LAST-METER PRECISION NAVIGATION FOR UAVS: A DIFFUSION-REFINED AERIAL VISUAL SERVOING APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we study the last-meter precision navigation for UAVs, e.g., autonomously reaching a target within the final 10 meters using monocular vision. This task is challenging due to scale ambiguity, rotation discontinuities, and the need for fine-grained spatial reasoning. Existing methods often fail under large viewpoint changes or lack generalization to unseen environments. To this end, we propose **DreamNav**, a coarse-to-fine diffusion-refined aerial visual servoing framework. In the first coarse-estimation stage, a robust regression policy employs a trigonometric parameterization to predict rotation by jointly modeling sine and cosine components, effectively mitigating optimization instabilities caused by angular periodicity. Given this coarse estimate, the second diffusion-refined stage utilizes a pre-trained world model to simulate future visual observations for candidate actions, selecting the trajectory that minimizes visual discrepancy with the target through a process of visual imagination. To support rigorous evaluation, we contribute **PairUAV**, a large-scale benchmark comprising 4.8 million image pairs across 72 scenes, curated from the University-1652 dataset. Extensive experiments show that DreamNav consistently improves coarse pose estimators built on different visual backbones and outperforms representative visual servoing, aerial localization, vision foundation, and generalist vision-language-action baselines, while maintaining strong zero-shot generalization to unseen scenes.

1 INTRODUCTION

The "last-meter" precision navigation scenario is critical for a wide range of real-world applications, including disaster search and rescue (patki et al., 2022; Rashid et al., 2020), swarm organization (Kramarić et al., 2025; Bu et al., 2024) and autonomous delivery (Hwang & Choe, 2019; Brunner et al., 2019). Despite the maturity of global navigation satellite systems (GNSS), their accuracy is typically limited to several meters and lacks the ability to resolve fine vertical distinctions or discriminate between closely spaced structures (e.g., windows or balconies)[1]. As a result, the primary challenge in aerial navigation has shifted from coarse global localization to fine-grained local positioning and precise terminal pose alignment (see Figure 1).

Despite recent advances in aerial navigation (Zhu et al., 2017; Deitke et al., 2020), two critical challenges remain unresolved in achieving robust last-meter navigation for UAVs. **(1)** There is a lack of large-scale datasets tailored to close-range aerial navigation, particularly those capturing fine-grained spatial variations at different distances and viewpoints. Existing benchmarks often focus on long-range localization or indoor flight, failing to model the visual dynamics encountered during terminal descent and alignment. **(2)** Accurately estimating subtle differences in relative pose remains highly challenging, especially under scale ambiguity (e.g., meter-level displacements) and rotational discontinuities (e.g., small angular deviations). This challenge is further exacerbated by varying lighting and viewpoint conditions, which demand robust, fine-grained spatial reasoning. This challenge, in turn, requires models to extract discriminative visual features and perform precise metric regression. Current approaches (Deuser et al., 2023) struggle under large viewpoint changes

[1] See official GPS performance standards(gps, 2020).

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.

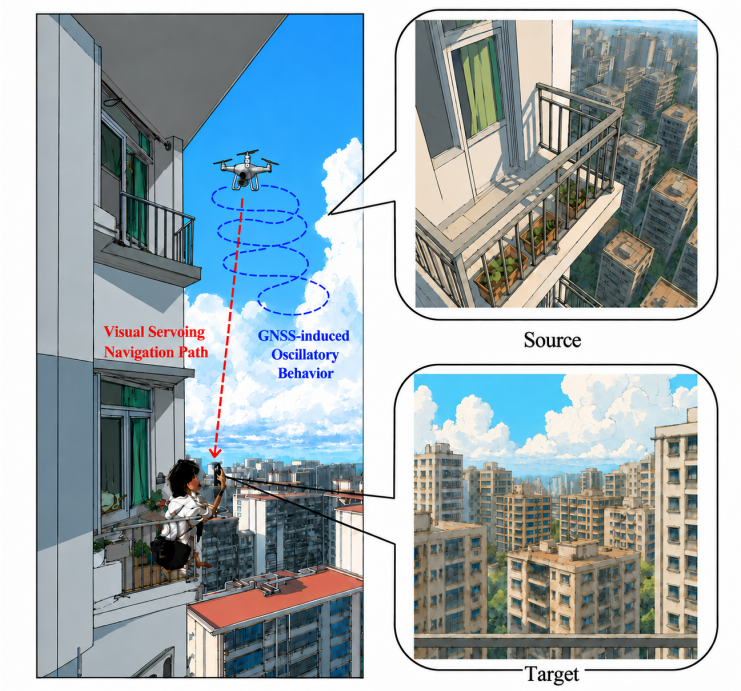

Figure 1: **Last-Meter Precision Navigation.** (*Left*) Many drone applications, such as package delivery, necessitate high-precision localization during the final approach, a phase where GNSS signals are often unreliable or unavailable, particularly for resolving vertical position. (*Right*) To overcome this challenge, we frame the task as a visual servoing problem, where the drone estimates the relative pose between its current view and a target image to guide its trajectory to the final destination based on visual refinement.

and exhibit poor generalization to unseen environments (see Section 5), primarily due to their reliance on direct regression without explicit modeling of visual dynamics.

(1) For the first challenge, we introduce PairUAV, the first large-scale open-world benchmark for image-driven UAV navigation. PairUAV spans 72 geographic locations and 1,652 building-centric targets, comprising over 4.8 million navigation instances that capture the semantic diversity and visual complexity of real environments. In each trial, the UAV is initialized at a source viewpoint and receives an egocentric RGB observation centered on the same scene-level visual reference as the target view. The dataset provides ground-truth relative heading and radial-range changes between the source and target viewpoints.

(2) For the second challenge, we argue that visual servoing offers a promising solution by enabling dynamic, feedback-driven refinement of positional accuracy. While visual servoing has been extensively studied in controlled environments, such as industrial automation and factory assembly systems (Zhu et al., 2017), and in robotic manipulation (Bar et al., 2025; Ren et al., 2025), its application to unmanned aerial vehicles (UAVs) in complex outdoor environments remains significantly underexplored. In particular, we propose DreamNav, a coarse-to-fine, diffusion-refined visual servoing framework. In Stage I, instead of regressing the rotation angle directly, we introduce a trigonometric loss that jointly constrains cosine and sine components to accurately regress the rotation angle, thereby avoiding angle-wrapping issues. Building on this loss, a DINO-R50-based backbone is fine-tuned to predict a rotation angle and the corresponding forward distance. In Stage II, we leverage a large-scale world model to imagine the visual outcome of each candidate action and select the one whose predicted view best matches the target image. Unlike direct prediction, which easily overfits to narrow, brittle cues (e.g., the edge of a square building) that may vanish or be misdetected in new scenes, imagination enforces image-wide consistency, requiring the policy to account for scene geometry, semantics, and viewpoint changes across the entire frame. This holistic constraint penalizes single-cue reliance, provides a built-in test-time self-verification signal, and thus improves robustness under distribution shift, yielding stronger generalization to unseen scenes.

To enable systematic evaluation, we establish a standardized protocol on the proposed PairUAV benchmark. We assess all methods on unseen test sets, which feature novel environments, to evaluate out-of-distribution generalization. Performance is quantified using three aggregate metrics averaged over all episodes: mean angular error, mean distance error, and success rate. We benchmark our method, DreamNav, against representative methods from four complementary families: visual servoing models (AI2THOR), aerial visual localization methods (Sample4Geo), vision foundation models (DINOv3-ViT7B), and generalist vision-language-action policies ($\pi_{0.5}$). We further

instantiate DreamNav with two different coarse visual backbones, DINO-ResNet-50 and MASt3R, to evaluate whether the proposed diffusion-based refinement generalizes across different first-stage predictors. In brief, our primary contributions are summarized as follows:

- **A Large-Scale Benchmark for Last-Meter Navigation.** We identify the lack of standardized evaluation for fine-grained, close-range UAV navigation, particularly in the critical final 10 meters. To fill this gap, we introduce **PairUAV**, a comprehensive benchmark featuring 4,817,232 image pairs across 72 diverse scenes, curated from real-world environments. It enables rigorous evaluation of visual servoing methods under scale ambiguity, viewpoint variation, and fine spatial reasoning.
- **A Coarse-to-Fine Diffusion-Refined Framework.** We propose **DreamNav**, a novel visual servoing framework that combines robust coarse estimation with diffusion-based refinement. In the first stage, a trigonometrically parameterized policy ensures stable rotation prediction. In the second stage, DreamNav leverages a pre-trained world model and a diffusion-inspired imagination process to simulate future observations and select optimal actions, effectively refining pose estimates through visual prediction.
- **Superior Accuracy and Generalization in Realistic Settings.** Extensive experiments show that DreamNav consistently improves coarse pose estimators built on different visual backbones. With MASt3R as the coarse estimator, the full model achieves an MAE_R of **3.11 m**, an MAE_H of **7.50**$^\circ$, and a success rate of **78.86%**. It outperforms representative visual servoing, aerial localization, vision foundation, and generalist vision-language-action baselines on unseen test scenes, while requiring substantially less task-specific training time than the strong $\pi_{0.5}$ model.

2 Related Work

Drone Navigation. Drone navigation has received increasing attention in recent years, propelled by the proliferation of Unmanned Aerial Vehicles (UAVs) as a powerful tool for gathering rich and diverse multimedia content. Their unique vantage point enables applications in aerial photography (Zhao & Zhu, 2023; Shao et al., 2022), cinematography (Pitas & Mademlis, 2022; Dang et al., 2022), and geo-localization (Liu & Li, 2019; Shi et al., 2019; 2020b;a; Shi & Li, 2022; Shi et al., 2022a;b; 2023; Song et al., 2023), among others. Early navigation methods often assume reliable GNSS and strong localization, an assumption that is extremely challenging in signal-denied settings. To address this, recent work guides UAVs to their targets by leveraging richer onboard inputs, a direction further empowered by advancements in computer vision and machine learning (Wang et al., 2024a;b; Zheng & Yang, 2022). For example, some approaches cast movement commands into semantic, instruction-following formulations (Xiao et al., 2025; Chu et al., 2024), while others specify the destination via detailed goal descriptors (Liu et al., 2023a). In this work, we guide a UAV for last-meter precision navigation using a single target-view image in lieu of GNSS. Unlike richer multimodal or map-based inputs, this single-image cue is lightweight and easy to acquire yet provides less explicit information, making the problem more challenging and practically relevant.

Vision Servoing Models. Visual servoing uses visual feedback in a closed loop to command robot motion. Two canonical formulations are Position-Based Visual Servoing (PBVS) and Image-Based Visual Servoing (IBVS). Recently, deep neural networks have shown strong performance in this area, benefiting from their capability in automated image and video analysis (Liu et al., 2023b; Li et al., 2021; Jin et al., 2022; Zheng & Yang, 2019; Wang et al., 2023; Ju et al., 2024). For instance, Zhu et al. (Zhu et al., 2017) first apply deep neural networks to IBVS, and subsequent work (Ren et al., 2025; Bar et al., 2025) expands the data and substantially improves performance. Our approach follows this line by injecting target image cues into a neural network to close the loop, but with two key differences. First, prior studies focus on ground navigation, whereas we address aerial navigation, a domain of growing importance for capturing data from previously inaccessible or hard-to-reach locations (Zhang et al., 2019; Cai et al., 2020; Liu et al., 2023b). To the best of our knowledge, this is the first visual-servoing system studied for aerial scenarios. Second, earlier approaches treat decisions as classification on a finite action set, whereas we learn a regressor that outputs continuous, parameterized control. Our method is similar to the above works in injecting target image cues into a neural network to close the loop, but differs in that we address aerial navigation and regress continuous, parameterized control rather than classify over a finite action set. It remains challenging to estimate the subtle relative-pose differences under scale ambiguity and rotational discontinuities.

Table 1: Comparison with the existing ObjectNav and Aerial Navigation benchmarks.

Benchmark	Agent	External Inputs	#DoF	#Instances
AI2-THOR (Zhu et al., 2017)	Robot	Target Image	-	2,176
RoboTHOR (Deitke et al., 2020)	Robot	Object Category	6 actions	731
UAV-ON (Xiao et al., 2025)	Drone	Target Description	4	11,000
AerialVLN (Liu et al., 2023a)	Drone	Movement Instruction	4	25,338
GeoText (Chu et al., 2024)	Drone	Target Description	-	276,045
PairUAV (ours)	Drone	Target Image	2	**4,817,232**

3 PAIRUAV DATASET

Task Definition. We formulate image-based terminal aerial navigation in a constrained 3D viewpoint space. A UAV receives monocular source and goal RGB images at 512×512 resolution, defined with respect to the same scene-level visual reference center. The camera pitch is fixed at $\psi = 45^\circ$, the horizontal field of view is constant, and the camera remains oriented toward that center. Each admissible viewpoint is therefore parameterized by range and heading coordinates $q = (r, \theta)$, rather than an arbitrary 6-DoF transformation. Given source and target viewpoints q_s and q_g, the relative command is $a = [\Delta\theta, \Delta r]$, where $\Delta\theta = \mathrm{wrap}(\theta_g - \theta_s)$ and $\Delta r = r_g - r_s$. Importantly, Δr is a range-coordinate change rather than the Euclidean source-to-target travel distance. The model estimates these changes from the two observations without requiring the global coordinates of the reference center.

Dataset Construction. We build upon University-1652 (Zheng et al., 2020), which spans 1,652 buildings across 72 universities. Monocular UAV views are rendered from Google Earth models at 54 sampled coordinates per building. We repurpose these assets from cross-view retrieval to precision navigation by generating all ordered pairs for each building, with supervision given by the heading and radial-range changes defined above. This yields 4,817,232 ordered navigation instances under the constrained 3D flight geometry.

Comparison with Other Datasets. Table 1 compares PairUAV with ObjectNav and aerial navigation benchmarks. PairUAV combines image-specified goals with continuous heading–range controls, complementing ground-robot navigation, language-conditioned UAV navigation, and retrieval-based evaluation.

4 METHOD

Overview. As shown in Figure 2, the proposed DreamNav comprises two stages. Stage I combines appearance and geometric cues to predict coarse changes in both heading and radial range. Stage II generates local candidates, synthesizes their observations via a diffusion model, and selects the candidate whose synthesized view best matches the target image.

4.1 COARSE-ESTIMATION STAGE

Given source and target images S and G, the coarse stage predicts $p = (\Delta\widehat{r}, \Delta\hat{\theta})$, the radial-range and heading changes around their shared visual reference center.

Displacement-Based Geometric Cue Extraction. Directly concatenating S and G with convolutional fusion can lose absolute position information due to approximate translation equivariance (Liu et al., 2018). We therefore replace the target image with a coordinate displacement field. A pretrained keypoint matcher produces correspondences $\mathcal{M} = \{(\mathbf{u}_i^S, \mathbf{u}_i^G, w_i)\}_{i=1}^N$, where w_i denotes confidence. We retain matches above threshold τ and diffuse their displacements $\Delta\mathbf{u}_i = \mathbf{u}_i^G - \mathbf{u}_i^S$ with a Gaussian kernel $\kappa_\sigma(r) = \exp(-r^2/(2\sigma^2))$:

$$D(\mathbf{p}) = \frac{\sum_i w_i \, \kappa_\sigma(\|\mathbf{p} - \mathbf{u}_i^S\|_2) \, \Delta\mathbf{u}_i}{\sum_i w_i \, \kappa_\sigma(\|\mathbf{p} - \mathbf{u}_i^S\|_2) + \varepsilon}. \tag{1}$$

Here $\mathbf{p}$ indexes image pixels and ε ensures numerical stability. The field shrinks toward zero in poorly matched regions.

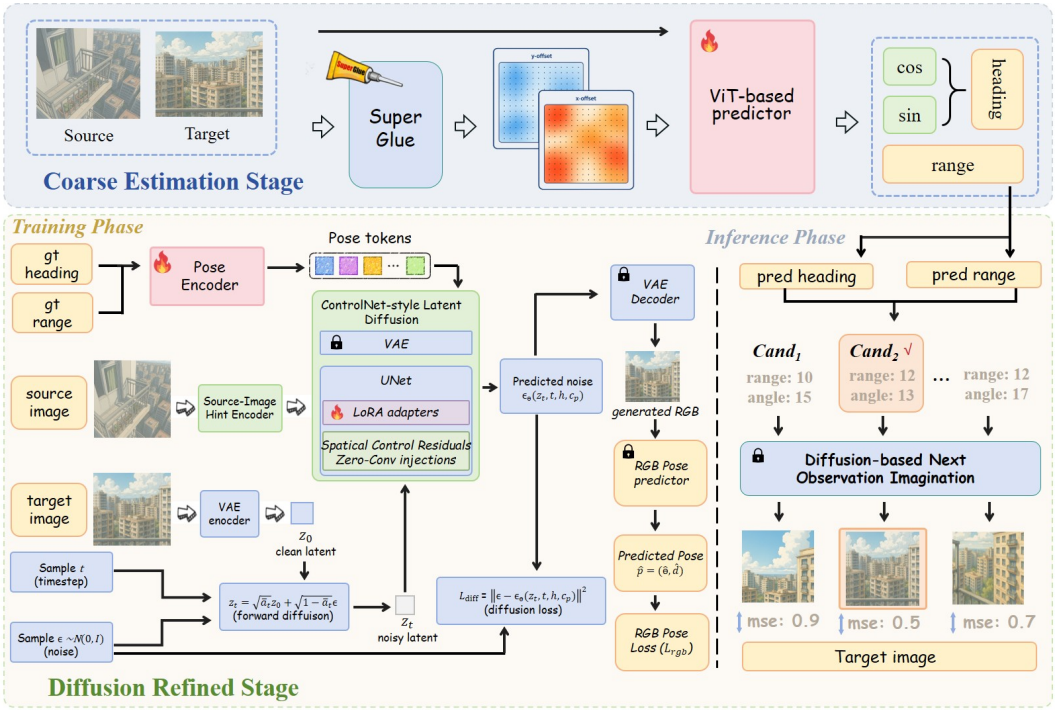

Figure 2: Overall architecture of **DreamNav**. *(Top)* Stage I regresses coarse heading and radial-range changes from source and target views. *(Bottom)* Stage II conditions a ControlNet-style latent diffusion model on the source image through a spatial hint pathway and on the viewpoint change through pose tokens. Ground-truth changes supervise target-view synthesis during training; at inference, the candidate whose synthesized view best matches the target is selected.

Dual-Cue Representation Fusion. The source image contributes appearance cues, while D contributes geometric cues. We normalize displacement by image size as $\widehat{D} = \beta[D_x/W, D_y/H]^\top$, with optional scale β, and normalize S to $[-1, 1]$. Concatenating them along the channel dimension yields $X \in \mathbb{R}^{H\times W\times 5}$, the input to the coarse pose estimator.

Pose Regression Head. A linear head after the visual backbone predicts range and heading. To compress the dynamic range while preserving fine resolution for small changes, we use signed logarithmic range regression:

$$\mathcal{L}_{\text{dist}} = \big(\hat{d} - \text{sign}(d)\log(1 + |d|/\gamma)\big)^2, \tag{2}$$

where $d = \Delta r$, γ is a scale factor, and $\hat{d}$ is the prediction in the transformed space. To avoid angular wrap-around discontinuities, we regress sine and cosine components:

$$\mathcal{L}_{\text{rot}} = (\hat{s} - \sin\Delta\theta)^2 + (\hat{c} - \cos\Delta\theta)^2. \tag{3}$$

At inference, these components reconstruct $\Delta\hat{\theta}$. The coarse-stage objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dist}} + \lambda\mathcal{L}_{\text{rot}}$, where λ balances the two terms.

4.2 Diffusion-Refined Stage

While coarse estimation provides a strong initialization, direct regression may still suffer from multi-modal visual ambiguity and local misalignment. We locally perturb the predicted heading–range pair, synthesize the visual consequence of each candidate, and select the best match to the target.

Heuristic Candidate Generation. Given $p_{\text{coarse}} = (\Delta\widehat{r}, \Delta\hat{\theta})$, we form $\mathcal{P} = \{(\Delta\widehat{r} + \delta_r, \Delta\hat{\theta} + \delta_\theta) : \delta_r \in \Omega_r, \delta_\theta \in \Omega_\theta\}$. The fixed offset sets are reported in the experimental setup. This local grid corrects small errors while keeping the search computationally tractable.

Pose Information Encoding. A trainable pose encoder preserves the periodic nature of heading using sinusoidal features $(\cos(\omega_k\Delta\hat{\theta}), \sin(\omega_k\Delta\hat{\theta}))_{k=1}^{K_\theta}$. For range, it concatenates the normalized

scalar with its sinusoidal expansion using frequencies ν_k, $k = 1, \ldots, K_d$. An MLP maps the concatenated heading and range features into pose tokens $\mathbf{c}_p = E_p(p)$ for cross-attention conditioning.

Next Observation Generation. We use a ControlNet-style latent diffusion model. The source image is processed by a hint block, $\mathbf{h} = H_\phi(S)$, to provide spatial conditioning. Given noisy latent $\mathbf{z}_t$, timestep t, pose tokens $\mathbf{c}_p$, and hint $\mathbf{h}$, ControlNet produces residuals $\mathbf{r}_l = Z_l(C_l(\mathbf{z}_t, t, \mathbf{c}_p, \mathbf{h}))$, where Z_l is a zero-initialized 1×1 convolution. These residuals are added to the corresponding UNet features and middle block. The model denoises in VAE latent space and decodes the result into the candidate observation $\widehat{I}(p)$. We freeze the VAE and most pretrained diffusion parameters, optimizing only the pose encoder, source-image hint pathway, and LoRA adapters in the ControlNet and UNet decoder/output blocks. This preserves the pretrained generative prior while adapting the model to pose-conditioned UAV view synthesis.

Training. Given target latent $\mathbf{z}_0$, timestep t, and Gaussian noise ϵ, the model predicts the added noise from $\mathbf{z}_t$ and the source-image and ground-truth pose conditions:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}_0, t, \epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{h}, \mathbf{c}_p)\|_2^2\right]. \tag{4}$$

A frozen RGB pose predictor $F_{\text{rgb}}(\widehat{I})$ supplies auxiliary supervision by estimating the generated image's heading and range. We encourage agreement with the conditioning pose using

$$\mathcal{L}_{\text{rgb}} = \mathcal{L}_\theta + \mathcal{L}_d + 0.001\mathcal{L}_n, \tag{5}$$

where the terms penalize heading discrepancy, range discrepancy, and a small normalization regularizer, respectively. Following the setting in Spetlik et al. (2024), we set the weight of the normalization term to 0.001. The final objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{rgb}}$.

Inference. For each candidate in $\mathcal{P}$, the diffusion model synthesizes an observation. We select the pose with the lowest pixel-wise MSE to the target:

$$p^* = \underset{p_k \in \mathcal{P}}{\operatorname{argmin}} \|\widehat{I}(p_k) - G\|_2^2. \tag{6}$$

5 Experiment

Dataset. We evaluate our approach on PairUAV using the scene partition of University-1652. The training set encompasses 33 scenes with 701 buildings; the test set contains the remaining 39 unseen scenes with 951 buildings. We use all 54×54 ordered image combinations per building to assess out-of-domain generalization.

Metrics. Following previous navigation works (Anderson et al., 2018), we report success rate (SR), mean absolute angle error (MAE_H), and mean absolute range error (MAE_R). SR measures the fraction of episodes ending within 10 m of the goal, using the three-dimensional Euclidean distance between the final position and target waypoint. Higher SR and lower MAE_H/MAE_R indicate better performance.

Implementation Details. Our method is implemented in PyTorch. For correspondence extraction in the coarse-estimation stage, we use a pretrained SuperGlue matcher (Sarlin et al., 2020) with the indoor weights. SuperPoint detects up to 1,024 keypoints using a keypoint confidence threshold of 0.005 and an NMS radius of 4 pixels. SuperGlue uses 20 Sinkhorn iterations, and matches with confidence below $\tau = 0.2$ are discarded. The displacement vectors of the remaining matches are scaled and rasterized at their source keypoint locations on a 224×224 grid. The resulting two-channel displacement field is concatenated with the 224×224 source image and supplied to the Stage-I pose estimator. We instantiate the coarse estimator with two different visual backbones, DINO-ResNet-50 and MASt3R, while keeping the remaining prediction and refinement pipeline unchanged. The heading is represented using sine and cosine components and optimized jointly with the range regression objective. Stage I requires approximately 2 hours of training, excluding offline feature extraction. For Stage II, we construct a local candidate set around the coarse prediction using heading offsets of $\{0^\circ, \pm 3^\circ, \pm 4^\circ, \pm 6^\circ, \pm 15^\circ, \pm 20^\circ\}$ and range offsets of $\{-0.1, 0, +0.1\}$. The diffusion generator and trainable modules follow Sec. 4.2. Training Stage II requires approximately 50 hours on a single NVIDIA GeForce RTX 5090 GPU.

Table 2: Comparison with competitive methods on PairUAV. MAE_R and MAE_H denote mean absolute range and heading errors; AVG is their average. Stage II improves both DreamNav variants.

Methods	Backbone	MAE_R (m) ↓	MAE_H (deg) ↓	MAE_{AVG} ↓	SR (%) ↑
AI2THOR (Zhu et al., 2017)	ResNet-50	44.96	89.99	67.48	14.32
Sample4Geo (Deuser et al., 2023)	ConvNeXt-B	23.98	90.07	57.03	6.89
DINOv3 (Siméoni et al., 2025)	ViT-7B/16	15.77	89.86	52.81	10.33
$\pi_{0.5}$ (Intelligence et al., 2025)	SigLIP So400M	4.38	8.25	6.32	70.66
DreamNav (Stage I)	DINO-R50	24.60	30.30	27.50	21.40
DreamNav (Stage II)	DINO-R50	23.00	28.10	25.60	22.40
DreamNav (Stage I)	MASt3R	3.12	10.98	7.05	67.70
DreamNav (Stage II)	MASt3R	**3.11**	**7.50**	**5.30**	**78.86**

5.1 Main Results

Baseline Methods. Given the absence of prior methods specifically tailored for image-goal aerial visual navigation with continuous parameter control, we benchmark against representative approaches from four adjacent domains to establish a comprehensive evaluation:

- **Vision-based servoing methods** with neural controllers and inputs comparable to ours, including AI2THOR (Zhu et al., 2017).
- **Aerial visual localization methods** in similar UAV settings, including Sample4Geo (Deuser et al., 2023), which is also built on the University-1652 benchmark (Zheng et al., 2020).
- **Vision foundation models**, where we use DINOv3 (Siméoni et al., 2025) as a strong off-the-shelf visual representation baseline within the same controller.
- **Generalist vision-language-action models**, where we include $\pi_{0.5}$ (Intelligence et al., 2025) as a strong embodied-policy baseline with extensive robot-navigation pretraining.

All methods are benchmarked on the testing set of the PairUAV dataset. For the visual servoing models, we adapt their task setting to single-step action prediction and subsequently fine-tune their pre-trained backbones. For feature representation-based models, we first extract features using their officially released weights and then add a prediction head for training and evaluation. For $\pi_{0.5}$, we adapt its action interface to the 2-DoF heading–range prediction space of PairUAV and train it under the same evaluation protocol.

Quantitative Results. The comparison results on the PairUAV dataset are reported in Table 2. We evaluate two instantiations of DreamNav with different Stage-I backbones, namely DINO-R50 and MASt3R. Across both variants, the diffusion-refined second stage consistently improves the coarse predictions, indicating that the proposed refinement mechanism is not tied to a specific visual backbone. Among the conventional visual servoing, localization, and representation-based baselines, AI2THOR, DINOv3, and Sample4Geo exhibit large heading errors, which substantially limit their navigation success. In contrast, the generalist vision-language-action model $\pi_{0.5}$ provides a considerably stronger baseline, achieving an MAE_R of $4.38\,\mathrm{m}$, an MAE_H of $8.25°$, an AVG error of 6.32, and an SR of 70.66%. For DreamNav-DINO-R50, the diffusion-refined stage reduces MAE_R from $24.60\,\mathrm{m}$ to $23.00\,\mathrm{m}$ and MAE_H from $30.30°$ to $28.10°$, while improving SR from 21.40% to 22.40%. A substantially stronger coarse prediction is obtained when MASt3R is used as the Stage-I backbone: DreamNav-MASt3R achieves an MAE_R of $3.12\,\mathrm{m}$, an MAE_H of $10.98°$, and an SR of 67.70% before refinement. The Stage-II diffusion refinement further reduces the heading error to $7.50°$ and increases SR to 78.86%, while maintaining a low range error of $3.11\,\mathrm{m}$. Compared with $\pi_{0.5}$, the strongest competing baseline, the full DreamNav-MASt3R model reduces AVG error from 6.32 to 5.30 and improves SR from 70.66% to 78.86%. Notably, the Stage-I MASt3R predictor alone does not surpass $\pi_{0.5}$ in either heading error or success rate; the advantage emerges after diffusion-based refinement. This result provides direct evidence that the second stage contributes complementary visual verification beyond the strong geometric prior provided by the coarse predictor.

Qualitative Results. We visualize Stage-II diffusion-based imagination in Figure 3 (a). We randomly sample examples from PairUAV and use the coarse prediction from Stage-I as the central anchor to form a 3×3 grid of conditioning inputs by applying systematic offsets to the heading

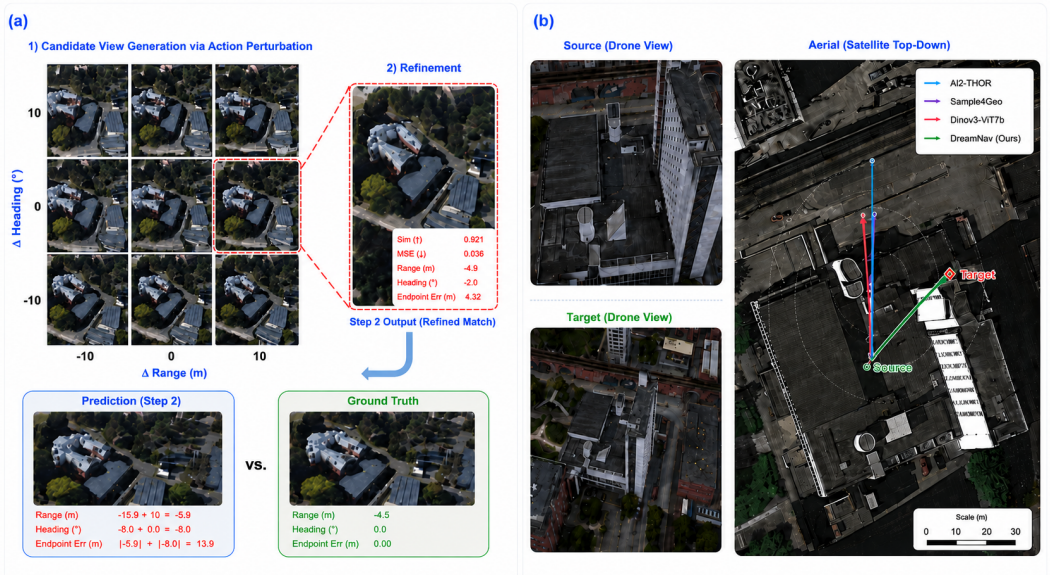

Figure 3: **Qualitative visualization of DreamNav.** (*a*) Stage-II next-view synthesis around the Stage-I prediction, illustrated with a 3×3 heading–range grid and compared with the ground truth. (*b*) Predictions projected onto an overhead satellite view. The red cross denotes the target; colored arrows show predicted navigation directions.

Table 3: Ablation studies on the test set of PairUAV. Trig. Reg. denotes Trigonometric Regression, while Match Data denotes correspondence-derived matching information.

Trig. Reg.	Dual-Cue Rep.	Refine Stage	MAE_R (m) ↓	MAE_H (deg) ↓	MAE_{AVG} ↓	SR (%) ↑
			20.25	22.26	21.26	22.50
✓			12.39	20.26	16.33	29.40
✓	✓		3.12	10.98	7.05	67.70
✓	✓	✓	**3.11**	**7.50**	**5.30**	**78.86**

and range. The synthesized next-views respond smoothly and consistently to these perturbations, where changing heading gradually rotates the rooftop orientation and facade layout, while moving forward/backward produces the expected scale change and parallax of nearby structures. Moreover, occlusion boundaries and relative depth ordering remain coherent across the grid, suggesting that the diffusion-refined stage captures geometry-aware viewpoint transformations rather than simply memorizing local textures. As illustrated in Figure 3 (b), our method predicts a direction that is well aligned with the true goal, allowing the UAV to reach a location close to the target. In contrast, existing baselines exhibit noticeable heading biases, leading to trajectories that drift toward incorrect orientations and accumulate large lateral errors. This qualitative evidence further shows that our trigonometric loss improves angular consistency, resulting in more accurate heading estimation and more reliable range regression.

Training Efficiency. In addition to navigation accuracy, we compare the training cost of DreamNav with the strongest competing baseline, $\pi_{0.5}$. Under our experimental setup, adapting and training $\pi_{0.5}$ on PairUAV requires approximately 80–81 hours. In comparison, DreamNav requires approximately 2 hours to train Stage I after feature extraction, while training the Stage-II diffusion refinement requires approximately 50 hours. Thus, despite achieving higher final navigation accuracy, DreamNav maintains a substantially lower task-specific training cost than $\pi_{0.5}$.

5.2 Ablation Studies

Effect of Key Components. As shown in Table 3, we investigate the contribution of the key components in DreamNav, including trigonometric rotation regression, correspondence-derived matching information, and the diffusion-refined stage. Starting from the baseline without these components, the model obtains an MAE_R of 20.25 m, an MAE_H of 22.26°, and an SR of 22.50%. Introducing trigonometric regression reduces MAE_R to 12.39 m and MAE_H to 20.26°, indicating that the periodic representation provides a more stable objective for heading estimation. When

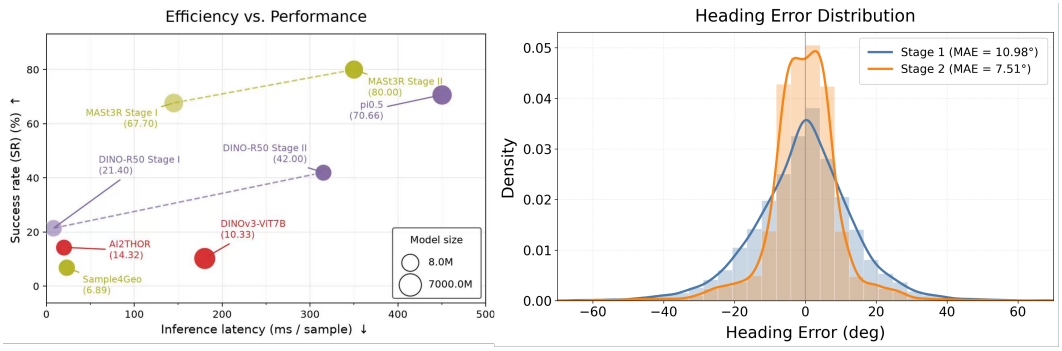

Figure 4: **Efficiency and heading-error analysis.** (*Left*) Inference latency versus navigation success rate, with marker size indicating model size. (*Right*) Stage II concentrates the signed heading errors more strongly around zero, reducing the heading MAE from 10.98° to 7.50°.

correspondence-derived matching information is further incorporated, the performance improves substantially: MAE_R decreases to 3.12 m, MAE_H to 10.98°, and SR increases to 67.70%. This result suggests that explicit cross-view geometric cues are particularly effective when combined with the stabilized angular representation provided by trigonometric regression.

Efficiency. We next ask whether the accuracy of DreamNav is bought with inference time. Stage 1 DreamNav costs XX ms/sample and reaches XX % SR; the coarse stage alone is cheap but far from deployment-grade accuracy. Stage 2, DreamNav, is the component that carries the accuracy. Starting from the same frozen backbone, adding the DreamNav navigator raises DINOv3-ViT7B from 10.33 % to 42.00 % (DINO-R50, 318 ms), i.e., +31.67 points for a 138 ms latency budget, and reaches 80.00 % with the MAST3R backbone at 350 ms. Compared with the largest model in the plot, DreamNav-MAST3R attains a higher success rate than $\pi_{0.5}$ (+8.20 points) while incurring $\sim$22 % less per-sample latency and using a smaller model, and it is the only point in the upper-left region of the latency–accuracy plane. We therefore place DreamNav on the Pareto frontier at this operating point rather than claim a uniform speed–accuracy dominance across all latency budgets.

Effect of Diffusion-Based Refinement. Adding the diffusion-refined stage further reduces MAE_H from 10.98° to 7.50° and increases SR from 67.70% to 78.86%, while maintaining a low range error of 3.11 m. Figure 4(*Right*) further summarizes the efficiency–performance trade-off and how refinement changes the prediction errors at the distribution level. The left panel shows that DreamNav achieves competitive navigation success while remaining substantially smaller than the largest foundation-model baseline. Compared with Stage I, the Stage-II distribution is noticeably narrower and more concentrated around zero, with reduced probability mass at larger heading deviations. This distributional contraction is consistent with the reduction in MAE_H, indicating that the improvement comes from broadly reducing heading errors across the test set rather than from a simple shift in signed error bias.

6 Conclusion

This paper introduces **DreamNav**, a coarse-to-fine framework for last-meter UAV navigation that pairs geometry-informed trigonometric pose regression with diffusion-based visual refinement, and **PairUAV**, a benchmark of 4.8 million image pairs across 72 scenes. Diffusion refinement reduces pose errors and improves navigation success even from inaccurate DINO-R50 coarse predictions, and consistently improves both backbones without closing the gap between them. With MASt3R as the coarse estimator, DreamNav achieves a range error of 3.11 m, a heading error of 7.50°, and a success rate of 78.86% on unseen PairUAV scenes, where both the geometry-based MASt3R estimator and the vision-language-action model $\pi_{0.5}$ outperform generic visual representation baselines. These results suggest that combining task-relevant pretrained representations with explicit visual verification is a promising route to image-goal navigation in unseen aerial environments.

Limitations. The refinement stage inherits the sampling overhead of diffusion models, requiring N denoising steps per refined pose. We confine this cost to the second stage rather than the framework: the geometry-informed coarse estimator remains a single forward pass and drives the navigation loop, and refinement is invoked once per episode instead of per control step. We report inference

latency as an efficiency reference, but make no real-time deployment claim. Because the sampler is a module boundary, few-step distillation and consistency objectives (Liu et al., 2024) can replace it without altering the geometry-informed regression or the visual verification, and we leave this acceleration, together with deployment on embedded UAV hardware, to future work.

AI USE STATEMENT

We used generative AI to assist with implementation, experimental workflows, manuscript editing, and feedback on methodology and experimental design. The authors reviewed all methods, code, analyses, claims, and interpretations; inspected and tested AI-assisted code; and checked scientific claims against experimental results and relevant literature. The authors made all final decisions and take responsibility for the complete work.

REFERENCES

Global positioning system (gps) standard positioning service (sps) performance standard. Technical report, U.S. Department of Defense, Washington, DC, April 2020. URL https://archive.gps.gov/technical/ps/2020-SPS-performance-standard.pdf. Executive Summary.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3674–3683, 2018.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025.

Gino Brunner, Bence Szebedy, Simon Tanner, and Roger Wattenhofer. The urban last mile problem: Autonomous drone delivery to your balcony. In *2019 international conference on unmanned aircraft systems (icuas)*, pp. 1005–1012. IEEE, 2019.

Yajun Bu, Ye Yan, and Yueneng Yang. Advancement challenges in uav swarm formation control: A comprehensive review. *Drones*, 8(7):320, 2024.

YuanQiang Cai, Dawei Du, Libo Zhang, Longyin Wen, Weiqiang Wang, Yanjun Wu, and Siwei Lyu. Guided attention network for object detection and counting on drones. In *ACM Multimedia*, 2020.

Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *ECCV (11)*, 2024.

Yuanjie Dang, Chong Huang, Peng Chen, Ronghua Liang, Xin Yang, and Kwang-Ting Cheng. Path-analysis-based reinforcement learning algorithm for imitation filming. *TMM*, 2022.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3164–3174, 2020.

Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16847–16856, 2023.

Jinsoo Hwang and Ja Young (Jacey) Choe. Exploring perceived risk in building successful drone food delivery services. *International Journal of Contemporary Hospitality Management*, 31(8): 3249–3269, 07 2019. ISSN 0959-6119. doi: 10.1108/IJCHM-07-2018-0558. URL https://doi.org/10.1108/IJCHM-07-2018-0558.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Pu Jin, Lichao Mou, Gui-Song Xia, and Xiao Xiang Zhu. Anomaly detection in aerial videos with transformers. *TGRS*, 2022.

Hao Ju, Shaofei Huang, Si Liu, and Zhedong Zheng. Video2bev: Transforming drone videos to bevs for video-based geo-localization. In *ICCV*, 2024.

Luka Kramarić, Niko Jelušić, Tomislav Radišić, and Mario Muštra. A comprehensive survey on short-distance localization of uavs. *Drones*, 9(3):188, 2025.

Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *CVPR*, 2021.

Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *CVPR*, 2019.

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31, 2018.

Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. Aerialvln: Vision-and-language navigation for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15384–15394, 2023a.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

Zhihao Liu, Yuanyuan Shang, Timing Li, Guanlin Chen, Yu Wang, Qinghua Hu, and Pengfei Zhu. Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark. *TMM*, 2023b.

Vinayak patki, Abolfazl Mehbodniya, Julian L. Webber, Arulmani kuppusamy, Mohd anul haq, Anil kumar, and Sathishkumar Karupusamy. Improving the geo-drone-based route for effective communication and connection stability improvement in the emergency area ad-hoc network. *Sustainable Energy Technologies and Assessments*, 53:102558, 2022. ISSN 2213-1388. doi: https://doi.org/10.1016/j.seta.2022.102558. URL https://www.sciencedirect.com/science/article/pii/S2213138822006087.

Ioannis Pitas and Ioannis Mademlis. Autonomous uav cinematography. In *ACM Multimedia*, 2022.

Md Tahmid Rashid, Daniel Yue Zhang, and Dong Wang. Socialdrone: An integrated social media and drone sensing system for reliable disaster response. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, pp. 218–227, 2020. doi: 10.1109/INFOCOM41043.2020.9155522.

Hao Ren, Yiming Zeng, Zetong Bi, Zhaoliang Wan, Junlong Huang, and Hui Cheng. Prior does matter: Visual navigation via denoising diffusion bridge models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12100–12110, 2025.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4938–4947, 2020.

Mingwen Shao, Chao Wang, Wangmeng Zuo, and Deyu Meng. Efficient pyramidal gan for versatile missing data reconstruction in remote sensing images. *TGRS*, 60:1–14, 2022.

Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *CVPR*, 2022.

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *NeurIPS*, 2019.

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *CVPR*, 2020a.

Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *AAAI*, 2020b.

Yujiao Shi, Xin Yu, Liu Liu, Dylan Campbell, Piotr Koniusz, and Hongdong Li. Accurate 3-dof camera geo-localization via ground-to-satellite image matching. *TPAMI*, 2022a.

Yujiao Shi, Xin Yu, Shan Wang, and Hongdong Li. Cvlnet: Cross-view feature correspondence learning for video-based camera localization. In *ACCV*, 2022b.

Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-dof ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer. In *ICCV*, 2023.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Zhenbo Song, Xianghui Ze, Jianfeng Lu, and Yujiao Shi. Learning dense flow field for highly-accurate cross-view camera localization. In *NIPS*, 2023.

Radim Spetlik, Denys Rozumnyi, and Jiří Matas. Single-image deblurring, trajectory and shape recovery of fast moving objects with denoising diffusion probabilistic models. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 6843–6852, 2024.

Chao Wang, Zhedong Zheng, Ruijie Quan, Yifan Sun, and Yi Yang. Context-aware pretraining for efficient blind image decomposition. In *CVPR*, 2023.

Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Tat-Seng Chua, Yi Yang, and Chenggang Yan. Multiple-environment self-adaptive network for aerial-view geo-localization. *Pattern Recognition*, 2024a.

Tingyu Wang, Zhedong Zheng, Zunjie Zhu, Yaoqi Sun, Yi Yang, and Chenggang Yan. Learning cross-view geo-localization embeddings via dynamic weighted decorrelation regularization. *TGRS*, 2024b.

Jianqiang Xiao, Yuexuan Sun, Yixin Shao, Boxi Gan, Rongqiang Liu, Yanjing Wu, Weili Gua, and Xiang Deng. Uav-on: A benchmark for open-world object goal navigation with aerial agents. *arXiv preprint arXiv:2508.00288*, 2025.

Haotian Zhang, Gaoang Wang, Zhichao Lei, and Jenq-Neng Hwang. Eye in the sky: Drone-based object tracking and 3d localization. In *ACM Multimedia*, 2019.

LiangLiang Zhao and MinLing Zhu. Ms-yolov7: Yolov7 based on multi-scale for object detection on uav aerial photography. *Drones*, 2023.

Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. *IJCAI*, 2019.

Zhedong Zheng and Yi Yang. Adaptive boosting for domain adaptation: Toward robust predictions in scene segmentation. *IEEE Transactions on Image Processing*, 31:5371–5382, 2022.

Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *ACM Multimedia*, 2020.

Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J. Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3357–3364, 2017. doi: 10.1109/ICRA.2017.7989381.