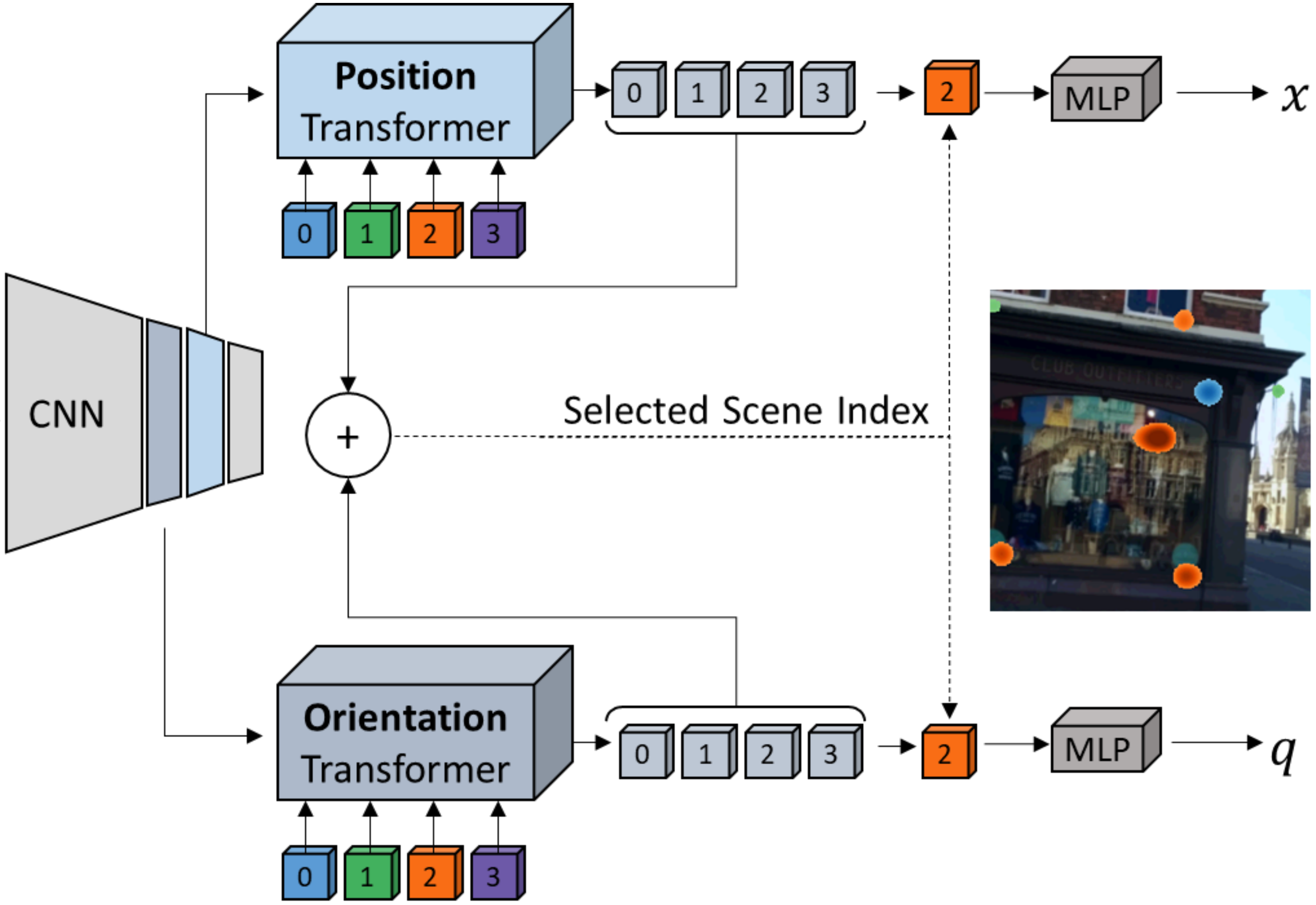
2021 ICCV OCTOBER 11-17 VIRTUAL

# Learning *Multi-Scene* Absolute Pose Regression with Transformers




Yoli Shavit Ron Ferens Yosi Keller

Bar-Ilan University
אוניברסיטת בר-אילן

# Camera Pose Regression

## Camera Pose Estimation


**Camera Pose**

$$p = (x, q)$$

$x \epsilon \mathbb{R}^3$ Position

$q \epsilon \mathbb{R}^4$ Orientation

**Input** **Output**

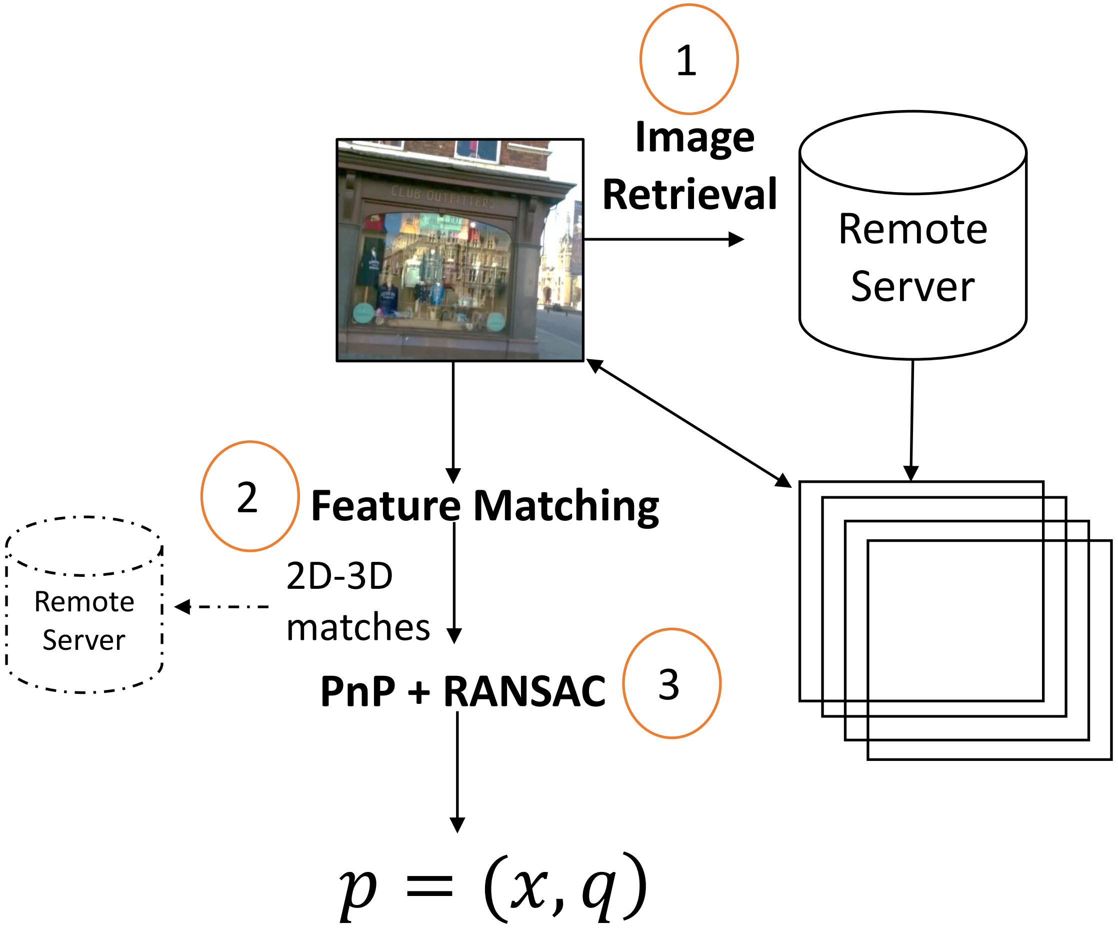
## Absolute Camera Pose Regression

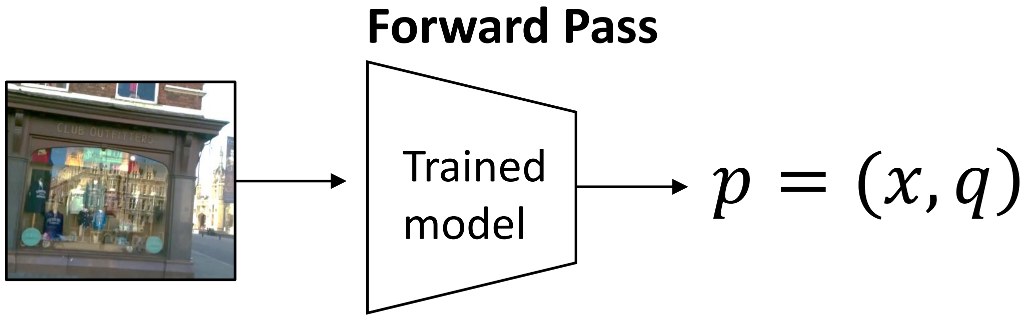
A learning-based method for solving the camera pose estimation problem

Trained model

$$p = (x, q)$$

# Camera Pose Estimation at Inference Time

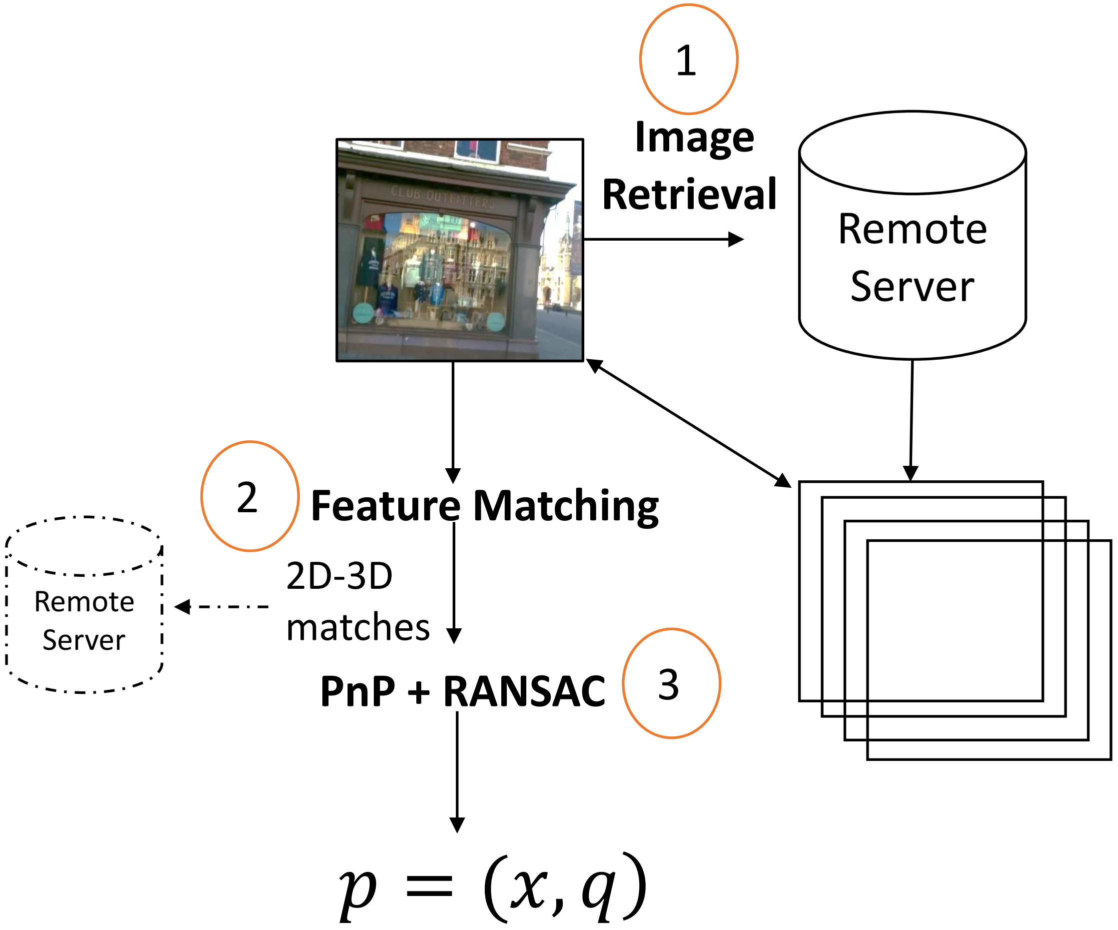



**Localization Pipelines**

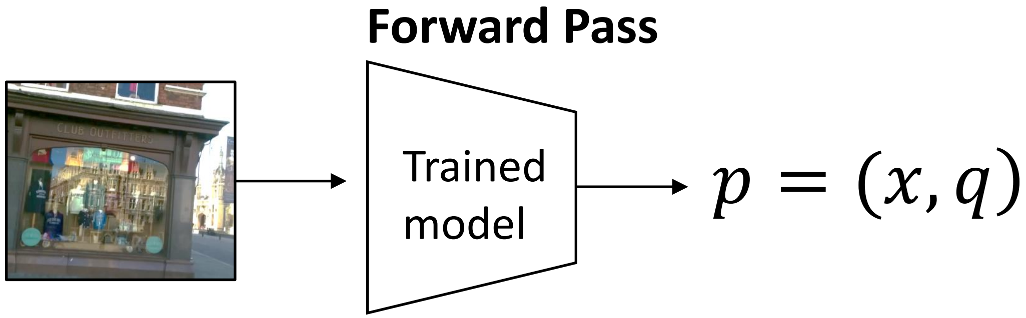
**Absolute Pose Regressors (APRs)**

# Camera Pose Estimation at Inference Time




Localization Pipelines

✓Fast (order of magnitude)
✓Light-weight
✓Standalone

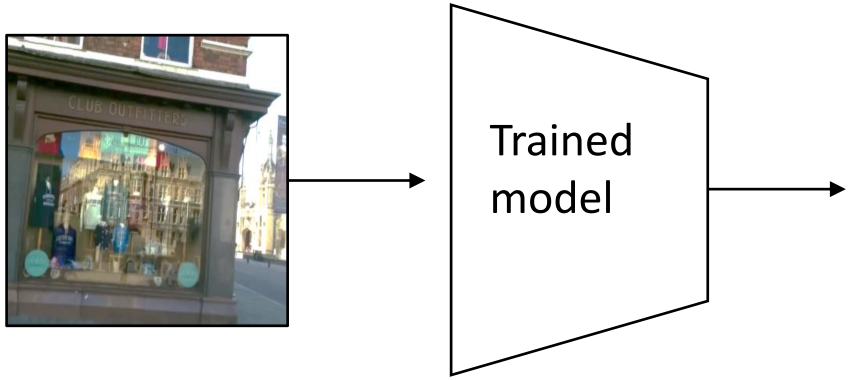



Absolute Pose Regressors (APRs)

# The Cons of *Single*-Scene APRs

- ✓ Fast
- ✓ Light-weight
- ✓ Standalone

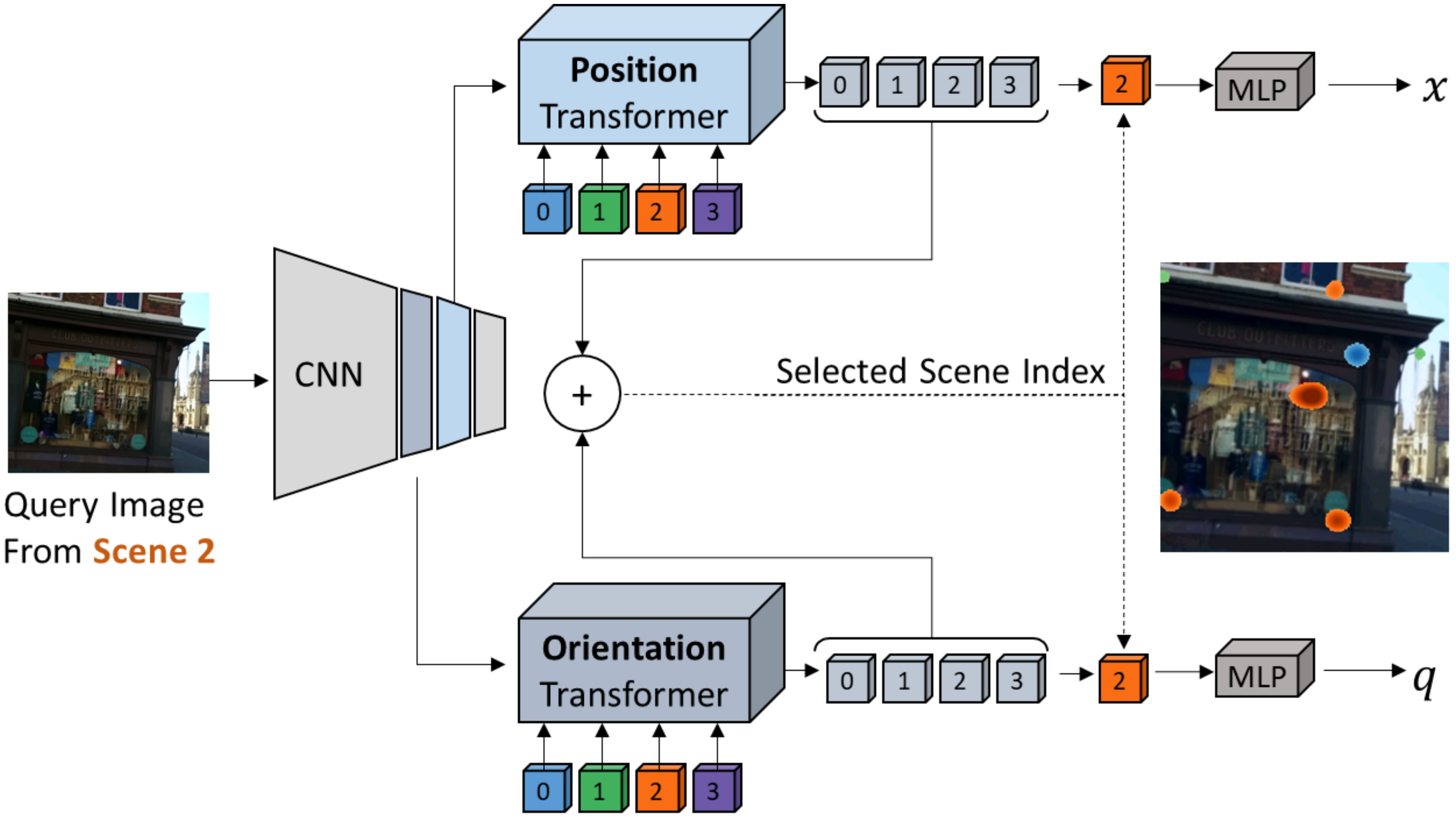
- x Less accurate
- x Trained **per scene**



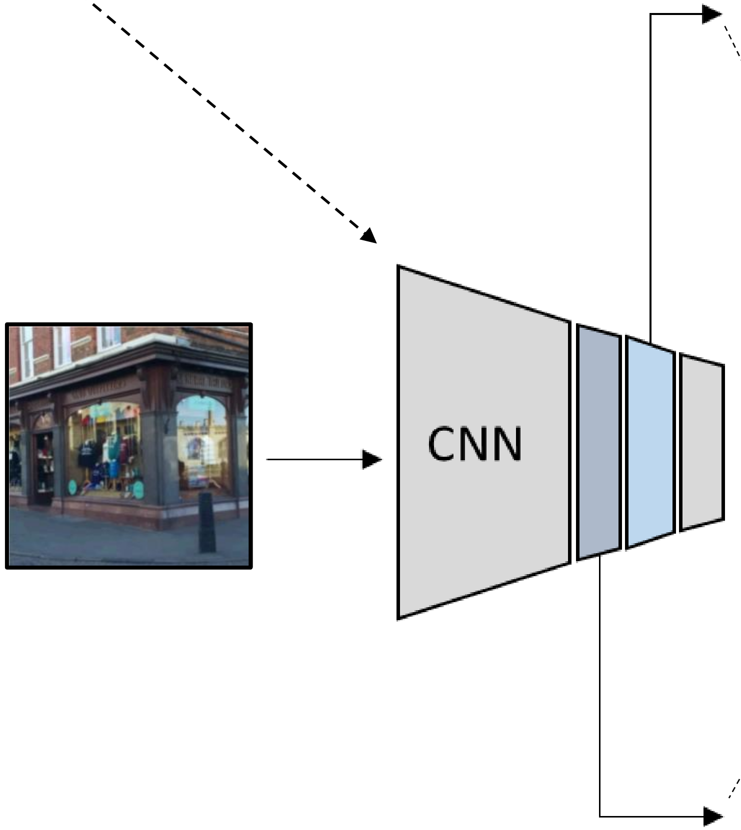

$$p = (x, q)$$

For localizing images from N scenes we need to train, deploy and choose from N models

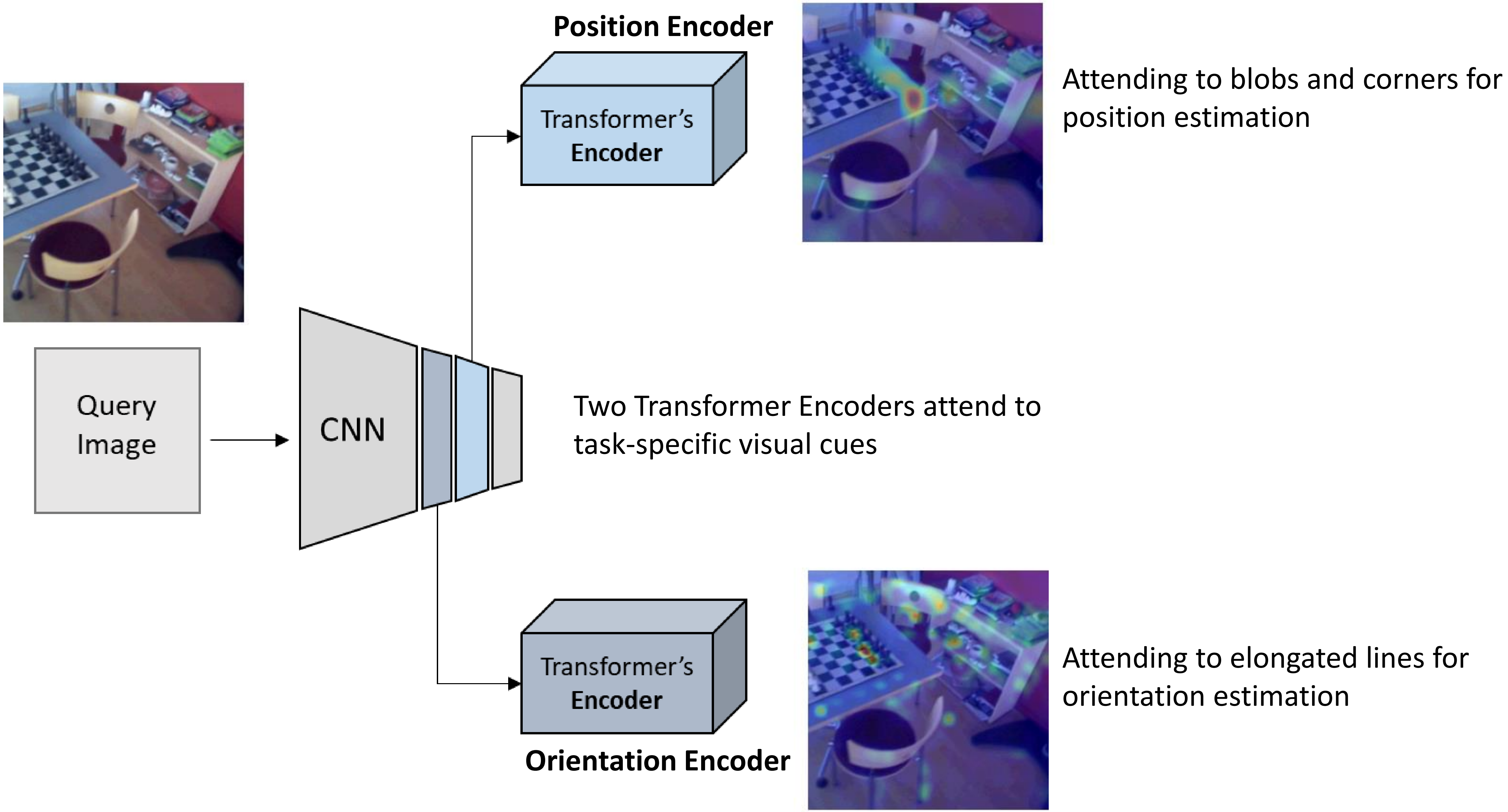
# Learning Multi-Scene Pose Regression with Transformers

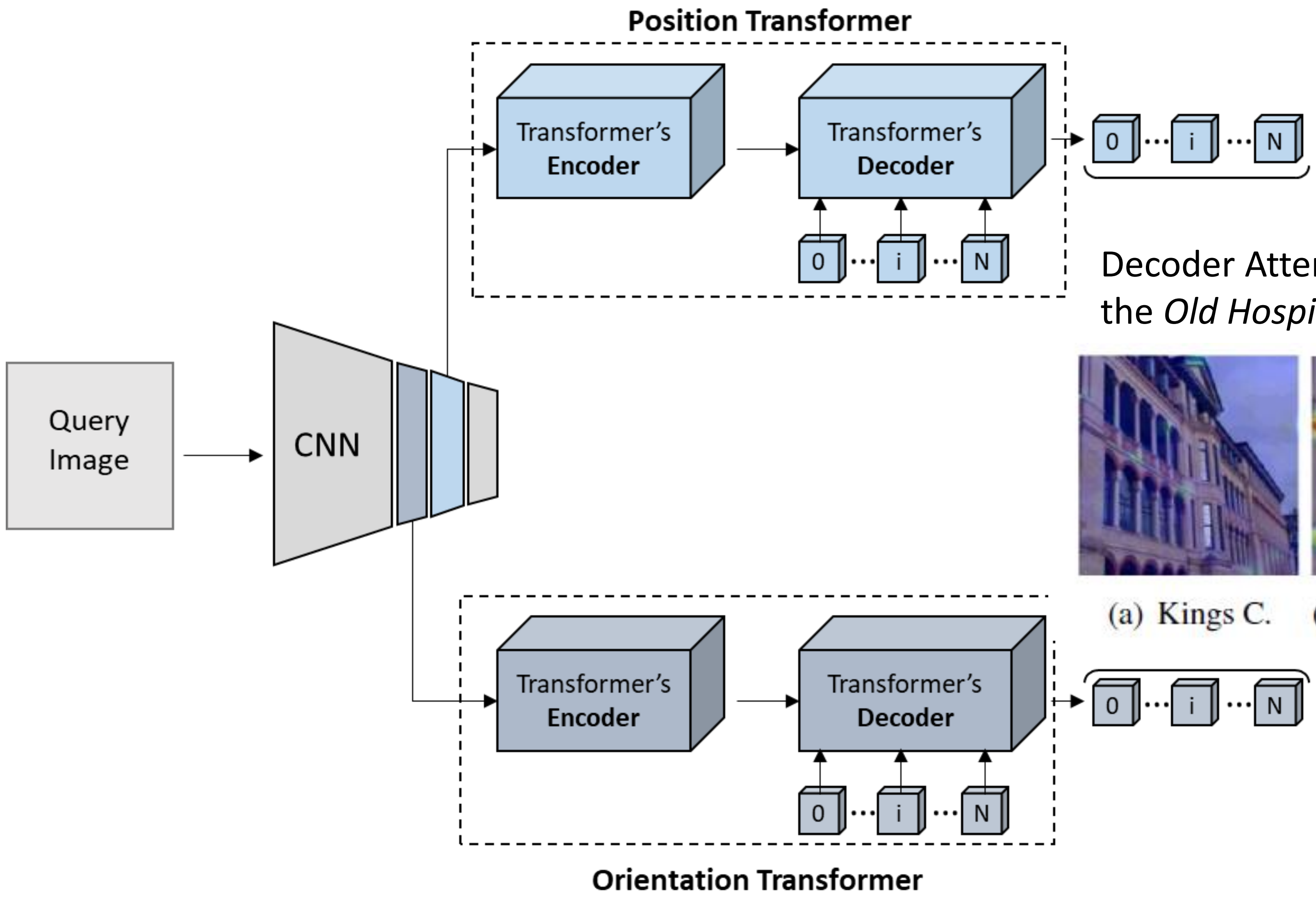


We extract visual features using a convolutional backbone and then encode, project and flatten activation maps



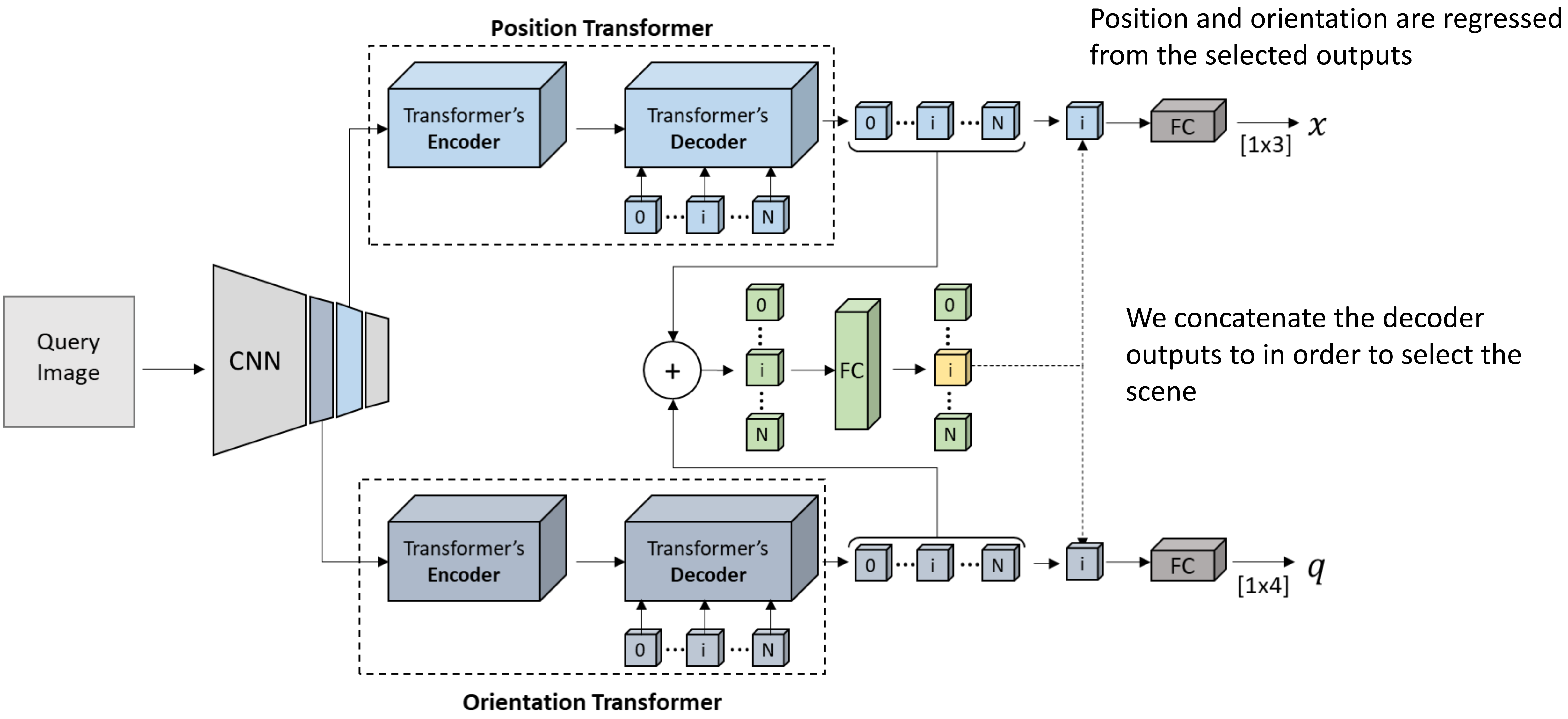

Since position and orientation are related to different visual cues, we extract activation maps at different resolutions

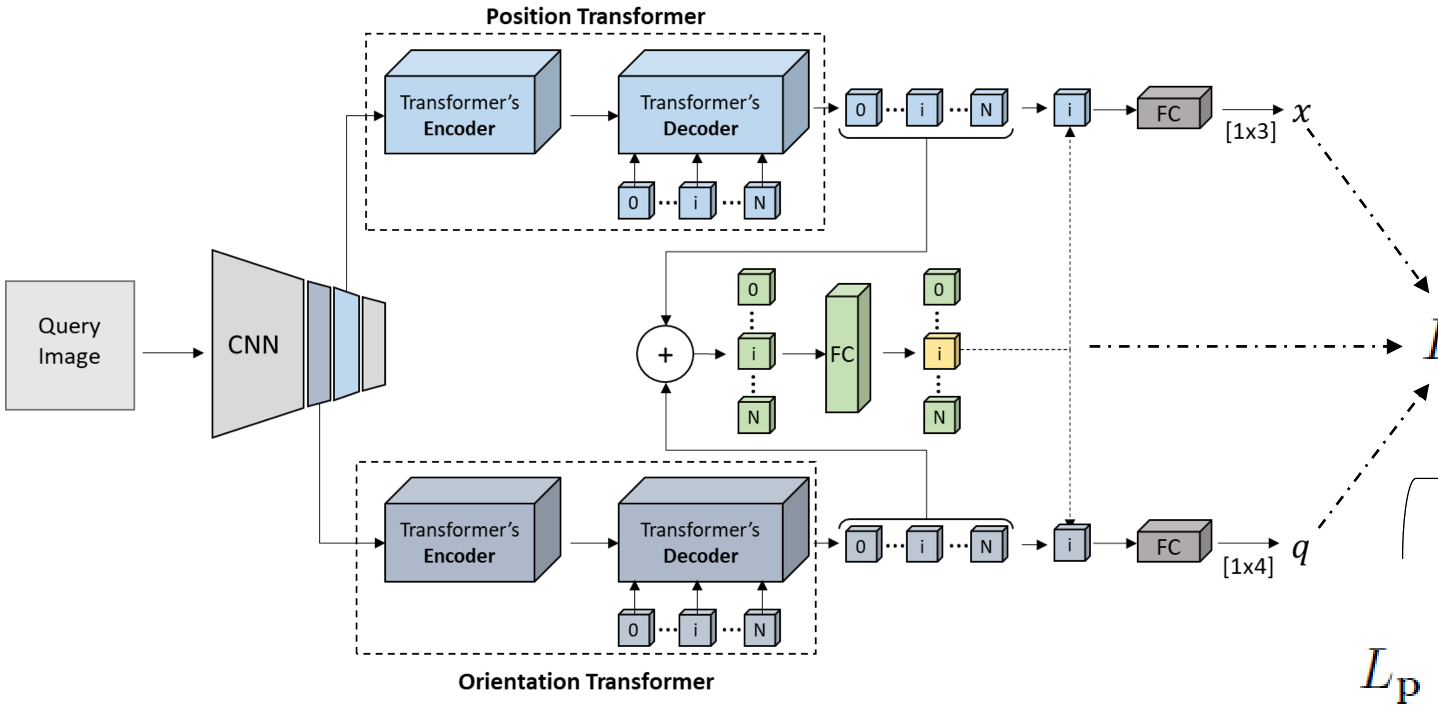

Position Encoder
Transformer's Encoder
Attending to blobs and corners for position estimation
Query Image
CNN
Two Transformer Encoders attend to task-specific visual cues
Transformer's Encoder
Orientation Encoder
Attending to elongated lines for orientation estimation




Decoder Attention maps when processing an image from the *Old Hospital* scene

(a) Kings C. (b) Old Hospital (c) Shop Facade (d) St. Mary


Position Transformer
Position and orientation are regressed from the selected outputs
Query Image
CNN
Transformer's Encoder
Transformer's Decoder
0
i
N
FC
x
[1x3]
+
We concatenate the decoder outputs to in order to select the scene
q
[1x4]
Orientation Transformer




$$L_{\mathbf{multi-scene}} = \underbrace{L_{\mathbf{p}}}_{\text{Learned Pose Loss (Kendall et al., 2017)}} + NLL(\mathbf{s}, \mathbf{s_0})$$

$$L_{\mathbf{p}} = \underbrace{L_{\mathbf{x}}}_{\text{Position Loss}} \exp(-s_{\mathbf{x}}) + s_{\mathbf{x}} + \underbrace{L_{\mathbf{q}}}_{\text{Orientation Loss}} \exp(-s_{\mathbf{q}}) + s_{\mathbf{q}}$$

$$L_{\mathbf{x}} = ||\mathbf{x}_0 - \mathbf{x}||_2$$

Position Loss

$$L_{\mathbf{q}} = |\mathbf{q}_0 - \frac{\mathbf{q}}{||\mathbf{q}||}||_2$$

Orientation Loss

$\hat{S}_x$ and $\hat{S}_q$ are learned parameters representing task uncertainty

# Comparison with MSPN

Median position and orientation errors of our method and
a recent multi-scene approach (MSPN) for the **CambridgeLandmarks** (top) and
**7Scenes** (bottom) datasets

| Method | K. College | Old Hospital | Shop Facade | St. Mary |
|---|---|---|---|---|
| MSPN [3] | 1.73/3.65 | 2.55/4.05 | 2.92/7.49 | 2.67/6.18 |
| MS-Transformer (ours) | **0.83/1.47** | **1.81/2.39** | **0.86/3.07** | **1.62/3.99** |

| Method | Chess | Fire | Heads | Office | Pumpkin | Kitchen | Stairs |
|---|---|---|---|---|---|---|---|
| MSPN[3] | **0.09**/4.76 | 0.29/10.5 | 0.16/13.1 | **0.16**/6.8 | 0.19/5.5 | **0.21**/6.61 | 0.31/11.63 |
| MS-Transformer (ours) | 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** |

# Comparison with Single-Scene APRs

Mean of median position and orientation errors and the respective ranks of our method, MSPN and state-of-the-art single-scene APRs on the CambridgeLandmarks (left) and 7Scenes (right) datasets.

| Method | Average [m/deg] | Ranks |
|---|---|---|
| Single-scene APRs | | |
| PoseNet [17] | 2.09/6.84 | 10/11 |
| BayesianPN [15] | 1.92/6.28 | 8/10 |
| LSTM-PN [35] | 1.30/5.52 | 2/9 |
| SVS-Pose [21] | 1.33/5.17 | 3/7 |
| GPoseNet [8] | 2.08/4.59 | 6/3 |
| PoseNet-Learnable [16] | 1.43/2.85 | 5/2 |
| GeoPoseNet [16] | 1.63/2.86 | 6/3 |
| MapNet [7] | 1.63/3.64 | 6/5 |
| IRPNet [29] | 1.42/3.45 | 4/4 |
| Multi-scene APRs | | |
| MSPN [3] | 2.47/5.34 | 11/8 |
| **MS-Transformer (Ours)** | **1.28/2.73** | **1/1** |

| Method | Average [m/deg] | Ranks |
|---|---|---|
| Single-scene APRs | | |
| PoseNet [17] | 0.44/10.4 | 10/11 |
| BayesianPN [15] | 0.47/9.81 | 11/8 |
| LSTM-PN [35] | 0.31/9.86 | 8/9 |
| GPoseNet [8] | 0.31/9.95 | 8/8 |
| PoseNet-Learnable [16] | 0.24/7.87 | 7/4 |
| GeoPoseNet [16] | 0.23/8.12 | 5/5 |
| MapNet [7] | 0.21/7.78 | 4/3 |
| IRPNet [29] | 0.23/8.49 | 5/7 |
| AttLoc [36] | 0.20/7.56 | 2/2 |
| Multi-scene APRs | | |
| MSPN [3] | 0.20/8.41 | 2/6 |
| **MS-Transformer (Ours)** | **0.18/ 7.28** | **1/1** |

# From Multi-Scene to Multi-Dataset

| APR Method | CambridgeLand. [m/deg] | 7Scenes [m/deg] |
|---|---|---|
| Single-scene [16] | 1.43/2.85 | 0.24/7.87 |
| Multi-scene (Ours) | 1.28/2.73 | 0.18/7.28 |
| Multi-dataset (Ours) | 1.50/ 2.57 | 0.22/6.78 |

Our method is able to learn multiple scenes from datasets with different scales and properties

# Robustness and Scalability

| Encoder/Decoder # Layers | Position [meters] | Orientation [degrees] |
|---|---|---|
| 2 | 0.19 | 7.48 |
| 4 | 0.18 | 6.94 |
| 6 | 0.18 | 7.28 |
| 8 | 0.18 | 6.92 |

| Transformer Dimension | Position [meters] | Orientation [degrees] |
|---|---|---|
| 64 | 0.18 | 8.06 |
| 128 | 0.19 | 7.56 |
| 256 | 0.18 | 7.28 |
| 512 | 0.18 | 7.19 |

| Backbone | Position [meters] | Orientation [degrees] |
|---|---|---|
| Resnet50 | 0.19 | 8.6 |
| EfficientNetB0 | 0.18 | 7.28 |
| EfficientNetB1 | 0.17 | 7.26 |

Our method maintains state-of-the-art performance across architectural choices

| Num. Scenes | Runtime [ms] | | Memory [Mb] | |
|---|---|---|---|---|
| Num. Layers | 2 | 6 | 2 | 6 |
| 1 | 18.8 | 34.6 | 40.8 | 74.6 |
| 4 | 18.8 | 35 | 40.8 | 74.6 |
| 7 | 19.2 | 35.2 | 40.8 | 74.6 |
| 10 | 19.2 | 35.2 | 40.8 | 74.6 |
| 100 | 19.6 | 35.4 | 41.0 | 74.8 |
| 500 | 21.0 | 41.0 | 41.8 | 75.6 |
| 1000 | 27.0 | 58.6 | 42.8 | 76.7 |

The memory footprint for a 1000 scenes with a single scene APR approach is **~5000Mb**

# Conclusion

**We propose a novel transformer-based approach for multi-scene absolute pose regression**

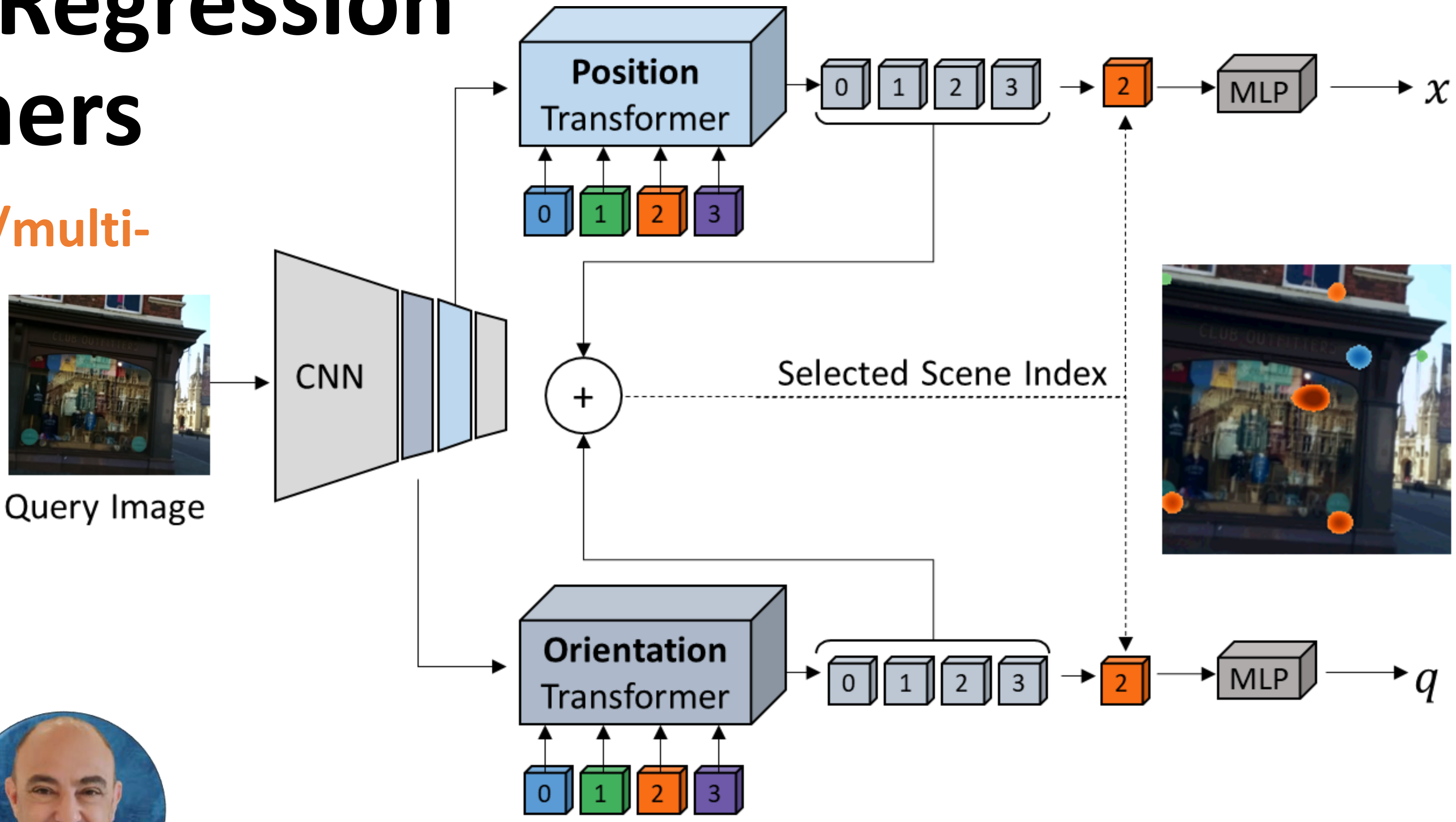
- Two Transformer Encoders separately attend to position- and orientation-informative image cues
- Two Transformer Decoders attend to scene-specific information

**Our approach is shown to provide a new state-of-the-art APR accuracy**

- Outperforming single and multi-scene APRs across indoor and outdoor benchmarks
- Demonstrating robustness to specific architecture choices

# Learning *Multi-Scene* Absolute Pose Regression with Transformers

https://github.com/yolish/multi-scene-pose-transformer





Yoli Shavit


Ron Ferens


Yosi Keller

Bar-Ilan University
אוניברסיטת בר-אילן