FAST 6DOF POSE ESTIMATION WITH SYNTHETIC TEXTURELESS CAD MODEL FOR MOBILE APPLICATIONS

Bowen Chen, Juhan Bae, Dibyendu Mukherjee

Epson Canada Ltd.

PROBLEM

Conventional methods for **6 Degrees-of-Freedom (DoF) pose estimation** require one or more of the following:

- Large-scale real training data
- Synthetic textured data
- Use of RGB-D for better inference if trained from only synthetic
- Multiple stages of regression or classification for accuracy when extended to multiple objects and large view-ranges

CONTRIBUTIONS

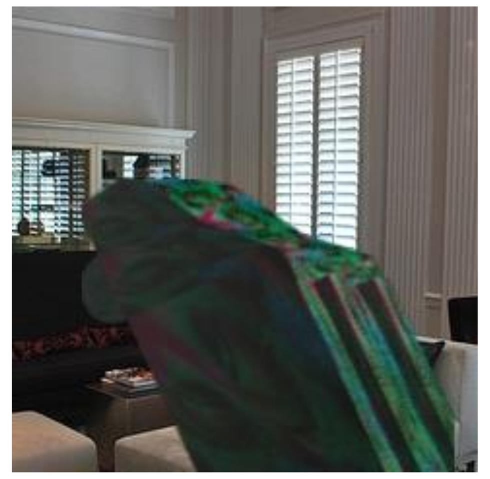
Propose two algorithms (**VIEWMOD**, **BBOX9**) that

- Only require **synthetic textureless CAD model** for training.
- Use of **only RGB** information during inference
- **Real-time inference** for mobile CPUs.

SYNTHETIC-TO-REAL DOMAIN ADAPTATION

1. Project 3D model to 2D image with **random textures and lighting**

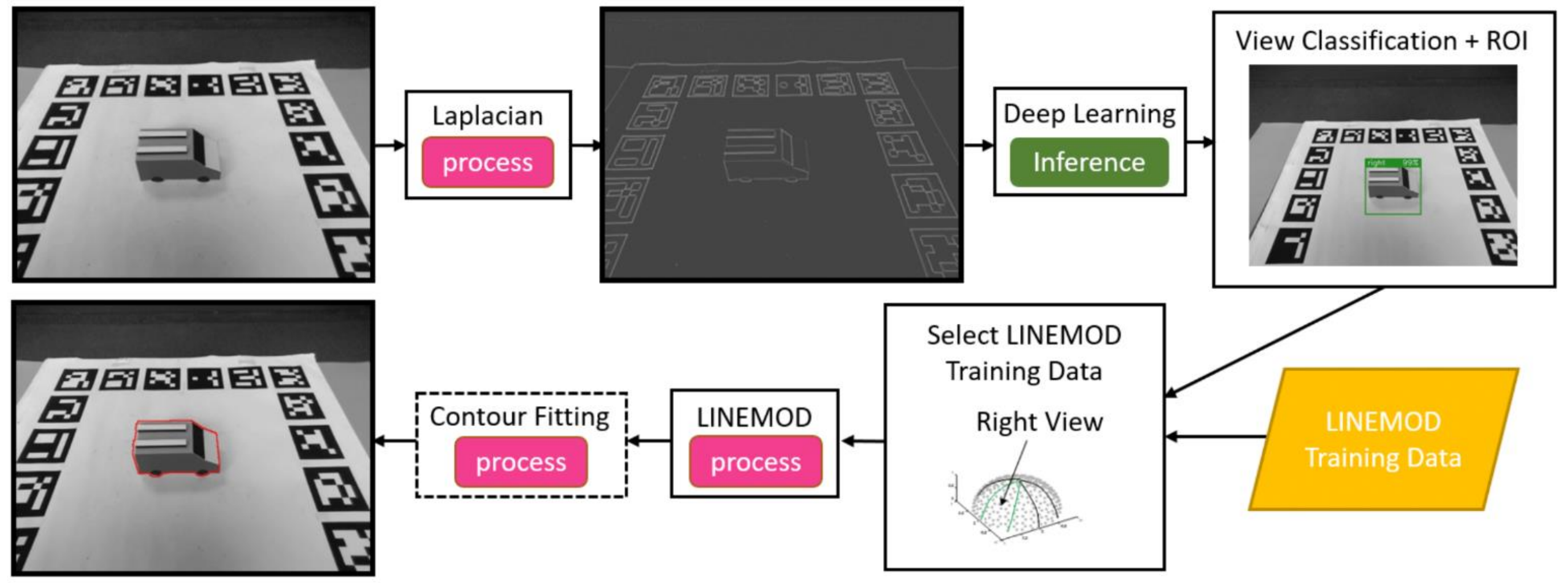
2. Apply **Random scaling**, **in-plane rotation**, and **background** with **noise** (Gaussian blur, motion blur, and additive noise)



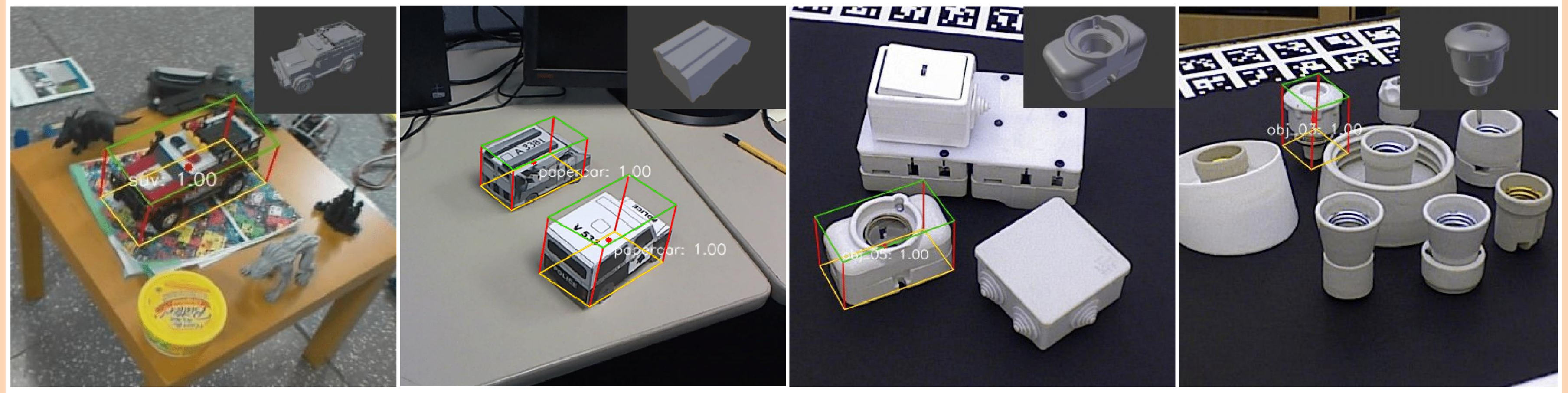
3. Apply **Laplacian Filter**

VIEWMOD

- Use 2D bounding box **detection** with **view-classification** followed by a **LINEMOD** [1] based pose estimation
- **Fast** and **accurate** two-stage inference with improved **interpretability** to detect failures.



BBOX9



Perform a one-stage direct regression of a **3D bounding box** surrounding the object, followed by a **PnP** routine to estimate the object's **6 DoF pose**.

$$\mathcal{L}(x,c,l,g) = \frac{1}{N}(\mathcal{L}_{\text{conf}}(x,c) + \alpha\mathcal{L}_{\text{loc}}(x,l,g))$$

$$\mathcal{L}_{\text{loc}}(x,l,g) = \sum_{i\in \text{Pos}}^{N} \sum_{m=1}^{18} x_{ij}^{k}\, \text{smooth}_{L1}(l_i^m - \hat{g}_j^m)$$

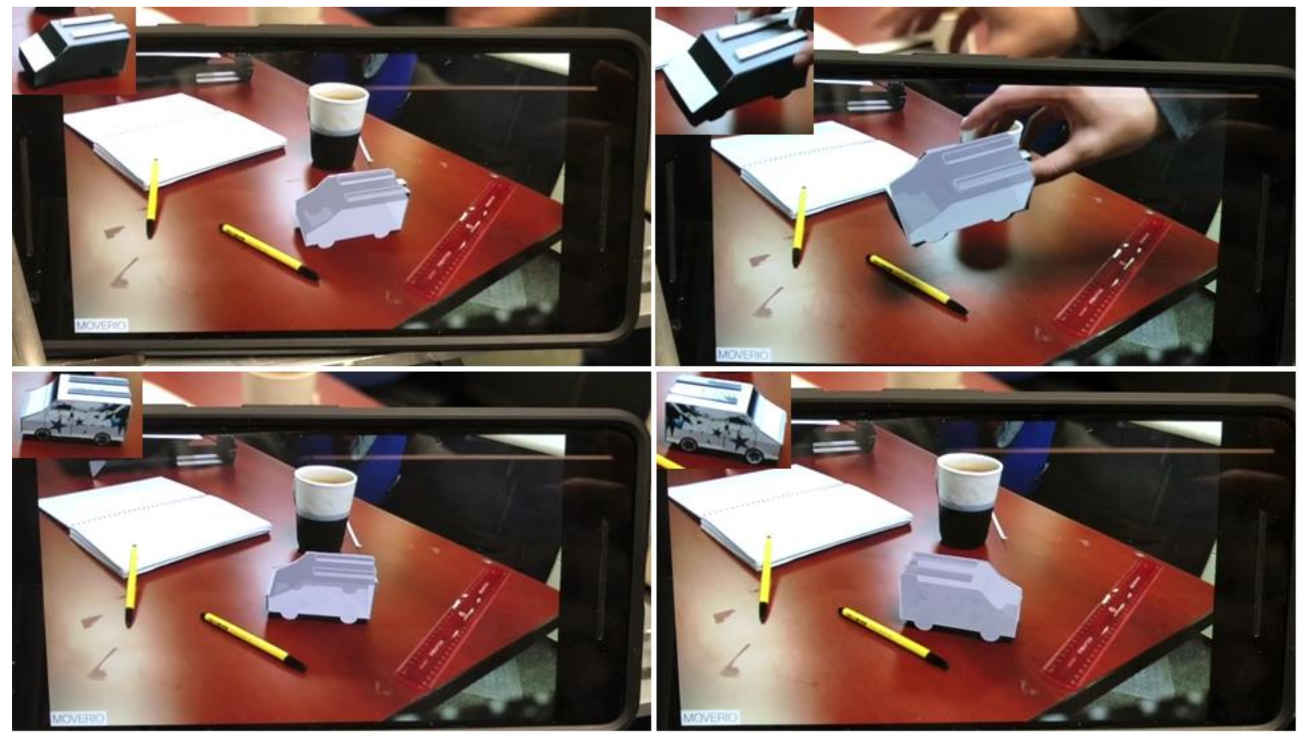
$$\hat{g}_j^m = \begin{cases} (g_j^m - d_i^{cx})/d_i^w, & \text{if } m \text{ is odd} \\ (g_j^m - d_i^{cy})/d_i^w, & \text{otherwise} \end{cases}$$

RESULTS

Scene ID: [Obj. IDs]	BB8 (**real training**) [2]	VIEWMOD (**textureless training**)		BBOX9 (**textureless training**)	
	>10% visibility	>10% visibility	>70% visibility	>10% visibility	>70% visibility
1: [2, 30]	50.8, 55.4	64.1, 66.0	**71.3**, **75.8**	44.0, 35.8	48.9, 40.7
2: [5, 6]	56.5, 55.6	81.0, 55.0	**90.7**, 62.0	75.4, 60.1	84.4, **67.8**
4: [5, 26, 28]	68.7, **53.3**, 40.6	68.0, 46.0, 56.7	**80.7**, 46.0, **64.2**	65.6, 37.7, 35.1	78.9, 37.7, 39.8
5: [1, 10, 27]	**39.6**, 69.9, 50.1	20.8, 69.7, 50.8	21.6, **77.6**, **56.7**	18.7, 56.7, 24.0	19.3, 63.3, 28.9
7: [1, 3, 13]	42.0, 61.7, **64.5**	42.5, 64.1, 18.5	**47.4**, 70.4, 21.3	41.5, 64.7, 12.4	46.2, **71.0**, 14.3
7: [14, 15]	**40.7**, **39.7**	34.1, 17.1	37.7, 20.6	28.2, 17.1	31.1, 20.6
7: [16, 17, 18]	**45.7**, 50.2, 83.7	33.1, 64.3, 76.7	38.5, **75.4**, **86.4**	21.0, 33.9, 71.8	24.4, 39.1, 81.9
Average	55.3	51.6	**58.0**	41.3	46.6

BB8 uses real RGB images while VIEWMOD and BBOX9 **only use synthetic textureless CAD models.**

MOBILE INFERENCE



VIEWMOD and **BBOX9** take **~200ms** per frame using CPU on Google Pixel 2, using a Tensorflow API.

CONCLUSION

We introduced an efficient and user-friendly 6DoF pose estimation framework for mobile applications:

- **Effective domain adaptation** strategy to use synthetic textureless CAD.
- **Real-time inference** for mobile CPUs.

Reference:

[1] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in ACCV. Springer, 2012, pp. 548–562

[2] M. Rad and V. Lepetit, "BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth," in ICCV, 2017, vol. 1, p. 5