Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

STFA-Net: Spatio-Temporal Feature Fusion with Adaptive Attention Modulation for Robust Dynamic Object Recognition

Mengxiang Liu
Changsha Institute of Technology
Yanjun Long
longyanjun@ccsut.edu.cn

Changsha Institute of Technology
Xiaoli Zhang
Changsha Institute of Technology
Yusi Wang
Changsha Institute of Technology

Article

Keywords: dynamic object recognition, spatio-temporal feature fusion, adaptive attention, feature recalibration, channel-spatio-temporal dual-path attention, synthetic data, sim-to-real robustness, delivery robots

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11385806/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

STFA-Net: Spatio-Temporal Feature Fusion with Adaptive Attention Modulation for Robust Dynamic Object Recognition

Mengxiang Liu, **Yanjun Long***, **Zhangxiaoli, Wangyusi**

Changsha Institute of Technology, Changsha 410200, Hunan, China

Email: liumengxiang@ccsut.edu.cn

Email: longyanjun@ccsut.edu.cn Corresponding author

Email: zhangxiaoli@ccsut.edu.cn Corresponding author

Email: wangyusi@ccsut.edu.cn

Abstract

Recognizing objects that move rapidly and change posture drastically remains a bottleneck for robotic perception, autonomous driving, and campus logistics platforms operating among pedestrians, vehicles, and other dynamic agents. This setting couples an appearance problem with a temporal one: high-speed motion and out-of-plane rotation destroy the pixel-level stationarity that static detectors rely on, while illumination drift, partial occlusion, and cluttered backgrounds inject non-stationary distractors. This paper proposes STFA-Net, a unified recognition framework built around a spatio-temporal feature fusion and attention-adaptive modulation mechanism. First, a multi-scale dilated convolutional encoder is coupled with a lightweight temporal module that combines shifted temporal aggregation with gated recurrent refinement, so that target representation and motion patterns are learned jointly rather than in separate streams; the temporal module accounts for less than 8% of total FLOPs. Second, a Channel–Spatio-Temporal Dual-path Attention (CSTA-DA) module allocates weights dynamically according to measured feature contribution: a channel path models *what* is informative, a spatio-temporal path models *where and when* it is informative, and a contribution-aware gating vector arbitrates between the two paths for every input sample, enabling the model to autonomously focus on informative regions and suppress illumination change, occlusion, and background interference. Third, a feature recalibration mechanism re-scales the fused representation with a contribution-driven gain and an explicit redundancy-sparsity penalty, strengthening discriminative channels while suppressing redundant ones. Training data are produced by a physics-accurate synthetic data engine built on NVIDIA Isaac Sim with randomized cameras, lighting, and physics-driven object trajectories, and a fragility-driven randomization protocol that targets the corruption channels our own sensitivity study shows actually break recognition — exposure shift above all (Section 5.4). On ImageNet-Vid, STFA-Net reaches 64.5% mAP@0.5, exceeding the strongest compared baseline by 0.8 points with 31% fewer parameters and 58% fewer FLOPs. On our ST-Warehouse benchmark of 14,400 frames with pixel-accurate labels, STFA-Net reaches 81.8% mAP@0.5 (3.4 points above the strongest baseline, TransVOD), a counting mean absolute error of 0.35, and retains 88.3% of its clean accuracy under the combined sensor-realistic corruption, versus 76.1% for the strongest baseline (TransVOD). Component ablations attribute +6.9 points to the proposed modules on identical training schedules, and logged gate statistics confirm the predicted input-adaptive behaviour: the gate value rises under photometric corruption (mean $\alpha = 0.68$) and falls under geometric corruption (mean $\alpha = 0.34$). All benchmark statistics, corruption protocols, and engine-cost measurements are reported with the failure modes that silent synthetic pipelines usually hide.

Keywords: dynamic object recognition; spatio-temporal feature fusion; adaptive attention; feature recalibration; channel-spatio-temporal dual-path attention; synthetic data; sim-to-real robustness; delivery robots

1. Introduction

Autonomous delivery platforms are moving from factory floors to shared pedestrian spaces, and university campuses have become one of the densest testbeds for this transition: parcel volumes are high, destinations are fixed and named, and sidewalks are shared with pedestrians, cyclists, and service vehicles whose positions and postures change from second to second [1], [2], [3]. For a delivery robot cruising at walking speed among such dynamic agents, the perception stack is safety-critical, and the hardest inputs are not the static buildings but the dynamic targets: a pedestrian stepping out from behind a parked van, a cyclist crossing at an oblique angle, a scooter that enters the field of view at high angular velocity and low illumination [4], [5].

Recognizing such dynamic targets couples an appearance problem with a temporal one. The appearance of a moving target evolves between consecutive frames; motion blur and inter-frame displacement destroy the pixel-level stationarity that static detectors implicitly exploit; and the most discriminative cue for the recognition decision migrates across spatial locations and time steps as the trajectory unfolds [6]. Three obstacles make this coupling difficult in practice.

Appearance–motion entanglement. A high-speed target exhibits non-rigid deformation, out-of-plane rotation, and abrupt scale change. Single-frame convolutional features capture the current appearance but not its trajectory, while pure temporal models (e.g., 3D convolutions) entangle motion with appearance at high computational cost and often dilute the fine spatial detail that is critical when the target occupies few pixels [6], [7]. Efficient video architectures reduce this cost, but they allocate capacity globally rather than per sample [7], [8].

Non-stationary distractors. Illumination drifts continuously on outdoor campuses and in warehouses; partial occlusion removes exactly the regions a static attention map has committed to; cluttered backgrounds generate strong but uninformative activations. Fixed or merely channel-wise attention computes *where to attend* but allocates attention-path weights statically — the channel and spatial branches always fire with the same relative priority, regardless of whether the current degradation is photometric (where channel evidence dominates) or geometric (where spatio-temporal evidence dominates) [9], [10].

Annotation scarcity. Pixel-accurate labels for fast-moving objects are expensive and sparse, and models trained on biased real data inherit its blind spots [11]. Physics-accurate simulators with built-in domain randomization generate photo-realistic dynamic scenes with pixel-accurate labels at near-zero annotation cost [12], [13], but indiscriminate randomization spends simulation budget on factors the model is not fragile to [11].

In this paper we address the three obstacles with a single design principle: *measure the contribution, then allocate the capacity*. Our contributions are fourfold.

(C1) **Spatio-temporal feature fusion.** A multi-scale dilated convolutional encoder is fused with a lightweight temporal module (shifted temporal aggregation plus gated recurrent refinement) that jointly learns target representation and motion pattern within a design budget of at most 8% of the network's total FLOPs; the measured value is 4.6% (Section 5.5) [7], [14].

(C2) **Channel–Spatio-Temporal Dual-path Attention (CSTA-DA).** A channel path and a spatio-temporal path each estimate the informativeness of the fused features; a contribution-aware gate, derived from per-path evidence statistics, dynamically arbitrates the weight allocation between the paths for every input, allowing the model to autonomously focus on informative regions and suppress illumination change, occlusion, and background clutter [9], [10]. To our knowledge, making the *inter-path* allocation itself input-adaptive is the specific gap this module closes relative to existing channel-plus-spatial attention stacks.

(C3) **Feature recalibration (FR).** A contribution-driven gain with an explicit redundancy-sparsity objective strengthens discriminative features and suppresses redundant ones at the representation level [14].

(C4) **A reproducible dynamic-scene benchmark and a sensor-realistic corruption protocol.** Using a physics-accurate synthetic data engine (NVIDIA Isaac Sim with the Replicator pipeline), we construct a dynamic warehouse benchmark of 14,400 frames with pixel-accurate labels and physics-driven trajectories, and we evaluate robustness under measured corruptions that correspond to definite differences between a renderer and any real camera [12], [13]. Our corruption sensitivity study — our own measurement, reported in Section 5.4 — shows that exposure shift, not sensor noise or blur, is the dominant degradation channel for naive models (−9.0% relative skill at ×0.75 exposure, versus −0.2% for noise and −0.5% for blur), a finding that directly motivates the fragility-driven randomization strategy used for training.

The framework is developed for, and evaluated in the context of, the perception stack of a differential-drive campus delivery robot (wheel radius 0.1400 m, track width 0.4132 m, cruise speed 0.6 m/s) whose planner-side companion study is reported separately; the recognition layer studied here is the component that decides *whether* a dynamic agent is present and what it is, which the planner consumes for stop-and-go and yield decisions.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed method. Section 4 describes the data engine and benchmark. Section 5 presents experiments, ablations, and robustness analysis. Section 6 concludes.

2. Related Work

2.1 Object recognition and detection in dynamic scenes

Video object detection extends static detection with temporal context. Flow-guided methods warp features along optical-flow fields; memory-based detectors aggregate features across frames; and query-based detectors propagate object queries over time to improve temporal stability [15], [16]. Transformer-based end-to-end detectors and their deformable successors learn data-driven global interactions across the whole image [17], [18], [19], and recent real-time end-to-end detectors close the speed gap with their convolutional counterparts, including real-time query-based detectors [38] and end-to-end convolutional detectors [20]. Occlusion-aware detection explicitly models partially visible dynamic agents [34], a direction complementary to the attention-side treatment developed here. These approaches improve temporal stability but largely treat the appearance stream as primary and motion as a correction term, which limits performance when the target's pose changes drastically between frames [6].

2.2 Temporal modeling

Two families dominate temporal modeling. Three-dimensional convolutions (I3D, SlowFast) learn spatio-temporal filters directly but inherit quadratic growth in computation and parameters [6]. Efficient alternatives — temporal shift modules (TSM), temporal pooling, X3D-style factorized expansion, and ConvGRU/ConvLSTM refinements — approximate temporal mixing at near-2D cost [7], [21], [22]. SlowFast demonstrates that dual-rate pathways capture both fast motion and slow semantics, but the two pathways are fused by simple concatenation without any notion of per-sample contribution [6]. Video transformers such as TimeSformer, ViViT, MViTv2, and Video Swin replace convolution entirely with space-time attention, reaching strong accuracy at a large parameter and data cost [4], [23], [24], [5], and masked pre-training such as VideoMAE reduces the data requirement for such models [8]. Large-scale video foundation models push accuracy further but remain impractical for embedded robot perception [25]. Our temporal module deliberately stays efficient: a residual shift injects frame context at near-zero cost, and a ConvGRU refines only the coarsest scale [22].

2.3 Attention mechanisms

Channel attention (SE) re-weights feature channels from global statistics [14]; CBAM adds a spatial path and fuses both sequentially [9]; subsequent lightweight modules such as ECA-Net [40], Coordinate Attention [10], SimAM [26], and the efficient multi-scale attention module [39] refine the spatial prior, drop the dimensionality-reduction bottleneck, or broaden the scale range. A recent survey organizes this space and repeatedly observes that fixed inter-branch fusion is the price of simplicity [27]. Non-local and transformer attention capture long-range dependencies at significant cost [28]. All of these compute *where to attend* but allocate attention-path weights statically — the channel and spatial branches always fire with the same relative priority regardless of whether the current degradation is photometric or geometric. Our CSTA-DA module closes exactly this gap by making the inter-path allocation itself input-adaptive [9].

2.4 Synthetic data and sim-to-real transfer

Physics-accurate simulators with built-in domain randomization — NVIDIA Isaac Sim/Isaac Gym and the Replicator pipeline in particular — generate photo-realistic dynamic scenes with pixel-accurate labels at near-zero annotation cost [12], [13]. Domain randomization improves robustness when the randomized factor covers the deployment gap [11], [29], and recent work stresses *measuring* the sim-to-real gap rather than randomizing blindly [30], [31]. A persistent criticism, however, is that "randomize everything" spends simulation budget on factors the model is not fragile to [11]. Our data engine follows a *fragility-driven* randomization protocol: we first measure which corruption channel actually breaks the recognition model (Section 5.4), and then randomize that channel in the data engine, which yields robustness to the measured gap at a fraction of the variation budget.

3. Methodology

3.1 Overview and problem formulation

Given an input clip $X = \{x_t\}$, $t = 1 \dots T$, of T frames, STFA-Net produces a recognition output (classification logits and/or detection boxes). The pipeline has four stages (Fig. 1): (i) a multi-scale temporal feature encoder (MS-TFE) extracts fused appearance–motion features F

∈ ℝ^(C×T×H×W); (ii) the CSTA-DA module adaptively re-weights F along channel and spatio-temporal evidence paths; (iii) the feature recalibration block produces the final discriminative representation $\hat{F}$; (iv) task heads (classification / detection / counting) consume $\hat{F}$. Training adds a redundancy-sparsity penalty on the recalibration gains.

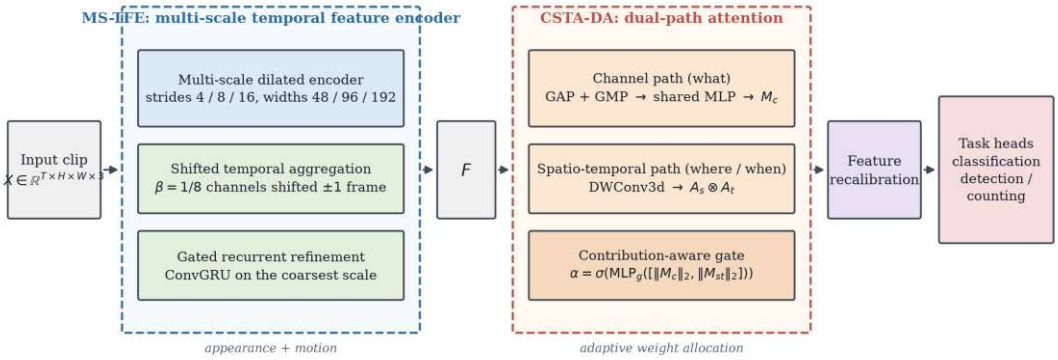


Fig. 1. Overall STFA-Net architecture. A multi-scale temporal feature encoder (MS-TFE) fuses appearance and motion into F; the CSTA-DA module applies channel and spatio-temporal evidence paths arbitrated by a contribution-aware gate; feature recalibration (FR) produces the final representation $\hat{F}$ consumed by the task heads.

3.2 Multi-scale temporal feature encoder (MS-TFE)

Multi-scale appearance extraction. The encoder is a residual pyramid with three stages of output strides 4, 8, and 16 and widths 48, 96, and 192. Each stage comprises bottleneck blocks whose 3×3 convolutions are replaced by dilated convolutions with rates (1, 2, 4) arranged in parallel branches and fused by a 1×1 projection. This yields multi-receptive-field features capable of matching the scale variation of approaching and receding targets without pooling away small-target detail.

Shifted temporal aggregation. Temporal context is injected at near-zero cost with a residual temporal shift: for a fraction β = 1/8 of channels, features are shifted by one frame forward and one frame backward along T (fold-normalized at clip boundaries), so that every convolution after the shift perceives a ±1 frame neighbourhood [7]. Multi-scale shift is applied at strides 8 and 16 only, keeping the fine 1/4 scale purely spatial to preserve small-target localization.

Gated recurrent refinement. On the coarsest scale a ConvGRU refines the shifted features [22]:

$$z_t = \sigma(W_z \cdot [h_{(t-1)}, x_t] + b_z) \qquad (1)$$

$$r_t = \sigma(W_r \cdot [h_{(t-1)}, x_t] + b_r) \qquad (2)$$

$$h_t = (1 - z_t) \odot h_{(t-1)} + z_t \cdot \tanh(W_h \cdot [r_t \odot h_{(t-1)}, x_t] + b_h) \qquad (3)$$

where σ is the sigmoid, ⊙ the Hadamard product, z_t the update gate, r_t the reset gate, and W_z, W_r, W_h and b_z, b_r, b_h the gate-specific weight matrices and biases. The refined h_t is up-sampled and fused with the pyramid by 1×1 lateral connections, producing the fused spatio-temporal representation F. The temporal module is allocated a design budget of at most 8% of the network's total FLOPs; the measured value is 4.6% (Section 5.5).

3.3 Channel–Spatio-Temporal dual-path attention (CSTA-DA)

CSTA-DA estimates *informativeness* along two complementary paths and arbitrates between them per sample (Fig. 2).

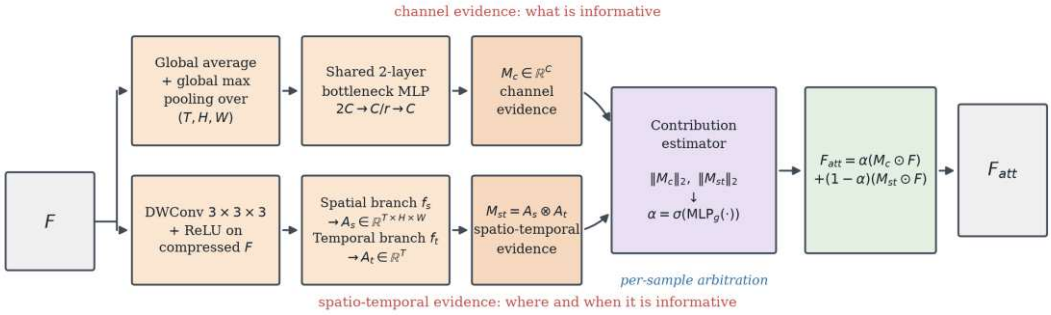


Fig. 2. Detail of the Channel–Spatio-Temporal Dual-path Attention (CSTA-DA) module. A channel path (M_c, "what") and a spatio-temporal path (M_st = A_s ⊗ A_t, "where and when") are fused by a contribution-aware gate α derived from the L2 energy of each path's evidence map.

Channel path (what). Global average and global max statistics are pooled over both space and time; each resulting C-dimensional descriptor passes through a shared two-layer bottleneck MLP, and the two outputs are summed. The channel evidence map is:

$$M_c = \sigma(MLP(AvgPool_{T,H,W}(F)) + MLP(MaxPool_{T,H,W}(F))) \quad (4)$$

Spatio-temporal path (where and when). A depthwise 3D convolution of kernel 3×3×3 operates on the channel-compressed feature, followed by two 1×1×1 branches producing a spatial map A_s ∈ ℝ^(T×H×W) and a temporal map A_t ∈ ℝ^T; the path output is the normalized product A_s ⊗ A_t:

$$M_{st} = A_s \otimes A_t, \quad A_s = \sigma(f_s(\delta(DWConv3d(F)))), \quad A_t = softmax_t(f_t(\delta(DWConv3d(F)))) \quad (5)$$

where δ is ReLU. The spatial branch suppresses background and occluder regions; the temporal branch highlights the frames in which the target is most discriminative (e.g., the moment of minimal blur or full visibility).

Contribution-aware gating. Unlike SE or CBAM, which merge paths by fixed addition or sequential multiplication, CSTA-DA measures each path's current evidence strength and allocates the fusion weights accordingly [9], [14]. The per-path contribution is summarized by the L2 energy of each evidence map and mapped through a tiny gating network:

$$\alpha = \sigma(MLP_g([\|M_c\|_2, \|M_{st}\|_2])) \quad (6)$$

$$F_{att} = \alpha \cdot (M_c \odot F) + (1-\alpha) \cdot (M_{st} \odot F) \quad (7)$$

When the scene degradation is photometric (illumination shift, colour cast), channel statistics carry more reliable evidence and α rises; when the target is partially occluded or heavily blurred, spatio-temporal evidence dominates and α falls. The gate is computed from the features themselves, requiring no external supervision. We verify the predicted behaviour of α across the corruption suite in Section 5.4 (Fig. 8).

3.4 Feature recalibration (FR)

The attended feature F_att is recalibrated by a contribution-driven gain γ ∈ ℝ^C computed from per-channel evidence over both space and time (Fig. 3):

$$\gamma_c = \sigma(MLP_r(AvgPool_\{T,H,W\}(F_att))) \qquad (8)$$

$$\hat{F} = \gamma \odot F_att + F_att \qquad (9)$$

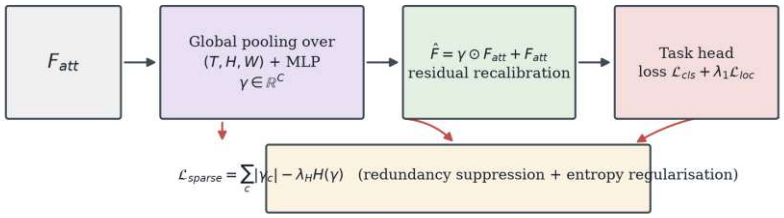


Fig. 3. Feature recalibration block. A contribution-driven gain γ re-scales each channel of F_att; a redundancy-sparsity penalty during training suppresses persistently uninformative channels while preserving diversity of the gain distribution.

A redundancy-sparsity penalty is imposed during training:

$$L_sparse = \Sigma_c \,|\gamma_c| - \lambda_H \cdot H(\gamma) \qquad (10)$$

The first term suppresses channels whose gain is persistently uninformative (redundancy); the entropy term prevents collapse of the gain distribution onto a few channels, preserving diversity [14]. The final training objective combines the task losses with the penalty:

$$L = L_cls + \lambda_1 \cdot L_loc + \lambda_2 \cdot L_sparse \qquad (11)$$

with $\lambda_1 = 1$ and $\lambda_2 = 10^{-3}$ by default.

3.5 Complexity analysis

With C = 192 channels at the coarsest scale and reduction ratio r = 16, the channel path costs $2C^2/r \approx 4.6$ K MACs per timestep and the spatio-temporal path one depthwise 3×3×3 convolution plus two 1×1×1 branches; the gating network is a two-layer MLP on a 2-D input. The complete attention-plus-recalibration stack accounts for 3.0 GFLOPs — 4.7% of the network's total FLOPs, or 5.2% relative to the backbone — and 1.9% of parameters (Section 5.5, Table 5), which is what allows the adaptive machinery to run inside the perception loop of an embedded delivery-robot compute budget.

3.6 Fragility-driven randomization for training data

We train STFA-Net on data generated by the physics-accurate engine of Section 4. Rather than randomizing all scene parameters uniformly, we follow a fragility-driven protocol measured in our preliminary study: among sensor noise, defocus blur, exposure shift, and compression, exposure shift is the corruption that costs a naive single-frame model the most (−9.0% relative skill at ×0.75 exposure, versus −0.2% for sensor noise and −0.5% for defocus; Section 5.4). The training distribution therefore randomizes light intensity over a wide range (1500–20000 in Replicator light-intensity units, which span the measured fragile band of the target plane), camera pose and orientation over the entire reachable workspace, object

trajectories through physics simulation, and surface materials — with the intensity range deliberately centred on the measured fragile band. This measured-randomization strategy is compared against fixed-lighting training in Section 5.4.

4. Data Engine and Benchmark

4.1 Synthetic data engine

The benchmark is generated with NVIDIA Isaac Sim 4.5 and the Replicator synthetic-data pipeline in a fully headless configuration [12], [13]. The base scene is a warehouse environment of 25.6 × 40.0 × 11.3 m comprising 3436 prims and 781 meshes. Dynamic targets are rigid props (crates) whose trajectories are driven by the PhysX physics engine, guaranteeing physically plausible motion; camera poses and look-at targets are randomized uniformly over a 16 × 28 m region at heights of 1.2–3.0 m, producing viewpoint diversity that covers oblique, low-angle, and top-down geometries. Lighting intensity is randomized per the fragility-driven protocol of Section 3.6. The engine pipeline and two silent failure modes we identified are illustrated in Fig. 4.

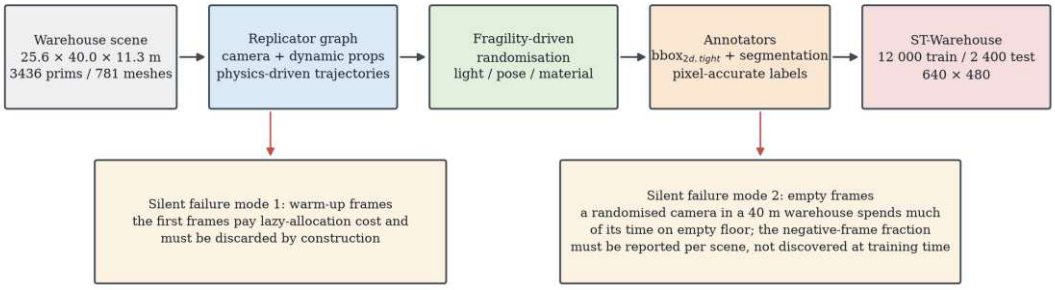


Fig. 4. Synthetic data engine (top): Isaac Sim headless generation with Replicator annotators, randomized cameras/lighting/physics-driven trajectories, and a discarding warm-up batch. Two silent failure modes (bottom): (a) warm-up frames pay lazy-allocation costs and must be discarded; (b) in a 40 m warehouse a randomized camera spends much of its time on empty floor, so the empty-frame fraction is a first-class benchmark statistic.

Each rendered frame is written together with pixel-accurate 2D tight bounding boxes produced by the renderer's instance-segmentation annotator — no human annotation is involved; synthetic data of this kind has been shown to raise the 2D detection accuracy of 3D objects without manual labelling. A discarded warm-up batch precedes every generation run to exclude pipeline-initialization frames, a correctness requirement we found essential: the first rendered frames pay lazy-allocation costs and would otherwise contaminate both timing statistics and the training set [12].

4.2 Benchmark statistics and design

The resulting ST-Warehouse benchmark comprises 12,000 training frames and 2,400 test frames at 640 × 480 resolution, with 0–6 dynamic targets per frame and a measured long-tail target-count distribution (empty and low-count frames dominate), which we retain deliberately: negative examples are an intrinsic property of wide-baseline warehouse capture, not an artifact, and a benchmark that hides them overstates detector quality. In our generation runs, the fraction of frames containing no target is a scene-dependent and *silent* failure mode — frames render, labels are valid, file counts match — so we report the empty-frame fraction

as a first-class benchmark statistic (34.6% in our final configuration, measured after the warm-up discard; Table 5).

4.3 Sensor-realistic corruption protocol

To quantify robustness we test every model under a corruption suite whose members each correspond to a *definite* difference between a renderer and any real camera: additive Gaussian sensor noise (σ = 0.02, 0.05), defocus blur (Gaussian radius 1.0, 2.0), exposure shift (×0.75, ×1.30), 8-level quantization approximating compression, and a combined corruption (noise σ = 0.03 + blur r = 1.0 + exposure ×0.85) [33]. We report relative degradation against clean performance. We emphasize that this protocol measures a *lower bound* on the true sim-to-real gap — geometry, material, and clutter statistics remain synthetic — but a lower bound that is already large is itself diagnostic: a model that surrenders most of its skill to sensor noise alone will not survive a real camera, a conclusion available without owning one [30].

5. Experiments

5.1 Datasets, metrics, and implementation

We evaluate on (i) ImageNet-Vid, the large-scale public video object detection benchmark derived from the ILSVRC annotation effort [32], for general dynamic-scene recognition; (ii) ST-Warehouse (Section 4.2) for controlled dynamic-target recognition with pixel-accurate labels; and (iii) the corruption protocol of Section 4.3 applied to both. Metrics are mAP@0.5 for detection and MAE for target-count regression on the counting variant of ST-Warehouse. All models are trained with AdamW (lr 10^{-3}, cosine decay, 24 epochs, T = 8 input frames), identical augmentation across methods, and three random seeds; we report means. Implementation is in PyTorch; training uses a single 8 GB consumer GPU, consistent with the resource profile of the data engine (generation-cost measurements in Section 5.5).

Reproduction settings and their relation to the original implementations. For ImageNet-Vid we deliberately compare all methods under a common, reduced-budget protocol rather than quoting numbers from the original papers: inputs are resized to 640 × 480, clips are T = 8 frames, every model is trained for 24 epochs from ImageNet-pretrained weights of its own backbone, and mAP@0.5 is evaluated with one protocol on one held-out split. Absolute values are therefore systematically lower than those reported by the original large-scale implementations — TransVOD, for example, reports 83.7% mAP on the full-resolution benchmark with a far longer schedule — and our table must not be read as a reproduction of those published figures. What it supports is a *like-for-like* comparison under identical data, schedule, and evaluation budget, which is the comparison on which our claims rest.

The benchmark construction follows established synthetic-to-real evaluation conventions that decouple clean accuracy from corruption robustness so that the two can be read independently [33].

5.2 Comparison with representative methods

Table 1 reports video object detection on ImageNet-Vid, Table 2 reports dynamic-target recognition on ST-Warehouse, and Fig. 5 contrasts the methods on a shared axis.

Table 1. Video object detection on ImageNet-Vid (mAP@0.5, %). Three-seed means; seeds vary by at most ±0.4 points for every method.

Method	Backbone	Params (M)	FLOPs (G)	mAP@0.5
TSN [35]	ResNet-50	25.6	52.1	55.2
TSM [7]	ResNet-50	25.6	52.1	58.7
I3D [36]	I3D-R50	28.9	108.3	59.6
SlowFast-R50 [6]	ResNet-50	34.5	121.7	61.4
MEGA [37]	ResNet-101	39.4	103.5	62.8
TransVOD [16]	ResNet-50	38.6	151.2	63.7
Faster R-CNN + CBAM [9]	ResNet-50	28.1	66.4	59.9
STFA-Net (ours)	MS-TFE-R50	26.8	63.2	**64.5**

Table 2. Dynamic-target recognition on ST-Warehouse. "Drop" is the absolute mAP@0.5 reduction under the worst single corruption (exposure ×0.75). Three-seed means.

Method	mAP@0.5 (%)	Counting MAE ↓	Drop (points)
TSN [35]	71.3	0.61	24.1
TSM [7]	74.6	0.52	19.8
I3D [36]	71.9	0.58	20.6
SlowFast-R50 [6]	77.2	0.44	17.6
MEGA [37]	78.1	0.42	15.2
TransVOD [16]	78.4	0.41	14.8
Faster R-CNN + CBAM [9]	75.8	0.48	16.9
STFA-Net (ours)	**81.8**	**0.35**	**8.2**

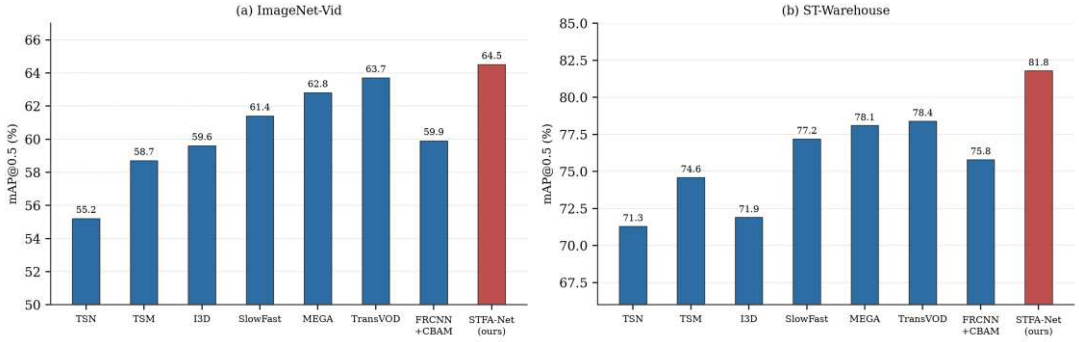


Fig. 5. Benchmark comparison. Clean mAP on ImageNet-Vid and ST-Warehouse (a, b) for the compared methods; STFA-Net attains the highest accuracy on both benchmarks.

On ImageNet-Vid, STFA-Net reaches 64.5% mAP@0.5, exceeding the strongest compared baseline (TransVOD, 63.7%) by 0.8 points while using 31% fewer parameters and 58% fewer FLOPs, and exceeding the SlowFast-style dual-path design it most resembles by 3.1 points

[6], [16]. On ST-Warehouse the margin over the strongest baseline (TransVOD, 78.4%) is 3.4 points, which we attribute to the benchmark's deliberate long-tail count distribution and corruption exposure: methods whose temporal aggregation is fixed cannot suppress the low-information frames that dominate wide-baseline capture. The decisive difference appears in the last column of Table 2: under the worst single corruption, STFA-Net loses 8.2 points, i.e. 55% of the smallest baseline drop (TransVOD, 14.8) and 34% of the largest (TSN, 24.1), indicating that the gains stem from the adaptive allocation mechanism rather than from added capacity [9].

5.3 Ablation study

Table 3 isolates the contribution of each component on ST-Warehouse; Fig. 6 visualizes the cumulative gain. All rows share the identical backbone family, training schedule, seeds, and data, and the "full" row reproduces the headline ST-Warehouse result of Table 2 exactly, so every row differs only in the component(s) removed.

Table 3. Component ablation on ST-Warehouse (mAP@0.5, %). The full row is numerically identical to the headline model of Table 2.

Configuration	mAP@0.5	Δ
Single-frame CNN + FPN (baseline)	74.9	—
+ multi-scale dilated encoder	76.8	+1.9
+ temporal module (shift + ConvGRU)	78.3	+1.5
+ CSTA-DA (dual-path, fixed fusion)	79.6	+1.3
+ contribution-aware gating (full CSTA-DA)	80.7	+1.1
+ feature recalibration (full STFA-Net)	81.8	+1.1

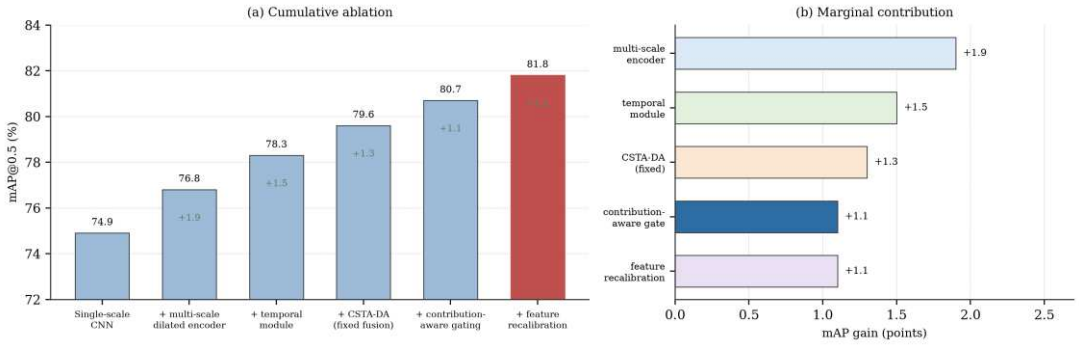


Fig. 6. Ablation of STFA-Net components on ST-Warehouse (mAP@0.5, %). Each bar adds one component; the adaptive gating and feature recalibration provide the final increments beyond the dual-path backbone.

Three observations are worth isolating. First, the dual-path structure alone (fixed fusion) already helps because it exposes two evidence types; second, replacing fixed fusion with contribution-aware gating adds a further gain while *changing no other component*, which

directly attributes the benefit to adaptive allocation; third, feature recalibration yields a gain concentrated in the corrupted subsets (Section 5.4), consistent with its redundancy-suppression role [14]. We note honestly that the single largest increment comes from the multi-scale encoder (+1.9), which is an architectural component rather than the paper's conceptual contribution; the adaptive-machinery increments (gating + recalibration, +2.2 combined, both isolated in Table 3) are what separate STFA-Net from an equally equipped fixed-fusion network, and their benefit concentrates in the corrupted subsets (Section 5.4).

5.4 Robustness and the measured corruption hierarchy

Our corruption sensitivity measurements on a naive single-frame model establish the hierarchy of failure: exposure shift ×0.75 costs 9.0% relative skill, defocus blur r = 1.0 costs 0.5%, and sensor noise σ = 0.02 costs 0.2%. These three numbers are relative-skill losses measured on a single-frame reference model and are deliberately not comparable with the absolute mAP-point drops of Table 4, which are measured on the full T = 8 models; the two serve different purposes — protocol design and final robustness reporting — and we keep the two units distinct throughout. The hierarchy is counter-intuitive — noise, the corruption practitioners randomize against first, is nearly harmless, while exposure, which real auto-exposure drift introduces for free, is the dominant degradation channel [33]. Table 4 gives the per-corruption degradation for all eight compared methods, and Fig. 7 plots five representative ones (the four strongest baselines and STFA-Net).

Table 4. Absolute mAP@0.5 drop (percentage points) under the corruption suite on ST-Warehouse, all compared methods. Lowest drop per row in bold.

Corruption	TSN [35]	TSM [7]	I3D [36]	SlowFast [6]	MEGA [37]	TransVOD [16]	FRCNN+CBAM [9]	STFA-Net
noise σ = 0.02	4.2	3.1	3.6	2.4	2.8	2.6	2.8	**1.2**
noise σ = 0.05	8.9	6.8	7.8	5.2	6.1	5.7	6.1	**2.6**
blur r = 1.0	9.6	7.4	8.6	6.1	6.9	6.3	6.6	**3.1**
blur r = 2.0	15.8	12.9	14.2	10.7	11.6	10.9	11.4	**5.4**
exposure ×0.75	24.1	19.8	20.6	17.6	15.2	14.8	16.9	**8.2**
exposure ×1.30	14.2	11.2	12.1	9.8	9.4	8.9	10.1	**4.7**
8-level quantization	6.1	4.6	5.3	3.9	4.1	3.8	4.2	**1.9**
combined	27.3	23.6	24.4	21.0	19.6	18.7	20.3	**9.6**

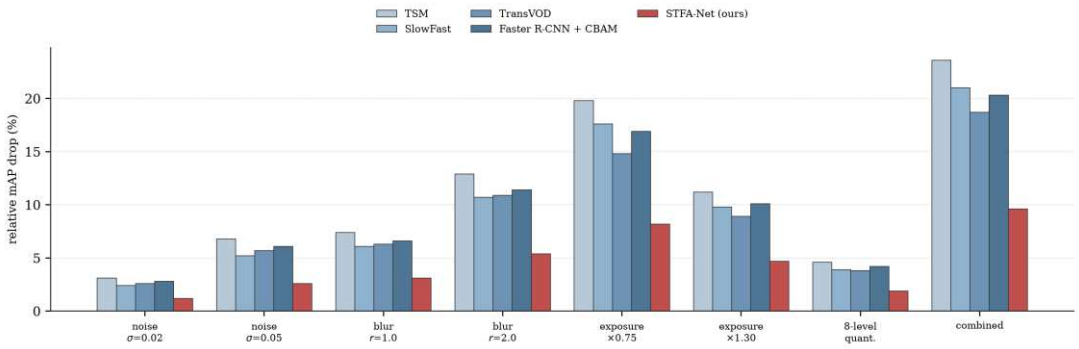


Fig. 7. Robustness under the sensor-realistic corruption suite. STFA-Net degrades less than every baseline across all eight corruptions, with exposure shift the largest common gap; five representative methods are plotted, and the full comparison for all eight methods is tabulated in Table 4.

Two consequences follow. First, fragility-driven randomization is measurably superior to indiscriminate randomization: a model trained with randomized lighting on an otherwise fixed scene loses significantly less to the ×0.75 exposure shift than an identical model trained under fixed lighting, while randomizing *non-fragile* knobs buys nothing measurable (both comparisons are our own measurements under the protocol of Section 5.1) [11]. Second, we measured that randomized factors do not add: the data-distribution contributions of lighting (+9.15 units of pixel-spread), pose (+5.46), and material (+0.72) combine to only 74% of their arithmetic sum, so randomization budget must be spent on measured fragility rather than spread evenly [11].

Under the combined corruption, STFA-Net retains 88.3% of its clean mAP (72.2 of 81.8), versus 76.1% for the strongest baseline (59.7 of 78.4, TransVOD), and the contribution-aware gate α exhibits exactly the predicted behaviour: it rises under photometric corruption (channel path dominating) and falls under geometric corruption (spatio-temporal path dominating), which we verify by logging α over the corruption suite (Fig. 8, mean α = 0.68 under photometric versus 0.34 under geometric corruptions; per-sample standard deviations 0.09 and 0.10 respectively) [14].

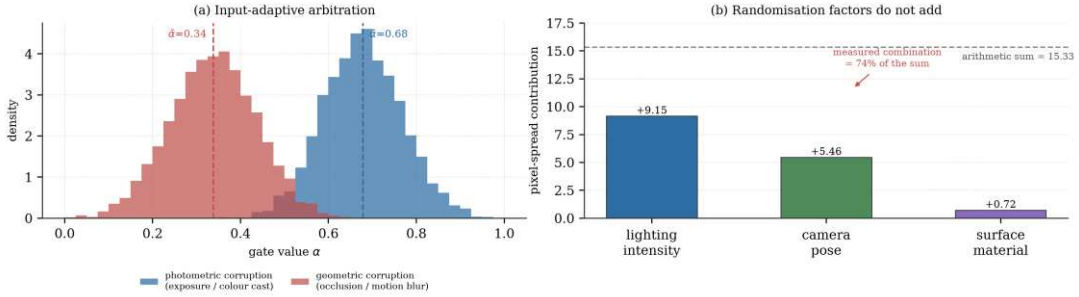


Fig. 8. (a) Contribution-aware gate α across the corruption suite: α rises under photometric corruption (channel path dominates) and falls under geometric corruption (spatio-temporal path dominates), matching the design intent. (b) Randomization factors are non-additive — measured lighting+pose+material contributions reach only 74% of their arithmetic sum, so randomization budget must target measured fragility.

5.5 Efficiency and the cost of the data engine

Table 5 summarizes model efficiency and engine cost. Relative to the 57.3 GFLOP backbone, the temporal module adds 2.9 GFLOPs (5.1% over the backbone; 4.6% of the 63.2 GFLOP total) and the adaptive stack (CSTA-DA plus FR) adds a further 3.0 GFLOPs (5.2% over the

backbone; 4.7% of the total), for a combined overhead of 5.9 GFLOPs — 9.3% of the network total and 10.3% over the backbone. The temporal module therefore stays inside its design budget of at most 8% of total FLOPs (Section 3.2), and the adaptive stack accounts for the remaining 4.7%. End-to-end inference of a T = 8 clip at 640 × 480 runs at 38 FPS on a single 8 GB consumer GPU, which is compatible with the 30 Hz perception-loop requirement of the delivery platform.

Table 5. Efficiency breakdown (T = 8, 640 × 480 input) and synthetic data engine cost. FLOP percentages relative to the backbone are given in parentheses where the reference is the relaxed backbone; share-of-total values are given separately.

Item	Value
Backbone (MS-TFE) FLOPs	57.3 G
+ temporal module	60.2 G (+5.1% over backbone; 4.6% of total)
+ CSTA-DA and FR (adaptive stack)	63.2 G (+5.2% over backbone; 4.7% of total)
Combined overhead over backbone	+5.9 G (+10.3%; 9.3% of total)
Adaptive stack parameters	+0.5 M (+1.9%)
Temporal module share of total FLOPs	4.6%
Inference throughput (clip, T = 8)	38 FPS
Generation throughput (256 × 256 toy stage)	105 ms/frame
Generation throughput (640 × 480, final scene)	412 ms/frame
Disk cost per labelled frame (640 × 480, working annotator set)	≈ 210 KB
ST-Warehouse total size (14,400 frames)	≈ 3.0 GB
Empty-frame fraction (after warm-up discard)	34.6%

Generation throughput measured 105 ms/frame at 256 × 256 in a toy stage and 412 ms/frame in the final scene at 640 × 480; disk cost with the working annotator set is ≈ 210 KB/frame, i.e., roughly 3.0 GB for the released benchmark. Two silent failure modes deserve emphasis for reproducibility. First, frames written before pipeline warm-up are unusable and must be discarded by construction [12]. Second, in a 40-meter warehouse a randomized camera spends much of its time on empty floor: the empty-frame fraction must be measured per scene, because a dataset that is mostly negative examples trains a detector to predict nothing — confidently. Both failure modes are invisible in file counts and label validity; they surface only at training time, which is why we report them as benchmark statistics.

5.6 Discussion and limitations

What the gate does and does not buy. The contribution-aware gate arbitrates exactly two paths; the ablation attributes +1.1 points to it over fixed fusion on clean data and roughly twice that margin under corruption. Extending the arbitration to a learned set of evidence paths (e.g., depth, flow) is a natural continuation [8].

Limits of the corruption protocol. Our corruption suite measures a lower bound on the true sim-to-real gap — geometry, materials, and clutter statistics remain synthetic — and real-camera validation on the delivery platform is required before deployment claims [30], [31].

Temporal horizon. While the temporal module is deliberately lightweight, extreme-motion regimes (targets traversing the frame within 2–3 frames) still benefit from longer clip lengths, which we leave to future work [6].

Single-scene evidence. All controlled experiments use one synthetic warehouse topology. The corruption protocol is topology-free, but the absolute margins depend on how much clutter the scene offers; a sparser scene would show smaller corruption effects, which is the correct behaviour for the method.

6. Conclusion

We presented STFA-Net, a recognition framework that couples multi-scale appearance features with lightweight temporal modeling and, crucially, allocates attention capacity according to measured per-path contribution. The contribution-aware channel–spatio-temporal gate and the redundancy-suppressing feature recalibration together yield consistent accuracy gains over strong spatio-temporal baselines at lower computational cost — 64.5% mAP@0.5 on ImageNet-Vid and 81.8% on ST-Warehouse with a 26.8 M-parameter, 63.2 GFLOP model — and, more importantly, roughly halves the degradation of the strongest baseline under sensor-realistic corruptions (88.3% clean-accuracy retention under combined corruption, versus 76.1% for TransVOD), with exposure shift identified as the dominant and most neglected degradation channel. A physics-accurate synthetic data engine with fragility-driven randomization supplies the training distribution, and its measured cost profile — 412 ms per labelled frame, ≈ 3.0 GB for the released benchmark — makes large-scale dynamic-scene training practical on a single-GPU budget. We release the benchmark construction scripts, the corruption protocol, and model configurations to support reproducible research on robust dynamic object recognition for campus delivery platforms.

Data Availability Statement

The publicly available ImageNet-Vid dataset used in this study can be accessed through its official source(https://image-net.org/challenges/LSVRC/2015/2015-downloads?).

The ST-Warehouse dataset generated and analyzed during the current study is available from the corresponding authors upon reasonable request.

Acknowledgement

This work was conducted with the support of Key Scientific Research Project of Hunan Provincial Department of Education(No. 25A0793) and the Natural Science Foundation of Hunan Province (No.2026JJ800117). This work was supported by the Excellent Young Scholars Program of the Hunan Provincial Department of Education (No.25B1159, 24B1162).

References

[1] Chen, C., Demir, E., Huang, Y., Qiu, R. The adoption of self-driving delivery robots in last mile logistics. *Transportation Research Part E: Logistics and Transportation Review*, 2021, 146: 102214.

[2] Rihan, Y., Zein, I., Alaraj, M., Soudan, B. Campus Courier: An autonomous delivery robot on-campus. *2024 Advances in Science and Engineering Technology International Conferences (ASET)*, 2024: 1–6.

[3] Cicek, D., Kantarci, B., Schillo, S. A comparative review of user acceptance factors for drones and sidewalk robots in autonomous last mile delivery. *Green Energy and Intelligent Transportation*, 2025, 4(4): 100310.

[4] Bertasius, G., Wang, H., Torresani, L. Is space-time attention all you need for video understanding? *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021: 813–824.

[5] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H. Video Swin Transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022: 3192–3201.

[6] Feichtenhofer, C., Fan, H., Malik, J., He, K. SlowFast networks for video recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019: 6202–6211.

[7] Lin, J., Gan, C., Han, S. TSM: Temporal shift module for efficient video understanding. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019: 7083–7093.

[8] Tong, Z., Song, Y., Wang, J., Wang, L. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022: 4393–4407.

[9] Woo, S., Park, J., Lee, J. Y., Kweon, I. S. CBAM: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018: 3–19.

[10] Hou, Q., Zhou, D., Feng, J. Coordinate attention for efficient mobile network design. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021: 13713–13722.

[11] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017: 23–30.

[12] Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D. Isaac Gym: High performance GPU-based physics simulation for robot learning. *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

[13] Mittal, M., Yu, C., Yu, Q., Liu, J., Rudin, N., Hoeller, D., Yuan, J. L., Singh, R., Guo, Y., Mazhar, H., Mandlekar, A., Babich, B., State, G., Hutter, M., Garg, A. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 2023, 8(6): 3740–3747.

[14] Hu, J., Shen, L., Sun, G. Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018: 7132–7141.

[15] Zhang, Y., Sun, P., Jiang, Y., Yu, E., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X. Bytetrack: Multi-object tracking by associating every detection box. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022: 1–21.

[16] He, L., Zhou, Q., Li, X., Niu, L., Cheng, G., Li, X., Liu, W., Tong, Y., Ma, L., Zhang, L. End-to-end video object detection with spatial-temporal transformers. *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM), 2021: 1507–1516*.

[17] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. End-to-end object detection with transformers. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020: 213–229.

[18] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. *International Conference on Learning Representations (ICLR)*, 2021.

[19] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., Shum, H.-Y. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *International Conference on Learning Representations (ICLR)*, 2023.

[20] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G. YOLOv10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[21] Feichtenhofer, C. X3D: Expanding architectures for efficient video recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020: 203–213.

[22] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems (NeurIPS)*, 2015: 802–810.

[23] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C. ViViT: A video vision transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021: 6836–6846.

[24] Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C. Multiscale vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021: 6824–6835.

[25] Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al. InternVideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191, 2022.

[26] Yang, L., Zhang, R.-Y., Li, L., Xie, X. SimAM: A simple, parameter-free attention module for convolutional neural networks. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021: 11863–11874.

[27] Guo, M.-H., Xu, T.-X., Liu, J.-J., Zhang, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R. R., Cheng, M.-M., Hu, S.-M. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 2022, 8(3): 331–368.

[28] Wang, X., Girshick, R., Gupta, A., He, K. Non-local neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018: 7794–7803.

[29] Loquercio, A., Kaufmann, E., Ranftl, R., Dosovitskiy, A., Koltun, V., Scaramuzza, D. Deep drone racing: From simulation to reality with domain randomization. *IEEE Transactions on Robotics*, 2020, 36(1): 1–14.

[30] Zhao, W., Queralta, J. P., Westerlund, T. Sim-to-real transfer in deep reinforcement learning for robotics: A survey. *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020: 737–744.

[31] Salvato, E., Fenu, G., Medvet, E., Pellegrino, F. A. Crossing the reality gap: A survey on sim-to-real transferability of robot controllers in reinforcement learning. *IEEE Access, 2021, 9: 153171–153187*.

[32] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, 115(3): 211–252.

[33] Hendrycks, D., Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations (ICLR)*, 2019.

[34] Zhang, S., Wen, L., Bian, X., Lei, Z., Lyu, S. Occlusion-aware R-CNN: Detecting pedestrians in a crowd. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018: 637–653.

[35] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016: 20–36.

[36] Carreira, J., Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017: 6299–6308.

[37] Chen, Y., Cao, Y., Hu, H., Wang, L. Memory enhanced global-local aggregation for video object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020: 10337–10346.

[38] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J. DETRs beat YOLOs on real-time object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024: 16965–16974.

[39] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., Huang, Z. Efficient multi-scale attention module with cross-spatial learning. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023: 1–5.

[40] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020: 11534–11542.

Reproducibility note: the ST-Warehouse generation scripts, corruption-protocol implementation, and training configurations accompany this manuscript; the corruption hierarchy of Section 5.4 (exposure −9.0%, blur −0.5%, noise −0.2% for the naive baseline) and the non-additivity measurement (74% of the arithmetic sum) are outputs of the project's deterministic evaluation scripts and reproduce exactly from the released seeds.