teacher (Hinton et al., 2015). Generalized distillation connects the two ideas by using privileged variables to construct an informative teacher whose predictions supervise a deployable student (Lopez-Paz et al., 2016).

Teacher-student paradigms differ in what gives the teacher an advantage and what signal is transferred to the student. In conventional distillation, the teacher's advantage typically comes from greater model capacity or an ensemble (Hinton et al., 2015). In privileged-information settings, the advantage instead comes from training-only annotations or modalities, such as semantic object at-

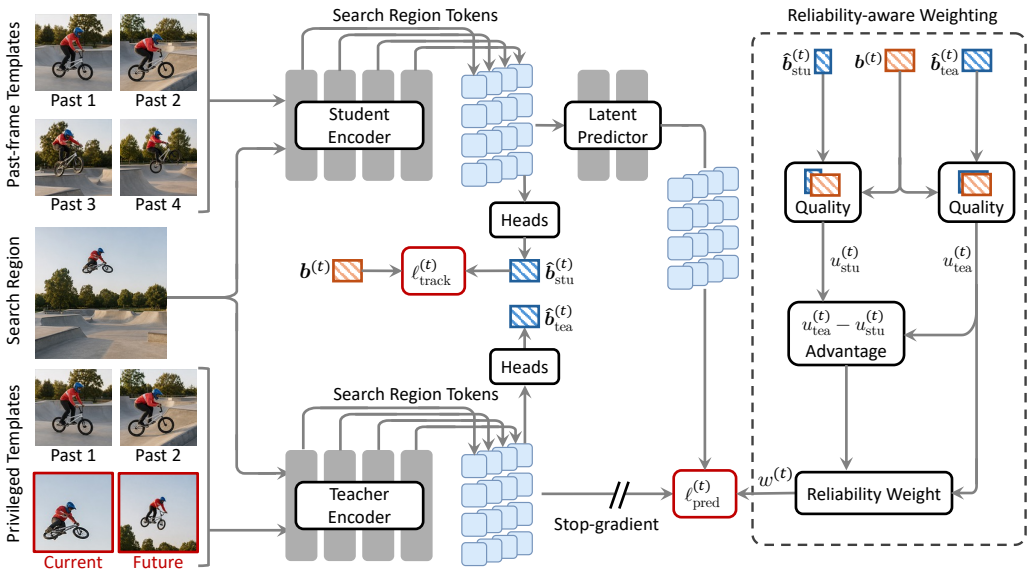


Figure 2: **Overview of PATT.** The student and privileged teacher process the same search crop with the same tracker architecture, model capacity, and four-slot template budget. The student uses only past-frame templates, whereas the teacher replaces two template slots with exact current- and future-frame target crops extracted from ground-truth bounding boxes. A latent predictor maps the student's multi-level search representations to the teacher's stop-gradient representations, defining the representation-prediction loss. A reliability weight modulates this loss by combining the teacher's relative localization advantage over the student and its absolute localization quality. At deployment, the teacher, predictor, reliability computation, and privileged crops are discarded, leaving only the student tracker.

tributes, bounding boxes (Motiian et al., 2016), or depth (Hoffman et al., 2016). The transferred signal may be an output distribution (Hinton et al., 2015) or an intermediate representation (Romero et al., 2015). A separate design choice is teacher stability. Mean Teacher (Tarvainen & Valpola, 2017) maintains the teacher as an EMA of the student to provide stable consistency targets, but parameter averaging alone does not create an information advantage.

Teacher-student transfer has also been explored in Siamese tracking (Shen et al., 2022). PATT differs from these settings by keeping the teacher and student at the same architecture and capacity: the teacher's advantage comes solely from exact current- and future-frame target crops available during training, and the transferred signal is its intermediate search representations.

3 Privileged Appearance Transfer for Tracking

In this section, we present PATT, a training framework that transfers privileged target-appearance information to a deployable tracker, while leaving its inference procedure unchanged. As shown in Figure 2, the teacher shares the student's tracker architecture and template budget, but replaces two past-frame templates with current- and future-frame target crops extracted from ground-truth bounding boxes. A latent predictor maps the student's multi-level search representations to the corresponding teacher representations, and a sample-wise reliability weight modulates this transfer. After training, all privileged inputs and auxiliary modules are discarded, leaving only the standard student tracker.

3.1 Template-Based Tracking Formulation

Student template construction. We first formulate the causal template input used by the student tracker. We represent a T-frame training video $\mathbf{I}$ and its ground-truth bounding boxes $\mathbf{B}$ as temporally ordered sequences:

$$\mathbf{I} = \left(\mathbf{I}^{(1)}, \ldots, \mathbf{I}^{(T)}\right), \qquad \mathbf{B} = \left(\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(T)}\right). \tag{1}$$

Let $\mathrm{crop}(\cdot,\cdot)$ denote the template-extraction operator of the underlying tracker. The target template extracted from frame i is

$$\boldsymbol{x}^{(i)} = \mathrm{crop}\left(\mathbf{I}^{(i)}, \boldsymbol{b}^{(i)}\right). \tag{2}$$

For search frame t, let r_k index the past frame used to extract the k-th student template, where $1 \le r_1 < \cdots < r_K < t$ and $K \ge 2$. The student's ordered template input is

$$\mathbf{X}_{\mathrm{stu}}^{(t)} = \left(\boldsymbol{x}^{(r_1)}, \ldots, \boldsymbol{x}^{(r_K)}\right). \tag{3}$$

During training, these past-frame templates are extracted using ground-truth bounding boxes. At inference, however, ground truth is available only for the initial template; any later template must be cropped from a predicted target box.

Student forward pass. Let $\boldsymbol{s}^{(t)}$ denote the search crop sampled from $\mathbf{I}^{(t)}$ around the ground-truth box $\boldsymbol{b}^{(t)}$. We use $\mathcal{L}$ for the set of encoder layers selected for representation transfer, with $L = |\mathcal{L}|$, and arrange their outputs in increasing order of encoder depth when stacked. Given the student templates and search crop, the tracker predicts a target box and produces search representations at the selected layers:

$$\left(\left\{\mathbf{Z}_{\mathrm{stu},l}^{(t)}\right\}_{l\in\mathcal{L}}, \hat{\boldsymbol{b}}_{\mathrm{stu}}^{(t)}\right) = \boldsymbol{f}_{\boldsymbol{\theta}}\left(\mathbf{X}_{\mathrm{stu}}^{(t)}, \boldsymbol{s}^{(t)}\right), \tag{4}$$

where $\mathbf{Z}_{\mathrm{stu},l}^{(t)} \in \mathbb{R}^{N\times D}$ contains the D-dimensional representations of the N student search tokens at layer l.

Standard tracking objective. The student parameters $\boldsymbol{\theta}$ are optimized using the tracker's standard localization objective:

$$\ell_{\mathrm{track}}^{(t)}(\boldsymbol{\theta}) = \ell_{\mathrm{loc}}\left(\boldsymbol{b}^{(t)}, \hat{\boldsymbol{b}}_{\mathrm{stu}}^{(t)}\right), \tag{5}$$

where the dependence on $\boldsymbol{\theta}$ follows from the forward pass in Equation (4). Additional head outputs and tracker-specific targets are left implicit. This objective provides direct supervision for localization, while the privileged teacher introduced in the next subsection supplies an additional representation-level signal at intermediate encoder layers.

3.2 Multi-Level Representation Prediction

We now construct a privileged teacher that accesses training-only information while keeping the tracker architecture and template budget fixed. The teacher provides multi-level search representations as prediction targets for the student, and its box prediction is used only to measure the reliability of this transfer.

Matched privileged templates. The oracle study in Figure 1 demonstrates the potential value of privileged target information: supplying the tracker with an exact current-frame target crop raises LaSOT AUC from 73.0% to 88.2%. This 15.2-point gain motivates using ground-truth bounding boxes during training to construct privileged target crops that are unavailable to a deployable tracker.

We instantiate this privilege with two teacher-only templates. For search frame t, the current-frame crop $\boldsymbol{x}^{(t)}$ depicts the target in the same frame as the search crop $\boldsymbol{s}^{(t)}$. A future-frame crop $\boldsymbol{x}^{(t')}$, with $t < t' \le T$, provides a later target appearance that may capture subsequent variation. Neither crop is a valid deployment-time input for localizing the target at frame t: the current crop requires the unknown target box at frame t, and the future crop additionally requires access to a later frame.

Rather than increasing the template budget, we form the teacher input by replacing the student's final two past-frame templates in Equation (3) with the privileged current- and future-frame crops:

$$\mathbf{X}_{\mathrm{tea}}^{(t)} = \left(\boldsymbol{x}^{(r_1)}, \ldots, \boldsymbol{x}^{(r_{K-2})}, \boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t')}\right). \tag{6}$$

The student and teacher otherwise share the same search crop, tracker architecture, model capacity, and template budget.

Teacher representation prediction. Given the privileged templates and the shared search crop, the teacher produces search representations at the selected encoder layers and predicts a target box:

$$\left(\left\{\mathbf{Z}_{\text{tea},l}^{(t)}\right\}_{l\in\mathcal{L}}, \hat{\boldsymbol{b}}_{\text{tea}}^{(t)}\right) = \boldsymbol{f}_{\bar{\boldsymbol{\theta}}}\left(\mathbf{X}_{\text{tea}}^{(t)}, \boldsymbol{s}^{(t)}\right), \tag{7}$$

where $\bar{\boldsymbol{\theta}}$ denotes the teacher parameters.

We transfer the teacher's intermediate search representations rather than imitating its predicted box. This is because box-level imitation would add another localization target on top of the direct ground-truth supervision in Equation (5), while exposing only the teacher's final localization outcome. Intermediate representations instead capture the token-level evidence shaped by the privileged current- and future-frame templates, providing complementary supervision for the student's encoder.

Following multi-level feature prediction (Mur-Labadia et al., 2026), we apply LayerNorm separately to the outputs of the L selected encoder layers in each branch, and concatenate them from shallow to deep along the feature dimension. This yields student and teacher representations $\mathbf{Z}_{\text{stu}}^{(t)}, \mathbf{Z}_{\text{tea}}^{(t)} \in \mathbb{R}^{N\times(D\times L)}$. Because privileged templates change how the teacher organizes search evidence, we do not force the student representations $\mathbf{Z}_{\text{stu}}^{(t)}$ to match the teacher representations $\mathbf{Z}_{\text{tea}}^{(t)}$ directly. Instead, we learn a Transformer predictor $\boldsymbol{g}_{\boldsymbol{\phi}} : \mathbb{R}^{N\times(D\times L)} \mapsto \mathbb{R}^{N\times(D\times L)}$ that maps $\mathbf{Z}_{\text{stu}}^{(t)}$ to $\mathbf{Z}_{\text{tea}}^{(t)}$.

Prediction objective and teacher update. During each student update, the teacher target is detached with the stop-gradient operator $\mathrm{sg}[\cdot]$, which is the identity in the forward pass and has zero derivative in the backward pass. Consequently, the representation-prediction loss is

$$\ell_{\text{pred}}^{(t)}(\boldsymbol{\theta}, \boldsymbol{\phi}; \bar{\boldsymbol{\theta}}) = \frac{1}{NDL}\left\|\boldsymbol{g}_{\boldsymbol{\phi}}\left(\mathbf{Z}_{\text{stu}}^{(t)}\right) - \mathrm{sg}\left[\mathbf{Z}_{\text{tea}}^{(t)}\right]\right\|_{1,1}, \tag{8}$$

where $\|\cdot\|_{1,1}$ sums absolute values over all matrix entries. The denominator averages the prediction error over search tokens, feature channels, and selected encoder layers. Gradients thus update the student and predictor, but not the teacher. Instead, we adjust the teacher parameters as an EMA of the student parameters:

$$\bar{\boldsymbol{\theta}} \leftarrow \rho\bar{\boldsymbol{\theta}} + (1-\rho)\boldsymbol{\theta}, \qquad \rho \in [0, 1). \tag{9}$$

The privileged templates and EMA play distinct roles: the templates determine what additional information the teacher observes, while EMA stabilizes the teacher's parameter trajectory and representation targets.

3.3 Reliability Weighting

Privileged inputs make the teacher more informative, but they do not necessarily make every teacher representation equally reliable as a prediction target. For a given sample, the teacher may still localize poorly, or it may be no better than the student. We therefore use localization quality as a training-time proxy for transfer reliability, weighting representation prediction by the teacher's relative advantage over the student and its absolute localization accuracy.

Reliability weight. For frame t, we measure the localization quality of each branch by its intersection over union (IoU) with the ground-truth bounding box:

$$u_{\text{stu}}^{(t)} = \mathrm{IoU}\left(\hat{\boldsymbol{b}}_{\text{stu}}^{(t)}, \boldsymbol{b}^{(t)}\right), \qquad u_{\text{tea}}^{(t)} = \mathrm{IoU}\left(\hat{\boldsymbol{b}}_{\text{tea}}^{(t)}, \boldsymbol{b}^{(t)}\right).$$

The reliability weight combines a relative advantage term with an absolute quality term:

$$w^{(t)} = \mathrm{sg}\left[\underbrace{\mathrm{clip}\left(\frac{u_{\text{tea}}^{(t)} - u_{\text{stu}}^{(t)}}{\tau}, 0, 1\right)}_{\text{advantage}} \underbrace{\left(u_{\text{tea}}^{(t)}\right)^{\gamma}}_{\text{quality}}\right], \qquad \tau > 0,\ \gamma \geq 0, \quad w^{(t)} \in [0, 1]. \tag{10}$$

The clipped advantage term suppresses transfer when the teacher is no more accurate than the student, increases linearly with the teacher's positive IoU margin, and saturates once this margin

reaches τ. However, a positive margin alone is insufficient: the teacher may outperform the student even when both predictions are poor. The quality term therefore discounts such cases according to the teacher's absolute IoU. With the convention $(u_{\text{tea}}^{(t)})^0 = 1$, setting $\gamma = 0$ recovers advantage-only weighting.

Overall training objective. We define the complete per-sample training objective as

$$\ell^{(t)}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \ell_{\text{track}}^{(t)}(\boldsymbol{\theta}) + \lambda w^{(t)} \ell_{\text{pred}}^{(t)}(\boldsymbol{\theta}, \boldsymbol{\phi}), \qquad \lambda \geq 0, \tag{11}$$

where λ controls the strength of privileged representation prediction relative to standard tracking supervision. The tracking loss is applied to every sample, while $w^{(t)}$ adjusts the representation-prediction loss according to the teacher's reliability. The stop-gradient in Equation (10) treats $w^{(t)}$ as a fixed sample-dependent coefficient during backpropagation. Without this detachment, the student could reduce the weighted prediction loss $w^{(t)}\ell_{\text{pred}}^{(t)}$ by changing its localization output to lower $w^{(t)}$, rather than by predicting the teacher representation more accurately. Detaching $w^{(t)}$ prevents this shortcut and restricts the weight to controlling the strength of representation supervision.

4 Experiments

In this section, we evaluate whether privileged appearance transfer improves a deployable tracker that uses only the student branch at inference. We first specify the training and inference protocols, then compare two model scales across seven benchmarks, and finally use controlled ablations to isolate the effects of privileged appearances, multi-level prediction, and reliability weighting.

4.1 Experimental Setup

We instantiate PATT-B256 and PATT-L256 in the one-stream Transformer tracking framework (Ye et al., 2022) using Base and Large Fast-iTPN encoders (Tian et al., 2024), respectively. We train on COCO (Lin et al., 2014), LaSOT (Fan et al., 2019), GOT-10k (Huang et al., 2021), TrackingNet (Müller et al., 2018), and VastTrack (Peng et al., 2024), and evaluate on seven benchmarks covering different tracking regimes. Only the student is retained at inference. More details of the model variants and the training, inference, and evaluation protocols are provided in Appendix A.

4.2 Main Results

Performance against long-horizon appearance change. LaSOT in Table 1 is the benchmark most directly aligned with our motivation: as sequences become longer, past templates are more likely to become stale, and the tracker must rely on imperfectly updated target evidence. Compared with temporal baselines such as ARPTrack (Liang et al., 2025), which pretrains appearance-motion evolution, and RELO (Chen et al., 2026), which adds reward-based localization training and temporal token propagation, PATT changes only the training signal while leaving the deployed tracker unchanged. This distinction is most visible on LaSOT. Both PATT variants achieve the best overlap and precision metrics within their scale groups, improving AUC over the strongest same-scale alternative by 2.1 points for PATT-B256 and 1.6 points for PATT-L256. The gains across both AUC and precision suggest that privileged appearance transfer helps the student encode search evidence that remains useful when historical templates no longer faithfully describe the target. On $\text{LaSOT}_{\text{ext}}$, the pattern is slightly less uniform: PATT-B256 leads all metrics in the base-scale group, while PATT-L256 ranks second to RELO-L256. Thus, the benefit extends to category shift, although strong online temporal adaptation remains competitive at larger scale.

Generalization across evaluation regimes. On TrackingNet, both PATT variants obtain the best results within their scale groups, though the margins are smaller. This indicates that privileged appearance transfer does not harm short-term tracking and can still improve the student representation, even when template staleness is less severe. The small gains are nevertheless meaningful because ARTrack (Wei et al., 2023) and ODTrack (Zheng et al., 2024) modify temporal inference through autoregressive trajectory modeling or online token propagation, whereas PATT adds no inference-time mechanism beyond the underlying tracker. On GOT-10k, PATT-L256 leads all criteria, while

Table 1: **Comparison on four large-scale tracking benchmarks.** Results are grouped by model scale, with base-scale trackers in the upper block and large-scale trackers in the lower block. AUC, P, P_{Norm}, AO, and SR denote area under the success curve, precision, normalized precision, average overlap, and success rate. Within each block, **bold** and underlined values indicate the best and second-best results, respectively.

Model	LaSOT			$\text{LaSOT}_{\text{ext}}$			TrackingNet			GOT-10k		
	AUC	P	P_{Norm}	AUC	P	P_{Norm}	AUC	P	P_{Norm}	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$
SwinTrack-B384 (Lin et al., 2022)	71.3	76.5	-	49.1	55.6	-	84.0	82.8	-	72.4	80.5	67.8
SeqTrack-B256 (Chen et al., 2023)	69.9	76.3	79.7	49.5	56.3	60.8	83.3	82.2	88.3	74.7	84.7	71.8
ARTrack-B256 (Wei et al., 2023)	70.4	76.6	79.5	46.4	52.3	56.5	84.2	83.5	88.7	73.5	82.2	70.9
OneTracker-B384 (Hong et al., 2024)	70.5	76.5	79.9	-	-	-	83.7	82.7	88.4	-	-	-
SUTrack-B224 (Chen et al., 2025)	73.2	80.5	83.4	53.1	60.5	64.2	85.7	85.1	90.3	77.9	87.5	78.5
ARPTrack-B256 (Liang et al., 2025)	72.6	78.5	81.4	52.0	58.7	62.9	85.5	85.3	90.0	77.7	87.3	74.3
RELO-B256 (Chen et al., 2026)	73.3	80.6	83.4	54.2	62.2	65.4	86.4	86.2	90.7	**80.5**	**90.5**	81.2
PATT-B256 (Ours)	**75.4**	**84.2**	**85.8**	**55.4**	**63.6**	**66.8**	**86.6**	**86.5**	**91.2**	79.7	89.2	**81.4**
MixFormer-L320 (Cui et al., 2022)	70.1	76.3	79.9	-	-	-	83.9	83.1	88.9	-	-	-
SimTrack-L224 (Chen et al., 2022)	70.5	-	79.7	-	-	-	83.4	-	87.4	69.8	78.8	66.0
ODTrack-L384 (Zheng et al., 2024)	74.0	82.3	84.2	53.9	61.7	65.4	86.1	86.7	91.0	78.2	87.2	77.3
LoRAT-L224 (Lin et al., 2024)	74.2	80.9	83.6	52.8	60.0	64.7	85.0	84.4	89.5	75.7	84.9	75.0
SUTrack-L224 (Chen et al., 2025)	73.5	80.9	83.3	54.0	61.7	65.3	86.5	86.7	90.9	81.0	90.4	82.4
LoRATv2-L224 (Lin et al., 2025)	74.4	81.2	83.8	-	-	-	85.4	-	-	76.9	86.3	76.4
RELO-L256 (Chen et al., 2026)	75.1	83.4	85.1	**57.5**	**66.7**	**69.1**	87.3	88.0	91.6	81.8	91.1	83.5
PATT-L256 (Ours)	**76.7**	**85.4**	**87.1**	56.7	64.6	68.2	**87.4**	**88.2**	**91.9**	**82.2**	**91.4**	**84.9**

PATT-B256 shows a more selective profile: it attains the best strict-overlap success rate $\text{SR}_{0.75}$, but remains behind RELO-B256 in AO and $\text{SR}_{0.5}$. Together, these results suggest that privileged appearance transfer generalizes beyond long sequences, with its clearest base-scale effect appearing under stricter localization criteria.

Transfer beyond the training benchmarks. We further evaluate on TNL2K (Wang et al., 2021b), NFS (Kiani Galoogahi et al., 2017), and UAV123 (Müller et al., 2016), none of which is used for training. PATT-L256 achieves AUCs of 64.8%, 72.1%, and 72.4%, respectively, ranking first on TNL2K and NFS and second on UAV123 among the compared trackers. Full comparisons and analysis are provided in Appendix B.

4.3 Ablation Studies

We ablate PATT-B256 by changing one design choice at a time and report AUC on LaSOT and $\text{LaSOT}_{\text{ext}}$ in Table 2. The variants test three questions: whether the gain comes from privileged target appearances, which form of teacher supervision is most effective, and how teacher targets should be weighted across samples.

Where does the gain originate? Rows 2–4 isolate the information available to the teacher. The student-only baseline and the past-only teacher perform almost identically, showing that a teacher branch alone does not explain the improvement. Adding the current-frame target crop recovers most of the full model's gain, while the future-frame crop closes the remaining gap. These results indicate that exact current-frame appearance is the dominant source of improvement, while future appearance provides a smaller but complementary signal about target variation.

What should be transferred? Rows 5–8 examine the form and depth of teacher supervision. Output imitation improves over the student-only baseline, but remains clearly below representation prediction, indicating that the teacher's final box does not capture all useful privileged information. Predicting representations from four encoder levels outperforms last-layer prediction, suggesting that privileged appearances shape search evidence at multiple depths. Using all layers provides little additional benefit, so a compact set of representative layers is sufficient. Directly matching teacher features without a latent predictor also performs poorly: because the student and teacher receive different template sets, their representations need not be identical. The predictor is therefore important because it allows the student to infer the privileged representation target without forcing exact feature equality.

Table 2: **Ablation study on LaSOT and LaSOT$_{\text{ext}}$.** Results report AUC for PATT-B256 and its training variants. The rows isolate the effects of privileged template information, representation prediction, and reliability weighting. Δ denotes the average AUC change relative to the full model across the two benchmarks.

#	Training variant	LaSOT	LaSOT$_{\text{ext}}$	Δ
1	Full model (PATT-B256)	75.4	55.4	–
2	Student-only baseline	73.0	52.7	-2.6
3	Past-only teacher	73.2	52.6	-2.5
4	Past+current teacher	75.1	55.0	-0.4
5	Output imitation	73.5	53.7	-1.8
6	Last-layer prediction	74.8	54.3	-0.9
7	All-layer prediction	75.5	55.2	-0.1
8	Direct matching (no predictor)	73.3	54.3	-1.6
9	Uniform weighting	75.0	54.6	-0.6
10	Quality-only weighting	75.1	55.2	-0.3
11	Advantage-only weighting	75.2	54.8	-0.4

Which teacher targets should be trusted? Rows 9–11 evaluate how the prediction loss is weighted across samples. Uniform weighting degrades performance, confirming that privileged inputs do not make every teacher target equally reliable. Quality-only weighting discounts poorly localized teachers, while advantage-only weighting emphasizes samples where the teacher improves over the student; both recover part of the lost performance. Combining the two gives the best result, as it selects teacher targets that are informative relative to the student and accurate in absolute terms.

5 Conclusion and Discussion

We have introduced PATT, a teacher-student framework that uses exact current- and future-frame target crops as privileged training context for visual tracking. Rather than imitating the teacher's final box prediction, the student learns to predict the teacher's multi-level search representations, with reliability weighting emphasizing teacher targets that are both accurate and more informative than the student's own prediction. After training, the teacher, predictor, reliability computation, and privileged crops are removed, leaving the standard student-only inference procedure unchanged. Across seven benchmarks and two model scales, PATT achieves its clearest gains in long-horizon tracking while remaining competitive in short-term and cross-benchmark evaluations.

These results also clarify the scope of the present study. The framework requires frame-level ground-truth bounding boxes to construct privileged target crops and introduces auxiliary teacher and predictor computation during training. It also uses a fixed design for privileged-view selection, replacing two past-frame templates with one current-frame crop and one future-frame crop. Thus, the experiments do not yet determine which privileged views are most informative, how they should be selected across a trajectory, or how much of the benefit remains when annotation density or training resources are limited.

A natural next step is to make privileged appearance transfer more principled within the current framework. Instead of using a fixed current-plus-future template layout, future work could select privileged views according to information gain, temporal diversity, occlusion state, teacher uncertainty, or consistency across multiple candidate crops. The reliability weight could likewise be extended beyond localization quality to account for how stable and informative each teacher representation is for the student.

More broadly, privileged supervision need not be limited to exact target crops. Dense masks, object parts, depth, optical flow, language descriptions, long-range reappearance cues, and offline trajectory context could provide additional training-only signals for learning representations that remain causal at deployment. This perspective also connects naturally to emerging practical tracking settings, including foundation-model-assisted tracking, memory-based trackers, multi-object tracking, and multi-camera tracking. In these settings, information from later frames, neighboring objects, or other cameras may be available during training or offline annotation, even when the deployed tracker must operate online. Using such information to train stronger trackers is a promising direction toward robust tracking under appearance change, occlusion, re-identification, and cross-view ambiguity.

AI USE STATEMENT

This work was conceived, designed, and carried out by the authors, with assistance from generative AI tools. The research question, the idea of privileged appearance transfer, the teacher-student design with multi-level representation prediction and reliability weighting, and the experimental protocol originated with the authors, who also implemented the method, analyzed the results, and wrote the manuscript. Every research decision was made by the authors.

Among the tasks with required disclosure, we used generative AI tools in three ways. First, after the authors had formulated the core idea and method, we used them as a discussion partner to check our reasoning and obtain feedback on the methodology and experimental design. Second, with the method and code structure designed by the authors, we used them to speed up the writing of implementation code. Third, we occasionally used them to translate excerpts of related literature. We did not use generative AI tools to generate synthetic datasets, formulate mathematical claims, provide ingredients for or write proofs, clean or reformat datasets, perform qualitative or thematic data analysis, or interpret results.

Among the tasks with recommended disclosure, we used generative AI tools to supplement the authors' own literature survey by searching for and identifying additional related work, to organize and format experimental results, to support experiment management and server scheduling, and to improve the clarity and readability of text written by the authors. No section of this paper was drafted by generative AI.

All AI-assisted work was reviewed by the authors. Suggestions arising from discussions with AI tools were adopted only after the authors evaluated them, and we compared the research ideas with relevant literature through a manual survey to assess originality and establish their relationship to prior work. AI-assisted code was inspected and tested by the authors for correctness and consistency with the intended method. Organized experimental results were manually checked against the raw outputs, and language edits were checked to preserve the intended scientific meaning. We take full responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

PATT advances representation learning for visual tracking, with potential applications in robotics, autonomous systems, and human-computer interaction. Our experiments use established public benchmarks, with privileged target crops derived from existing ground-truth bounding-box annotations. In applications involving people, the use of visual tracking should respect privacy and informed consent. We encourage responsible deployment through appropriate human oversight and validation in the intended application setting.

REPRODUCIBILITY STATEMENT

We describe PATT in Section 3, including the template-based tracking formulation, the construction of privileged templates, and multi-level prediction of teacher search representations. We also specify the training objectives, the update rule, and the reliability weighting mechanism. Section 4.1 and Appendix A document the experimental setup, including model variants and the training, inference, and evaluation protocols. Controlled ablations in Section 4.3 isolate the contributions of the main design choices. To support independent reproduction and further research, we will release our code as open source, together with trained models and the configurations used for training and evaluation.

REFERENCES

Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. ARTrackV2: Prompting autoregressive tracker where to look and how to describe. In *CVPR*, pp. 19048–19057, 2024.

Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional Siamese networks for object tracking. In *ECCV Workshops*, pp. 850–865, 2016.

Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, pp. 6182–6191, 2019.

Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qiuhong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: A simplified architecture for visual object tracking. In *ECCV*, pp. 375–392, 2022.

Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, pp. 8126–8135, 2021.

Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. SeqTrack: Sequence to sequence learning for visual object tracking. In *CVPR*, pp. 14572–14581, 2023.

Xin Chen, Ben Kang, Wanting Geng, Jiawen Zhu, Yi Liu, Dong Wang, and Huchuan Lu. SUTrack: Towards simple and unified single object tracking. In *AAAI*, pp. 2239–2247, 2025.

Xin Chen, Chuanyu Sun, Jiao Xu, Houwen Peng, Dong Wang, Huchuan Lu, and Kede Ma. RELO: Reinforcement learning to localize for visual object tracking. In *ICML*, 2026.

Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. MixFormer: End-to-end tracking with iterative mixed attention. In *CVPR*, pp. 13608–13618, 2022.

Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate tracking by overlap maximization. In *CVPR*, pp. 4660–4669, 2019.

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 *CVPR*, pp. 5374–5383, 2019.

Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. LaSOT: A high-quality large-scale single object tracking benchmark. *IJCV*, 129(2):439–461, 2021.

Xiaokun Feng, Shiyu Hu, Xuchen Li, Dailing Zhang, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. ATCTrack: Aligning target-context cues with dynamic target states for robust vision-language tracking. In *ICCV*, pp. 19850–19861, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *CVPR*, pp. 826–834, 2016.

Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, et al. OneTracker: Unifying visual object tracking with foundation models and efficient tuning. In *CVPR*, pp. 19079–19091, 2024.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE TPAMI*, 43(5):1562–1577, 2021.

Yuqing Huang, Liting Lin, Weijun Zhuang, Zhenyu He, and Xin Li. Drift-resilient temporal priors for visual tracking. In *CVPR*, pp. 6847–6856, 2026.

Daniel Kahneman, Anne Treisman, and Brian J. Gibbs. The reviewing of object files: Object-specific integration of information. *Cogn Psychol.*, 24(2):175–219, 1992.

Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for Speed: A benchmark for higher frame rate object tracking. In *ICCV*, pp. 1125–1134, 2017.

Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with Siamese region proposal network. In *CVPR*, pp. 8971–8980, 2018.

Shiyi Liang, Yifan Bai, Yihong Gong, and Xing Wei. Autoregressive sequential pretraining for visual tracking. In *CVPR*, pp. 7254–7264, 2025.

Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. SwinTrack: A simple and strong baseline for Transformer tracking. In *NeurIPS*, pp. 16743–16754, 2022.

Liting Lin, Heng Fan, Zhipeng Zhang, Yaowei Wang, Yong Xu, and Haibin Ling. Tracking meets LoRA: Faster training, larger model, stronger performance. In *ECCV*, pp. 300–318, 2024.

Liting Lin, Heng Fan, Zhipeng Zhang, Yuqing Huang, Yaowei Wang, Yong Xu, and Haibin Ling. LoRATv2: Enabling low-cost temporal modeling in one-stream trackers. In *NeurIPS*, pp. 119714–119732, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *ICLR*, 2016.

Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pp. 674–679, 1981.

Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *ICCV*, pp. 13444–13454, 2021.

Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *CVPR*, pp. 8731–8740, 2022.

Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Information bottleneck learning using privileged information for visual recognition. In *CVPR*, pp. 1496–1505, 2016.

Matthias Müller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for UAV tracking. In *ECCV*, pp. 445–461, 2016.

Matthias Müller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, pp. 300–317, 2018.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, and Nicolas Ballas. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. In *ECCV*, pp. 671–689, 2026.

Liang Peng, Junyuan Gao, Xinran Liu, Weihong Li, Shaohua Dong, Zhipeng Zhang, Heng Fan, and Libo Zhang. VastTrack: Vast category visual object tracking. In *NeurIPS*, pp. 130797–130818, 2024.

Zenon W. Pylyshyn and Ron W. Storm. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spat Vis.*, 3(3):179–197, 1988.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *ICLR*, 2015.

Jianbing Shen, Yuanpei Liu, Xingping Dong, Xiankai Lu, Fahad Shahbaz Khan, and Steven C. H. Hoi. Distilled Siamese networks for visual tracking. *IEEE TPAMI*, 44(12):8896–8909, 2022.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pp. 1195–1204, 2017.

Yunjie Tian, Lingxi Xie, Jihao Qiu, Jianbin Jiao, Yaowei Wang, Qi Tian, and Qixiang Ye. Fast-iTPN: Integrally pre-trained Transformer pyramid network with token migration. *IEEE TPAMI*, 46(12):9766–9779, 2024.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Netw.*, 22(5–6):544–557, 2009.

Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*, pp. 1571–1580, 2021a.

Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *CVPR*, pp. 13763–13773, 2021b.

Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *CVPR*, pp. 9697–9706, 2023.

Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, pp. 2411–2418, 2013.

Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal Transformer for visual tracking. In *ICCV*, pp. 10448–10457, 2021.

Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, pp. 341–357, 2022.

Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhiyi Mo, Shengping Zhang, and Xianxian Li. ODTrack: Online dense temporal token learning for visual tracking. In *AAAI*, pp. 7588–7596, 2024.

A Implementation Details

This appendix details the experimental setup summarized in Section 4.1.

Table 3: **Model configurations and inference cost.** Parameter counts and floating-point operations (FLOPs) outside parentheses are for the deployed tracker, while values in parentheses indicate the additional training-only cost of the representation predictor. Speed is reported in frames per second (FPS) on a single NVIDIA RTX 4090 GPU.

Model	# Params (M)	FLOPs (G)	Speed (FPS)
PATT-B256	70 (+23)	39 (+6)	50
PATT-L256	247 (+25)	131 (+6)	32

Model variants. We instantiate PATT in the one-stream Transformer tracking framework of OSTrack (Ye et al., 2022). PATT-B256 and PATT-L256 use the Base and Large Fast-iTPN encoders (Tian et al., 2024), respectively. Table 3 reports their deployed costs. Both variants use a 256×256 search crop, 128×128 templates, a patch stride of 16, $K = 4$ template slots, and the center-based prediction head of OSTrack. The training-only representation predictor comprises 12 Transformer layers with a hidden dimension of 384 and 12 attention heads.

Training protocol. We train on the training splits of COCO (Lin et al., 2014), LaSOT (Fan et al., 2019), GOT-10k (Huang et al., 2021), TrackingNet (Müller et al., 2018), and VastTrack (Peng et al., 2024). For each training example, we sample ordered frames $r_1 < r_2 < r_3 < r_4 < t < t'$ from one video. The student uses target crops from frames $r_1, \ldots, r_4$ as templates and the crop from frame t as the search input. The teacher uses the same search crop but replaces the final two past-frame templates with ground-truth target crops from frames t and t'. We obtain template and search crops by expanding their ground-truth bounding boxes by factors of 2 and 4, respectively, and further apply translation and scale jitter to the search crop. We train with AdamW for 10 epochs using a minibatch size of 64 and an initial learning rate of 10^{-4}, decayed by a factor of 10 after epoch 8. We predict representations from $L = 4$ evenly spaced encoder layers and update the teacher by EMA with $\rho = 0.99925$. The loss hyperparameters are $\lambda = 3$, $\tau = 0.1$, and $\gamma = 0.5$.

Inference protocol. At test time, only the student f_{θ} is retained; the teacher, predictor, reliability computation, and privileged templates are removed. Following a common template-update protocol (Yan et al., 2021), the student keeps the ground-truth first-frame template and fills the remaining three slots with dynamic templates cropped from its own predictions. Dynamic templates are refreshed every 25 frames when prediction confidence exceeds 0.80.

Evaluation protocol. We evaluate on seven benchmarks covering different tracking regimes. LaSOT (Fan et al., 2019) and LaSOT$_{\text{ext}}$ (Fan et al., 2021) evaluate long-horizon appearance change and cross-category generalization. TrackingNet (Müller et al., 2018) and GOT-10k (Huang et al., 2021) test large-scale short-term tracking. TNL2K (Wang et al., 2021b), NFS (Kiani Galoogahi et al., 2017), and UAV123 (Müller et al., 2016) assess transfer beyond the training datasets. We use the area under the success curve (AUC) as the primary metric. LaSOT, LaSOT$_{\text{ext}}$, and TrackingNet additionally report precision (P) and normalized precision (P_{Norm}), while GOT-10k reports average overlap (AO) and success rates at IoU thresholds 0.5 and 0.75.

B TRANSFER BEYOND THE TRAINING BENCHMARKS

Table 4: **Cross-benchmark transfer on held-out datasets.** Results report AUC on TNL2K, NFS, and UAV123, none of which is used for training.

Model	TNL2K	NFS	UAV123
TrDiMP (Wang et al., 2021a)	-	66.5	67.5
TransT (Chen et al., 2021)	50.7	65.7	69.1
SimTrack (Chen et al., 2022)	55.6	-	71.2
OSTrack (Ye et al., 2022)	55.9	66.5	70.7
SeqTrack-L256 (Chen et al., 2023)	56.9	66.9	69.7
ARTrack-L384 (Wei et al., 2023)	60.3	67.9	71.2
LoRAT-L224 (Lin et al., 2024)	61.1	66.0	**73.3**
RELO-L256 (Chen et al., 2026)	<u>63.6</u>	71.3	71.4
PATT-B256 (Ours)	62.9	<u>71.6</u>	70.4
PATT-L256 (Ours)	**64.8**	**72.1**	<u>72.4</u>

Table 4 further evaluates transfer on TNL2K, NFS, and UAV123, none of which is used for training. PATT-L256 achieves the highest AUC on TNL2K and NFS, improving over the strongest non-PATT result by 1.2 points and 0.8 points, respectively, while ranking second on UAV123 with an AUC of 72.4. This consistent held-out performance supports the view that privileged target appearances improve the learned representation rather than merely fitting the training benchmarks. The base-scale model is less uniform: PATT-B256 is competitive on TNL2K and ranks second on NFS, but does not improve on UAV123. The contrast between PATT-B256 and PATT-L256 suggests that model capacity may affect how fully the student can absorb training-only privileged targets. Overall, the results show that PATT provides its strongest and most consistent gains at larger scale.