NEURAL INFORMATION PROCESSING SYSTEMS

SUNG KYUN KWAN UNIVERSITY(SKKU)

NeurIPS 2024

Activating Self-Attention for Multi-Scene Absolute Pose Regression

Miso Lee, Jihwan Kim, Jae-Pil Heo

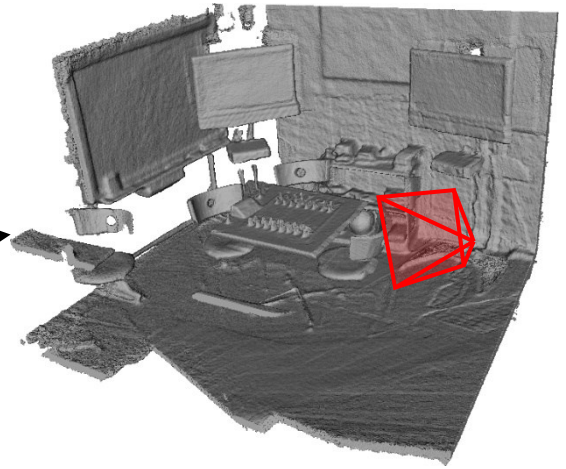
Visual Computing Lab

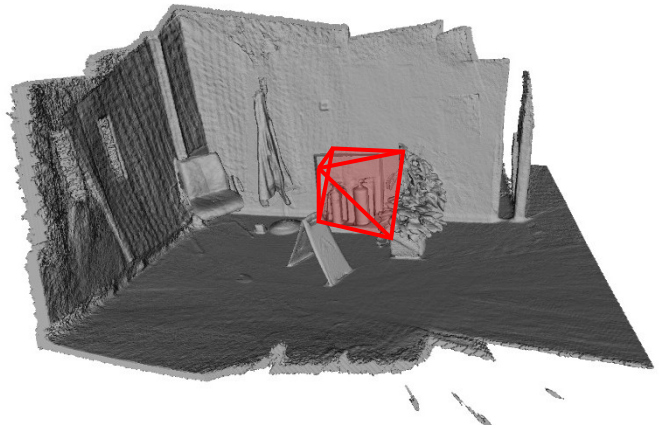
Sungkyunkwan University

Multi-Scene Absolute Pose Regression

input
RGB image

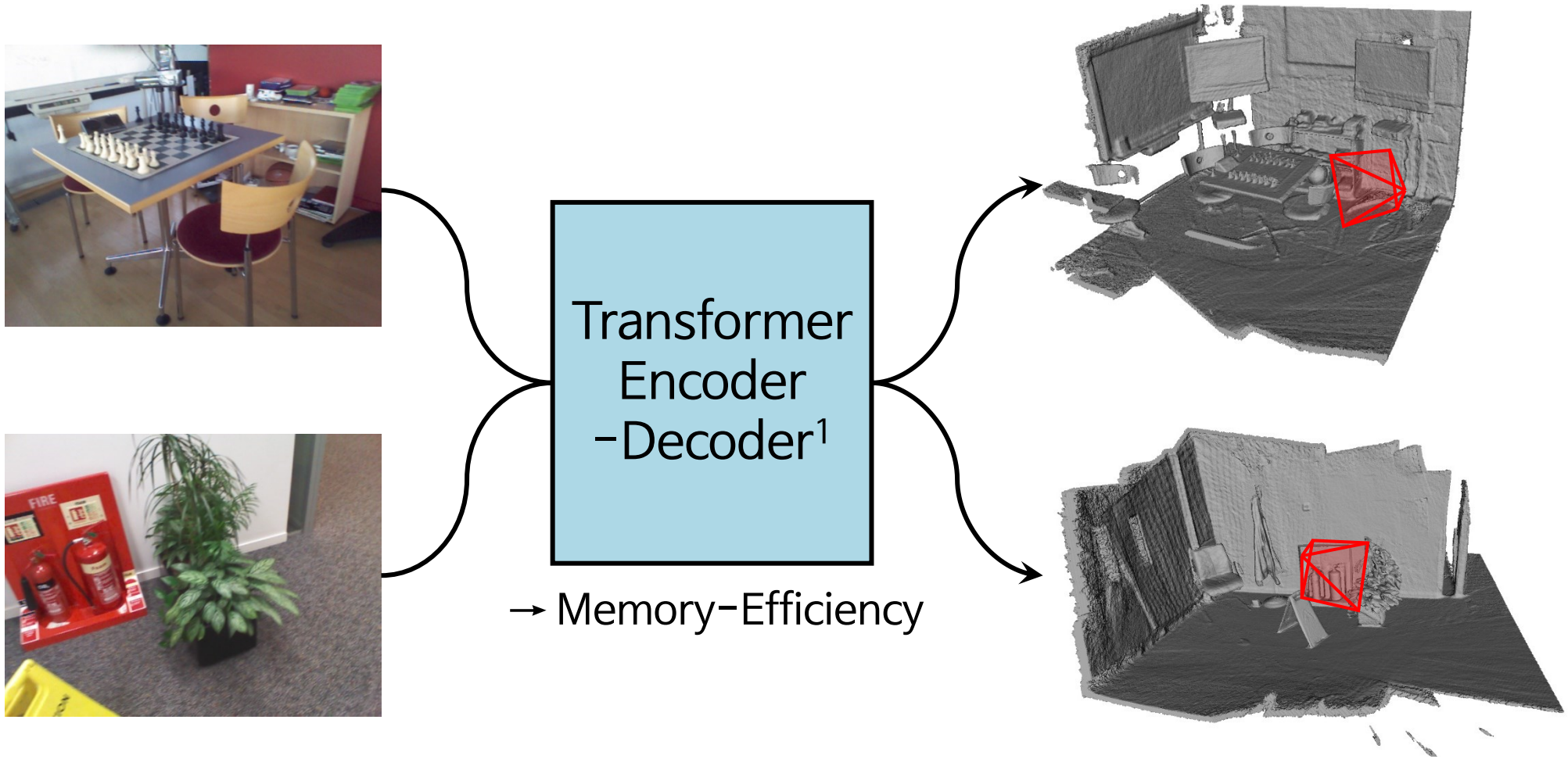
End-to-End
Model





output
camera pose [R|t]

Multi-Scene Transformer



Q. Do all components of the model effectively utilize memory?

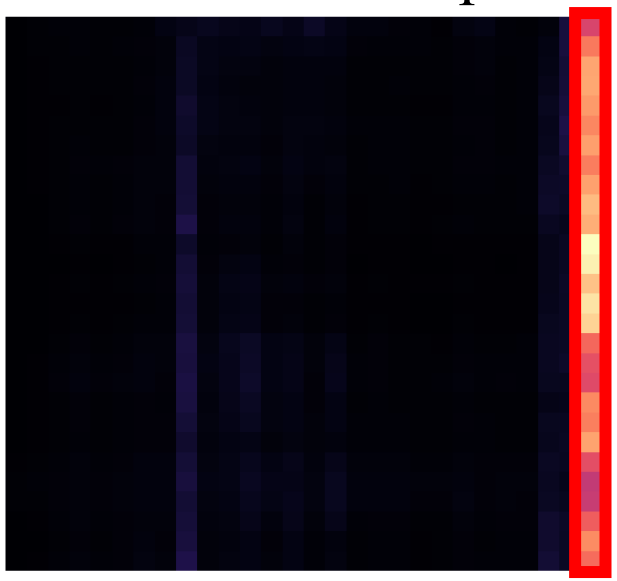
1) Learning Multi-Scene Absolute Pose Regression with Transformers. Shavit, et al. ICCV 2021.

Problem Definition

✓ Collapse of the self-attention

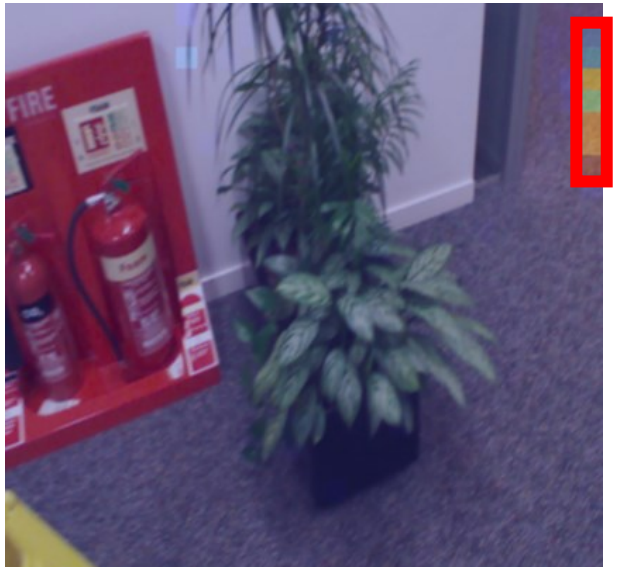
- The transformer encoder was deactivated due to attention collapse.
- This not only produces useless representations and waste of capacity but also causes significant training difficulties such as gradient vanishing and training instability[2].
- General solutions did not correct the problem in this task.

Attention Map



All queries attend to few keys.

Attended Keys

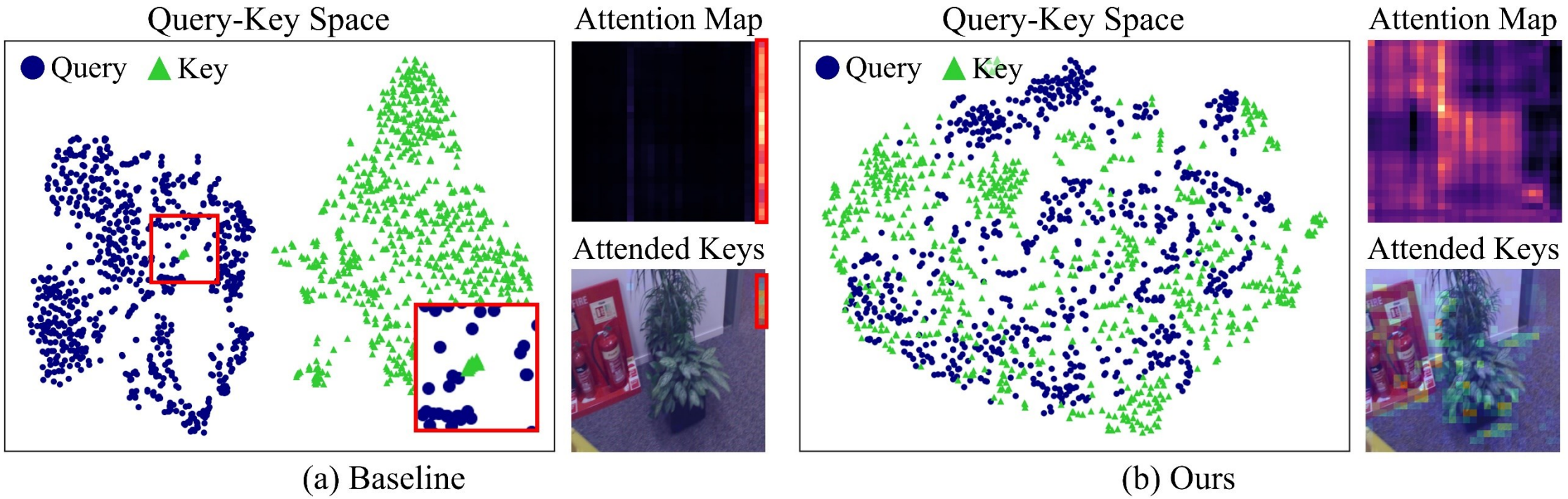


2) Attention is not all you need: Pure attention loses rank doubly exponentially with depth. Dong, et al. ICML 2021.

Main Contribution

✓ Why self-attention does not work?

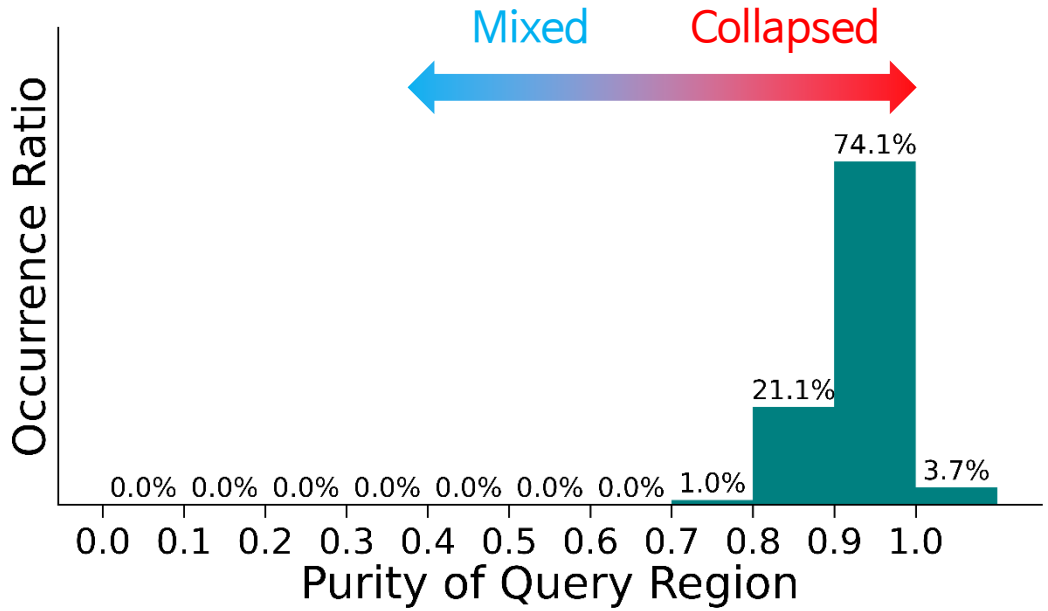
- Analysis 1: Distortion of Query-Key Embedding Space
- Analysis 2: Undertrained Positional Embedding
- Proposed Solution: Query-Key Alignment Loss & Fixed Positional Encoding

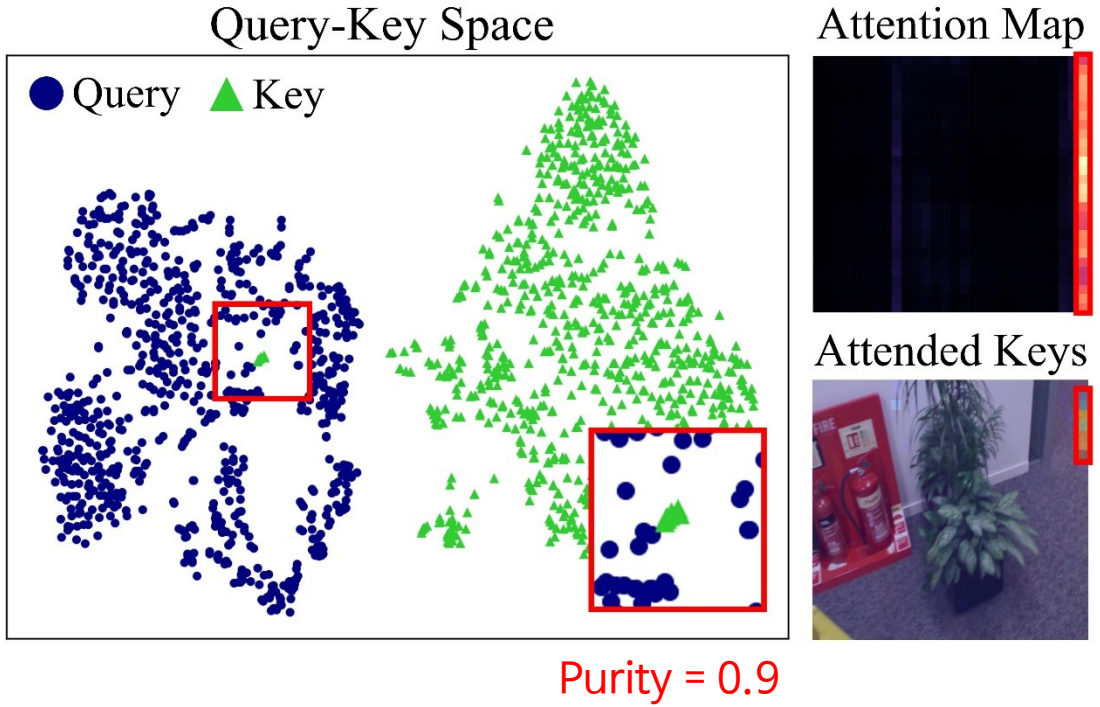


Distortion of Query-Key Space

- ✓ Queries and keys are separated while a small subset of keys are blended into the query region.
  - The input of query and key is the same in self-attention, but the projection results are separated except few keys.
  - This leads to the attention collapse as all queries are considered similar to those few keys.
  - We statistically demonstrate that this phenomenon is predominant throughout the whole dataset.

✓ Purity $\mathcal{P} = \frac{1}{|\hat{\boldsymbol{Q}}|}|\hat{\boldsymbol{Q}} \cap \boldsymbol{Q}|$





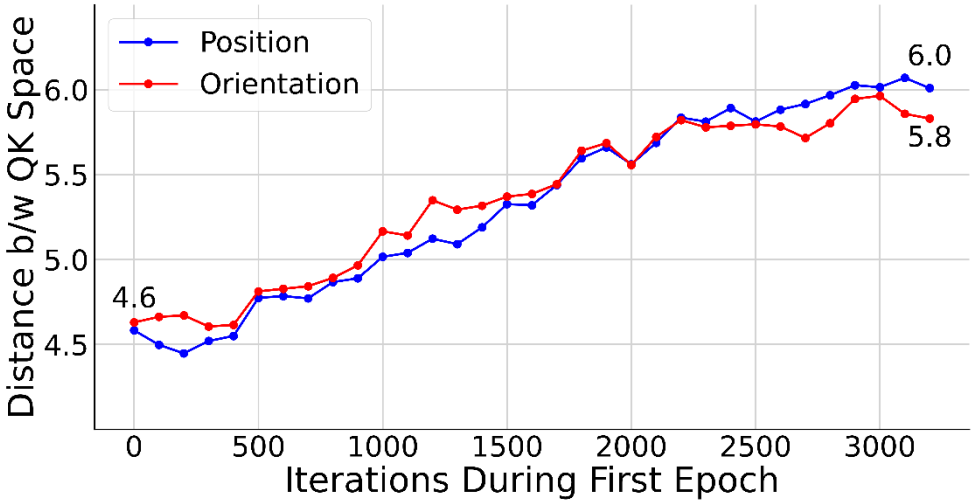
Deactivated Self-Attention

✓ The model fails to capture useful self-relation for the task; just skip.

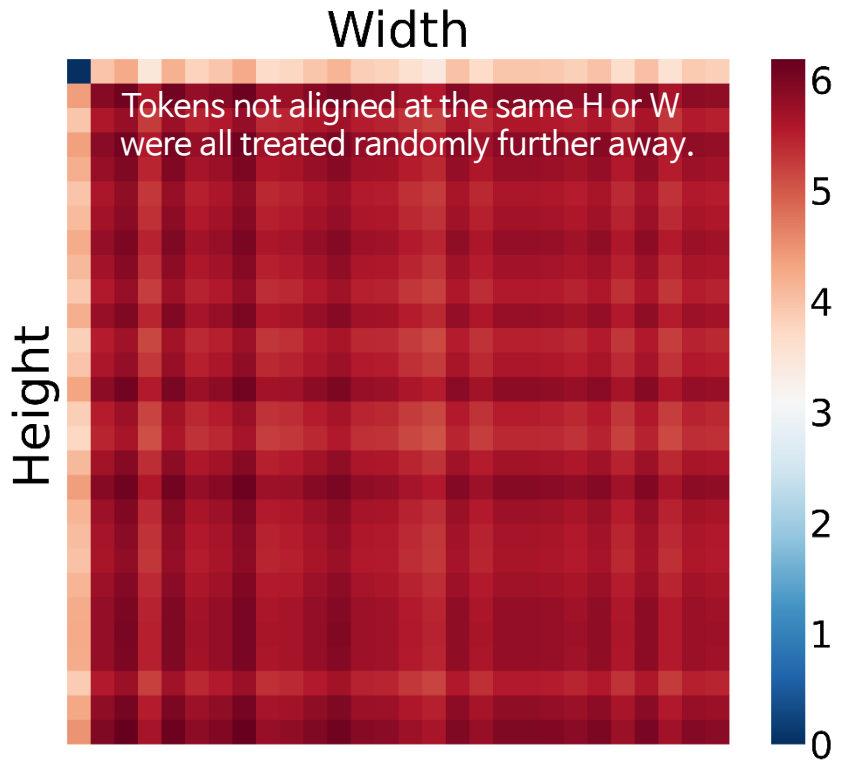
- Ablation Study

	Outdoor [8]	Indoor [27]
MST [20]	1.28m/2.73°	0.18m/7.28°
MST w/o encoder SA	1.21m/2.84°	0.18m/7.49°

- Increase Tendency of Distance b/w Queries & Keys



- Undertrained Positional Encoding

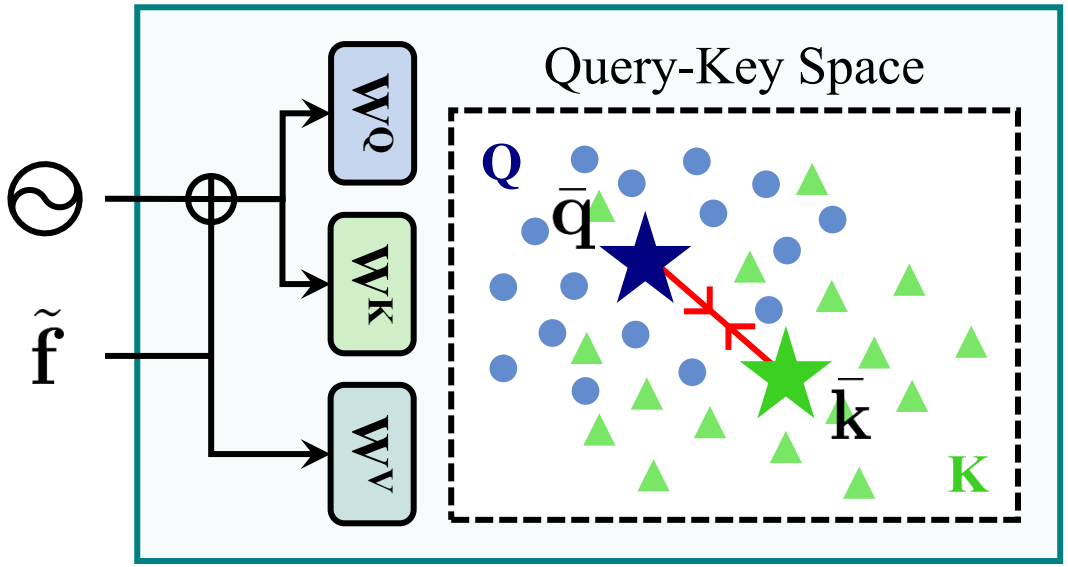


L2 Distance b/w Top-Left Token and Others

Proposed Solution

✓ Query-Key Alignment Loss

- Make sure that queries and keys do not become alienated from each other
- Encourage their interaction by embedding in close proximity



$$\rightarrow \mathcal{L}_{\mathrm{QKA}} = \frac{1}{L}\frac{1}{H}\sum_{l=1}^{L}\sum_{h=1}^{H} \|\bar{\boldsymbol{q}}_h^l - \bar{\boldsymbol{k}}_h^l\|_2$$

✓ Fixed Sinusoidal Positional Encoding

- Guide the model to stably learn self-relationships from the beginning of the training

Experimental Result

✓ Comparative Analysis

Table 2: Comparative analysis of MS-APRs on the Cambridge Landmarks dataset (outdoor localization) [8]. We report the median position/orientation errors in meters/degrees.

Method	King's College	Old Hospital	Shop Facade	St. Mary	Average
MSPN [19]	1.73/3.65	2.55/4.05	2.92/7.49	2.67/6.18	2.47/5.34
MST [20]	**0.83**/1.47	1.81/2.39	0.86/3.07	1.62/3.99	1.28/2.73
+Ours	0.88/**1.29**	**1.55/1.87**	**0.79/2.51**	**1.57/3.50**	**1.19/2.29**

Table 4: Comparative analysis of the baseline and ours. We report the localization recall at several thresholds on Cambridge Landmarks and 7Scenes datasets.

Method	Cambridge Landmarks [8]				7Scenes [27]			
	(1m, 5°)	(1m, 10°)	(2m, 5°)	(2m, 10°)	(0.2m, 5°)	(0.2m, 10°)	(0.3m, 5°)	(0.3m, 10°)
MST [20]	32.6	35.8	60.4	68.2	28.8	50.2	34.4	63.5
+Ours	**35.8**	**38.5**	**65.5**	**72.7**	**32.6**	**52.6**	**39.5**	**67.1**

Experimental Result

✓ Comparative Analysis

- We conjecture that the task difficulty contributes to the issue; the model should extrapolate the camera pose from a single RGB image across multiple scenes. It is required to regularize the model in a more direct manner to utilize self-attention for the task.
- Even advanced PE methods are not suitable to the task as they mainly depend on learnable parameters and relative position. We assume that absolute, purified positional clues, i.e., not randomly initialized ones, are important to stabilize the training of the embedding space for Multi-Scene Absolute Pose Regression.

Table 5: Comparison with alternative methods for collapsed SA and positional encoding methods. We report the average of the median position/orientation errors on 7scenes dataset.

(a) Alternative Methods for Collapsed SA

Method	7Scenes [27]
Improved SN [23]	0.18m/7.04°
$1/\sqrt{L}$-scaling [24]	0.18m/6.87°
σReparam [22]	0.19m/6.81°
QK Alignment	**0.17m/6.64°**

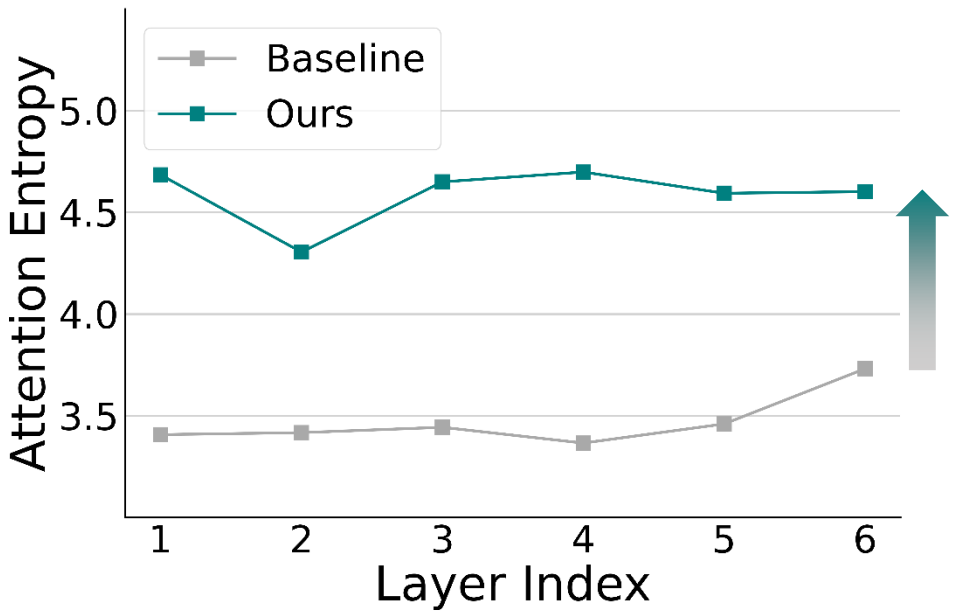
(b) Alternative Positional Encoding Methods

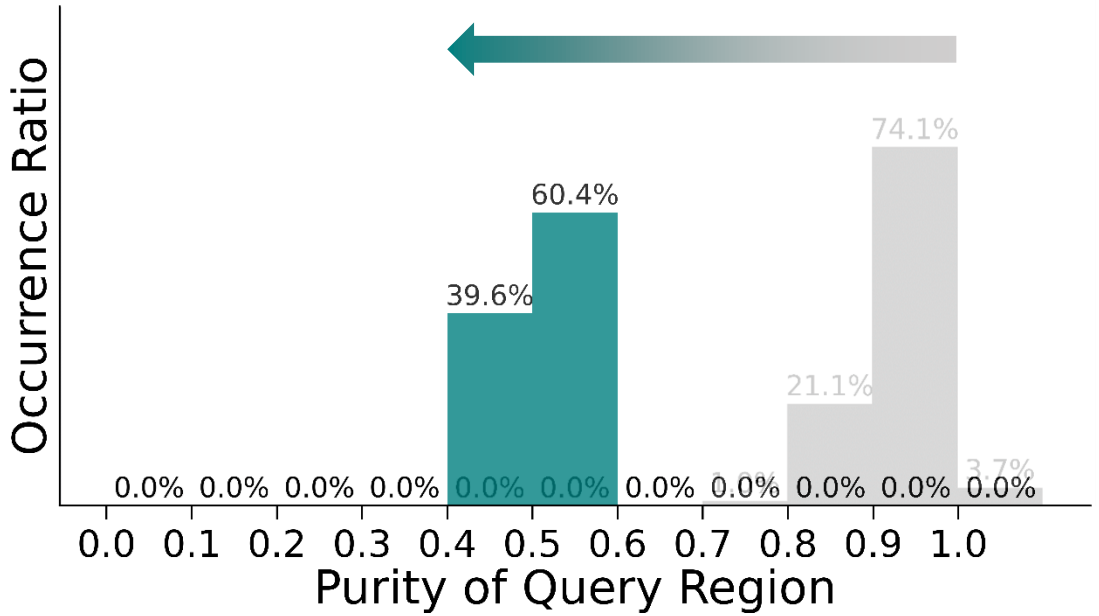
Method	7Scenes [27]
T5 PE [25]	0.18m/6.97°
Rotary PE [26]	0.18m/6.94°
Fixed PE	**0.17m/6.64°**

Experimental Result

✓ Quantitative Analysis

- Attention Entropy[3]: Higher entropy indicates greater training stability and higher representation capacity.
- Purity of Query Region
  - 1.0: Query region is composed only with queries.
  - ~1.0: Few keys are blended in the query region.
  - 0.5: Queries and keys are mixed together.

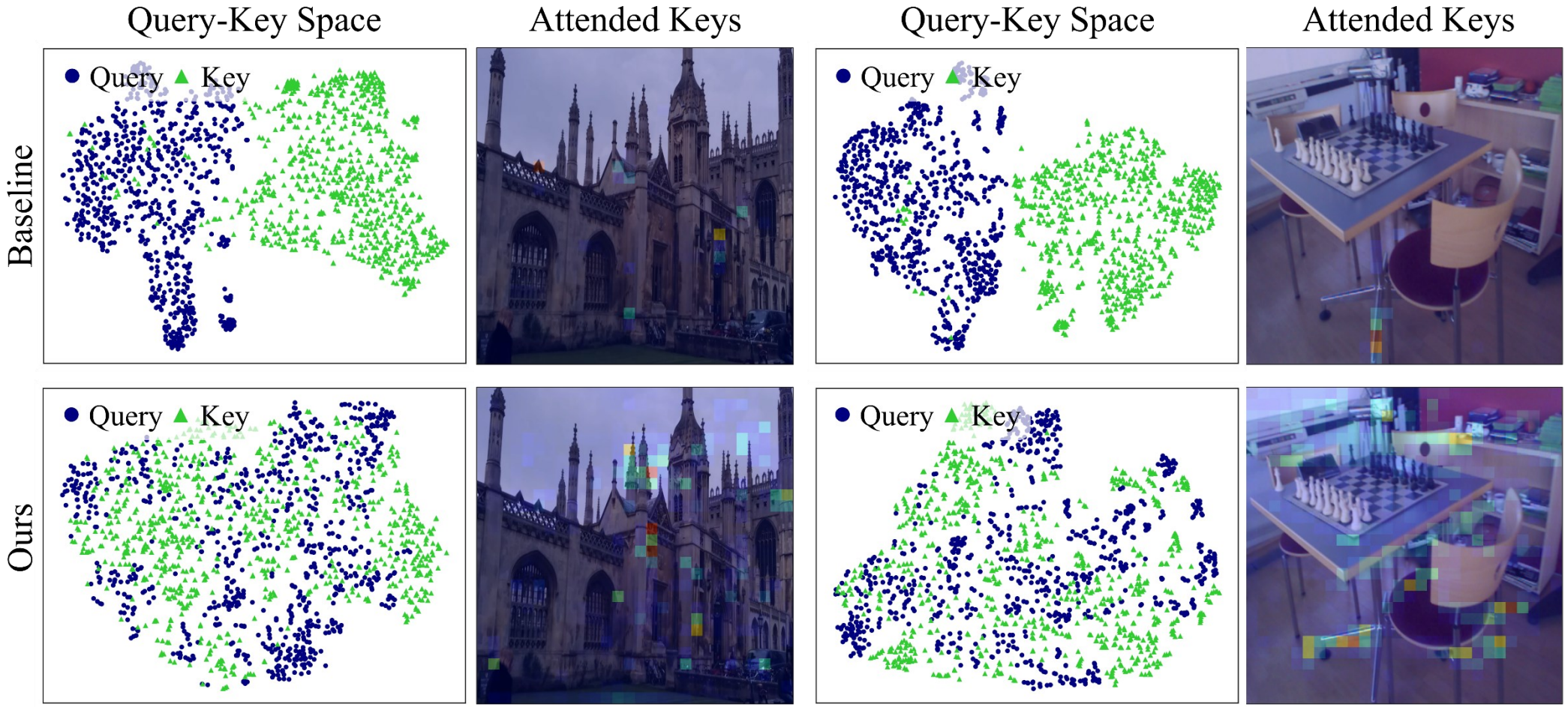




3) Stabilizing Transformer Training by Preventing Attention Entropy Collapse. Zhai, et al. ICML 2023.

Experimental Result

✓ Qualitative Analysis



Take Home Message

- ✓ Stronger Regularization Against Attention Collapse for Challenging Task
- ✓ Effectiveness of the Self-Attention Mechanism in Scene Understanding
- ✓ Importance of Appropriate Positional Clues in Self-Attention in Scene Understanding