Image and Vision Computing 162 (2025) 105674

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

CoHAtNet: An integrated convolutional-transformer architecture with hybrid self-attention for end-to-end camera localization

Hussein Hasan [a],*, Miguel Angel Garcia [b], Hatem Rashwan [a], Domenec Puig [a]

[a] Department of Computer Engineering and Mathematics, University Rovira i Virgili, Tarragona, 43007, Spain
[b] Department of Electronic and Communications Technology, Autonomous University of Madrid, Madrid, Spain

ARTICLE INFO

Implementation Link: https://github.com/Husseinhhameed/CoHAtNet

Keywords:
Camera localization
Hybrid CNN-transformers
CoAtNet
Hybrid self-attention

ABSTRACT

Camera localization refers to the process of automatically determining the position and orientation of a camera within its 3D environment from the images it captures. Traditional camera localization methods often rely on Convolutional Neural Networks, which are effective at extracting local visual features but struggle to capture long-range dependencies critical for accurate localization. In contrast, Transformer-based approaches model global contextual relationships appropriately, although they often lack precision in fine-grained spatial representations. To bridge this gap, we introduce CoHAtNet, a novel Convolutional Hybrid-Attention Network that tightly integrates convolutional and self-attention mechanisms.

Unlike previous hybrid models that stack convolutional and attention layers separately, CoHAtNet embeds local features extracted via Mobile Inverted Bottleneck Convolution blocks directly into the Value component of the self-attention mechanism of Transformers. This yields a hybrid self-attention block capable of dynamically capturing both local spatial detail and global semantic context within a single attention layer. Additionally, CoHAtNet enables modality-level fusion by processing RGB and depth data jointly in a unified pipeline, allowing the model to leverage complementary appearance and geometric cues throughout.

Extensive evaluations have been conducted on two widely-used camera localization datasets: 7-Scenes (RGB-D) and Cambridge Landmarks (RGB). Experimental results show that CoHAtNet achieves state-of-the-art performance in both translation and orientation accuracy. These results highlight the effectiveness of our hybrid design in challenging indoor and outdoor environments. This makes CoHAtNet a strong candidate for end-to-end camera localization tasks.

1. Introduction

Camera localization, also known as camera pose estimation, involves determining the six-degrees-of-freedom (6DoF) pose of a camera from a single captured image. This task plays a fundamental role in a wide range of applications, including robotics, autonomous navigation, augmented reality (AR), and virtual reality (VR). Due to its potential for low-cost and scalable deployment, camera localization has attracted increasing attention as an alternative to more expensive sensing technologies such as LiDAR [1].

Traditional approaches, such as image retrieval and 3D structure matching, heavily rely on large, densely annotated datasets and computationally intensive feature matching algorithms. These methods are often brittle in real-world conditions, particularly in scenes with clutter, repetitive structures, or limited texture.

The emergence of deep learning has led to the adoption of Convolutional Neural Networks (CNNs) for direct pose regression. CNNs excel at extracting local visual features but inherently struggle to model long-range dependencies and global context, both of which are crucial for precise localization in complex environments. In recent years, vision Transformers [2] have emerged as a ground-breaking technology in computer vision, revolutionizing object detection, image classification, and scene understanding tasks. Unlike CNNs, which are limited by their local receptive fields, vision Transformers leverage self-attention mechanisms to capture global dependencies and contextual relationships within an image. This global perspective has proven particularly

* Corresponding author.
E-mail addresses: husseinhasanhameed.al-sinayyid@estudiants.urv.cat (H. Hasan), miguelangel.garcia@uam.es (M.A. Garcia), hatem.abdellatif@urv.cat (H. Rashwan), domenec.puig@urv.cat (D. Puig).

https://doi.org/10.1016/j.imavis.2025.105674
Received 10 November 2024; Received in revised form 14 June 2025; Accepted 16 July 2025
Available online 26 July 2025

valuable for tasks such as camera localization, where understanding spatial relationships across the entire scene is critical. However, while vision Transformers excel at modeling long-range dependencies, they may struggle with fine-grained local feature extraction.

To leverage the strengths of both paradigms, hybrid architectures have been proposed. CoAtNet [3] is one such model that stacks convolutional layers in the early stages and Transformer layers in the later stages, aiming to balance spatial detail and semantic context. While effective in image classification tasks, CoAtNet exhibits limitations in geometric vision problems such as camera localization. Specifically, the separation of convolution and self-attention impedes the joint modeling of local and global features. Moreover, its self-attention mechanism applies simple linear projections to compute the Value component, which inadequately captures spatial structure.

In this work, we introduce CoHAtNet, a novel Convolutional Hybrid-Attention Network that extends CoAtNet with a tightly integrated attention mechanism. Rather than stacking convolutional and Transformer stages separately, CoHAtNet embeds Mobile Inverted Bottleneck Convolution (MBConv) blocks [4] right into the Value branch of self-attention blocks. This hybrid formulation allows local spatial details and global semantic relationships to be modeled concurrently within each Transformer stage, resulting in more expressive and geometry-aware representations.

This architectural fusion proves especially beneficial in scenes with clutter, occlusion, or challenging lighting—conditions where CNNs or Transformers alone tend to underperform. Furthermore, CoHAtNet supports multi-modal learning by processing RGB and depth information in a unified pipeline. This early and consistent fusion of appearance and geometric cues enhances robustness across diverse environments. To the best of our knowledge, CoHAtNet is the first end-to-end model to explore such hybridization specifically for the task of 6DoF camera localization. The two main contributions of this work are summarized below:

- **Hybrid Self-Attention Mechanism:** A novel attention design that incorporates MBConv-derived convolutional features into the Transformer's Value component—enabling simultaneous modeling of local and global features—is integrated within a unified processing pipeline for RGB and depth data, allowing the model to jointly learn appearance and geometry for improved pose estimation.
- **Unified CNN-Transformer Architecture:** We extend CoAtNet by fusing convolution and self-attention directly within each Transformer stage, enhancing the model's ability to capture detailed and contextual cues simultaneously.

We evaluate CoHAtNet on two widely used benchmarks: the Microsoft 7-Scenes RGB-D dataset and the Cambridge Landmarks RGB dataset. The proposed method achieves state-of-the-art accuracy in both translation and rotation measures.

This paper is organized as follows. Section 2 introduces the fundamental concepts related to vision-based camera localization, including image-based methods (2.1), structure-based methods (Section 2.2), and regression-based methods (Section 2.3). Then, we summarize recent work related to Absolute Pose Regression (Section 2.3.1), Relative Pose Regression (Section 2.3.2), Scene Coordinate Regression (Section 2.3.3), and Multi-Scene Absolute Pose Regression (Section 2.3.4). Section 3 describes the proposed hybrid self-attention scheme. The original CoAtNet model is first summarized in Section 3.1. Based on it, the proposed CoHAtNet neural model is described in Section 3.2, including the new hybrid Transformer block (Section 3.2.1). The loss function utilized to train the model is introduced in Section 3.3, whereas implementation details are described in Section 3.4. The experimental evaluation is presented in Section 4, including the 7-Scenes dataset (Section 4.2.1), and the Cambridge Landmarks dataset (Section 4.2.2), in addition to a comparison with the state-of-the-art in camera localization. Section 4.2.3 presents an analysis of the attention heatmaps generated by CoHAtNet, the original CoAtNet, and several competitive baselines, providing deeper insights into the spatial focus and structural differences across architectures. Section 4.2.4 conducts a parameter sensitivity analysis to evaluate the robustness of CoHAtNet under variations in key hyper-parameters. Finally, conclusions and future lines are given in Section 5.

2. Related work

Vision-based camera localization utilizes visual data to determine the spatial position and orientation of a camera within a scene. Given an input image $\mathbf{I}$ captured by the camera from an unknown viewpoint, the camera pose can be expressed as:

$$\mathbf{p} = (\mathbf{T}, \theta) \tag{1}$$

where $\mathbf{T} = (x, y, z)$ denotes the translation vector, and θ represents the 3D orientation of the camera. Broadly speaking, vision-based camera localization methods can be categorized as image-based methods, structure-based methods, and regression-based methods [1]. They are summarized below.

2.1. Image-based camera localization

The classical approach to camera localization is constituted by image-based methods that rely on image retrieval techniques for estimating the camera pose. A query image is matched to a database of reference images with known 6DoF poses [5]. The camera pose is determined either by selecting the most visually similar reference image or by interpolating poses from the Nearest Neighbors [6]. However, the density and coverage of the database heavily influences accuracy: sparse databases can lead to significant deviations between the estimated and actual poses, thereby limiting their suitability for applications requiring high precision.

2.2. Structure-based camera localization

Structure-based localization approaches establish correspondences between the 2D position of every pixel from the query image, and the 3D coordinates of the respective scene's point. Key-point detection is achieved by using local descriptors, followed by the application of a Perspective-n-Point algorithm (PnP), often combined with Random Sample Consensus (RANSAC) [7]. Traditional structure-based approaches require matching in a full 3D map. Therefore, the search space is very large and often computationally expensive [8]. More recently, deep learning techniques have been applied in conjunction with structural approaches. For example, MatchFormer [9] interleaves self-attention for feature extraction and cross-attention for feature matching. Then, RANSAC with matched correspondences is applied for camera localization. EAAINet [10] contains a Global Affinity Aggregation Module and an Element-wise Attention Module to improve feature extraction and scene parsing, followed by a RANSAC-based PnP algorithm. ALNet [11] employs a Local Discrepancy Perception Module, followed by an Adaptive Channel Attention Module to distinguish similar image patches and integrate multilevel features. RANSAC-based PnP is finally applied for accurate camera localization. Although structure-based methods perform well in small-scale environments, they face difficulties when applied to large-scale scenes, as the complexity of matching and maintaining a robust correspondence increases significantly. Additionally, in complex scenes, where repetitive local features are present, structure-based methods may struggle to correctly match features, leading to localization failures. This issue can be particularly challenging in environments with similar or ambiguous features.

2.3. Regression-based camera localization

Regression-based methods apply deep neural networks to directly regress the camera's 6DoF from visual data. These methods can be classified into four broad categories: Absolute Pose Regression, Relative Pose Regression, Scene Coordinate Regression, and Multi-Scene pose Regression.

2.3.1. Absolute Pose Regression

Absolute Pose Regression methods focus on directly regressing the camera's 6DoF from a single captured image. A foundational contribution in this area is PoseNet [12], which uses an end-to-end CNN to localize the camera pose from an input image. Although its performance initially did not surpass structure-based approaches, subsequent developments have significantly enhanced its capabilities. For example, Bayesian PoseNet [13] introduced a Bayesian framework to quantify pose uncertainty and improve robustness under real-world conditions.

Camera localization also improved with the advent of Transformers [14]. Transformers differ from convolutional networks in that they apply self-attention mechanisms to dynamically capture long-range dependencies and contextual information from the entire image. This global perspective enhances feature representation, improves robustness to changes in appearance, and enables better generalization across diverse environments. AtLoc [15] applies absolute pose regression by leveraging an attention mechanism that drives the neural network towards the most stable geometric features. The neural architecture has three primary components: Visual Encoder, Attention Module, and Pose Regressor. HyperPose [16] is an attention hyper-network for camera localization. The main network comprises an EfficientNet-B0 model as the backbone for processing the input query image and producing intermediate activation maps. Parallel to the main network, the hyper-network is designed to predict the weights for the regression heads of the main network. Recently, Neural Feature Synthesizer [17] improves absolute pose regression by incorporating implicit geometric constraints during the evaluation stage. In contrast to traditional methods, which mostly rely on 2D operations, it enhances pose estimation accuracy by encoding 3D geometric features.

2.3.2. Relative pose regression

The relative pose regression approach estimates the relative 6DoF motion between consecutive or non-consecutive frames [18]. One of the early models for relative pose regression was Relative NN [19], which introduces an end-to-end approach for regressing the relative pose between two cameras. The method employs a Siamese Hybrid-CNN architecture that integrates a pre-trained AlexNet network with two branches. Later, GL-Net [20] applies a Graph Neural Network (GNN) for multi-frame relative pose estimation, by allowing information exchange between non-consecutive frames. This enhances robustness. In turn, AnchorNet [21] employs uniformly distributed anchor points across the scene to estimate their relative positions in the query image. This anchor-based methodology facilitates the computation of camera poses by interpolating predicted anchor positions. UA-Fusion [22] fuses the geometric solver and a deep neural network through an uncertainty-based fusion framework. A self-attention graph neural network captures relationships between key points, enabling end-to-end training. This leverages both geometric and learned uncertainties for camera localization. The maic drawback of relative pose regression methods is the limited performance in feature-poor environments, such as long corridors or open spaces, where being able to distinguish between different frames becomes a challenge.

2.3.3. Scene coordinate regression

Scene coordinate regression techniques focus on the estimation of 3D scene coordinates directly from input images, which is an essential requirement for camera localization tasks that involve mapping image pixels to their corresponding 3D locations in a pre-constructed environment model. These methods are particularly effective when accurate 3D reconstructions of the environment are available. For instance, DSAC (Differentiable SAmple Consensus) applies a probabilistic selection mechanism to make the traditional RANSAC algorithm differentiable. This allows its integration into a deep learning pipeline for end-to-end training [23]. Building on this approach, the Single-View Scene Coordinate Regression (SV-SCR) method further simplifies the pipeline by using a fully-convolutional neural network that directly regresses 3D scene coordinates from a single image, achieving accurate localization without relying on pre-constructed 3D models [24].

2.3.4. Multi-scene Absolute Pose Regression

Multi-scene absolute pose regression methods are intended to generalize camera pose estimation over multiple environments with a single deep neural network rather than training different models over each scene. The general approach involves a unified network architecture that can classify the type of scene corresponding to the input image, and subsequently adapt its internal parameters to estimate the camera pose accurately. For example, Multi-Scene PoseNet [25] is a model that first determines the specific scene of the input image using a shared CNN backbone. This backbone produces an activation map for classifying the scene and pose regression. It must be trained on multiple scenes to be effective. To address this gap, MS Transformer [26] proposes a method using two separate Transformers blocks, as encoders for positional and orientational features. It then applies decoders to generate scene-specific pose predictions. The MS-Transformer [26] model improves upon traditional absolute pose regression methods by allowing a single model to handle multiple scenes simultaneously. Similarly, coarse-to-fine multi-scene pose regression [27] has been proposed. It applies the Transformer architecture in combination with coarse-to-fine classification-regression in order to improve localization accuracy across multiple scenes.

Fig. 1 summarizes the limitations of the camera localization methods discussed above. While regression-based approaches have shown remarkable progress, they face limitations in terms of accuracy, generalization, and interpretability [28]. To bridge these gaps, we propose a hybrid model that leverages CNNs' local feature extraction capabilities and Transformers' global contextualization capabilities. This integration makes the model more powerful in learning geometrical features, resulting in improved accuracy and robustness in camera localization tasks for both indoor and outdoor environments.

3. Hybrid convolution-transformer architecture

3.1. Original CoAtNet model

CoAtNet [3] is a hybrid neural network that sequentially combines convolutional layers and Transformer blocks to benefit from both local inductive biases and global contextual modeling. Its architecture is structured in five stages (S0–S4), as shown in Fig. 2. The network begins with a convolutional stem (Stage S0) composed of two 3×3 convolutional layers, each followed by batch normalization and GELU activation. This stage reduces spatial dimensions while expanding channel capacity to prepare features for deeper processing.

Stages 1 and 2 apply MBConv blocks with residual connections to extract local spatial features efficiently. In Stages 3 and 4, CoAtNet transitions to Transformer layers, modeling long-range dependencies across the entire feature map. The final features are globally pooled and passed through a fully-connected layer for downstream prediction. The depth and width of these stages vary across CoAtNet variants (CoAtNet-0 to CoAtNet-4), enabling a flexible tradeoff between accuracy and

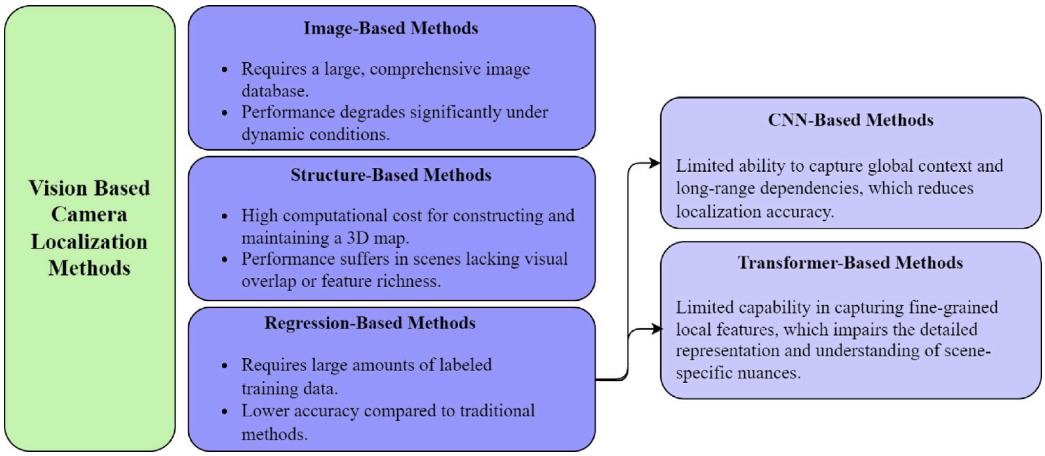

Fig. 1. Overview of limitations of various camera localization methods [1,28,29].

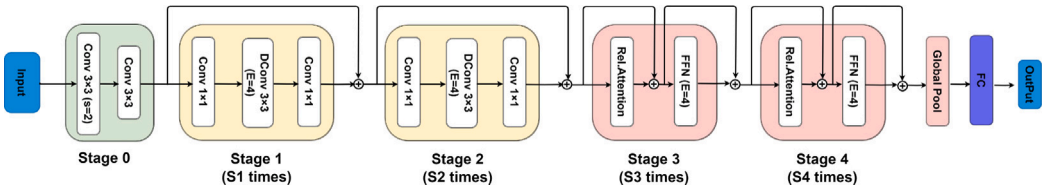

Fig. 2. Overview of the CoAtNet architecture (Stages S0 to S4). CoAtNet variants (CoAtNet-0 to CoAtNet-4) vary in depth and width to balance efficiency and accuracy. For instance, CoAtNet-0 starts with 2 layers and 64 channels in S0, while CoAtNet-4 scales this to 192. Channel capacity in the final stage (S4) increases from 768 to 1536 across variants, with CoAtNet-1 to CoAtNet-3 providing intermediate scaling [3].

efficiency. For instance, CoAtNet-0 begins with 2 layers and 64 channels in the stem (Stage S0), while CoAtNet-4 scales this to 192 channels. In the final Transformer stage (S4), channel capacity ranges from 768 in CoAtNet-0 to 1536 in CoAtNet-4. Intermediate variants progressively scale both depth and width, bridging lightweight and high-capacity configurations.

Despite its success for image classification, CoAtNet exhibits critical limitations in geometric vision tasks such as 6DoF camera localization. The strict separation between convolution and self-attention stages limits the capacity to simultaneously capture local detail and global structure. Furthermore, its self-attention mechanism relies on linear projections for computing the Value component, which lacks spatial sensitivity. These constraints hinder CoAtNet's ability to model the fine-grained geometric relationships essential for precise pose estimation. We tackle these limitations by integrating the convolutional and self-attention mechanisms within a same block. In particular, a lightweight convolutional branch produces a spatially-aware Value tensor, enabling the joint encoding of local geometry and global context. This fusion improves performance on geometry-sensitive tasks like 6-DoF pose estimation with minimal computational overhead.

3.2. Proposed CoHAtNet model

As previously discussed, accurate 6DoF camera localization requires capturing both fine-grained local features and long-range global context [30]. While CNNs effectively extract local patterns such as textures and edges [31], they often lack global representational capacity. Conversely, Transformers excel at modeling global dependencies via self-attention but may lose spatial precision needed for fine-grained geometric reasoning [32]. Furthermore, incorporating depth information is critical, as it provides explicit geometric cues that enhance the model's ability to infer 3D structure and disambiguate visual correspondences—factors essential for precise pose estimation. Recent studies demonstrate that utilizing depth data significantly improves pose estimation accuracy, particularly in challenging scenarios involving occlusion or poor lighting conditions [33].

To address these complementary limitations, we propose CoHAtNet, a hybrid architecture that unifies convolutional and attention-based representations within a single end-to-end framework. As illustrated in Fig. 3, CoHAtNet follows the five-stage architecture of CoAtNet, beginning with a convolutional stem (Stage S0) comprising two 3×3 convolutional layers with Batch Normalization and GELU activation. Stages 1 and 2 employ MBConv blocks with residual connections to efficiently extract spatially localized geometric features.

However, unlike CoAtNet, which separates convolution and attention into distinct stages, CoHAtNet introduces a novel hybrid Transformer block in Stages 3 and 4. Within each of these blocks, local features derived from MBConv blocks are directly injected into the Value (V) branch of the self-attention mechanism. This design allows local and global cues to interact within the same computational layer, enabling the network to model both structural detail and semantic relationships more effectively than prior approaches.

CoHAtNet also supports multi-modal input processing by integrating RGB and depth information into a unified stream. When depth is available, the network receives a 4-channel input—concatenating RGB with depth—and fuses them early in the pipeline. This modality-aware design enables joint propagation of appearance and geometric cues, enhancing robustness in cluttered, repetitive, or texture-sparse environments.

A detailed formulation of the hybrid Transformer block is provided in Section 3.2.1.

3.2.1. Architecture of hybrid transformer blocks

The core architectural innovation of CoHAtNet lies in its hybrid Transformer block, which enhances the self-attention mechanism by integrating local spatial information directly into the computation of the Value (V) component.

Fig. 4 contrasts this hybrid design with the standard Transformer block employed in CoAtNet. In both architectures, the input feature map $\mathcal{F} \in \mathbb{R}^{H \times W \times C}$ — where H and W are spatial dimensions and C is the number of channels — is reshaped into a sequence $X \in \mathbb{R}^{HW \times C}$.

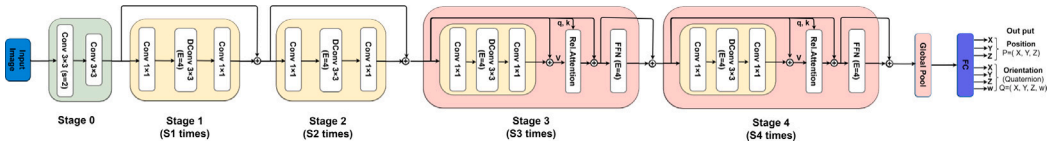

Fig. 3. Proposed CoHAtNet architecture that hybridizes MBConv blocks for local feature extraction and self-attention for global context modeling.

The Query (Q) and Key (K) matrices are then obtained through standard linear projections: $Q = XW_Q$ and $K = XW_K$. These are used to compute the attention weights via the standard scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2}$$

In conventional Transformers, the Value (V) matrix is also generated through a linear projection: $V = XW_V$. However, CoHAtNet departs from this paradigm. Instead of a linear layer, it employs an MBConv block to process the original feature map $\mathcal{F}$, producing a spatially enriched representation $\mathcal{C} = \text{MBConv}(\mathcal{F})$, which is reshaped into $V \in \mathbb{R}^{HW \times C}$.

The MBConv block, adapted from EfficientNet [4], is designed to efficiently capture local geometric structures. It expands the input channels via a 1×1 pointwise convolution, applies a 3×3 depthwise convolution to model spatial dependencies, and projects the result back using another 1×1 convolution. A Squeeze-and-Excitation (SE) module further enhances this representation by adaptively recalibrating channel-wise responses. The operation can be formally expressed as:

$$F_{\text{MBConv}} = F + \text{BN}\left(\text{SE}\left(\text{DWConv}\left(\text{BN}(F * W_e)\right)\right) * W_p\right), \tag{3}$$

where F is the input feature map, W_e and W_p are learnable 1×1 convolution weights, and $*$ denotes convolution.

After obtaining Q, K, and V, the attention output is computed using (2). This output is then fused with the original input via residual connection and finally normalized:

$$\widehat{\mathcal{F}} = \text{LN}\left(\mathcal{F} + \text{Attention}(Q, K, V)\right) \tag{4}$$

In this formulation, Q and K provide global context, while V contributes with detailed local features. As a result, each position in $\widehat{\mathcal{F}}$ reflects both its MBConv-enhanced local representation and its global semantic relevance determined by self-attention. This tight integration of local and global information fosters more geometry-aware representations, making the architecture particularly effective for tasks such as precise 6DoF camera pose estimation.

3.3. Homography-based loss function

Traditional loss functions for camera localization, including PoseNet loss [12], homoscedastic uncertainty loss, and geometric reprojection loss [34], combine translation and rotation errors into a single measure. The PoseNet and homoscedastic uncertainty losses require challenging multi-objective tuning, while the geometric reprojection loss can be unstable due to its reliance on ground-truth 3D scene points. To overcome these limitations, we applied a homography-based loss function proposed in [35], which integrates homographies across multiple virtual planes within a scene, providing a more stable and interpretable approach.

The loss function measures the error between the ground-truth and the estimated camera poses by considering the homographies induced by planes parallel to the ground-truth sensor plane. The error is quantified as the difference between the identity matrix $\mathbf{I}$ and the homography matrix $\mathbf{H}$, computed from the ground-truth and estimated poses. This error is expressed using the Frobenius norm:

$$L_H = \frac{1}{x_{\max} - x_{\min}} \int_{x_{\min}}^{x_{\max}} \|\mathbf{I} - \mathbf{H}\|_F^2 \, dx \tag{5}$$

where $x_{\min}$ and $x_{\max}$ represent the minimum and maximum distances to the planes containing the observations, and $\mathbf{H}$ is the homography matrix that maps points between the ground-truth and the estimated views. This formulation allows for an effective approximation of the reprojection error, promoting convergence as the estimated pose aligns with the ground-truth.

3.4. Implementation details

We implemented the proposed CoHAtNet model using PyTorch with the Adam optimizer [36]. The initial learning rate was 0.0001, and batch size 16. During training, a ReduceLROnPlateau scheduler was adopted, which was set to drop the learning rate by a factor of 0.1 when the validation loss failed to improve over ten epochs. The input images were resized to 256×256 pixels for both training and validation. We trained the model using K-fold cross-validation with K = 5 splits, and 150 training epochs per split. To evaluate the impact of model complexity, five different configurations of CoHAtNet were generated. These configurations were designed to align with the varying depths and channel widths found in the different CoAtNet configurations, providing a direct comparison between our proposed CoHAtNet architecture and the different CoAtNet variants. We used the homography loss function (5), with x_{min} and x_{max} set based on the 2.5th and 97.5th percentiles of the depth distribution of the dataset [35]. Training and testing were performed on an NVIDIA-A100 GPU (see Table 1).

4. Experimental results

4.1. Dataset

To evaluate the proposed CoHAtNet model, we used two widely recognized benchmark datasets: 7-Scenes [37] and Cambridge Landmarks [12]. These datasets encompass various challenging scenarios in both indoor and outdoor environments. The 7-Scenes dataset consists of RGB-D image sequences captured in different indoor scenes: Chess, Fire, Heads, Office, Pumpkin, Red Kitchen, and Stairs. It provides a total of 26,000 training images and 17,000 testing images, The dataset has an average spatial extent of 4×3 m, making it suitable for evaluating models in small-scale indoor environments.

In contrast, the Cambridge Landmarks dataset encompasses sequences of RGB images taken from large-scale outdoor locations, including King's College, Old Hospital, St. Mary's Church, and Shop Facade. This dataset includes 8380 training images and 4841 testing images, with average spatial extent of 100×500 m, This dataset presents diverse challenges for camera localization in large-scale outdoor settings.

4.2. Evaluation

4.2.1. 7-Scenes dataset

The evaluation results of the proposed CoHAtNet models on the 7-Scenes dataset are summarized in Tables 2 and 3. We trained CoHAtNet with both RGB (3-channel input) and RGB-D (4-channel input) data. For comparison purposes, we also trained the CoAtNet-0 model using both RGB and RGB-D data. The performance for each scene is reported in terms of mean translation error (in meters) and mean orientation error (in degrees). The "Avg" column shows the mean performances across all scenes.

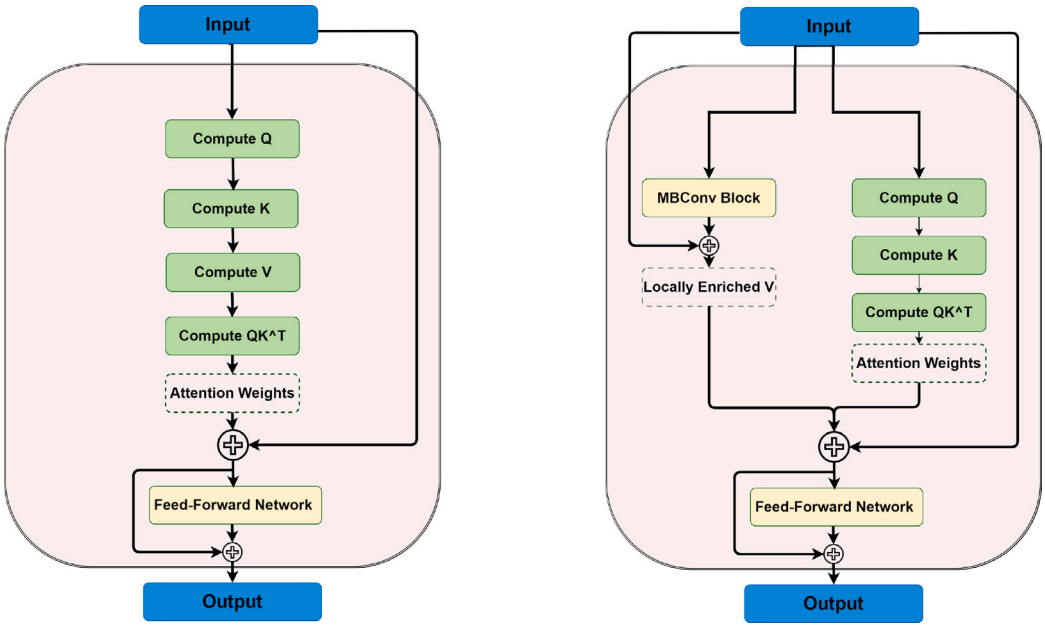

Fig. 4. Comparison between the standard Transformer block (*left*) used in CoAtNet, and the proposed hybrid Transformer block (*right*) used in CoHAtNet. In the standard block, Query (Q), Key (K), and Value (V) are computed via linear projections. In the hybrid version, V is enriched with local features using an MBConv block, allowing CoHAtNet to integrate global self-attention with spatial context for improved localization.

Table 1
Architecture details of CoHAtNet variants, showing the number of blocks (L) and channels (D) for each stage.

Stage	CoHAtNet-0	CoHAtNet-1	CoHAtNet-2	CoHAtNet-3	CoHAtNet-4
S0: Conv Stem	L = 2, D = 64	L = 2, D = 64	L = 2, D = 128	L = 2, D = 192	L = 2, D = 192
S1: MBConv	L = 2, D = 96	L = 2, D = 96	L = 2, D = 128	L = 2, D = 192	L = 2, D = 192
S2: MBConv+SE	L = 3, D = 192	L = 6, D = 192	L = 6, D = 256	L = 6, D = 384	L = 12, D = 384
S3: HTransformer	L = 5, D = 384	L = 14, D = 384	L = 14, D = 512	L = 14, D = 768	L = 28, D = 768
S4: HTransformer	L = 2, D = 768	L = 2, D = 768	L = 2, D = 1024	L = 2, D = 1536	L = 2, D = 1536
Final output size = 7					

Table 2
Translation and rotation errors of CoAtNet-0 and CoHAtNet models on the 7-Scenes dataset (classes: “Chess”, “Fire”, “Heads”, “Office”).

Method	Chess		Fire		Heads		Office	
CoAtNet-0 (RGB)	11 cm	4.27°	12 cm	5.12°	14 cm	7.45°	10 cm	8.76°
CoAtNet-0 (RGB-D)	7 cm	1.94°	7 cm	2.01°	6 cm	3.63°	6 cm	3.21°
CoHAtNet-0 (RGB)	3 cm	0.65°	2 cm	0.69°	3 cm	1.89°	2 cm	0.97°
CoHAtNet-1(RGB)	3 cm	0.61°	2 cm	0.72°	2 cm	1.73°	2 cm	0.90°
CoHAtNet-2 (RGB)	3 cm	0.63°	2 cm	0.66°	2 cm	1.51°	2 cm	0.45°
CoHAtNet-3 (RGB)	2 cm	0.59°	2 cm	0.61°	2 cm	1.46°	2 cm	0.75°
CoHAtNet-4 (RGB)	2 cm	0.55°	2 cm	0.58°	2 cm	1.37°	1 cm	0.71°
CoHAtNet-0 (RGB-D)	2 cm	0.60°	2 cm	0.59°	2 cm	1.64°	2 cm	0.81°
CoHAtNet-1 (RGB-D)	2 cm	0.52°	1 cm	0.64°	2 cm	1.33°	2 cm	0.85°
CoHAtNet-2 (RGB-D)	2 cm	0.51°	1 cm	0.58°	2 cm	1.25°	2 cm	0.62°
CoHAtNet-3 (RGB-D)	2 cm	0.52°	1 cm	0.53°	2 cm	1.19°	1 cm	0.60°
CoHAtNet-4 (RGB-D)	1 cm	0.49°	1 cm	0.51°	2 cm	1.02°	1 cm	0.56°

Table 3
Translation and rotation errors of CoAtNet-0 and CoHAtNet models on the 7-Scenes dataset (classes: “Pumpkin”, “Kitchen”, “Stairs”; and average of 7 classes).

Method	Pumpkin		Kitchen		Stairs		Avg	
CoAtNet-0 (RGB)	11 cm	5.77°	13 cm	4.71°	16 cm	7.17°	12 cm	6.17°
CoAtNet-0 (RGB-D)	8 cm	1.26°	7 cm	1.74°	12 cm	1.87°	7 cm	2.23°
CoHAtNet-0 (RGB)	4 cm	0.93°	2 cm	1.12°	4 cm	1.01°	3 cm	1.03°
CoHAtNet-1 (RGB)	3 cm	0.91°	2 cm	0.95°	3 cm	0.92°	2 cm	0.96°
CoHAtNet-2 (RGB)	3 cm	0.88°	2 cm	0.89°	2 cm	0.90°	2 cm	0.84°
CoHAtNet-3 (RGB)	2 cm	0.81°	2 cm	0.76°	2 cm	0.84°	2 cm	0.83°
CoHAtNet-4 (RGB)	1 cm	0.63°	2 cm	0.61°	2 cm	0.74°	2 cm	0.74°
CoHAtNet-0 (RGB-D)	3 cm	0.81°	2 cm	1.0°	4 cm	1.1°	2 cm	0.93°
CoHAtNet-1 (RGB-D)	2 cm	0.66°	2 cm	0.73°	3 cm	0.93°	2 cm	0.80°
CoHAtNet-2 (RGB-D)	2 cm	0.58°	1 cm	0.63°	2 cm	0.58°	2 cm	0.67°
CoHAtNet-3 (RGB-D)	1 cm	0.50°	2 cm	0.51°	2 cm	0.52°	1 cm	0.62°
CoHAtNet-4 (RGB-D)	1 cm	0.49°	2 cm	0.52°	1 cm	0.46°	1 cm	0.57°

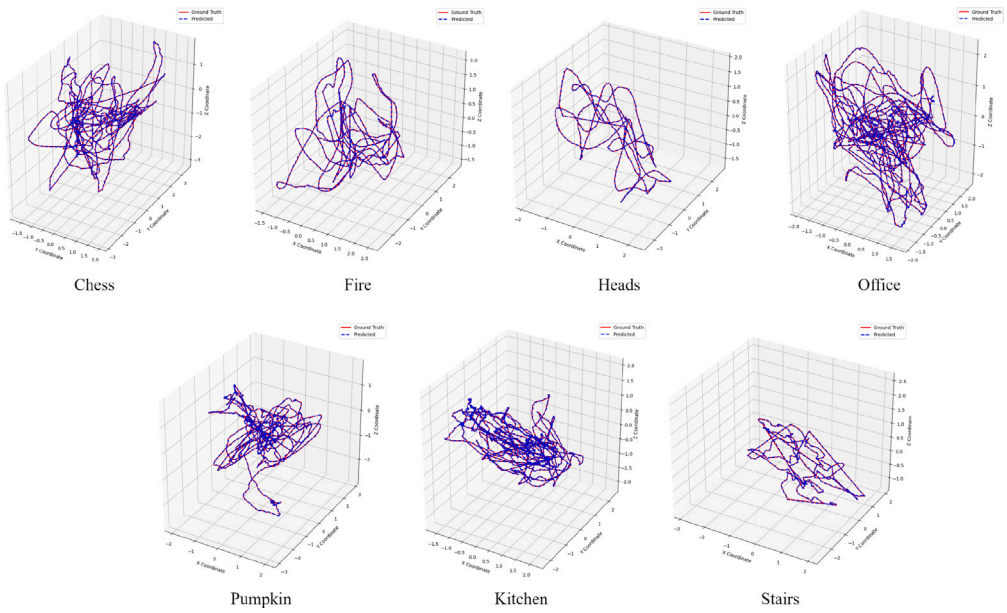

Fig. 5. Visualization of camera trajectories on the 7-Scenes dataset: red lines indicate ground-truth trajectories, while blue dots represent trajectories estimated by CoHAtNet.

The results indicate that all configurations of the CoHAtNet model perform competitively across all the scenes for both datasets. Particularly, CoHAtNet-4 (RGB-D) achieved the best overall accuracy of translation and orientation error, with other variants performing closely behind. Compared to the original CoAtNet model, CoHAtNet showed significantly improved localization performance. In addition, we visualized the ground-truth trajectories alongside the camera trajectories predicted by our model on the 7-Scenes dataset. As shown in Fig. 5, the red lines correspond to the ground-truth trajectories, whereas the blue dots indicate the estimated trajectories. It can be observed that the predicted trajectories align well with the ground truth.

To further validate the effectiveness of the proposed CoHAtNet model, we compared its performance with several state-of-the-art camera localization methods. The results of these comparisons for the seven classes of the 7-Scenes dataset, as well as their average, are summarized in Tables 4 and 5. The best translation and orientation result achieved by the alternative methods is highlighted in green. In turn, those cases in which the proposed CoHAtNet model achieved similar or better results than the best alternative method are highlighted in orange. These experimental results on 7-Scenes show that CoHAtNet consistently outperforms CNN-based methods, such as PoseNet [12], Bayesian PoseNet [13], and LSTM PoseNet [38] across all scenes. Moreover, CoHAtNet also demonstrates superior robustness compared to recent Transformer-based methods, such as TransPoseNet [39], TransCamP [40], HyperPose [16], and TransBoNet [41]. The closest competitors are EAAINET [10] and ALNET [11], which are not end-to-end neural models (the deep network estimates 3D positions and then a Perspective-n-Point solver is applied), and DFNet+NeFeS [42]. However, it is important to highlight that the authors of DFNet+NeFeS stated that they observed imperfections in the ground-truth (GT) poses of the 7-Scenes dataset, caused by asynchronous data between the RGB and the associated depth sequences. As a result, they retrained their model using a different ground-truth generated with COLMAP [30], a Structure-from-Motion technique with higher computational costs. To ensure a fair comparison, we have also included in Tables 4 and 5 the results obtained by the public implementation of DFNet+NeFeS trained on the original 7-Scenes GT. We have marked in red the cases where there is a significant difference between the results reported in [42] and those we obtained with the 7-Scenes GT.

4.2.2. Cambridge Landmarks dataset

The results obtained for the Cambridge Landmarks dataset are provided in Table 6. They demonstrate the robustness of CoHAtNet for large-scale datasets by only using RGB data. Notice the significant advantage of CoHAtNet-0 against the corresponding CoAtNet-0 regarding both translation and orientation errors. The best results are obtained by CoHAtNet-4. In turn, Table 7 shows the comparison between CoHAtNet-4 and various state-of-the-art methods. CoHAtNet-4 yields the best average performance across all fours scenes, with a translation error of 30 cm, and orientation error of 0.57°. These results are better than the closest competitor, DFNet+NeFeS [42]. This highlights the robustness and effectiveness of CoHAtNet for outdoor localization tasks, even when compared to more complex models. Other state-of-the-art methods, such as EFRNet-VL [47] and MS-Transformer [26], also fall behind CoHAtNet in terms of performance. For example, EFRNet-VL yields a translation error of 1.39 m, and an orientation error of 4.13° on average, much higher than CoHAtNet-4. The ability of CoHAtNet to maintain low error rates in a variety of scenes further demonstrates its reliability and adaptability to diverse environments.

4.2.3. Analysis of attention heatmaps

To gain deeper insights into the CoHAtNet architecture, we visualized and thoroughly analyzed the attention heatmaps generated by its final Transformer block. This analysis compares both RGB and RGB-D versions of CoHAtNet against three state-of-the-art deep networks: the original CoAtNet [3], MS-Transformer [26], and ALNet [11]. Both ALNet and MS-Transformer exclusively utilize Transformer-based attention mechanisms without convolutional integration, whereas CoHAtNet uniquely leverages hybrid convolutional-attention blocks.

Figs. 6 and 7 illustrate representative attention distributions for two distinct scenarios: the indoor "chess" scene from the 7-Scenes dataset, and the outdoor "college" view from Cambridge Landmarks, respectively.

In the indoor setting (Fig. 6), CoAtNet broadly disperses attention across multiple bright, reflective surfaces, such as walls, monitors, and desks, leading to a diluted focus on the chessboard, which is the primary object of interest. The incorporation of depth slightly reduces attention spread but still fails to adequately isolate key geometric features. MS-Transformer, which uses pure self-attention, focuses into multiple fragmented regions that obscure the coherent global structure

Table 4
Evaluation of different methods on the 7-Scenes dataset (classes: "Chess", "Fire", "Heads", "Office") [43–46].

Method	Date	Chess		Fire		Heads		Office	
PoseNet [12]	2015	0.32 m	8.12°	0.47 m	14.4°	0.29 m	12.0°	0.48 m	7.68°
Bayesian PoseNet [13]	2016	0.37 m	7.24°	0.43 m	13.7°	0.31 m	12.0°	0.48 m	8.04°
LSTM PoseNet [38]	2017	0.24 m	5.77°	0.34 m	11.9°	0.21 m	13.7°	0.30 m	8.08°
BranchNet [43]	2017	0.18 m	5.17°	0.34 m	8.99°	0.20 m	14.15°	0.30 m	7.05°
AtLoc [15]	2019	0.10 m	4.07°	0.25 m	11.4°	0.16 m	11.8°	0.17 m	5.34°
VMLoc [44]	2021	0.10 m	3.70°	0.25 m	10.5°	0.15 m	10.8°	0.16 m	5.08°
TransPoseNet [39]	2021	0.08 m	5.68°	0.24 m	10.6°	0.13 m	12.7°	0.17 m	6.34°
TransCamP [40]	2021	0.08 m	1.97°	0.27 m	8.26°	0.12 m	9.66°	0.12 m	2.37°
DFNet [45]	2022	0.05 m	1.88°	0.17 m	6.45°	0.06 m	6.36°	0.08 m	2.48°
EAAINET [10]	2023	0.02 m	0.58°	0.02 m	0.70°	0.01 m	0.93°	0.02 m	0.65°
HyperPose [16]	2023	0.08 m	6.29°	0.22 m	11.2°	0.11 m	12.7°	0.17 m	7.53°
TransAPR [46]	2023	0.08 m	3.40°	0.21 m	8.41°	0.14 m	9.51°	0.17 m	5.52°
TransBoNet [41]	2024	0.11 m	4.48°	0.25 m	12.46°	0.18 m	14.00°	0.20 m	5.08°
MS-Transformer [26]	2024	0.11 m	5.00°	0.24 m	9.45°	0.13 m	11.8°	0.18 m	5.92°
EFRNet-VL [47]	2024	0.11 m	3.96°	0.26 m	11.11°	0.17 m	12.87°	0.17 m	5.46°
DFNet+NeFeS [42]	2024	0.02 m	0.57°	0.02 m	0.74°	0.02 m	1.28°	0.02 m	0.56°
DFNet+NeFeS [42] 7S-GT	2024	0.01 m	2.94°	0.02 m	5.16°	0.01 m	6.27°	0.02 m	5.26°
ALNET [11]	2024	0.02 m	0.66°	0.02 m	0.81°	0.01 m	0.97°	0.02 m	0.68°
CoHAtNet-4 (RGB)	Proposed	0.02 m	0.55°	0.02 m	0.58°	0.02 m	1.37°	0.01 m	0.71°
CoHAtNet-4 (RGB-D)	Proposed	0.01 m	0.49°	0.01 m	0.51°	0.02 m	1.02°	0.01 m	0.56°

Table 5
Evaluation of different methods on the 7-Scenes dataset (classes: "Pumpkin", "Kitchen", "Stairs", and average of 7 classes).

Method	Date	Pumpkin		Kitchen		Stairs		Avg	
PoseNet [12]	2015	0.47 m	8.42°	0.59 m	8.64°	0.47 m	13.8°	0.45 m	9.94°
Bayesian PoseNet [13]	2016	0.61 m	7.08°	0.58 m	7.54°	0.48 m	13.1°	0.47 m	9.81°
LSTM PoseNet [38]	2017	0.33 m	7.0°	0.37 m	8.83°	0.40 m	13.7°	0.31 m	9.85°
BranchNet [43]	2017	0.27 m	5.10°	0.33 m	7.40°	0.38 m	10.26°	0.29 m	8.30°
AtLoc [15]	2019	0.21 m	4.37°	0.23 m	5.42°	0.26 m	10.5°	0.19 m	7.55°
VMLoc [44]	2021	0.20 m	4.01°	0.21 m	5.01°	0.24 m	10.0°	0.18 m	7.01°
TransPoseNet [39]	2021	0.17 m	5.6°	0.19 m	6.75°	0.30 m	7.02°	0.18 m	7.8°
TransCamP [40]	2021	0.16 m	2.49°	0.19 m	2.64°	0.28 m	9.01°	0.17 m	5.2°
DFNet [45]	2022	0.10 m	2.78°	0.22 m	5.45°	0.16 m	3.29°	0.12 m	3.71°
EAAINET [10]	2023	0.03 m	0.92°	0.03 m	1.00°	0.04 m	1.16°	0.03 m	0.84°
HyperPose [16]	2023	0.16 m	6.66°	0.17 m	8.48°	0.26 m	10.8°	0.16 m	9.09°
TransAPR [46]	2023	0.18 m	4.07°	0.19 m	4.65°	0.23 m	8.45°	0.17 m	6.29°
TransBoNet [41]	2024	0.19 m	4.77°	0.17 m	5.35°	0.30 m	13.04°	0.20 m	8.45°
MS-Transformer [26]	2024	0.19 m	4.62°	0.17 m	5.97°	0.26 m	7.92°	0.18 m	7.24°
EFRNet-VL [47]	2024	0.20 m	4.48°	0.22 m	5.95°	0.25 m	10.28°	0.20 m	7.73°
DFNet+NeFeS [42]	2024	0.02 m	0.55°	0.02 m	0.57°	0.05 m	1.28°	0.02 m	0.79°
DFNet+NeFeS [42] 7S-GT	2024	0.02 m	3.71°	0.02 m	5.37°	0.02 m	9.32°	0.02 m	5.40°
ALNET [11]	2024	0.03 m	1.01°	0.03 m	0.99°	0.04 m	0.96°	0.03 m	0.87°
CoHAtNet-4 (RGB)	Proposed	0.01 m	0.63°	0.02 m	0.61°	0.02 m	0.74°	0.02 m	0.74°
CoHAtNet-4 (RGB-D)	Proposed	0.01 m	0.49°	0.02 m	0.52°	0.01 m	0.46°	0.01 m	0.57°

Table 6
Performance of CoHAtNet models on the Cambridge Landmarks dataset (classes: "King's college", "Old hospital", "Shop facade", "Church", and average of 4 classes).

Method	King's college		Old hospital		Shop facade		Church		Avg	
CoAtNet-0 (RGB)	71 cm	2.84°	81 cm	2.25°	25 cm	2.88°	49 cm	3.86°	56 cm	2.95°
CoHAtNet-0 (RGB)	39 cm	1.12°	59 cm	1.17°	19 cm	1.48°	41 cm	1.53°	39 cm	1.32°
CoHAtNet-1 (RGB)	38 cm	0.93°	59 cm	1.02°	17 cm	1.17°	39 cm	0.99°	38 cm	1.02°
CoHAtNet-2 (RGB)	38 cm	0.67°	57 cm	0.73°	17 cm	0.91°	37 cm	0.72°	37 cm	0.75°
CoHAtNet-3 (RGB)	37 cm	0.63°	55 cm	0.70°	15 cm	0.88°	34 cm	0.62°	35 cm	0.70°
CoHAtNet-4 (RGB)	31 cm	0.48°	45 cm	0.67°	16 cm	0.43°	31 cm	0.70°	30 cm	0.57°

of the chessboard, making it difficult to capture spatially coherent patterns. Similarly, ALNet disproportionately converges attention onto irrelevant objects like the monitor bezel, neglecting critical spatial context. Conversely, CoHAtNet effectively consolidates attention into a singular, cohesive region precisely bounded by the chessboard and adjacent informative features. The integration of depth further sharpens the attention map, closely adhering to the accurate contours of chess pieces and board edges. This precision reflects the efficacy of the depth-gated Value branch in suppressing planar and peripheral clutter, thereby amplifying attention towards genuine 3-D geometric boundaries.

In the outdoor scenario (Fig. 7), baseline models exhibit distinct but suboptimal attention behaviors. CoAtNet oscillates its focus erratically between non-informative sky regions and foreground foliage, substantially neglecting architecturally significant facade structures. MS-Transformer, while identifying key features such as rooflines, produces fragmented, disconnected streaks of attention, impairing spatial continuity. ALNet, on the other hand, narrowly targets isolated prominent structures like the tower spire, disregarding horizontally extensive structures such as the cornice line, which is crucial for vertical pose determination. In sharp contrast, CoHAtNet evenly distributes its attention across multiple critical landmarks including the tower, roof edges, and the shadow-boundary on the ground plane. Importantly, it

Table 7
Evaluation of different methods on the Cambridge Landmarks dataset ("college", "hospital", "shop", "church", and average of all classes) [48,49].

Method	Date	college		hospital		shop		church		Avg	
PoseNet [12]	2015	1.92 m	5.40°	2.31 m	5.38°	1.46 m	8.08°	2.65 m	8.48°	2.08 m	6.83°
Bayesian PoseNet [13]	2016	1.74 m	4.06°	2.57 m	5.14°	1.25 m	7.54°	2.11 m	8.38°	1.91 m	6.28°
LSTM PoseNet [38]	2017	0.99 m	3.65°	1.51 m	4.29°	1.18 m	7.44°	1.52 m	6.68°	1.30 m	5.57°
DeepDSAIR [48]	2019	0.86 m	1.11°	1.64 m	2.47°	1.01 m	4.14°	1.80 m	6.25°	1.32 m	3.49°
TransPoseNet [39]	2021	0.60 m	2.43°	1.45 m	3.08°	0.55 m	3.49°	1.09 m	4.99°	0.92 m	3.49°
TransCamP [40]	2021	0.42 m	0.36°	1.01 m	1.83°	0.24 m	1.56°	1.55 m	2.56°	0.81 m	1.58°
CoordiNet [49]	2021	0.70 m	0.92°	0.97 m	2.08°	0.69 m	3.74°	1.32 m	3.56°	0.92 m	2.58°
DFNet [45]	2022	0.73 m	2.37°	2.0 m	2.98°	0.67 m	2.21°	1.37 m	4.03°	1.19 m	2.09°
HyperPose [16]	2023	0.56 m	2.40°	1.41 m	2.91°	0.54 m	3.37°	0.98 m	4.86°	0.87 m	3.43°
TransAPR [46]	2023	0.59 m	0.86°	1.42 m	2.29°	0.54 m	2.18°	1.21 m	3.16°	0.15 m	6.71°
MS-Transformer [26]	2024	0.80 m	2.86°	1.84 m	3.82°	0.80 m	3.66°	1.17 m	4.10°	1.15 m	3.41°
EFRNet-VL [47]	2024	1.25 m	3.05°	2.49 m	3.55°	1.19 m	4.54°	1.02 m	5.38°	1.39 m	4.13°
DFNet+NeFeS [42]	2024	0.37 m	0.54°	0.52 m	0.88°	0.15 m	0.53°	0.37 m	1.14°	0.35 m	0.77°
CoHAtNet-4 (RGB)	Proposed	0.31 m	0.48°	0.45 m	0.67°	0.16 m	0.43°	0.31 m	0.70°	0.30 m	0.57°

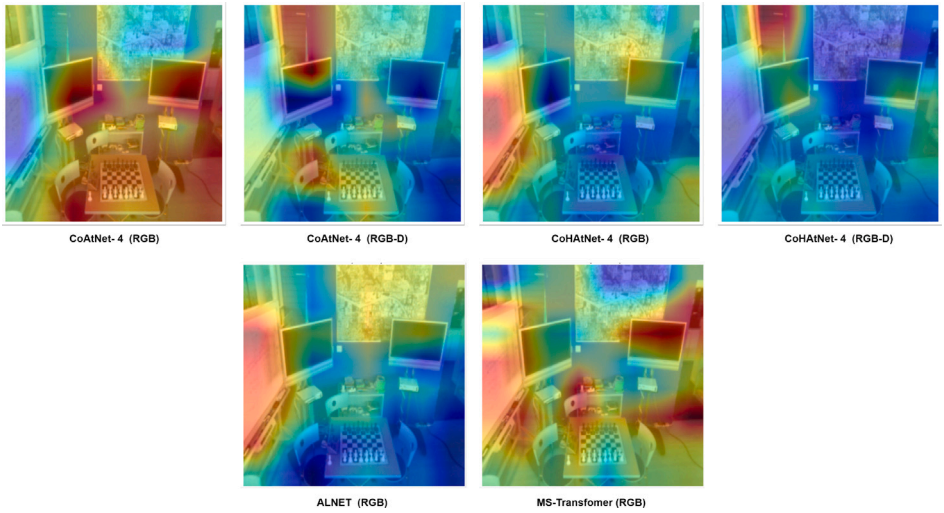

Fig. 6. Attention heatmaps for the "chess" scene (7-Scenes). CoHAtNet (RGB/RGB-D) isolates the chessboard and informative pieces, whereas CoAtNet [3], MS-Transformer [26], and ALNet [11] leave substantial activation on background surfaces.

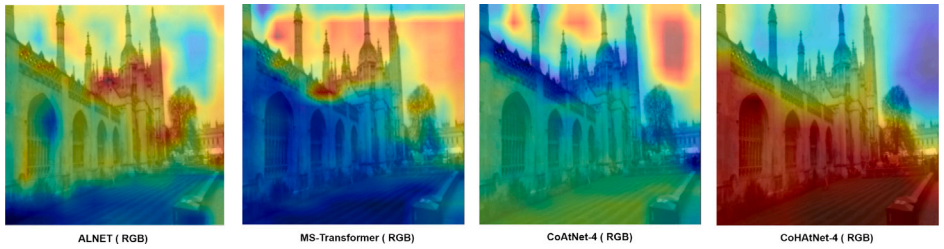

Fig. 7. Attention heatmaps for the "college" scene (Cambridge Landmarks). CoHAtNet (RGB) distributes attention coherently over the tower, roof cornice, and ground plane, while the baselines either fragment the map or focus on a single landmark.

avoids "sky-collapse", a common deficiency among purely attention-based models, demonstrating robust handling of contextual information and yielding lower pose estimation errors for expansive scenes.

Overall, these detailed observations elucidate the reasons behind CoHAtNet's superior localization accuracy compared to state-of-the-art alternatives. While CoHAtNet demonstrates robust performance, particularly in RGB-D settings, attention modulation efficacy can degrade when depth maps are compromised by noise or low resolution, notably in environments featuring reflective or transparent surfaces. Additionally, deeper model variants, such as CoHAtNet-4, consistently outperform shallower models, albeit at increased computational costs. These deeper configurations may thus face deployment challenges in scenarios demanding real-time inference or those with strict computational constraints.

4.2.4. *Parameter sensitivity*

To verify that *CoHAtNet* does not depend on brittle hyper-parameter tuning, we carried out a targeted sensitivity study on the *Chess* scene of 7-Scenes with the mid-capacity CoHAtNet-3 backbone. We varied one hyper-parameter at a time while keeping all others at their default values, and report the mean localization error averaged over three random seeds.

Table 8 examines three influential knobs — the MBConv expansion ratio r, the depth-wise kernel size k, and the base learning rate η — using the RGB-D variant of CoHAtNet-3. Across every perturbation, translation error fluctuates by at most 0.005 m and rotation error by at most 0.05° relative to the baseline, evidencing a broad performance plateau. Notably, halving the expansion ratio ($r = 2$), enlarging the kernel up to 9×9, or varying the learning rate between 0.5× and 5× its

Table 8
Sensitivity of CoHAtNet-3 (RGB-D) to selected hyper-parameters on the *Chess* scene. Each cell lists mean translation and rotation error (m/°); default settings are bold.

Symbol	Hyper-parameter	Default	Test 1	Test 2	Test 3
r	MBConv expansion ratio	**4**	2: 0.022 m 0.55°	6: 0.021 m 0.54°	8: 0.024 m 0.56°
k	Depth-wise kernel size	**3**	5: 0.021 m 0.54°	7: 0.023 m 0.55°	9: 0.025 m 0.57°
η	Learning rate (Adam)	1×10^{-4}	5×10^{-5}: 0.021 m 0.53°	2×10^{-4}: 0.022 m 0.60°	5×10^{-4}: 0.025 m 0.57°

default value hardly affects accuracy. These results demonstrate that the proposed hybrid architecture delivers practical robustness without the need for delicate hyper-parameter tuning.

5. Conclusions

We have introduced CoHAtNet, a novel hybrid neural architecture that integrates convolutional and self-attention mechanisms within a unified block structure. This design allows the network to concurrently model fine-grained spatial features and long-range semantic dependencies, effectively bridging the limitations of traditional CNNs and pure Transformer-based models in visual localization tasks.

We applied CoHAtNet as a fully end-to-end neural solution to the camera localization problem, and evaluated its performance on two benchmark datasets: the indoor 7-Scenes (RGB-D) and the large-scale outdoor Cambridge Landmarks (RGB). The model was tested using both RGB and RGB-D modalities. In both cases, CoHAtNet significantly outperformed existing CNN-based and Transformer-based approaches in terms of translation and rotation accuracy. Notably, the RGB-D version consistently achieved the best results by jointly leveraging appearance and depth cues through early fusion. The hybrid self-attention mechanism, which combines MBConv-derived local spatial representations with global Transformer-based semantics, proved critical for improving robustness across cluttered, textureless, or repetitive environments.

Beyond academic benchmarks, CoHAtNet shows strong potential for real-time deployment in a wide range of practical scenarios. Its ability to produce precise and reliable pose estimations makes it suitable for applications such as AR, autonomous robotics, mobile visual SLAM, and wearable navigation systems, especially in GPS-denied environments or complex indoor scenes. These applications can benefit from both RGB-only and RGB-D configurations of CoHAtNet, depending on the hardware and environmental constraints.

Although CoHAtNet demonstrates excellent performance, further improvements are possible. Future work will focus on optimizing deeper CoHAtNet variants for real-time applications through efficient attention mechanisms, model compression techniques, and lightweight architectural modifications suitable for edge devices. In addition, we plan to extend the model's capabilities to handle dynamic or multi-scene environments, and to generalize it towards related tasks, such as visual odometry, multi-view pose estimation, and full SLAM integration. Enhancing training efficiency through self-supervised learning, incorporating additional modalities such as inertial or LiDAR data, and improving resilience to sensor noise or occlusion are also key directions that could broaden CoHAtNet's applicability in real-world deployments.

CRediT authorship contribution statement

Hussein Hasan: Writing – original draft, Software. **Miguel Angel Garcia:** Writing – review & editing, Validation, Supervision, Software. **Hatem Rashwan:** Supervision. **Domenec Puig:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The Spanish Government partly supported this research through Project TED2021-130081B-C21, and Project PDC2022-133383-I00.

Data availability

An implementation of the proposed method is publicly available on GitHub: https://github.com/Husseinhhameed/CoHAtNet.

References

[1] M. Xu, Y. Wang, B. Xu, J. Zhang, J. Ren, Z. Huang, S. Poslad, P. Xu, A critical analysis of image-based camera pose estimation techniques, Neurocomputing 570 (2024) 127125, http://dx.doi.org/10.1016/j.neucom.2023.127125, URL https://www.sciencedirect.com/science/article/pii/S0925231223012481.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021, arXiv:2010.11929. URL https://arxiv.org/abs/2010.11929.

[3] Z. Dai, H. Liu, Q.V. Le, M. Tan, Coatnet: Marrying convolution and attention for all data sizes, Adv. Neural Inf. Process. Syst. 34 (2021) 3965–3977.

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2018, pp. 4510–4520, http://dx.doi.org/10.1109/CVPR.2018.00474, URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00474.

[5] F. Radenović, G. Tolias, O. Chum, Fine-tuning CNN image retrieval with no human annotation, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2019) 1655–1668, http://dx.doi.org/10.1109/TPAMI.2018.2846566.

[6] P.-E. Sarlin, C. Cadena, R.Y. Siegwart, M. Dymczyk, From coarse to fine: Robust hierarchical localization at large scale, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018, pp. 12708–12717, URL https://api.semanticscholar.org/CorpusID:54460261.

[7] P. Wang, B. Jiao, P. Yao, X. Wei, A. Zhang, A robust direct linear transformation for camera pose estimation using points, Image Vis. Comput. 141 (2024) 104883, http://dx.doi.org/10.1016/j.imavis.2023.104883, URL https://www.sciencedirect.com/science/article/pii/S0262885623002573.

[8] L. Liu, H. Li, Y. Dai, Efficient global 2D-3D matching for camera localization in a large-scale 3D map, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 2391–2400, URL https://api.semanticscholar.org/CorpusID:5658428.

[9] Q. Wang, J. Zhang, K. Yang, K. Peng, R. Stiefelhagen, MatchFormer: Interleaving attention in transformers for feature matching, in: Computer Vision – ACCV 2022: 16th Asian Conference on Computer Vision, Macao, China, December 4–8, 2022, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2023, pp. 256–273, http://dx.doi.org/10.1007/978-3-031-26313-2_16.

[10] K. Dai, T. Xie, K. Wang, Z. Jiang, D. Liu, R. Li, J. Wang, EAAINet: An element-wise attention network with global affinity information for accurate indoor visual localization, IEEE Robot. Autom. Lett. 8 (6) (2023) 3166–3173, http://dx.doi.org/10.1109/LRA.2023.3261703.

[11] H. Gao, K. Dai, K. Wang, R. Li, L. Zhao, M. Wu, ALNet: An adaptive channel attention network with local discrepancy perception for accurate indoor visual localization, Expert Syst. Appl. 250 (2024) 123792, http://dx.doi.org/10.1016/j.eswa.2024.123792, URL https://www.sciencedirect.com/science/article/pii/S0957417424006584.

[12] A. Kendall, M.K. Grimes, R. Cipolla, PoseNet: A convolutional network for real-time 6-DOF camera relocalization, in: 2015 IEEE International Conference on Computer Vision, ICCV, 2015, pp. 2938–2946, URL https://api.semanticscholar.org/CorpusID:12888763.

[13] A. Kendall, R. Cipolla, Modelling uncertainty in deep learning for camera relocalization, in: 2016 IEEE International Conference on Robotics and Automation, ICRA, 2015, pp. 4762–4769, URL https://api.semanticscholar.org/CorpusID:3240382.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023, arXiv:1706.03762. URL https://arxiv.org/abs/1706.03762.

[15] B. Wang, C. Chen, X.C. Lu, P. Zhao, N. Trigoni, A. Markham, AtLoc: Attention guided camera localization, Proc. the AAAI Conf. Artif. Intell. 34 (06) (2020) 10393–10401, http://dx.doi.org/10.1609/aaai.v34i06.6608, URL https://ojs.aaai.org/index.php/AAAI/article/view/6608.

[16] R. Ferens, Y. Keller, HyperPose: Camera pose localization using attention hypernetworks, 2023, arXiv arXiv:2303.02610. URL https://api.semanticscholar.org/CorpusID:257365499.

[17] S. Chen, Y. Bhalgat, X. Li, J.-W. Bian, K. Li, Z. Wang, V.A. Prisacariu, Neural refinement for absolute pose regression with feature synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 20987–20996.

[18] J.L. Charco, A.D. Sappa, B.X. Vintimilla, H.O. Velesaca, Camera pose estimation in multi-view environments: From virtual scenarios to the real world, Image Vis. Comput. 110 (2021) 104182, http://dx.doi.org/10.1016/j.imavis.2021.104182, URL https://www.sciencedirect.com/science/article/pii/S0262885621000871.

[19] I. Melekhov, J. Ylioinas, J. Kannala, E. Rahtu, Relative camera pose estimation using convolutional neural networks, in: J. Blanc-Talon, R. Penne, W. Philips, D. Popescu, P. Scheunders (Eds.), Advanced Concepts for Intelligent Vision Systems, Springer International Publishing, Cham, 2017, pp. 675–687.

[20] F. Xue, X. Wu, S. Cai, J. Wang, Learning multi-view camera relocalization with graph neural networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 11372–11381, URL https://api.semanticscholar.org/CorpusID:221707829.

[21] S. Saha, G. Varma, C.V. Jawahar, Improved visual relocalization by discovering anchor points, in: British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018, BMVA Press, 2018, p. 164, URL http://bmvc2018.org/contents/papers/0962.pdf.

[22] B. Zhuang, M. Chandraker, Fusing the old with the new: Learning relative camera pose with geometry-guided uncertainty, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 32–42, URL https://api.semanticscholar.org/CorpusID:233289634.

[23] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, C. Rother, DSAC — Differentiable RANSAC for camera localization, in: 2017 IEEE IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 2492–2500, URL https://api.semanticscholar.org/CorpusID:4001530.

[24] E. Brachmann, C. Rother, Learning less is more - 6D camera localization via 3D surface regression, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 4654–4662, URL https://api.semanticscholar.org/CorpusID:4302093.

[25] H. Blanton, C. Greenwell, S. Workman, N. Jacobs, Extending absolute pose regression to multiple scenes, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2020, pp. 170–178, http://dx.doi.org/10.1109/CVPRW50498.2020.00027.

[26] Y. Shavit, R. Ferens, Y. Keller, Learning single and multi-scene camera pose regression with transformer encoders, Comput. Vis. Image Underst. 243 (2024) 103982, http://dx.doi.org/10.1016/j.cviu.2024.103982, URL https://www.sciencedirect.com/science/article/pii/S1077314224000638.

[27] Y. Shavit, R. Ferens, Y. Keller, Coarse-to-fine multi-scene pose regression with transformers, IEEE Trans. Pattern Anal. Mach. Intell. 45 (12) (2023) 14222–14233, http://dx.doi.org/10.1109/TPAMI.2023.3310929.

[28] T. Sattler, Q. Zhou, M. Pollefeys, L. Leal-Taixé, Understanding the limitations of CNN-based absolute camera pose regression, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 3297–3307, http://dx.doi.org/10.1109/CVPR.2019.00342.

[29] S. Mokssit, D.B. Licea, B. Guermah, M. Ghogho, Deep learning techniques for visual SLAM: A survey, IEEE Access 11 (2023) 20026–20050, http://dx.doi.org/10.1109/ACCESS.2023.3249661.

[30] E. Brachmann, M. Humenberger, C. Rother, T. Sattler, On the limits of pseudo ground truth in visual camera re-localisation, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 6198–6208, http://dx.doi.org/10.1109/ICCV48922.2021.00616.

[31] X. Zhao, L. Wang, Y. Zhang, et al., A review of convolutional neural networks in computer vision, Artif. Intell. Rev. 57 (2024) 99, http://dx.doi.org/10.1007/s10462-024-10721-6.

[32] S. Jamil, M. Jalil Piran, O.-J. Kwon, A comprehensive survey of transformers for computer vision, Drones 7 (5) (2023) http://dx.doi.org/10.3390/drones7050287, URL https://www.mdpi.com/2504-446X/7/5/287.

[33] J. Guan, Y. Hao, Q. Wu, S. Li, Y. Fang, A survey of 6dof object pose estimation methods for different application scenarios, Sensors 24 (4) (2024) 1076, http://dx.doi.org/10.3390/s24041076, URL https://www.mdpi.com/1424-8220/24/4/1076.

[34] A. Kendall, R. Cipolla, Geometric loss functions for camera pose regression with deep learning, in: 2017 IEEE IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 6555–6564, URL https://api.semanticscholar.org/CorpusID:13141802.

[35] C. Boittiaux, R. Marxer, C. Dune, A. Arnaubec, V. Hugel, Homography-based loss function for camera pose regression, IEEE Robot. Autom. Lett. 7 (2022) 6242–6249, URL https://api.semanticscholar.org/CorpusID:248291230.

[36] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, CoRR arXiv:1412.6980. URL https://api.semanticscholar.org/CorpusID:6628106.

[37] B. Glocker, S. Izadi, J. Shotton, A. Criminisi, Real-time RGB-D camera relocalization, in: International Symposium on Mixed and Augmented Reality, International Symposium on Mixed and Augmented Reality (ISMAR), ISMAR, IEEE, 2013, pp. 173–179, URL https://www.microsoft.com/en-us/research/publication/real-time-rgb-d-camera-relocalization/.

[38] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, D. Cremers, Image-based localization using LSTMs for structured feature correlation, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2016, pp. 627–637, URL https://api.semanticscholar.org/CorpusID:15053207.

[39] Y. Shavit, R. Ferens, Y. Keller, Paying attention to activation maps in camera pose regression, 2021, arXiv:2103.11477. URL https://arxiv.org/abs/2103.11477.

[40] X. Li, H. Ling, TransCamP: Graph transformer for 6-DoF camera pose estimation, 2021, arXiv arXiv:2105.14065. URL https://api.semanticscholar.org/CorpusID:235253755.

[41] X. Song, H. Li, L. Liang, W. Shi, G. Xie, X. Lu, X. Hei, TransBoNet: Learning camera localization with transformer bottleneck and attention, Pattern Recognit. 146 (2024) 109975, http://dx.doi.org/10.1016/j.patcog.2023.109975, URL https://www.sciencedirect.com/science/article/pii/S0031320323006738.

[42] S. Chen, Y. Bhalgat, X. Li, J.-W. Bian, K. Li, Z. Wang, V.A. Prisacariu, Neural refinement for absolute pose regression with feature synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 20987–20996.

[43] J. Wu, L. Ma, X. Hu, Delving deeper into convolutional neural networks for camera relocalization, in: 2017 IEEE International Conference on Robotics and Automation, ICRA, 2017, pp. 5644–5651, http://dx.doi.org/10.1109/ICRA.2017.7989663.

[44] K. Zhou, C. Chen, B. Wang, M.R.U. Saputra, N. Trigoni, A. Markham, VMLoc: Variational fusion for learning-based multimodal camera localization, Proc. the AAAI Conf. Artif. Intell. 35 (7) (2021) 6165–6173, http://dx.doi.org/10.1609/aaai.v35i7.16767, URL https://ojs.aaai.org/index.php/AAAI/article/view/16767.

[45] S. Chen, X. Li, Z. Wang, V.A. Prisacariu, DFNet: Enhance absolute pose regression with direct feature matching, in: S. Avidan, G. Brostow, M. Cissé, G.M. Farinella, T. Hassner (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham, 2022, pp. 1–17.

[46] C. Qiao, Z. Xiang, Y. Fan, T. Bai, X. Zhao, J. Fu, TransAPR: Absolute camera pose regression with spatial and temporal attention, IEEE Robot. Autom. Lett. 8 (8) (2023) 4633–4640, http://dx.doi.org/10.1109/LRA.2023.3286123.

[47] J. Wang, H. Yu, X. Lin, Z. Li, W. Sun, N. Akhtar, EFRNet-VL: An end-to-end feature refinement network for monocular visual localization in dynamic environments, Expert Syst. Appl. 243 (2024) 122755, http://dx.doi.org/10.1016/j.eswa.2023.122755, URL https://www.sciencedirect.com/science/article/pii/S0957417423032578.

[48] M. Abolfazli Esfahani, K. Wu, S. Yuan, H. Wang, DeepDSAIR: Deep 6-DOF camera relocalization using deblurred semantic-aware image representation for large-scale outdoor environments, Image Vis. Comput. 89 (2019) 120–130, http://dx.doi.org/10.1016/j.imavis.2019.06.014, URL https://www.sciencedirect.com/science/article/pii/S0262885619300976.

[49] A. Moreau, N. Piasco, D.V. Tsishkou, B. Stanciulescu, A.d. Fortelle, CoordiNet: uncertainty-aware pose regressor for reliable vehicle localization, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2021, pp. 1848–1857, URL https://api.semanticscholar.org/CorpusID:232290863.