Institute for Infocomm Research
I²R

CREATING GROWTH, ENHANCING LIVES

# 6D Pose Estimation with Correlation Fusion

Yi Cheng, Hongyuan Zhu, Ying Sun, Cihan Acar, Wei Jing,
Yan Wu, Liyuan Li, Cheston Tan, Joo-Hwee Lim

**Institute for Infocomm Research,**
**Agency for Science, Technology and Research (A*STAR), Singapore**

# Outlines

Introduction

Objective

Methodology

Results

Robotic Grasping Experiments

Conclusion

CREATING GROWTH, ENHANCING LIVES

# Introduction

- 6D pose estimation, which aims to predict the 3D rotation and translation from object space to camera space, is useful in 3D object detection and recognition, robot grasping and manipulation.

- Limitations of existing methods:
  - RGB-only methods ignore complementary information from depth modality, which are vulnerable to heavy occlusion and poor illumination.
  - RGB-D based methods fail to adequately exploit the consistent and complementary information between the RGB and depth modalities.

# Objective

- To address the aforementioned limitations, we propose a novel Correlation Fusion (CF) framework which models the feature correlation within and between RGB and depth modalities to improve the final performance of 6D pose estimation.

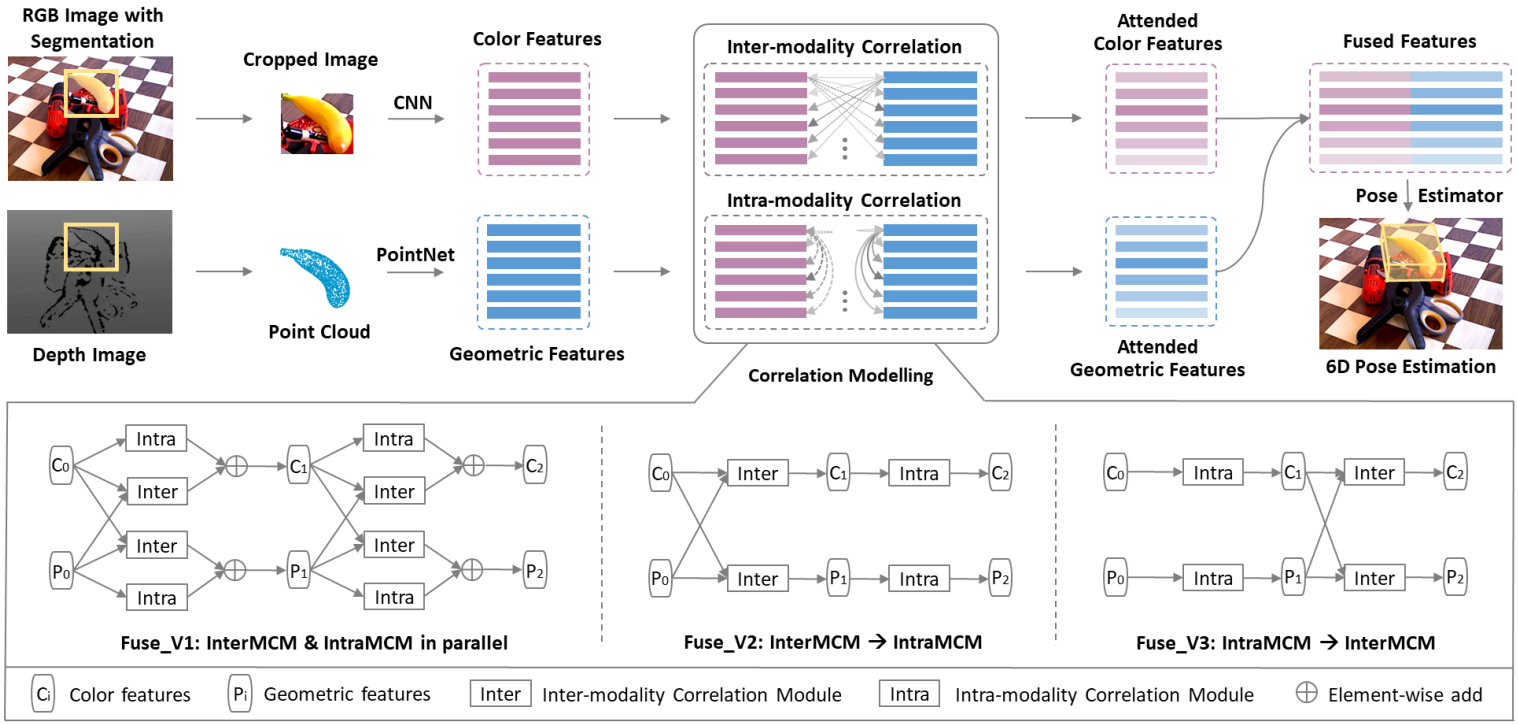
- Our contribution can be summarized as follows:
  - We propose the intra- and inter-correlation modules to exploit the consistent and complementary information within and between RGB and depth modalities for 6D pose estimation.
  - We explore multiple strategies for fusing the intra- and inter-modality information flow to learn discriminative multi-modal features.
  - We demonstrate that the proposed method can achieve the state-of-the-art performance on widely-used benchmark datasets for 6D pose estimation, including LineMOD and YCB-Video datasets.
  - We showcase that our method can benefit robot grasping tasks by providing an accurate estimation of object pose.

# Methodology

- The overall framework includes three stages:
  - **Semantic Segmentation and Feature Extraction**. We first segment the target objects in the image with an existing semantic segmentation architecture, and then generate the color and geometric features with the predicted segmentation maps.
  - **Multi-modality Correlation Learning**. It consists of Intra-modality Correlation Modelling (IntraMCM), Inter-Modality Correlation Modelling (InterMCM) and Multi-modality Fusion Strategies.
  - **Iterative Pose Refinement**. A refiner network is employed for iteratively refining the predicted object pose.

# Methodology

➢ The framework is illustrated as below:



# Results

➢ Comparison with existing state-of-the-art methods on YCB-Video dataset:

TABLE I

THE 6D POSE ESTIMATION ACCURACY ON **YCB-VIDEO DATASET** IN TERMS OF THE **ADD(-S)** <**2CM AND THE AUC OF ADD(-S)**. THE OBJECTS WITH BOLD NAME ARE CONSIDERED AS SYMMETRIC. ALL THE METHODS USE RGB-D IMAGES AS INPUT.(**BEST ZOOM-IN AND VIEW IN PDF.**)

| Methods | PoseCNN [5] | | DenseFusion [21] | | OURS (IntraMCM) | | OURS (InterMCM) | | OURS (Fuse_V1) | | OURS (Fuse_V2) | | OURS (Fuse_V3) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | AUC | <2cm | AUC | <2cm | AUC | <2cm | AUC | <2cm | AUC | <2cm | AUC | <2cm | AUC | <2cm |
| 002_master_chef_can | 68.06 | 51.09 | 73.16 | 72.56 | 87.61 | 88.37 | 86.94 | 88.07 | 86.18 | 86.28 | **92.47** | **98.71** | 87.24 | 86.18 |
| 003_cracker_box | 83.38 | 73.27 | 94.21 | 98.50 | 94.80 | **99.54** | 93.69 | 99.08 | 91.79 | 98.50 | **95.45** | 98.62 | 95.20 | 99.19 |
| 004_sugar_box | **97.15** | 99.49 | 96.50 | **100.00** | 93.73 | **100.00** | 95.06 | **100.00** | 95.68 | **100.00** | 96.69 | 99.92 | 96.19 | 99.58 |
| 005_tomato_soup_can | 81.77 | 76.60 | 85.42 | 82.99 | 91.50 | 95.42 | 90.23 | 93.19 | **92.73** | 95.56 | 92.02 | **95.76** | 91.51 | 95.56 |
| 006_mustard_bottle | **98.01** | 98.60 | 94.61 | 96.36 | 92.27 | 98.04 | 93.10 | 98.60 | 89.66 | 91.04 | 94.82 | 97.48 | 95.29 | **99.16** |
| 007_tuna_fish_can | 83.87 | 72.13 | 81.88 | 62.28 | 80.86 | 69.69 | 86.18 | 84.58 | 85.94 | 83.45 | **88.85** | 84.15 | 85.27 | **86.31** |
| 008_pudding_box | **96.62** | **100.00** | 93.33 | 98.60 | 91.69 | 97.13 | 91.83 | 98.60 | 91.76 | 99.07 | 93.16 | 98.60 | 94.10 | 98.13 |
| 009_gelatin_box | **98.08** | 100.00 | 96.68 | 100.00 | 95.35 | 100.00 | 95.06 | 100.00 | 95.92 | 100.00 | 95.68 | 100.00 | 97.28 | 100.00 |
| 010_potted_meat_can | 83.47 | 77.94 | 83.54 | 79.90 | 85.01 | 83.55 | 83.77 | 80.81 | 84.07 | 82.90 | **86.19** | 83.94 | 86.03 | **84.07** |
| 011_banana | 91.86 | 88.13 | 83.49 | 88.13 | 84.70 | 81.79 | 90.71 | 98.68 | 88.73 | 98.15 | **92.57** | **98.94** | 86.84 | 88.92 |
| 019_pitcher_base | **96.93** | 97.72 | 96.78 | 99.47 | 95.76 | 98.02 | 96.55 | **100.00** | 96.07 | **100.00** | 95.43 | 98.42 | 95.97 | 99.65 |
| 021_bleach_cleanser | **92.54** | **92.71** | 89.93 | 90.96 | 87.93 | 83.19 | 89.10 | 83.28 | 90.19 | 89.70 | 88.99 | 86.20 | 89.00 | 83.28 |
| **024_bowl** | 80.97 | 54.93 | **89.50** | 94.83 | 88.70 | **97.78** | 87.00 | 84.24 | 86.32 | 90.64 | 86.06 | 94.33 | 89.08 | 95.81 |
| 025_mug | 81.08 | 55.19 | 88.92 | 89.62 | 91.84 | 92.77 | 92.00 | 94.97 | 91.06 | 91.98 | **93.51** | 94.81 | 93.44 | **96.38** |
| 035_power_drill | **97.66** | **99.24** | 92.55 | 96.40 | 92.05 | 95.65 | 86.60 | 90.35 | 85.05 | 87.70 | 82.89 | 84.77 | 93.52 | 98.20 |
| **036_wood_block** | 87.56 | 80.17 | **92.88** | **100.00** | 91.44 | 98.35 | 90.16 | **100.00** | 91.46 | 99.59 | 92.32 | 99.59 | 92.35 | 98.76 |
| 037_scissors | 78.36 | 49.17 | 77.89 | 51.38 | **91.28** | 86.37 | 78.98 | 67.40 | 79.25 | 64.70 | 90.15 | **89.50** | 88.38 | 86.74 |
| 040_large_marker | 85.26 | 87.19 | 92.95 | **100.00** | 93.55 | **100.00** | 93.84 | **100.00** | **94.10** | **100.00** | 93.91 | 99.85 | 93.82 | 99.85 |
| **051_large_clamp** | **75.19** | 74.86 | 72.48 | **78.65** | 71.27 | 78.51 | 72.14 | 77.95 | 70.18 | 75.70 | 70.31 | 76.69 | 73.22 | **78.65** |
| **052_extra_large_clamp** | 64.38 | 48.83 | 69.94 | 75.07 | 70.11 | **76.83** | **73.74** | 75.51 | 69.71 | 75.22 | 69.53 | 74.49 | 70.80 | 76.25 |
| **061_foam_brick** | **97.23** | 100.00 | 91.95 | 100.00 | 94.36 | 100.00 | 94.15 | 100.00 | 93.08 | 100.00 | 94.62 | 100.00 | 94.89 | 100.00 |
| MEAN | 86.64 | 79.87 | 87.55 | 88.37 | 88.85 | 91.48 | 88.61 | 91.21 | 88.04 | 90.96 | 89.79 | **93.08** | **89.97** | 92.89 |

# Results

- Comparison with existing state-of-the-art methods on LINEMOD dataset:

TABLE II

THE 6D POSE ESTIMATION ACCURACY ON THE **LINEMOD DATASET** IN TERMS OF THE **ADD(-S)** METRIC. THE OBJECTS WITH BOLD NAME (GLUE AND EGGBOX) ARE CONSIDERED AS SYMMETRIC. ALL THE METHODS USE RGB-D IMAGES AS INPUT.

| | SSD6D [16] | BB8 [13] | DenseFusion [21] | OURS (IntraMCM) | OURS (InterMCM) | OURS (Fuse_V1) | OURS (Fuse_V2) | OURS (Fuse_V3) |
|---|---|---|---|---|---|---|---|---|
| ape | 65 | 40.4 | 92.3 | 94.9 | 95.2 | 94.8 | **95.6** | 95.4 |
| bench | 80 | 91.8 | 93.2 | 93.7 | 94.0 | 96.1 | **96.9** | 96.1 |
| camera | 78 | 55.7 | 94.4 | 97.5 | 95.6 | 96.0 | **97.9** | 97.5 |
| can | 86 | 64.1 | 93.1 | 95.4 | 95.7 | 92.2 | **96.0** | 95.0 |
| cat | 70 | 62.6 | 96.5 | 98.4 | 98.8 | **99.2** | 97.8 | 99.1 |
| driller | 73 | 74.4 | 87.0 | 92.2 | 92.7 | 91.4 | **95.6** | 94.7 |
| duck | 66 | 44.3 | 92.3 | **96.2** | 95.1 | 95.7 | 95.7 | 95.8 |
| **eggbox** | **100** | 57.8 | 99.8 | **100.0** | 99.6 | **100.0** | 99.9 | 99.9 |
| **glue** | **100** | 41.2 | **100.0** | 99.8 | 99.8 | 99.8 | 99.7 | 99.8 |
| hole | 49 | 67.2 | 92.1 | 95.2 | 95.6 | 95.8 | 96.7 | **97.1** |
| iron | 78 | 84.7 | 97.0 | 95.8 | 96.2 | 97.4 | 97.8 | **98.4** |
| lamp | 73 | 76.5 | 95.3 | 95.4 | 96.3 | 96.5 | **97.0** | 96.8 |
| phone | 79 | 54.0 | 92.8 | 97.3 | **97.5** | 95.6 | 97.0 | 97.4 |
| MEAN | 77 | 62.7 | 94.3 | 96.3 | 96.3 | 96.2 | **97.2** | 97.1 |

# Results

- We present some qualitative results on the examples from YCB-Video dataset, for both DenseFusion and our proposed method.

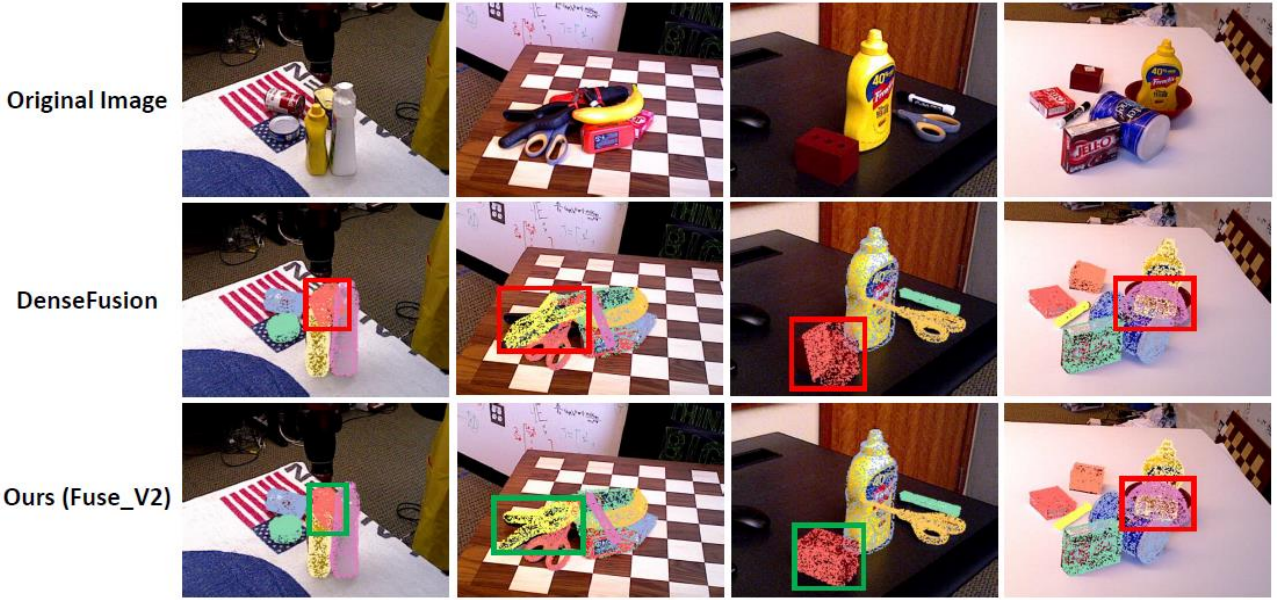



Fig. 4. **Visualizations of results on the YCB-Video Dataset.** The first row is original RGB image, the second row is from DenseFusion, and third row is our proposed method Fuse_V2. The red boxes show the cases with poor pose estimation while the green boxes shows the ones with good pose estimation.

# Robotic Grasping Experiments

- We also carry out robotic grasping experiments in both simulation and real world to demonstrate that our method is effective in assisting robots to correctly grasp objects by providing accurate object pose estimation.
- We compare the proposed method with DenseFusion in Gazebo simulation environment:

TABLE III
SUCCESS RATE FOR THE GRASPING EXPERIMENTS WITH ROBOTIC ARM IN SIMULATION ENVIRONMENT OF GAZEBO.

| Success Attempts (%) | tomato_soup_can | mustard_bottle | banana | bleach_cleanser |
|---|---|---|---|---|
| DenseFusion [21] | 80.0 | 70.0 | 55.0 | 65.0 |
| Ours | 90.0 | 85.0 | 75.0 | 80.0 |

# Conclusion

- In this paper, we have proposed a novel Correlation Fusion framework with intra- and inter-modality correlation learning for 6D object pose estimation. The IntraMCM module is designed to learn prominent modality-specific features and the InterMCM module is to capture complementary modality features. Subsequently, multiple fusion schemes are explored to further improve the performance on 6D pose estimation. Extensive experiments conducted on YCB Video dataset, LINEMOD dataset and a real-world robot grasping task demonstrate the superior performance of our method to several benchmarking methods.

CREATING GROWTH, ENHANCING LIVES

Institute for Infocomm Research

I²R

# THANK YOU

www.a-star.edu.sg