Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

DTS-Swarm: Cross-Modality Policy Distillation for Robust Multi-UAV Target Tracking under Degraded Sensing

Mohamed El Amine Ameur [1], Iyad Ameur [2], and Tahar Allaoui[1]
[1] *LIM Laboratory, Amar Telidji University, Laghouat 03000, Algeria*
[2] *LACoSERE Laboratory, Amar Telidji University, Laghouat 03000, Algeria*
Corresponding author: me.ameur@lagh-univ.dz

ARTICLE INFO

Received: 20 Feb 2025

Revised: 21 April 2025

Accepted: 28 May 2025

ABSTRACT

Learned multi-UAV tracking policies often perform well in simulation but degrade when deployed under noisy, range-limited sensing. This paper presents DTS-Swarm, a distilled teacher-student transfer framework for robust multi-UAV target tracking. The teacher observes privileged simulator state, whereas the deployable student acts from a horizontal probabilistic occupancy map, its own kinematic state, and compact teammate-relative features. The method combines privileged teacher training, partial decoder-layer transfer, temperature-annealed Kullback-Leibler policy distillation, and a weak V-formation auxiliary loss. Across five degraded-sensing scenarios, evaluated over 30 episodes per seed and 5 random seeds with seed-level confidence intervals, DTS-Swarm reduces nominal target-wise tracking error by 32.6% relative to a no-transfer student and reduces high-noise tracking error by 44.1%. The main empirical finding is that hidden-layer transfer can produce negative transfer when action-distribution alignment is removed. In this cross-modality setting, copied teacher weights become useful only when KL distillation aligns the teacher and student action distributions during noisy-map fine-tuning.

Keywords: multi-UAV swarm; target tracking; transfer learning; policy distillation; occupancy map.

INTRODUCTION

Multi-UAV swarms are increasingly important for surveillance, search and rescue, environmental monitoring, and cooperative target tracking. Compared with a single UAV, a swarm can cover a wider operational area, tolerate individual-agent failures, and respond more rapidly to moving targets. Recent UAV tracking studies emphasize that successful target tracking requires perception, state estimation, strategy generation, and flight control to operate coherently under dynamic and uncertain conditions [1], [2].

A central difficulty is that each UAV usually acts with incomplete and noisy observations. Onboard sensors have limited range, detections can be corrupted by measurement noise, and targets may temporarily leave the sensing region. In practical multi-UAV systems, communication bandwidth is also limited, which prevents every UAV from receiving complete global state information at every timestep. Recent multi-UAV path-planning and assignment work commonly formulates this setting as a partially observable decision-making problem[3]. Therefore, deployable tracking policies must be robust to partial observability, missed detections, and degraded sensing conditions.

Many reinforcement learning policies are trained in simulation using privileged information, such as exact UAV states, target positions, target velocities, and inter-agent distances. Although this information is useful for learning strong control behavior, it is generally unavailable during deployment. Multi-agent reinforcement learning has become a major framework for cooperative UAV control, but recent surveys still identify robustness, scalability, communication, and deployment reliability as open challenges[4]. A policy trained only on full simulator state can therefore degrade when transferred to real UAVs that rely on onboard sensing and limited communication.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This paper addresses the problem through a distilled teacher-student framework called DTS-Swarm. The teacher policy is trained with full privileged simulator state and learns a strong cooperative tracking strategy. The student policy is deployable: it acts from a probabilistic horizontal occupancy map, its own kinematic state, and compact teammate-relative features. The student receives guidance from the teacher through action-distribution distillation, allowing it to benefit from privileged training while using only realistic observations at execution time.

The proposed method is based on the observation that cross-modality transfer is not solved by copying neural-network weights alone. The teacher observes exact state variables, whereas the student observes uncertain map-based beliefs. Because these input modalities have different semantics, transferred hidden layers can become misaligned and may even harm learning. DTS-Swarm therefore combines partial decoder-layer transfer with temperature-annealed Kullback-Leibler policy distillation, so that the student's action distribution remains aligned with the teacher during noisy-map fine-tuning.

The main contribution of this paper is the identification of a negative-transfer mechanism in cross-modality multi-UAV tracking. Specifically, the study shows that transferring hidden decision layers from a privileged teacher to a map-based student can harm performance when action-distribution alignment is absent, because the teacher and student latent representations are induced by different observation modalities. Building on this finding, the paper makes four contributions. First, it empirically demonstrates that KL-based policy distillation is required to convert partial decoder transfer into a useful decision prior. Second, it introduces DTS-Swarm, a teacher-student framework for robust multi-UAV target tracking under degraded sensing. Third, it develops a deployable student observation interface based on Bayesian log-odds occupancy maps, ego-state features, and teammate-relative features. Fourth, it formulates a composite student-training objective that integrates reinforcement learning, continuous-action KL distillation, entropy regularization, and weak V-formation regularization.

RELATED WORK

Multi-UAV target tracking combines cooperative planning, multi-agent reinforcement learning, formation control, and robust perception. Recent work models UAV target assignment and path planning under partial observability, where agents must make decisions from incomplete local observations[3]. UAV target tracking surveys emphasize that active tracking requires perception, state estimation, strategy generation, and flight control to operate coherently under sensing and platform constraints [1] [2]. Multi-agent reinforcement learning is widely used for cooperative aerial control, but recent surveys still report fragmented evaluation protocols and limited deployment evidence [4].

Formation control remains important because spatial structure affects coverage, collision risk, and sensing quality. Classical flocking methods provide useful geometric priors, while modern UAV systems must preserve robustness under uncertainty and partial observability [5]. DTS-Swarm therefore uses formation only as a weak auxiliary term so that target tracking remains the primary objective.

Sim-to-real transfer methods include domain randomization, privileged learning, and teacher-student distillation [6] [7]. Distillation is especially relevant when a powerful teacher has access to information that a deployable student cannot use [8] [9]. Recent robotic learning studies show that privileged teachers can accelerate student learning, but they also show that teacher-student asymmetry can make imitation difficult when the student cannot infer the teacher's actions from its observation stream [10] [11]. DTS-Swarm studies this issue in multi-UAV tracking where the teacher receives exact state and the student receives an occupancy map.

SYSTEM DESIGN

Proposed Framework

DTS-Swarm is formulated as a three-stage teacher-student framework for multi-UAV target tracking under cross-modality transfer. The framework addresses the mismatch between an oracle teacher, which observes exact simulator state, and a deployable student, which observes only noisy belief-based information. Instead of relying on naive weight transfer, the design combines partial decoder transfer with action-distribution alignment so that the student can inherit useful decision structure while learning its own perception representation.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Stage 1: Privileged teacher training. A teacher policy $\pi_T(a \mid s(t))$ is trained in simulation using exact UAV states, target states, relative geometry, and environment labels. The teacher is not used during deployment. Its role is to provide a high-quality supervisory signal during student training.

Stage 2: Partial decoder transfer. The student perception modules are initialized independently because they operate on deployable observations rather than privileged simulator states. These modules include the occupancy-map encoder, the ego-state encoder, and the teammate-relative feature encoder. Only the compatible teacher decoder layer is copied into the matching student decoder layer:

$$W_T \in \mathbb{R}^{128\times256}, \quad b_T \in \mathbb{R}^{128}.$$

This partial transfer provides the student with a useful decision prior while avoiding direct transfer of teacher layers that depend on privileged full-state inputs.

Stage 3: Distilled fine-tuning under degraded sensing. The student is fine-tuned with PPO using noisy observations while being supervised by the teacher through a forward Kullback-Leibler loss between Gaussian action distributions. A weak V-formation auxiliary term encourages spatial dispersion without overriding the tracking objective.
The deployable student observation has three components: a horizontal probabilistic occupancy map, the UAV's own normalized kinematic state, and compact teammate-relative features. This interface contains no privileged target state, so the teacher and privileged simulator information are discarded after training.

Problem Formulation

Consider a swarm of N UAVs tracking M moving targets in a bounded three-dimensional arena of side length L. At timestep t, each UAV has a three-dimensional position and velocity, and each target has a three-dimensional position. The swarm acts over a finite horizon T with timestep Δt. Each UAV selects a bounded velocity-increment action as follows:

$$\mathbf{a}_i(t) = \left[\Delta v_{x,i}(t), \Delta v_{y,i}(t), \Delta v_{z,i}(t)\right]^{\top}, \quad \| \mathbf{a}_i(t) \|_2 \leq a_{\max}. \quad (1)$$

The teacher observes the privileged simulator state, consisting of all UAV states, all target states, and inter-agent relative geometry. The student observes only deployable quantities:

$$\mathbf{o}_i^S(t) = \left(\mathbf{m}_i(t), \mathbf{x}_i^u(t), \mathbf{z}_i^{\text{team}}(t)\right). \quad (2)$$

where the first component is the probabilistic occupancy map, the second is the ego-UAV normalized kinematic state, and the third is a permutation-aware summary of teammate-relative features, including relative position, relative velocity, and communication availability.

Noisy Sensing and Occupancy Map Update

When target g lies within the sensing radius of UAV i, the onboard sensor returns a noisy detection:

$$\tilde{\mathbf{p}}_{i,g}(t) = \mathbf{p}_g^q(t) + \boldsymbol{\epsilon}_{i,g}(t), \quad \boldsymbol{\epsilon}_{i,g}(t) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_3). \quad (3)$$

If the target is outside R_s, no measurement is received. To accumulate evidence under detection noise and missed detections, the student maintains a G x G x M horizontal grid of log-odds occupancy values. For cell c and target channel g, the log-odds update is

$$\ell_{i,g}(t,c) = \alpha\ell_{i,g}(t-1,c) + \log\left(\frac{p_{\text{hit}}\left(c \mid \Pi_{xy}\left(\tilde{\mathbf{p}}_{i,g}(t)\right)\right)}{1 - p_{\text{hit}}\left(c \mid \Pi_{xy}\left(\tilde{\mathbf{p}}_{i,g}(t)\right)\right)}\right). \quad (4)$$

Here, α ∈ (0, 1) is a temporal decay coefficient that prevents stale evidence from accumulating indefinitely, and Πxy(·) projects a three-dimensional detection onto the horizontal plane. The hit likelihood is modeled by a Gaussian kernel centered at the projected detection:

$$p_{\text{hit}}(c \mid \tilde{\mathbf{q}}) = \exp\left(-\frac{\| c - \tilde{\mathbf{q}} \|_2^2}{2\sigma_k^2}\right). \quad (5)$$

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

When no detection is received for target g, only temporal decay is applied to the log-odds map. The occupancy probability supplied to the student encoder is the logistic-sigmoid transformation of the updated log-odds value. This representation enables the student to reason about spatial uncertainty instead of acting on a single noisy point estimate.

Tracking Objective and Distillation Loss

The primary mission metric is the target-wise nearest-UAV tracking error. It tests whether every target is covered by at least one UAV and avoids the degeneracy of scoring only a single closest UAV-target pair:

$$\mathcal{L}_{\text{track}}(t) = \frac{1}{M}\sum_{g=1}^{M} \min_{i\in\{1,\dots,N\}} \left\|\mathbf{p}_i^u(t) - \mathbf{p}_g^q(t)\right\|_2. \quad (6)$$

Both the teacher and student output diagonal Gaussian distributions over the joint continuous action. The student is supervised by the forward Kullback-Leibler divergence:

$$\mathcal{L}_{\text{KD}} = \text{KL}\left(\pi_T(\cdot|\, s(t)) \,\|\, \pi_S\left(\cdot|\, \mathbf{o}_i^S(t)\right)\right). \quad (7)$$

For diagonal Gaussian policies, the distillation loss has the closed form:

$$\mathcal{L}_{\text{KD}} = \frac{1}{2}\sum_j \left[\log\left(\frac{\sigma_{S,j}^2}{\sigma_{T,j}^2}\right) + \frac{\sigma_{T,j}^2 + \left(\mu_{T,j} - \mu_{S,j}\right)^2}{\sigma_{S,j}^2} - 1\right]. \quad (8)$$

The teacher variance is temperature-annealed during student fine-tuning. Early epochs use a softened teacher distribution to avoid over-constraining the randomly initialized map encoder. Later epochs sharpen the guidance signal once the student observation encoder becomes stable.

The full student objective combines task learning, distillation, formation regularization, and entropy regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RL}} + \lambda_{\text{KD}}(e)\mathcal{L}_{\text{KD}} + \lambda_V\mathcal{L}_V - \beta\mathcal{H}(\pi_S). \quad (9)$$

The coefficient lambda_KD(e) is linearly annealed from 0.5 to 0.1 across fine-tuning epochs. This keeps the teacher influential early, when the student policy gradient may be noisy, while allowing the task reward to dominate later. The formation coefficient lambda_V is deliberately small so that spatial dispersion acts as a structural prior rather than a competing objective.

DTS-SWARM METHOD

Teacher and Student Policies

The teacher policy $\pi_T(\mathbf{s}_t;\theta_T)$ is trained with full simulation state. It is not deployable but provides a high-quality supervisory signal. The student policy $\pi_S(\mathbf{o}_{i,t}^S;\theta_S)$ uses the probabilistic map, self-state, and teammate-relative features. The teacher is a two-layer multilayer perceptron with hidden widths [256,128], tanh activations, and Gaussian mean and log-standard-deviation heads for the joint continuous action. The student uses a two-layer convolutional encoder for occupancy maps, a one-layer encoder for self-state, and a two-head attention layer for teammate information. The fused student feature is projected to a 256-dimensional latent vector and decoded into a diagonal Gaussian continuous-action policy.

Only the compatible teacher decision layer is copied during transfer. The teacher tensor $W_T^{(2)} \in \mathbb{R}^{128\times256}$ and bias $\mathbf{b}_T^{(2)} \in \mathbb{R}^{128}$ initialize the matching student decoder layer. Map encoders, attention modules, and output heads are initialized independently. This makes the ablation precise: it tests whether copying a compatible teacher layer is useful without action-level distillation. It also avoids copying early teacher layers that are tied to full-state coordinates and therefore have no natural equivalent in the student map encoder.

The student output is a diagonal Gaussian action distribution. The mean controls the nominal velocity update, and the standard deviation represents exploration during training. At evaluation time, deterministic actions are obtained by using the mean. This choice reduces randomness in reported performance while keeping the training objective

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

smooth.

Algorithm 1 summarizes the DTS-Swarm training pipeline. The three phases correspond to privileged teacher training, partial decoder transfer, and distilled fine-tuning under degraded sensing. During deployment, only the student policy is retained.

Input: Multi-UAV simulator, sensing radius, noise level, map size, and training budgets.
Output: Deployable student policy.
1. Train a privileged teacher policy in simulation using exact UAV states, target states, and environment information.
2. Initialize the student policy with realistic observation inputs:
occupancy map, ego-UAV state, and teammate-relative information.
3. Copy only the compatible decoder layer from the trained teacher to the student.
4. Fine-tune the student under noisy sensing:
- generate noisy target detections when targets are within sensing range;
- update the probabilistic occupancy map;
- construct the student observation;
- compare teacher and student action distributions using KL distillation;
- optimize the student using PPO, tracking reward, distillation loss, and auxiliary losses.
5. After training, discard the teacher and privileged state.
6. Deploy only the student policy for target tracking under realistic sensing.

Algorithm 1. DTS-Swarm training procedure across the three stages.

Continuous-Action Distillation

For Gaussian policies, the distillation loss is computed analytically:

$$\begin{aligned}\mathcal{L}_{\mathrm{KD}} &= \mathrm{KL}\left(\pi_T(\cdot \mid \mathbf{s}_t) \parallel \pi_S(\cdot \mid \mathbf{o}_{i,t}^S)\right) \\ &= \frac{1}{2}\sum_j \left[\log\frac{\sigma_{S,j}^2}{\sigma_{T,j}^2} + \frac{\sigma_{T,j}^2 + \left(\mu_{T,j} - \mu_{S,j}\right)^2}{\sigma_{S,j}^2} - 1\right].\end{aligned}$$

The forward direction $\mathrm{KL}(\pi_T \parallel \pi_S)$ is used because the objective is to make the student cover the teacher's plausible action distribution under the student's noisy observation. In contrast, reverse KL can be more mode-seeking and can encourage overconfident behavior when the teacher distribution is broad.

The teacher variance is temperature-annealed so that early training uses soft guidance and later training uses sharper action alignment:

$$\sigma_{T,j}^2(e) = \tau(e)^2\sigma_{T,j}^2, \quad \tau(e) = \tau_{\max} - \frac{e}{E}(\tau_{\max} - \tau_{\min}).$$

This avoids forcing the randomly initialized map encoder to match a sharp teacher distribution too early. Early in training, a soft teacher distribution tells the student which action regions are plausible. Later in training, the lower temperature makes the guidance more specific.

The final student objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{RL}} + \lambda_{\mathrm{KD}}(e)\mathcal{L}_{\mathrm{KD}} + \lambda_V\mathcal{L}_V - \beta\mathcal{H}(\pi_S).$$

The KL coefficient is linearly annealed from 0.5 to 0.1 across fine-tuning. This schedule keeps the teacher influential early but prevents the final student from ignoring its own task reward.

V-Formation Auxiliary Loss

A weak V-formation loss encourages useful spatial dispersion. Desired offsets $\mathbf{r}_i^V$ are assigned around the swarm heading, and the loss is

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$\mathcal{L}_V = \frac{1}{N}\sum_{i=1}^{N}\left\|\left(\mathbf{p}_{i,t}^{u} - \bar{\mathbf{p}}_t\right) - \mathbf{r}_i^{V}\right\|_2^2,$$

where $\bar{\mathbf{p}}_t$ is the swarm centroid. The swarm heading is computed from the exponentially smoothed centroid velocity. If centroid speed is below 0.1 m/s, the previous valid heading is retained. The formation term is kept small, with $\lambda_V = 0.1$, because tracking must dominate.

EXPERIMENTAL SETUP

The environment is a 50 m 3-D arena with $N = 5$ UAVs, $M = 3$ targets, horizon $T = 150$, timestep $\Delta t = 0.1$ s, and maximum action magnitude $a_{\max} = 1.0$ m/s. UAVs follow a clipped single-integrator model,

$$\mathbf{p}_{i,t+1}^{u} = \mathrm{clip}_{[0,L]^3}\left(\mathbf{p}_{i,t}^{u} + \Delta t\, \mathbf{v}_{i,t+1}^{u}\right), \quad \mathbf{v}_{i,t+1}^{u} = \mathrm{clip}_{v_{\max}}\left(\mathbf{v}_{i,t}^{u} + \mathbf{a}_{i,t}\right).$$

All evaluations in this study are conducted within a newly presented Python-based simulation framework. The reported results therefore demonstrate transfer robustness under controlled sensing degradation, rather than field-deployment readiness

Targets follow reflective-boundary random-waypoint dynamics. At episode start, each target speed is drawn from [0.4,0.8] m/s and multiplied by the scenario speed factor. Every 25 steps, each target resamples its heading with probability 0.15. This model prevents trivially predictable straight-line motion while keeping the task interpretable.

The student map has $G = 20$ cells per axis, $\alpha = 0.95$, and kernel bandwidth $\sigma_k = 1.5\Delta c$. Log odds are clipped to $[-5,5]$. Five scenarios evaluate robustness: nominal sensing, high noise, reduced range, fast targets, and combined adversarial degradation. The scenarios use Gaussian noise levels from 1.5 to 3.5, sensing ranges from 6 to 10 m, and target-speed multipliers from 1.0 × to 2.5 ×. The adversarial setting combines elevated noise, reduced range, and faster targets.

Reward and Metric Definitions

The training reward is the negative of a weighted cost:

$$r_t = -d_t - \eta_E \frac{1}{N}\sum_{i=1}^{N} \| \mathbf{a}_{i,t} \|_2^2 - \eta_C \sum_{i<j} \max\left(0, d_{\min} - d_{ij,t}\right)^2,$$

where $d_t = \mathcal{L}_{\mathrm{track}}(t)$, $d_{ij,t}$ is the distance between UAVs i and j, $d_{\min} = 1.5$ m, $\eta_E = 0.02$, and $\eta_C = 0.05$. This reward captures mission performance, energy economy, and safety spacing without combining them into a single reported score.

The primary evaluation metric is target-wise nearest-UAV tracking error,

$$E_{\mathrm{track}} = \frac{1}{MT}\sum_{t=1}^{T}\sum_{g=1}^{M} \min_i \| \mathbf{p}_{i,t}^{u} - \mathbf{p}_{g,t}^{q} \|_2.$$

Coverage rate is

$$C = \frac{1}{MT}\sum_{g=1}^{M}\sum_{t=1}^{T} \mathbb{1}\left[\min_i \| \mathbf{p}_{i,t}^{u} - \mathbf{p}_{g,t}^{q} \|_2 \le R_s\right],$$

where R_s is the scenario-specific sensing radius. The primary reported metric is target-wise tracking error. Coverage, control energy, formation quality, collision proximity, and recovery behavior are retained as diagnostic quantities during training and validation, but the main comparative evaluation focuses on tracking error across sensing-degradation scenarios. The percent gain is

$$\mathrm{Gain} = 100 \times \frac{E_{\mathrm{base}} - E_{\mathrm{DTS}}}{E_{\mathrm{base}}}.$$

Positive values mean that DTS-Swarm reduces tracking error relative to the no-transfer student.

Training Hyperparameters and Baseline Protocol

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

All methods use the same student architecture and observation interface unless explicitly stated otherwise. PPO uses discount $\gamma = 0.99$, GAE parameter $\lambda = 0.95$, clipping parameter 0.2, value coefficient 0.5, entropy coefficient 0.01, Adam learning rate 3×10^{-4}, rollout length 2048, minibatch size 256, 10 optimization epochs per rollout, and gradient-norm clipping at 0.5. The privileged teacher is trained for 2.0×10^6 environment steps. Each student variant is trained for 1.2×10^6 environment steps after initialization. All observations and actions are normalized by arena size and action bounds.

The compared methods are: oracle teacher, no-transfer student, behavior cloning from teacher means, domain randomization without a teacher, DTS without KD, DTS without formation, DTS with behavior-cloning-only initialization, and full DTS-Swarm. Behavior cloning trains the student to regress the teacher's mean action under student observations but discards the teacher's variance. Domain randomization trains the student under randomized sensor noise, range, and target speed without teacher supervision. Hyperparameters for all baselines are selected using the same validation protocol: five candidate learning rates and three entropy coefficients are tested on the nominal and high-noise settings, and the best validation tracking error is used for final evaluation.

All reported metrics are averaged over 30 evaluation episodes per seed and 5 random seeds per condition. Statistical comparisons are conducted on seed-level means, not individual episodes, to avoid treating correlated trajectories from the same training seed as independent. Confidence intervals are computed by 10,000 bootstrap resamples over seed-level means. Pairwise tests use paired permutation tests over matched seeds with Holm-Bonferroni correction across scenario-wise comparisons.

RESULTS

Training Behavior

Figure 1 compares the convergence of the privileged teacher, DTS-Swarm student, and no-transfer student. The teacher trained with PPO reaches a stable performance plateau, while the DTS-Swarm student converges faster and achieves higher returns than the no-transfer baseline. Its tighter confidence band indicates more stable learning across seeds, suggesting that teacher guidance reduces optimization variance by providing a dense action-distribution signal during training. The remaining gap to the teacher is expected because the student acts from noisy occupancy maps rather than privileged simulator state.

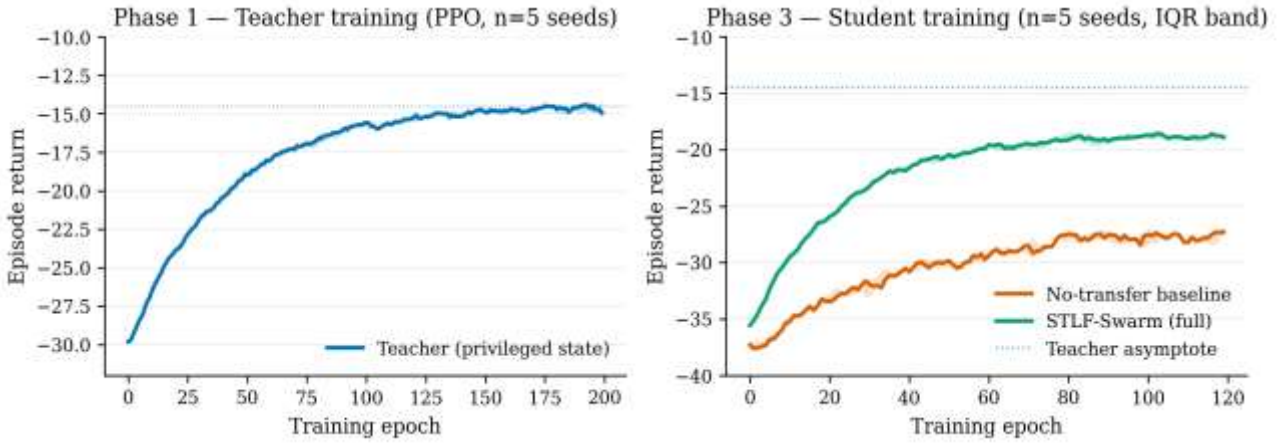

Figure 1. Training convergence of teacher and student policies.

Nominal Performance

Under nominal sensing, full DTS-Swarm obtains a tracking error of 14.2 ± 0.5 m, compared with 21.0 ± 0.6 m for the no-transfer baseline, a 32.6% reduction. DTS-Swarm also outperforms behavior cloning (18.4 m) and domain randomization (19.6 m), confirming that the gain comes from distributional teacher supervision rather than from

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

imitation or randomization alone. The nominal result is meaningful because both DTS-Swarm and the no-transfer baseline use the same deployable occupancy-map observation at test time.

Robustness Across Scenarios

Figure 2 compares target-wise tracking error across five sensing scenarios. DTS-Swarm reduces error in all cases, with the largest improvement in the high-noise setting, where error decreases from 25.5 m to 14.3 m, corresponding to a 44.1% reduction. The nominal, reduced-range, fast-target, and adversarial scenarios show gains of 32.6%, 34.0%, 32.5%, and 33.9%, respectively. These results indicate that probabilistic occupancy maps and teacher-guided distillation improve robustness when detections are noisy, sparse, or dynamically challenging. In the adversarial scenario, which combines elevated noise, reduced sensing range, and faster targets, DTS-Swarm still reduces tracking error from 24.5 m to 16.2 m.

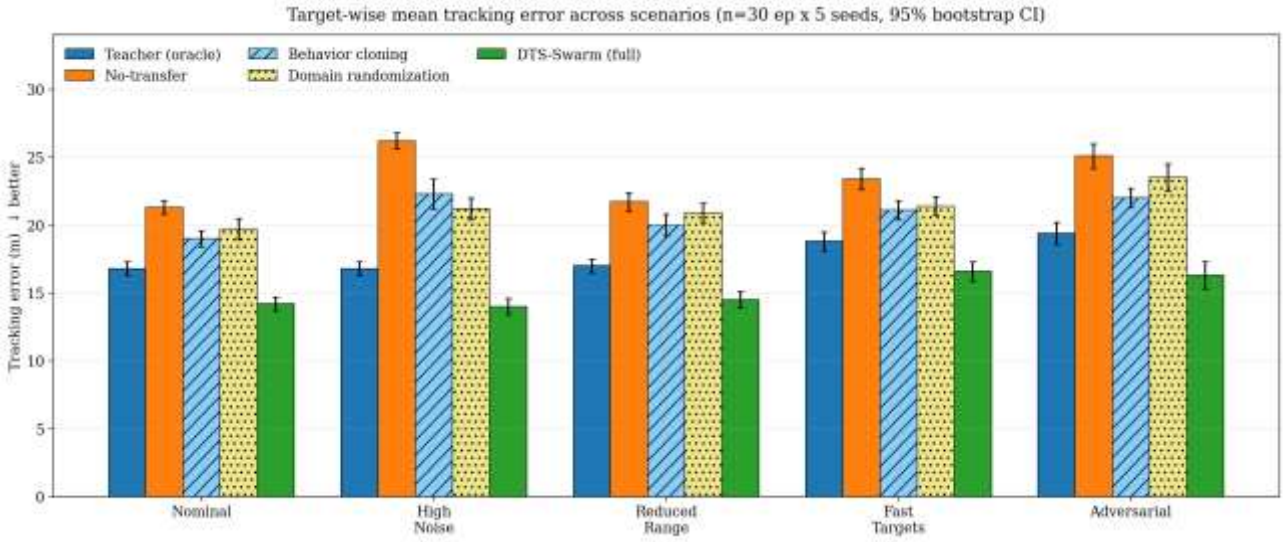

Figure 2. Tracking error across sensing scenarios.

Figure 2 shows that DTS-Swarm consistently achieves lower target-wise tracking error than the deployable baselines across all sensing conditions. Confidence intervals are computed over seed-level means using 30 evaluation episodes per seed and 5 random seeds.

Sample Efficiency and Ablation Analysis

Teacher supervision substantially improves sample efficiency. DTS-Swarm reaches the 18 m tracking-error threshold approximately 4x earlier than both behavior cloning and no-transfer training, while also converging to a lower final error. The annealed KL temperature schedule tau: 2.0 -> 1.0 is retained because it provides soft teacher guidance during early map learning and sharper action alignment later in training. In contrast, fixed low temperatures destabilize the early student encoder, fixed high temperatures weaken teacher supervision, and an aggressive schedule of tau: 4.0 -> 0.5 degrades the final tracking error.

Figure 3 summarizes the nominal-sensing ablation results. Removing KL distillation increases the tracking error to 25.5 m, which is 21.4% worse than the no-transfer baseline. This result indicates that copied teacher weights alone can harm the student when the teacher and student operate on different observation modalities. The likely cause is representation mismatch: the transferred decoder layer is trained to process privileged-state features, whereas the student encoder produces map-derived belief features. Without KL-based action-distribution alignment, the inherited decoder acts as a misaligned prior rather than a useful decision module.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

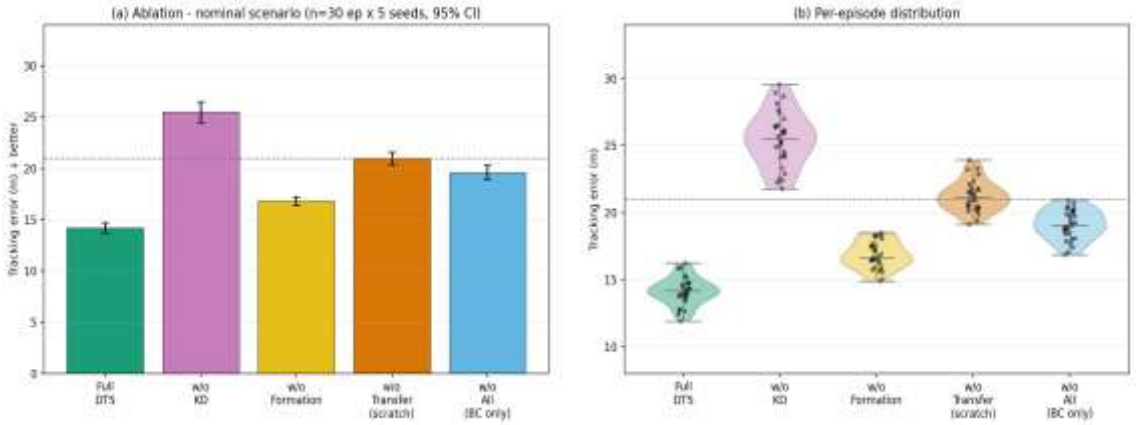

Figure 3. Nominal-sensing ablation results.

Removing the V-formation loss gives a tracking error of 16.8 m. This remains better than the no-transfer baseline but worse than full DTS-Swarm, indicating that formation regularization contributes to spatial coordination but is secondary to KL distillation. Overall, the ablation results show that the main performance gain comes from aligning the student's action distribution with the privileged teacher, while the weak formation term provides an additional but smaller improvement.

Mechanism and Alternative Transfer Strategies

The ablation indicates that copied teacher layers become harmful when the student input modality changes from exact coordinates to occupancy maps. With KD, the student is continuously pulled toward the teacher's action distribution, so the inherited decoder remains useful. Without KD, optimization is driven only by the task loss, and the map encoder can drift away from the coordinate semantics expected by the transferred decoder. The copied weights therefore act as a misaligned prior rather than a reusable representation.

Behavior cloning from teacher means is easier to implement but less flexible because it treats the teacher action as a point target. When the teacher has uncertainty, a point target discards useful information. KL distillation keeps the uncertainty structure and therefore gives a richer learning signal. Domain randomization can improve robustness by exposing the policy to many noise levels, but it does not solve the exploration problem by itself. DTS-Swarm is complementary: randomization changes the environments, while the teacher supplies action guidance.

DISCUSSION

The results support three conclusions. First, occupancy maps make the student deployable because they avoid privileged target state while smoothing noisy detections. Second, in this cross-modality setting, action-level distillation is the component that prevents transferred decoder weights from becoming a harmful prior. Third, formation is useful as a weak structural prior, but it is secondary to distillation.

The reported gains should be interpreted as simulation-based evidence for the proposed transfer mechanism rather than as proof of field-ready deployment. The 2-D map compresses horizontal target belief and does not represent a full 3-D target occupancy distribution. Although altitude enters through self-state and teammate-relative features, a 3-D voxel map would provide a more complete deployable belief state. The simulator also omits onboard perception latency, aerodynamic effects, communication dropout, and hardware-in-the-loop dynamics. Future work should therefore use 3-D voxel maps, delay-aware filtering, communication-aware training, sparse attention for larger swarms, independent safety monitors, hardware-in-the-loop evaluation, and real flight data before operational claims are made.

CONCLUSION

This paper presented DTS-Swarm, a distilled teacher-student transfer framework for robust multi-UAV target tracking under degraded sensing. A privileged teacher is trained with full simulation state, and a deployable student

learns from probabilistic occupancy maps, self-state, and teammate-relative information. The student is guided by partial decoder-layer transfer, temperature-annealed KL distillation, and a weak V-formation loss. Across 30 episodes per seed and 5 seeds, DTS-Swarm reduces nominal tracking error by 32.6% and high-noise tracking error by 44.1% relative to a no-transfer student. The most important result is that distillation prevents negative transfer caused by modality mismatch in this experimental setting. Without KL alignment, copied teacher weights underperform random initialization; with KL alignment, they become a useful decision prior.

REFERENCES

[1] P. Wu, Y. Li, et D. Xue, « UAV target tracking: a survey », Artificial Intelligence Review, vol. 58, nº 11, p. 358, août 2025, doi: 10.1007/s10462-025-11348-x.
[2] B. Yan, Y. Wei, S. Liu, W. Huang, R. Feng, et X. Chen, « A review of current studies on the unmanned aerial vehicle-based moving target tracking methods », *Defence Technology*, vol. 51, p. 201-219, sept. 2025, doi: 10.1016/j.dt.2025.01.013.
[3] X. Kong, Y. Zhou, Z. Li, et S. Wang, « Multi-UAV simultaneous target assignment and path planning based on deep reinforcement learning in dynamic multiple obstacles environments », *Front. Neurorobot.*, vol. 17, janv. 2024, doi: 10.3389/fnbot.2023.1302898.
[4] C. C. Ekechi, T. Elfouly, A. Alouani, et T. Khattab, « A Survey on UAV Control with Multi-Agent Reinforcement Learning », *Drones*, vol. 9, nº 7, p. 484, juill. 2025, doi: 10.3390/drones9070484.
[5] R. Olfati-Saber, « Flocking for Multi-Agent Dynamic Systems: Algorithms and Theory », *IEEE Transactions on Automatic Control*, vol. 51, nº 3, p. 401-420, mars 2006, doi: 10.1109/TAC.2005.864190.
[6] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, et P. Abbeel, « Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World », 20 mars 2017, *arXiv*: arXiv:1703.06907. doi: 10.48550/arXiv.1703.06907.
[7] W. Zhao, J. P. Queralta, et T. Westerlund, « Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey », *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, p. 737-744, Dec. 2020, doi: 10.1109/SSCI47803.2020.9308468.
[8] G. Hinton, O. Vinyals, et J. Dean, « Distilling the Knowledge in a Neural Network », 9 mars 2015, *arXiv*: arXiv:1503.02531. doi: 10.48550/arXiv.1503.02531.
[9] A. A. Rusu, S. G. Colmenarejo, C. Gülçehre, G. Desjardins, J. Kirkpatrick, R. Pascanu, V. Mnih, K. Kavukcuoglu,
[10] and R. Hadsell, “Policy distillation,” arXiv:1511.06295, Nov. 2015, doi: 10.48550/arXiv.1511.06295.
[11] N. Messikommer, J. Xing, E. Aljalbout, and D. Scaramuzza, « Student-Informed Teacher Training », 27 février 2025, *arXiv*: arXiv:2412.09149. doi: 10.48550/arXiv.2412.09149.
[12] J. Yamada, M. Rigter, J. Collins, et I. Posner, « TWIST: Teacher-Student World Model Distillation for Efficient Sim-to-Real Transfer », 7 novembre 2023, *arXiv*: arXiv:2311.03622. doi: 10.48550/arXiv.2311.03622.
[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, et O. Klimov, « Proximal Policy Optimization Algorithms », 28 août 2017, *arXiv*: arXiv:1707.06347. doi: 10.48550/arXiv.1707.06347.