Fraunhofer Heinrich Hertz Institute
Computer Vision & Graphics
Vision & Imaging Technologies Department

Paper Number: 0899

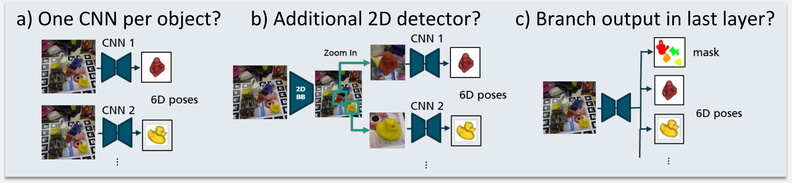
CASAPose: Class-Adaptive and Semantic-Aware Multi-Object Pose Estimation

Niklas Gard, Anna Hilsmann, Peter Eisert

1. Problem Definition

- Pose estimation of known 3D shapes in monocular images.
- Usecases require efficient simultaneous estimation of multiple objects' poses.
- Training only on synthetic data, as annotation of complex scenes with multiple objects is complicated. This requires the bridging of a domain gap.

State-of-the-Art Multi-object Pose Estimation

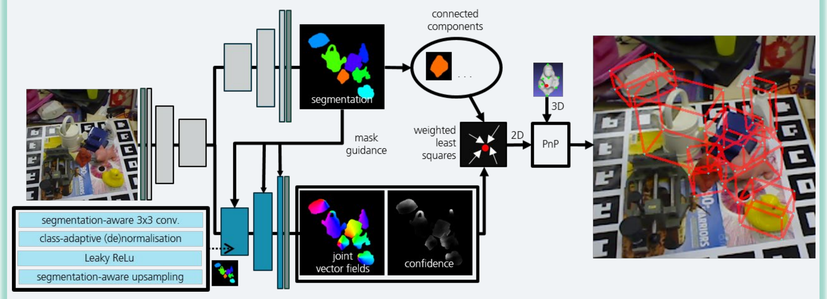


a) Multiple evaluations per image with limited knowledge per model.
b) Two-stage method (extra training) with additional processing (crop and scale).
c) High dimensional output which increases with every added object. Difficult to train (GPU memory) and performance issues.

Our Contribution

CASAPose is a simple and fast method for estimating the pose of multiple objects simultaneously using a single network.

2. CASAPose Architecture



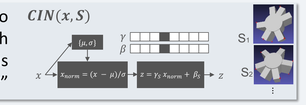
Decoder guidance: Segmentation locally influences keypoint prediction.
End-to-end differentiable: 2D keypoint projections directly calculated and evaluated.
Reduced network size: Increases by one output and few weights per object.

3. Network Components

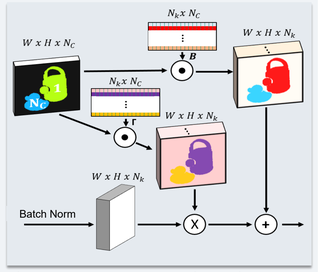
Object-adaptive Local Weights

- **Conditional Instance Normalisation (CIN):** Learned (de)normalisation (affine trans-formation) depends on object class (S_i).

"How to know which object class is present?"



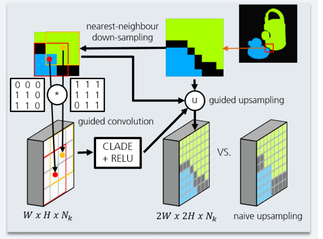
- **Class-adaptive (de)normalisation (CLADE) [7]** : Learned (de)normalisation depends on semantic class at a specific location.



The segmentation is used to select modulation parameters from the weight matrices Γ and B.

Semantically Guided Decoder

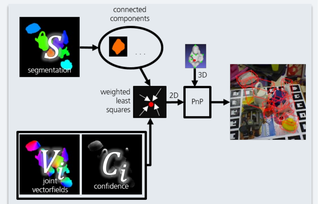
- **Segmentation-aware convolution**: Evaluates semantic mask at filter position and considers only features from the same region.
- **Segmentation-aware upsampling**: Enlarges feature map without losing alignment with segmentation.



Guided operations improve the alignment between features and mask.

Differentiable Keypoint Regression (DKR)

- Directly optimises common intersection point in loss function. Replaces non-differentiable RANSAC-voting [8].
- **Weighted linear equation system**: Sum of squared distances minimised with respect to cost function D.
- **Weighting by learned confidence C**: Recognise relevant part of an object to localise keypoint.



$$\hat{k}_{i,j} = \underset{k}{\operatorname{argmin}}\, D(k; S_j, V_i, C_i, P)$$

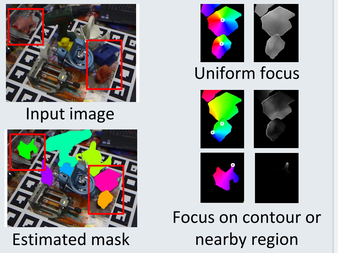
i : Index of keypoint
j : Index of object
$k_{i,j}$: 2D keypoint
P : pixel locations

Keypoint location is calculated from estimated vectors V and confidence C by solving weighted least squares with Moore-Penrose inverse.

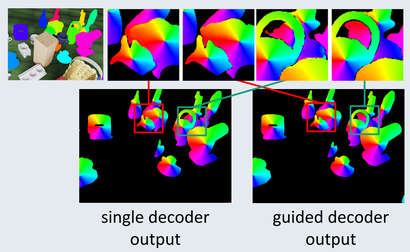
Implementation Details

- Resnet-18 backbone.
- Guidance with ground truth mask during training.
- Connected component filtering reduces influence of segmentation clutter during inference (DKR).

4. Visual Results



Depending on the keypoint location, the confidence map learns to focus on the entire mask or a specific area.



Semantic guidance minimises artefacts if occlusions are present. This improves the accuracy of the pose estimates.

5. Experiment Design

- Physically-based rendering (*pbr*) images of BOP challenge.
- Comparison with methods not using annotated real data.
- Train single network for multiple objects.
- Datasets: LINEMOD (**LM**), Occluded LINEMOD (**LM-O**), HomebrewedDB* (**HB**)
- **ADD/S** and Projection 2D metric.

* Sequences with Linemod objects

Image from pbr dataset with pose annotation.

6. Results

Ablation Study

Architecture	LM-O	LM	pbr (synth.)
Base	22.2	49.8	42.2
Base + C	26.3	55.7	47.5
Base + C/GCU	26.7	59.0	49.0
Base + DKR	28.9	59.2	52.5
Base + C + DKR	29.9	64.7	56.1
Base + C/GCU + DKR	**32.7**	**68.1**	**57.6**

Base: single decoder merged vector field
C: split decoder with CLADE guidance
C/GCU: split decoder with full semantic guidance
DKR: differentiable keypoint regression

Capacity	LM*	LM-O
13 objects	60.4	32.7
8 objects	59.2	35.9
2 x 4 objects	**64.5**	**38.7**

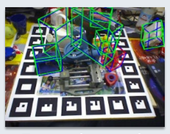
ADD/S Recall with respect to the network component (top) and object capacity (bottom).

- All introduced network components further improve accuracy.
- DKR improves result also for simplest architecture by a large margin.
- Access to silhouette (C/GCU) is beneficial to bridge domain gap.

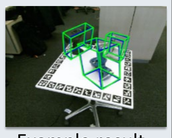
* 8 LM-O objects

Comparison with the State-of-the-Art

Method	Data	Single-stage	Result LM-O (ADD/S Recall)
DSC-PoseNet [1]	pbr + RGB	-	24.8
PyraPose [2]	pbr	✓	28.1
Self6D [3]	pbr + RGBD	-	32.1
DAKDN [4]	pbr + RGB	-	33.7
SD-Pose [5]	pbr	-	34.6
Ours	**pbr**	**✓**	**35.9**



Example result LM-O



Example result HB

- **Performance on LM-O**: (see above). Further comparison with single-stage multi-object method EPOS [6] : ours is multiple times faster with the same accuracy.
- **Performance on LM**: Increase of 7.4% compared to next best single-stage multi-object method [2].
- **Performance on HB**: The 13-object model without retraining surpasses next best method DAKDN[4] by 35%.

7. Conclusions

- **Class-adaptiveness and semantic awareness** improve the performance of a multi object 6D pose estimator.
- **Local feature processing** minimises interference between overlapping regions in a reduced output space. Use one network to estimate poses of multiple objects.
- **Differentiable Keypoint Regression** reduces domain gap.
- **Next step**: add single-stage instance awareness.
- Custom layers could be integrated also in other pose estimation architectures.

[1] Yang, Zongxin. "DSC-PoseNet: Learning 6dof object pose estimation via dual-scale consistency." CVPR 2021.
[2] Thalhammer, Stefan et al. " PyraPose: Feature pyramids for fast and accurate object pose estimation under domain shift." ICRA 2021.
[3] Wang, Gu et al. "Self6d: Self-supervised monocular 6d object pose estimation." ECCV 2020.
[4] Li, Zhigang et al. "Keypoint- graph-driven learning framework for object pose estimation." CVPR 2021.
[5] Li, Zhigang et al. "SD-Pose: Semantic decomposition for cross-domain 6d object pose estimation." AAAI 2021.
[6] Hodan, Tomas et al. "EPOS: Estimating 6d pose of objects with symmetries." CVPR 2020.
[7] Tan, Zhentao, et al. "Efficient Semantic Image Synthesis via Class-Adaptive Normalization." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
[8] Peng, Sida, et al. "Pvnet: Pixel-wise voting network for 6dof pose estimation." CVPR 2019.

www.hhi.fraunhofer.de
niklas.gard@hhi.fraunhofer.de

BMVC 2022
Supported by: Federal Ministry for Economic Affairs and Climate Action, on the basis of a decision by the German Bundestag