2025 IEEE Intelligent Vehicles Symposium (IV)
June 22-25, 2025. Cluj-Napoca, Romania

PoseViTNet: Multi-Scene Absolute Pose Regression Using Vision Transformers

Asmaa Loulou[1,2] and Mustafa Unel[1]

Abstract— **Accurate camera pose estimation is crucial for autonomous driving and vehicle networking. Traditional pipelines based on geometric models and feature matching struggle in dynamic, featureless environments which are common in many environments. Inspired by the success of vision transformers (ViT), our approach uses a ViT backbone with an attention-based mask to extract a global image descriptor, which is then passed through fully connected layers for pose regression. The multi-headed self-attention in ViT helps the model learn scene layouts and focus on relevant features. We introduce an attention mask to improve performance in challenging scenes, especially dynamic or featureless ones. We compare three backbones: ViT (multi-headed self-attention throughout), ConViT (self-attention in the last two layers, gated positional self-attention elsewhere), and ResNet (pure convolution). We evaluate our model on two commonly used benchmarks for outdoor and indoor localization and we show that our model which uses ViT backbone achieves the state of the art results for both indoor and outdoor multi-scene absolute localization benchmarks.**

I. INTRODUCTION

Estimating camera pose is essential in computer vision applications like autonomous driving and navigation. Traditionally, this is done through geometric methods like feature extraction, PnP algorithms, and image retrieval. Recently, deep learning has been integrated with these methods to improve accuracy and robustness. For example, neural networks can detect keypoints, generate descriptors, or perform matching. Other approaches, like hierarchical localization, use neural networks for image retrieval to find correspondences, which are then mapped to 2D-3D correspondences and used in a PnP method with RANSAC. While these methods achieve state-of-the-art accuracy, they are slow and require accurate 3D models and hyperparameter tuning.

Alternatively, end-to-end deep learning approaches or Absolute Pose Regressors (APRs) are trained to regress the absolute pose with one forward pass using one query image. Such methods use a CNN backbone as a feature extractor with a multi layer perceptron (MLP) head to regress the pose and do not require knowledge of the 3D structure of the scene. Although these models are at least one magnitude faster than the 3D based methods, APRs are also an order of magnitude less accurate [1]. Moreover, APRs are usually trained per scene which means that they suffer from limited generalization. Recently, some methods have addressed the generalization problem of APRs. Blanton et al. proposed a two-stage network with a shared CNN backbone and scene-specific fully connected layers [2]. However, the model still required per-scene weights and failed to achieve SOTA results. Shavit et al. proposed a transformer-based approach where they used two transformers to process feature maps from a CNN backbone [1]. Scene-specific queries were encoded to generate latent representations, followed by a classifier and MLP head for pose regression. Lee et al. studied the generalization problem of APRs from a different perspective and tried to prevent the distortion of query-key space and encourage the model to find global relations by self-attention by utilizing an auxiliary loss that aligns queries and keys [3]. However, both models relied on a CNN backbone and required a classifier with predefined scene classes.

To the best of our knowledge, all existing absolute pose regression models have relied on convolutional neural networks (CNNs) either as the primary backbone or as preliminary feature extractors, in order to finally generate one or two global latent feature vectors for pose estimation. In this work, we propose a new vision transformer based absolute camera pose regression architecture (PoseViTNet) and an attention-based mask. We conduct multiple experiments to demonstrate that the multi-headed self attention mechanism in Vision Transformers (ViT) facilitates the model's capacity to transcend the locality that is introduced by CNNs. Moreover, ViT's inherent absence of inductive biases compared to CNNs proves advantageous in generating more effective latent global feature vectors which will result in better overall pose estimates. We evaluate our approach on two widely used datasets, consisting of multiple indoor and outdoor scenes. The main contributions of this work are summarized as follows:

- We introduce a fully ViT based architecture for absolute camera pose regression in both indoor and outdoor scenes. We do not use CNN based features nor a classifier at any stage in our architecture.
- We introduce an attention-based mask designed to enhance the model's ability to concentrate on crucial regions within the image.
- We show that introducing soft convolutional inductive bias into ViT (ConViT) actually degrades the results which highlights the importance of the multi-headed self

*This work was supported in part by Scientific and Technological Research Council of Turkey (TUBİTAK) and in part by Ford Otosan under Grant 118C061.

[1]Asmaa Loulou and Mustafa Unel are with the Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkey (aloulou@sabanciuniv.edu; munel@sabanciuniv.edu)

[2]Asmaa Loulou is with Sancaktepe Engineering Center, Ford Otosan, Istanbul 34885, Turkey

2025 IEEE Intelligent Vehicles Symposium (IV) | 979-8-3315-3803-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/IV64158.2025.11097720

attention mechanism.
- We show that the ViT features explicitly contain the scene layout and can identify the parts of image most relevant for pose estimation in each scene.

II. RELATED WORK

There are many methods available for camera pose estimation. These methods can be divided into different categories depending on their algorithmic properties.

Absolute Pose Regression was initially introduced by Kendall et al. [4], drawing inspiration from the success of convolutional neural networks (CNNs) in various computer vision tasks. Typically, a CNN backbone pre-trained on large datasets like ImageNet is fine-tuned with a multi-layer perceptron (MLP) head to regress both translation and orientation from a query image [5], [6]. Walsh et al. incorporated LSTM layers to reduce dimensionality and overfitting [7], but scene-specific training remained a limitation. Chen et al. [8] improved accuracy through direct matching supervised by photometric similarity between the query and rendered images, achieving SOTA results. However, their approach suffers from photometric inconsistencies and domain gaps, as rendered images may differ from real ones. A later work [9] addressed these issues using feature-space matching but still required per-scene training.

Multi-scene Absolute Pose Regression Blanton et al. proposed a two-stage network with a shared CNN backbone and scene-specific fully connected layers [2]. The CNN generates a global image descriptor for scene prediction and pose estimation. One branch indexes a database of scene-specific weights, while the other transforms the descriptor into pose parameters. However, this approach fails to achieve SOTA results and still requires a scene-specific classifier. Shavit et al. proposed a transformer-based architecture [1], where encoders aggregate activation maps, and decoders transform latent features into pose predictions. Despite using self-attention, this method still relies on a CNN backbone and a classifier. Lee et al. studied the generalization problem of APRs from a different perspective and tried to prevent the distortion of query-key space and encourage the model to find global relations by self-attention [3].

Vision Transformers. Transformers are the defacto standards for sequence modeling in natural language processing due to their simplicity and computational efficiency.Recently, following the work of Dosovitskiy et al. [10], which demonstrated the scalability and effectiveness of Vision Transformers (ViTs) with large-scale data, ViTs have gained increasing attention in a variety of computer vision tasks, including semantic segmentation [11] and image classification [10], [11]. D'Ascoli et al. introduced ConViT [12] where they altered the multi-headed attention mechanism used in the original ViT in order to softly introduce a convolutional inductive bias into ViT. They used gated positional self attention in the first 10 layers in order to mimic the locality of the convolutional layers and then give each head the freedom to escape locality by adjusting a gating parameter that regulate the attention paid to position versus the attention paid to content. Different versions of ViTs have been introduced and many training schemes have been proposed.

We use the ViT architecture [10] both as a backbone and to generate an attention based mask that will further enhance our network ability to learn challenging dynamic or featureless scenes. We show that ViT-based architectures can be used to perform multi-scene absolute pose regression and we show their superior performance compared to both ConViT [12] and ResNet. We further show that our architecture does not need to learn a certain distribution over N scenes and thus needs no classifier of any kind. We propose a single network trained in an end to end fashion that is able to regress the pose through an arbitrary number of scenes.

III. MULTI-SCENE ABSOLUTE POSE REGRESSION WITH VISION TRANSFORMERS

PoseViTNet estimates the camera pose in a single forward pass, given a single query image as input. The pose is represented as $p = [x, q]$. $x \in \mathbb{R}^3$ represents the absolute camera position in the world coordinates and $q \in \mathbb{R}^4$ is a quaternion that represents the camera orientation. Similar to [4], [13], [14], we represent rotation using quaternions because they are 4-D vectors that are easily mapped to reasonable rotation matrices after being normalized to unit length.

A. Network Architecture

The network architecture is shown in Figure 1. ViT uses attention instead of convolution to encode input images [10]. An input image $I \in \mathbb{R}^{H\times W\times C}$ is reshaped into fixed-size patches $x \in \mathbb{R}^{N\times(P^2C)}$, where (P, P) is the patch resolution, and $N = HW/P^2$. These patches are flattened and embedded into D_{emb} dimensions via a trainable linear projection $E \in \mathbb{R}^{(P^2C)\times D_{emb}}$. A learnable class token x_{class} is concatenated with the embedded sequences, and positional embedding $E_{pos} \in \mathbb{R}^{N+1\times D_{emb}}$ is added.

$$z_0 = [x_{class}; x_1E; ..., x_NE] + E_{pos} \tag{1}$$

The resulting embeddings are then passed through transformer encoder which consists of alternating layers $\ell = 0, 1, 2..., L$ of multi-headed self attention (MSA) and MLP blocks with layer normalization (LN) applied before each block and residual connection after each block [10], as shown by Equations (2) and (3).

$$z'_\ell = MSA(LN(z_{\ell-1})) + z_{\ell-1} \tag{2}$$

$$z_\ell = MLP(LN(z'_\ell)) + z'_\ell \tag{3}$$

In ViT, the attention for each head is calculated as shown in Equation (4):

$$A = softmax(\frac{QK^T}{\sqrt{D_h}}) \tag{4}$$

There are three learnable matrices W_{key}, W_{qry}, $W_{val} \in \mathbb{R}^{D_{emb}\times D_h}$ and $Q = z_\ell W_{qry}$, $K = z_\ell W_{key}$ where z_ℓ is the embedded patches from the input image and each patch is of dimension D_{emb}. The attentions from the heads are

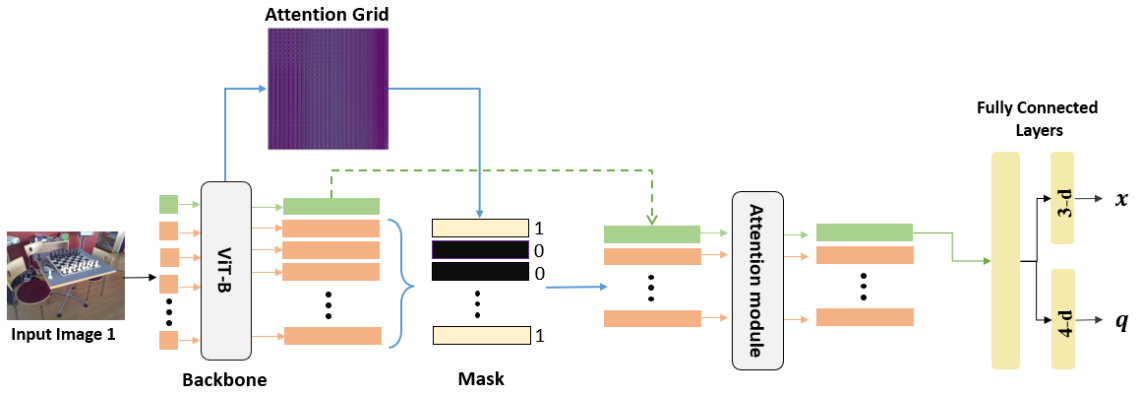

Fig. 1: **PoseViTNet architecture**: After passing the image to the ViT backbone, we use the attention grid to generate a mask and filter out the patches' embedding that includes dynamic object or redundant features. The remaining embeddings and the class token are passed through a smaller attention module. Fully connected layers are used to regress the rotation and translation.

concatenated and the final MSA is calculated as shown in Equation (5):

$$MSA(X) = concat_{h\in[N_h]}[A^h z_\ell W^h_{val}]W_{out} + b_{out} \quad (5)$$

where N_h are the available heads, $W_{out} \in \mathbb{R}^{D_{emb}\times D_{emb}}$ and $b_{out} \in \mathbb{R}^{D_{emb}}$ are matrices used to preserve dimension.

We further propose to learn a mask in order to eliminate the patches of the image which may harm the overall pose estimation such as patches which are featureless or which include moving objects. The attention grids $A^h \in \mathbb{R}^{N\times N}$ generated using Equation 4 where N is the number of patches and h is the head includes the attention performed by each patch on every other patch. Thus, we calculate the mean of the grids over the number of heads. Then we calculate the mean of the overall attention generated on each patch by the other patches and generate an overall attention vector V_a. We finally apply a softmax function in order to generate a probability distribution overall the patches in the vector $V_a \in \mathbb{R}^{N\times 1}$ as shown in Equation (6):

$$V_a = softmax(\frac{1}{N}\sum_{n=1}^{N}(\frac{1}{N_h}\sum_{h=1}^{N_h}A^h)) \quad (6)$$

then we apply a threshold r on every single value v_a within the vector V_a in order to generate the final mask as shown in Equation 7:

$$mask(v_a) = \begin{cases} 0 & v_a \le r \\ 1 & v_a > r \end{cases} \quad (7)$$

The mask is calculated through Equations (6) and (7) using only the patches of the image without the extra learnable sequence x_{class} which is shown in green in Figure 1. We set the threshold to 0.0012. Then this mask is applied on the final image embedding z_ℓ in order to filter out the patches which have the lowest attention values. The remaining embeddings are concatenated with the extra learnable sequence x_{class} and passed through one more attention module. The attention module consist of the same alternating layers of multi-headed self attention (MSA) and MLP blocks which are used within the ViT backbone (Equations 2 and 3). While there are 12 attentional layers in the ViT backbone, we use 6 attentional layers within the attention module in order to reduce the size of the overall model and prevent overfitting. The extra learnable sequence $z_0^0 = x_{class}$ serves as an image representation at the output of the transformer encoder. During training, the multi-headed self-attention mechanism ensures this sequence attends to all embedded patches. The regression head is attached to the resulting state of this sequence after the attention module, as shown in Figure 1. Section IV.G shows that attaching the regression head to a pooled vector from all patches, in addition to the extra sequence, leads to worse results.

B. Loss function

Joint training of orientation and translation regressors has been shown to improve overall performance [4]. Accordingly, we employ stochastic gradient descent to minimize a combined loss function comprising both translation and rotation errors:

$$\mathcal{L}(I) = \|\hat{x} - x\| \exp(-s_q) + s_q + \left\|\frac{\hat{q}}{|\hat{q}|} - q\right\| \exp(-s_x) + s_x \quad (8)$$

where $\hat{x}$ and $\hat{q}$ are the estimated translation and rotation vectors. x and q are ground truth translation and rotation vectors. We combine the two losses using the camera pose loss function suggested by Kendal et al. [4] s_q and s_x are learnable parameters controlling the balance between the orientation and position losses.

IV. Experiments

A. Datasets and Experimental Setup

We evaluate our method on the **Cambridge Landmarks** [4] and **7-Scenes** [18] datasets, representing outdoor and indoor environments, respectively. Cambridge includes

TABLE I: **PoseViTNet indoor localization results for 7Scenes dataset.** We report the median position/orientation error in meter/degrees for each method.

	Method	chess	fire	heads	office	pumpkin	red kitchen	stairs	Average[m/deg]
APR	PoseNet [4]	0.32, 6.60	0.41, 14.4	0.29, 12.0	0.48, 7.20	0.49, 8.10	0.49, 8.34	0.48, 13.1	0.45, 9.94
	LSTM-Pose [7]	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
	Geometric PoseNet [15]	0.13, 4.48	0.27, 11.3	0.17, 13.0	0.19, 5.55	0.26, 4.75	0.23, 5.35	0.35, 12.4	0.23, 8.12
	AtLoc [16]	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
	IRPNet [6]	0.13, 5.64	0.25, 9.67	0.15, 13.1	0.24, 6.33	0.22, 5.78	0.30, 7.29	0.34, 11.6	0.23, 8.49
	HyperPose [17]	**0.08**, 6.29	0.22, 11.2	**0.11**, 12.7	0.17, 7.53	**0.16**, 6.66	0.17, 8.48	0.26, 10.8	**0.17**, 9.09
M-APR	MSPN [2]	0.09, 4.76	0.29, 10.5	0.16, 13.1	**0.16**, 6.80	0.19, 5.50	0.21, 6.61	0.31, 11.6	0.20, 8.41
	MS-Transformer [1]	0.11, 4.66	0.24, 9.60	0.14, 12.2	0.17, 5.66	0.18, 4.44	0.17, 5.94	0.26, 8.45	0.18, 7.28
	MS-APR [3]	0.10, 4.15	0.24, 8.79	0.14, 11.6	0.17, 5.28	0.17, **3.48**	0.17, 5.62	**0.22**, 7.58	0.17, 6.64
	PoseViTNet (ours)	0.10, **2.84**	**0.21**, **7.26**	0.13, **8.27**	**0.16**, **3.95**	**0.16**, 4.11	0.21, **3.64**	0.28, **7.03**	0.18, **5.30**

TABLE II: **PoseViTNet outdoor localization results for Cambridge Landmarks dataset.** We report the median position/orientation error in meter/degrees for each method.

	Method	Kings College	Old Hospital	Shop Facade	St. Marys C.	Average[m/deg]
APR	PoseNet [4]	1.66, 4.86	2.62, 4.90	1.41, 7.18	2.45, 7.96	2.04, 6.24
	LSTM-Pose [7]	0.99, 3.65	1.51, 4.29	1.18, 7.44	1.52, 6.68	1.30, 5.52
	Geometric PoseNet [15]	0.88, 1.04	3.20, 3.29	0.88, 3.78	1.57, 3.32	1.63, 2.85
	IRPNet [6]	1.18, 2.19	1.87, 3.38	0.72, 3.47	1.87, 4.94	1.42, 3.45
	HyperPose [17]	**0.56**, 2.40	**1.41**, 2.91	0.54, 3.37	**0.98**, 4.86	**0.87**, 3.43
M-APR	MSPN [2]	1.73, 3.65	2.55, 4.05	2.92, 7.49	2.67, 6.18	2.47, 5.34
	MS-Transformer [1]	0.83, 1.47	1.81, 2.39	0.86, 3.07	1.62, 3.99	1.28, 2.73
	MS-APR [3]	0.88, 1.29	1.55, **1.87**	0.79, 2.51	1.57, 3.50	1.19, 2.29
	PoseViTNet (ours)	0.62, **0.84**	1.82, 2.32	**0.53**, **1.88**	1.09, **2.85**	1.02, **1.97**

six urban scenes ($\sim$900–5500 m^2), of which four are used. 7-Scenes contains seven office scenes ($\sim$1–10 m^2).

Images are resized to 224×224, normalized, and mean-centered. We use a ViT backbone pre-trained with DINO [11]. Training uses a batch size of 8 and an initial learning rate of 10^{-5}, decayed twice every 50k steps. All experiments are conducted on a single NVIDIA GeForce RTX 3080 GPU with 16 GB of memory.

B. Comparison with state of the art relative pose regressors

We compare our network to the available multi-scene and single scene APRs. To the best of our knowledge the only available multi-scene APRs are MS-Transformer [1], MSPN [2] and MS-APR [3]. Our goal is to show the ability of our network to be trained for multi-scene and achieve good results despite using a simple loss function and no further fine-tuning. Thus, we do not include methods that use a more advanced loss function to train the model such as photometric error [9], [8], [19] or that utilize additional data at inference time such as the relative pose methods [20], [21], [22]. We group the methods into 2 main categories: (i) Absolute Pose Regression (APR), and (ii) Multi-scene Absolute Pose Regression (M-APR). While APR methods are trained for each scene, M-APR methods are trained once for all scenes in each dataset.

1) Indoor Results: Table I shows that on 7-scenes dataset, our network increased the orientation accuracy with an average of 20.2% compared to MS-APR [3] while decreasing the position accuracy with only an average of 5.9%. Compared to Hyperpose [17] which is trained per scene, our network also increased the orientation accuracy with an average of 41.7% while decreasing the position accuracy with only an average of 5.9%. Our network achieves the best rotation accuracy compared to both multi-scene APRs and single-scene APRs.

2) Outdoor Results: Table II shows that on Cambridge landmarks dataset, our network increased the orientation accuracy with an average of 13.9% and increased the translation accuracy with an average of 14.2% compared to MS-APR [3]. Compared to Hyperpose [17] which is trained per scene, our network also increased the orientation accuracy with an average of 42.6% while decreasing the position accuracy with only an average of 25.3%. Our network achieves both the best rotation and translation accuracy compared to multi-scene APRs and achieves the best rotation and the second best translation compared to all APRs.

C. Attention Maps Visualization and Interpretation

ViTs focus on the most relevant areas of an image for the task at hand. As shown in [11], attention maps can capture the scene layout and object boundaries after training. Similarly, we visualize attention maps from the last block of the ViT backbone. After training, the model ignores moving objects and focuses on key details of buildings or landmarks. Figure 2 shows attention maps for 3 scenes (a,b,c) from the Cambridge Landmarks dataset, highlighting landmarks while ignoring the background and moving objects. For instance, in the shop facade scene, more attention is placed on the building than on people in front of it. Similarly, for 3 scenes (d,e,f) from the 7-scenes indoor dataset, the model focuses on common objects across frames and less on background elements like walls.

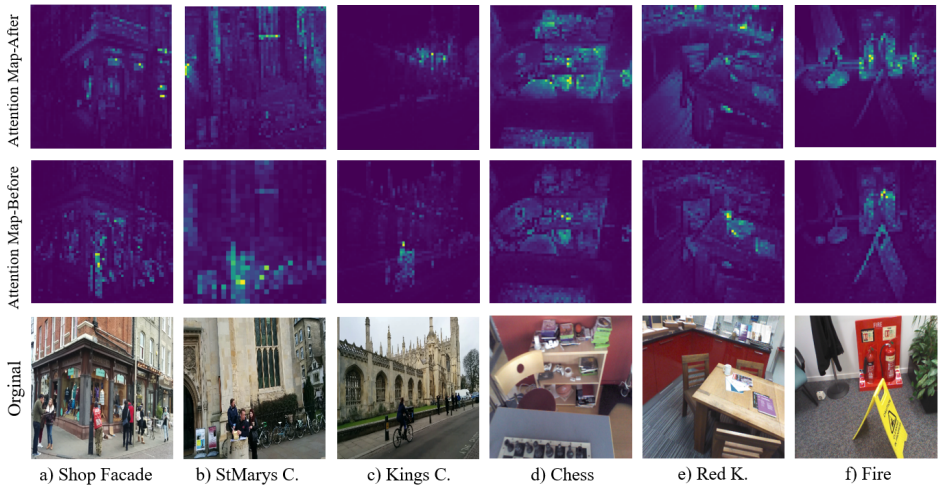

Fig. 2: **Attention Maps for different scenes**

Fig. 3: **Attention Maps at ViT backbone and attention module.**

TABLE III: **Attention Module Importance.** We report the median position/orientation error in meter/degrees.

Method	Cambridge landmarks	7-scenes
PoseViTNet w/o attention module	1.04, 2.42	0.19, 7.01
PoseViTNet	**1.02**, **1.97**	**0.18**, **5.30**

D. Importance of the mask and the attention module

We investigate the importance of the mask and attention module by visualizing attention maps from the last self-attention block. The mask helps the network filter out irrelevant patches, such as moving objects or featureless surfaces, which could harm pose estimation. Figure 3 shows attention maps from the ViT backbone and attention module after training. After removing such patches, the model focuses on the main landmark, ignoring moving objects and featureless surfaces like walls. Table III shows that the pose estimation accuracy (both orientation and position) improves with the attention module and mask.

E. Ablation study: The backbone architecture

In order to further study the architecture of our model we performed an ablation study on the backbone using the Cambridge Landmarks dataset. We considered 5 different backbones: two ViT based architectures, one ConViT based architecture and two completely convolutional based. In order to be able to test the backbone separately without the mask and the attention module.

ConViT [12] uses gated positional self attention (GPSA) instead of multi-headed self attention (MSA). The GPSA layers are initialized to mimic the locality of convolutional layers as it uses a relative positional encodings within the attention to allow the network to focus on position. In GPSA, the attention between two patches i and j is calculated as shown in Equation (9) and the final output for each head is calculated as shown in Equation (10):

$$A_{ij}^{h} = (1-\sigma(\lambda_h))softmax(Q_i^h K_j^{hT}) + \sigma(\lambda_h)softmax(v_{pos}^{hT} r_{ij}) \quad (9)$$

$$GPSA_h(X) = Normalize[A^h]XW_{val}^h \quad (10)$$

where σ is the sigmoid function and λ_h is a gating parameter that allows the network to decide between focusing on position or content. v_{pos}^h is a trainable embedding and r_{ij} is a relative positional encoding depends on the distance between i and j. ConViT uses GPSA in the first 10 layers and MSA in the last 2 layers which means that ConVit introduce soft inductive biases to the overall model. For the ViT and ConViT backbones, we attached the regression head to the learnable sequence output of the transformer encoder, and for the convolutional backbone, we attached it to the feature vector obtained by average pooling the last layer. Different backbones have different feature vector sizes, so we adjusted the fully connected layer input size accordingly. All models were pretrained on ImageNet. As shown in Table IV, ViT-based architectures outperform ConViT and ResNet models in both orientation and position. ViT-B slightly outperforms

TABLE IV: **Different backbones**. We report the median position/orientation error in meter/degrees.

Backbone	FC input	FC output	Position	Orientation
ViT-S	384	768	1.06	2.57
ViT-B	**768**	**1536**	**1.04**	**2.42**
ConViT	768	1536	1.11	2.64
Resnet-18	512	1024	1.79	4.35
Resnet-50	2048	4096	4.35	7.20

TABLE V: **Global image descriptor.** We report the median translation/rotation error in meter/degrees.

Global image descriptor	Position	Orientation
pooled vector from all patches	3.90	2.13
class token	**1.97**	**1.01**

ViT-S. These results highlight the benefit of using a ViT backbone, which leverages multi-headed self-attention to attend to image patches from the first layer, avoiding locality issues in ConViT. ConViT generalizes better than ResNet in multi-scene scenarios.

F. Ablation study: The global image descriptor

We explore the impact of utilizing the class token alone as a global image descriptor, that will be passed through the fully connected layers, versus using a vector pooled from both the class token and filtered embeddings after the mask and the attention module. Table V demonstrates superior results when utilizing only the class token.

V. CONCLUSIONS

In this work we introduced a new camera pose regression network using a ViT backbone with an attention-based mask, without CNNs or a classifier. The multi-headed self-attention mechanism in ViT enables the network to focus on important image parts for pose estimation and avoid locality. We showed that our mask effectively filters out irrelevant moving objects, achieving state-of-the-art accuracy in multi-scene absolute pose regression for both indoor and outdoor benchmarks. Specifically, our model improved orientation accuracy with an average of 20.2% while decreasing the position accuracy with only an average of 5.9% for the outdoor dataset, and improved the orientation accuracy with an average of 13.9% and position accuracy with an average of 14.2% for the indoor dataset. Future work will explore advanced loss functions to further improve accuracy.

REFERENCES

[1] Y. Shavit, R. Ferens, and Y. Keller, "Learning multi-scene absolute pose regression with transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2713–2722.

[2] H. Blanton, C. Greenwell, S. Workman, and 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.

[3] M. Lee, J. Kim, and J.-P. Heo, "Activating self-attention for multi-scene absolute pose regression," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=rM24UUgZg8

[4] A. Kendall, M. Grimes, and 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.

[5] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, "Image-based localization using hourglass networks," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 870–877.

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

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

[8] S. Chen, Z. Wang, and V. Prisacariu, "Direct-posenet: Absolute pose regression with photometric consistency," in *2021 International Conference on 3D Vision (3DV)*, 2021, pp. 1175–1185.

[9] S. Chen, X. Li, Z. Wang, and V. A. Prisacariu, "Dfnet: Enhance absolute pose regression with direct feature matching," in *Computer Vision – ECCV 2022*. Cham: Springer Nature Switzerland, 2022, pp. 1–17.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[11] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9630–9640.

[12] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2286–2296.

[13] S. En, A. Lechervy, and F. Jurie, "Rpnet: An end-to-end network for relative camera pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[14] C. Yang, Y. Liu, and A. Zell, "Relative camera pose estimation using synthetic data with domain adaptation via cycle-consistent adversarial networks," *Journal of Intelligent & Robotic Systems*, vol. 102, no. 4, pp. 1–17, 2021.

[15] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6555–6564.

[16] B. Wang, C. Chen, C. X. Lu, P. Zhao, N. Trigoni, and A. Markham, "Atloc: Attention guided camera localization," *arXiv preprint arXiv:1909.03557*, 2019.

[17] R. Ferens and Y. Keller, "Hyperpose: Camera pose localization using attention hypernetworks," *arXiv preprint arXiv:2303.02610*, 2023.

[18] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi, "Real-time rgb-d camera relocalization," in *International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, October 2013.

[19] X. Wu, H. Zhao, S. Li, Y. Cao, and H. Zha, "Sc-wls: Towards interpretable feed-forward camera re-localization," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2022.

[20] M. Ding, Z. Wang, J. Sun, J. Shi, and P. Luo, "Camnet: Coarse-to-fine retrieval for camera re-localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2871–2880.

[21] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, "Camera relocalization by computing pairwise relative poses using convolutional neural network," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, 2017.

[22] M. O. Turkoglu, E. Brachmann, K. Schindler, G. J. Brostow, and A. Monszpart, "Visual camera re-localization using graph neural networks and relative pose supervision," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 145–155.