INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Improving Visual Relocalization by Discovering Anchor Points

Soham Saha, Girish Varma, C.V. Jawahar
Center for Visual Information Technology, Kohli Center for Intelligent Systems, IIIT Hyderabad

BMVC 2018
Newcastle upon Tyne

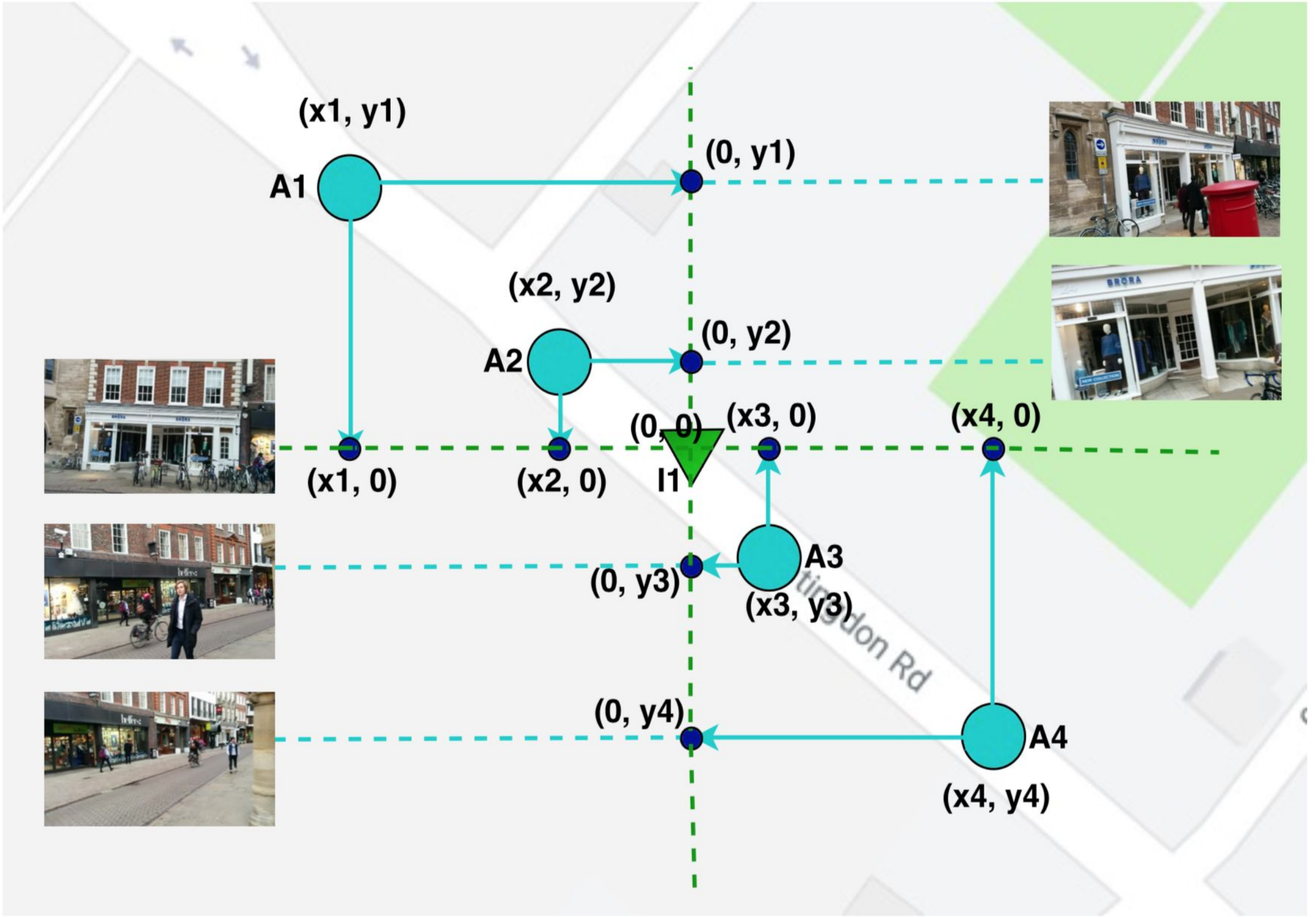
Flipkart

MOTIVATION

- The visual relocalization problem is an essential component of many practical systems like autonomous navigation, augmented reality, drone navigation.
- An independent source of location information is essential for safety and applicability in these areas.
- We propose to identify the location and camera pose of a scene from only its monocular image.

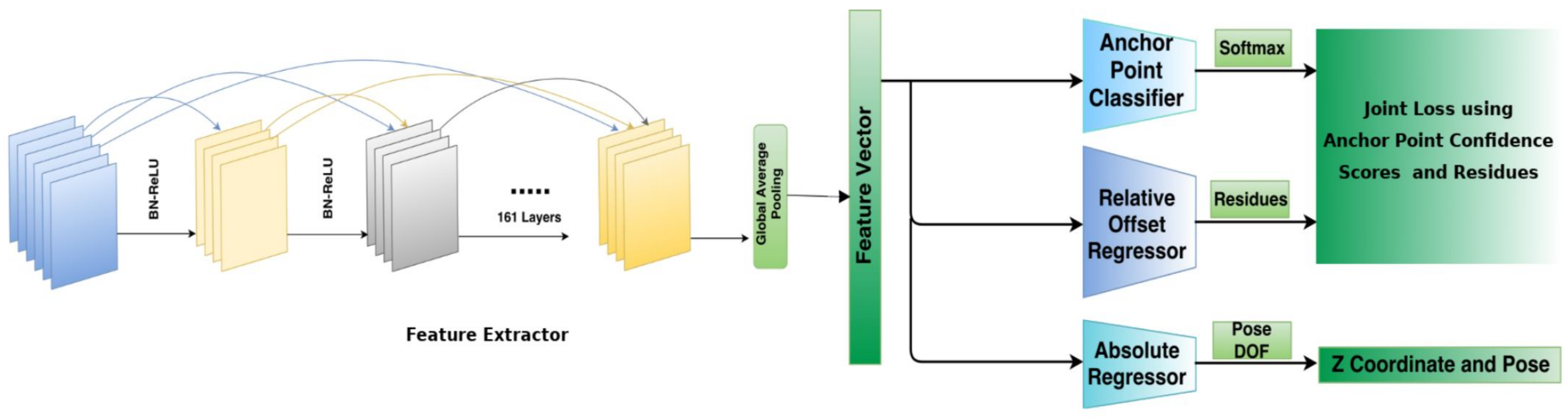
ANCHOR POINT BASED VISUAL RELOCALIZATION

- Recent deep neural network based approaches aim to directly regress the 6-DOF relative to the real world coordinates.
- Humans generally identify their location relative to other locations or landmarks. Taking inspiration from this, we propose an end to end trainable model.
- We define certain landmarks as anchor points and predict relative distances from them.



- The green triangle denotes the current scene for which the 6-DOF need to be predicted.
- The blue circles are the predefined anchor points.
- We formulate the solution as predicting the relative offsets for the current scene from each of the anchor points.

NETWORK ARCHITECTURE



Our architecture of a feature extractor which branches out into 3 heads:

Anchor point Classifier: Outputs the probabilities of belonging to each anchor point.

Relative Offset Regressor: Predicts the X,Y offsets of the current scene from each of the anchor points. [Groundtruth calculated as a pre-processing step]

Absolute Regressor: Predicts the Z coordinates and the 3 remaining coordinates for camera pose.

DISCOVERING RELEVANT ANCHOR POINTS

We propose a confidence based loss function which automatically discovers the most relevant anchor point which must be considered in order to minimize the error in prediction.

$$\underbrace{\alpha_1 H(C_i,\hat{C}_i)}_{\text{Cross entropy loss component}} + \alpha_2 \textstyle\sum_i \underbrace{\left[\left(X_i-\hat{X}_i\right)^2+\left(Y_i-\hat{Y}_i\right)^2\right] C_i}_{\text{X,Y offsets from a scene weighted by confidence.}} + \alpha_3 \textstyle\sum_i \underbrace{\left[\left(Z_i-\hat{Z}_i\right)\right]^2}_{\text{Z-coordinate Regressor}} + \underbrace{\left\|P_i-\frac{\hat{P}_i}{\|\hat{P}_i\|}\right\|^2}_{\text{Pose Regressor}}$$

- The first component of this loss function calculates the predicted probabilities of the anchor points from and nearest anchor point.
- The second component of the loss regresses the <X,Y> coordinates of the scene relative to all the anchor points and weighs them with the predicted probabilities for those anchor points.
- We treat these predicted probabilities as **confidence scores.**
- The 3rd and the 4th components are the regressors for the Z-coordinate and the pose.

RESULTS

Quantitative Results

Scene	Area or Volume	Active Search (SIFT) [20]	Posenet Spatial LSTM[28]	Posenet sigma^2 weight [11]	Posenet Geom. Rep. [11]	Ours (DenseNet) (cross entropy)	Ours (DenseNet) (w/o cross entropy)	**Ours (GoogleNet) (w/o cross entropy)**
Great Court	$8000m^2$	-	-	$7.00m, 3.65^\circ$	$6.83m, 3.47^\circ$	$5.85m, 3.61^\circ$	$4.64m, 3.42^\circ$	**$5.89m, 3.53^\circ$**
King's College	$5600m^2$	$0.42m, 0.55^\circ$	$0.99m, 3.65^\circ$	$0.99m, 1.06^\circ$	$0.88m, 1.04^\circ$	$0.55m, 0.97^\circ$	$0.57m, 0.88^\circ$	**$0.79m, 0.95^\circ$**
Old Hospital	$2000m^2$	$0.44m, 1.01^\circ$	$1.51m, 4.29^\circ$	$2.17m, 2.94^\circ$	$3.20m, 3.29^\circ$	$1.45m, 3.16^\circ$	$1.21m, 2.55^\circ$	**$2.11m, 3.05^\circ$**
Shop Facade	$875m^2$	$0.12m, 0.40^\circ$	$1.18m, 7.44^\circ$	$1.05m\ 3.97^\circ$	$0.88m, 3.78^\circ$	$0.49m, 2.42^\circ$	$0.52m, 2.27^\circ$	**$0.77m, 3.25^\circ$**
St. Mary's Church	$4800m^2$	$0.19m, 0.54^\circ$	$1.52m, 6.68^\circ$	$1.49m, 3.43^\circ$	$1.57m, 3.32^\circ$	$1.12m, 2.84^\circ$	$1.04m, 2.69^\circ$	**$1.22m, 3.02^\circ$**
Street	$50000m^2$	$0.85m, 0.83^\circ$	-	$20.7m, 25.7^\circ$	$20.3m, 25.5^\circ$	$8.19m, 25.5^\circ$	$7.86m, 24.2^\circ$	**$11.8m, 24.3^\circ$**
Chess	$6m^2$	$0.04m, 1.96^\circ$	$0.24m, 5.77^\circ$	$0.14m, 4.50^\circ$	$0.13m, 4.48^\circ$	$0.06m, 3.95^\circ$	$0.06m, 3.89^\circ$	**$0.08m, 4.12^\circ$**
Fire	$2.5m^2$	$0.03m, 1.53^\circ$	$0.34m, 11.9^\circ$	$0.27m, 11.8^\circ$	$0.27m, 11.3^\circ$	$0.16m, 10.4^\circ$	$0.15m, 10.3^\circ$	**$0.16m, 11.1^\circ$**
Head	$1m^2$	$0.02m, 1.45^\circ$	$0.21m, 13.7^\circ$	$0.18m, 12.1^\circ$	$0.17m, 13.0^\circ$	$0.08m, 10.7^\circ$	$0.08m, 10.9^\circ$	**$0.09m, 11.2^\circ$**
Office	$7.5m^2$	$0.09m, 3.61^\circ$	$0.30m, 8.08^\circ$	$0.20m, 5.77^\circ$	$0.19m, 5.55^\circ$	$0.11m, 5.24^\circ$	$0.09m, 5.15^\circ$	**$0.11m, 5.38^\circ$**
Pumpkin	$5m^2$	$0.08m, 3.10^\circ$	$0.33m, 7.00^\circ$	$0.25m, 4.82^\circ$	$0.26m, 4.75^\circ$	$0.11m, 3.18^\circ$	$0.10m, 2.97^\circ$	**$0.14m, 3.55^\circ$**
Red Kitchen	$18m^2$	$0.07m, 3.37^\circ$	$0.37m, 8.83^\circ$	$0.24m, 5.52^\circ$	$0.23m, 5.35^\circ$	$0.08m, 4.83^\circ$	$0.08m, 4.68^\circ$	**$0.13m, 5.29^\circ$**
Stairs	$7.5m^2$	$0.03m, 2.22^\circ$	$0.40m, 13.7^\circ$	$0.37m, 10.6^\circ$	$0.35m, 12.4^\circ$	$0.13m, 10.1^\circ$	$0.10m, 9.26^\circ$	**$0.21m, 11.9^\circ$**

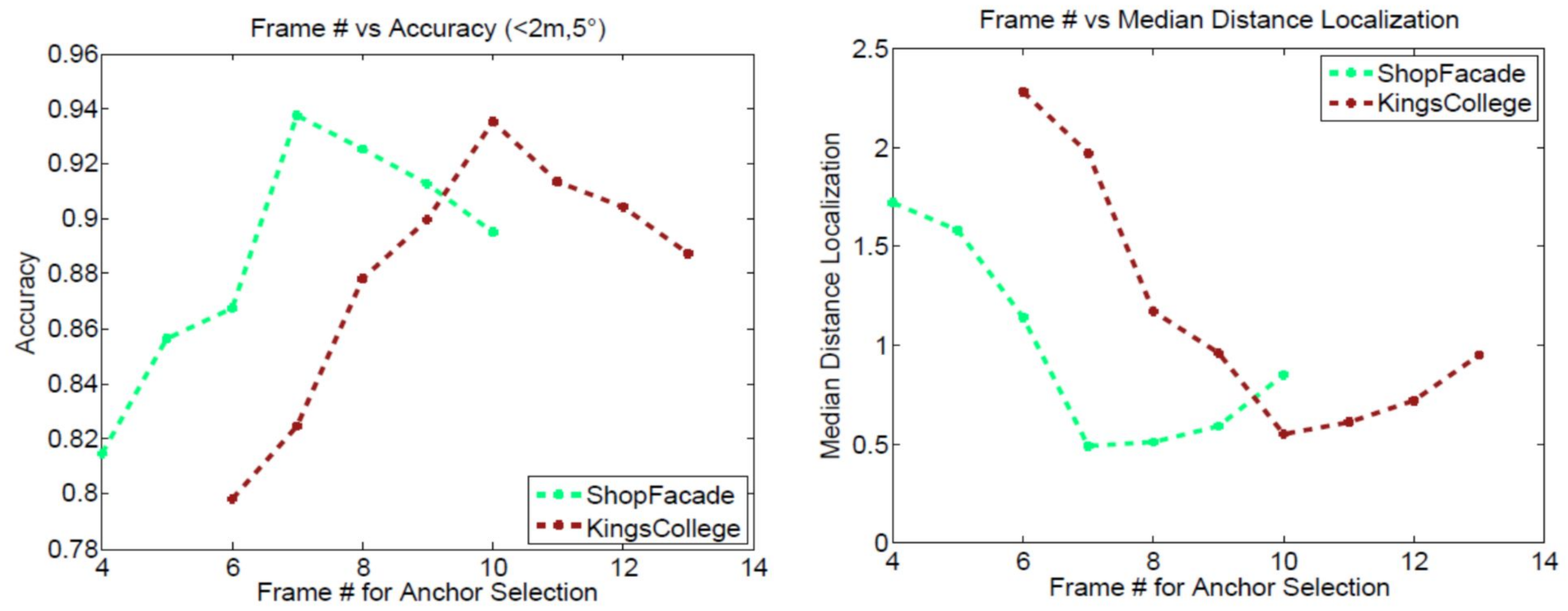
- Our model performs better than the previously best performing variation of PoseNet on the Cambridge Landmarks and the 7 Scenes dataset.
- Performance improves when we allow the network to learn the most relevant anchor point by not enforcing the cross entropy from the nearest anchor point. [rightmost column]

Scene	**Median Dist. (DenseNet + DOF Regressor)**	**Median Dist. (Our method)**	**Accuracy (DenseNet + DOF Regressor)**	**Accuracy (Our Method)**
Shop Facade	1.32m	**0.52m**	82.64%	**93.76%**
King's College	1.45m	**0.57m**	81.80%	**93.52%**

Comparison with DenseNet feature extractor followed by simple regression acts as a control for our method. The improvement is not solely dependent on the quality of feature extractor.

Scene	**DenseNet (Feature Extractor)**		**GoogleNet (Feature Extractor)**		**MobileNet (Feature Extractor)**	
	Performance	**FLOPs**	**Performance**	**FLOPs**	**Performance**	**FLOPs**
Kings College	$0.57m, 0.88^\circ$	5998 M	$0.79m, 0.95^\circ$	760 M	$0.67m, 0.94^\circ$	569 M
Shop Facade	$0.52m, 2.27^\circ$		$0.77m, 3.25^\circ$		$0.60m, 2.31^\circ$	

Comparing performance vs FLOPs for different feature extractors. A trade-off is observed between the two parameters which is essential for deployment in a real-time system.



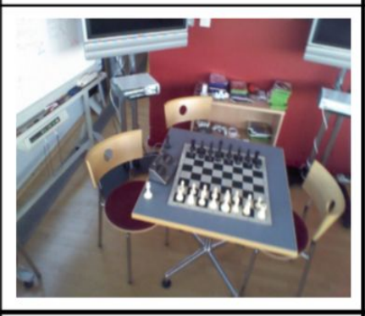
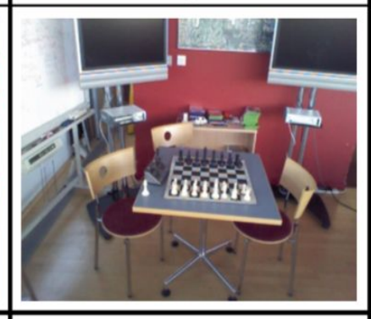
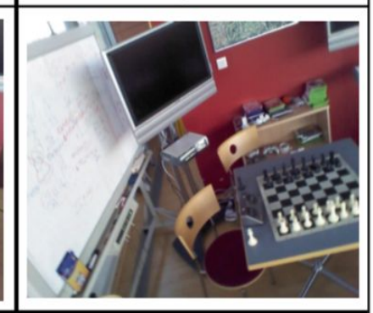
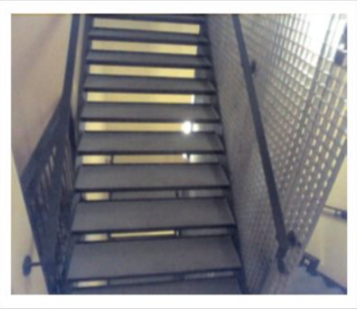
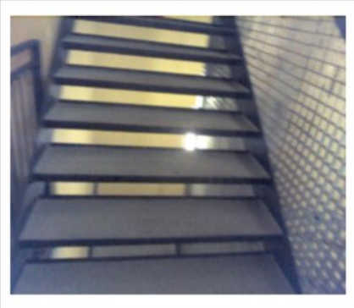
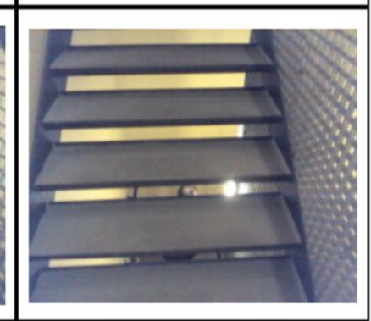
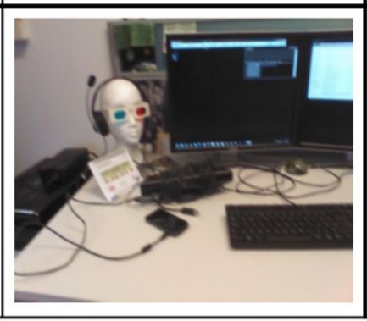
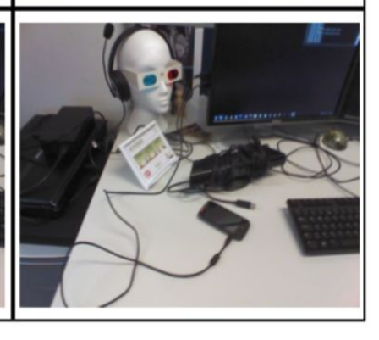
We vary the frame number for anchor point definition for every scene. Plotting this hyperparameter against accuracy and median distance error in localization helps us to finalize on an optimum value for it.

Qualitative Results

Scene	Input Frame	Nearest Anchor Point	Learned Anchor Point
Great Court			
King's College			
Old Hospital			

We contrast the nearest anchor point and the learned anchor point for an input query in the Cambridge Landmarks dataset and the 7 Scenes dataset.
In certain cases, a non occluded frame is learnt to be the reference anchor point, as seen [left] for the Old Hospital scene.

Scene	Input Frame	Nearest Anchor Point	Learned Anchor Point
Shop Facade			
St. Mary's Church			
Street			

Scene	Input Frame	Nearest Anchor Point	Learned Anchor Point
Chess			
Stairs			
Heads			

soham.saha@research.iiit.ac.in, girish.varma@iiit.ac.in, jawahar@iiit.ac.in