DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion

Teammate: 罗一飞 许左 王宋然 相昊阳 李东睿 陈钊楷

2024.5.12

AncoraSIR.com

Motivation and Main Problem

High-level description of problem being solved

- Meet speed requirements
- exhibit robustness when handling objects of various shapes and textures even under conditions of heavy occlusion, sensor noise, and changes in lighting
- Save money

AncoraSIR.com

Motivation and Main Problem

why prior approaches didn't already solve? & Key insights

- *Estimating Pose from RGB Images*

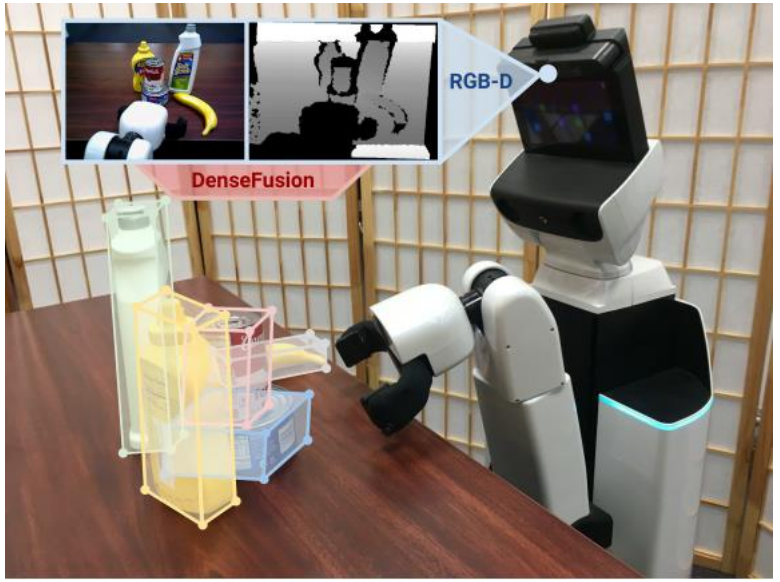
  Becomes unreliable in low-texture or low-resolution inputs.

- Estimating Pose from Depth/Point Clouds

  High computational cost.

- Estimating Pose from RGB-D Data

  Difficult to accurately estimate pose.



AncoraSIR.com

Problem Setting

Problem formulation, key definitions and notations

- **Problem formulation**

  I. Integrating color and depth information obtained by RGB cameras

  II. The neural network architecture integrates iterative fine-tuning, eliminating dependence on post-processing ICP steps.

  Key definitions

  I. End-to-End: A system designed to work seamlessly from start to finish.

  II. 6D pose: includes three dimensional position coordinates (x, y, z) and three dimensional rotation angles (pitch, yaw, roll).

  III. Dense Fusion: This approach integrates RGB and depth image data, utilizing a deep learning network to extract features and perform pose estimation.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Related Work & Limitations of Prior Work

Pose from RGB images

- Classical method
  - "The moped framework: Object recognition and pose estimation for manipulation," (A. Collet, M. Martinez, and S. S. Srinivasa,2011)
  - "Single im age 3d object detection and pose estimation for grasping,"(M. Zhu, K. G. Derpanis, Y. Yang, S. Brahmbhatt, M.Zhang, C. Phillips, M. Lecce, and K. Daniilidis, 2014)
- Learning to predict 2D key points
  - "Learning 6d object pose estimation using 3d object coordinates" (2014)
  - "6-dof object pose from semantic keypoints"(G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis, 2017)
  - "Real-Time Seamless Single Shot 6D Object Pose Prediction"(B. Tekin, S. N. Sinha, and P. Fua, 2018)

- High computing costs and real-time challenges
- It excels with rich textures and high-resolution inputs, but may become unreliable with low texture or low-resolution inputs.

AncoraSIR.com

Related Work & Limitations of Prior Work

Pose from depth/point cloud

- Performing 6D pose estimation directly on 3D point cloud data
  - "Pointnet: Deep learning on point sets for 3d classification and segmentation,"(C. R. Qi, H. Su, K. Mo, and L. J. Guibas . 2016)
  - "Voxelnet: End-to-end learning for point cloud based 3d object detection,".(Y. Zhou and O. Tuzel. 2017)
- Estimating poses using voxelization input through 3D ConvNets
  - "Sliding shapes for 3d object detection in depth images"(S. Song and J. Xiao, 2014)
  - "Deep sliding shapes for amodal 3d object detection
  - in rgb-d images"(2016)

- These methods are computationally expensive, for exampleeach frame takes nearly 20 seconds.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Related Work & Limitations of Prior Work

Setting poses from RGB-D data

- PoseCNN estimates 6D pose directly from image data
- "Posecnn:A convolutional neural network for 6d object pose estimation in cluttered scenes," Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, 2017)
- Extract 3D features from input RGB-D data, and perform grouping and hypothesis verification
- "Learning 6d object pose estimationusing 3d object coordinates,"(E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, 2014)
- "Deep learning of local rgb-d patches for 3d object detection and 6d pose estimation," (2016)

- These methods typically rely on expensive post-processing steps to fully utilize 3D input.
- These features may be hard coded or learned through optimizing alternative objects.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Proposed Approach

Innovation of the framework and Algorithm

- Use the 2D information learned for this task in the embedding space to increase the information of each 3D point, and use this new color depth space to estimate the 6D pose.

- Attitude estimation can improve accuracy through differentiable iterative fine-tuning modules. This module can be trained synchronously with the main architecture, with a short time consumption.

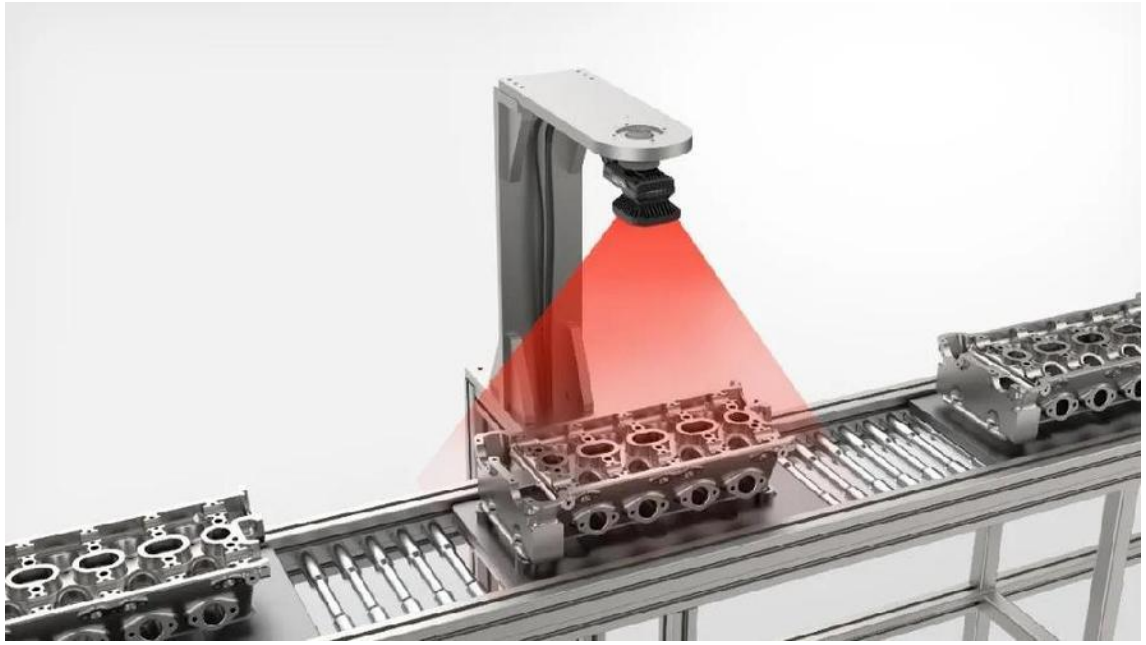
AncoraSIR.com

SUSTech
Southern University of Science and Technology

Theory

Model

- Our goal is to estimate the 6D pose of a set of known we represent 6D poses as ho-mogeneous transformation matrix
- $p \in SE(3)$. In other words, a 6D pose is composed by a rotation $R \in SO(3)$ and a translation $t \in R3$, $p = [R|t]$.

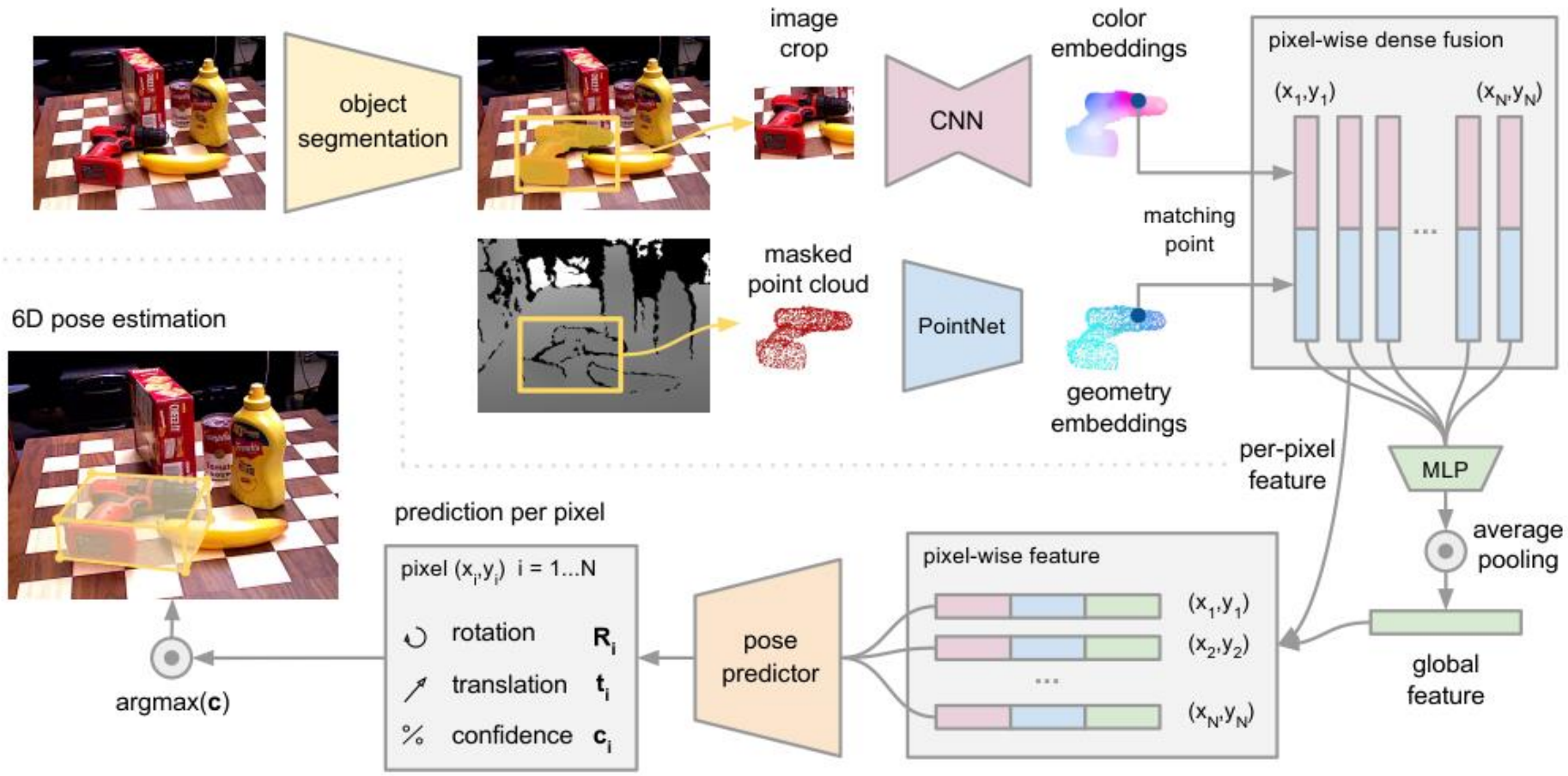


AncoraSIR.com

SUSTech
Southern University of Science and Technology

Theory

Overview of our 6D pose estimation model

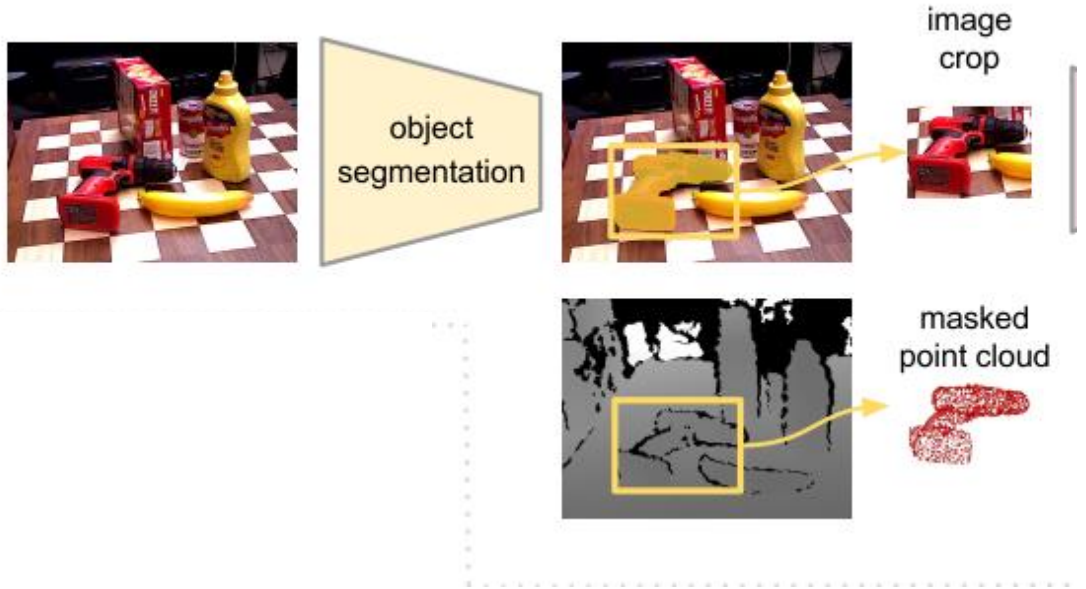


AncoraSIR.com

Theory

Semantic Segmentation

- Our semantic segmentation network is an encoder-decoder architecture that takes an image as input and gener-ates an N +1-channelled semantic segmentation map. Each channel is a binary mask where active pixels depict objects of each of the N possible known classes.



AncoraSIR.com

Theory

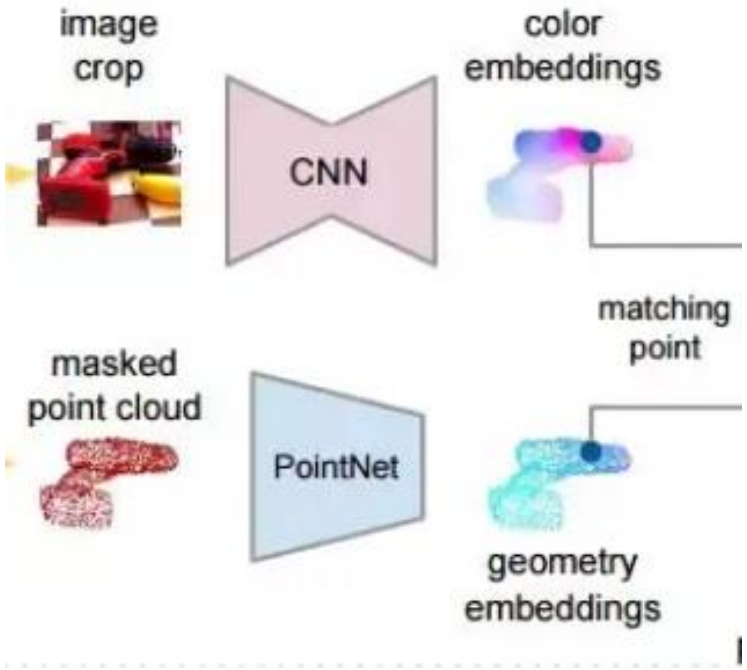
Dense Feature Extraction

- Dense color image feature embedding:

The image embedding network is a CNN-based encoder-decoder architecture that maps an image of size H × W × 3 into a H × W × drgb embedding space.

- Dense 3D point cloud feature embedding:

we first convert the segmented depth pixels into a 3D point cloud using the known camera intrinsics, and then use a PointNet-like architecture to extract geometric features

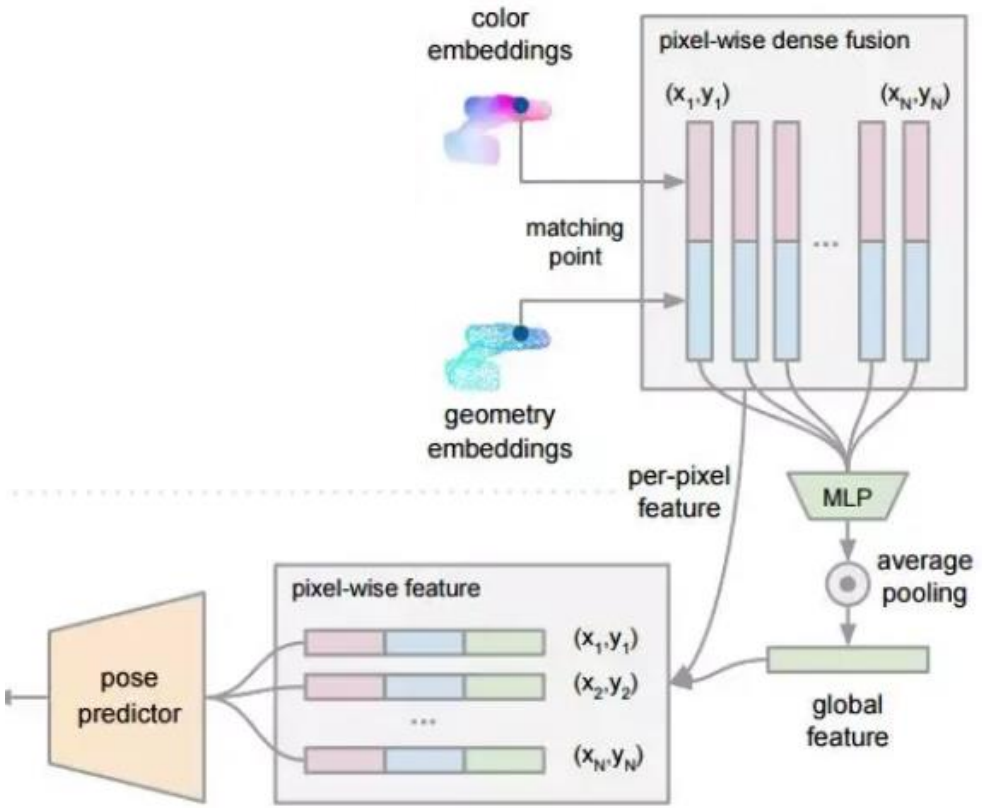


AncoraSIR.com

Theory

Pixel-wise Dense Fusion

- We associate geometric features with corresponding image feature pixels to obtain dense fusion features.
- These paired features are concatenated and fed into another network to generate a fixed-size global feature vector.
- Dense fusion is obtained for each feature by combining dense fusion features with the global feature vector.
- Each pixel-level feature is input into the final network to predict the 6D pose of the object.



AncoraSIR.com

Theory

6D Object Pose Estimation

- minimize for the

prediction per dense-pixel (asymmetric):

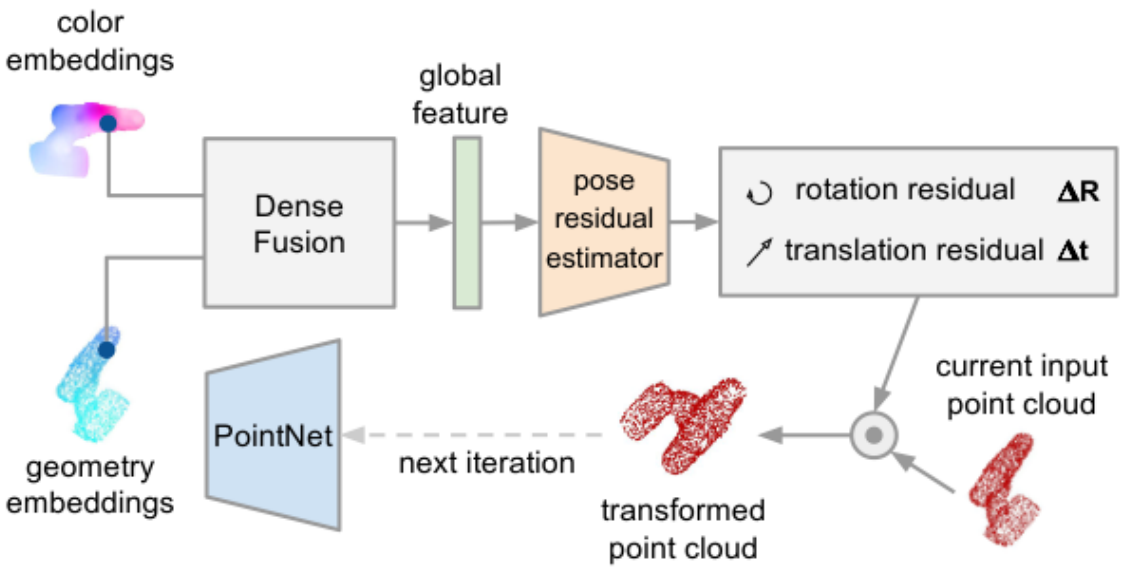
$$L_i^p = \frac{1}{M} \sum_j ||(Rx_j + t) - (\hat{R}_i x_j + \hat{t}_i)||$$

- minimize for the prediction per dense-pixel(symmetric):

$$L_i^p = \frac{1}{M} \sum_j \min_{0<k<M} ||(Rx_j + t) - (\hat{R}_i x_k + \hat{t}_i)||$$

- minimize for the prediction:

$$L = \frac{1}{N} \sum_i (L_i^p c_i - w \log(c_i)),$$

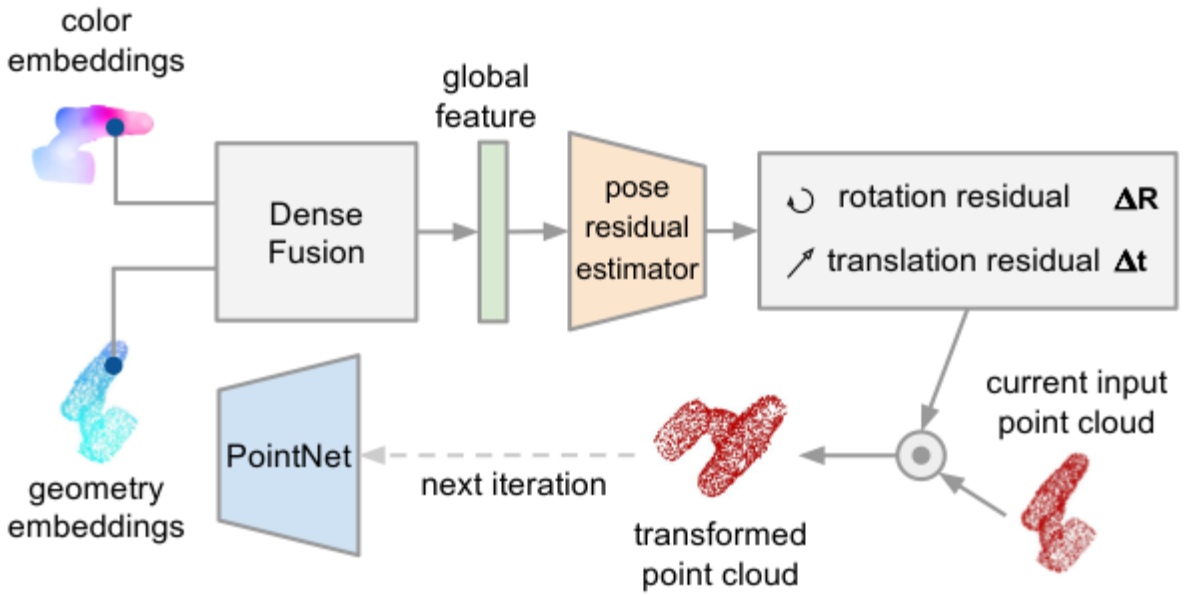


AncoraSIR.com

Theory

Iterative Refinement

- We employ a dedicated pose residual estimation network to refine the pose estimation from the main network. At each iteration, we fuse the image features and point cloud geometric features to obtain a more precise pose estimation. The final pose estimation is composed of the results obtained from multiple iterations.



AncoraSIR.com

Experiments

In order to verify the more accurate effect and real-time performance of the densefusion model, the author set up a series of experiments to verify the performance and effect of the model

AncoraSIR.com

Datasets and Evaluation Metrics

YCB-Video Dataset: This dataset comprises RGB-D images from various videos, with clearly annotated 6D poses of objects within these images. It is used to assess the performance of DenseFusion in handling everyday objects.

Table 1. Quantitative evaluation of 6D pose (ADD-S[40]) on YCB-Video Dataset. Objects with bold name are symmetric.

	PointFusion [41]		PoseCNN+ICP [40]		Ours (single)		Ours (per-pixel)		Ours (iterative)	
	AUC	<2cm	AUC	<2cm	AUC	<2cm	AUC	<2cm	AUC	<2cm
002_master_chef_can	90.9	99.8	95.8	100.0	93.9	100.0	95.2	100.0	**96.4**	100.0
003_cracker_box	80.5	62.6	92.7	91.6	90.8	98.4	92.5	99.3	**95.5**	**99.5**
004_sugar_box	90.4	95.4	**98.2**	100.0	94.4	99.2	95.1	100.0	97.5	100.0
005_tomato_soup_can	91.9	96.9	94.5	96.9	92.9	96.7	93.7	96.9	**94.6**	96.9
006_mustard_bottle	88.5	84.0	**98.6**	100.0	91.2	97.8	95.9	100.0	97.2	100.0
007_tuna_fish_can	93.8	99.8	**97.1**	100.0	94.9	100.0	94.9	100.0	96.6	100.0
008_pudding_box	87.5	96.7	**97.9**	100.0	88.3	97.2	94.7	100.0	96.5	100.0
009_gelatin_box	95.0	100.0	**98.8**	100.0	95.4	100.0	95.8	100.0	98.1	100.0
010_potted_meat_can	86.4	88.5	**92.7**	**93.6**	87.3	91.4	90.1	93.1	91.3	93.1
011_banana	84.7	70.5	**97.1**	99.7	84.6	62.0	91.5	93.9	96.6	**100.0**
019_pitcher_base	85.5	79.8	**97.8**	100.0	86.9	80.9	94.6	100.0	97.1	100.0
021_bleach_cleanser	81.0	65.0	**96.9**	99.4	91.6	98.2	94.3	99.8	95.8	**100.0**
024_bowl	75.7	24.1	81.0	54.9	83.4	55.4	86.6	69.5	**88.2**	**98.8**
025_mug	94.2	99.8	95.0	99.8	90.3	94.7	95.5	**100.0**	**97.1**	**100.0**
035_power_drill	71.5	22.8	**98.2**	**99.6**	83.1	64.2	92.4	97.1	96.0	98.7
036_wood_block	68.1	18.2	87.6	80.2	81.7	76.0	85.5	93.4	**89.7**	**94.6**
037_scissors	76.7	35.9	91.7	95.6	83.6	75.1	96.4	**100.0**	**95.2**	**100.0**
040_large_marker	87.9	80.4	97.2	99.7	91.2	88.6	94.7	99.2	**97.5**	**100.0**
051_large_clamp	65.9	50.0	**75.2**	74.9	70.5	77.1	71.6	78.5	72.9	**79.2**
052_extra_large_clamp	60.4	20.1	64.4	48.8	66.4	50.2	69.0	69.5	**69.8**	**76.3**
061_foam_brick	91.8	100.0	**97.2**	100.0	92.1	100.0	92.4	100.0	92.5	100.0
MEAN	83.9	74.1	93.0	93.2	88.2	87.9	91.2	95.3	**93.1**	**96.8**

LineMOD Dataset: This is a standard dataset widely used for object pose estimation, featuring images of objects with low texture and precise pose annotations.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

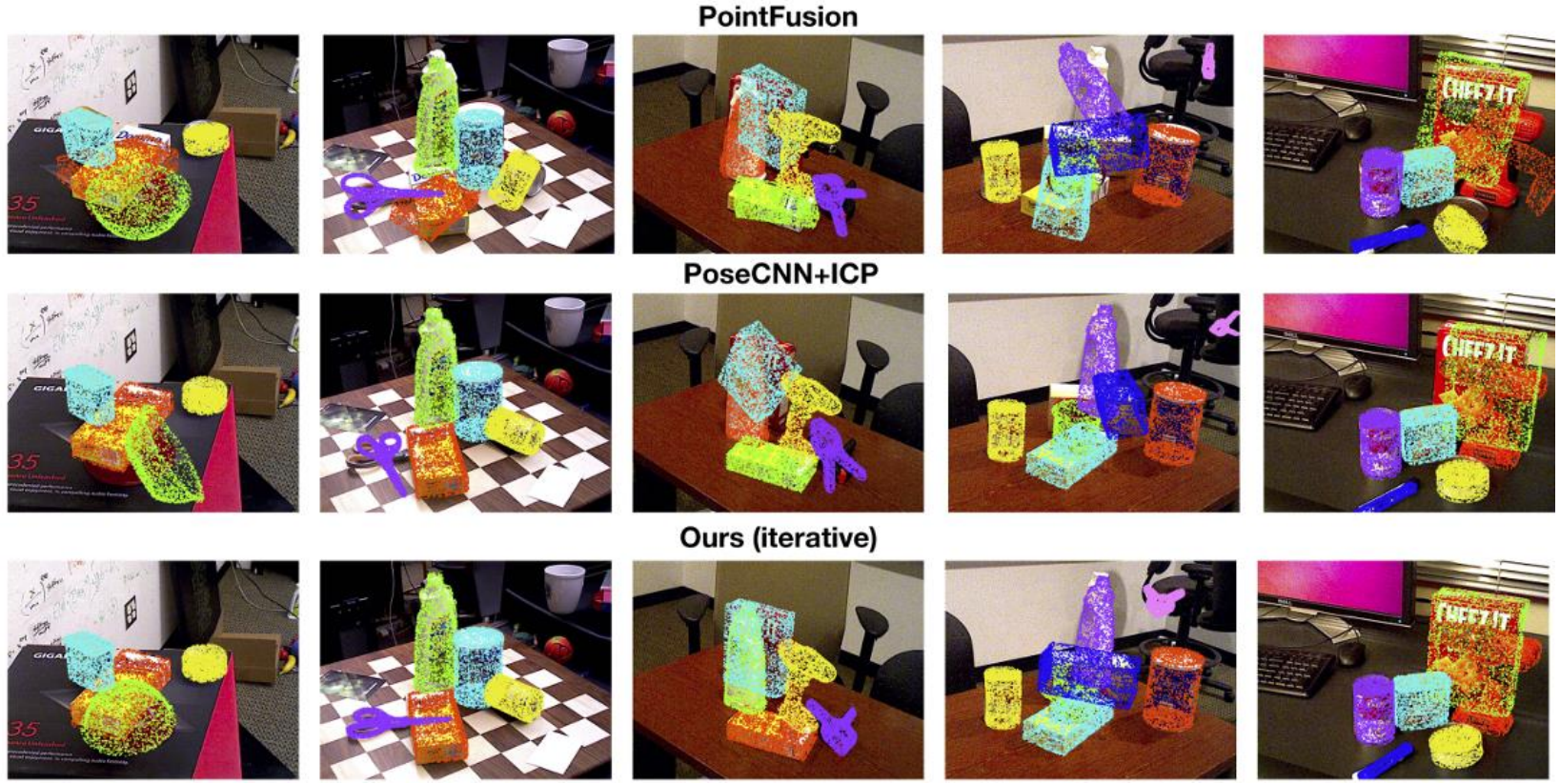
Comparison with Existing Technologies

DenseFusion's performance is compared with other advanced technologies such as PoseCNN and PointFusion, analyzing accuracy and real-time capabilities on standard datasets.

- **Occlusion Handling Capability**: The model's ability to handle partially occluded objects is specially analyzed by comparing the pose estimation accuracy when parts of objects are visible.
- **Processing Speed**: The time DenseFusion takes to estimate the pose of a single object is measured, demonstrating its suitability for real-time applications.
- **Impact of Iterative Refinement**: The gradual improvement in pose estimation accuracy through iterative refinement steps is observed, validating the contribution of this process to overall performance.

Comparison with Existing Technologies

- 【DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion】 https://www.bilibili.com/video/BV1fg4y187um/?share_source=copy_web&vd_source=3fbda0f34adfb8a51c1f2203d2e770aa



AncoraSIR.com

Real-World Application Testing

- **Robotic Grasping Experiment**:
  - **Experimental Setup**: DenseFusion is deployed on a real robotic platform, where the robot is used to locate and grasp objects.
  - **Task Execution**: The robot attempts to grasp objects placed in various positions and orientations, recording the success rate to verify the accuracy and practicality of the pose estimation.
- **Environmental Adaptability**: DenseFusion is tested under different background and lighting conditions to assess its robustness across diverse environments.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Thanks !

AncoraSIR.com