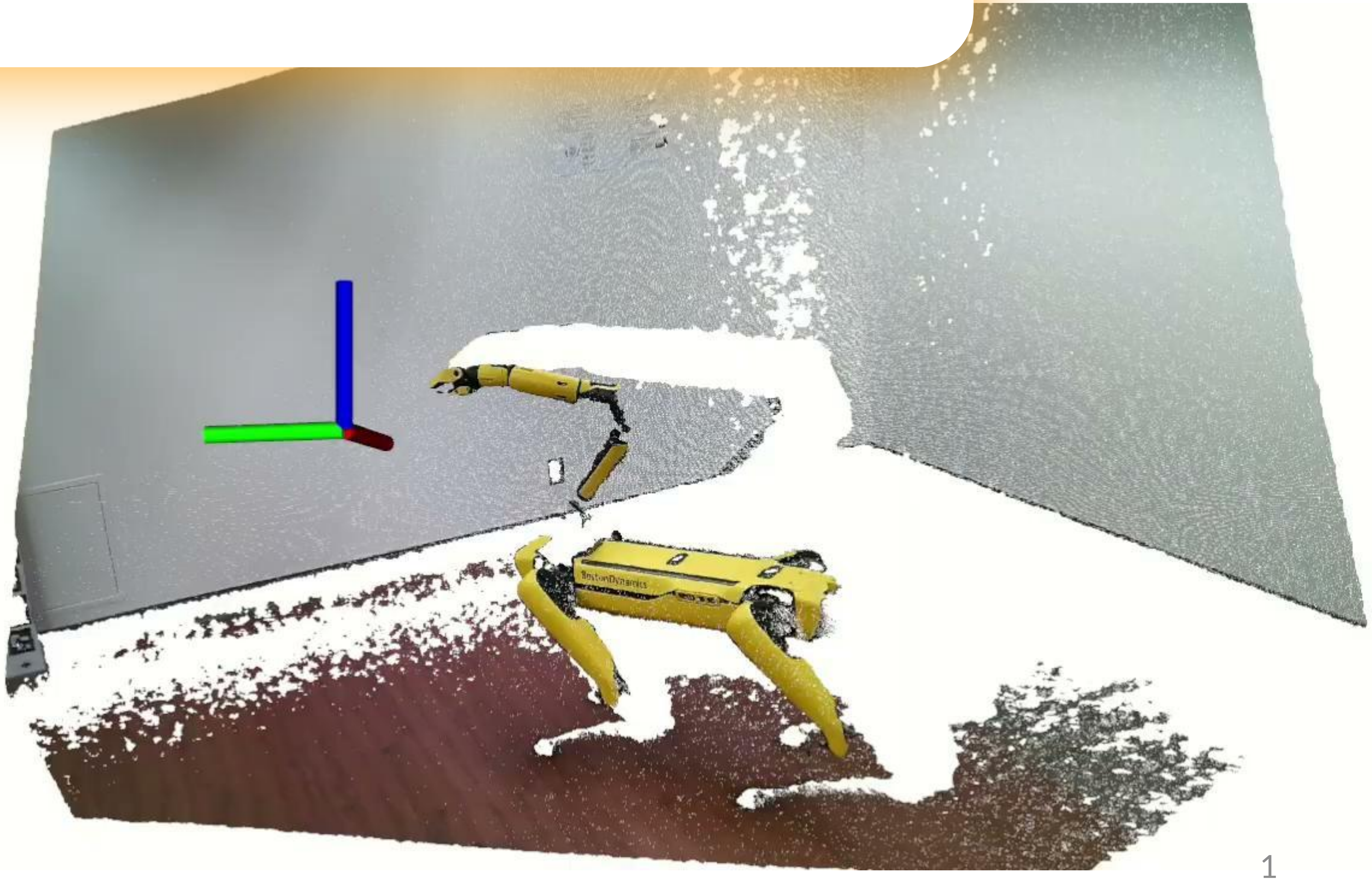
DeepRob

[Student] Lecture 1

by Tzu-Hsien Lee, Rammesh Adhav Saravanan , Fidan Mahmudova

RGB-D Networks and Manipulation

University of Minnesota



P4 is released - Due Nov 13th

- Instructions available on the webpage
  - Here: https://rpm-lab.github.io/CSCI5980-F24-DeepRob/projects/project4/
  - Uses [PROPS Pose Estimation Dataset]()
- Implement PoseCNN
- Autograder is available.
- Due Wednesday, November 13th, 11:59 PM CT

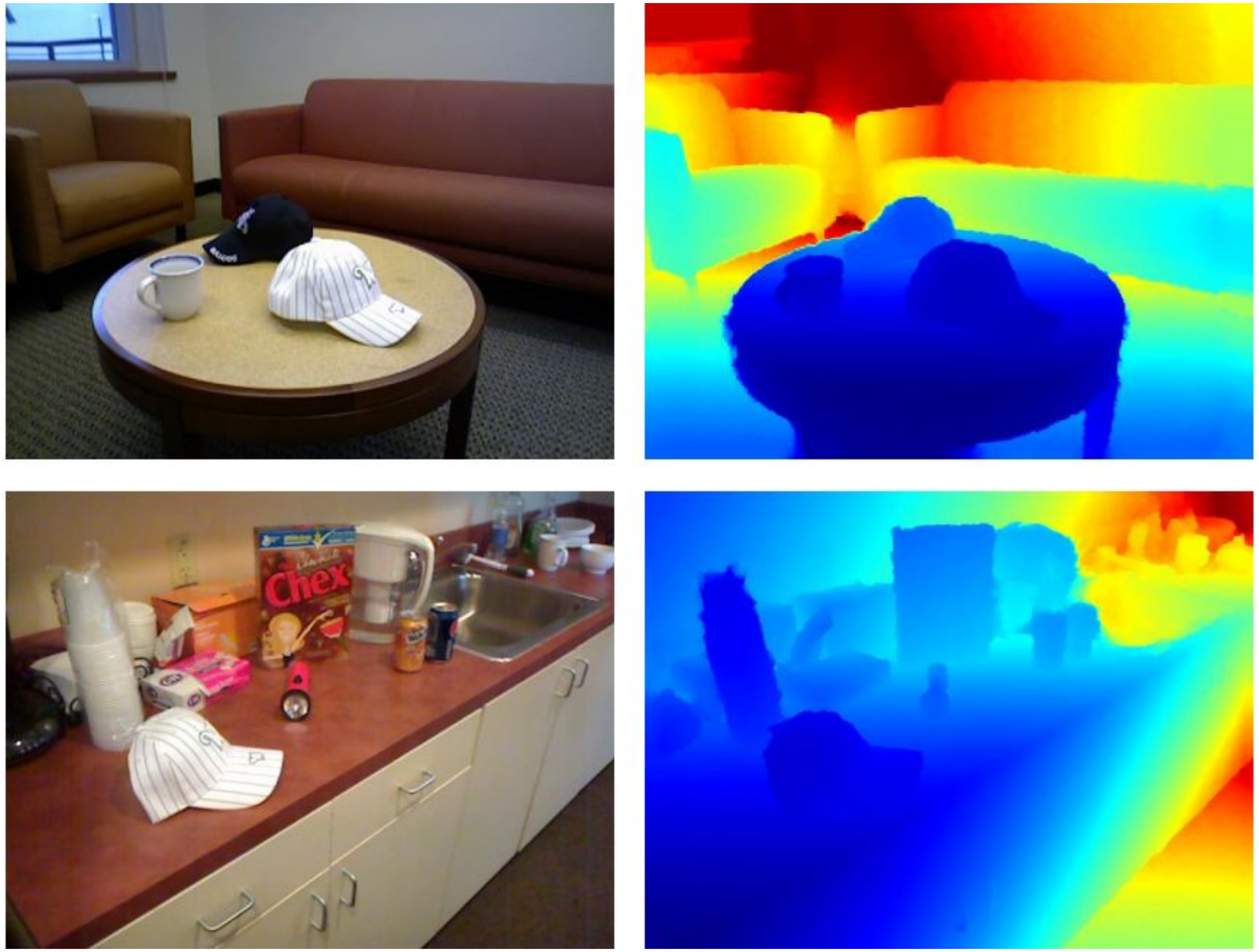
Team task - Data viz - Due Nov 6th

As a first step to narrowing down your final project, I want you to start researching the data **X** that will be used.

Please upload a ***video*** showing all the data streams in your project that will be used to train the deep-learning models **y=f(X)**.

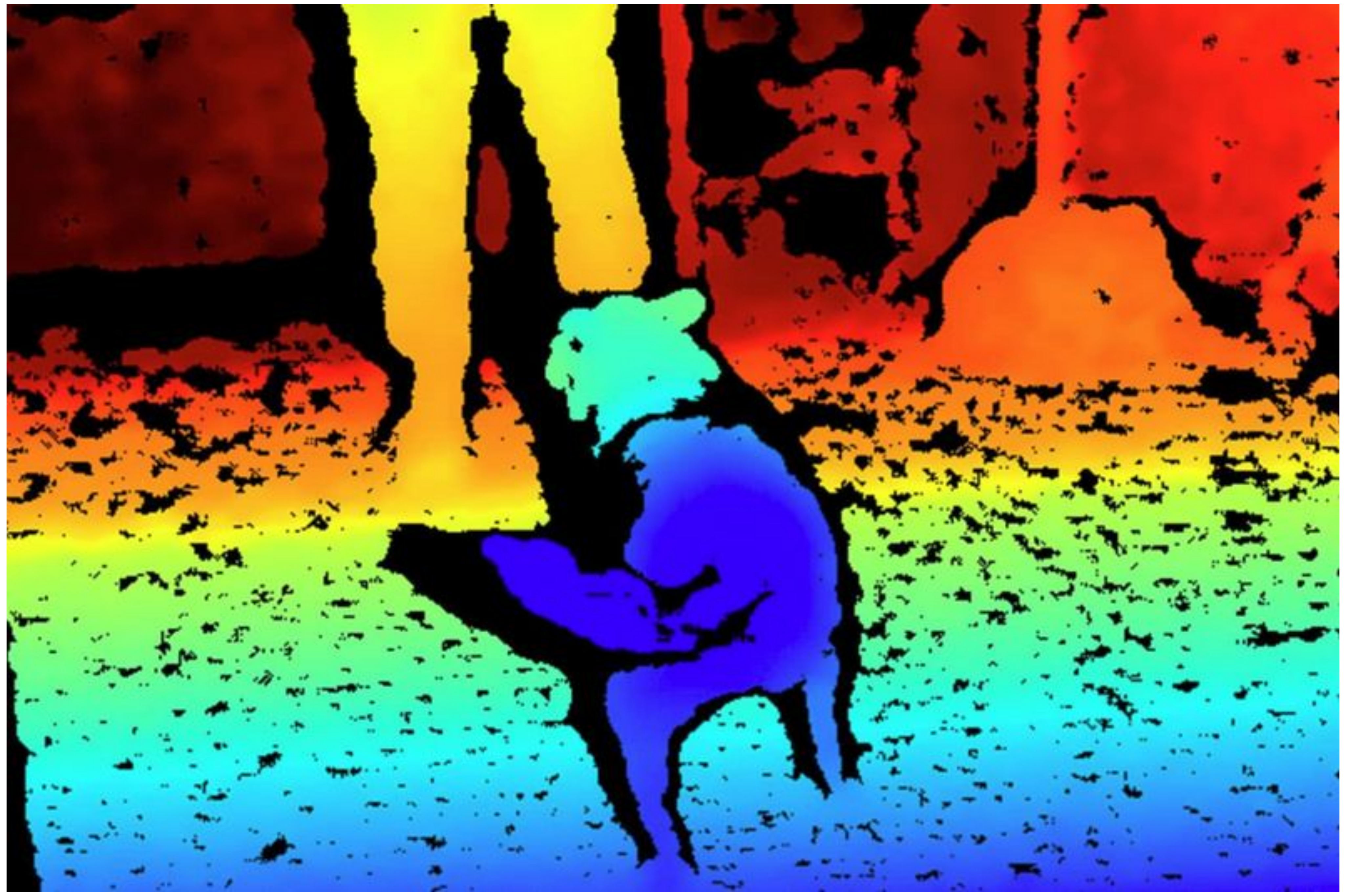
- If you use an existing dataset for your project, I expect your video to contain samples of these sensor observations and the correct labels.
- If you are using a simulator, I expect you to collect the data from the simulator and then show the data streams that will be used for training your model.
- The same goes for real-world experiments as well.

What is RGB-D data?



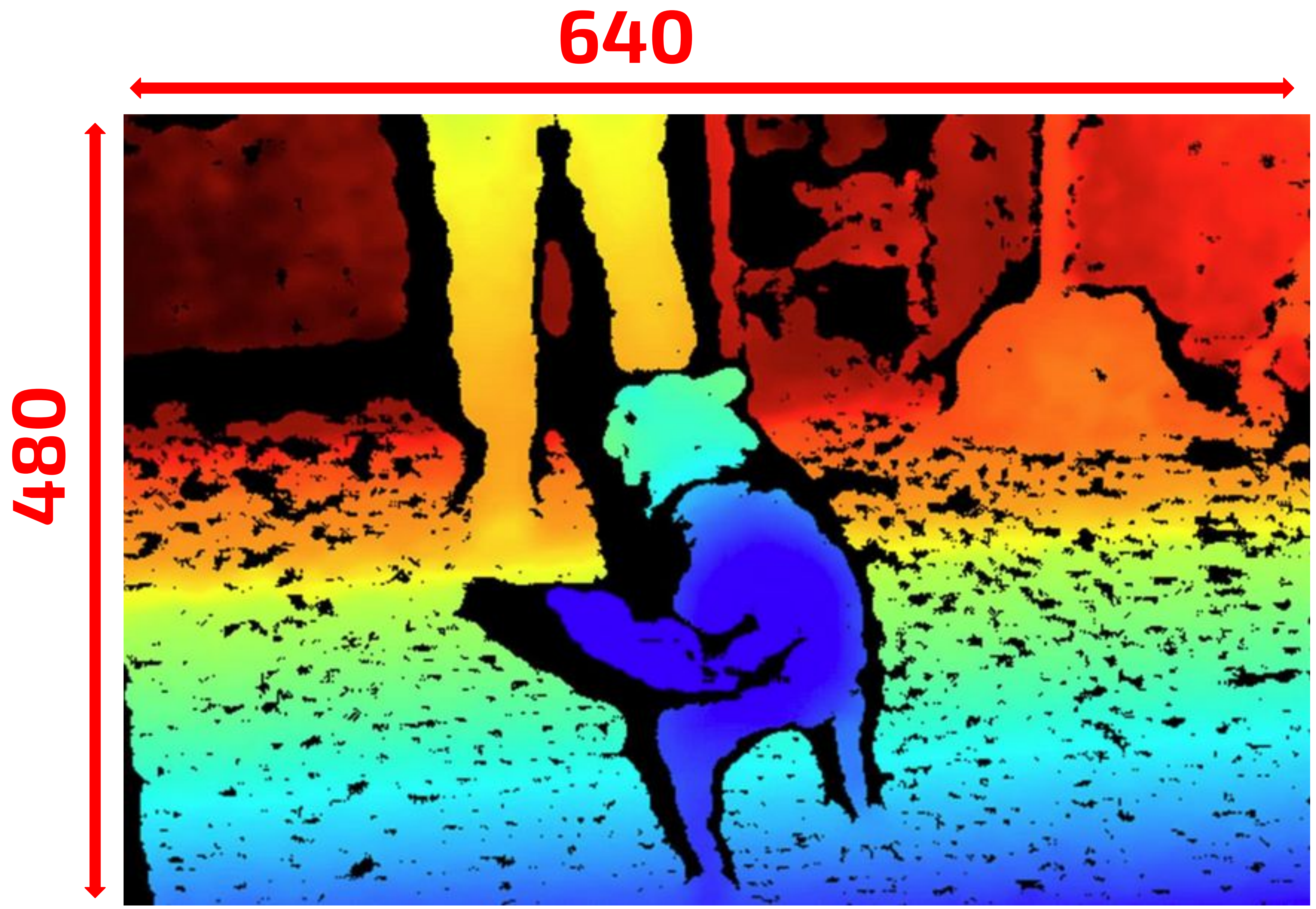
Source: Fu, H., Xu, D., Lin, S., & Liu, J. *Object-based RGBD Image Co-segmentation with Mutex Constraint.*

Organized and Unorganized point clouds



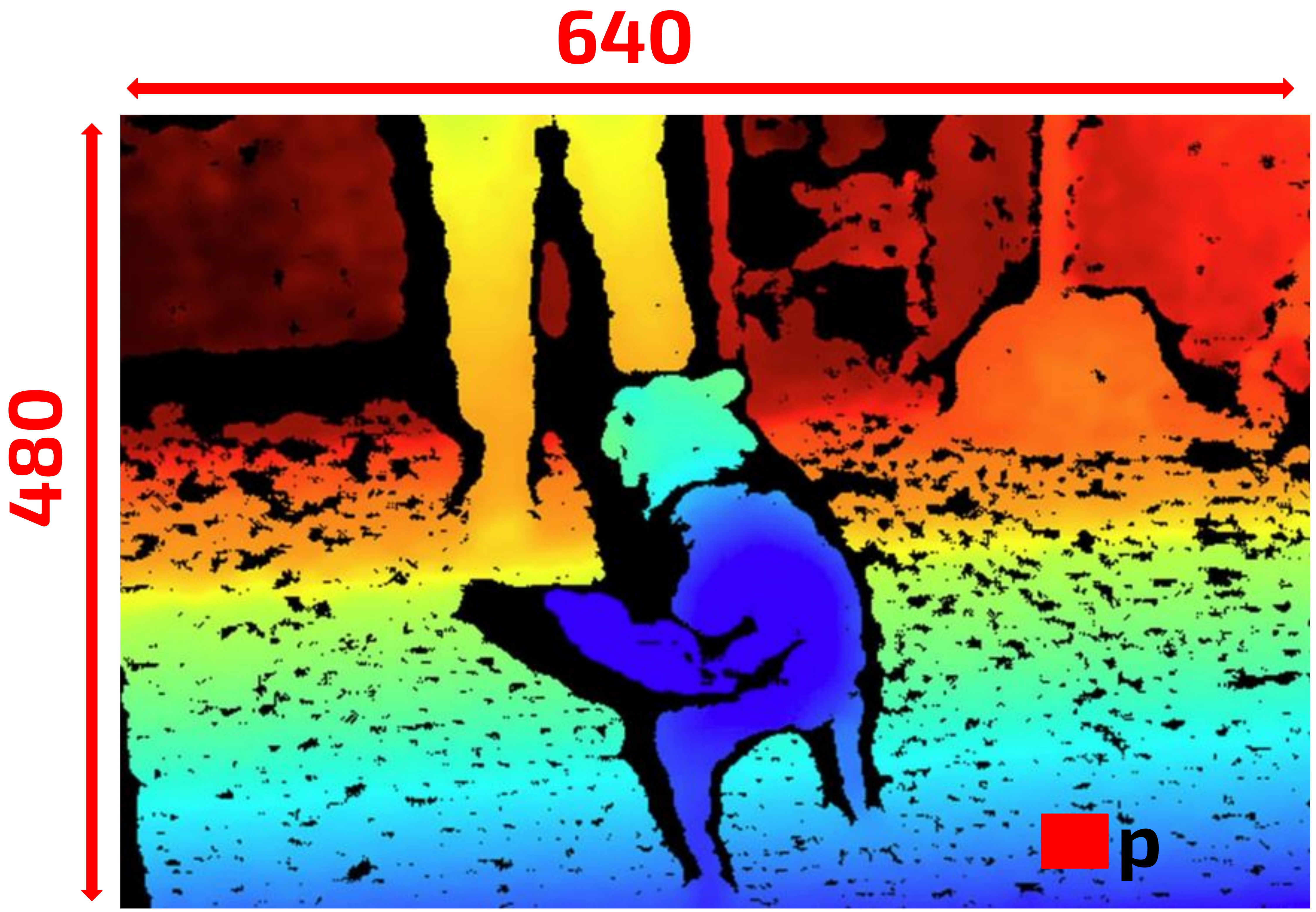
Source : https://www.ac3filter.net/what-is-a-stereo/

Organized and Unorganized point clouds



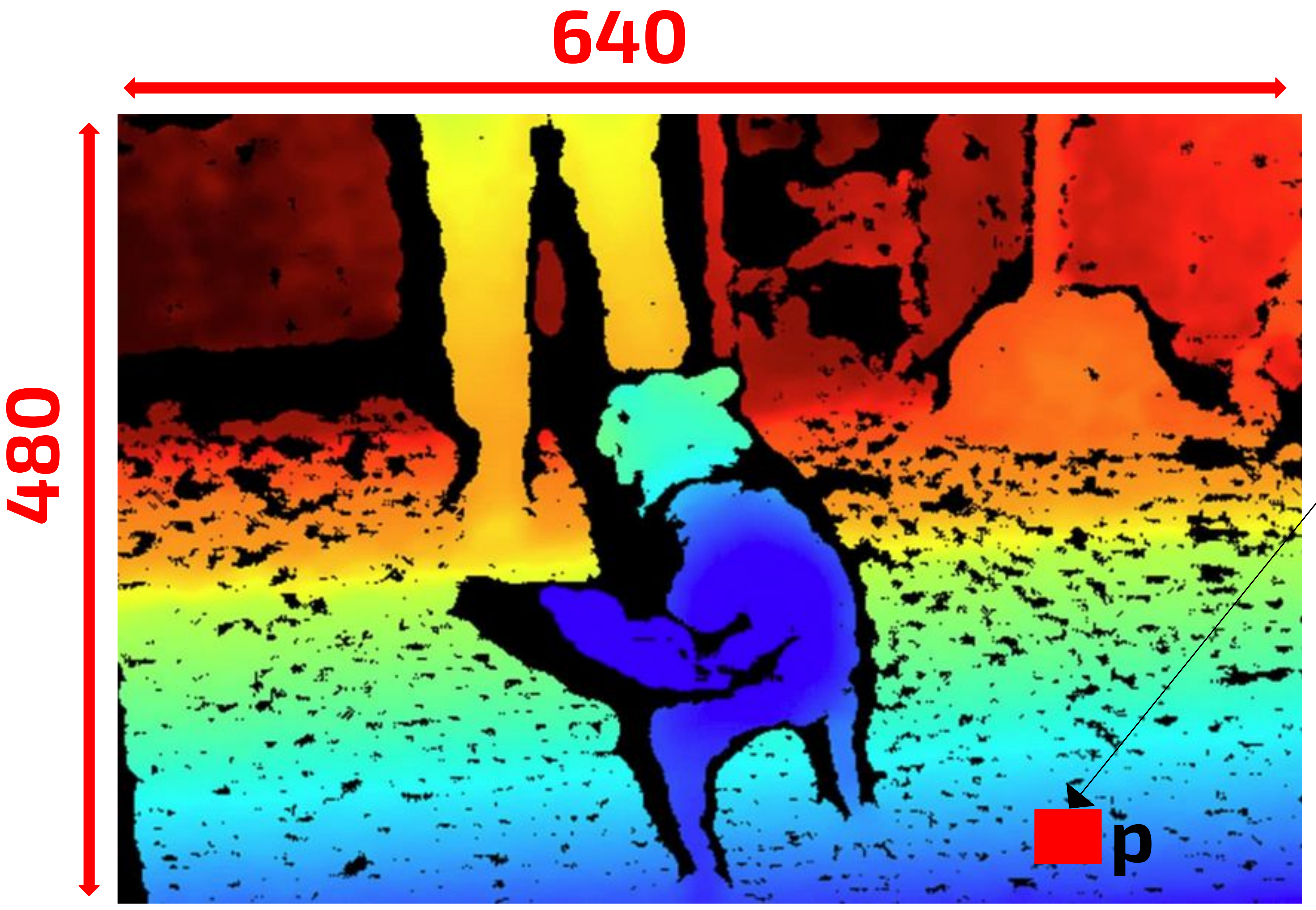
Source : https://www.ac3filter.net/what-is-a-stereo/

Organized and Unorganized point clouds



Source : https://www.ac3filter.net/what-is-a-stereo/

Organized and Unorganized point clouds



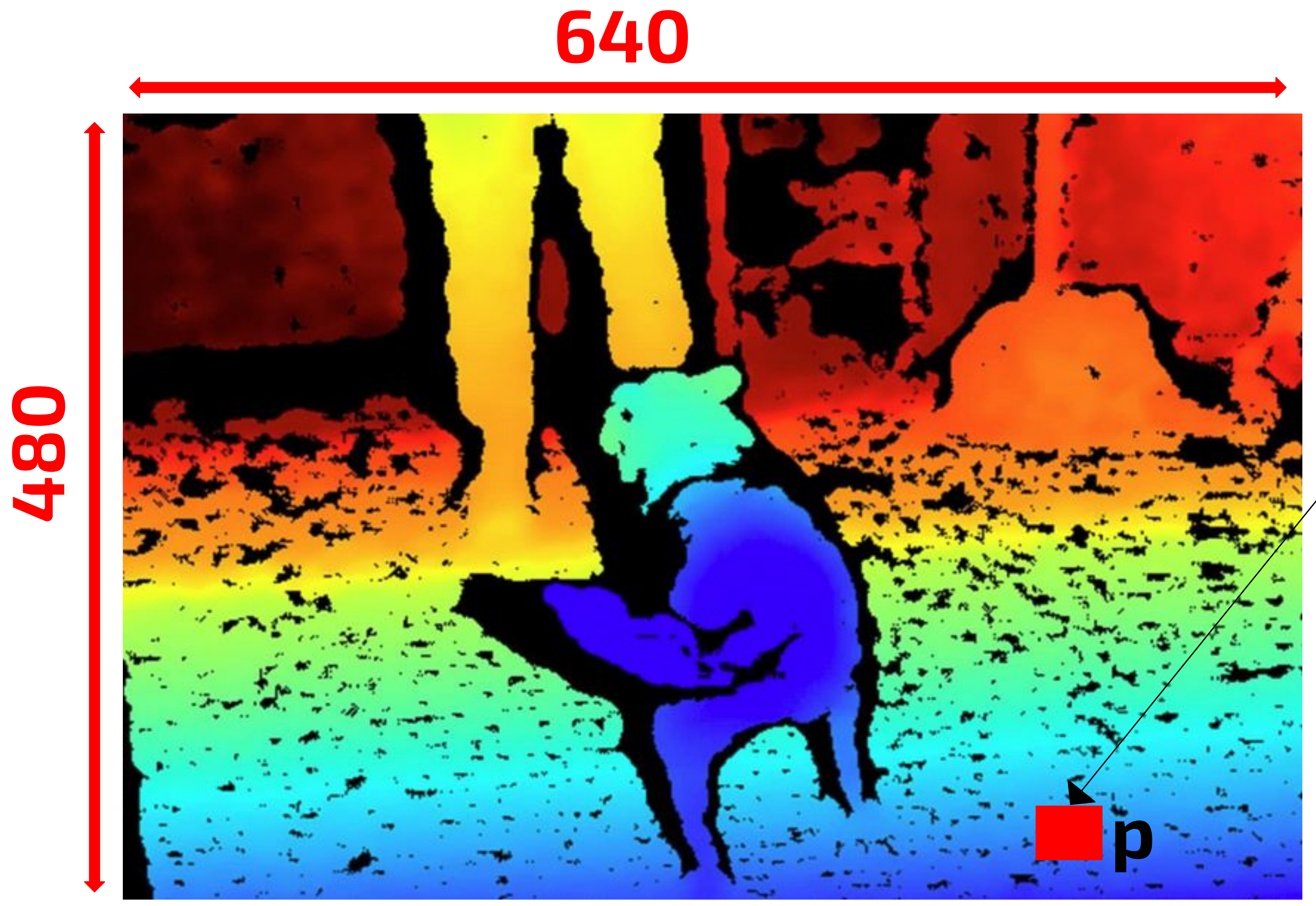
Source : https://www.ac3filter.net/what-is-a-stereo/

p = (620 , 5)

idx = 640 * 5 + 620

idx = 3820

Organized and Unorganized point clouds



Source : https://www.ac3filter.net/what-is-a-stereo/

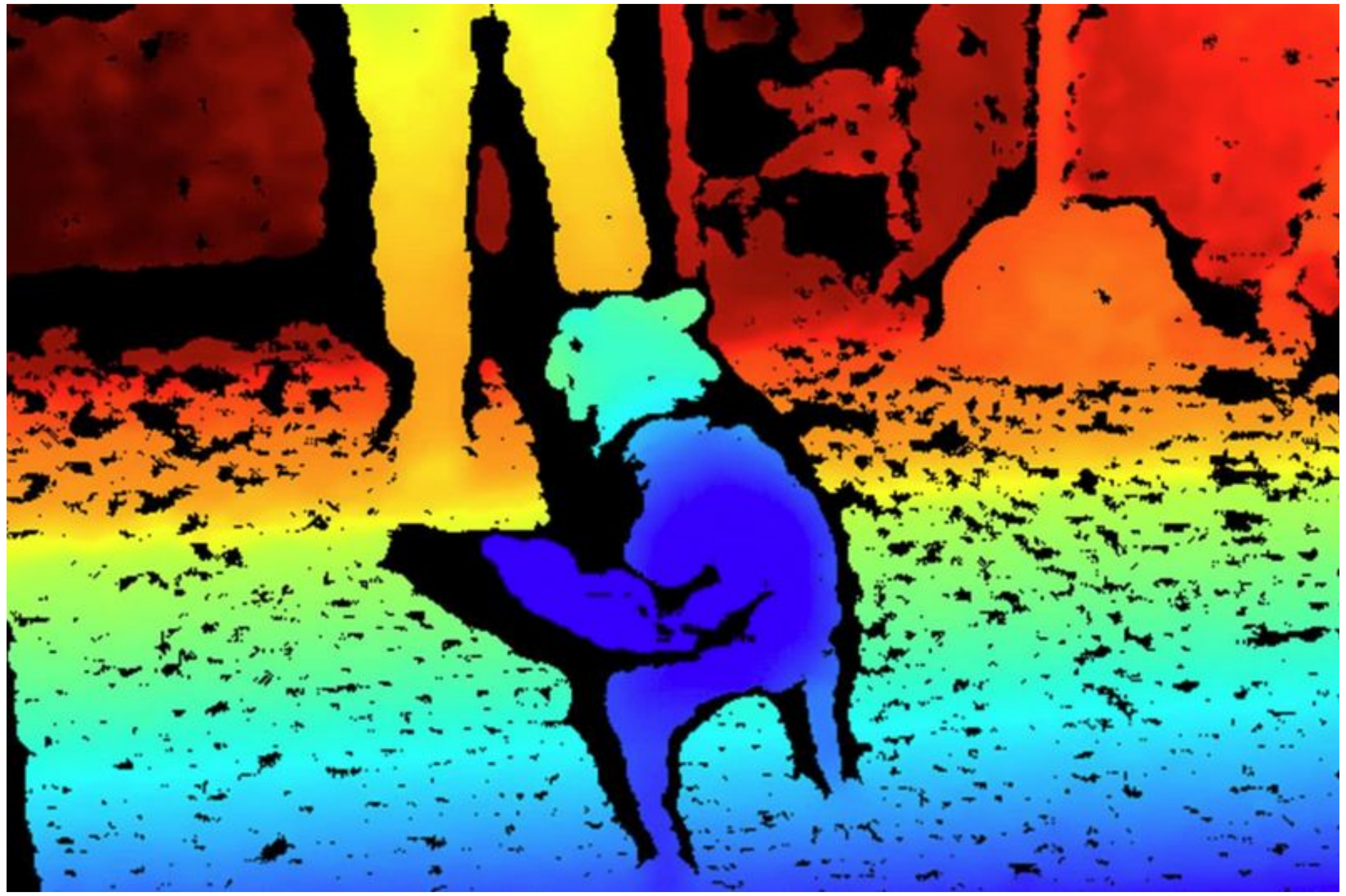
p = (620 , 5)

idx = 640 * 5 + 620

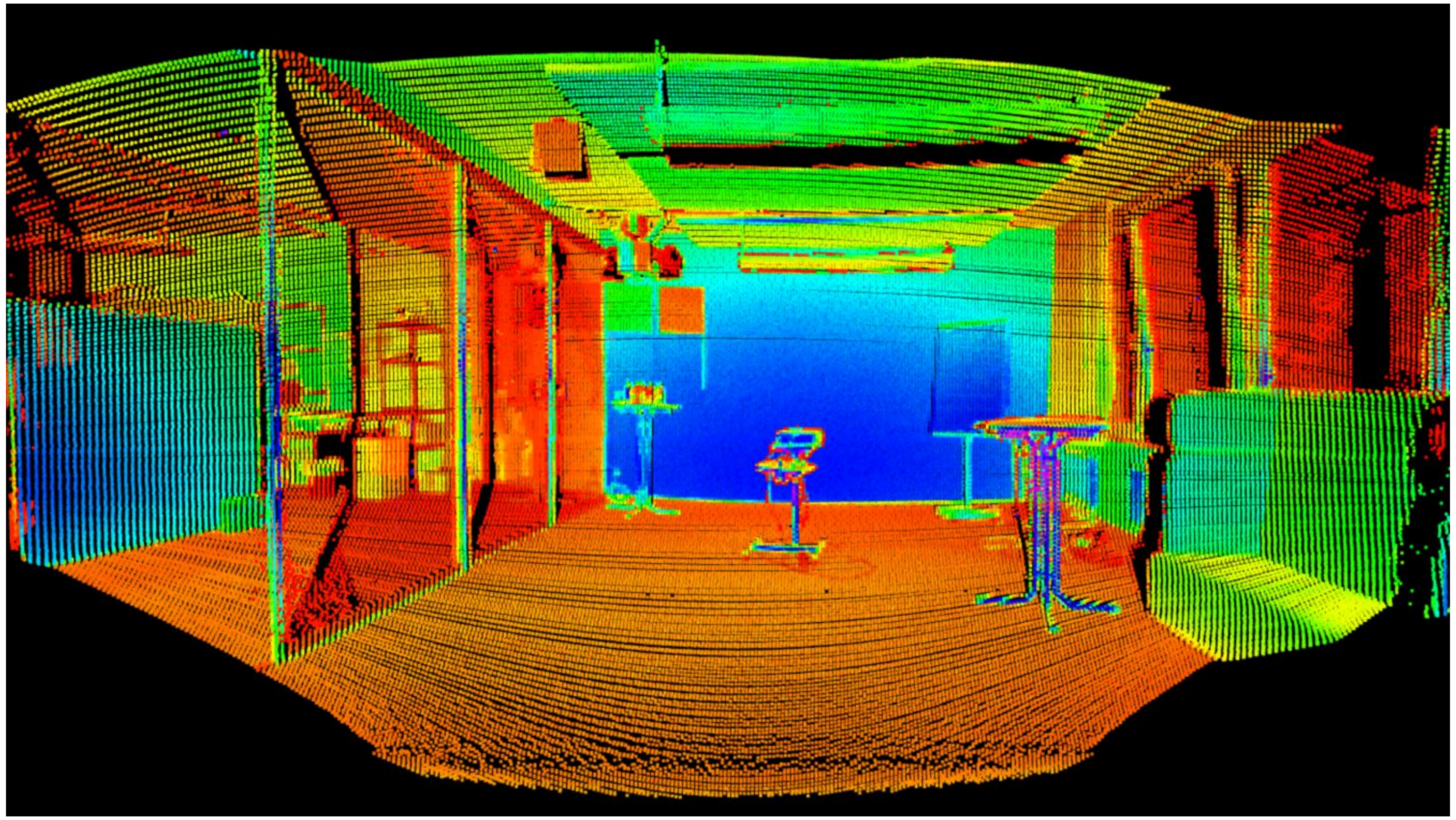
idx = 3820

point_cloud[idx] =[X, Y, Z , R, G, B]

Organized and Unorganized point clouds



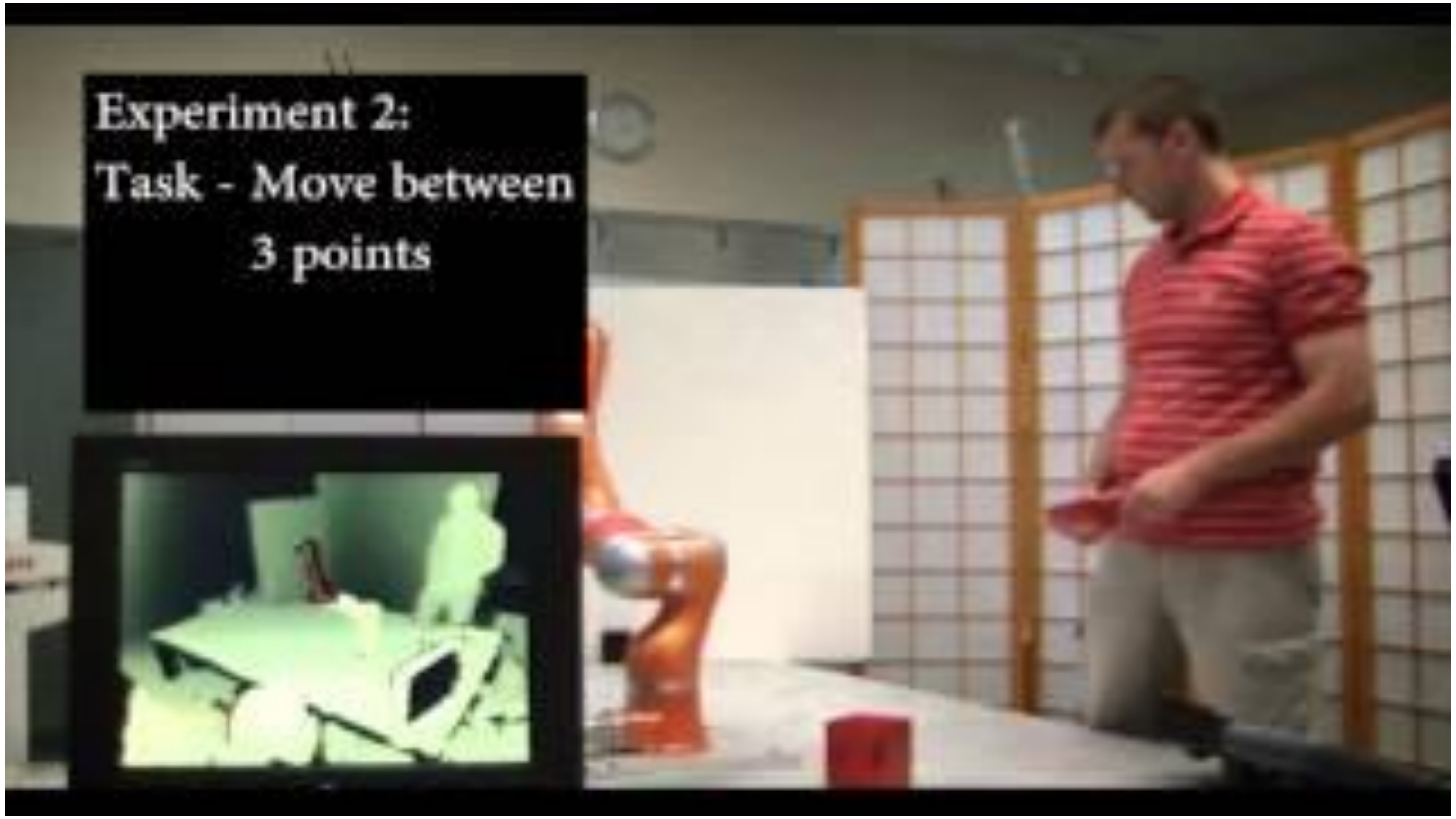
Source : https://www.ac3filter.net/what-is-a-stereo/



Source: :https://www.blickfeld.com/blog/understanding-lidar-specifications/

Why depth matters in manipulation?

1. *Safe and Strategic Movement Planning*



Source : Flacco, F., Kröger, T., De Luca, A., & Khatib, O. (2012). *A depth space approach to human-robot collision avoidance.*

Why depth matters in manipulation?

2. Accurate Object Grasping

Source : https://www.youtube.com/watch?v=ry0mqY5I-04

Foundations for RGB-D based Robot Grasp Manipulation: Traditional Techniques Before Deep Learning

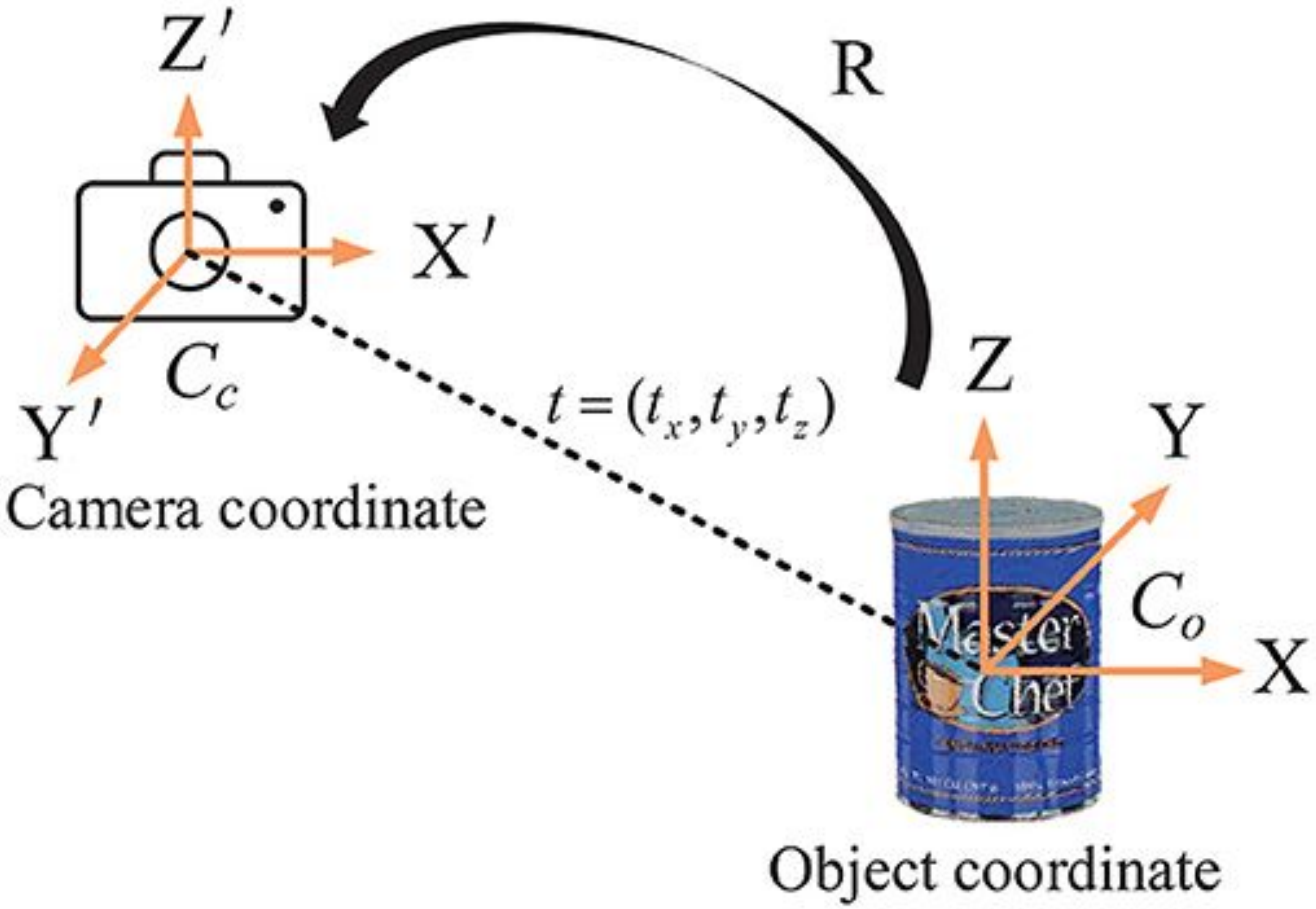
Foundations for RGB-D based Robot Grasp Manipulation:
Traditional Techniques Before Deep Learning
1. Traditional Pose Estimation Techniques

Foundations for RGB-D based Robot Grasp Manipulation :

Traditional Methods Before Deep Learning :

1. Traditional Pose Estimation Methods
2. Traditional Feature Extraction Methods

Pose Estimation



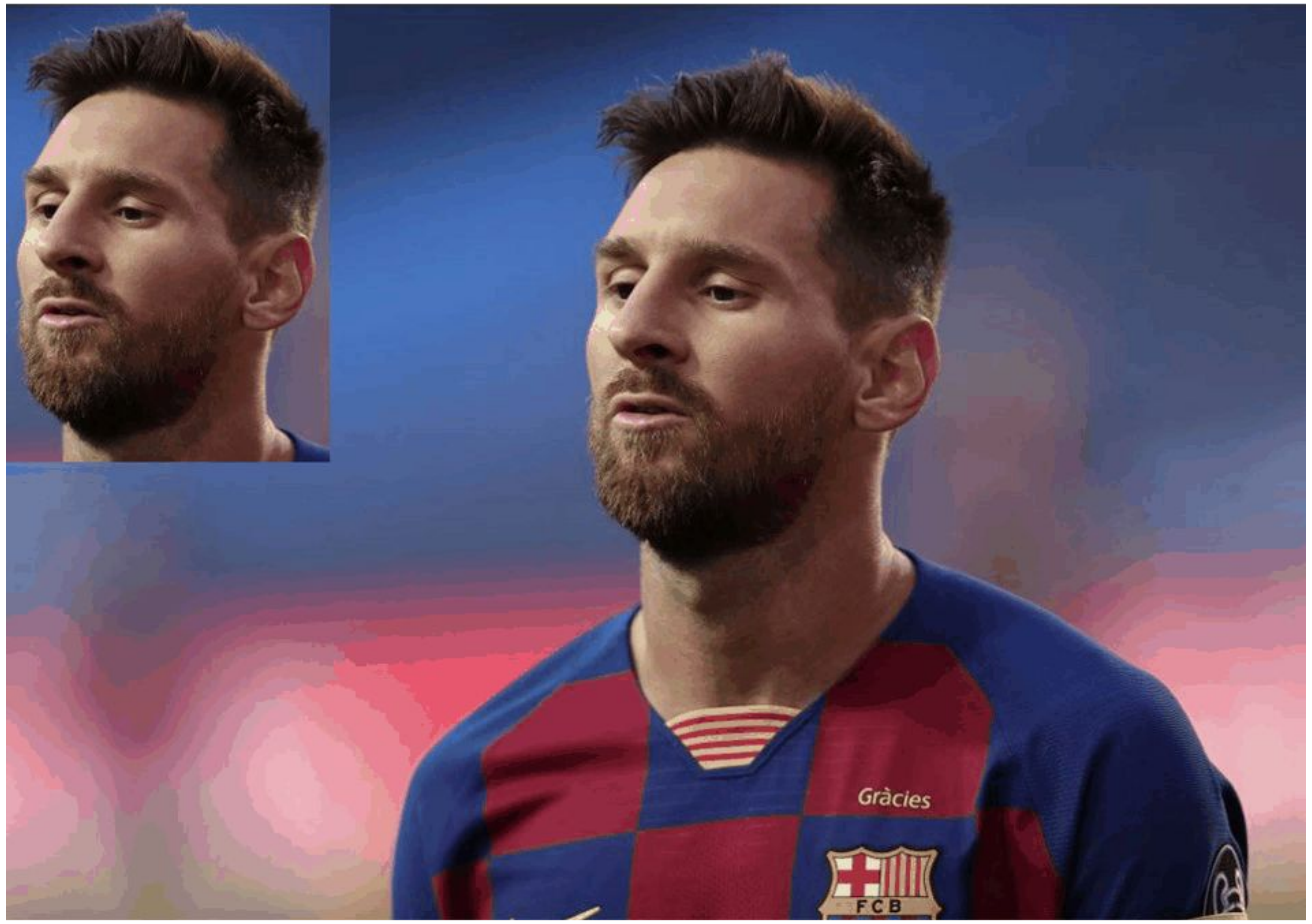
$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix}$$

$$\begin{bmatrix} P_{camera} \\ 1 \end{bmatrix} = T \begin{bmatrix} P_{object} \\ 1 \end{bmatrix}$$

Source : https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2020.616775/full

Traditional Pose Estimation Methods

- ***Template Matching using RGB data***



Source : https://datahacker.rs/014-template-matching-using-opencv-in-python/

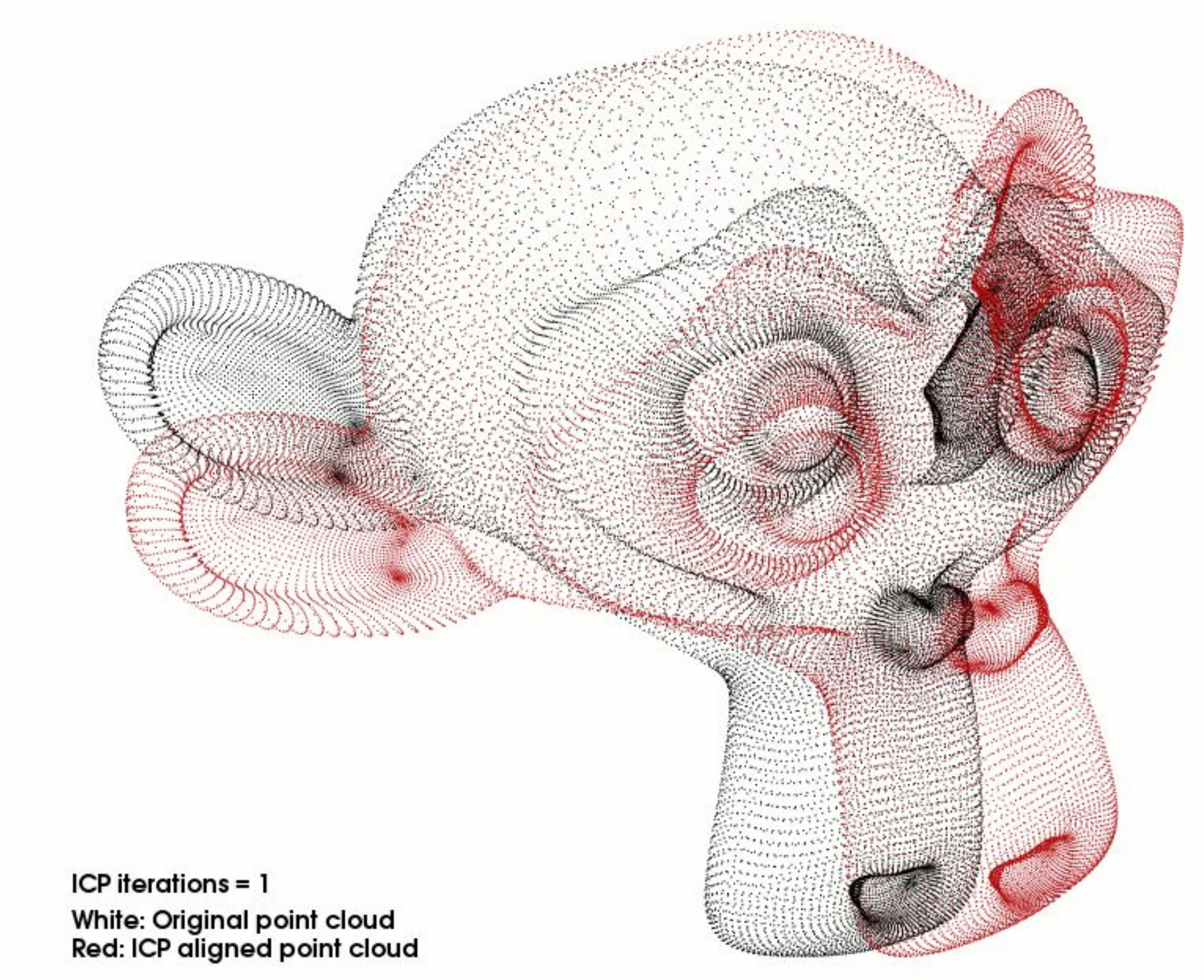
- ***Template Matching using RGB-D data***

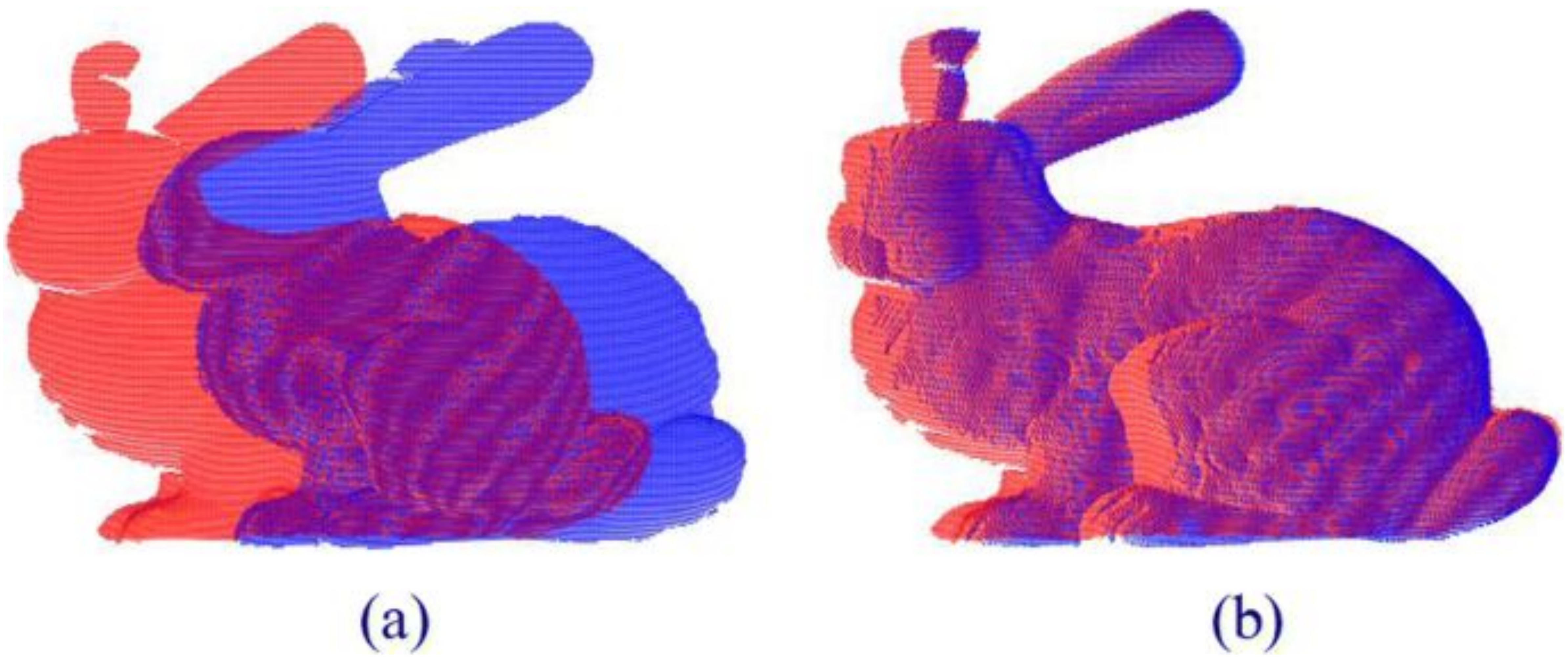
1. Enhanced Object Localization

2. Robust Matching Process

Traditional Pose Estimation Methods

- ***ICP (Iterative Closest Point)***



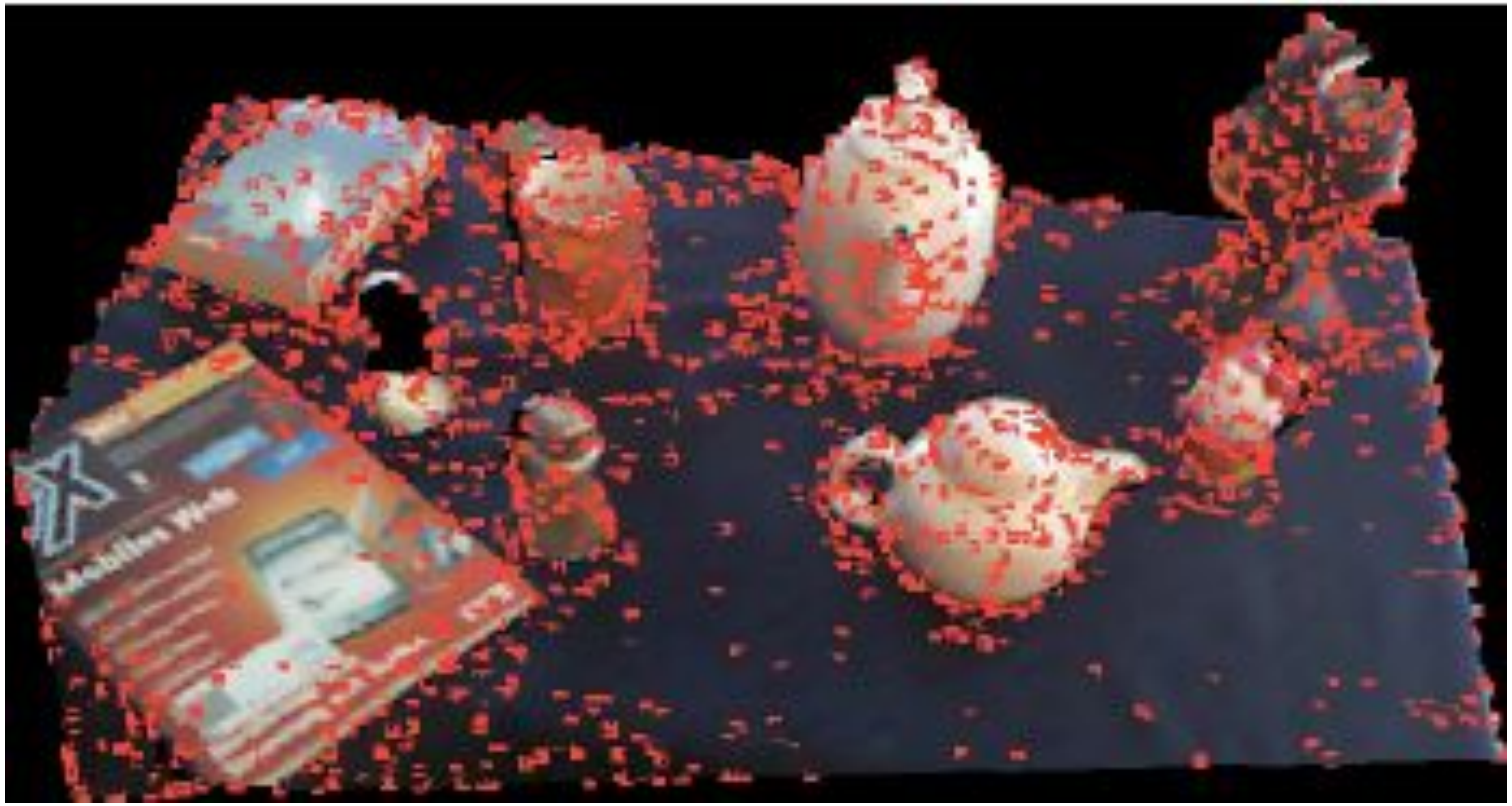


Source:https://unibe-cas-assignment.readthedocs.io/en/late st/assignment.registration.html

Source : Wan, T., Du, S., Xu, Y., Xu, G., Li, Z., Chen, B., & Gao, Y. (2019). RGB-D point cloud registration via infrared and color camera.

Traditional Feature Extraction Methods

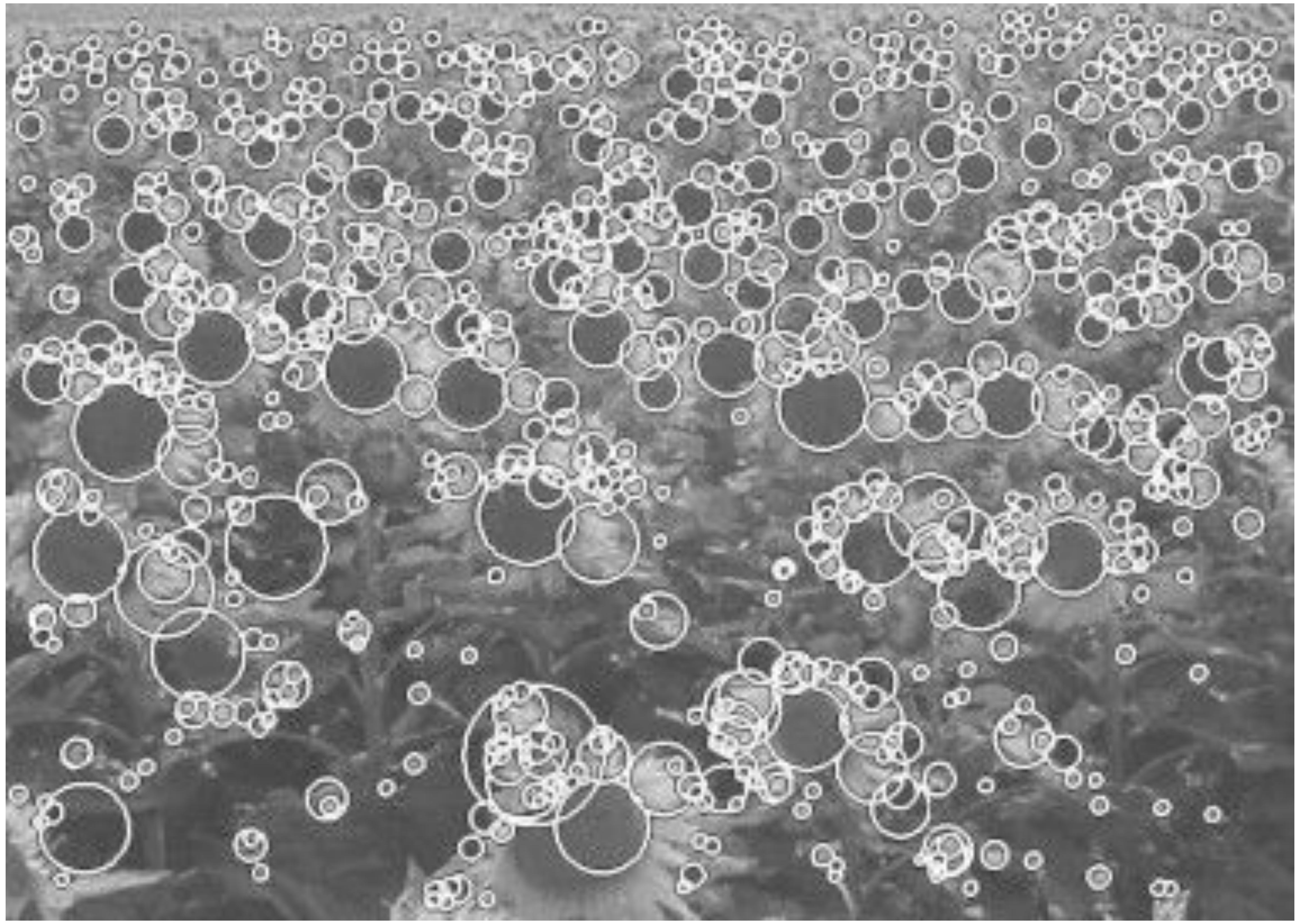
- ***3D SIFT (Scale Invariant Feature Transformation)***



Source: "Comparison of 3D Interest Point Detectors and Descriptors for Point Cloud Fusion," *ISPRS Annals of the Photogrammetry, Remote Sensing, and Spatial Information Sciences*, vol. II-3, Sept. 2014,

Traditional Feature Extraction Methods

- ***3D SURF (Speeded-Up Robust Features)***



Source: Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, 2008,

Why to move to Deep Learning?

1. Sensitivity to Variations

Why to move to Deep Learning?

1. Sensitivity to Variations

2. Handcrafted Features

Why to move to Deep Learning?

1. Sensitivity to Variations

2. Handcrafted Features

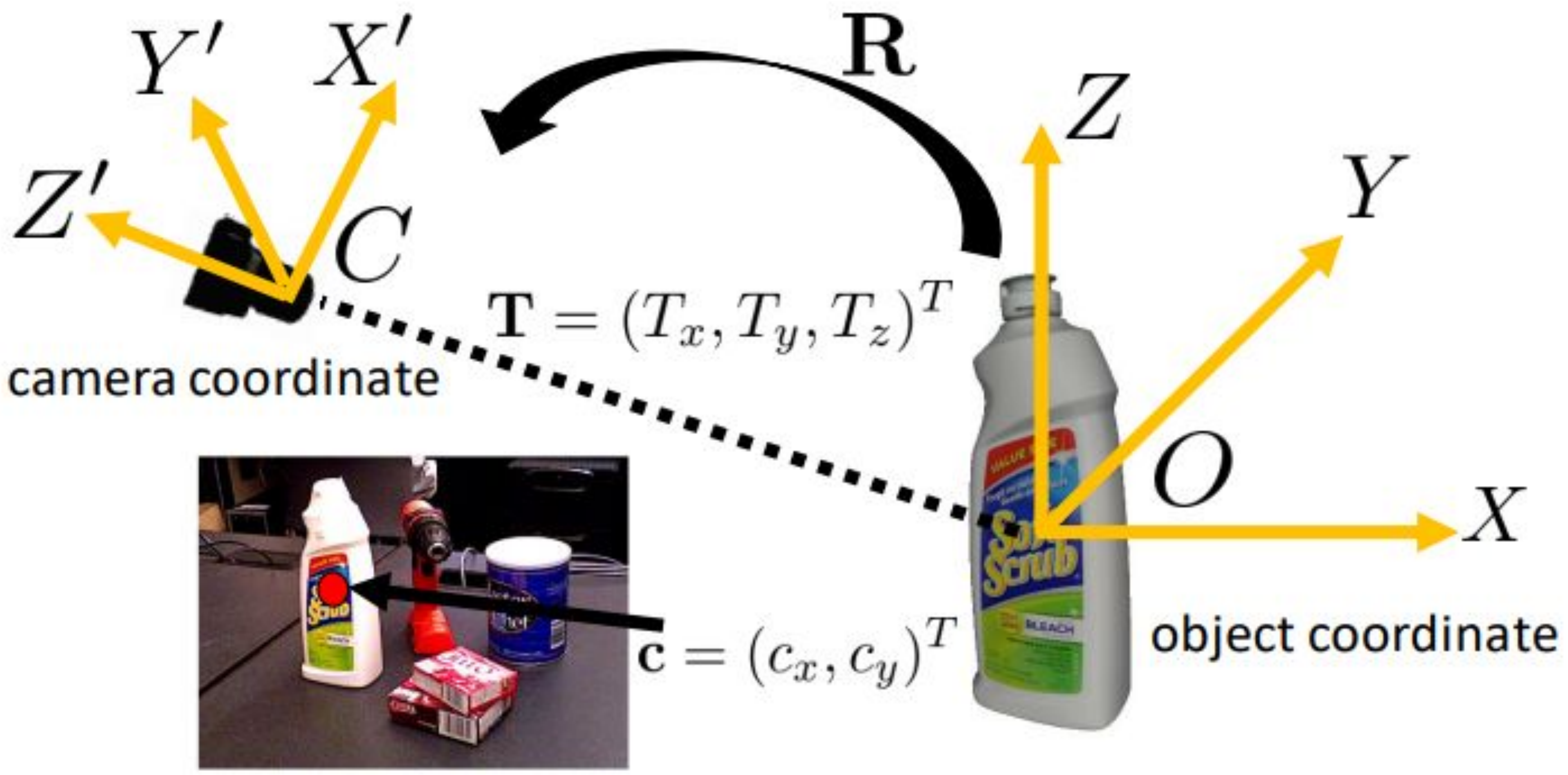
3. Complexity in Scaling

Let’s talk about deep neural network architectures for RGB-D with some examples:

- PoseCNN
- RGB-D Salient Object Detection

PoseCNN

Object Pose Estimation



1. 3D Translation

2. 3D Rotation

Transformation from object to camera coordinate system

Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

Limitations of Existing Works

1. Feature - based methods:

   a. Texture-less objects

Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

Limitations of Existing Works

2. Template - based methods:

   a. Occlusion of objects

Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

Limitations of Existing Works

3. Image pixel to 3D coordinates mapping:

   a. Symmetrical objects.

Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

Objectives of PoseCNN:

1. Develop a CNN-based 6D Pose Estimation Model Robust to Occlusions
2. Collect a Large-Scale RGB-D Dataset with pose annotation for Model Training
3. Define a Training Loss Function for Symmetrical Objects

CNN-based 6D Pose Estimation Model

1) PoseCNN:

- Takes only RGB image as input for estimating 3D translation and rotation

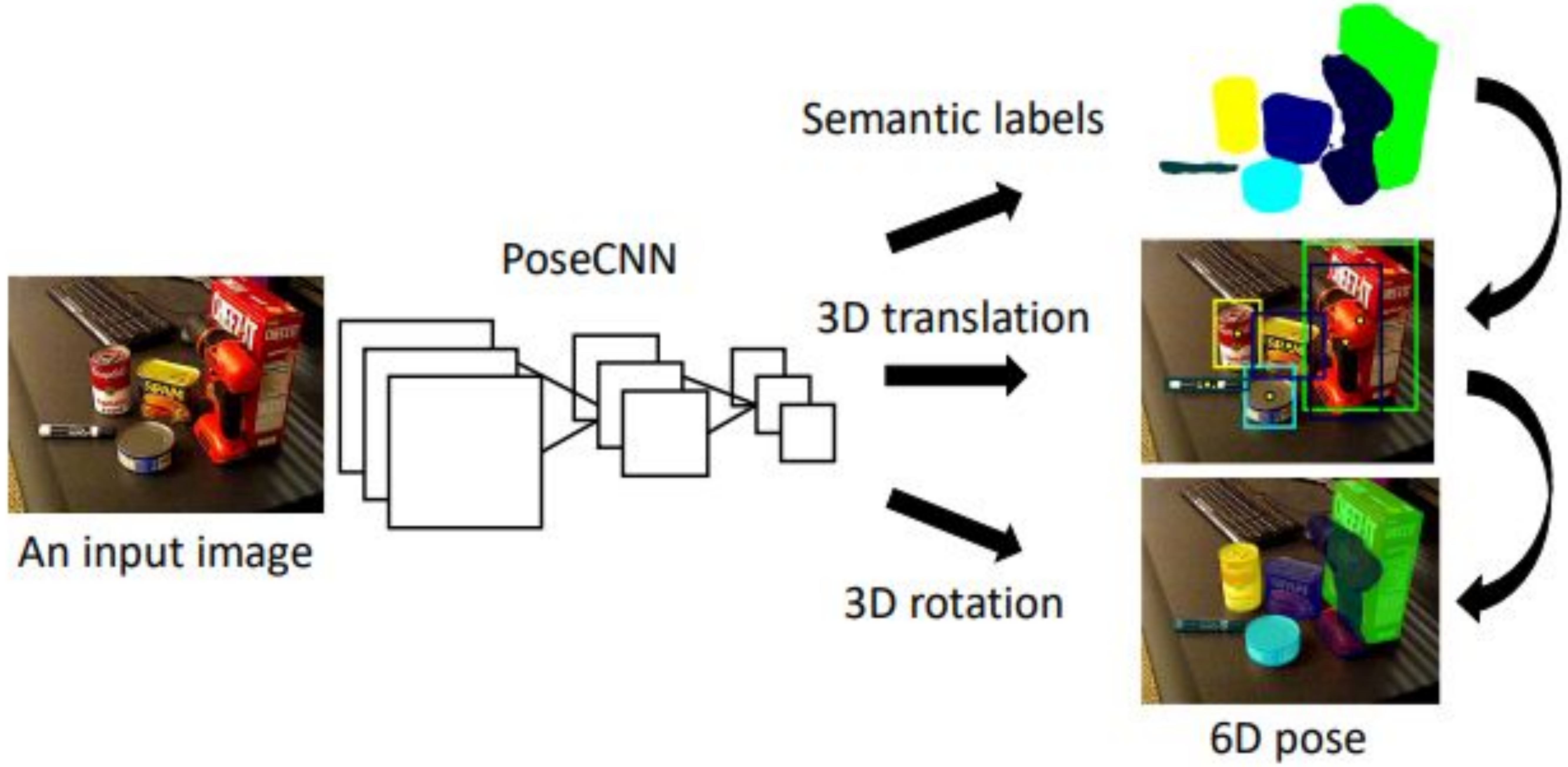


Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

CNN-based 6D Pose Estimation Model

2) PoseCNN + Iterative closest point(ICP):

- Takes both RGB image as input for estimating 3D translation and rotation and uses depth for refinement.

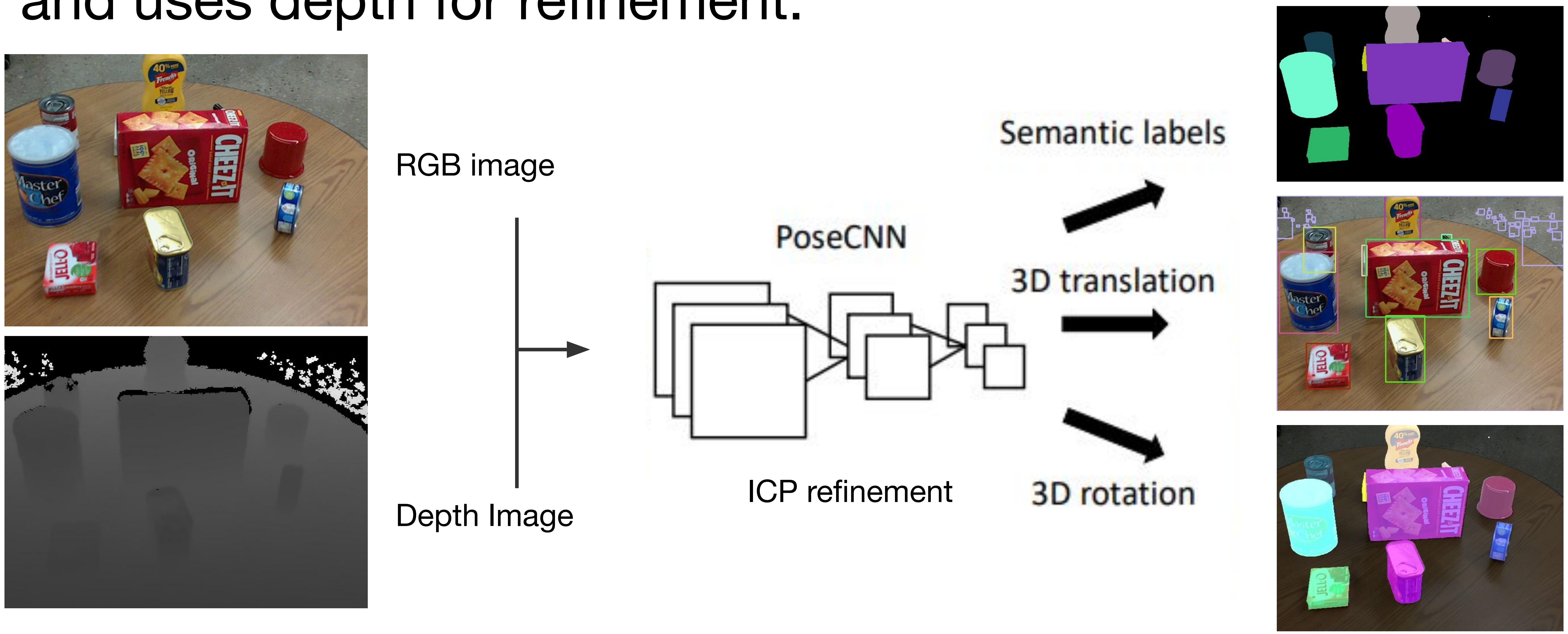


Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018. and RPM lab props dataset - https://drive.google.com/file/d/15rhwXhzHGKtBcxJAYMWJG7gN7BLLhyAq/view

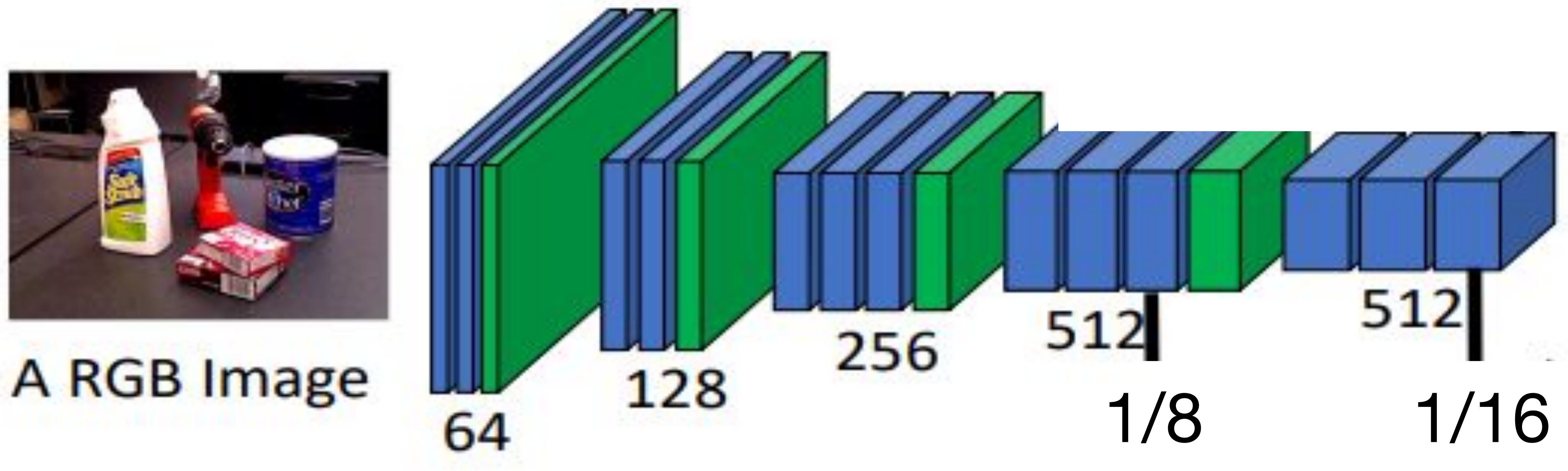
Dataset

The collected dataset has:

- 3D models (with set of 3D points)
- RGB images
- Depth images
- 6D pose annotations

Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018. and RPM lab props dataset - https://drive.google.com/file/d/15rhwXhzHGKtBcxJAYMWJG7gN7BLLhyAq/view

Feature extraction



1/8 1/16

original image size

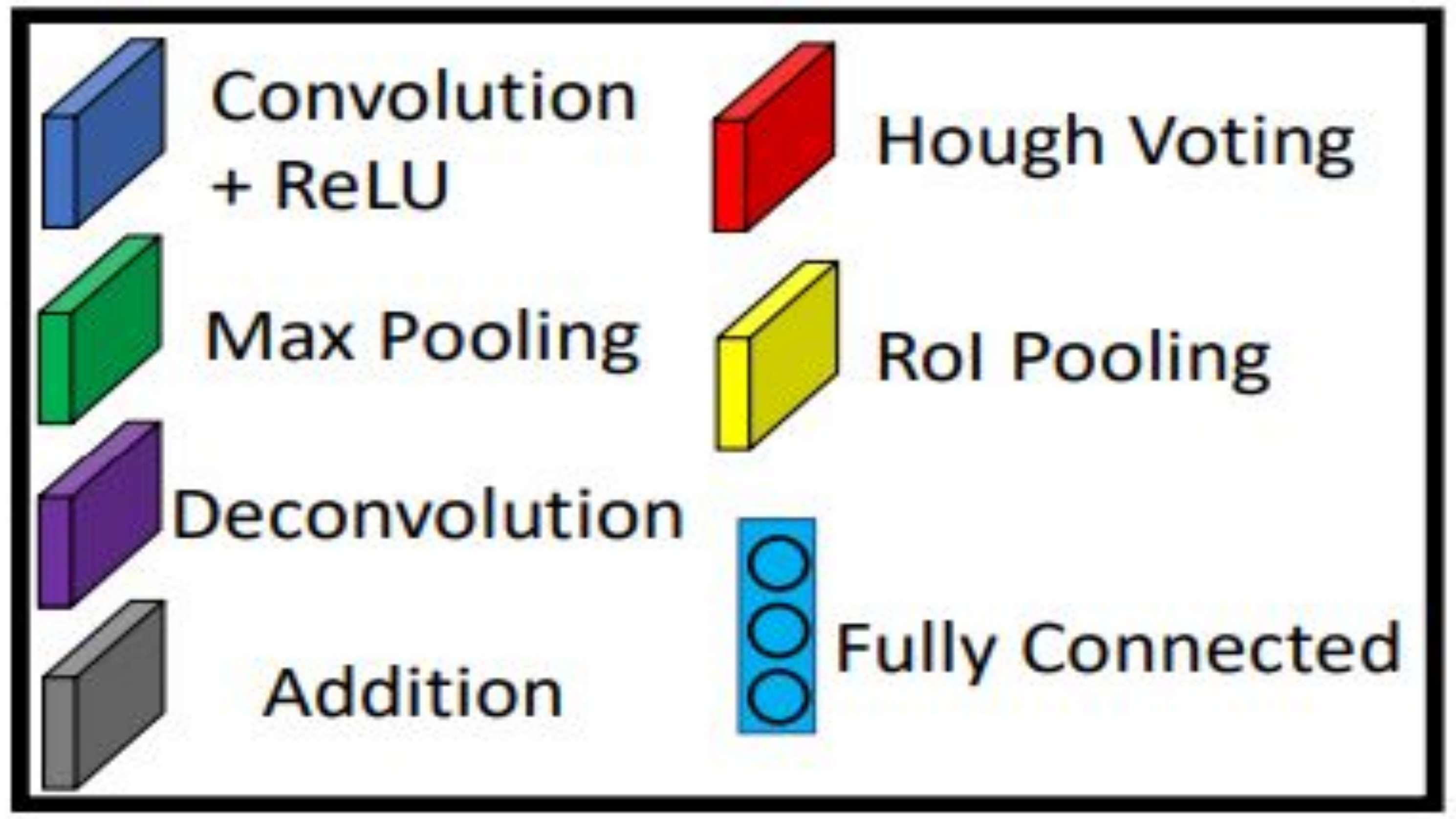
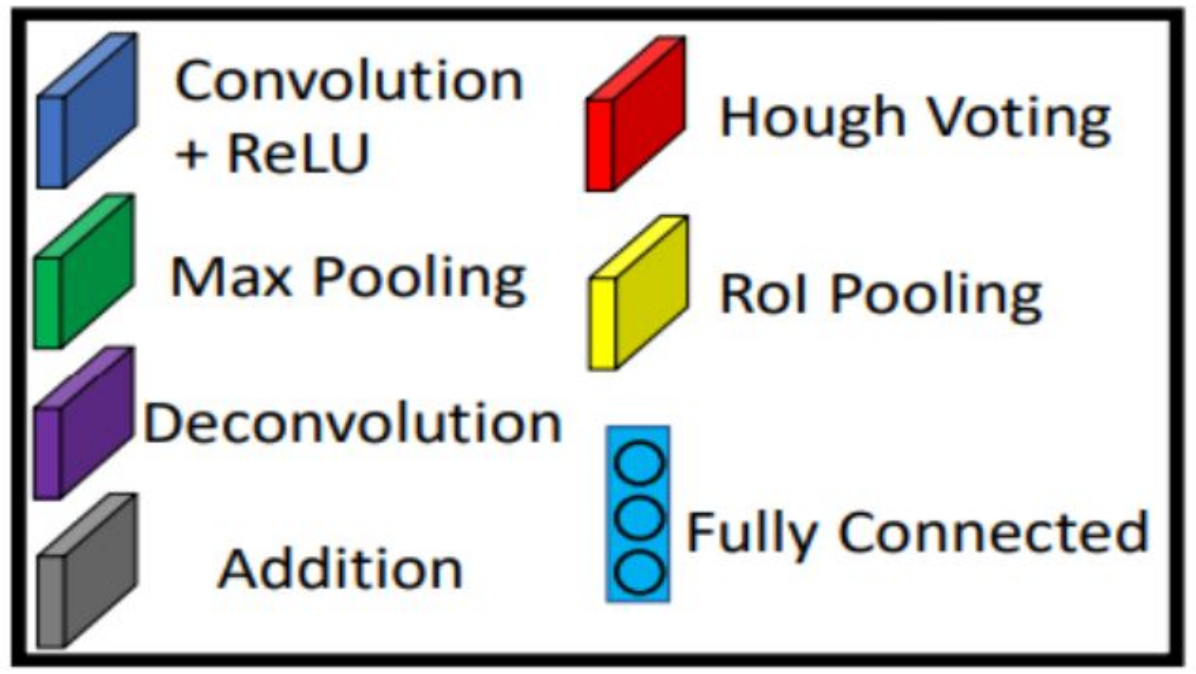


Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

Semantic Labels

1. Feature embedding

2. Softmax score for each pixel



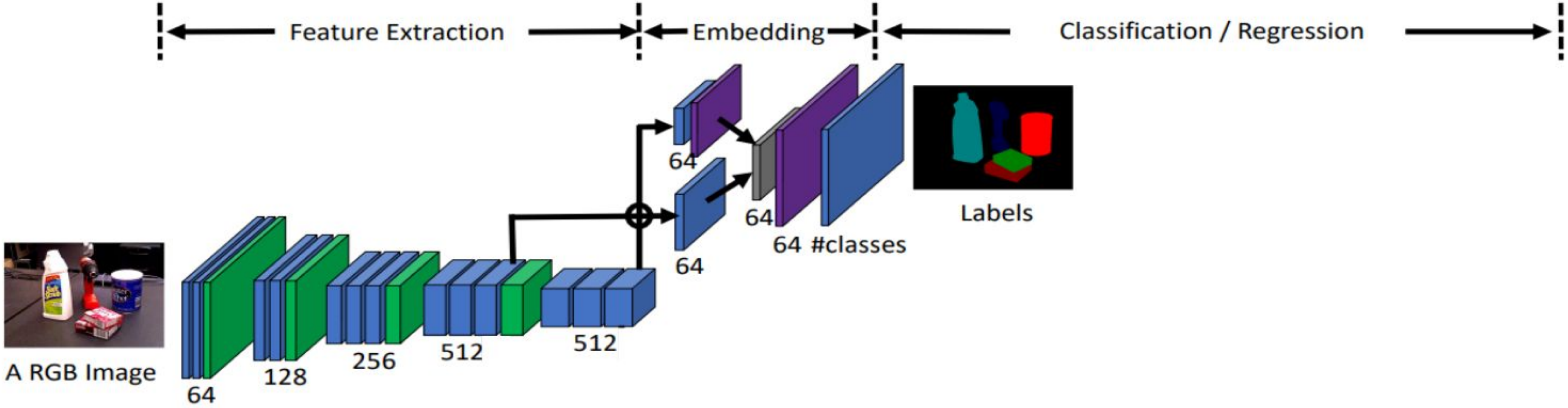


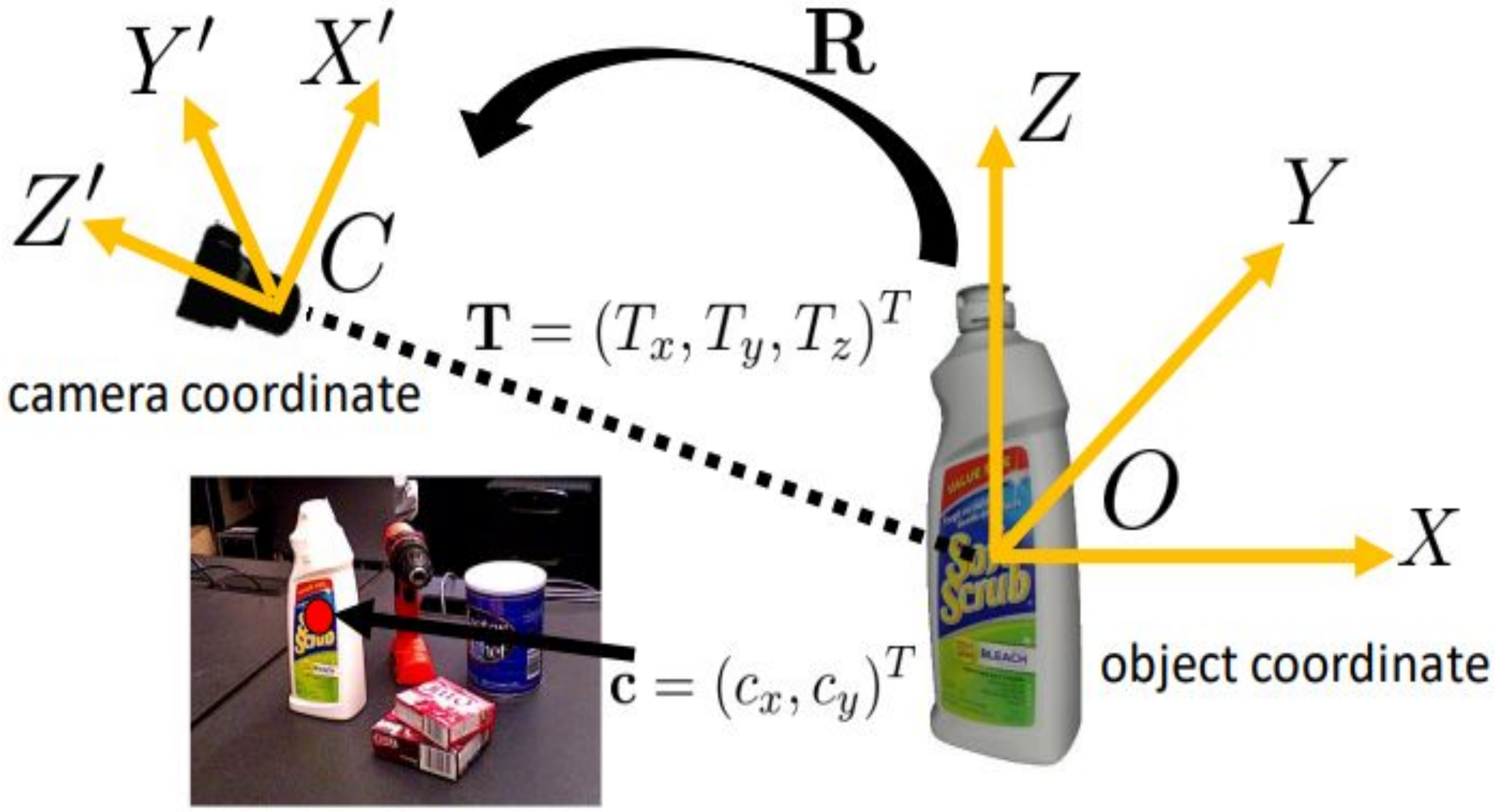
Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

3D Translation Estimation

Required Output ： $\mathbf{T} = (T_x, T_y, T_z)^T$

Method:

1. Object center: $$\begin{bmatrix} c_x \\ c_y \end{bmatrix} = \begin{bmatrix} f_x \frac{T_x}{T_z} + p_x \\ f_y \frac{T_y}{T_z} + p_y \end{bmatrix}$$



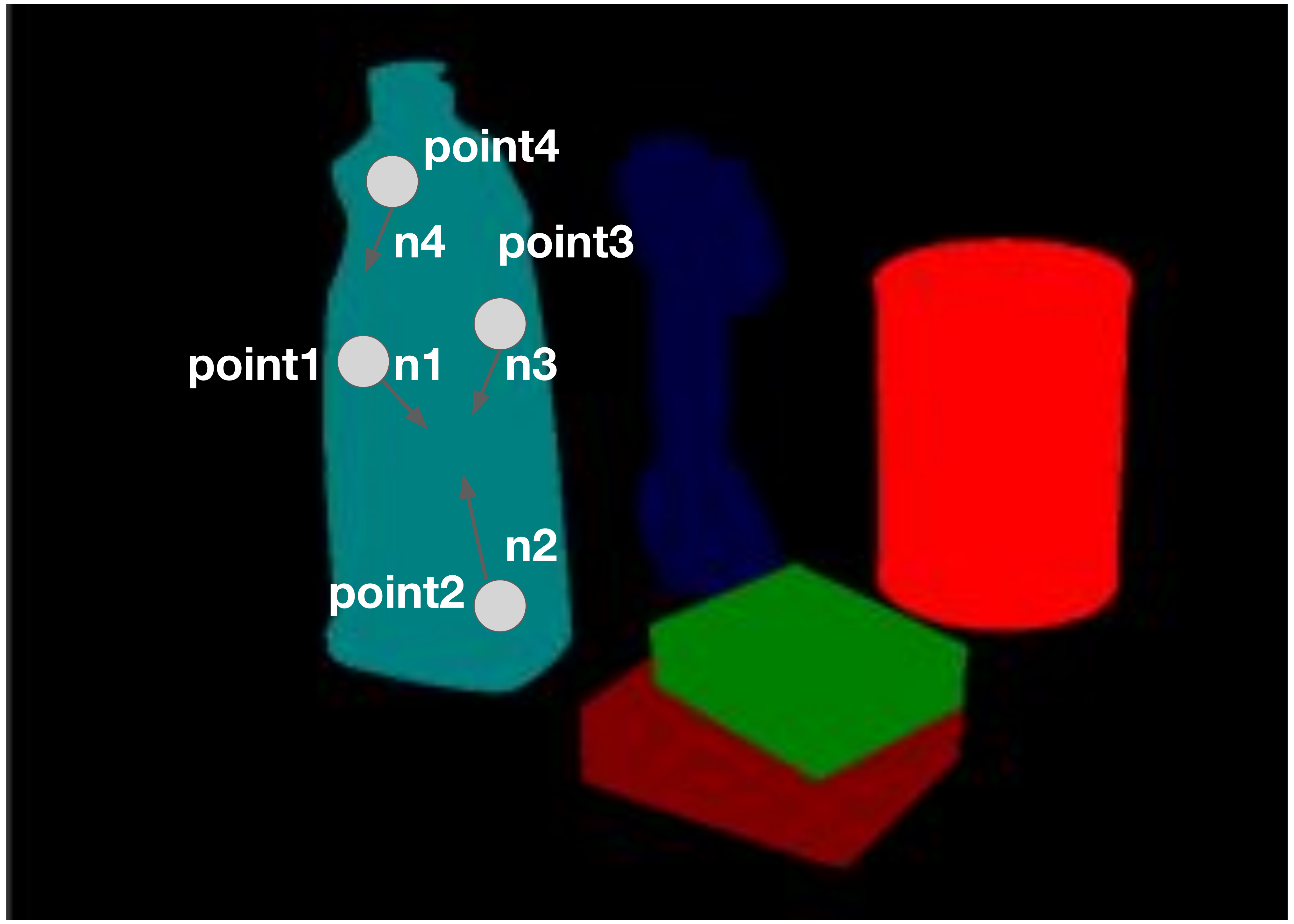
f_x and f_y $\longrightarrow$ Focal lengths of the camera

p_x and p_y $\longrightarrow$ Principle points

Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

3D Translation Estimation

2. Find Object center: $$(x, y) \rightarrow \left(n_x = \frac{c_x - x}{\|\mathbf{c} - \mathbf{p}\|}, n_y = \frac{c_y - y}{\|\mathbf{c} - \mathbf{p}\|}, T_z \right)$$



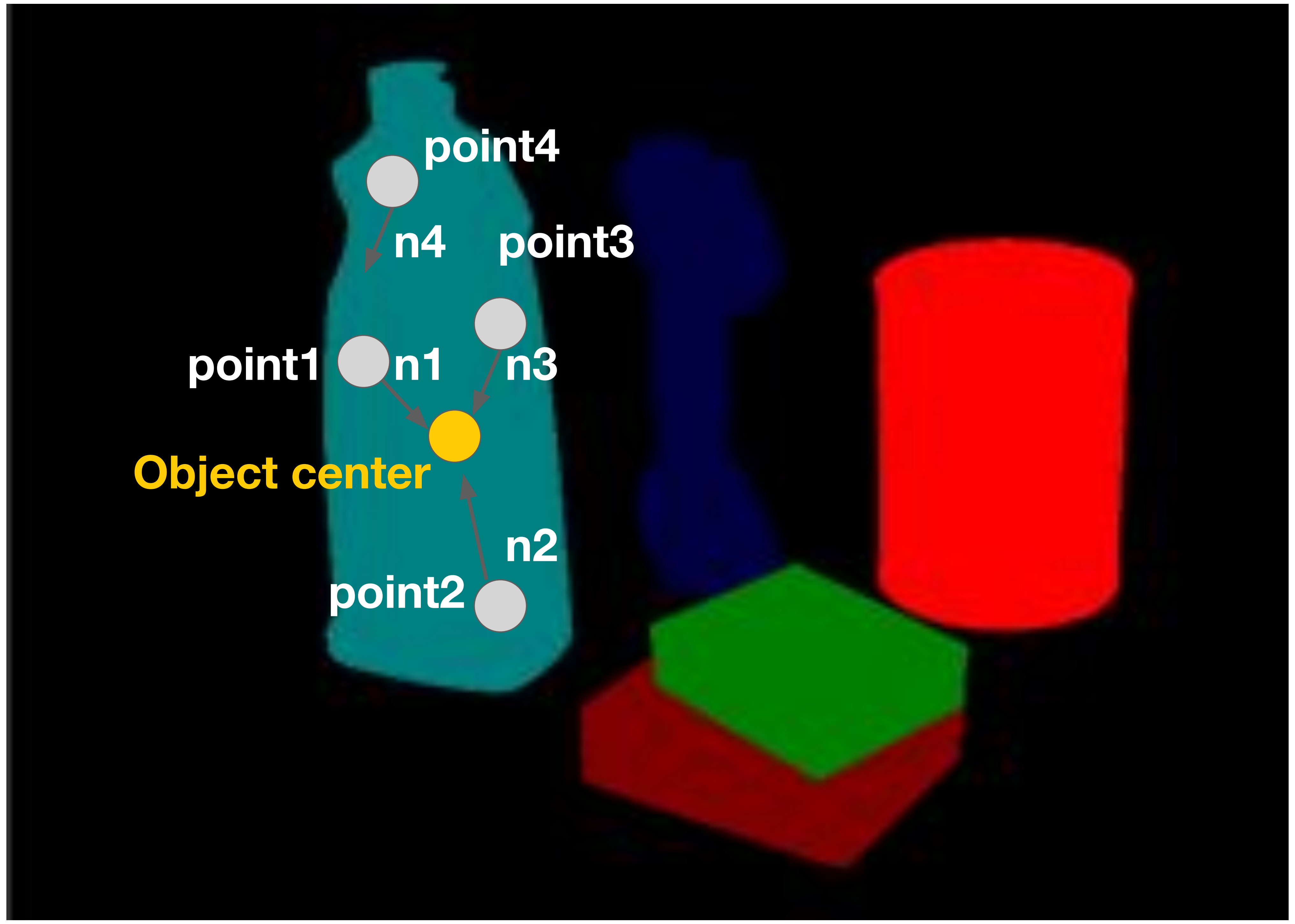
n_x and n_y $\longrightarrow$ Unit length vector

(x, y) $\longrightarrow$ Pixel in each object

Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

3D Translation Estimation

2. Find Object center: $(x, y) \rightarrow \left(n_x = \frac{c_x - x}{\|\mathbf{c} - \mathbf{p}\|}, n_y = \frac{c_y - y}{\|\mathbf{c} - \mathbf{p}\|}, T_z\right)$



Object center identified based on voting

Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

3D Translation Estimation

3. Training the model to estimate n_x, n_y and T_z :

- n_x, n_y are utilized to identify c_x, c_y
- Then T_x, T_y and T_z can be predicted

Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

3D Rotation Estimation

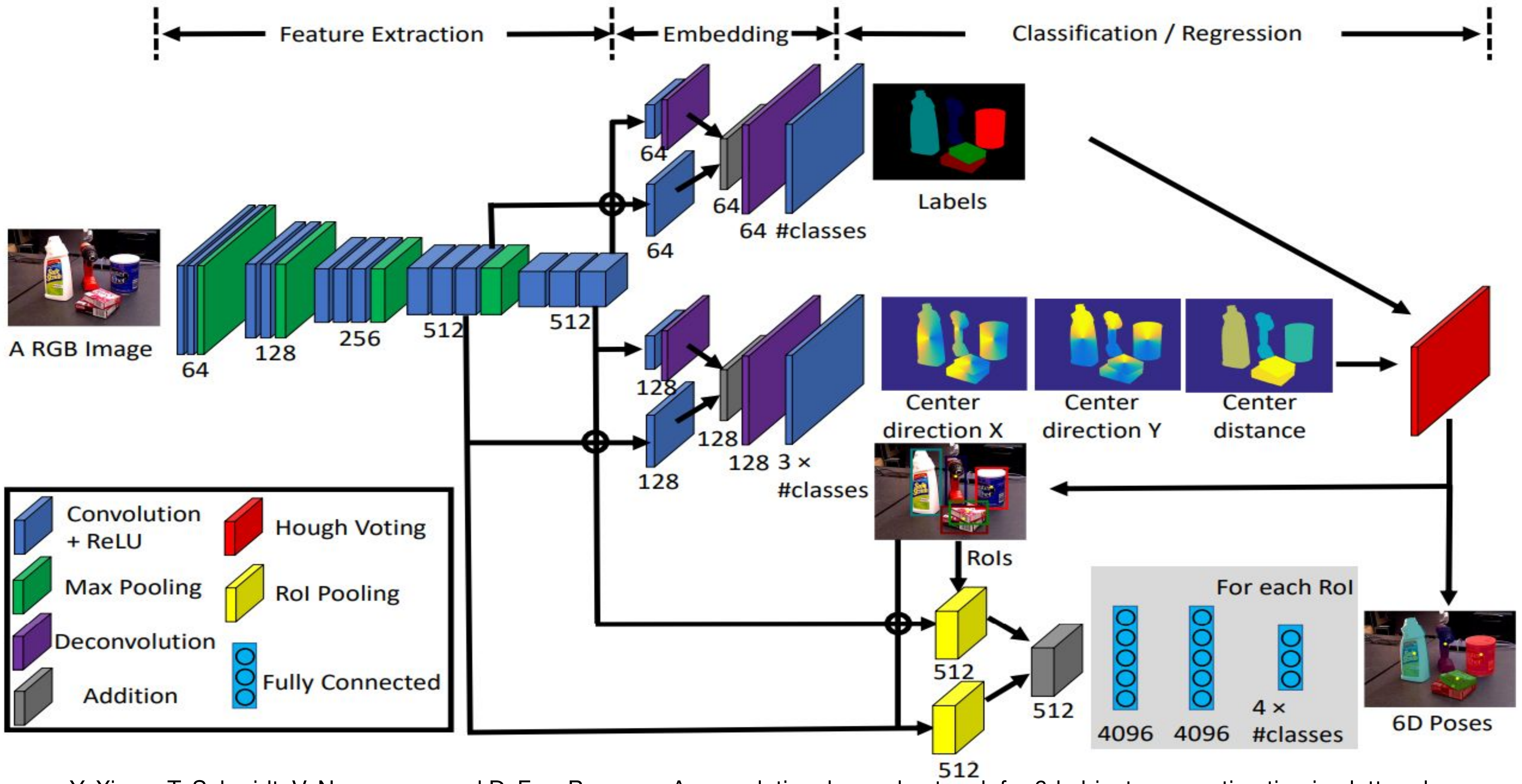


Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

3D Rotation Estimation

1. Pose loss: $$\mathrm{PLoss}(\tilde{\mathbf{q}}, \mathbf{q}) = \frac{1}{2m} \sum_{\mathbf{x} \in \mathcal{M}} \| R(\tilde{\mathbf{q}})\mathbf{x} - R(\mathbf{q})\mathbf{x} \|^2$$

2. ShapeMatch loss: $$\mathrm{SLoss}(\tilde{\mathbf{q}}, \mathbf{q}) = \frac{1}{2m} \sum_{\mathbf{x}_1 \in \mathcal{M}} \min_{\mathbf{x}_2 \in \mathcal{M}} \| R(\tilde{\mathbf{q}})\mathbf{x}_1 - R(\mathbf{q})\mathbf{x}_2 \|^2$$

$R(\tilde{\mathbf{q}}) \longrightarrow$ Predicted Rotation matrix

$R(\mathbf{q}) \longrightarrow$ Ground truth Rotation matrix

$\mathcal{M} \longrightarrow$ Set of 3D model points

$m \longrightarrow$ Number of points

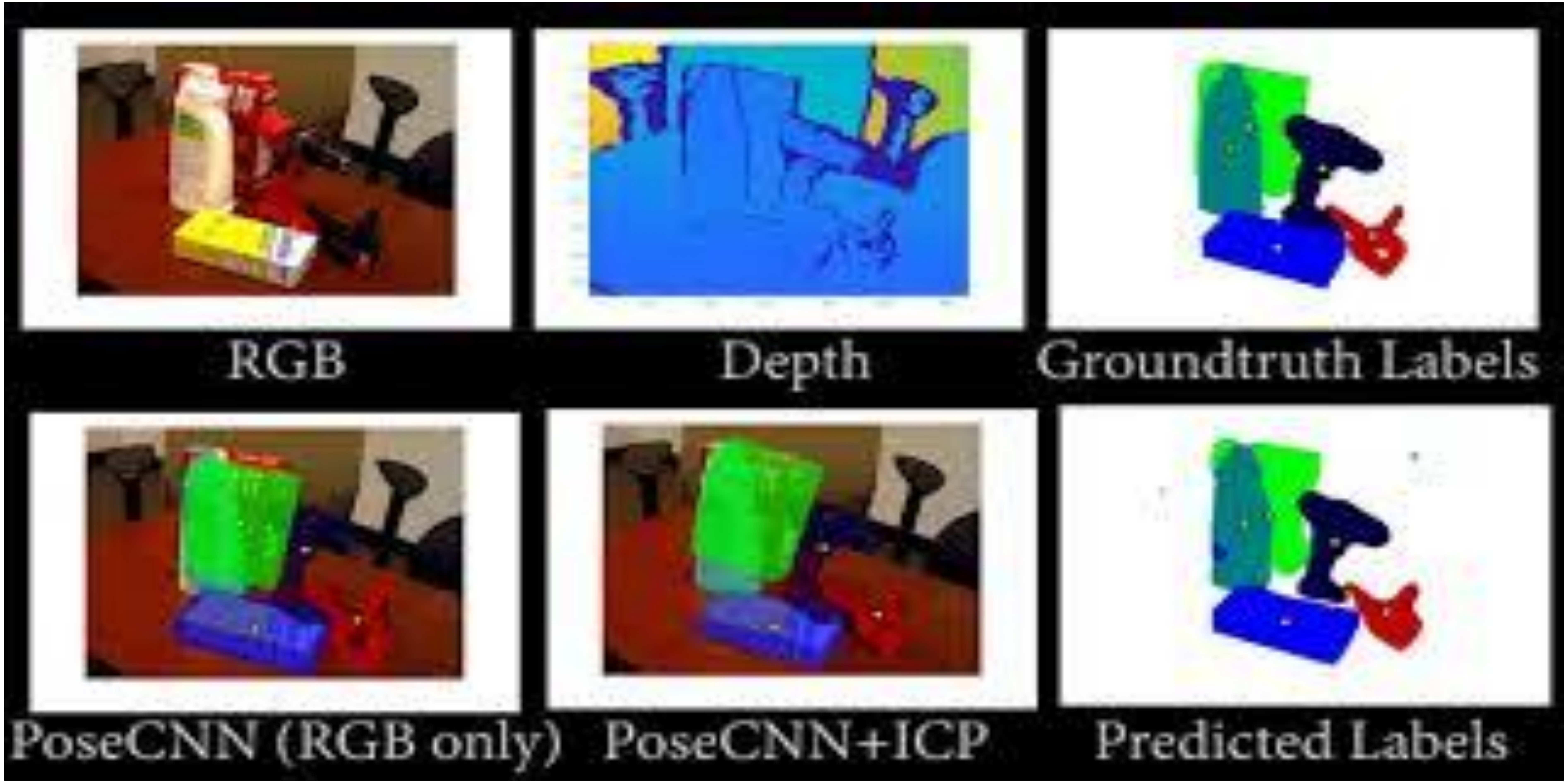
Results

Results for OccludedLINEMOD dataset

Method	Michel et al. [21]	Hinterstoisser et al. [14]	Krull et al. [17]	Brachmann et al. [3]	Ours PoseCNN Color	Ours PoseCNN+ICP
Ape	80.7	**81.4**	68.0	53.1	9.6	76.2
Can	88.5	**94.7**	87.9	79.9	45.2	87.4
Cat	**57.8**	55.2	50.6	28.2	0.93	52.2
Driller	**94.7**	86.0	91.2	82.0	41.4	90.3
Duck	74.4	**79.7**	64.7	64.3	19.6	77.7
Eggbox	47.6	65.5	41.5	9.0	22.0	**72.2**
Glue	73.8	52.1	65.3	44.5	38.5	**76.7**
Holepuncher	**96.3**	95.5	92.9	91.6	22.1	91.4
MEAN	76.7	76.3	70.3	56.6	24.9	**78.0**

Image source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

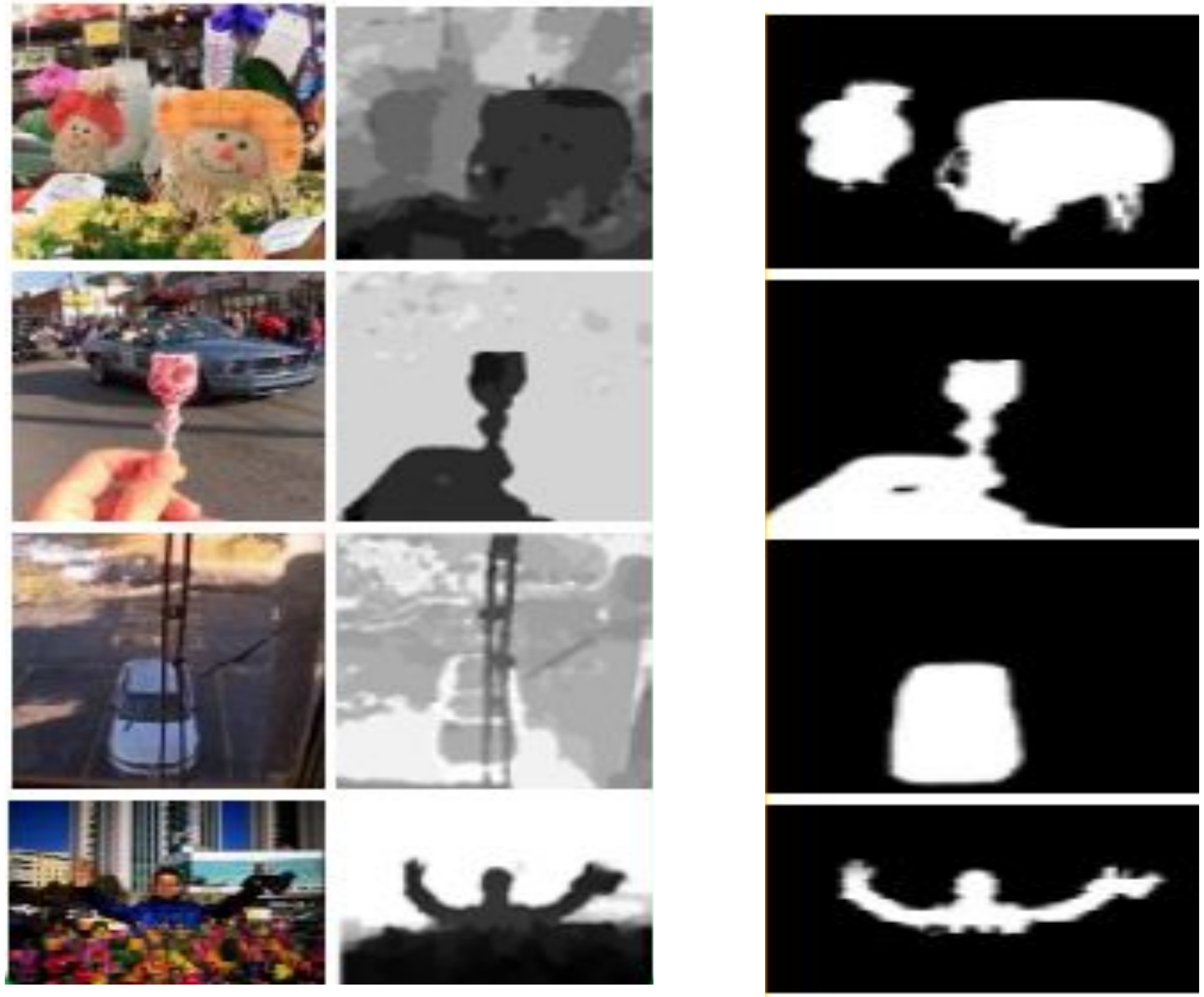
Results



source: Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018.

RGB-D for Salient Object Detection

RGB-D Salient Object Detection



1. RGB image

2. Depth image

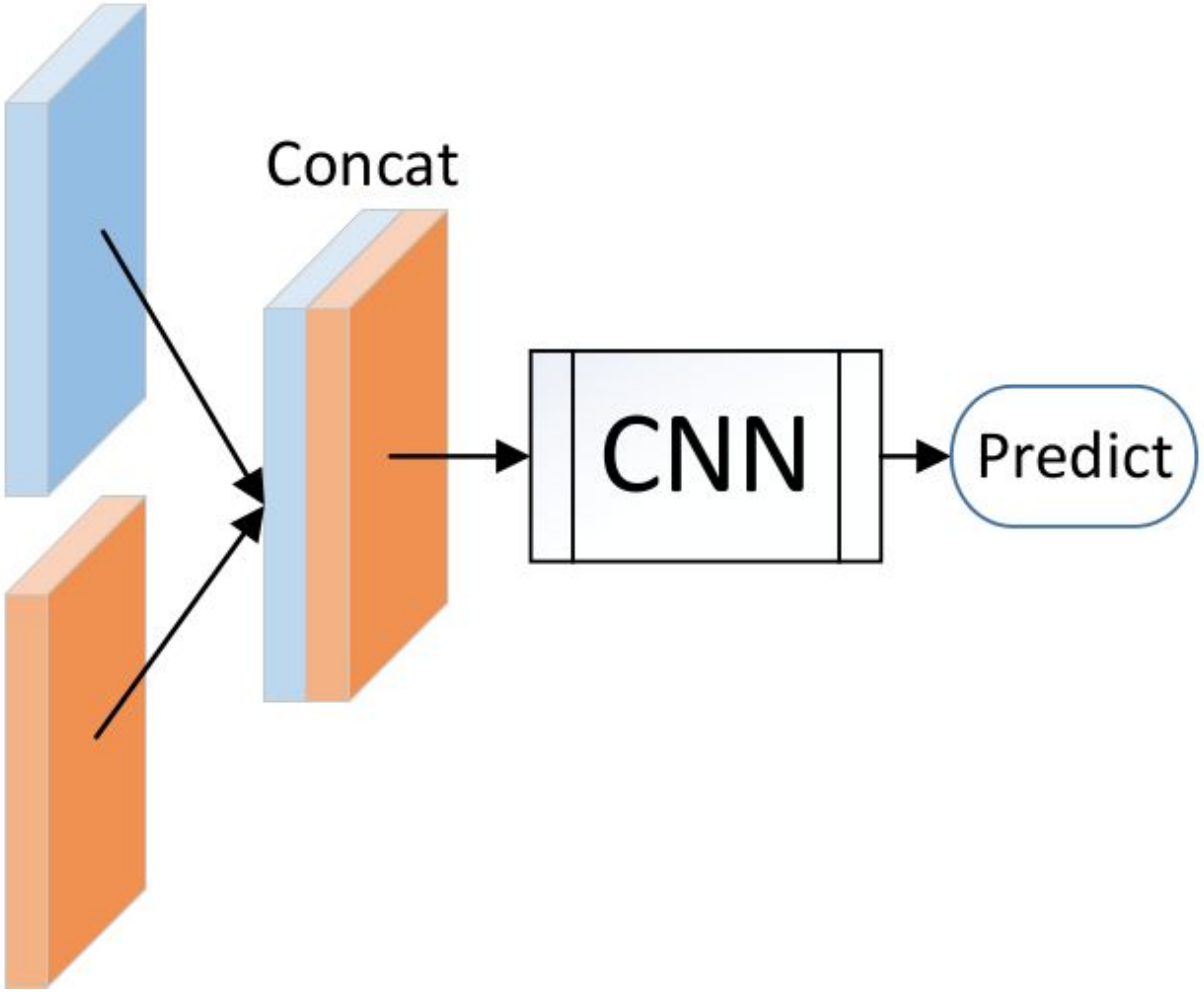
Detection of Salient Objects

source: Fu, K., Fan, D. P., Ji, G. P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. IEEE transactions on pattern analysis and machine intelligence, 44(9), 5541-5559.

Limitations of Existing Works

1. Fusion strategy:

   a. Early fusion

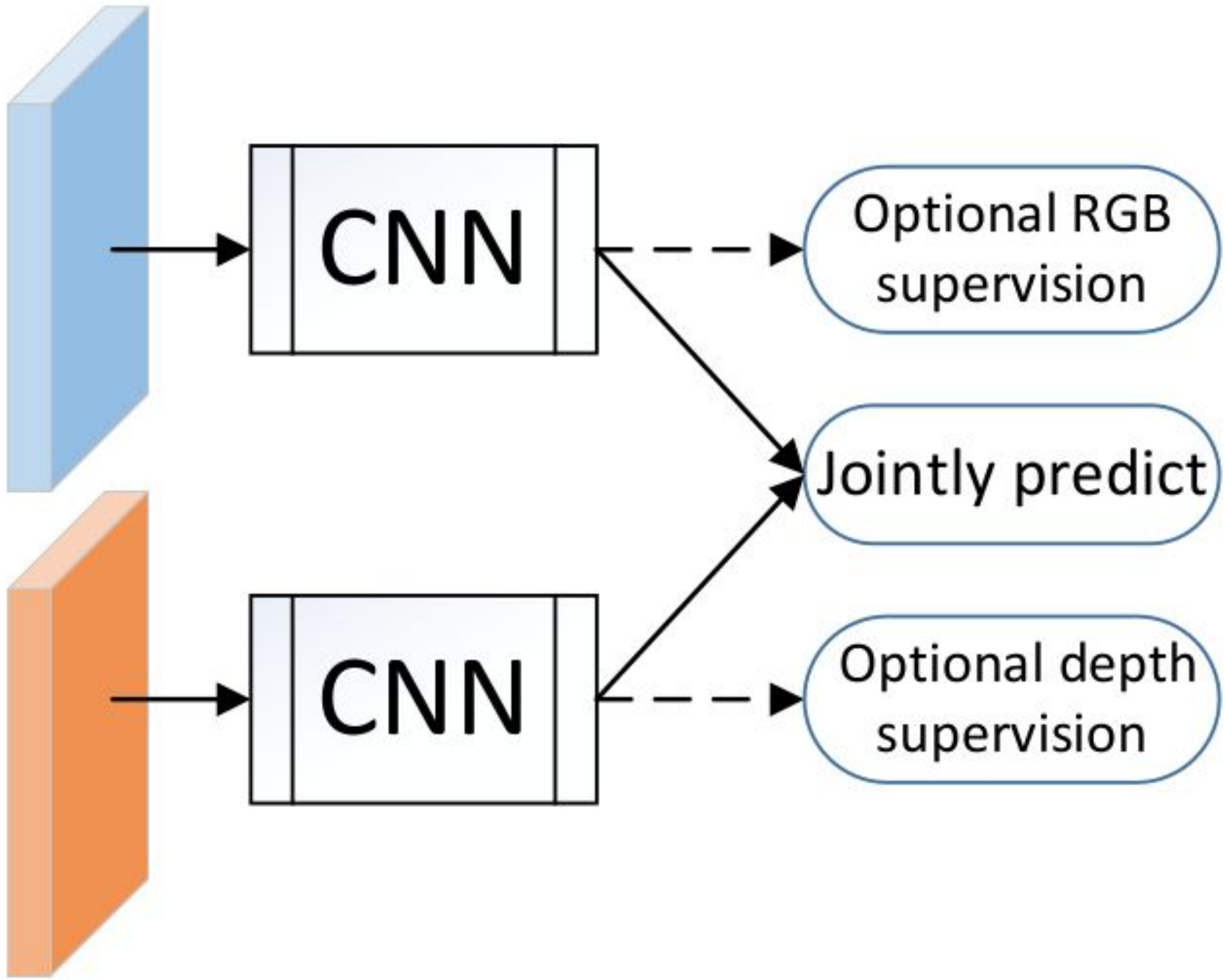


Source: Fu, K., Fan, D. P., Ji, G. P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. IEEE transactions on pattern analysis and machine intelligence, 44(9), 5541-5559.

Limitations of Existing Works

1. Fusion strategy:

   b. Late fusion

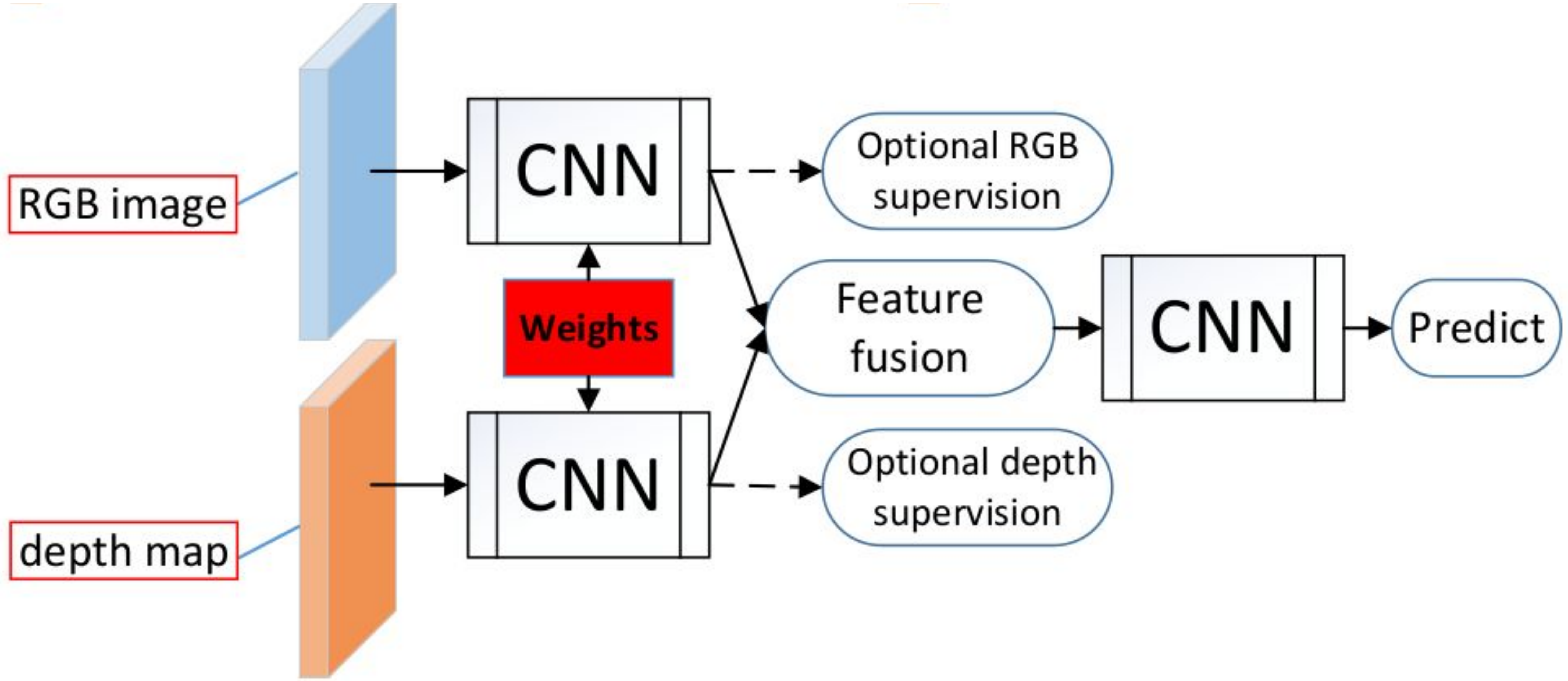


Source: Fu, K., Fan, D. P., Ji, G. P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. IEEE transactions on pattern analysis and machine intelligence, 44(9), 5541-5559.

Limitations of Existing Works

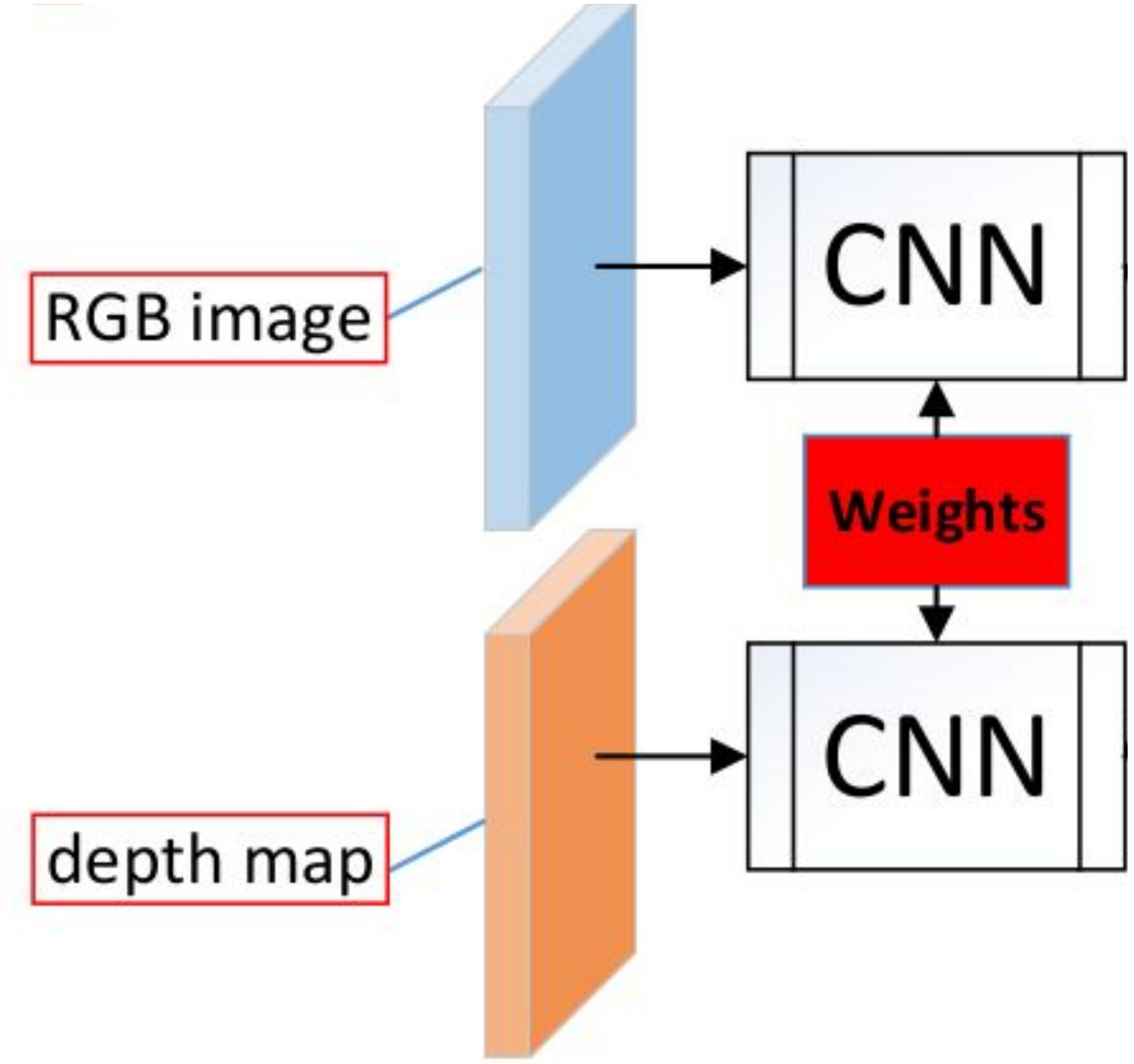
1. Fusion strategy:

c. Middle fusion



Source: Fu, K., Fan, D. P., Ji, G. P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. IEEE transactions on pattern analysis and machine intelligence, 44(9), 5541-5559.

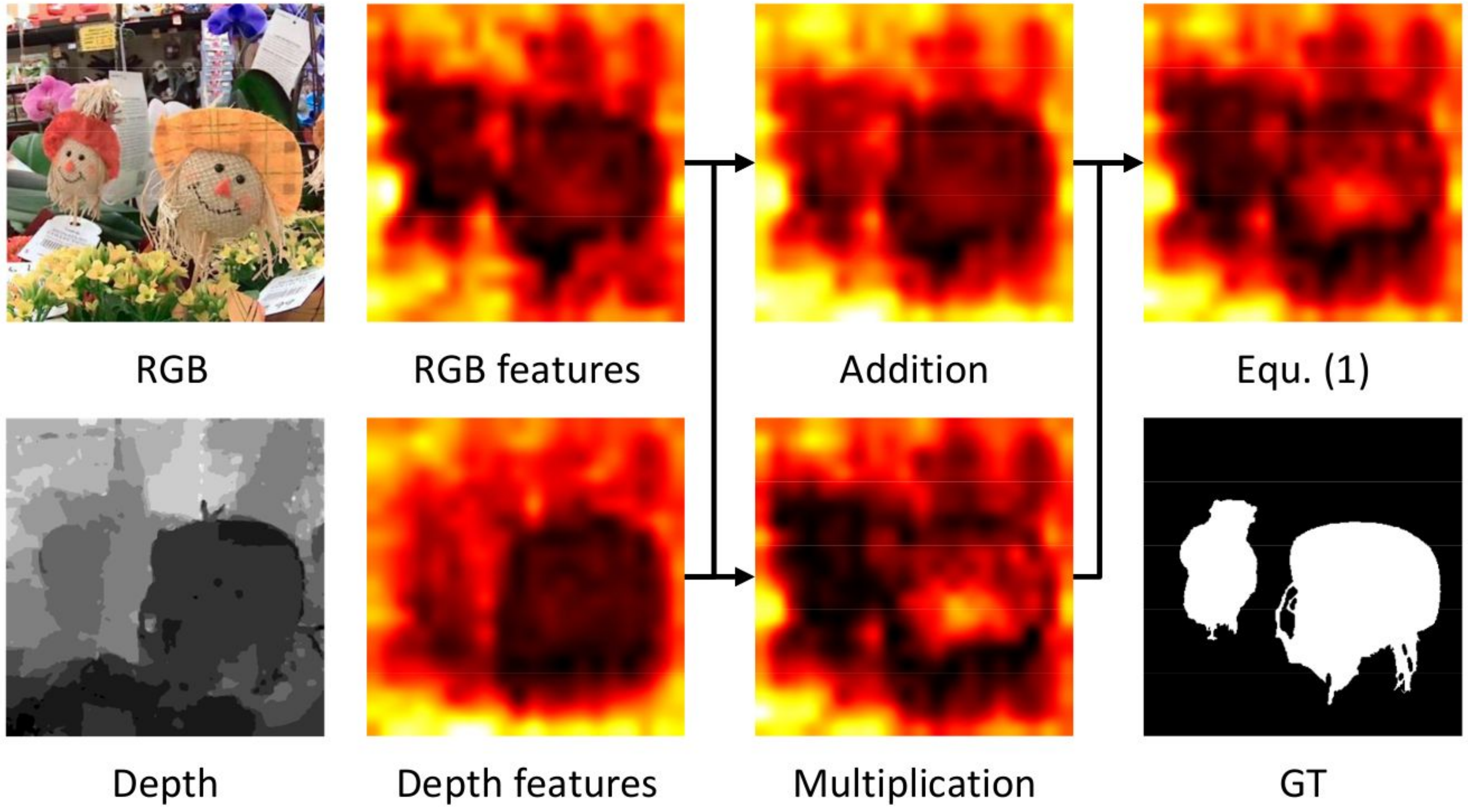
Joint Learning



Siamese network :
Process two different inputs in parallel with shared weights

source: Fu, K., Fan, D. P., Ji, G. P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. IEEE transactions on pattern analysis and machine intelligence, 44(9), 5541-5559.

Densely Cooperative Fusion

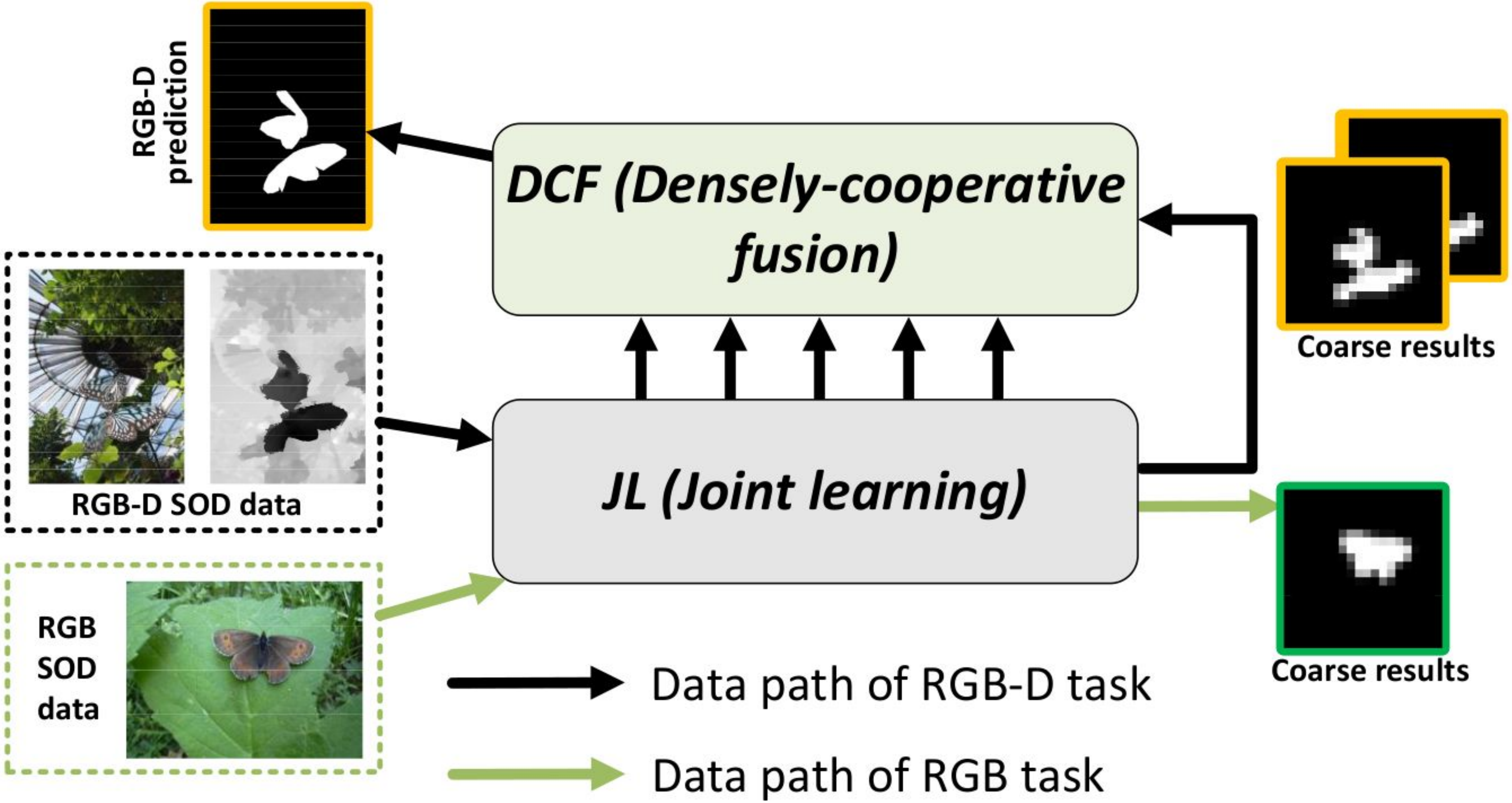


$$CM(\{X_{rgb}, X_d\}) = X_{rgb} \oplus X_d \oplus (X_{rgb} \otimes X_d)$$ Equ. (1)

$CM \longrightarrow$ Cross modal fusion

source: Fu, K., Fan, D. P., Ji, G. P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. IEEE transactions on pattern analysis and machine intelligence, 44(9), 5541-5559.

Joint learning and Densely Cooperative Fusion



source: Fu, K., Fan, D. P., Ji, G. P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. IEEE transactions on pattern analysis and machine intelligence, 44(9), 5541-5559.

Loss function

$$\mathcal{L}_{\text{total}} = \mathcal{L}_f\left(S^f, G\right) + \lambda \sum_{x \in \{rgb,d\}} \mathcal{L}_g(S^c_x, G)$$

$G \longrightarrow$ Ground truth

$S^f \longrightarrow$ Final prediction of model

$S^c_{rgb} \longrightarrow$ Coarse RGB prediction

$S^c_d \longrightarrow$ Coarse Depth prediction

source: Fu, K., Fan, D. P., Ji, G. P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. IEEE transactions on pattern analysis and machine intelligence, 44(9), 5541-5559.

Loss function

$$\mathcal{L}(S,G) = -\sum[G_i \log(S_i) + (1-G_i)\log(1-S_i)]$$

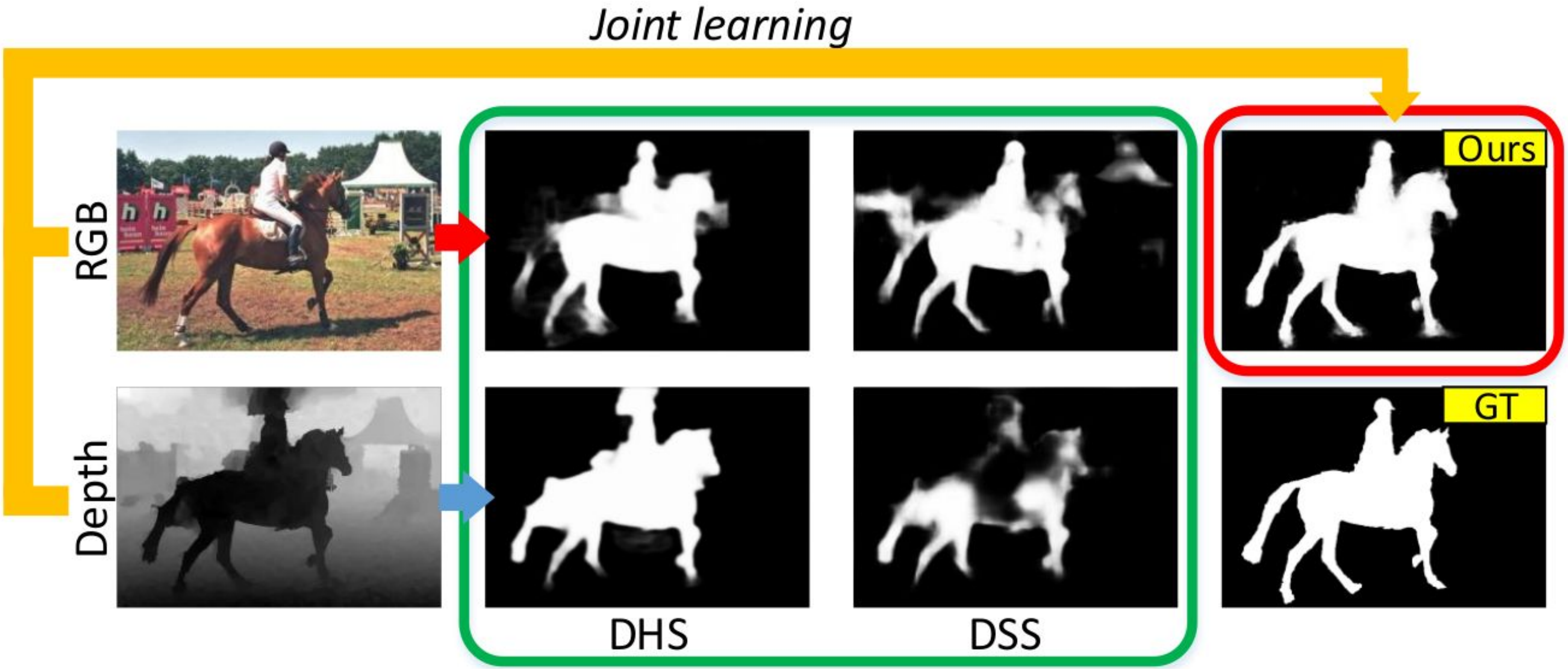
$\mathcal{L}(S,G) \longrightarrow$ Cross-entropy loss

$i \longrightarrow$ Pixel Index

$$S \in \{S^c_{rgb}, S^c_d, S^f\}$$

source: Fu, K., Fan, D. P., Ji, G. P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. IEEE transactions on pattern analysis and machine intelligence, 44(9), 5541-5559.

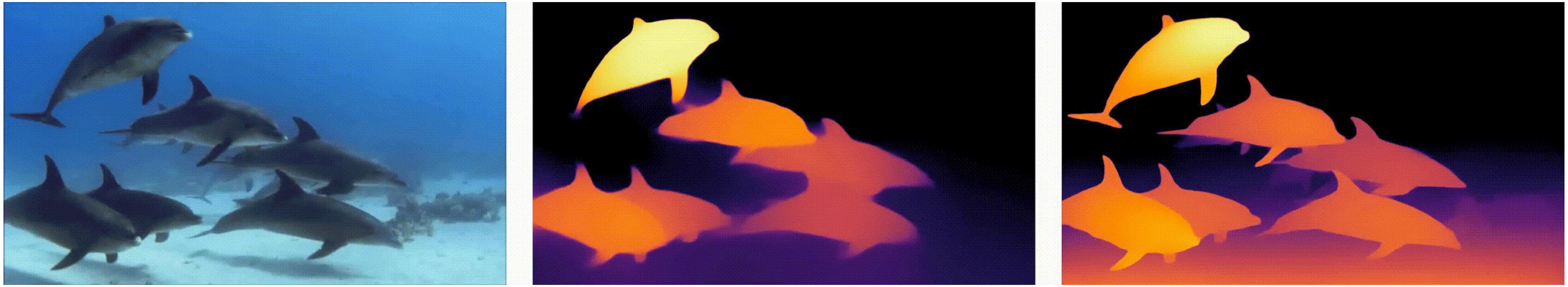
Results



Source: Fu, K., Fan, D. P., Ji, G. P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. IEEE transactions on pattern analysis and machine intelligence, 44(9), 5541-5559.

Depth from Single Image?

Depth from Single Image?



Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Monocular Depth Estimation (MDE)

Relative Depth Estimation → Relative Distance: 0~1

Metric Depth Estimation → Actual Distance: meters

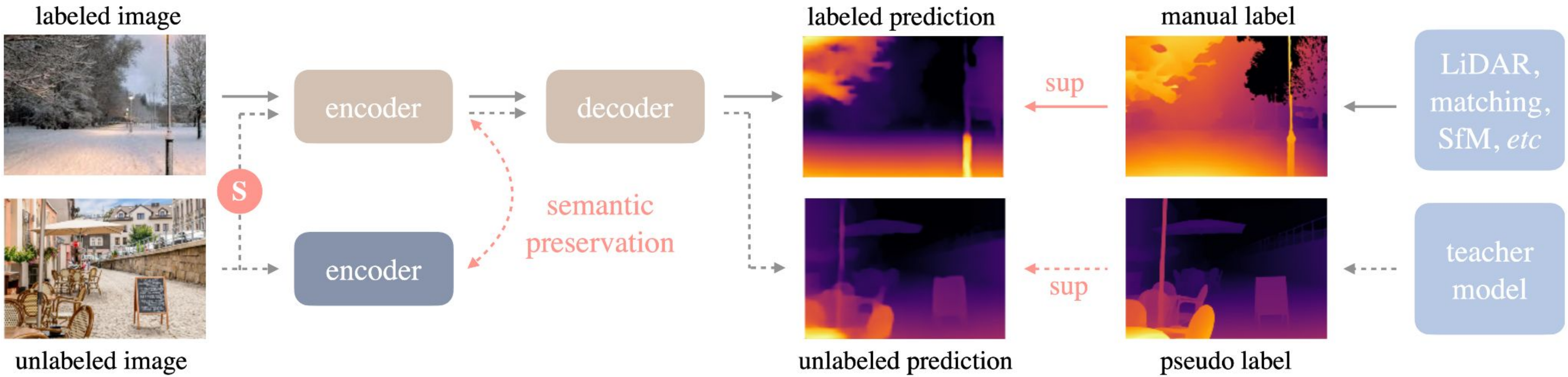
Ranftl, René, et al. "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer." IEEE transactions on pattern analysis and machine intelligence 44.3 (2020): 1623-1637.

Depth Anything

Dedicate to solving the generalization of MDE

Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Anything Pipeline



Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Anything Pipeline

Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Anything Pipeline

Input

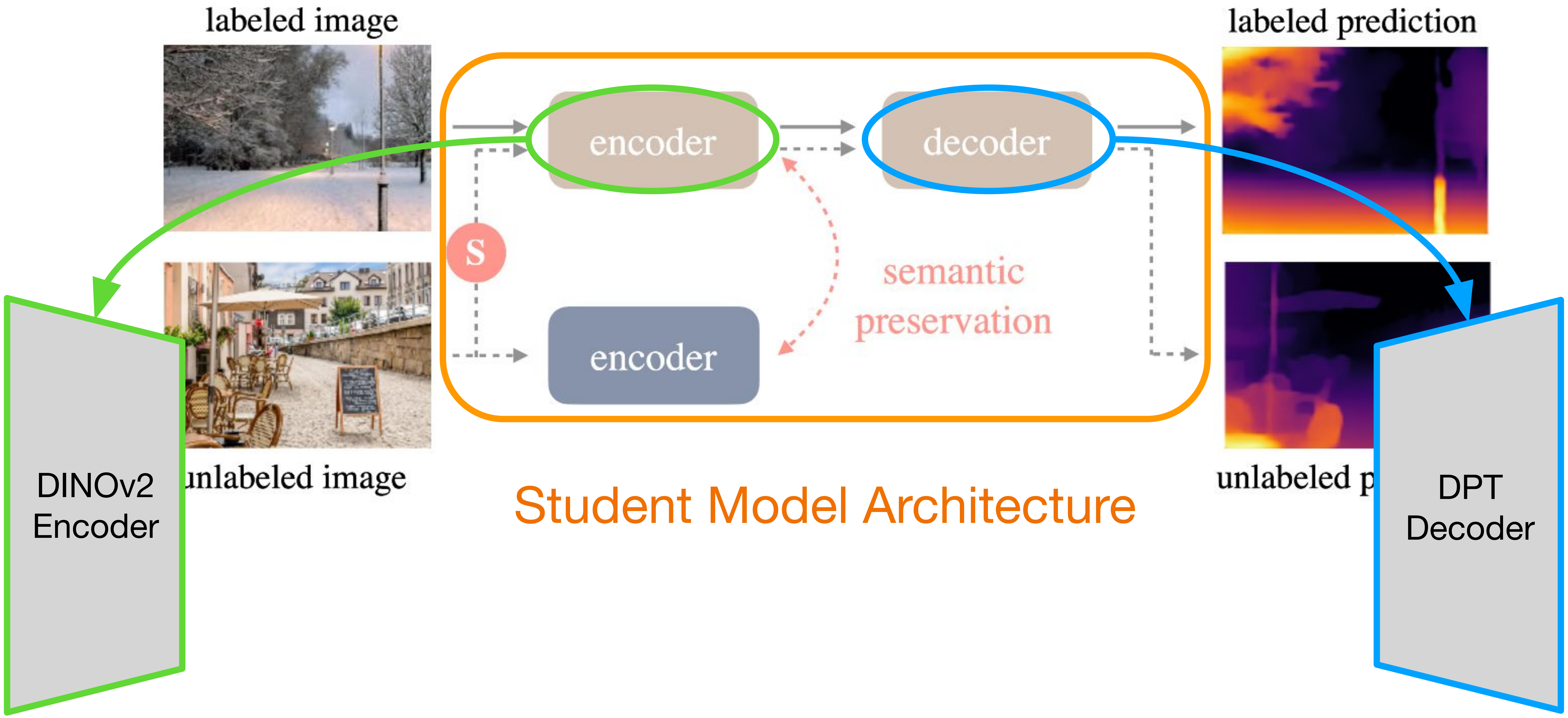
Label

Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Anything Pipeline

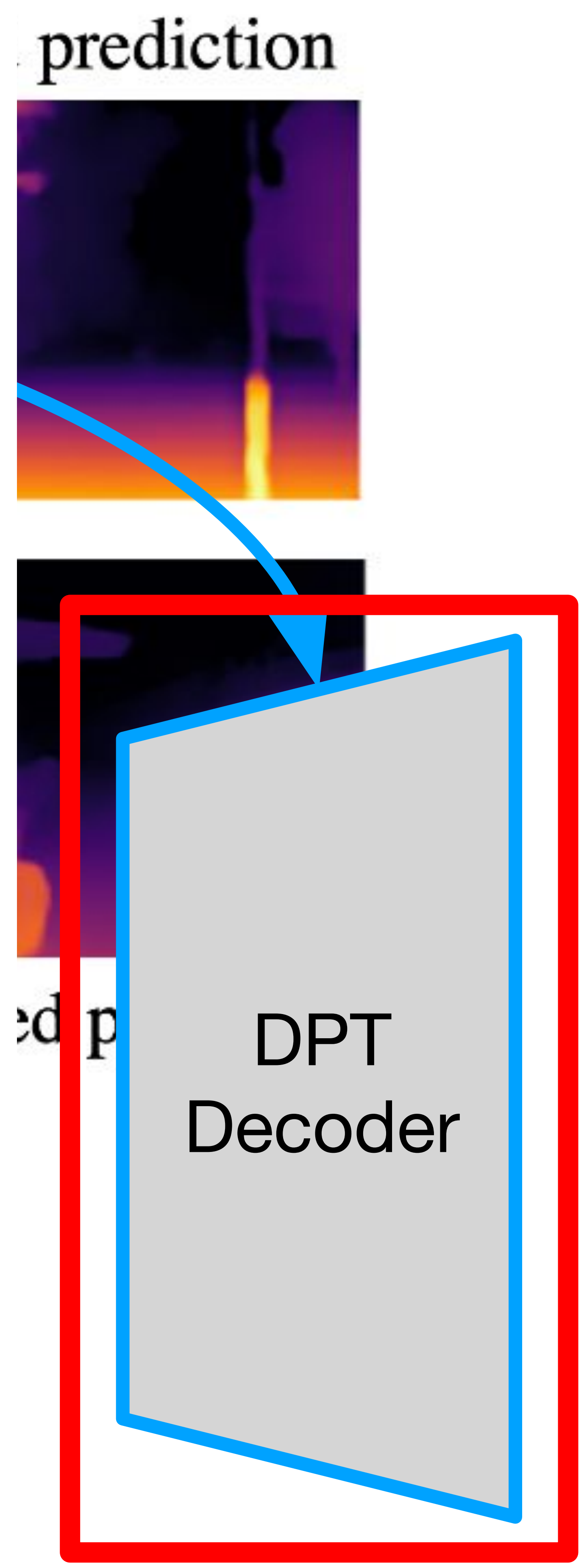
Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Anything Model Architecture



Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Anything Model Architecture

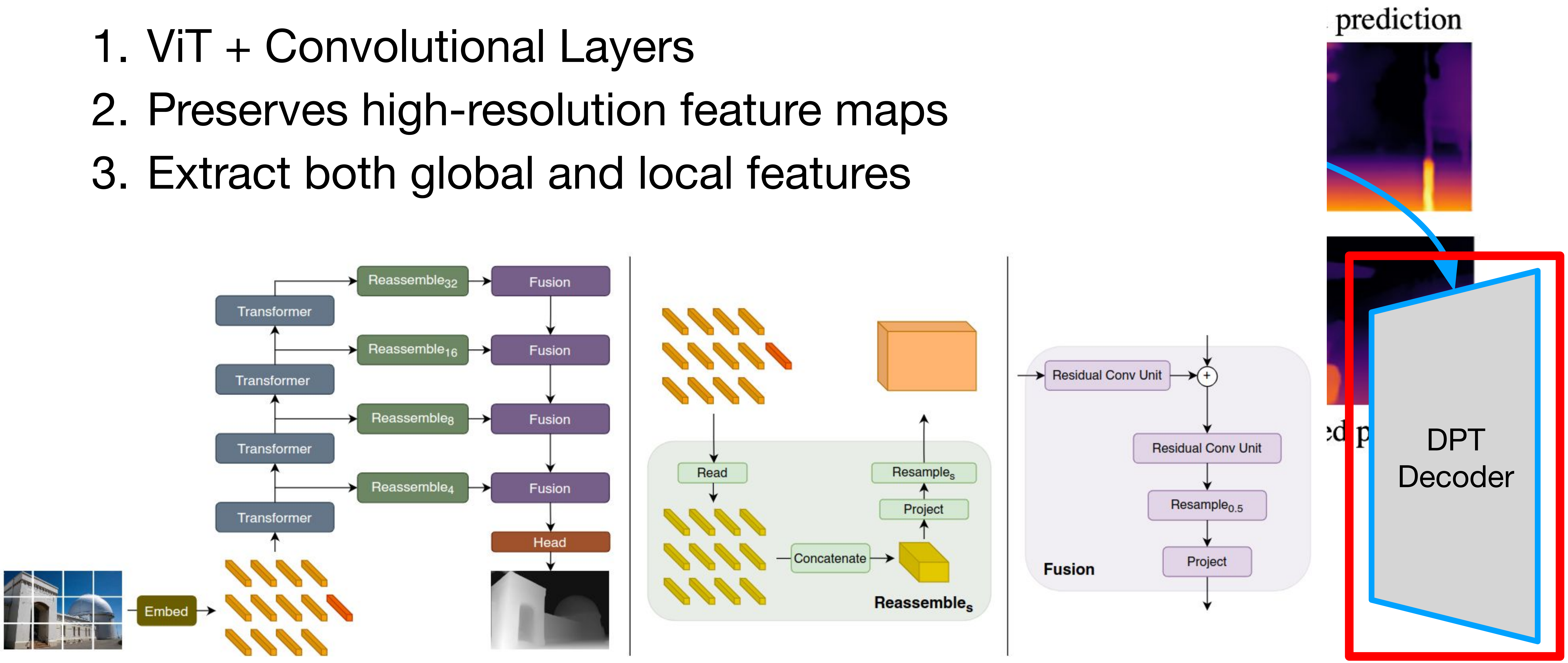


What is DPT?

Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Dense Prediction Transformer (DPT)

1. ViT + Convolutional Layers
2. Preserves high-resolution feature maps
3. Extract both global and local features

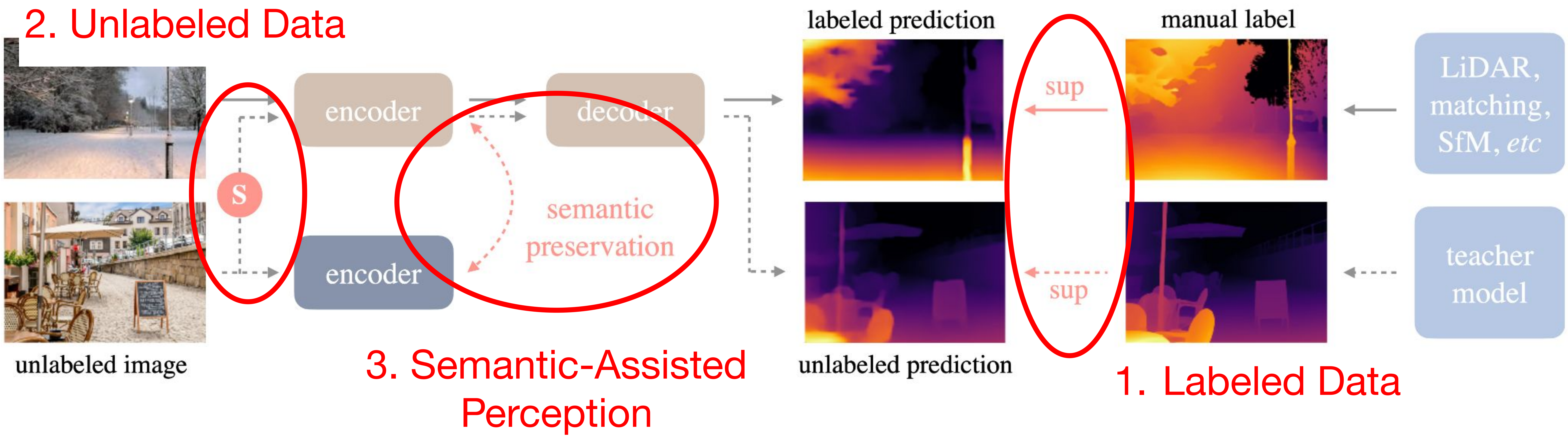


Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Anything Pipeline

Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Anything



Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Labeled Data: Affine-Invariant Loss

Predicted Depth

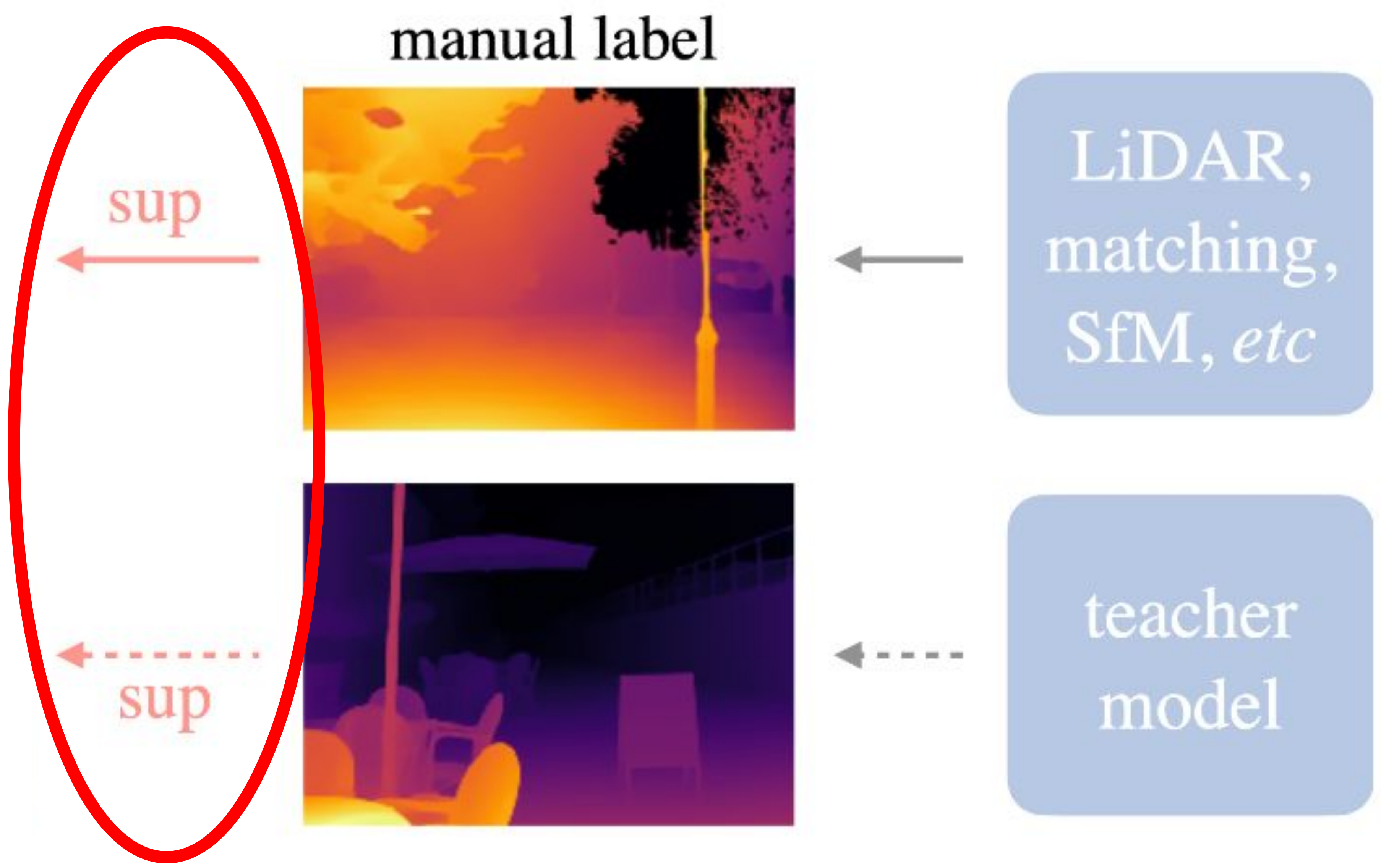
GT Depth

$$\mathcal{L}_l = \frac{1}{HW}\sum_{i=1}^{HW}\rho(d_i^*, d_i)$$

absolute error loss:

$$\rho(d_i^*, d_i) = |\hat{d}_i^* - \hat{d}_i|$$

Shifted & Scaled



1. Labeled Data

Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Labeled Data: Affine-Invariant Loss

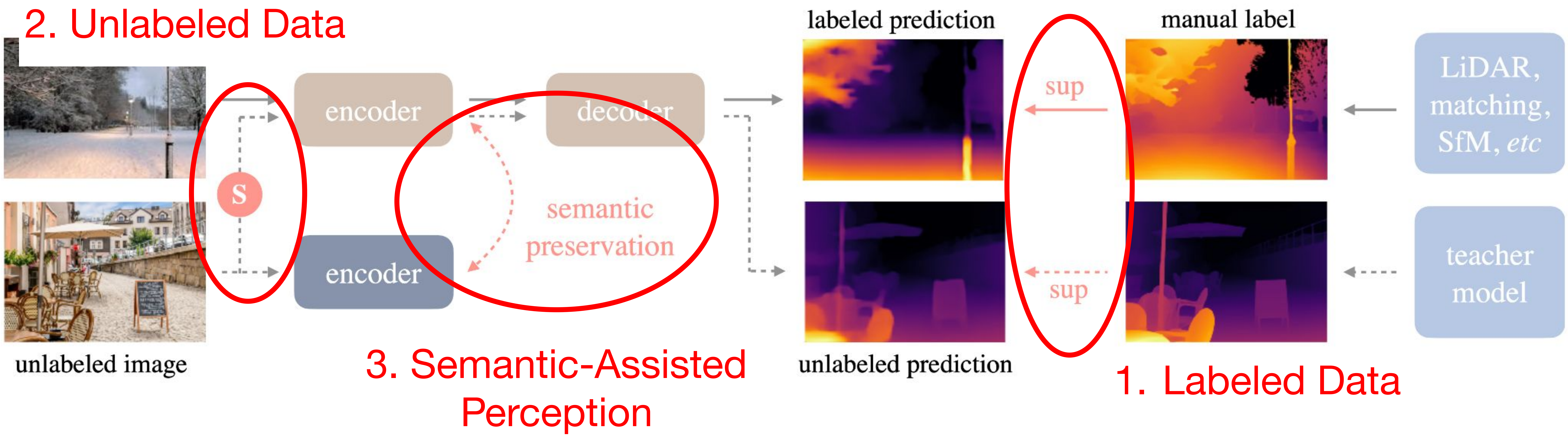
$$\mathcal{L}_l = \frac{1}{HW} \sum_{i=1}^{HW} \rho(d_i^*, d_i)$$

$$\rho(d_i^*, d_i) = |\hat{d}_i^* - \hat{d}_i|$$

$$\hat{d}_i = \frac{d_i - t(d)}{s(d)} \qquad t(d) = \text{median}(d), \quad s(d) = \frac{1}{HW} \sum_{i=1}^{HW} |d_i - t(d)|$$

Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Anything



Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Unlabeled Data

2. Unlabeled Data

Problem: Failed to gain improvement at first.

Hypothesis: Teacher and Student Model behave similar.

Solution: Challenge student model with strong perturbations.

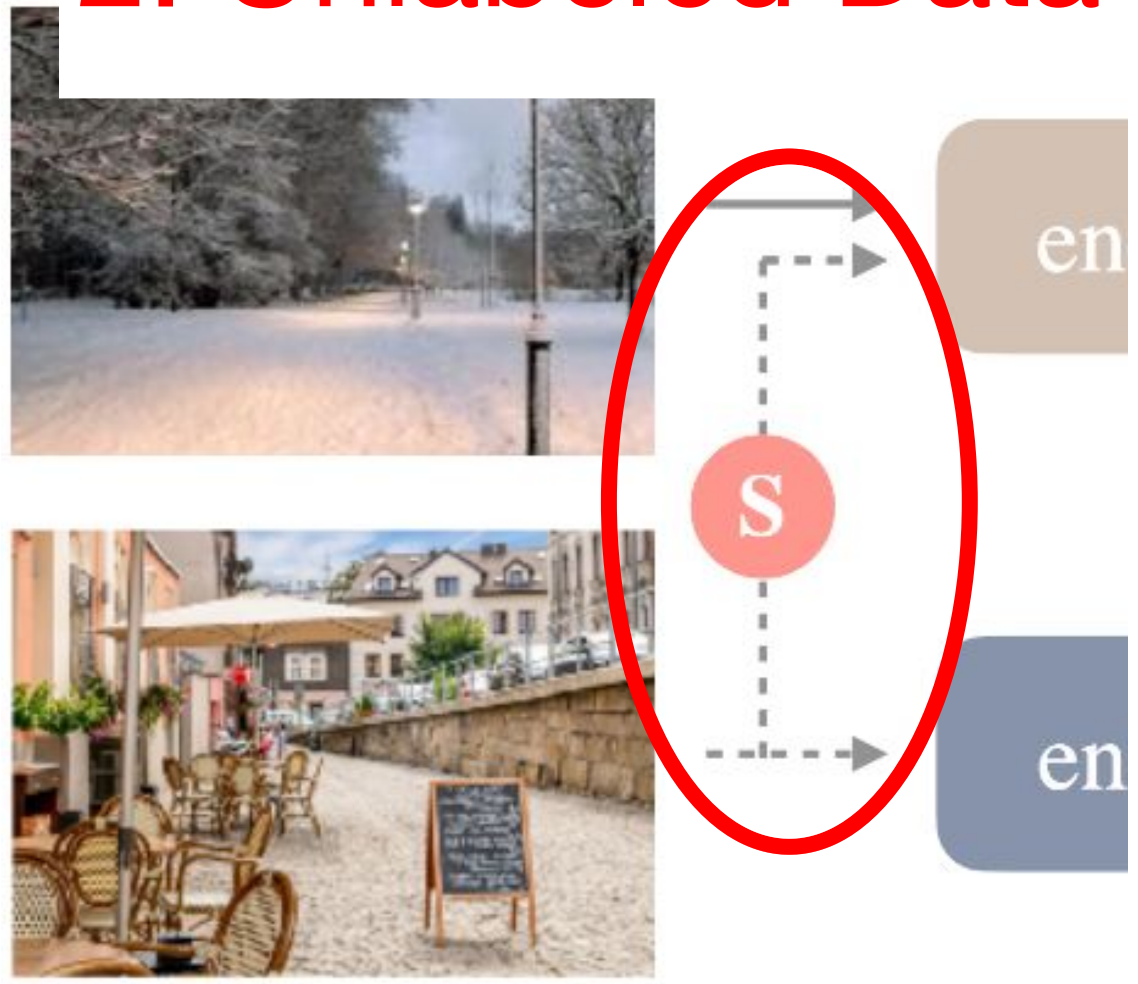
1. Strong Color Distortions
   a. Color Jittering
   b. Gaussian Blurring
2. Strong Spatial Distortions: CutMix

Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Unlabeled Data Loss

Affine-Invariant Loss

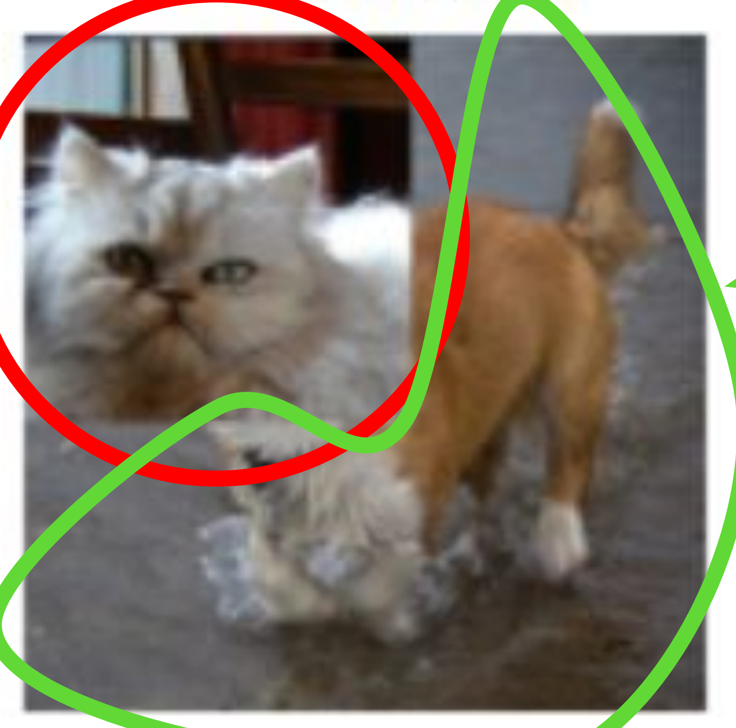
2. Unlabeled Data



unlabeled image

$$\mathcal{L}_u = \frac{\sum M}{HW}\mathcal{L}_u^{M} + \frac{\sum(1-M)}{HW}\mathcal{L}_u^{1-M}.$$

CutMix



Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

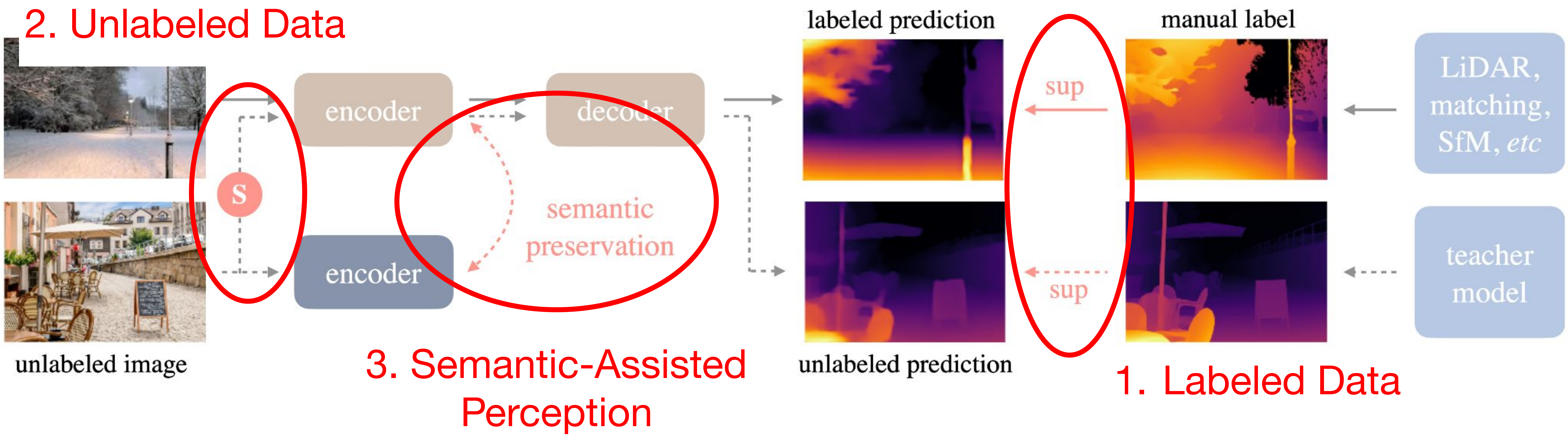
Unlabeled Loss

$$\mathcal{L}_u = \frac{\sum M}{HW}\mathcal{L}_u^M + \frac{\sum(1-M)}{HW}\mathcal{L}_u^{1-M}.$$

$$\mathcal{L}_u^M = \rho\big(S(u_{ab}) \odot M,\, T(u_a) \odot M\big),$$
$$\mathcal{L}_u^{1-M} = \rho\big(S(u_{ab}) \odot (1-M), T(u_b) \odot (1-M)\big)$$

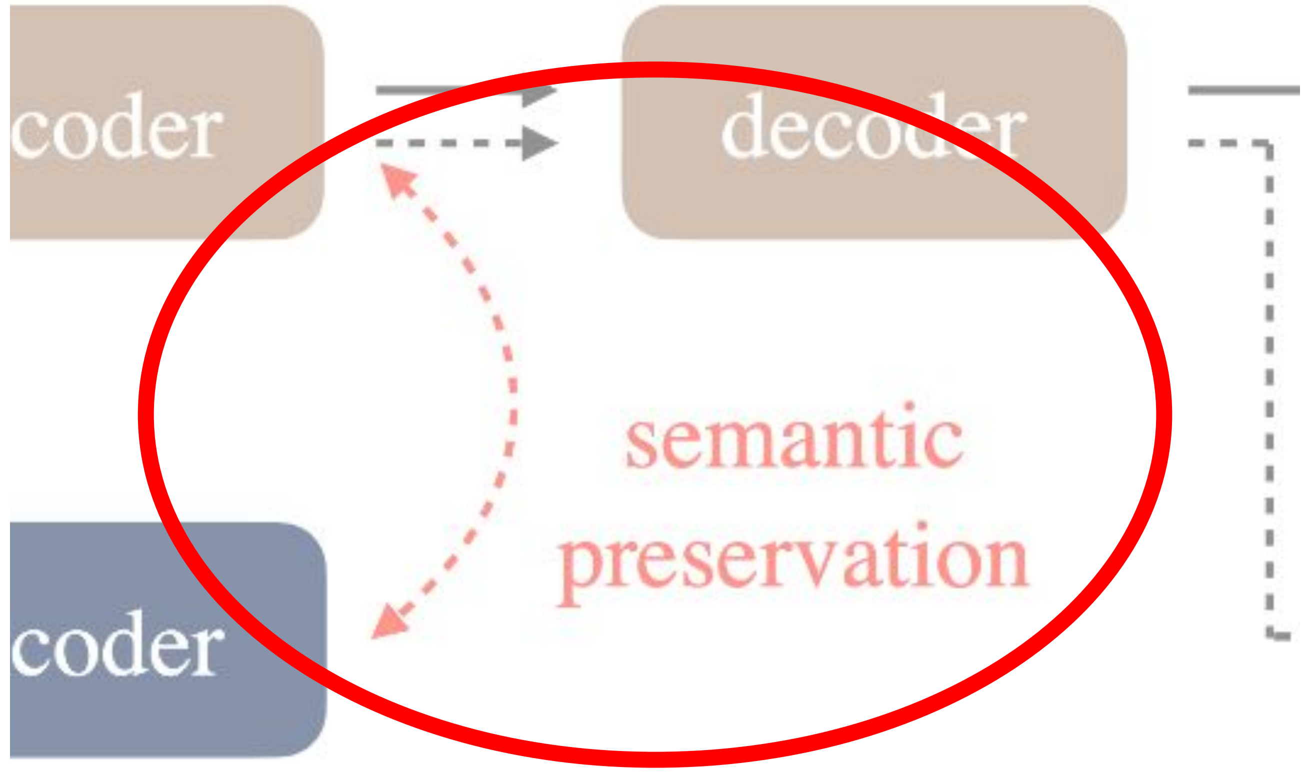
Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Anything



Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Depth Anything

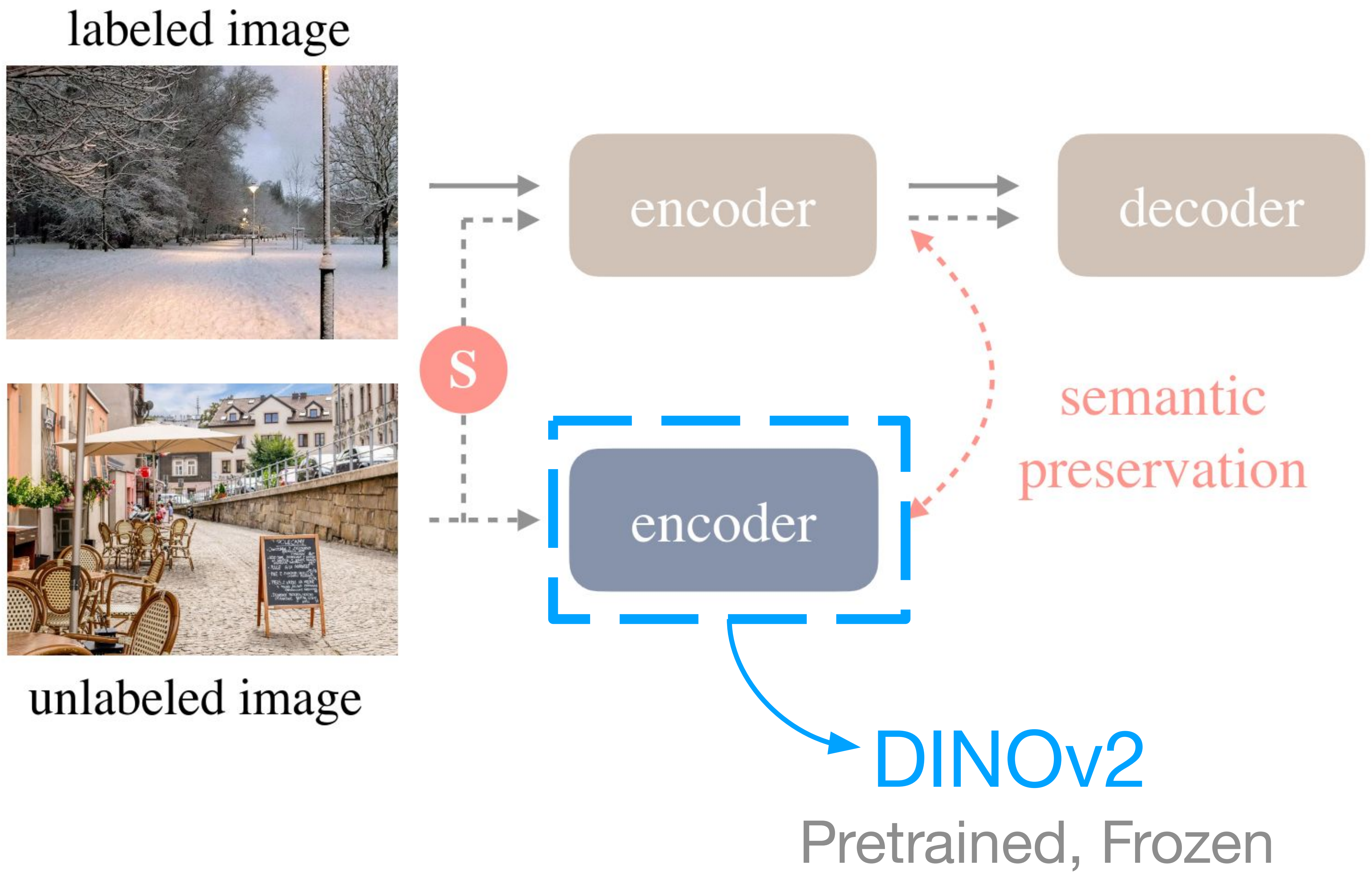


3. Semantic-Assisted Perception

Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Semantic-Assisted Perception

1. Combat the potential noise in pseudo depth label.
2. Transfer DINOv2's strong semantic capability



Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Feature Alignment Loss

feature from DINOv2

$$\mathcal{L}_{feat} = 1 - \frac{1}{HW} \sum_{i=1}^{HW} \cos(f_i, f_i')$$

feature from student model

Set a tolerance margin α:

DINOv2 produce similar feature for same object, but different part can be of varying depth.

Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

3 types of Loss

1. Affine-Invariant Loss: Labeled Data

$$\mathcal{L}_l = \frac{1}{HW}\sum_{i=1}^{HW}\rho(d_i^*, d_i)$$

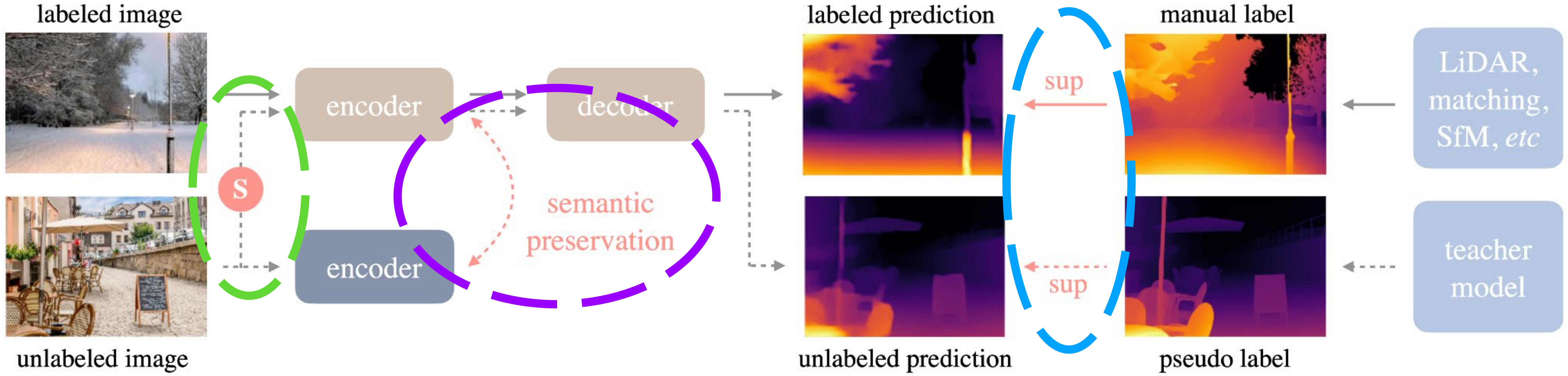
2. Unlabeled Loss: Unlabeled Data

$$\mathcal{L}_u = \frac{\sum M}{HW}\mathcal{L}_u^M + \frac{\sum(1-M)}{HW}\mathcal{L}_u^{1-M}.$$

3. Feature Alignment Loss: Semantic-Assisted Perception

$$\mathcal{L}_{feat} = 1 - \frac{1}{HW}\sum_{i=1}^{HW}\cos(f_i, f_i')$$

$$\boxed{\mathcal{L}_l} + \boxed{\mathcal{L}_u} + \boxed{\mathcal{L}_{feat}}$$



Depth Anything v2

Problem with v1 fine-grained Detail

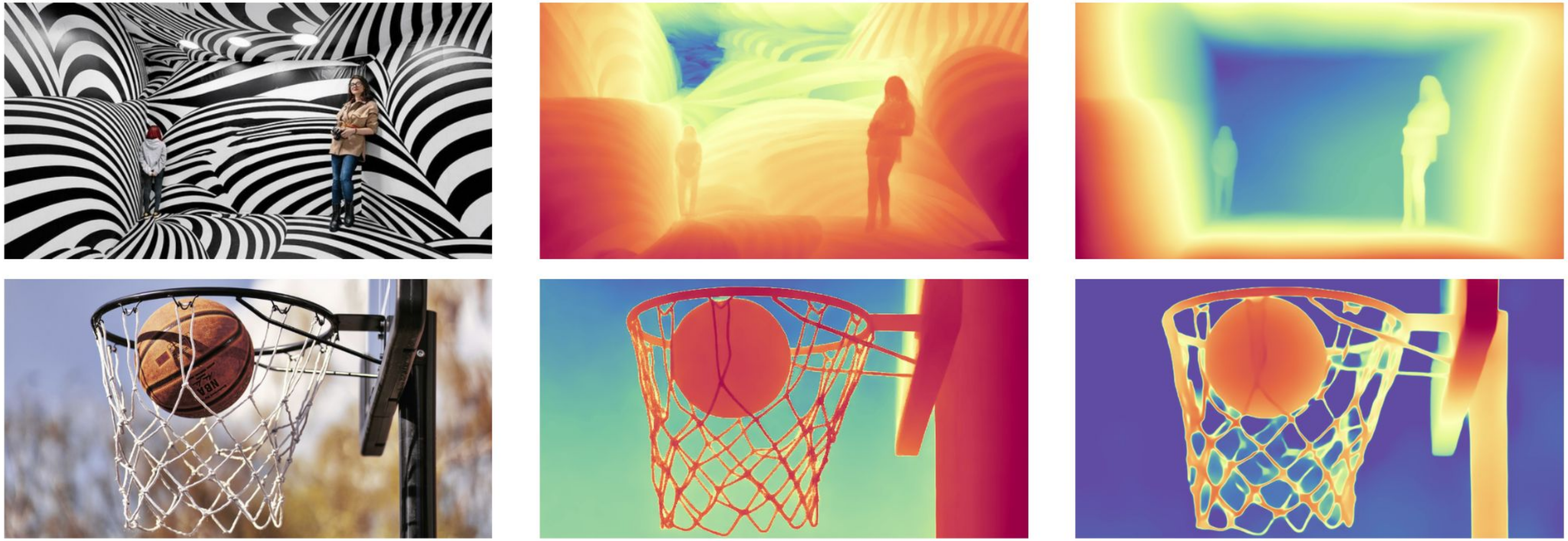
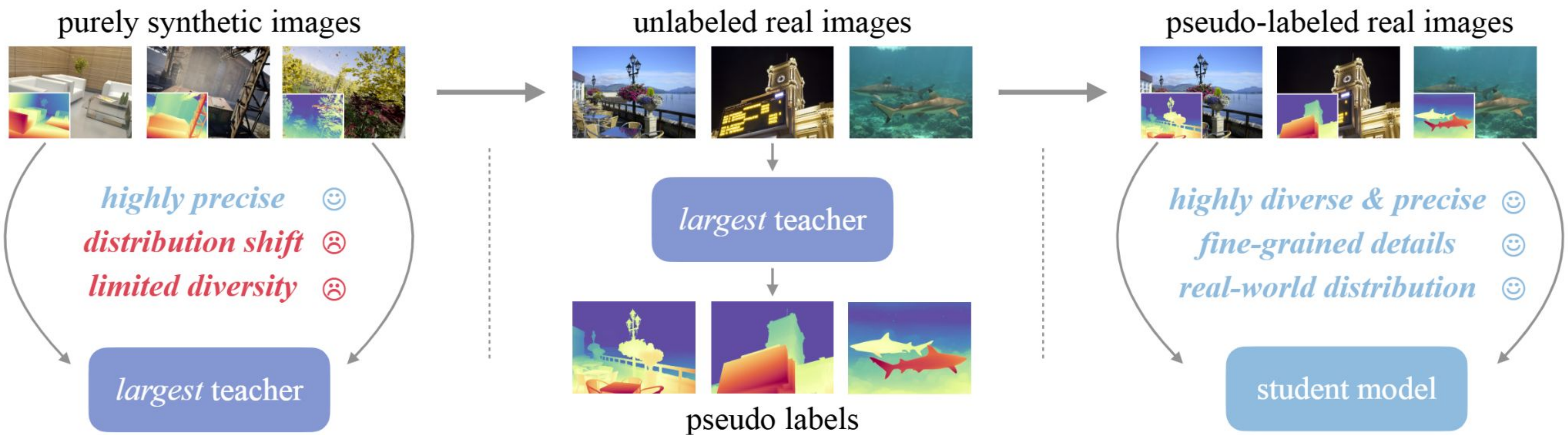


Image | Marigold [31] | Depth Anything V1 [89]

Yang, Lihe, et al. "Depth Anything V2." arXiv preprint arXiv:2406.09414 (2024).

Depth Anything v2



1. Replacing real images with synthetic images
2. Scaling up teacher model's capacity
3. Teach student model with large-scale real images

Yang, Lihe, et al. "Depth Anything V2." arXiv preprint arXiv:2406.09414 (2024).

Depth Anything v2

Image v1 v2

Yang, Lihe, et al. "Depth Anything V2." arXiv preprint arXiv:2406.09414 (2024).

Next Lecture:
Student Lecture 2
PointNets and 3D Networks

DeepRob

[Student] Lecture 1

by Tzu-Hsien Lee, Rammesh Adhav Saravanan , Fidan Mahmudova

RGB-D Networks and Manipulation

University of Minnesota