IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

TransAPR: Absolute Camera Pose Regression with Spatial and Temporal Attention

Chengyu Qiao[1], Zhiyu Xiang[†2], Yuangang Fan[1], Tingming Bai[1], Xijun Zhao[3] and Jingyun Fu[1]

***Abstract*—Visual relocalization aims to estimate the absolute camera pose from an image or sequential images. Recent works tackle this problem by exploiting deep neural networks to regress camera poses. However, spatial and temporal clues from sequential images still remain underexplored, resulting in inaccurate poses and large outliers. In this work, we introduce a novel vision Transformer based absolute pose regression model, TransAPR, to tackle this problem. Upon the traditional CNN backbone, we design Transformer based spatial and temporal fusion modules respectively to realize sufficient feature interaction among the neighboring images in the sequence. A hierarchical feature aggregation (HFA) module is further designed to aggregate multi-scale and multi-level features in the pose regressor. Benefiting from these delicate designs, our model is able to generate reliable image representations for absolute pose regression, resulting in more robust localization under challenging environments. We conduct extensive experiments on various indoor and outdoor datasets and show that our method achieves state-of-the-art performance.**

***Index Terms*—Localization, Deep Learning for Visual Perception**

I. INTRODUCTION

VISUAL relocalization is the task aiming at recovering the 3D position and orientation of a camera from one or sequential images. It plays a critical role in a variety of applications including robotics, autonomous driving, and virtual/augmented reality. However, the task is still challenging due to the complexity induced by large perspective, illumination or dynamic changes associated with the camera and scenes.

Traditionally, visual relocalization has been tackled using 3D geometry, which establishes 2D-to-3D correspondences to a prebuilt map and leverages them statistically to estimate the camera pose [1], [2]. Kendall et al. [3] propose the first work that directly regresses the absolute camera pose with a deep neuro-network without a prebuilt map. Their network, called PoseNet, provides a novel end-to-end approach to regress the absolute pose from a single image. Following PoseNet, more single image absolute pose regression (APR) based methods [4]–[6] are proposed. They either use deeper architecture or design different loss functions to improve the performance. Despite their promising progress, these single-image based methods tend to have frequent localization failure in real scenes who have similar visual appearance or dynamic objects.

Compared with a single image, using sequential images for pose regression can be beneficial due to the additional temporal and spatial clues embedded in the images. Along this idea, some recent methods adopt sequential images as input and introduce the temporal constraint by adding visual odometry (VO) [7], [8], Long-Short Term Memory Networks (LSTMs) [9] or Graph Neural Networks (GNNs) [10]. However, adding one or two neighboring frames to the input seems not to give too much help [7], [8]. Methods with LSTM or GNN usually take more images as the input to strengthen the temporal constraint, which increases the memory and time cost.

Recently, Transformer [11] shows promising performance in many computer vision tasks [12], [13] with its strong capability of aggregating features through the self-attention mechanism. MS-Trans [14] first introduced Transformer into the APR task to tackle the pose regression problem over multiple scenes. It employs two transformers with complete encoder-decoder structure for separately regressing translation and orientation output. With only a single image as input, it achieves superior performance than previous works. However, it still suffers some pose outliers due to the lack of sufficient temporal constraints.

In this work, we propose a novel Transformer-based model, namely, TransAPR, for absolute pose estimation with sequential images. Upon the feature extracted from a traditional CNN backbone, Transformer based spatial- and temporal-wise feature interaction modules are separately designed and connected serially to generate robust representations of the sequential images. The spatial-wise module is responsible for enhancing the convolutional features from the backbone and generates global representations for each image. The temporal-wise feature interaction module further strengthens the features by exploiting the temporal relations of neighboring images. Both modules are based on transformers to strengthen the interaction among features. A hierarchical feature aggregation (HFA) block is further designed in the pose regressor for better orientation prediction. By facilitating information propagation within and between frames with Transformers, more robust

Manuscript received: March 27, 2023; Revised May 20, 2023; Accepted June 8, 2023.

This paper was recommended for publication by Editor Sven Behnke upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the Key Research & Development Plan of Zhejiang Province under grant No.2021C01196 and NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization under grant No.U1709214.

[1] Chengyu Qiao, Yuangang Fan, Tingming Bai and Jingyun Fu are with the College of Information and Electronic Engineering, Zhejiang University, Hangzhou 310027, China 3140104437, ygfan, incredibai, fujingyun@zju.edu.cn

[2] Zhiyu Xiang is with the Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking, Zhejiang University, Hangzhou 310027, China xiangzy@zju.edu.cn

[3] Xijun Zhao is with China North Artificial Intelligence & Innovation Research Institute, China North Vehicle Research Institute, Beijing 100072, China

[†] The corresponding author is Zhiyu Xiang.

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

features for absolute pose regression are obtained. We evaluate our approach on three well-known public datasets acquired from indoor and large scale outdoor environments. The results show that our proposed TransAPR outperforms previous absolute pose regression methods and achieves state-of-the-art performance.

In summary, our main contributions are as follows:

- A novel Transformer based absolute pose regression framework which can produce robust pose estimation with only 3 successive sequential images is proposed.
- Transformer-based spatial-wise and temporal-wise feature interaction modules are designed, which can fully utilize spatial and temporal information to enhance image representations for absolute pose regression.
- A hierarchical feature aggregation module is designed to further incorporate multi-scale and multi-level information for orientation prediction.
- Our method achieves state-of-the-art performance on multiple indoor and outdoor datasets containing various challenging conditions.

II. Related Work

A. DNN-based Absolute Camera Pose Regression

Many works have been proposed to regress the absolute pose with deep neural networks. Compared with the structure-based [15], [16] and the image retrieval-based counterparts [17], [18], DNN-based APR methods have the obvious advantage of not requiring any prebuilt map or stored image database during runtime. PoseNet [3] first proposes to regress the camera pose from a single image by appending an MLP head to a GoogLeNet backbone. Subsequent works further extend this approach by adding uncertainty of the estimated poses [19] and a learning weighted loss [4]. These methods are prone to produce large pose errors when facing images acquired from distant positions but with similar appearance. To overcome this problem, sequential image based regression methods are developed to enhance the temporal constraints of the resulting poses. MapNet [7] utilizes pose transformation from visual odometry (VO) as a constraint to strengthen the pose consistency between consecutive frames. Localization performance is further improved by introducing LSTMs [8], [9] or GNNs [10] to the network with strong information exchange among different frames. Recent efforts additionally exploit the attention mechanism [20] or learn neural representations of camera poses in the learned neural space [6]. MS-Trans [14] first utilizes Transformers with complete encoder-decoder structure to regress camera poses over multiple scenes from a single image. ORGMapNet [21] proposes an object relation map to enhance the semantic feature on pose estimation. E-PoseNet [22] introduces a rotation-translation equivariant backbone to encode more geometric information about the image. Different from previous methods, our proposed TransAPR models images as tokens input to Transformers and can extract highly effective spatial-temporal features for robust pose regression. Different from MS-Trans where full encoder-decoder Transformer structures are used for single image pose prediction, we apply transformers with only encoder parts and feed them with sequential images to achieve spatial-wise and temporal-wise feature interaction.

B. Transformers

Transformer is first proposed for the sequence-to-sequence machine translation task [11]. Self-attention, the core mechanism in Transformers, enables the model to aggregate information from the input sequence in a global manner. Very recently, Transformers are showing outstanding performance in solving various computer vision tasks including object detection [23], image recognition [12], segmentation [13], [24] and place recognition [25]. Following DETR [23] and ViT [12], most works model an image as a sequence of patches and feed them into Transformers as tokens for downstream tasks. However, directly applying Transformers to absolute pose regression in a similar way do not exploit the temporal information among sequential images. Considering this difference, we take compressed image representations as tokens and utilize Transformer to extract distinctive features on both spatial and temporal dimensions.

III. Method

This section introduces our TransAPR method in detail. Figure 1 illustrates the architecture of the proposed network, which consists of a CNN backbone, feature interaction modules and a pose regressor. Given sequential images of length T, their corresponding features are extracted individually by the weight-sharing backbone. These features are then fed into the Transformer for spatial-wise feature interaction within the frame and the results are pooled to obtain the image-level features. The temporal-wise feature interaction module further utilizes Transformers to exchange temporal information among different images. Finally, the pose regressor maps the aggregated features to the camera poses $\{\mathbf{t}_i, \mathbf{r}_i\}_1^T$, where $\{\mathbf{t}_i\}_1^T \subseteq \mathbb{R}^3$ are the camera positions and $\{\mathbf{r}_i\}_1^T \subseteq \mathbb{R}^3$ are the orientations represented by the logarithm of the quaternion.

A. Backbone

Following MapNet [7], we adopt ResNet34 [26] as the backbone due to its efficiency and performance. Specifically, the weights of the backbone are initialized with the ResNet34 pretrained on the ImageNet dataset [27]. The backbone is further modified by removing the layers after the last conv layer. Given a sequence with T frames of resolution $H_0 \times W_0$ as input, the backbone generates a feature block $\{\mathbf{F}_i\}_1^T \subseteq \mathbb{R}^{H \times W \times C}$ for each frame. Meanwhile, for each image, the pyramid scale feature maps output by the last three residual blocks are saved for later use.

B. Spatial-wise Feature Interaction

After obtaining the features for each frame, the transformer-based spatial-wise feature interaction module is applied to enhance the convolutional features via self-attention mechanism. Given a feature map $\mathbf{F}_i$, a 1×1 convolution is first applied to reduce the dimension from C to d. Then the spatial dimension of the new feature map is flattened, resulting in a 2D feature

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

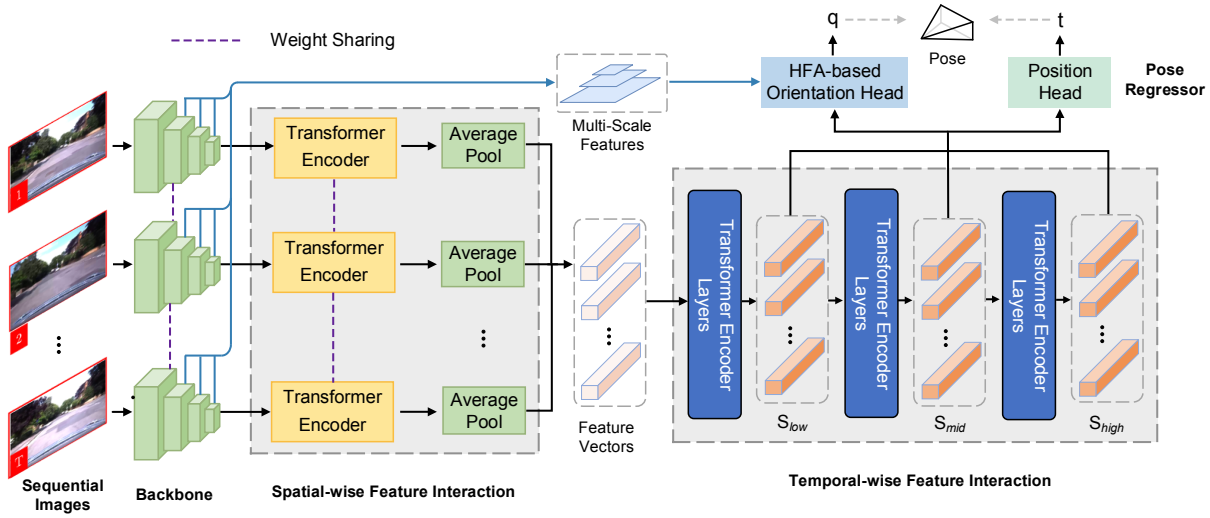


Fig. 1: The network architecture of TransAPR. It contains four main components: 1) a CNN backbone that extracts features for each image in the sequence; 2) a transformer based spatial-wise feature interaction module; 3) a transformer based temporal-wise feature interaction module; 4) a HFA based pose regressor that maps the aggregated features to camera poses.

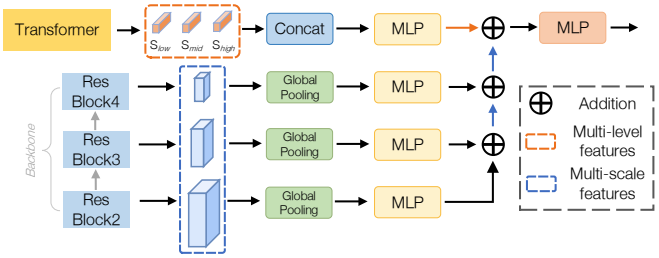


Fig. 2: The structure of the hierarchical feature aggregation **(HFA)** based orientation head. In the upper branch, multi-level features from the temporal Transformer are concatenated along the channel dimension and fed into an MLP. The bottom part of the HFA takes the multi-scale feature maps extracted from the CNN backbone as input, and outputs the processed features with a global average pooling layer followed by an MLP. Finally, all the resulting features are added together and fed into a MLP to produce orientation regression.

map $\mathbf{F}'_i$ of size $H \cdot W \times d$. Fixed positional encodings are further supplemented to the feature to preserve the spatial position information of the image. We adopt the positional encoding from the original Transformer [11]. Specifically, we use $d/2$ sine and cosine functions with different frequencies to encode the coordinates of the feature:

$$PE_{(pos,i)} = \begin{cases} \sin(pos \cdot \omega_k), & \text{for } i = 2k, \\ \cos(pos \cdot \omega_k), & \text{for } i = 2k+1; \end{cases} \quad (1)$$

where $\omega_k = 1/10000^{2k/\frac{d}{2}}$, pos is the feature position in the corresponding spatial dimension. Then, the group of raw patch embeddings input to the Transformer, denoted as $\mathbf{P}_i^0$, is given by:

$$\mathbf{P}_i^0 = \mathbf{F}'_i + \mathbf{E}_{\mathbf{F}_i} \in \mathbb{R}^{H \cdot W \times d} \quad (2)$$

where $\mathbf{E}_{\mathbf{F}_i}$ is the positional encoding of feature $\mathbf{F}_i$.

Next, a Transformer encoder is applied to model the similarities among all the patch-level features for an image. We adopt a standard architecture of the Transformer encoder, which consists of L stacked layers of multi-head self-attention (MHA) modules. In each layer l $(l = 1, 2, \cdots, L)$, a residual connection is employed around each module, followed by layer normalization (LN) [28]:

$$\mathbf{P}_i^{l'} = LN(MHA(\mathbf{P}_i^{l-1}) + \mathbf{P}_i^{l-1}) \in \mathbb{R}^{H \cdot W \times d} \quad (3)$$

$$\mathbf{P}_i^{l} = LN(MLP(\mathbf{P}_i^{l'}) + \mathbf{P}_i^{l'}) \in \mathbb{R}^{H \cdot W \times d} \quad (4)$$

The feature $\mathbf{P}_i^L$ output from the last encoder layer is the group of patch embeddings updated by the attention mechanism. To further decrease the dimension of the obtained features, $\mathbf{P}_i^L$ is passed through a global average pooling layer, resulting in a higher level image representation $\widehat{\mathbf{F}}_i \in \mathbb{R}^d$.

C. Temporal-wise Feature Interaction

Given the features enhanced in the spatial domain, the temporal-wise feature interaction module is applied to fully exchange information among frames and generate more robust features for the APR task. First, we concatenate the spatial features of each frame to form the sequence level feature $\widehat{\mathbf{F}} \in \mathbb{R}^{T \times d}$. In order to preserve the temporal order, we use the position embedding as in Eq.(1) with $\omega_k = 1/10000^{2k/d}$. The full input to the Transformer is formed by adding the positional embedding to the sequence level features $\widehat{\mathbf{F}}$, as:

$$\mathbf{S}_{\widehat{\mathbf{F}}}^0 = \widehat{\mathbf{F}} + \mathbf{E}_{\widehat{\mathbf{F}}} \in \mathbb{R}^{T \times d} \quad (5)$$

where $\mathbf{E}_{\widehat{\mathbf{F}}}$ is the positional encoding of feature $\widehat{\mathbf{F}}$.

The Transformer encoder within the module also contains L layers to realize temporal-wise feature interaction among the frame-level features. Eq.(3)-(4) are then applied again for each layer, but instead of spatial-wise attention, this time the temporal-wise feature attention is computed. In order to effectively integrate information across multiple levels, frame features from the first, middle and last 2 layers of the Transformer, denoted as $\mathbf{S}_{low}$, $\mathbf{S}_{mid}$ and $\mathbf{S}_{high}$, respectively, are drawn out for subsequent pose regression.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Method	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Average	Ranks
Single-scene image-based APRs									
PoseNet15 [3]	0.32/8.12	0.47/14.4	0.29/12.0	0.48/7.68	0.47/8.42	0.59/8.64	0.47/13.8	0.44/10.4	17/17
PoseNet16 [19]	0.37/7.24	0.43/13.7	0.31/12.0	0.48/8.04	0.61/7.08	0.58/7.54	0.48/13.1	0.47/9.81	18/15
PoseNet17 [4]	0.14/4.50	0.27 11.80	0.18/12.10	0.20/5.77	0.25/4.82	0.24/5.52	0.37/10.60	0.24/7.87	14/9
PoseNet+log q [7]	0.11/4.29	0.27/12.13	0.19/12.15	0.19/6.35	0.22/5.05	0.25/5.27	0.30/11.29	0.22/8.07	11/11
LSTM-Pose [29]	0.24/5.77	0.34/11.9	0.21/13.7	0.30/8.08	0.33/7.00	0.37/8.83	0.40/13.7	0.31/9.85	16/16
MLFBPPose [30]	0.12/5.82	0.26/11.99	**0.14**/13.54	0.18/8.24	0.21/7.05	0.22/8.14	0.38/10.26	0.22/9.29	11/14
AtLoc [20]	0.10/4.07	0.25/11.4	0.16/11.8	0.17/5.34	0.21/4.37	0.23/5.42	0.26/10.5	0.20/7.56	6/7
IRPNet [31]	0.13/5.64	0.25/9.67	0.15/13.10	0.24/6.33	0.22/5.78	0.30/7.29	0.34/11.6	0.23/8.49	13/13
NeuralR-Pose [6]	0.12/4.83	0.27/8.91	0.16/12.84	0.19/6.64	0.22/5.45	0.24/6.10	0.29/10.70	0.21/7.92	9/10
E-PoseNet [22]	**0.08/2.57**	**0.21**/11.0	0.16/10.3	**0.15**/6.80	**0.16**/3.82	0.20/6.81	0.24/9.92	**0.17**/7.32	**1**/5
Multi-scene APRs									
MSPN [32]	0.09/4.76	0.29/10.50	0.16/13.10	0.16/6.80	0.19/5.50	0.21/6.61	0.31/11.63	0.20/8.41	6/12
MS-Trans [14]	0.11/4.66	0.24/9.6	**0.14**/12.19	0.17/5.66	0.18/4.44	0.17/5.94	0.26/8.45	0.18/7.28	3/3
Single-scene sequence-based APRs									
VidLoc [8]	0.18/NA	0.26/NA	**0.14**/NA	0.26/NA	0.36/NA	0.31/NA	0.26/NA	0.25/NA	15/-
MapNet [7]	**0.08**/3.25	0.27/11.69	0.18/13.25	0.17/5.15	0.22/4.02	0.23/4.93	0.30/12.08	0.21/7.77	9/8
ORGMapNet [21]	0.09/3.60	0.26/9.49	0.15/12.81	0.20/**4.96**	0.18/5.04	0.22/5.68	0.27/9.54	0.20/7.30	6/4
LSG [9]	0.09/3.28	0.26/10.92	0.17/12.70	0.18/5.45	0.20/**3.69**	0.23/4.92	0.23/11.3	0.19/7.47	4/6
AtLoc+ [20]	0.10/3.18	0.26/10.8	**0.14**/11.4	0.17/5.16	0.20/3.94	**0.16**/4.90	0.29/10.2	0.19/7.08	4/2
TransAPR (Ours)	**0.08**/3.40	**0.21/8.41**	**0.14/9.51**	0.17/5.52	0.18/4.07	0.19/**4.65**	**0.23/8.45**	**0.17/6.29**	**1/1**

TABLE I: Localization results for the 7Scenes dataset. We report the median translation/rotation errors in meters/degrees and the rankings of the average median errors. The best results are highlighted.

D. Hierarchical Feature-based Pose Regression

The pose regressor maps the obtained features from the images to the position and orientation respectively. Different from the most existing methods that take the same feature to the position and orientation head for pose regression, we feed different features to these two heads for better performance. Inspired by MS-Trans [14], we believe different feature granularities are required for the position and orientation prediction. For the position prediction, effective high level features are required to discriminate the current image from the others with similar appearance. However, for the orientation regression, only high level features are not sufficient as they are highly abstract and insensitive to the small orientation change. Multi-scale features are more suitable for this task. MS-Trans [14] employs two separate transformers with full encoder-decoder structure for position and pose regression. However, they take only one image as input and do not consider the temporal constraints of the position. In our design, the high level features after temporal interaction from the transformer are suitable for position prediction. We directly concatenate the three groups of frame features from the Transformer along the channel dimension, the resulting feature vector $\mathbf{S} = \{\mathbf{S}_{low}, \mathbf{S}_{mid}, \mathbf{S}_{high}\} \in \mathbb{R}^{T\times 3d}$ is then fed into a fully connected layer to predict the position $\{\mathbf{t}_i\}_1^T \subseteq \mathbb{R}^3$ for each image. For orientation, we propose a hierarchical feature aggregation (HFA) block to better integrate information from different frames and spatial scales. The structure of the HFA block is illustrated in Figure 2. Given the pyramid features extracted from the CNN backbone, we feed the obtained features from different scales into global pooling layers followed by MLPs to ensure the same dimensionality. Then a bottom-up element-wise addition is applied to obtain the embeddings containing multi-scale visual clues. In another branch, the multi-level features from the Transformer $\mathbf{S} = \{\mathbf{S}_{low}, \mathbf{S}_{mid}, \mathbf{S}_{high}\}$ are concatenated along the channel dimension and fed into a MLP. We merge the feature vectors from the two branches via element-wise addition and feed the fused embeddings into a fully connected layer to predict the orientations $\{\mathbf{r}_i\}_1^T \subseteq \mathbb{R}^3$.

E. Sequence-based Loss Function

Similar to MapNet [7], our TransAPR also minimizes the loss of the absolute pose for each image and the loss of relative poses between image pairs:

$$L_{seq} = \sum_{i=1}^{T} d(\mathbf{p}_i, \mathbf{p}_i^*) + \sum_{i,j=1, i\neq j}^{T} d(\mathbf{v}_{ij}, \mathbf{v}_{ij}^*) \tag{6}$$

where $\mathbf{p}_i$ and $\mathbf{p}_i^*$ are the predicted and the ground-truth absolute poses, respectively. $\mathbf{v}_{ij}$ is the relative pose between the predicted poses of the image pair $(\mathbf{I}_i, \mathbf{I}_j)$. $d(\cdot)$ is the distance function defined as:

$$d(\mathbf{p}_i, \mathbf{p}_i^*) = \|\mathbf{t}_i - \mathbf{t}_i^*\|_1 e^{-\beta} + \beta + \|\mathbf{r}_i - \mathbf{r}_i^*\|_1 e^{-\gamma} + \gamma \tag{7}$$

where β and γ are the learnable parameters used to balance translation and rotation losses, and they are optimized jointly with other parameters in the neural network during training.

IV. EXPERIMENTS

A. Implementation Details

The input to the APR model is monocular RGB sequential images of length 3. As in MapNet [7], all images are resized to the length of 256 on the shortest side and normalized by

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Method	Val Split				Test Split		
	Score ↑	DCRE(0.15) ↑	(Δt,$\Delta\theta$) ↓	Outlier(0.5) ↓	Score ↑	DCRE(0.15) ↑	Outlier(0.5) ↓
Image-Retrieval methods							
NetVLAD [17]	0.673	0.125	(0.93m, 31.44°)	0.452	0.706	0.137	0.431
DenseVLAD [33]	0.604	0.124	(0.98m, 32.26°)	0.520	0.635	0.136	0.501
APRs							
PoseNet [7]	0.624	0.149	(0.89m, 41.89°)	0.525	0.436	0.064	0.628
MapNet [7]	0.682	0.192	(0.84m, 37.43°)	0.510	0.407	0.067	0.66
AtLoc+ [20]	0.594	0.124	(0.90m, 40.64°)	0.53	0.506	0.088	0.582
TransAPR (Ours)	**0.85**	**0.229**	(**0.75m**, **26.04°**)	**0.379**	**0.978**	**0.303**	**0.325**

TABLE II: Comparison on the **val and test** split of the RIO10 benchmark w.r.t. Score (1 + DCRE(0.05) - Outlier(0.5)), DCRE error, pose error and outlier ratio. The best results are in bold.

Method	King's College $5.6 \times 10^3 m^2$	Old Hospital $2.0 \times 10^3 m^2$	Shop Facade $8.8 \times 10^3 m^2$	St. Mary $4.8 \times 10^3 m^2$	Average	Ranks
PoseNet [3]	1.92m, 5.40°	2.31m, 5.38°	1.46m, 8.08°	2.65m, 8.48°	2.09m, 6.84°	8/8
BayesianPN [19]	1.74m, 4.06°	2.57m, 5.14°	1.25m, 7.54°	2.11m, 8.38°	1.92m, 6.28°	7/7
LSTM-PN [29]	0.99m, 3.65°	1.51m, 4.29°	1.18m, 7.44°	1.52m, 6.68°	1.30m, 5.52°	4/6
MapNet [7]	1.07m, 1.89°	1.94m, 3.91°	1.49m, 4.22°	2.00m, 4.53°	1.63m, 3.64°	6/5
IRPNet [31]	1.18m, 2.19°	1.87m, 3.38°	0.72m, 3.47°	1.87m, 4.94°	1.42m, 3.45°	5/4
MS-Trans [14]	0.83m, 1.47°	1.81m, 2.39°	0.86m, 3.07°	1.62m, 3.99°	1.28m, 2.73°	3/3
E-PoseNet [22]	0.95m, 1.63°	1.43m, 2.64°	0.60m, 2.78°	**1.00m**, **3.16°**	1.00m, 2.55°	2/2
TransAPR (Ours)	**0.59m**, **0.86°**	**1.42m**, **2.29°**	**0.54m**, **2.18°**	1.21m, **3.16°**	**0.94m**, **2.12°**	**1/1**

TABLE III: Median pose errors on the Cambridge dataset. The best results are in bold.

pixel mean subtraction and standard deviation division. The ResNet34 [26] component in our network is pretrained on ImageNet and the feature dimension C is 512. For Transformers, both encoders consist of 6 encoder layers with 8 attention heads for feature aggregation. The latent embedding dimension d of the Transformer is 512. To obtain multi-level features, we select the output from the second, fourth and sixth Transformer layer in the temporal module as $\mathbf{S}_{low}$, $\mathbf{S}_{mid}$ and $\mathbf{S}_{high}$.

The whole TransAPR model is implemented with PyTorch [34] on a single GTX 2080Ti GPU. β and γ are set to 0 and 3 for initialization, and dropout rate is 0.5. The Adam [35] with weight decay of 5×10^{-4} is utilized to optimize the parameters with batch size of 32 for 300 epochs in total. The learning rate of the backbone is set to 1×10^{-5} during training, while the learning rate for the other parts is 1×10^{-4} at the beginning and divided by 10 at the 200-th epoch. For the Oxford RobotCar dataset, random ColorJitter is applied during training with brightness, contrast and saturation values set to 0.7 and hue to 0.5. It is effective for improving the model's generalization ability to illumination and weather changes in outdoor scenarios.

To illustrate the performance of our proposed TransAPR, we compare our model with previous absolute pose regression methods. Both single-image-based [3], [4], [19] and sequence-based methods [7]–[9] are considered.

B. Datasets

Two indoor and two outdoor datasets are used for experiments.

7Scenes. The 7-Scenes dataset [36] is an indoor dataset containing seven different scenes recorded by a Kinect sensor. For each scene multiple sequences with 500 to 1000 frames in length are captured in a room. Images with some repetitive textures and textureless regions are included in this dataset, which can cause problems to the APR algorithms.

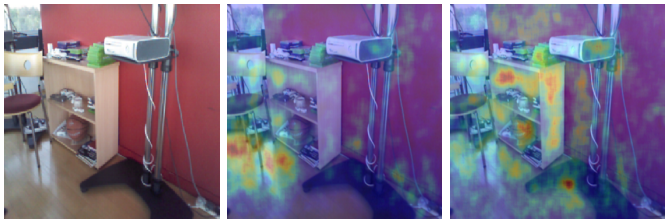


Image MapNet TransAPR (Ours)

Fig. 3: An example image from Chess and the corresponding activation maps, which shows our TransAPR focus on a larger area of the image than MapNet.

RIO10. The RIO10 dataset [37] is captured by a mobile phone in ten different indoor environments. Each scenario is scanned multiple times over a period up to one year. As a long-term relocalization benchmark, the RIO10 dataset contains many challenging situations, e.g., objects and illumination changes in long-time spans. This increases the difficulties of the relocalization task at hand.

Cambridge. The Cambridge dataset [3] is an outdoor dataset containing six different scenes captured by a mobile phone. Dynamic pedestrians and vehicles along with changing lighting and weather conditions bring challenges for the APR algorithms. We select four scenes that are widely benchmarked by APRs for comparison.

Oxford RobotCar. The Oxford RobotCar dataset [38] contains image sequences acquired over 100 repetitions of a continuous route through the central Oxford city over a year, collected by a camera mounted on a vehicle. It contains much longer trajectories and larger areas than the previous three benchmarks. Following MapNet, we evaluate the performance of our method on LOOP and FULL scenarios.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Method	N	Scene				Average
		LOOP1 $8.8 \times 10^4 m^2$	LOOP2 $8.8 \times 10^4 m^2$	FULL1 $1.2 \times 10^6 m^2$	FULL2 $1.2 \times 10^6 m^2$	
PoseNet [4]	1	25.29m, 17.45°	28.81m, 19.62°	125.6m, 27.1°	131.06m, 26.05°	77.69m, 22.56°
AtLoc [20]	1	8.61m, 4.58°	8.86m, 4.67°	29.6m, 12.4°	48.2m, 11.1°	23.8m, 8.19°
MapNet [7]	3	9.84m, 3.96°	8.76m, 3.46°	41.40m, 12.50°	59.30m, 14.81°	29.83m, 8.68°
ORGMapNet [21]	3	-	-	14.41m, 3.73°	36.61m, 9.49°	-
LSG [9]	7	9.07m, 3.31°	9.19m, 3.53°	31.65m, 4.51°	53.45m, 8.60°	25.84m, 4.99°
AtLoc+ [20]	3	7.82m, 3.62°	7.24m, 3.60°	21.0m, 6.15°	42.6m, 9.95°	19.67m, 5.83°
GRNet [10]	8	7.76m, **2.54**°	8.15m, **2.57**°	17.35m, 3.47°	37.81m, 7.55°	17.77m, **4.03**°
TransAPR (Ours)	3	**6.74m**, 3.07°	**7.00m**, 3.12°	**7.96m**, **2.87**°	**34.50m**, **7.26**°	**14.05m**, 4.08°

TABLE IV: Mean pose errors on the RobotCar dataset. "N" indicates the number of images input to the network at one time. The best results are highlighted.

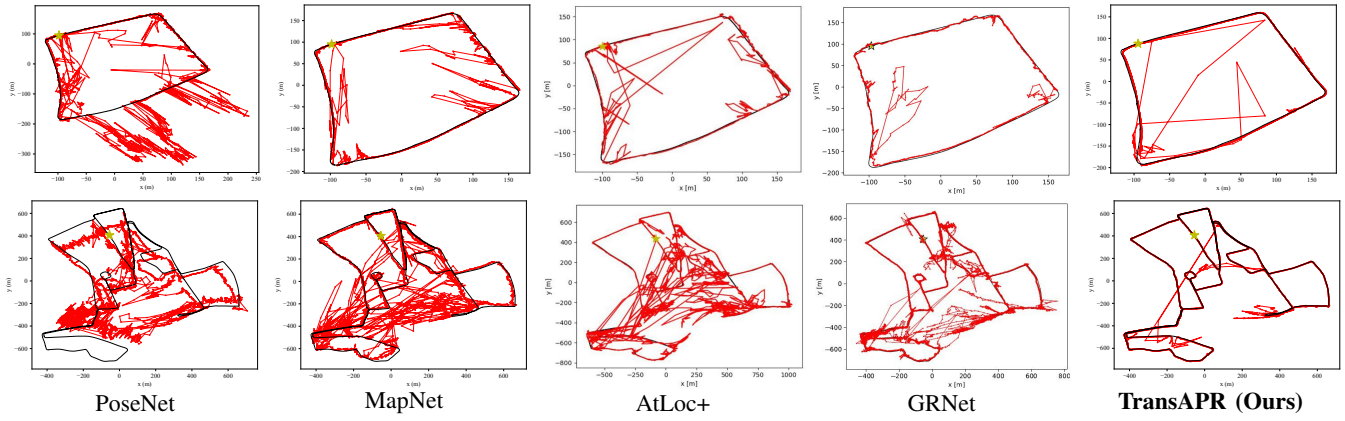


Fig. 4: Comparison of the trajectories on the LOOP1 (top) and FULL1 (bottom) in the RobotCar dataset. The red and black lines represent the predicted and groundtruth positions, respectively.

Fig. 5: An example scene selected from Oxford RobotCar and its corresponding activation maps. Compared with MapNet, TransAPR focuses on a wider area in the image and largely ignores the dynamic objects such as moving vehicles.

C. Experiments on the 7Scenes Dataset

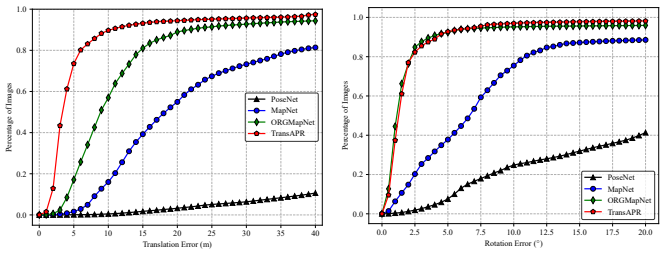
To fully illustrate the performance of TransAPR, we compare our method to previous image-based and sequence-based methods, as well as to existing multi-scene APRs. The quantitative results are shown in Table I, where we can see our method ranks first on the average median pose errors across all scenes. Compared with the single image based APR methods, our method performs the best in most of the sequences, especially in scenes with repetitive (Stairs) and weak texture (Fire). This should be attribute to the introduction of temporal constraints embedded in the sequential images, which can alleviate more visual ambiguity than regressing pose from a single image. Moreover, our TransAPR outperforms the other sequence-based methods by a large margin on the average orientation error. We believe the improvement comes from the specially designed HFA block that effectively aggregates information across multiple Transformer layers and multi-scale visual clues from CNN.

To intuitively understand the reason behind the improvements, we visualize the attention maps of an image taken from the Chess scene. As shown in Figure 3, benefiting from the spatial and temporal attention, TransAPR focuses on larger regions with structural and textural information on the image. In contrast, MapNet focuses on smaller and more concentrated regions, which means that it does not fully utilize the information of the whole image and is susceptible to local region changes.

D. Experiments on the RIO10 Dataset

Containing indoor images recorded over a year, the RIO10 dataset validates the potential of our method in handling changing environments. In the RIO10 dataset, we further adopt their proposed new metric Dense Correspondence Re-Projection Error (DCRE), which is calculated by the average displacement between the 2D correspondences. We compare our algorithm with other APRs as well as some representative image retrieval methods. As shown in Table II, our method outperforms previous APR and image retrieval methods by a large margin. It demonstrates the robustness of our proposed TransAPR on the dynamic indoor environments.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.



Translation error on FULL Rotation error on FULL

Fig. 6: Cumulative distributions of the translation and rotation errors of PoseNet, MapNet, ORGMapNet and our method on the Oxford RobotCar FULL scene. X-axis represents the error and y-axis is the proportion of frames with error less than the value.

E. Experiments on the Cambridge Dataset

We compare our approach with previous APR methods. As shown in Table 3, our method outperforms previous networks in pose errors consistently, achieving the smallest position and orientation errors across all scenes. Especially when the size of the scene increases (King's College and Shop Facade), our method shows more obvious performance improvements.

F. Experiments on the RobotCar Dataset

Table IV presents the quantitative comparison of our method against previous APR methods on the outdoor RobotCar Dataset. The testing routes cover a large scale of the road with seasonal and weather changes. For the LOOP route, our method outperforms previous methods and ranks the first in the translation error. The mean translation error is considerably reduced from $7.76m$ to $6.74m$ on LOOP1, and from $7.24m$ to $7.00m$ on LOOP2. Compared with the LOOP route, the FULL sequences contain a larger area of $1.2 \times 10^6 m^2$ with a total length of $9562m$, which brings greater challenges to the relocalization models. Our proposed TransAPR takes the first place on both position and orientation errors in these two FULL sequences. It is worth noting that compared with some sequence-based APR methods such as LSG and GRNet, our proposed TransAPR aggregates features from only 3 images instead of at least 7. The accuracy gain with fewer images must be attribute to the novel Transformer-based feature interaction modules.

To qualitatively illustrate the effectiveness of our algorithm, we separately plot the trajectories recovered by PoseNet, MapNet, AtLoc+, GRNet and our method, shown in Figure 4. As an image-based APR method, PoseNet produces lots of outliers due to the appearance similarity. The sequence-based methods MapNet and AtLoc+ can better handle this problem by introducing the relative pose between consecutive frames as constraints. However, their estimation results still contain a large number of outliers, especially on the longer FULL sequences. GRNet further reduces the number of outliers by modeling the temporal constraints with graph neuro-network, but there are still many large position drifts on the trajectory. In contrast, our approach significantly reduces the number of outliers and recovers much clearer trajectories than previous methods. An example scene image and its corresponding activation maps are shown in Figure 5. Thanks to the Transformer based spatial attention, TransAPR focuses on a larger region of the image and produces more robust features to disturbances such as illumination changes. The adoption of temporal attention further makes TransAPR focus more on geometrically meaningful regions such as static points and edges rather than dynamic objects, which is especially vital for dynamic traffic scenes.

	Spatial-wise Module	Temporal-wise Module	HFA Block		Average
			multi-scale	multi-level	[m/deg]
(a)					33.00/9.54
(b)	✓				22.89/8.15
(c)		✓			18.52/7.15
(d)	✓	✓			15.19/5.92
(e)	✓	✓	✓		15.01/4.32
(f)	✓	✓		✓	14.25/5.19
(g)	✓	✓	✓	✓	**14.05/4.08**

TABLE V: Ablation experiments on the RobotCar dataset. We report the average of mean translation and rotation errors across all scenes.

We additionally compute the cumulative distribution errors on the FULL route to further compare the pose estimation performance in Figure 6. The left part shows that TransAPR significantly improves position accuracy compared to PoseNet, MapNet, and ORGMapNet. The right part further validates the efficacy of TransAPR to retain orientation accuracy.

G. Ablation Study

To illustrate the effectiveness of each proposed module in TransAPR, we conduct ablation studies on the RobotCar dataset and the results are shown in Table V. We compute the mean translation and rotation errors for each scene and report the average value across all scenes. (a) does not add any extra modules and the features extracted from the CNN backbone are directly fed into pose regression heads, while (b) and (c) introduce the spatial-wise and temporal-wise feature interaction modules, respectively. We observe that both modules are important in the sense that both translation and rotation errors are reduced with introducing either module. Combining the both modules achieves the better performance, as shown in (d). Based on (d), integrating multi-scale visual clues in the HFA block can further improve the orientation accuracy effectively, while integrating multi-level features is beneficial for both translation and orientation prediction. The best performance is achieved by combining both multi-scale and multi-level features to HFA, as the final configuration of the proposed model.

V. CONCLUSIONS

In this paper, we proposed a novel absolute pose regression model TransAPR, which can deeply exploit the spatial and temporal clues among multiple images through self-attention of Transformers. The spatial-wise feature interaction module enhances the spatial relationship of features within one frame. The temporal-wise feature interaction module is able to allow sufficient information exchange among neighboring frames. The multi-scale and multi-level features are fused by the HFA block, generating more effective image representations for orientation prediction. Extensive experiments on various public datasets demonstrate that our model is able to achieve superior performance on challenging scenes.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

REFERENCES

[1] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac-differentiable ransac for camera localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6684–6692.

[2] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl *et al.*, "Back to the feature: Learning robust camera localization from pixels to pose," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3247–3257.

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

[4] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5974–5983.

[5] M. Cai, C. Shen, and I. Reid, "A hybrid probabilistic model for camera relocalization," 2019.

[6] Y. Zhu, R. Gao, S. Huang, S.-C. Zhu, and Y. N. Wu, "Learning neural representation of camera pose with matrix representation of pose shift via view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9959–9968.

[7] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Mapnet: Geometry-aware learning of maps for camera localization," 2017.

[8] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. Wen, "Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6856–6864.

[9] F. Xue, X. Wang, Z. Yan, Q. Wang, J. Wang, and H. Zha, "Local supports global: Deep camera relocalization with sequence enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2841–2850.

[10] F. Xue, X. Wu, S. Cai, and J. Wang, "Learning multi-view camera relocalization with graph neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 11 372–11 381.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8741–8750.

[14] Y. Shavit, R. Ferens, and Y. Keller, "Learning multi-scene absolute pose regression with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2733–2742.

[15] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.

[16] E. Brachmann and C. Rother, "Learning less is more-6d camera localization via 3d surface regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4654–4662.

[17] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[18] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-scale location recognition and the geometric burstiness problem," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1582–1590.

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

[20] B. Wang, C. Chen, C. X. Lu, P. Zhao, N. Trigoni, and A. Markham, "Atloc: Attention guided camera localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 06, 2020, pp. 10 393–10 401.

[21] C. Qiao, Z. Xiang, X. Wang, S. Chen, Y. Fan, and X. Zhao, "Objects matter: Learning object relation graph for robust absolute pose regression," *Neurocomputing*, vol. 521, pp. 11–26, 2023.

[22] M. A. Musallam, V. Gaudillière, M. O. del Castillo, K. Al Ismaeil, and D. Aouada, "Leveraging equivariant features for absolute pose regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6876–6886.

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[24] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[25] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zhen, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," *arXiv preprint arXiv:2201.02001*, 2022.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[29] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using lstms for structured feature correlation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 627–637.

[30] X. Wang, X. Wang, C. Wang, X. Bai, J. Wu, and E. R. Hancock, "Discriminative features matter: Multi-layer bilinear pooling for camera localization," in *British Machine Vision Conference*. York, 2019.

[31] Y. Shavit and R. Ferens, "Do we really need scene-specific pose encoders?" in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 3186–3192.

[32] H. Blanton, C. Greenwell, S. Workman, and N. Jacobs, "Extending absolute pose regression to multiple scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 38–39.

[33] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1808–1817.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2930–2937.

[37] J. Wald, T. Sattler, S. Golodetz, T. Cavallari, and F. Tombari, "Beyond controlled environments: 3d camera re-localization in changing indoor scenes," in *European Conference on Computer Vision*. Springer, 2020, pp. 467–487.

[38] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.