2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

EffLoc: Lightweight Vision Transformer for Efficient 6-DOF Camera Relocalization

Zhendong Xiao[1], Changhao Chen[2], Shan Yang[1], *Wu Wei[1]

***Abstract*— Camera relocalization is pivotal in computer vision, with applications in AR, drones, robotics, and autonomous driving. It estimates 3D camera position and orientation (6-DoF) from images. Unlike traditional methods like SLAM, recent strides use deep learning for direct end-to-end pose estimation. We propose EffLoc, a novel efficient Vision Transformer for single-image camera relocalization. EffLoc's hierarchical layout, memory-bound self-attention, and feed-forward layers boost memory efficiency and inter-channel communication. Our introduced sequential group attention (SGA) module enhances computational efficiency by diversifying input features, reducing redundancy, and expanding model capacity. EffLoc excels in efficiency and accuracy, outperforming prior methods, such as AtLoc and MapNet. It thrives on large-scale outdoor car-driving scenario, ensuring simplicity, end-to-end trainability, and eliminating handcrafted loss functions.**

I. INTRODUCTION

Camera relocalization (i.e. camera pose regression) focuses on the retrieval of the 3D position and orientation (6-DoF) of a camera based on the input images. It plays a crucial role in intelligent systems, ranging from augmented reality (AR) [1]/mixed reality (MR), delivery drones, robotics to autonomous driving [2]. Camera localization approaches have historically leaned on image structure and feature to match visual observation against a map [3], which establishes dense correspondences between 2D pixels and 3D points within the scene. Subsequently, the camera pose is estimated by employing Perspective-n-Point (PnP) solver [4] or the Kabsch algorithm with RANSAC [5]. These conventional relocalization methods fundamentally rely on a matching procedure, encompassing the comparison of a query image against a database of reference images [6]. The computational and storage requirements of these techniques correlate directly with the volume of sample points within the database. Furthermore, the efficacy of these methods is inherently intertwined with the quality of the matching process predicated upon the similarity score.

Deep learning-based camera relocalization methods can achieve end-to-end pose estimation directly from images via deep neural networks. For example, PoseNet [7] uses a convolutional neural network (CNN) based encoder to extract features from a single image as vector embeddings,

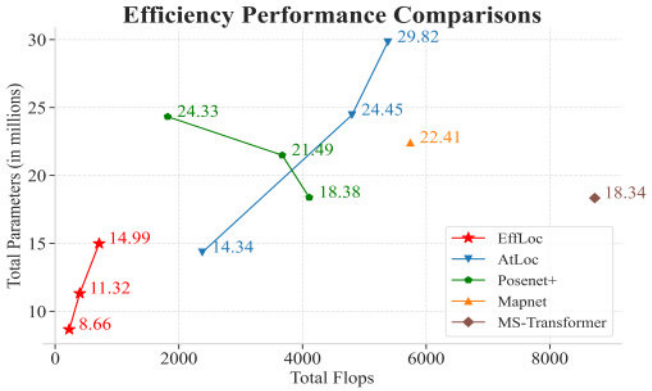


Fig. 1: A comparison between our proposed EffLoc model and other deep learning based relocalization models. The x-axis denotes Total Floating-Point Operations Per Second (FLOPs), while the y-axis represents the total number of parameters. The parameter count is labeled for each corresponding point on the graph. Our EffLoc models exhibit superior efficiency and computational complexity, attaining the lowest Flops and parameter counts.

which are subsequently transformed to 6-DoF pose. Other end-to-end leaning approaches such as [8] utilize an implicit map database to store scene information and eliminate the complex handcrafted feature engineering. Traditionally, deep learning based pose estimation heavily relies on Convolutional Neural Networks (CNNs) for feature extraction, which operate within localized pixel neighborhoods. However, Vision Transformers (ViTs) represent a recent breakthrough by segmenting images into patches and utilizing position embeddings to capture global dependencies. Unlike CNNs, ViTs establish meaningful correlations among spatially distant image regions, crucial for real-world relocalization tasks with large datasets. In addition, CNN-based visual localization models suffer from accuracy limitations and lack robust generalization due to challenges like lighting variations, occlusions, and dynamic objects. In contrast, ViTs, when trained with image-poses pairs, align better with map libraries, eliminating scale drift and cumulative errors. The emerging Light-weight Vision Transformers [10] offer computational efficiency and improved robustness in complex, resource-constrained real-world scenarios.

In addressing the challenges posed by existing CNN-based visual relocalization methods, we introduce EffLoc, a lightweight Vision Transformer framework for efficient 6-DoF camera relocalization. Figure 2 presents a modular overview of the EffLoc. Our hierarchical architecture integrates memory-bound self-attention and inter-channel communication for improved memory efficiency [12] [13]. We

[1]Zhendong Xiao, Shan Yang and Wu Wei are with Control Science and Engineering,School of Automation Science and Engineering, South China University of Technology, Guangzhou, Guangdong Province, China auxiao2022@mail.scut.edu.cn

[2]Changhao Chen is with the Colloge of Intelligence Science and Technology, National University of Defense Technology, Changsha, China changhao.chen66@outlook.com

Corresponding author: Wu Wei

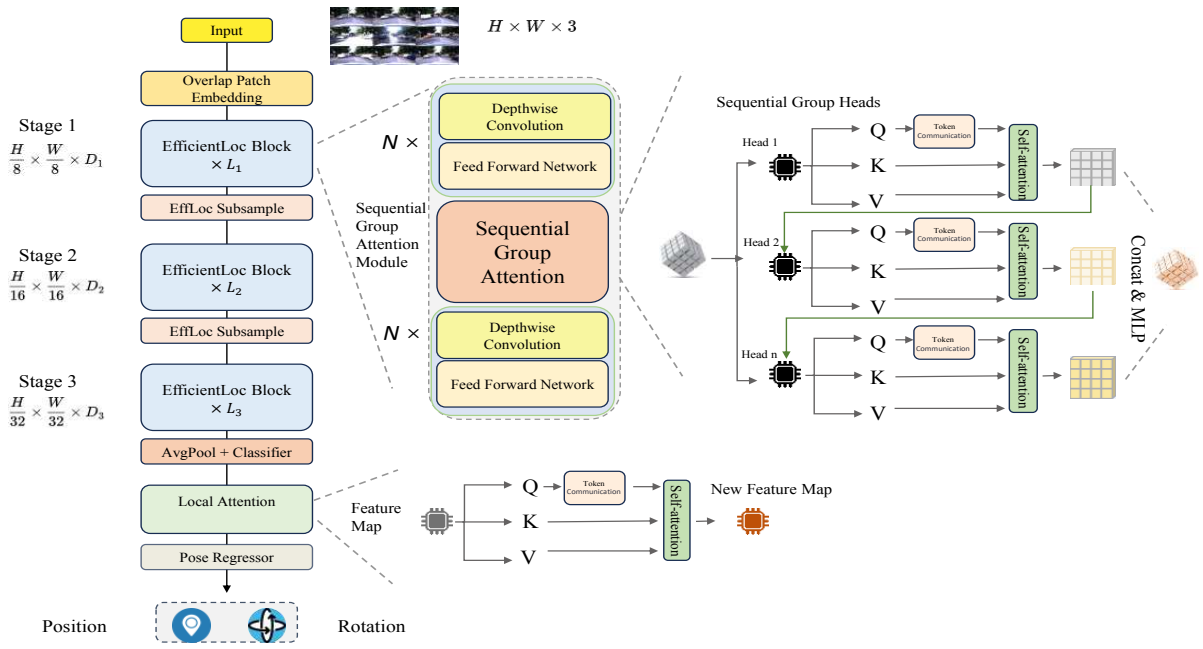


Fig. 2: An overview of EffLoc's hierarchical framework and its modules. The left column showcases the overall layout. The middle column highlights Sequential Group Attention Module and Sequential Group Heads (SGH). The right column details how SGH integrates outputs across heads. The bottom presents attention feature map and the pose regressor overview for feature-to-pose transformation.

also propose the sequential group attention (SGA) module, enhancing computational efficiency by diversifying input features across attention heads and redistributing parameters. Our model surpasses previous techniques, excelling in accuracy and efficiency trade-off for image-based 6-DoF camera relocalization, even on large-scale outdoor datasets. Its simplicity, end-to-end trainability, and elimination of handcrafted loss functions mark its strengths. Figure 1 demonstrates the superior performance of our model in contrast to prior techniques, excelling in the trade-off between accuracy and efficiency in image-based 6-DoF camera relocalization. Notably adept at processing extensive outdoor datasets, our model shines with its streamlined architecture, end-to-end trainability, and elimination of the necessity for manually engineered geometric loss functions.

In summary, the contributions of this work are three-fold:

- We propose EffLoc, a novel light-weight end-to-end Vision Transformer architecture 6-DoF camera relocalization using single images that can generalize to large-scale real-word environments.
- We present a simple and effective sequential group module that remarkably enhances the latency/accuracy trade-off for image-based 6-DoF camera relocalization. By introducing diverse channel-wise feature splits across attention heads, this module effectively reduces redundant attention computations, leading to notable memory efficiency gains.
- We introduce a novel parameterization approach that involves reconfiguring the original Vit QKV (Query, Key, Value) ratios for camera pose estimation. This optimization substantially enhances computation and memory efficiency, resulting in a significant reduction in both FLOPs (86.8%) and memory usage (49.7%) compared to AtLoc [16]. Notably, the impact of Q and K in the third block is substantially diminished through this reconfiguration process.

II. RELATED WORK

A. Deep learning for Camera Relocalization

Deep learning has revolutionized camera relocalization, employing Convolutional Neural Networks (CNNs) to manage variations in illumination, viewpoint, and object occlusion [17], [18]. Kendall [19] first demonstrated end-to-end pose estimation using CNNs, eliminating intermediate steps like feature matching [20]. Furthermore, Deep Neural Network (DNN)-based camera localization methods obviate the need for manual construction of a map or a database of landmark features [21]. This approach evolved to include Recurrent Neural Networks (RNNs) and Bayesian CNNs for spatial-temporal accuracy and uncertainty estimation [22], [23]. CMRNet [46] and CMRNet++ [47], integrate deep learning with geometric approaches to taskle image-based relocalization within LiDAR maps. However, challenges with hand-tuned scale factors and obstacles led to innovations like PosenetV2's geometric reprojection loss [20] and BranchNet's locator with two branches for position and angular deviations[24]. Attention mechanisms in Atloc [16] and additional reconstruction branches in MMLNet [25] further enhanced the relationship between 2D images and 3D scenes. Camera Pose Auto Encoders (PAEs) introduced lightweight test-time optimization [26]. Our EffLoc model, with its hierarchical layout of self-attention and feedforward

layers, outperforms previous approaches with less computation costs, faster convergence velocity and achieves the best results in common benchmarks.

B. Vision Transformers

Transformers, introduced by Vaswani [11], have excelled in NLP and Computer Vision, surpassing RNNs in NLP [27]. Vision Transformers (ViTs), introduced by Touvron [28], segment images into patches with position encoding, achieving high performance on datasets like JFT-300M [29]. However, original ViTs faced optimization, efficiency, and training challenges [27] [28]. Lightweight transformers prioritize computational efficiency for real-time applications and constrained devices [14]. MobileNet [32] uses depthwise separable convolutions, TinyBERT [33] distills compact BERT [34] for constrained deployment, MobileViT V1 [35] explores channel, spatial factorizations, hierarchical token-to-token attention [36], adaptive token mixing [37]. ParC-Net [38] introduces patch-aware adaptive receptive fields, and EfficientVIT [39] innovates cascaded group attention. Next-ViT [10] addresses sparse attention, redundancy, complexity, memory. Our work integrates lightweight vision transformers, efficient parameterization into camera relocalization, demonstrating hierarchical group attention's robust key feature correlation, offering deployment-friendly solutions for accurate, high-speed camera relocalization.

III. LIGHTWEIGHT-TRANSFORMER BASED EFFICIENT CAMERA RELOCALIZATION

This section presents an **Eff**icient Lightweight-Transformer based Camera Re**Loc**alization (EffLoc) approach, to learn 6-DoF camera poses from single images.

A. Overlap Patch Embedding for Feature Extraction

Vision Transformers (ViTs) are designed to capture holistic contextual information across images, distinct from Convolutional Neural Networks (ConvNets) that emphasize local features. Effective pose regression relies on extracting features from single images. In standard ViTs, input images are split into non-overlapping 16×16 patches, linearly projected into the encoder's input dimension using a learned weight matrix [27]. Our approach involves dividing images into overlapping patches, embedding each using a conventional ConvNet with layered convolutions [40]. This Overlap Patch Embedding enhances fine-grained localization by capturing local details and spatial sensitivity from neighboring patches. Past studies [17] [38] demonstrate classical convolutional networks' effectiveness. [16] underscores ResNet34's efficacy, a 34-layer residual network, as a base for camera pose estimation. Residual networks like ResNet34 train deeper layers, addressing gradient vanishing and memory inefficiency issues associated with cross-memory access, which is computationally expensive. Hence, we opt for a lightweight Vision Transformer (EfficientVit) as EffLoc's backbone. EfficientVit's weights are initialized from a pretrained ImageNet-1K model [42], optimized for image classification. Given an image $I \in R^{C\times H\times W}$, the features $X \in R^{C}$ can be extracted via the overlap patch embedding (Ope):

$$X = \mathrm{Ope}(I). \tag{1}$$

B. Hierarchical Layout to Enhance Memory Efficiency

Here, we introduce hierarchical layout with competitive convergence capability that enhances memory efficiency and mitigates attention computation redundancy. Specifically, it applies a single less memory-bound self-attention layer $\mathscr{L}_i^{SAL}$ with linear complexity and a depthwise convolution layer for spatial integration, which is nested between the feed forward layer $\mathscr{L}_i^{FFL}$. The computation can be expressed as follows:

$$X_{i+1} = \prod^{N} \mathscr{L}_i^{FFL}\left(\mathscr{L}_i^{SAL}\left(\prod^{N} \mathscr{L}_i^{FFL}(X_i)\right)\right), \tag{2}$$

where X_i represents the complete input feature for the $i-th$ position. The hierarchical layout converts X_i into X_{i+1} with N feed forward layers, both prior to and subsequent to the self-attention layer. This layout reduces the memory footprint while preserving crucial information and optimizes the utilization of model parameters resulting from self-attention layers. By facilitating more efficient multiple feed forward network layers communication between different feature channels, the model improve the velocity of convergence, and thus reduce the computational burden.

C. Sequential Group Attention to Aggregate Features

The redundancy of attention heads in multi-head self-attention is a significant issue that leads to computational inefficiency [15]. We introduce Sequential Group Attention (SGA) as attention module into our proposed EffLoc. Each head receives different subsets of the complete features, thereby effectively decomposing the attention computation across multiple heads in the attention module. Mathematically, this attention mechanism can be expressed as:

$$\widetilde{X}_{ij} = \mathrm{Attn}\left(X_{ij}\omega_{ij}^{\mathrm{Q}}, X_{ij}\omega_{ij}^{\mathrm{K}}, X_{ij}\omega_{ij}^{\mathrm{V}}\right), \tag{3}$$

where the j-th attention head performs self-attention computation over X_{ij}. The projection layers ω_{ij}^{Q}, ω_{ij}^{K}, and ω_{ij}^{V} transform the input feature split into distinct subspaces. Finally, the self-attention of input features X can be written as:

$$\widetilde{X}_{i+1} = Softmax\left(Concat\left[\widetilde{X}_{ij}\right]_{j=1:n}\right)\omega_i^{\mathrm{L}}, \tag{4}$$

where n is the total number of attention heads, i.e., $X_i = [X_{i1}, X_{i2}, \ldots, X_{in}]$ and $1 \le j \le n$. The linear layer ω_i^{L} projects the concatenated output features back to the dimension consistent with the input, ensuring dimensional aggregation coherence. The SoftMax function is applied to normalize the attention scores and convert them into a probability distribution within the range $[0, 1]$.

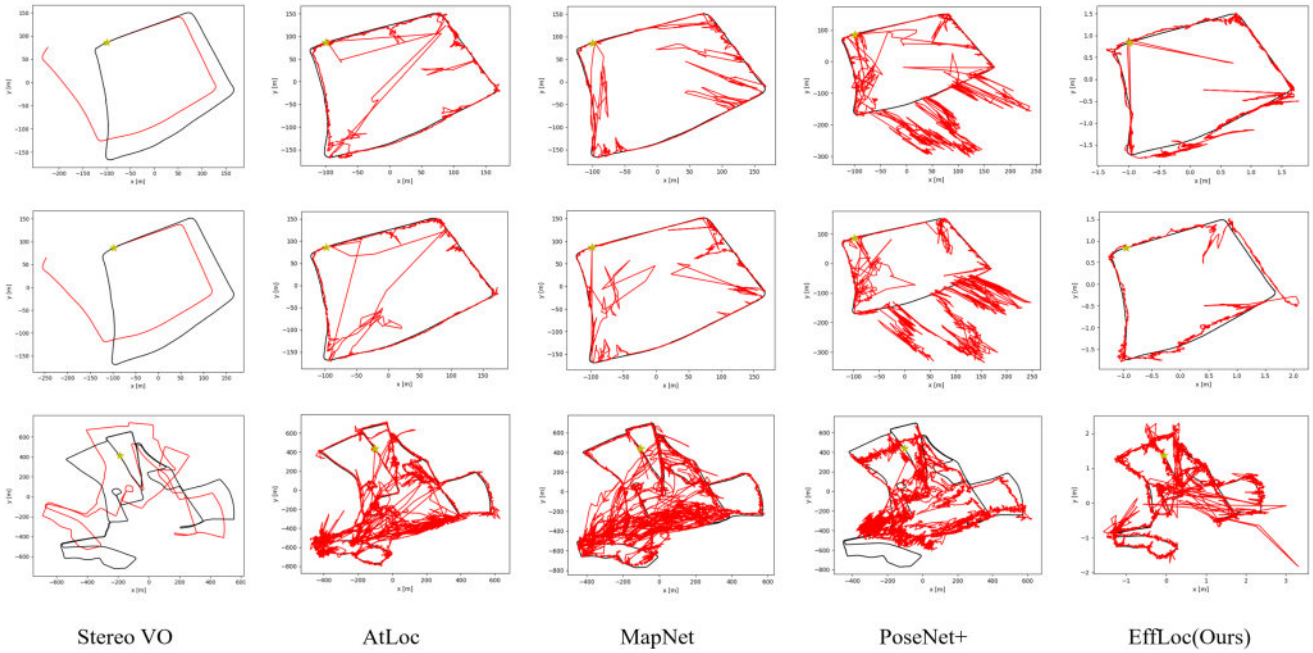


Fig. 3: Trajectories on LOOP1 (top), LOOP2 (middle), and FULL1 (bottom) of Oxford RobotCar. The black lines depict the ground truth trajectories, and the red lines represent the trajectory predictions. A yellow star denotes the starting point in each trajectory.

D. Sequential Group Heads to Improve Feature Representation

While utilizing feature splits instead of the complete features for each attention head enhances efficiency and reduces computational overhead, we aim to further enhance its capacity by encouraging the Q, K, and V layers to learn projections on feature representations that contain more comprehensive information. Figure 2 showcases sequential group heads (SGH) in to calculate the attention map which combine the output of each attention head with the subsequent head to iteratively refine the feature representations:

$$SGH(X_{ij}) = X_{ij} + \widetilde{X}_{i(j-1)}, 1 < j \leq n, \tag{5}$$

where $\mathrm{SGH}(X_{ij})$ is the summation of the $(j-1)$-th head output $\widetilde{X}_{i(j-1)}$ and the j-th position input X_{ij}. Sequential Group Heads enhances the feature representations by adding the output of each head to the subsequent head. An extra token interaction layer, utilizing depth-wise convolution, is incorporated before each feed-forward layer. This approach implements inductive bias to enhance the representation of local and global features.

E. Learning Camera Pose

Our work is built upon prior works in VidLoc [43] pose estimation method, which regresses 6-DoF camera pose from Sequential Group heads guided features $\mathrm{SGH}(X_{ij})$ through Multilayer Perceptrons (MLPs):

$$[p,q] = MLPs\left(SGH\left(X_{ij}\right)\right). \tag{6}$$

Here p represented by the 3D camera position and a 4D unit quaternion q for orientation. The parameters inside the sequential group attention modules are optimized with L1 Loss via the following loss function [41]:

$$loss\left(I\right) = |p-\hat{p}|_1 e^{-\alpha} + \alpha + |\log q - \log \hat{q}|_1 e^{-\beta} + \beta. \tag{7}$$

Here, α and β balance position and rotation losses. The logarithm of a unit quaternion, denoted as $\log q$, offers a three-dimensional, minimally parameterized representation. This characteristic enables direct utilization of L1 distance as the loss function without normalization. The L1 loss reduces outlier impact, enhancing robustness to atypical observations and promoting parameter and feature sparsity. This encourages feature selection and the allocation of zero weights to irrelevant or non-significant features.

Specifically, the unit quaternion $q = (u, v)$ is represented with a scalar u for the real part and a three-dimensional vector v for the imaginary part, defined as:

$$\log \mathbf{q} = \begin{cases} \frac{\mathbf{v}}{\|\mathbf{v}\|} \cos^{-1} u, & \text{if } \|\mathbf{v}\| \neq 0 \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{8}$$

Quaternions are commonly used for camera pose regression due to their continuous and differentiable orientation representation. By normalizing 4D quaternions to unit length, any 3D rotation can be mapped to valid unit quaternions. However, quaternion non-uniqueness arises: $-q$ and q can represent the same rotation due to the two hemispheres. To ensure uniqueness, this study constrains quaternions to a single hemisphere.

IV. EXPERIMENTS

A. Experiment Setup

1) Implementation Details: For consistent network training, images are rescaled via cropping to 256 × 256 pixels

Dataset	PoseNet+		MapNet		AtLoc		**EffLoc (Ours)**	
—	Median	Mean	Median	Mean	Median	Mean	Median	Mean
LOOP1	6.91m, 2.06°	25.39m, 17.49°	5.79m, 1.54°	8.79m, **3.53°**	5.68m, 2.23°	8.61m,4.58°	**5.23m, 2.18°**	**7.58m,** 3.72°
LOOP2	5.83m, 2.05°	28.89m, 19.65°	4.93m, 1.67°	9.81m, 3.86°	5.05m, 2.01°	8.86m, 4.67°,	**4.76m,** 2.06°	**7.89m, 4.19°**
FULL1	107.8m, 23.5°	125.6m, 28.10°	17.91m, 6.68°	41.2m, 13.5°	11.1m, 5.28°	29.6m, 12.4°	**10.28m, 4.98°**	**27.23m, 11.41°**
FULL2	101.9m, 21.1°	131.1m, 26.55°	20.34m, 6.39°	59.5m, 14.7°	12.2m, 4.63°	48.2m, 11.1°	**11.12m, 4.17°**	**44.82m, 9.87°**
Average	55.61m, 12.2°	77.75m, 22.95°	12.24m, 4.07°	29.8m, 8.79°	8.54m, 3.54°	23.8m, 8.19°	**7.85m, 3.35°**	**21.88m, 7.40°**

TABLE I: Camera Relocalization results of the LOOP and FULL trajectories on the Oxford Robotcar dataset. Median and mean errors of position and rotation are calculated for each trajectory using Posenet+, MapNet, AtLoc, and our proposed EffLoc.

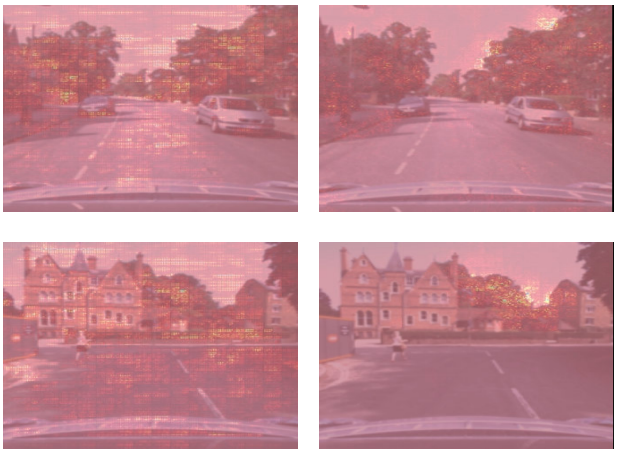
Sequence	Time	Tag	Distance	Mode
–	2014/06/26 8:53	Cloudy	1120m	Training
–	2014/06/26 9:24	Cloudy	1120m	Training
LOOP1	2014/06/23 15:41	Sunny	1120m	Testing
LOOP2	2014/06/23 15:36	Sunny	1120m	Testing
–	2014/11/28 12:07	Cloudy	9562m	Training
–	2014/12/02 15:30	Cloudy	9562m	Training
FULL1	2014/12/09 13:21	Cloudy	9562m	Testing
FULL2	2014/12/12 10:45	Cloudy	9562m	Testing

TABLE II: Training and testing datasets from the Oxford Robot-Car dataset. The testing datasets in LOOP sequence recorded under direct sunlight, whereas FULL datasets are captured under cloudy conditions.

using random and central strategies. Input images are then normalized within a -1 to 1 intensity range. Pretrained model experiments rely on ImageNet-1K [42] data. Model construction employs PyTorch 1.11.0 and Timm 0.6.13, training from scratch for 340 epochs on an Nvidia V100 GPU. We use AdamW [44] optimizer with a cosine learning rate scheduler. During Oxford Robot-car dataset training, we apply the ColorJitter augmentation technique, adjusting brightness, contrast, saturation, and hue (0.7, 0.7, 0.7, 0.5). This augmentation enhances the model's ability to generalize across weather and time variations, making EfficientLoc robust across real-world scenarios (Figure 4). The initial learning rate is 1×10^{-3} with the following hyperparameters: weight decay of 3.5×10^{-2}, mini-batch size of 64, dropout rate of 0.5, and weight initializations of $\alpha = -5.0$ and $\beta = -1.0$.

2) Oxford RobotCar Datasets and Baselines: The Oxford Robot-Car dataset [45] offers a wealth of data with over 100 iterations of a 10km consistent route in central Oxford. This extensive dataset was captured biweekly for more than a year, encompassing a diverse array of environmental conditions, including weather variations, traffic dynamics, pedestrians, construction activities, and roadwork scenarios. Comprising images from six car-mounted cameras, the dataset integrates LIDAR, GPS, INS measurements, and stereo visual odometry (VO). Notably, the dataset presents dynamic elements, including mobile and stationary vehicles, cyclists, and pedestrians, thereby posing significant challenges for vision-based relocalization tasks.

To ensure an equitable comparison, we adopt the evaluation strategy established by MapNet [41]. Our experiments focus on two distinct subsets from this dataset: LOOP (1120m) and FULL (9562m), segregated based on route lengths. For comprehensive information regarding these sequences, consult Table II.



(a) AtLoc (b) EffLoc(Ours)

Fig. 4: Saliency maps of a representative scene from the Oxford Robot-Car dataset illustrate EffLoc's adeptness in guiding the attention of the lightweight transformer towards geometrically resilient objects (e.g., distant skyline and trees edges on the right). This contrasts with environmental dynamics (e.g., the road in the top figure and moving pedestrians in the bottom), as observed in comparison with AtLoc. This emphasis contributes to enhanced global localization robustness.

B. Visual Reloclization in Car-Driving Scenario

The Oxford Robot-car dataset presents significant challenges due to its prolonged data collection period, necessitating a relocalization model with high robustness and adaptability. Table I displays a comparison of our proposed methods against PoseNet+, MapNet, and Atloc. Compared with PoseNet+, EffLoc demonstrates remarkable improvements on both LOOP sequences and FULL sequences. Additionally, we conducted experiments involving Ms-transformer [9] on the Oxford Robot-car dataset. However, their model lacks effective generalization in large-scale testing scenarios, leading to ineffective results.The mean position accuracy is enhanced from 25.39m to 7.58m on LOOP1 and from 28.89m to 7.89m on LOOP2. EffLoc reduces the mean rotation error from 17.49° to 3.72° on LOOP1 and from 19.65° to 4.19° on LOOP2. EffLoc achieves the largest performance gains on FULL1 and FULL2 dataset, surpassing PoseNet+ by 78.3% and 65.8%, respectively. EffLoc exhibits impressive 33.9% and 24.7% improvements in FULL1 and FULL2 routes compared with MapNet by using only a single image. Additionally, when compared to AtLoc, EffLoc achieves an outstanding accuracy in all cases by a large margin. Figure 3 presents the trajectory predictions of LOOP1 (top), LOOP2 (middle) and FULL1 (bottom) using Stereo VO, AtLoc,

Model	EffLoc-XS	EffLoc-Small	EffLoc
LOOP1	23.42m,16.82°	21.51m,14.67°	**7.58m,4.12°**
LOOP2	27.45m,20.19°	25.39m,15.23°	**7.89m,4.19°**
FULL1	109.23m,24.41°	43.23m,18.77°	**27.23m,11.41°**
FULL2	122.79m,20.47°	57.41m,11.79°	**44.82m,9.87°**
Average	70.72m,20.47°	36.89m,15.12°	**21.88m,7.40°**
	Architecture details		
$\{D1, D2, D3\}$	$\{128, 240, 320\}$	$\{128, 256, 384\}$	$\{192, 288, 384\}$
$\{L1, L2, L3\}$	$\{1, 2, 3\}$	$\{1, 2, 3\}$	$\{1, 3, 4\}$
$\{H1, H2, H3\}$	$\{4, 3, 4\}$	$\{4, 4, 4\}$	$\{3, 3, 4\}$

TABLE III: Architecture details of EffLoc model variants in ablation study of EffLoc are reported. We calculate the mean errors of position and rotation and the average of different size EffLoc models. Di, Li, and Hi refer to the width, depth, and number of heads in the i-th stage.

MapNet, PoseNet+, and EffLoc. Despite Stereo VO's smooth predicted trajectories, it experiences substantial drifts as the route length extends. PoseNet+ tends to generate numerous outliers as a result of its strong reliance on local similarities. However, EffLoc significantly reduced these outliers. EffLoc guided by Sequential Group Attention (SGA) captures both local and global geometric features of the diverse and challenging real-world scenarios, enabling robust pose prediction.

C. Efficiency Analysis

To validate the efficiency of SGA mechanism in our EffLoc, we compare against AtLoc, Posenet+Mapnet, and Ms-Transformer models (Figure 1). EffLoc displays superior compute (Flops, Parameters) vs. accuracy trade-off. Total Flops and Parameters are Torchstat-derived. EffLoc excels with 710.95M Flops, 14.99M parameters, notably outperforming AtLoc (29.82M parameters, 5380M Flops). EffLoc-Small enhances efficiency with 11.32M parameters, 397.68M Flops. EffLoc-XS excels, with 8.66M parameters, 227.68M Flops.

EffLoc's efficiency benefits are clear, using 49.7% fewer Flops with 9.6% lower position and rotation error vs. AtLoc. Figure 5 illustrates convergence rate over 50 epochs, crucial for efficiency assessment. EffLoc's Sequential Group Heads refine feature representation, boosting speed 32.9% compared to AtLoc. Figure 6 highlights EffLoc's superior total memory usage versus Atloc, Posenet+, Mapnet, MS-Transformer. Hierarchical layout with memory-bound self-attention and optimized parameters lead to EffLoc's 2.7× memory use reduction, fitting resource constraints.

In contrast to AtLoc's global feature focus, EffLoc adeptly balances local-global feature fusion, maintaining accuracy across real-world applications.

D. Ablation Study

In this section, we perform an ablation study to assess the influence of distinct architectural components in the EffLoc model using the Oxford RobotCar dataset. Architectural details are summarized in Table III. We train three models with varying width, depth, and attention heads for 300 epochs, examining the accuracy-complexity trade-off. To ensure fairness, other modules remain consistent. Table III compares EffLoc, EffLoc-Small, and EffLoc-XS. EffLoc-Small has constrained channel widths (128, 240, 320) in early stages(i)

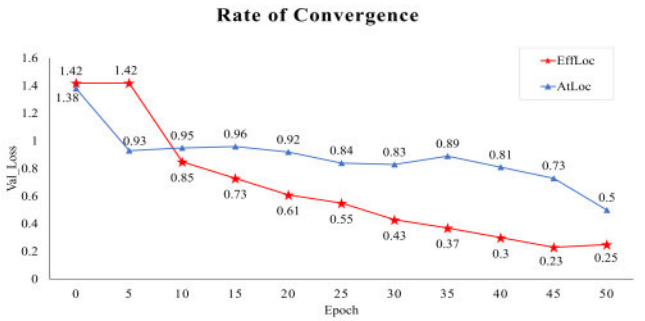


Fig. 5: Convergence velocity performances between EffLoc and AtLoc. The red line (EffLoc) of rate of convergence measures the faster speed converges to the optimal solutions as the epochs increase.

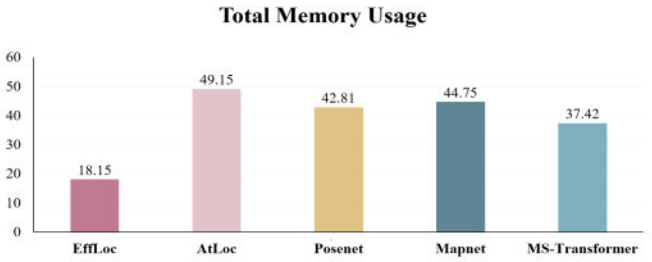


Fig. 6: Memory Usage Comparisons with models. EffLoc optimizes memory consumption while maintaining good performance(18.15MB).

and fewer blocks (1, 2, 3) initially, reducing redundancy and memory use. EffLoc-Small's hierarchy aids faster convergence but incurs slightly higher mean position (70.72m to 36.89m) and rotation errors (20.47° to 15.12°). EffLoc's optimized parameters strike a balance between accuracy and efficiency in camera relocalization. Vital modules feature increased channels, preserving crucial feature information through higher-dimensional learning. Smaller models exhibit improved efficiency yet with marginal accuracy decline. Larger models suit unconstrained resources. EffLoc's design aims to achieve an accuracy-efficiency tradeoff, adaptable to diverse practical scenarios. In conclusion, our ablation study emphasizes the importance of choosing appropriate width, depth, and attention heads in EffLoc for camera relocalization. Evaluation across model sizes underscores optimal accuracy-efficiency balance for real-world performance.

V. CONCLUSIONS

Due to the dynamic and complex nature of real-world long-distance scenes, camera relocalization poses significant challenges in the field of computer vision. Our proposed approach, EffLoc, is based on lightweight transformers and has demonstrated notable improvements in relocalization accuracy and model convergence speed, accompanied by reduced FLOPs and memory usage. In the future, we plan to focus on enhancing EffLoc's robustness and adaptability in dealing with dynamic and complex scenes.

REFERENCES

[1] R. Castle, G. Klein, and D. W. Murray, "Video-rate localization in multiple maps for wearable augmented reality," *2008 12th IEEE International Symposium on Wearable Computers*, Pittsburgh, PA, USA, 2008, pp. 15-22.

[2] Royer, E., Lhuillier, M., Dhome, M. *et al.* "Monocular Vision for Mobile Robot Localization and Autonomous Navigation." *Int J Comput Vision 74*, 2007, pp. 237–260.

[3] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM." In *IEEE Transactions on Robotics,* vol. 37, no. 6, Dec. 2021, pp. 1874-1890.

[4] R. Elvira, J. D. Tardós and J. M. M. Montiel, "ORBSLAM-Atlas: a robust and accurate multi-map system," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* Macau, China, 2019, pp. 6253-6259.

[5] E. Brachmann *et al.,* "DSAC — Differentiable RANSAC for Camera Localization," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* USA, 2017, pp. 2492-2500.

[6] Wang, S., Kang, Q., She, R., Tay, W. P., Hartmannsgruber, A., & Navarro Navarro, D. "RobustLoc: Robust Camera Pose Regression in Challenging Driving Environments." *Proceedings of the AAAI Conference on Artificial Intelligence,* 2023, pp. 6209-6216.

[7] A. Kendall, M. Grimes and R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization," *2015 IEEE International Conference on Computer Vision (ICCV),* Santiago, Chile, 2015, pp. 2938-2946.

[8] S. Wang, R. Clark, H. Wen and N. Trigoni, "DeepVO: Towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks," *2017 IEEE International Conference on Robotics and Automation (ICRA),* Singapore, 2017, pp. 2043-2050.

[9] Y. Shavit, R. Ferens and Y. Keller, "Learning Multi-Scene Absolute Pose Regression with Transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV),* Canada, 2021, pp. 2713-2722.

[10] J. Li *et al.* "Next-ViT: Next generation vision transformer for efficient deployment in realistic industrial scenarios." In *Proceedings of the European conference on computer vision (ECCV),* 2022.

[11] A. Vaswani *et al.*, "Attention is All You Need," *Advances in Neural Information Processing Systems(NIPS),* 2017.

[12] Andrei Ivanov, Nikoli Dryden, Tal Ben-Nun, Shigang Li, and Torsten Hoefler. "Data movement is all you need: A case study on optimizing transformers." In *Proceedings of the Fourth Conference on Machine Learning and Systems (MLSys)*, 2021.

[13] Michel P, Levy O, Neubig G. "Are Sixteen Heads Really Better than One?" *Advances in Neural Information Processing Systems(NIPS)*, 2019.

[14] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV),* Montreal, QC, Canada, 2021, pp. 9992-1000.

[15] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell, "Rethinking the Value of Network Pruning," *International Conference on Learning Representations(ICLR),* 2019.

[16] Bing Wang *et al.* "Atloc: Attention guided camera localization." In *Proceedings of the AAAI Conference on Artificial Intelligence,* volume 34, pages 10393–10401, 2020.

[17] L. Liu, H. Li, and 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),* Oct. 2017, pp. 2391–2400.

[18] D. M. Chen *et al.*, "City-scale landmark identification on mobile devices," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 737-744.

[19] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," *2016 IEEE International Conference on Robotics and Automation (ICRA),* Sweden, 2016, pp. 4762-4769

[20] A. Kendall and R. Cipolla, "Geometric Loss Functions for Camera Pose Regression with Deep Learning," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Honolulu, HI, USA, 2017, pp. 6555-6564.

[21] T. Sattler, Q. Zhou, M. Pollefeys and L. Leal-Taixé, "Understanding the Limitations of CNN-Based Absolute Camera Pose Regression," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Long Beach, CA, USA, 2019, pp. 3297-3307.

[22] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck and D. Cremers, "Image-Based Localization Using LSTMs for Structured Feature Correlation," *2017 IEEE International Conference on Computer Vision (ICCV),* Venice, Italy, 2017, pp. 627-637.

[23] K. Shridhar, F. Laumann, and M. Liwicki, "A comprehensive guide to bayesian convolutional neural network with variational inference." *arXiv preprint arXiv:1901.02731,* 2019

[24] C. Chen *et al.*, "Selective Sensor Fusion for Neural Visual-Inertial Odometry." In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Long Beach, CA, USA, Jun. 2019, pp. 10534–10543.

[25] Wang, J., Qi, Y. "Deep 6-DoF camera relocalization in variable and dynamic scenes by multitask learning." *Machine Vision and Applications*, pp. 34-37, 2023.

[26] Shavit, Y., Keller, Y. "Camera Pose Auto-encoders for Improving Pose Regression." In *Proceedings of the European conference on computer vision (ECCV),* vol 13670, 2022.

[27] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *ICLR,* 2021.

[28] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. Jegou, H.. "Training data-efficient image transformers & distillation through attention." *Proceedings of the 38th International Conference on Machine Learning (PMLR),* pp. 10347-10357, 2021.

[29] "Proceedings of IEEE International Conference on Computer Vision." In *Proceedings of IEEE International Conference on Computer Vision (ICCV),* Jun, 1995.

[30] Li Y, Yuan G, Wen Y, Hu J, Evangelidis G, Tulyakov S, *et al.* "EfficientFormer: Vision Transformers at MobileNet Speed." In *Advances in Neural Information Processing Systems(NIPS),* 2022

[31] J. Zhang *et al.*, "MiniViT: Compressing Vision Transformers with Weight Multiplexing," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* New Orleans, LA, USA, 2022, pp. 12135-12144.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks." In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,* 2018, pp. 4510–4520.

[33] Xiaoqi Jiao *et al.* "Tinybert: Distilling bert for natural language understanding." In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 4163-4174, 2020.

[34] Kenton, J. D. M. W. C., & Toutanova, L. K. "Bert: Pre-training of deep bidirectional transformers for language understanding." In *Proceedings of naacL-HLT, Vol. 1, p. 2,* June, 2019

[35] Mehta S, Rastegari M. "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer." *International Conference on Learning Representations(ICLR),* 2022

[36] S. Mehta and M. Rastegari, "Separable Self-attention for Mobile Vision Transformers." In *Transactions on Machine Learning Research (TMLR),* 2023

[37] S. N. Wadekar and A. Chaurasia, "MobileViTv3: Mobile-Friendly Vision Transformer with Simple and Effective Fusion of Local, Global and Input Features." *International Conference on Learning Representations(ICLR),* 2023

[38] Zhang, H., Hu, W., Wang, X. "ParC-Net: Position Aware Circular Convolution with Merits from ConvNets and Transformer." In *Proceedings of the European Conference on Computer Vision (ECCV),* November, 2022.

[39] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu and Y. Yuan, "EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Canada, pp. 14420-14430, 2023

[40] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early Convolutions Help Transformers See Better." *Advances in Neural Information Processing Systems(NIPS),* 2021.

[41] S. Brahmbhatt, J. Gu, K. Kim, J. Hays and J. Kautz, "Geometry-Aware Learning of Maps for Camera Localization," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, USA, pp. 2616-2625, 2018

[42] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Miami, FL, USA, pp. 248-255, 2009

[43] R. Clark, S. Wang, A. Markham, N. Trigoni and H. Wen, "VidLoc: A Deep Spatio-Temporal Model for 6-DoF Video-Clip Relocalization,"

2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), USA, pp. 2652-2660, 2017

[44] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization." *International Conference on Learning Representations(ICLR),* 2019.

[45] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *Int. J. Robot. Res.,* vol. 36, no. 1, pp. 3–15, Jan, 2017.

[46] D. Cattaneo, M. Vaghi, A. L. Ballardini, S. Fontana, D. G. Sorrenti and W. Burgard, "CMRNet: Camera to LiDAR-Map Registration," *2019 IEEE Intelligent Transportation Systems Conference (ITSC),* New Zealand, pp. 1283-1289, 2019

[47] D. Cattaneo, D. G. Sorrenti and A. Valada, "CMRNet: Camera to LiDAR-Map Registration," *2020 IEEE International Conference on Robotics and Automation (ICRA) Workshop on Emerging Learning and Algorithmic Methods for Data Association in Robotics,* France, 2020