The Building Blocks of Learning-based Monocular Visual Odometry for Underwater Environments

Olaya Álvarez-Tuñón, Yury Brodskiy, Stella Graßhof, Andrzej Wasowski

***Abstract*—Geometry-based visual odometry has long been the de facto standard due to its solid mathematical foundations, but its performance degrades in underwater environments affected by turbidity, scattering, and low contrast. Learning-based methods offer increased robustness by leveraging higher-level visual representations that are more robust to such degradations.**

We revisit monocular VO as a composition of subproblems: pixel correspondence, depth estimation, and pose optimization. We propose a modular framework that combines learning-based and geometry-based components. Our approach employs neural networks for dense optical flow and monocular depth prediction, jointly estimating per-pixel uncertainties. To ensure these uncertainties are reliable and comparable across modules, we introduce a conformal prediction framework for uncertainty calibration under distribution shift.

The calibrated uncertainties are integrated into a geometry-based pose graph optimization, improving robustness and convergence. The resulting system enables flexible, modular VO design and performs robustly in visually degraded underwater conditions.

***Index Terms*—Deep Learning for Visual Odometry, Underwater Visual Odometry.**

I. Introduction

Monocular visual odometry (VO) remains a fundamental yet challenging problem, particularly in environments where visual conditions deviate from standard conditions, such as underwater scenes with poor visibility, non-uniform lighting, and complex geometry. While classical geometric methods provide strong inductive biases and robustness in well-structured scenarios, they often struggle in low-texture or highly dynamic conditions. Conversely, learning-based approaches have shown impressive performance by leveraging data-driven representations, but they frequently suffer from limited generalization and poorly calibrated uncertainty under domain shift.

In this work, we revisit monocular VO as a composition of three fundamental subproblems: pixel correspondence, depth estimation, and camera pose optimization. Rather than relying exclusively on either geometric or learning-based paradigms, we propose a modular framework that integrates both. Building upon the MACVO framework [11], originally introduced for stereo VO, we extend it to the monocular setting by incorporating learning-based modules for dense optical flow and monocular depth estimation. Each module additionally predicts per-pixel uncertainty, enabling a principled integration of learned perception into a geometry-based optimization pipeline.

Main funding source(s): the Innovation Fund Denmark, project DeepODO (Deep visual odometry for underwater intervention drones).

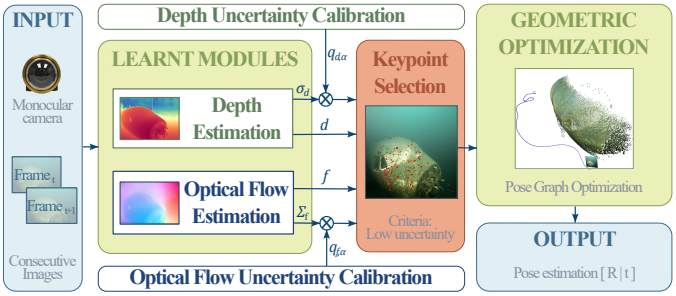


Fig. 1: The proposed framework processes consecutive monocular camera inputs to jointly estimate optical flow and per-pixel depth, along with their associated uncertainties. The predicted uncertainties are scaled by a calibration parameter. Subsequently, keypoints with low uncertainty are selected and passed to a geometric optimization module, which retrieves the camera's pose estimation.

A central challenge in such hybrid systems is the consistency and reliability of uncertainty estimates across modules. As shown in our experiments, learned uncertainty can become overconfident under domain shift, particularly in underwater environments, leading to suboptimal weighting of measurements during optimization. To address this, we introduce a conformal prediction-based framework for uncertainty calibration, which aligns the uncertainty distributions of the depth and optical flow modules. An overview of the framework is shown in Fig. 1. This calibration enables more reliable weighting of residuals within a geometry-based pose graph optimization, improving robustness, convergence, and overall system stability. The proposed uncertainty calibration framework is motivated by the observation that while learning-based components provide strong performance in-distribution and under moderate domain shift, their uncertainty estimates degrade in more challenging conditions.

By combining learned perception with calibrated uncertainty and geometric optimization, the framework aims to retain the strengths of both paradigms while mitigating their individual limitations.

The main contributions of this work are summarized as follows:

- A modular framework for monocular visual odometry that combines learning-based perception with geometry-based optimization.
- Learning-based modules for dense optical flow and monocular depth estimation with associated per-pixel

uncertainty prediction.
- A conformal prediction-based methodology for uncertainty calibration across heterogeneous network outputs, enabling improved and more robust pose estimation.

We evaluate the proposed approach across multiple datasets with varying levels of domain shift, including surface, public underwater, and proprietary underwater data. The results demonstrate that while the proposed framework achieves state-of-the-art performance in standard settings, its main advantage lies in improved robustness and stability under challenging conditions, highlighting the importance of uncertainty calibration in learned monocular VO systems.

II. State of Art

Visual odometry and SLAM methods can be broadly categorized into geometry-based and learning-based approaches. Geometry-based methods rely on explicit geometric formulations that optimize camera poses jointly with the reconstruction of the 3D scene, yielding interpretable and often accurate results under favorable conditions.

In contrast, learning-based VO approaches leverage deep neural networks to model motion directly from image sequences. Early work such as PoseNet [9] introduced convolutional architectures for absolute pose regression, establishing a paradigm where VO systems are decomposed into a matching module and a pose regression module. The matching stage is typically implemented using optical flow networks. Initial approaches employed encoder–decoder architectures, while more recent methods utilize feature pyramids and iterative refinement, achieving significant improvements in accuracy and efficiency [13, 14, 8, 7]. These advances have been integrated into modern VO pipelines such as DeepVO [16], TartanVO [17], and NeRF-SLAM [12].

Pose regression in learning-based VO is addressed using a variety of architectures, including recurrent neural networks (e.g., LSTM in DeepVO and GRU in DROID-SLAM), convolutional backbones (e.g., ResNet variants), and implicit neural representations such as NeRF. While these approaches effectively capture temporal dependencies and high-dimensional features, they often lack explicit geometric constraints and may struggle to generalize beyond the training distribution.

To address these limitations, hybrid approaches have emerged that integrate learned perception with geometric optimization. Systems such as DROID-SLAM [15] exemplify this paradigm, where dense correspondences are iteratively refined using a learned update operator, but still constrained by multi-view geometry. Similarly, methods like DF-VO [20] combine learned depth and flow with geometric pose estimation. These approaches retain the robustness of learned features while enforcing multi-view consistency through geometric constraints.

More recently, frameworks such as MAC-VO [11] extend this hybrid paradigm by explicitly incorporating uncertainty-aware learned measurements into the optimization pipeline. In MAC-VO, learned modules predict both geometric quantities and associated confidence estimates, which are used to guide feature selection and weighting during optimization. This represents a shift from treating learned outputs as deterministic inputs toward a probabilistic interpretation of network predictions.

In this context, MAC-VO represents an important step toward uncertainty-aware visual odometry, but it is limited to stereo settings and does not fully address the challenges of monocular reconstruction, where scale ambiguity and increased uncertainty play a significant role. In this work, we build upon this paradigm and propose a monocular visual odometry framework with calibrated uncertainty, enabling an integration of learned predictions into geometric optimization under the increased ambiguity of monocular setups.

III. Uncertainty Calibration

The proposed framework combines multiple deep learning modules that output both predictions and associated uncertainty estimates. However, these uncertainties are not directly comparable across models, as each network learns its own scale and error representation. This results in inconsistent weighting within the geometric optimizer and can degrade convergence. To mitigate this, we adopt a conformal calibration strategy that rescales each model's predicted variance based on empirical error statistics computed on a calibration dataset [1].

A. Nonconformity score

The nonconformity score evaluates the inconsistency between a prediction and its estimated uncertainty. Given a prediction $\hat{y}_i$, ground truth y_i and predicted variance $\hat{\sigma}_i^2$, the nonconformity score is defined as:

$$a_i = \frac{|e_i|^2}{\hat{\sigma}_i^2}, \quad e_i = \hat{y}_i - y_i. \tag{1}$$

Here, e_i represents the prediction error, while a_i quantifies how well the predicted uncertainty $\hat{\sigma}_i^2$ accounts for this error. For depth, $||e_i||^2 = (\hat{d}_i - d_i)^2$, while for optical flow $||e_i||^2 = ||(\hat{f}_i - f_i)||_2^2$. This yields a dimensionless measure of consistency between predicted uncertainty and observed error.

B. Quantile Calibration

Given a calibration set, the empirical α-quantile of the nonconformity scores is obtained as:

$$q_\alpha = \mathrm{Quantile}_\alpha\big(\{a_i\}_{i=1}^N\big), \tag{2}$$

which captures the global miscalibration of the predicted variances, representing the threshold below which α fraction of the nonconformity scores lie. It serves as a scaling factor to align the predicted uncertainties with the empirical distribution of errors.

C. Variance Scaling

The calibrated variance is obtained by scaling the original predicted variance $\hat{\sigma}_i^2$ with the quantile q_α:

$$\sigma_{i,\mathrm{cal}}^2 = q_\alpha \hat{\sigma} i^2. \tag{3}$$

This scaling ensures that, for approximately α fraction of samples, the squared prediction error satisfies:

$$|e_i|^2 \lesssim \sigma_{i,\text{cal}}^2, \tag{4}$$

thereby enforcing consistency between uncertainty and empirical error.

IV. Modules/Networks for Monocular Visual Odometry

Visual odometry relies on establishing correspondences across image frames and recovering their 3D structure. In stereo setups, depth can be directly inferred through geometric triangulation, given a known baseline. In contrast, monocular configurations lack explicit geometric constraints and therefore require additional sources of information to recover scale and depth.

In this work, we address this limitation by incorporating a learning-based monocular depth estimation module to infer the 3D structure of the scene. Combined with dense optical flow for pixel correspondence, this enables the reconstruction of 3D point trajectories from monocular input. Given these correspondences and their associated 3D projections, camera motion can be estimated using standard geometric optimization techniques, such as factor graph-based pose estimation.

A. Optical Flow

Given two consecutive images, the objective is to estimate the pixel-wise motion field $\hat{f} \in \mathbb{R}^2$ and its associated uncertainty $\hat{\Sigma}_f$ between two frames. For this purpose, we employ Flowformer, a state-of-the-art architecture for optical flow and uncertainty estimation, as proposed in MACVO [11].

FlowFormer [6] first constructs a matching cost volume and then processes it through an encoder–decoder architecture. The encoder compresses the cost volume into a low-dimensional cost embedding using a transformer-based module. The decoder integrates transformer blocks with a convolutional gated recurrent unit, which predicts optical flow increments Δf that are iteratively refined across multiple passes until producing the final optical flow. For uncertainty estimation, the same architecture is repurposed and retrained. Here, the decoder predicts incremental variance in log space, $\Delta \log \sigma$, which is exponentiated after refinement to yield the final uncertainty estimate $\hat{\Sigma}_f = diag(\sigma_u^2, \sigma_v^2)$.

B. Depth estimation

Given a single image, the objective is to estimate the pixel-wise depth $\hat{d} \in \mathbb{R}$ and, where applicable, its associated uncertainty $\hat{\sigma}_d$. To estimate the scene geometry from monocular images, we benchmark two state-of-the-art models: Depth Anything v2 (DAv2) and Depth Anything v3 (DAv3). Both models employ a DINOv2 (ViT) encoder, chosen for its ability to capture global scene context, object-level shape priors, and camera-aware semantic information. In both architectures, depth prediction is produced through a DPT-based decoder, which transforms camera-aware tokens from the encoder back into spatial feature maps and ultimately into (inverse) depth outputs.

a) DAv2: [19] combines a transformer-based DINOv2 encoder with a DPT decoder. Training follows a teacher–student scheme: (1) a teacher model based on DINOv2-G is trained using synthetic images, and (2) a student model using DINOv2-Small is then trained on real images to improve generalization. DAv2 predicts a single dense depth map but does not estimate uncertainty or confidence.

To incorporate uncertainty into the framework, we experimentally define the depth uncertainty $\hat{\sigma}_d$ as directly proportional to the predicted pixel depth $\hat{d}$, scaled by a factor of 0.1. This heuristic ensures that the uncertainty estimate grows with depth, reflecting the increased ambiguity typically associated with farther points in monocular depth estimation.

b) Depth Anything v3 (DAv3): DAv3 [10] extends the DAv2 architecture while enabling multi-view reasoning within a unified framework. The DINOv2 encoder is modified to accept an arbitrary number of input images without architectural changes. This is achieved via a cross-view self-attention mechanism: early transformer layers perform standard within-view attention, while deeper layers alternate between within-view and cross-view attention. This allows the model to aggregate geometric cues across multiple viewpoints. On the decoding side, DAv3 adopts a Dual-DPT head that jointly predicts dense depth maps and per-pixel ray directions. The predicted rays correspond to 3D directions from the camera center through each image pixel, thereby providing a geometrically consistent representation of the scene. Additionally, DAv3 estimates a per-pixel confidence map derived from the uncertainty of the ray predictions. This confidence, generated using an exponential activation, is strictly positive and unbounded, reflecting the internal consistency of the predicted rays with the implicit camera projection model. To align with the requirements of pose graph optimization, confidence values are converted to uncertainty measures by taking their inverse, ensuring a monotonically decreasing relationship.

C. Factor graphs for monocular VO

To integrate optical flow and depth estimates into a monocular visual odometry framework, we construct a factor graph using pose-to-point factors [4]. For each pair of consecutive frames, we use the pixel correspondences established by the optical flow $\hat{f} \in \mathbb{R}$ and the depth estimates $\hat{d} \in \mathbb{R}$ to project 2D pixel locations into 3D space. Specifically, the depth $\hat{d}$ is back-projected to obtain a 3D point $P \in \mathbb{R}^3$ in the camera frame of the reference image, using the camera intrinsics. This 3D point is used as a landmark in the factor graph.

The pose-to-point factor enforces geometric consistency by constraining the reprojection of P into the second frame to match the pixel location predicted by the optical flow. If the network provides an uncertainty estimate $\hat{\Sigma}_f$ for the optical flow or $\hat{\sigma}_d$ for the depth, these are incorporated into the factor's noise model to weight the constraint appropriately.

V. Experiments

We benchmark the proposed method across multiple datasets covering both underwater and surface conditions. We

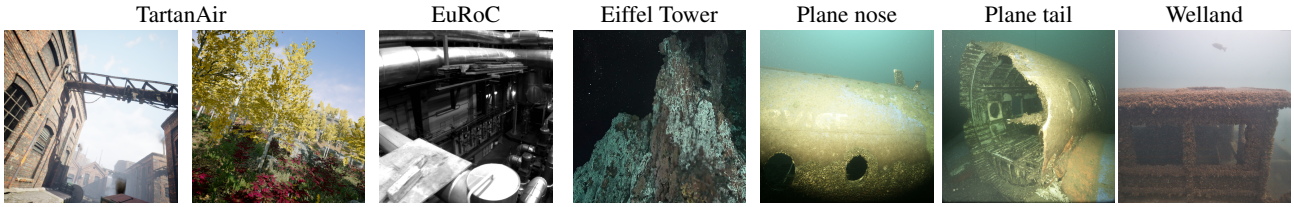


Fig. 2: Sample images from the evaluated datasets. TartanAir is a simulated dataset featuring sequences with both structured and unstructured elements. EuRoC comprises recordings of structured indoor environments. Eiffel Tower showcases natural seabed formations. Plane Nose and Plane Tail depict distinct sections of a sunken aircraft. Welland presents a sunken boat.

compare three different setups of our framework with the state-of-the-art network TartanVO [17], which serves as a strong learning-based baseline.

A. Visual Odometry Datasets

To evaluate the performance of visual odometry algorithms both in and out of underwater environments, experiments were conducted across a range of datasets. These datasets were selected to provide a comprehensive benchmark, encompassing surface, public underwater, and proprietary underwater recordings.

For surface conditions, two datasets were utilized. The **TartanAir** dataset [18] was partially used for training, allowing for an assessment of model behavior on within-distribution data. The **EuRoC** dataset [3], although recorded indoors, offers a robust benchmark due to its variability in speed and lighting conditions, as indicated by the difficulty levels assigned to each sequence (easy, medium, difficult).

In the underwater domain, public datasets were employed to facilitate comparative analysis and public benchmarking. The **Eiffel Tower** dataset [2] provides a deep-sea benchmark for long-term visual localization, featuring images from four visits to the same hydrothermal vent edifice over five years.

Additionally, proprietary underwater datasets were recorded using a Voyis Discovery Stereo camera, which delivers high-quality, high-resolution imagery. The **EIVA Plane Datasets** documents a sunken aircraft at the Vobster Quay diving facility in Radstock, UK, with comprehensive coverage and close-up views. The **EIVA Welland Dataset**, recorded at the Welland Scuba Park in Ontario, Canada, features a sunken boat, similarly captured with full coverage and close-up sequences. An overview of all datasets is shown in Fig. 2.

B. Uncertainty calibration

We perform uncertainty calibration using the TartanAir dataset [18], which provides ground truth for both dense optical flow and depth. This makes it particularly suitable for calibrating heterogeneous uncertainty sources within a unified framework. For calibration, we adopt a quantile-based approach, setting $\alpha = 0.90$ (the 90th quantile). This choice ensures that the calibrated uncertainties are conservative yet robust, covering 90% of the observed errors in the calibration dataset. The resulting quantile values for optical flow and depth are $q_{f,90} = 4.675$ and $q_{d,90} = 1.249$, respectively.

C. Quantitative metrics

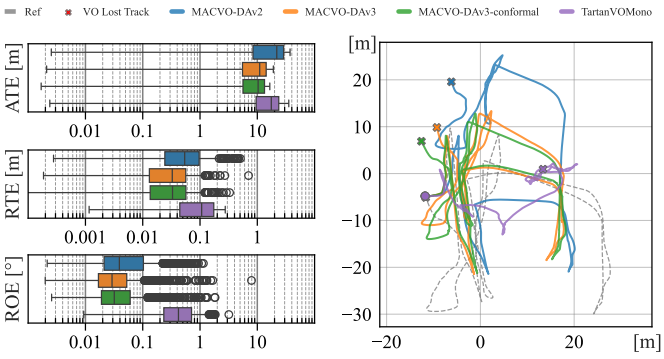
We evaluate trajectory accuracy using standard metrics in visual odometry and SLAM: Absolute Trajectory Error (ATE), Relative Translation Error (RTE), and Relative Orientation Error (ROE). ATE measures the global consistency between the estimated and ground-truth trajectories after alignment, while RTE and ROE quantify local errors in translation and rotation by comparing relative motions over short intervals.

All metrics are computed using the evo [5] evaluation toolkit, which provides a standardized implementation of trajectory alignment and error computation. Following common practice in monocular settings, we perform scale alignment between the estimated and ground-truth trajectories prior to evaluation to account for scale ambiguity.

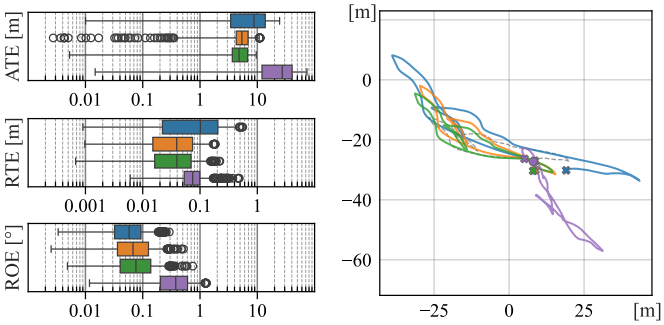
D. Experimental Analysis

The results are best understood by considering the level of domain shift across datasets. In the within-distribution setting (TartanAir), the learned models perform as expected, with MACVO-DAv3 consistently achieving the best performance across most sequences. The conformal variant further improves global consistency in several cases (e.g., *abandoned factory*, *amusement*, *ocean*; see Fig. 3), confirming that aligning uncertainty between depth and optical flow benefits optimization. However, even within this regime, failure cases such as *seansons forest* show that the learned uncertainty of DAv3 is not always reliable, allowing MACVO-DAv2 to remain competitive due to its simpler, geometry-driven uncertainty.

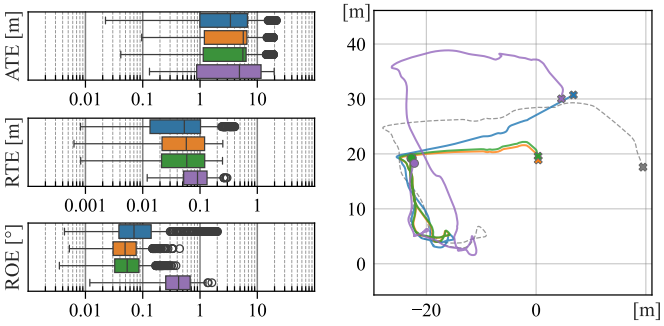
Under moderate domain shift (EuRoC), the learned models generalize well. MACVO-DAv3 provides the most accurate local motion estimates, while MACVO-DAv3-conformal improves global trajectory consistency, as seen in Fig. 4. This indicates that the proposed conformal calibration is effective when the uncertainty remains reasonably aligned with the data distribution. In strong out-of-distribution conditions, the behavior becomes more nuanced. On the *Eiffel Tower* dataset (Fig. 5), MACVO-DAv2 achieves the best global and local trajectory accuracy, outperforming both DAv3 variants and TartanVOMono. This contrasts with the surface datasets and highlights that the learned uncertainty in DAv3 becomes less reliable under significant domain shift, while the depth-based uncertainty of DAv2 provides more stable weighting. This trend is further reinforced in the proprietary underwater



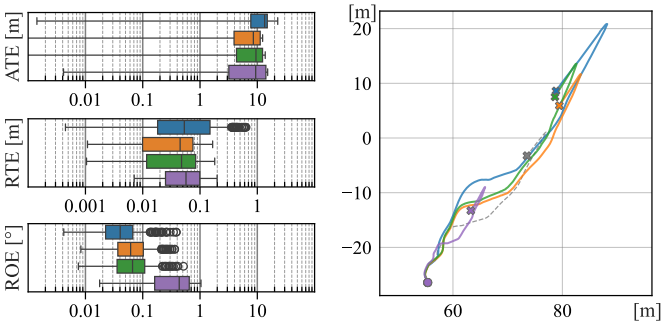
(a) Abandoned factory.



(b) Amusement.

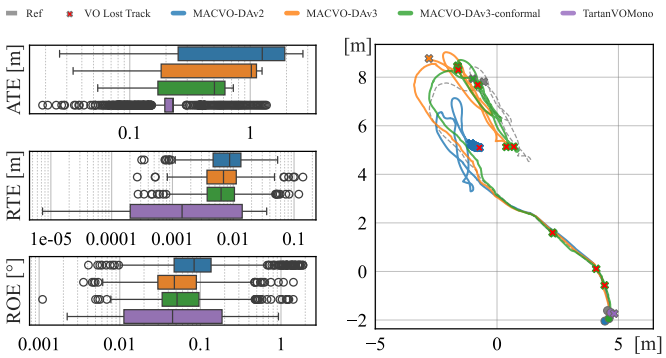


(c) Ocean.

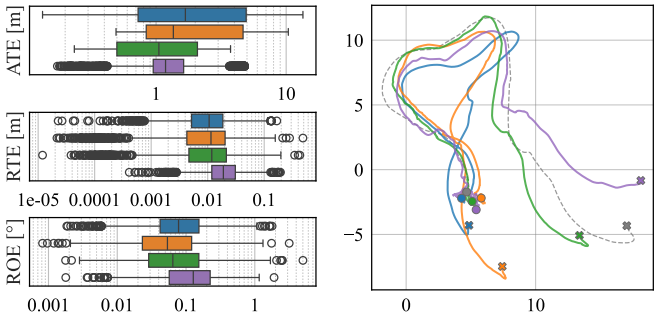


(d) Seasons forest.

Fig. 3: Results on TartanAir dataset. For each sequence, the error boxplot is shown on the left and the reference-versus-estimated XY comparison on the right.

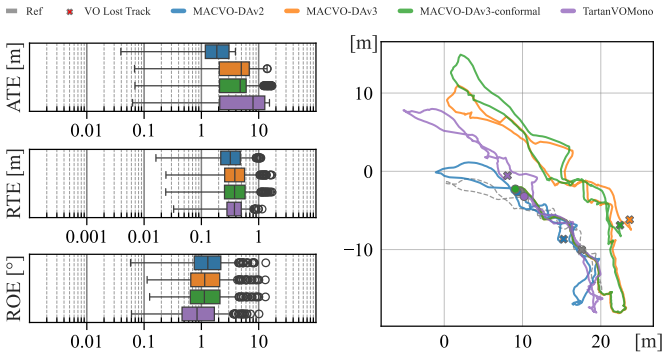


(a) MH01 Easy



(b) MH04 Difficult

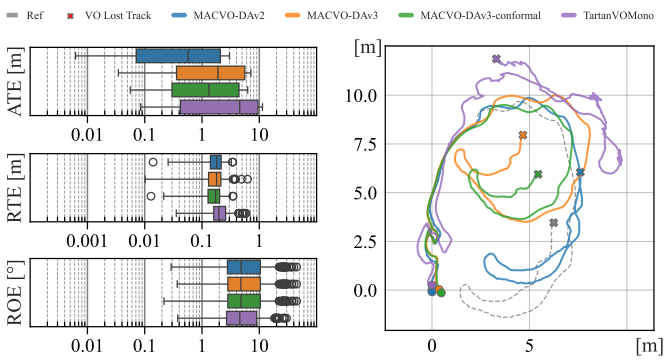
Fig. 4: Evaluation results on the EuRoC dataset. For clarity, only the first 1,200 poses are shown.



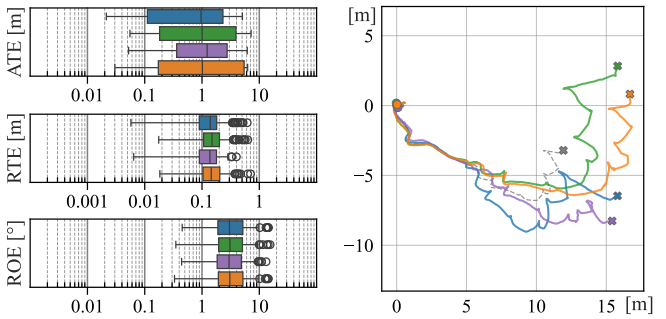
(a) 2020 sequence.

Fig. 5: Evaluation results on the Eiffel Tower dataset. For clarity, only a representative subset of 400 poses is displayed.

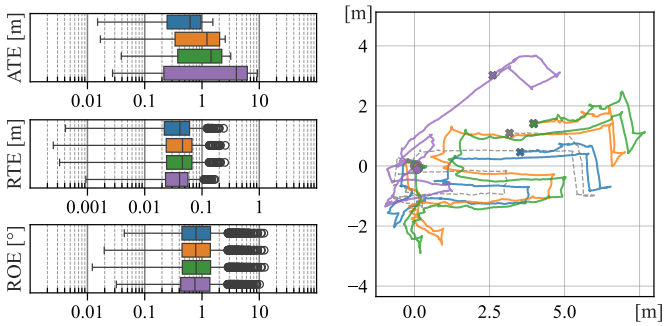
datasets, as shown by the results in Fig. 6. Here, MACVO-DAv3 tends to be overconfident, reducing its effectiveness during optimization, whereas MACVO-DAv2 often yields more stable results. The results also reveal a strong dependence on scene geometry: close-range observations of curved objects (e.g., *EIVA plane nose*) lead to scale inconsistencies and trajectory drift, while observing similar structures from a distance (*EIVA plane tail*) improves alignment. In predominantly planar environments (e.g., *Welland*), all methods become significantly more stable, with reduced ambiguity in depth estimation. Over-



(a) Plane nose.



(b) Plane tail.



(c) Welland.

Fig. 6: Evaluation results on EIVA's proprietary datasets. For clarity, only the first 400 poses of the trajectory are displayed.

all, the proposed framework demonstrates strong performance and generalization, with MACVO-DAv3 providing the best accuracy in-distribution and under moderate domain shift, and MACVO-DAv3-conformal improving robustness through better uncertainty alignment. However, the results clearly show that under strong domain shift, such as in underwater environments, uncertainty estimation becomes the limiting factor, with simpler heuristics (MACVO-DAv2) sometimes outperforming learned uncertainty. Improving the robustness of uncertainty estimation across domains and geometric conditions remains a key direction for future work.

E. Discussion

Across all datasets, three main conclusions emerge. First, MACVO-DAv3 provides the strongest overall performance in both within-distribution and moderately out-of-distribution scenarios, benefiting from improved learned representations. Second, the proposed conformal uncertainty calibration in MACVO-DAv3-conformal enhances robustness and often improves global trajectory consistency, particularly when uncertainty estimates are reasonably calibrated. Third, under strong domain shifts, learned uncertainty can become overconfident, reducing its effectiveness in optimization. In these cases, simpler uncertainty models, such as the depth-based heuristic used in MACVO-DAv2, can yield more stable results.

Overall, the results validate the proposed framework for modular monocular visual odometry and highlight the importance of properly calibrated uncertainty when combining learned depth and optical flow. Improving uncertainty calibration under domain shift remains a key future direction.

VI. Conclusion

In this work, we proposed a modular framework for monocular visual odometry that combines learning-based perception with geometry-based optimization. By decomposing the problem into optical flow, depth estimation, and pose optimization, the framework enables flexible integration of learned components while retaining the robustness of geometric methods.

A central contribution of this work is the explicit distinction between two forms of uncertainty handling: network-predicted uncertainty and offline conformal calibration. The former is learned jointly with the perception modules and reflects the model's internal confidence, while the latter is calculated and applied afterwards to correct miscalibration by aligning predicted uncertainty with empirical errors. Our results show that although learned uncertainty is effective in-distribution, it can become overconfident under domain shift. In contrast, conformal calibration provides a principled way to re-scale uncertainty estimates, improving their consistency across modules and enhancing optimization stability.

The experimental results demonstrate that the proposed approach achieves strong performance across diverse datasets. In within-distribution and moderately out-of-distribution scenarios, MACVO-DAv3 provides the best accuracy, while the conformal variant improves global trajectory consistency. Under strong domain shift, particularly in underwater environments, the limitations of learned uncertainty become evident, and the benefits of uncertainty calibration, although still present, are mitigated.

Future work will focus on further bridging the gap and exploring the benefits of learned and calibrated uncertainty. A first direction is to investigate calibration without retraining, deploying the conformal metrics obtained from a small set of samples from the target dataset. A second direction is to explore fine-tuning strategies, where both the network predictions and their uncertainty estimates are jointly adapted to new domains. Comparing these approaches will provide valuable insights into the trade-offs between generalization, robustness, and data requirements.

REFERENCES

[1] Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning **16**(4), 494–591 (2023)

[2] Boittiaux, C., Dune, C., Ferrera, M., Arnaubec, A., Marxer, R., Matabos, M., Van Audenhaege, L., Hugel, V.: Eiffel tower: A deep-sea underwater dataset for long-term visual localization. The International Journal of Robotics Research **42**(9), 689–699 (2023)

[3] Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M.W., Siegwart, R.: The euroc micro aerial vehicle datasets. The International Journal of Robotics Research **35**(10), 1157–1163 (2016)

[4] Dellaert, F., Kaess, M.: Factor graphs for robot perception. Foundations and Trends® in Robotics **6**(1-2), 1–139 (2017)

[5] Grupp, M.: evo: Python package for the evaluation of odometry and slam. https://github.com/michaelgrupp/evo (2017), accessed: 2026-03-27

[6] Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: Flowformer: A transformer architecture for optical flow. In: European conference on computer vision. pp. 668–685. Springer (2022)

[7] Hui, T.W., Loy, C.C.: Liteflownet3: Resolving correspondence ambiguity for more accurate optical flow estimation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 169–184. Springer (2020)

[8] Hui, T.W., Tang, X., Loy, C.C.: Liteflownet: A lightweight convolutional neural network for optical flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8981–8989 (2018)

[9] Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: Proceedings of the IEEE international conference on computer vision. pp. 2938–2946 (2015)

[10] Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)

[11] Qiu, Y., Chen, Y., Zhang, Z., Wang, W., Scherer, S.: Mac-vo: Metrics-aware covariance for learning-based stereo visual odometry mac-vo. github. io. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 3803–3814. IEEE (2025)

[12] Rosinol, A., Leonard, J.J., Carlone, L.: Nerf-slam: Real-time dense monocular slam with neural radiance fields. arXiv preprint arXiv:2210.13641 (2022)

[13] Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8934–8943 (2018)

[14] Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)

[15] Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Advances in Neural Information Processing Systems **34**, 16558–16569 (2021)

[16] Wang, S., Clark, R., Wen, H., Trigoni, N.: Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In: 2017 IEEE international conference on robotics and automation (ICRA). pp. 2043–2050. IEEE (2017)

[17] Wang, W., Hu, Y., Scherer, S.: Tartanvo: A generalizable learning-based vo. arXiv preprint arXiv:2011.00359 (2020)

[18] Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916. IEEE (2020)

[19] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)

[20] Zhan, H., Weerasekera, C.S., Bian, J.W., Garg, R., Reid, I.: Df-vo: What should be learnt for visual odometry? (2021)