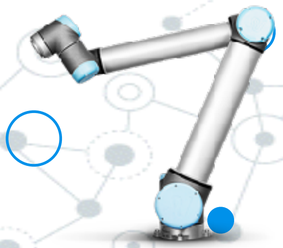
# Multi-step Pick-and-Place Tasks Using Object-Centric Dense Correspondences

**Chun-Yu Chai**, Keng-Fu Hsu, and Shiao-Li Tsao

2019/11/06


國立交通大學
National Chiao Tung University



# Pick-and-Place Tasks

- Sequential pick and place (Block stacking)

# Pick-and-Place Tasks

- Sequential pick and place (Block stacking)

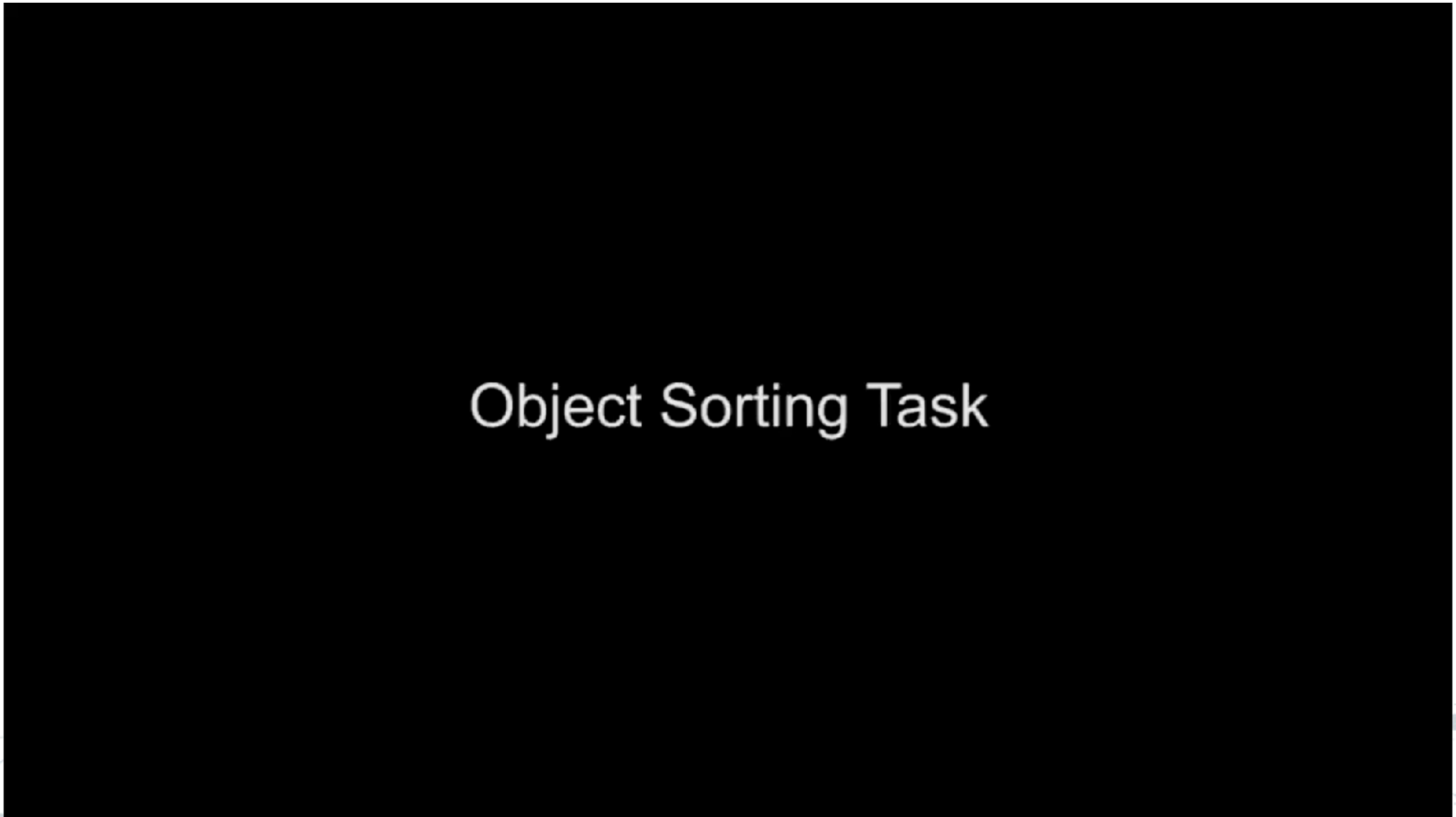
# Pick-and-Place Tasks

- Sequential pick and place (Block stacking)


| Block Stacking | $N_b = 3$ | $N_b = 4$ | $N_b = 5$ |
|---|---|---|---|
| Task Success Rate | 100% (30/30) | 96.67% (29/30) | 93.33% (28/30) |

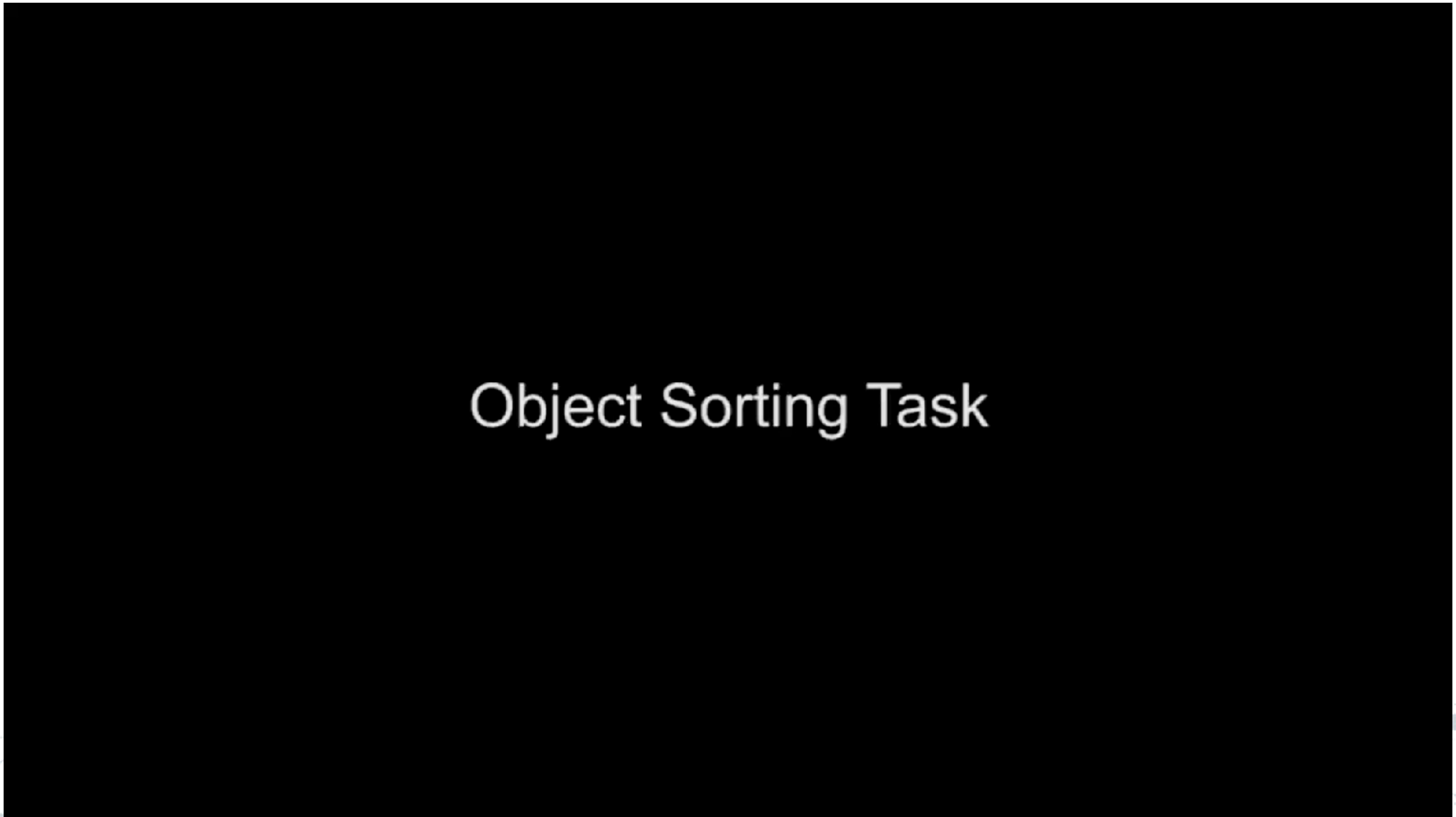
# Pick-and-Place Tasks

- Object sorting  (Cube/Carton/Doll)



# Pick-and-Place Tasks

- Object sorting  (Cube/Carton/Doll)



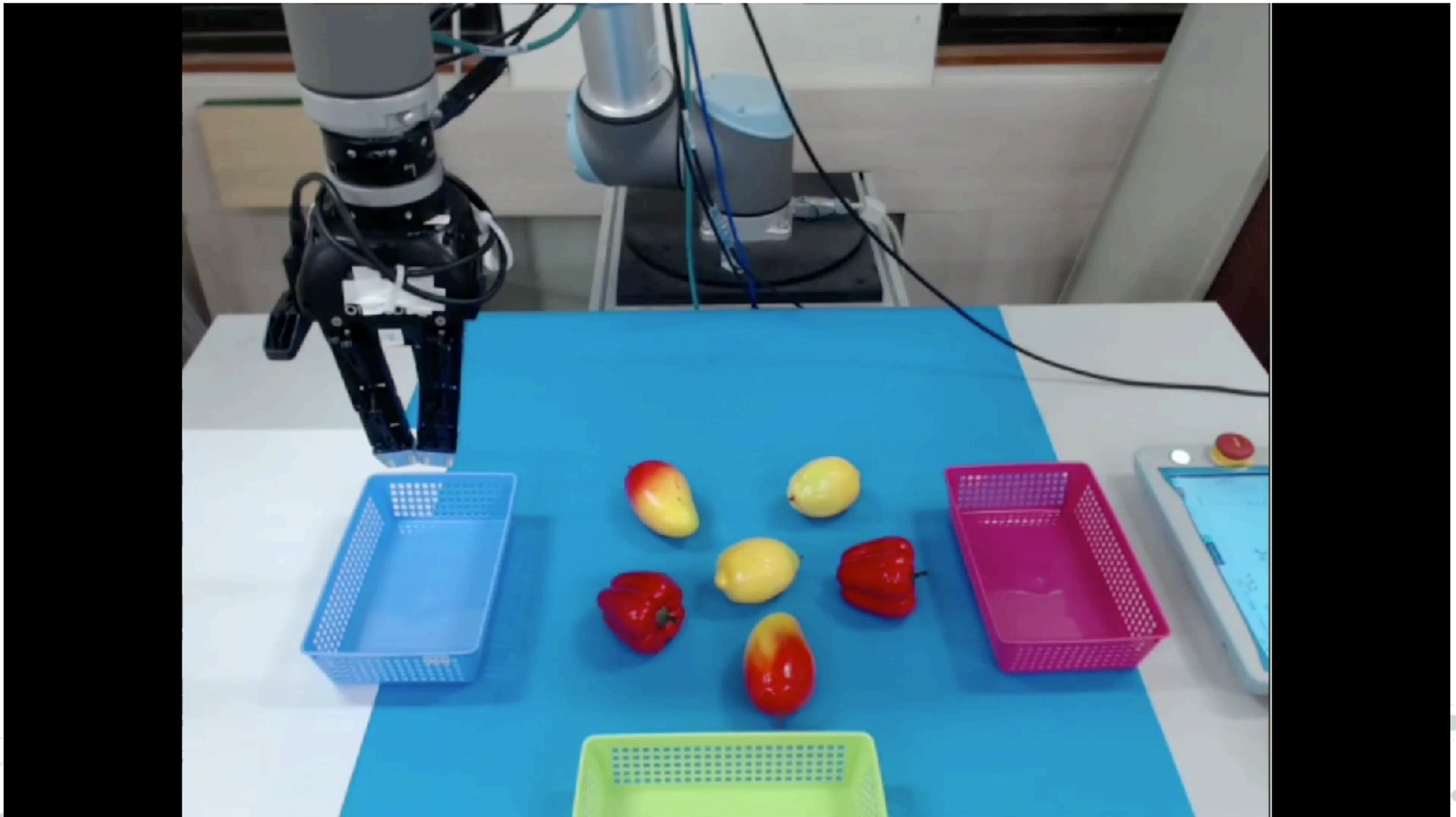
# Pick-and-Place Tasks

- Object sorting (Cube/Carton/Doll)


| Object Sorting | Carton, Fish Doll and Cube |
|---|---|
| Correct Placement Rate | 97.41% (263/270) |

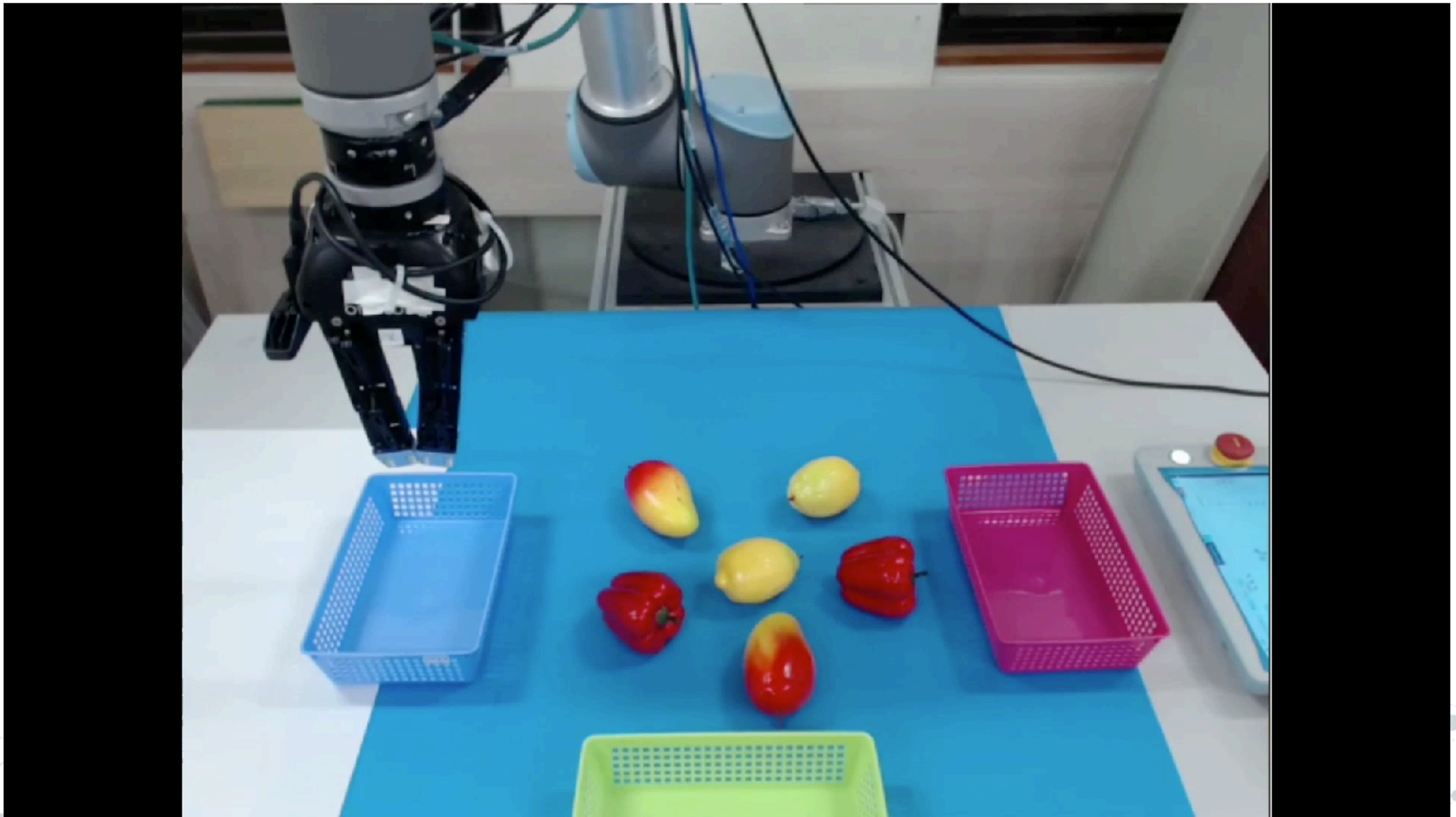
# Pick-and-Place Tasks

- Object sorting  (Mango/Pepper/Lemon)



# Pick-and-Place Tasks

- Object sorting  (Mango/Pepper/Lemon)

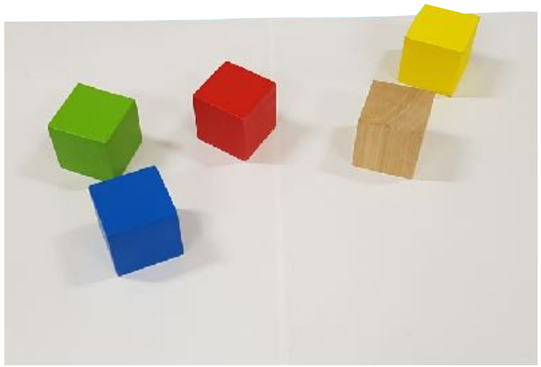


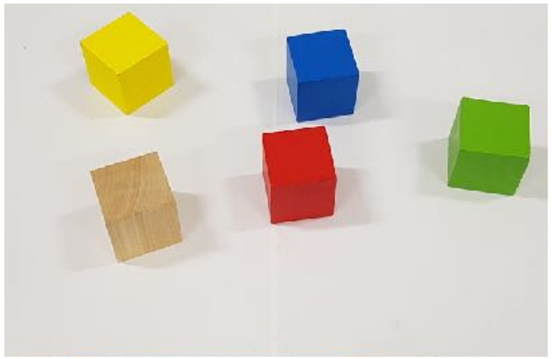
# Difficulties for Pick-and-Place Tasks

# Difficulties for Pick-and-Place Tasks

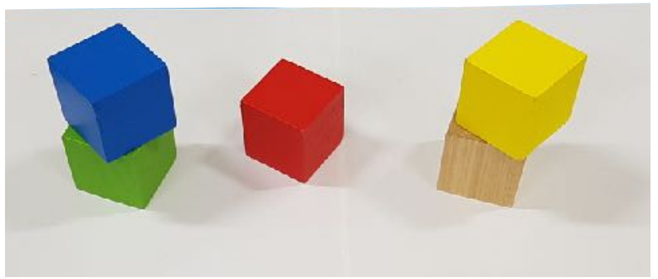
- Source and target combinations

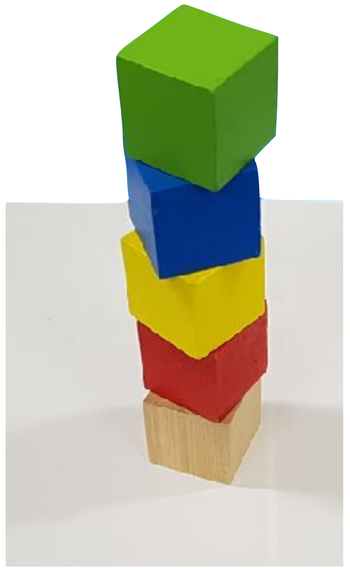
Source





Target

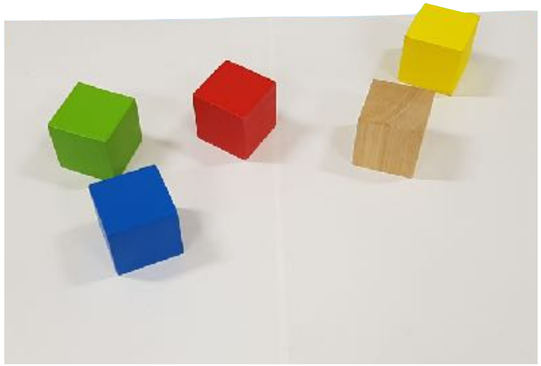


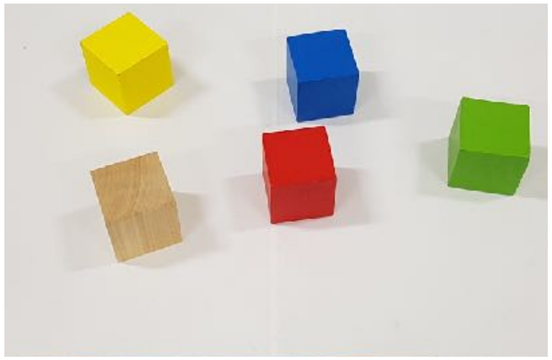


# Difficulties for Pick-and-Place Tasks

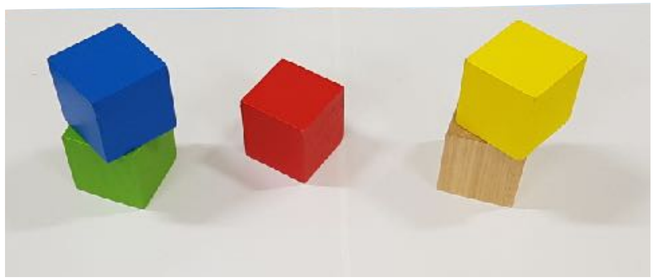
- Source and target combinations

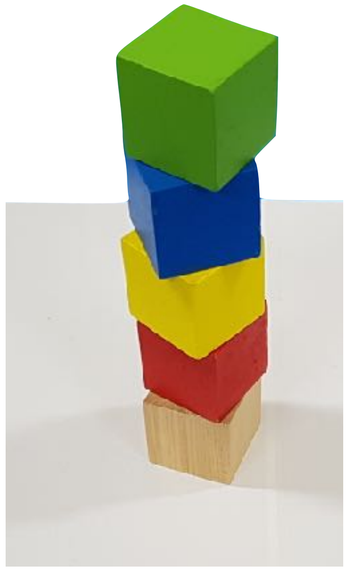
Source





Target





- Variations of object appearance

“

How to effectively assign the robot to perform a new pick-and-place task **without sophisticated re-programming or training ?**

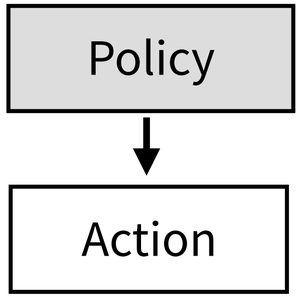
# Policies for Pick-and-Place Tasks

# Policies for Pick-and-Place Tasks

- Policy-level generalization
  - Duan et al, 2017 Nair et al, 2018

    Zhu et al, 2018 Tremblay et al, 2018

    Xu et al, 2017

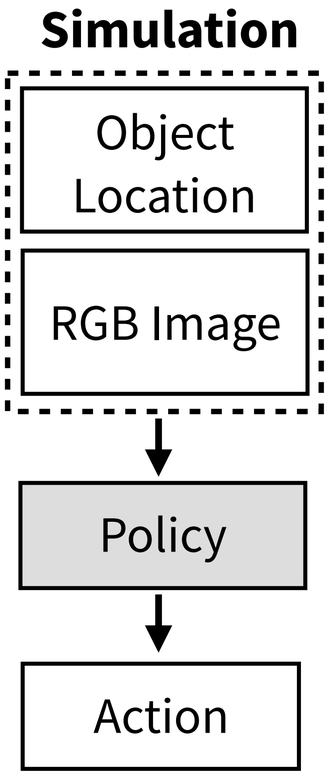


# Policies for Pick-and-Place Tasks

- Policy-level generalization
  - Duan et al, 2017 Nair et al, 2018

    Zhu et al, 2018 Tremblay et al, 2018

    Xu et al, 2017

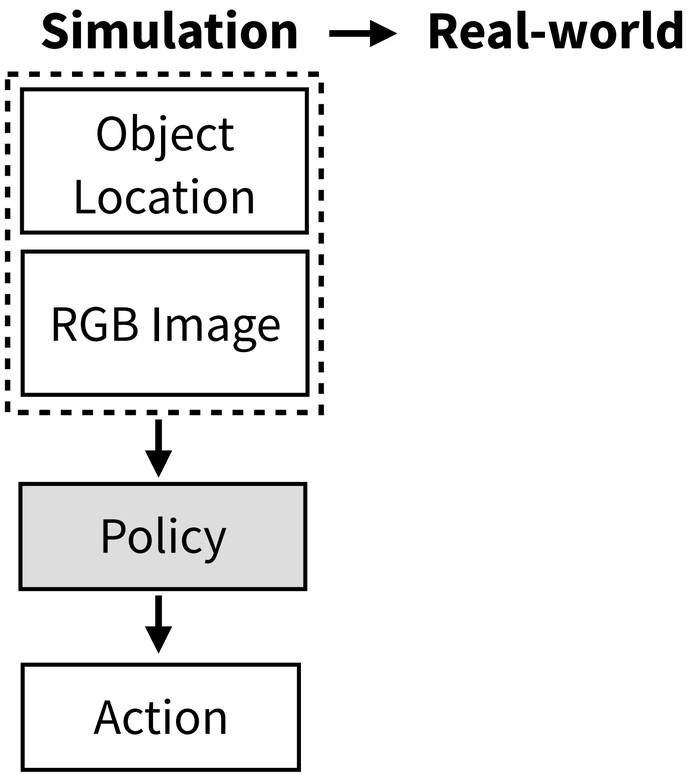


# Policies for Pick-and-Place Tasks

- Policy-level generalization
  - Duan et al, 2017 Nair et al, 2018

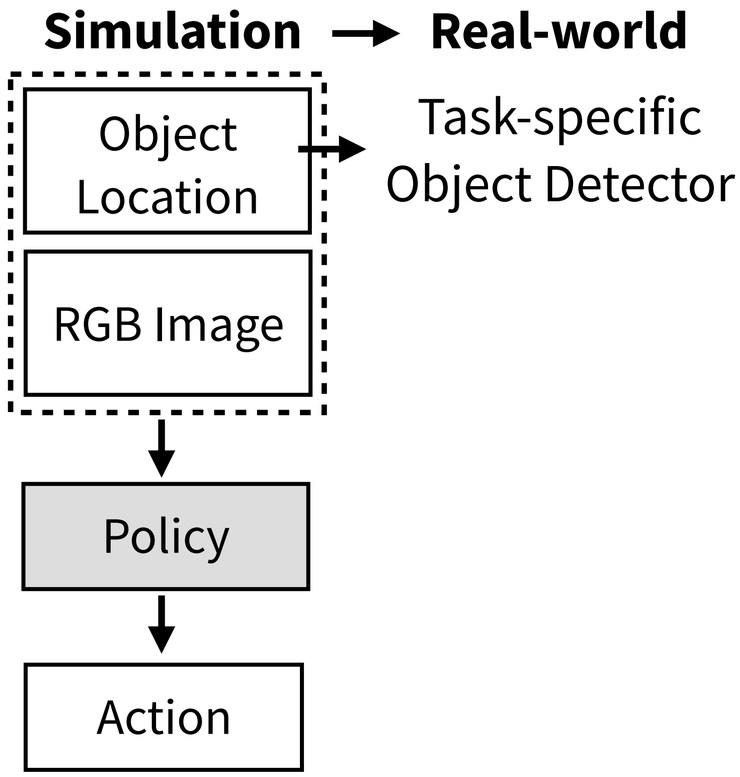
    Zhu et al, 2018 Tremblay et al, 2018

    Xu et al, 2017



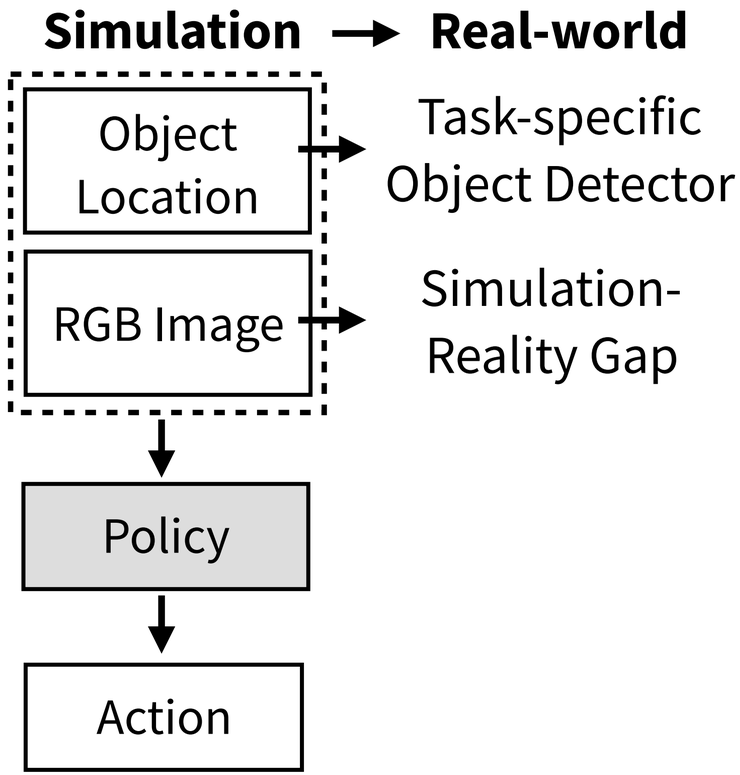
# Policies for Pick-and-Place Tasks

- Policy-level generalization
  - Duan et al, 2017 Nair et al, 2018
  - Zhu et al, 2018 Tremblay et al, 2018
  - Xu et al, 2017



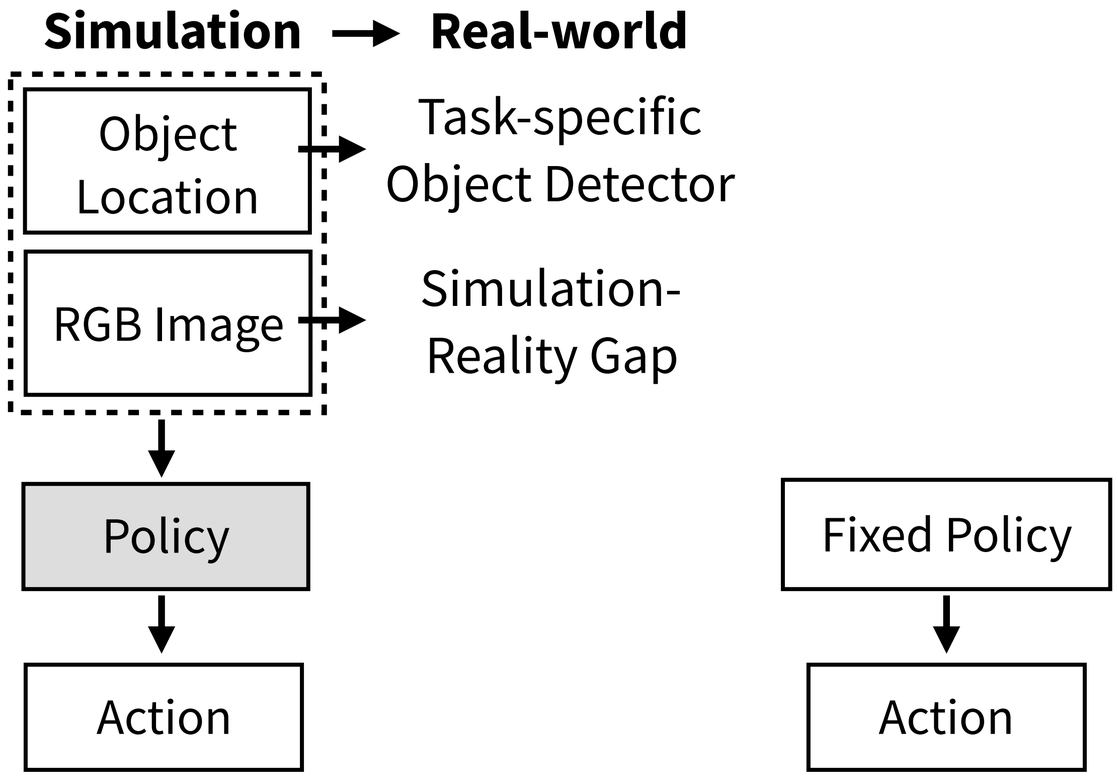
# Policies for Pick-and-Place Tasks

- Policy-level generalization
  - Duan et al, 2017 Nair et al, 2018
  - Zhu et al, 2018 Tremblay et al, 2018
  - Xu et al, 2017



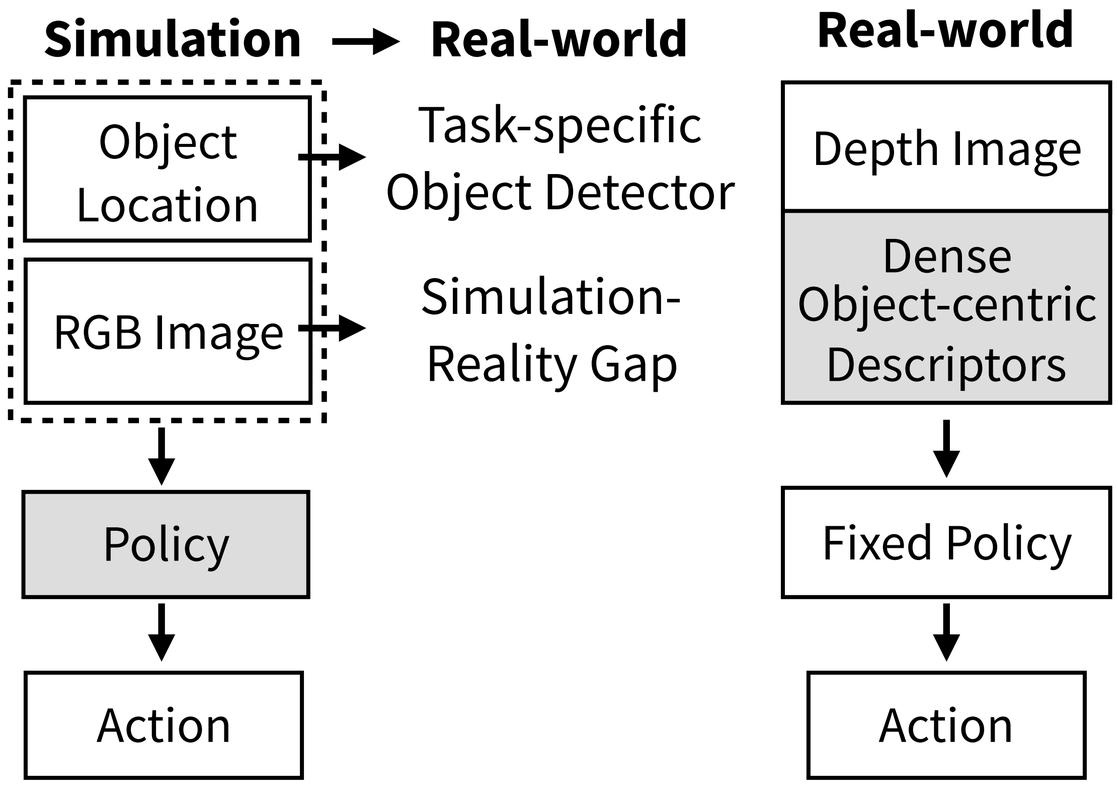
# Policies for Pick-and-Place Tasks

- Policy-level generalization
  - Duan et al, 2017 Nair et al, 2018
  - Zhu et al, 2018 Tremblay et al, 2018
  - Xu et al, 2017
- **Object-level generalization**



# Policies for Pick-and-Place Tasks

- Policy-level generalization
  - Duan et al, 2017 Nair et al, 2018
  - Zhu et al, 2018 Tremblay et al, 2018
  - Xu et al, 2017
- **Object-level generalization**
  - Florence et al, 2018



# Multiclass Dense Object Nets

# Learning Objectives-1

# Learning Objectives-1

- **Intra-class variations**: geometric correspondence[1]

[1] P. R. Florence, L. Manuelli, and R. Tedrake, "**Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation**." In CoRL, 2018.

# Learning Objectives-1

- **Intra-class variations**: geometric correspondence[1]
  - Find a specific geometry point for an object in different orientations

[1] P. R. Florence, L. Manuelli, and R. Tedrake, “**Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation**.” In CoRL, 2018.

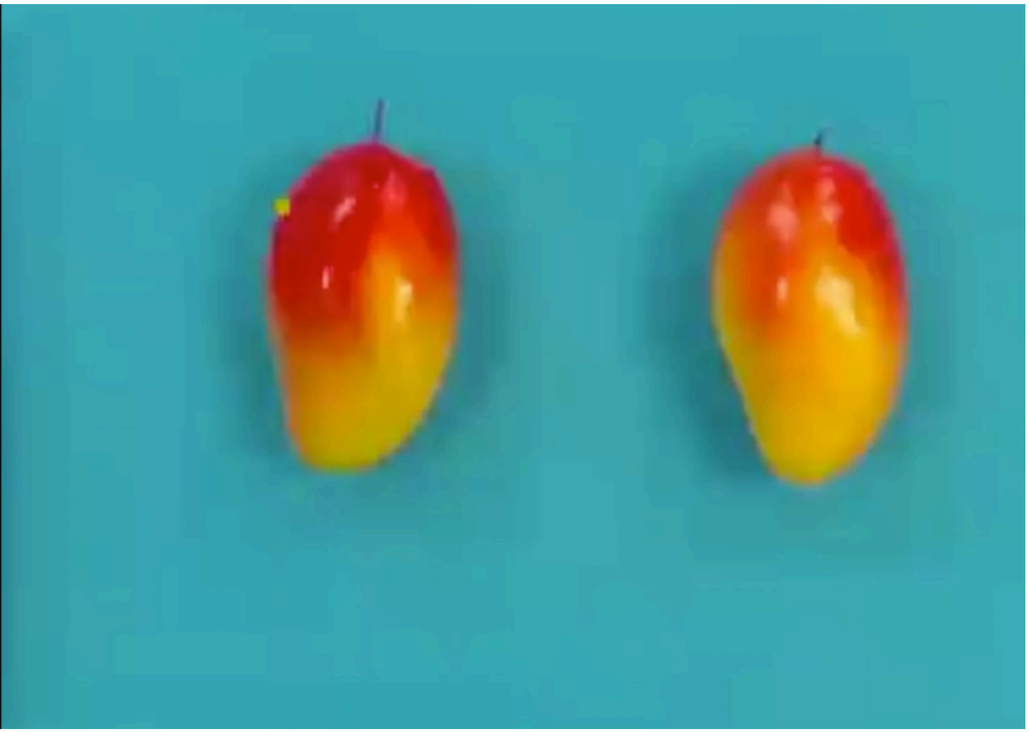
# Learning Objectives-1

- **Intra-class variations**: geometric correspondence[1]
  - Find a specific geometry point for an object in different orientations
  - Find similar points using the (inherently learned) intra-class consistency

[1] P. R. Florence, L. Manuelli, and R. Tedrake, "**Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation**." In CoRL, 2018.

# Learning Objectives-1

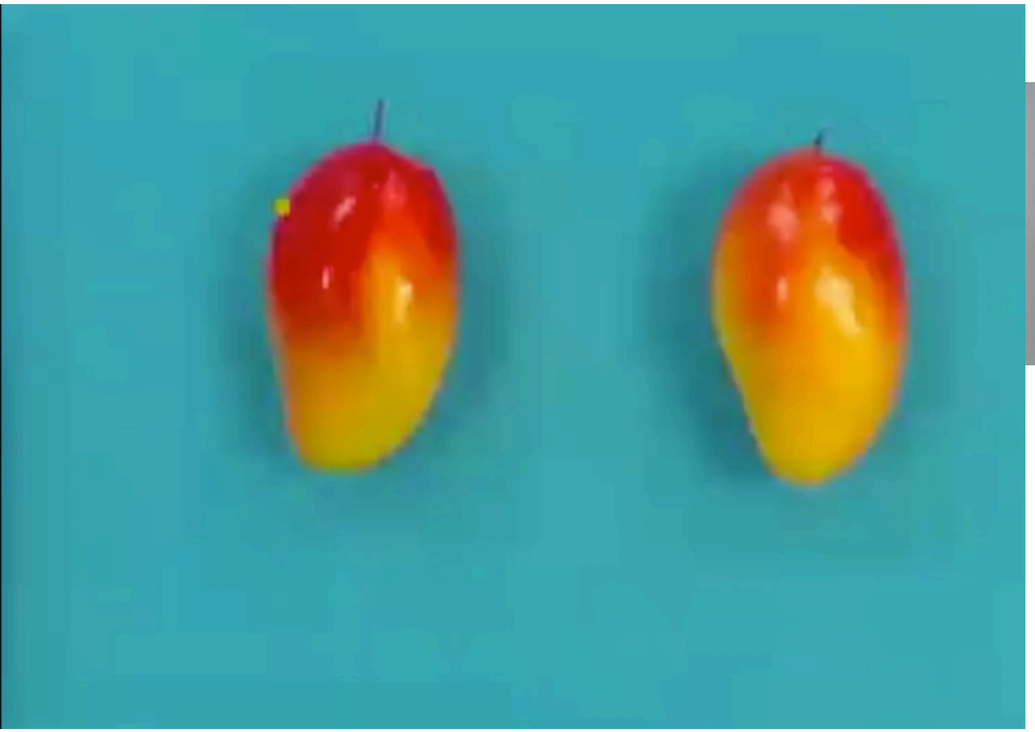
- **Intra-class variations**: geometric correspondence[1]
  - Find a specific geometry point for an object in different orientations
  - Find similar points using the (inherently learned) intra-class consistency



[1] P. R. Florence, L. Manuelli, and R. Tedrake, "**Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation**." In CoRL, 2018.

# Learning Objectives-1

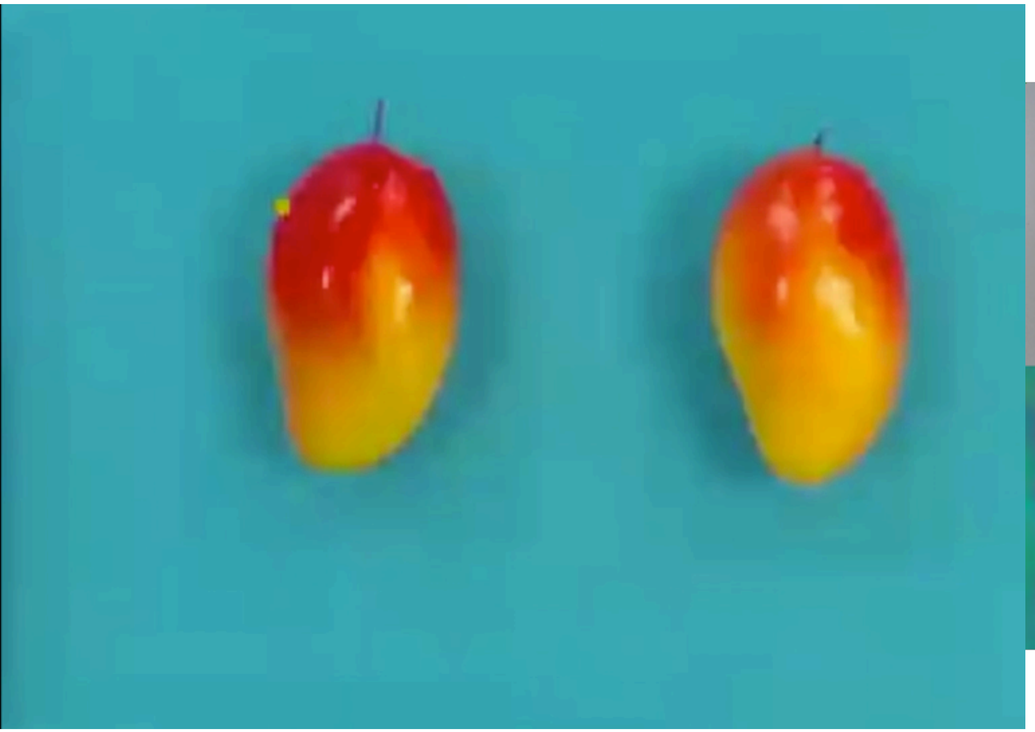
- **Intra-class variations**: geometric correspondence[1]
  - Find a specific geometry point for an object in different orientations
  - Find similar points using the (inherently learned) intra-class consistency

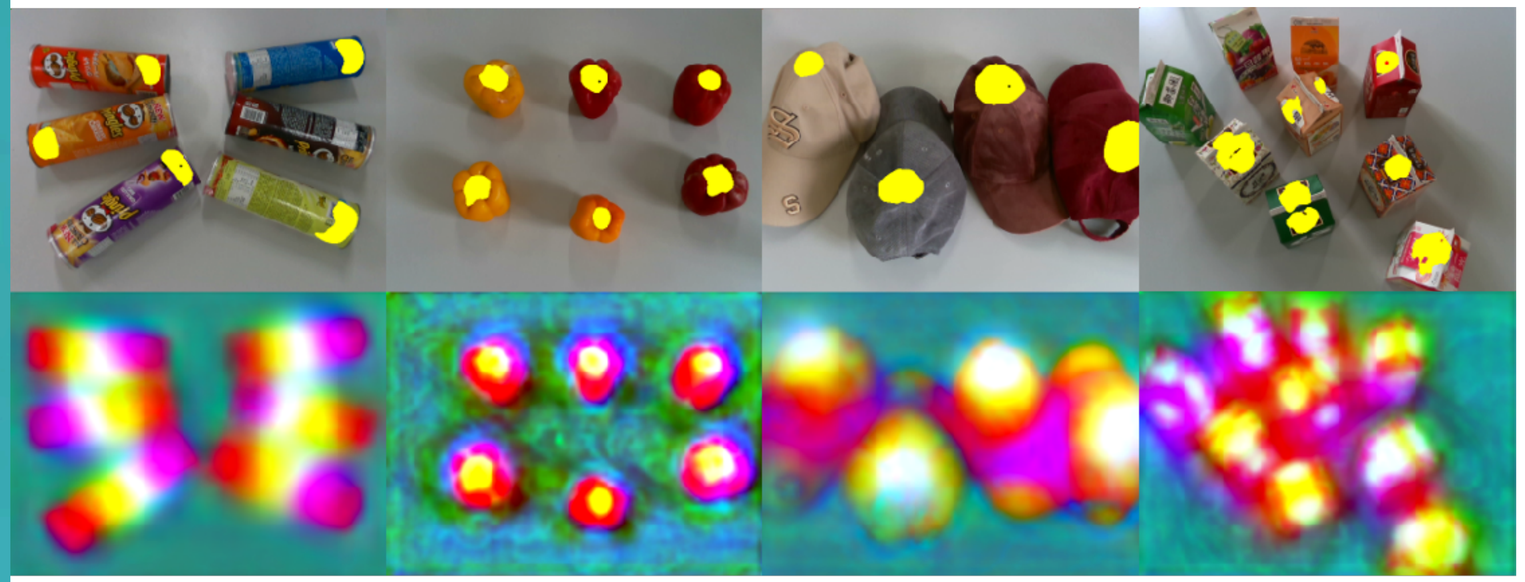


[1] P. R. Florence, L. Manuelli, and R. Tedrake, "**Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation**." In CoRL, 2018.

# Learning Objectives-1

- **Intra-class variations**: geometric correspondence[1]
  - Find a specific geometry point for an object in different orientations
  - Find similar points using the (inherently learned) intra-class consistency





[1] P. R. Florence, L. Manuelli, and R. Tedrake, "**Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation**." In CoRL, 2018.

# Learning Objectives-2

# Learning Objectives-2

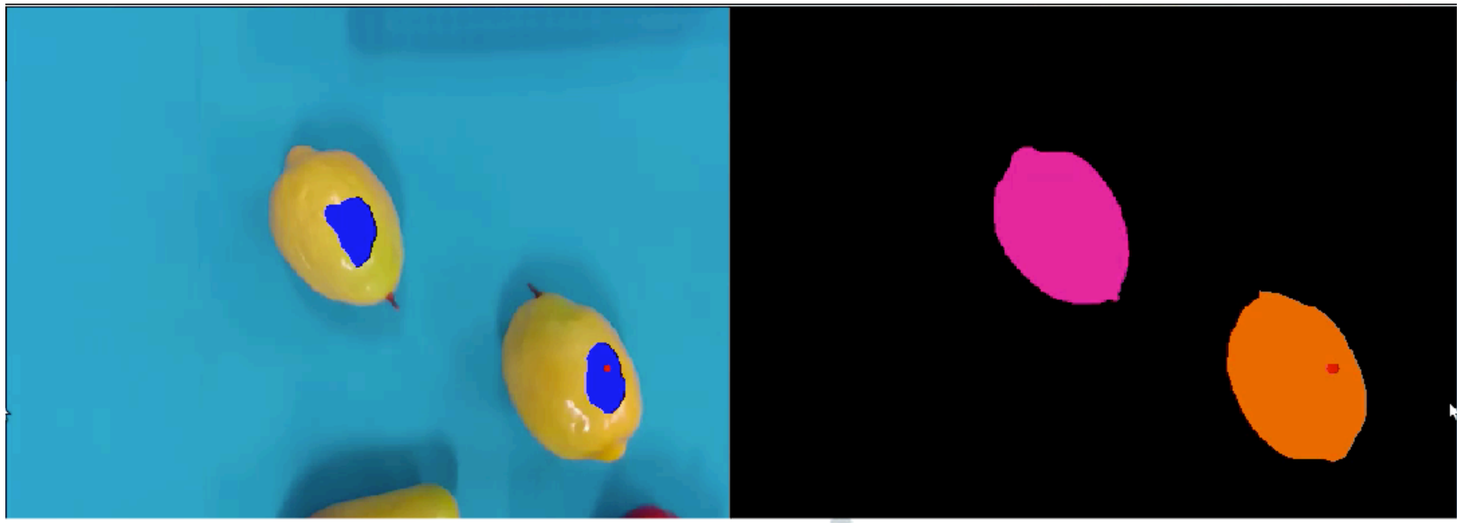
- **Inter-class separation**: semantic segmentation

# Learning Objectives-2

- **Inter-class separation**: semantic segmentation
  - Perform instance segmentation (e.g., Watershed algorithm)

# Learning Objectives-2

- **Inter-class separation**: semantic segmentation
  - Perform instance segmentation (e.g., Watershed algorithm)

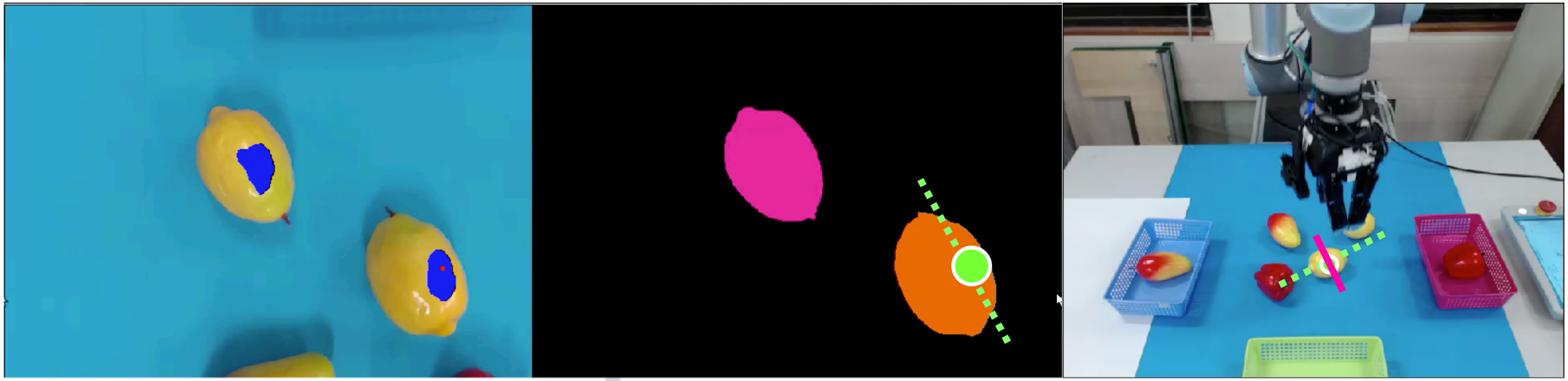


# Learning Objectives-2

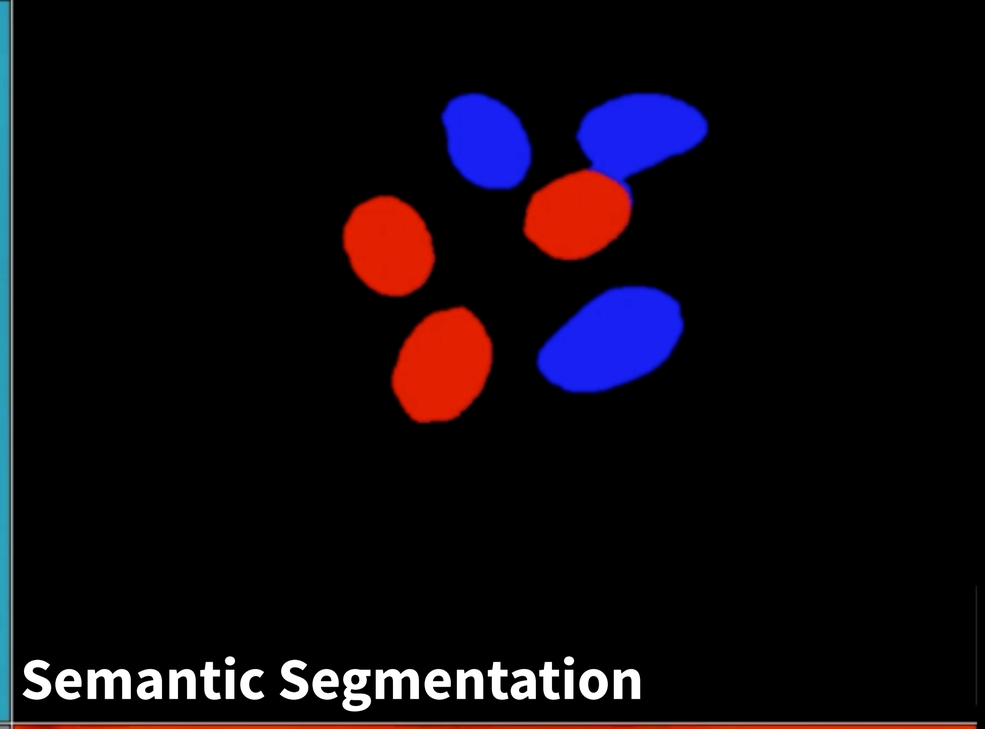
- **Inter-class separation**: semantic segmentation
  - Perform instance segmentation (e.g., Watershed algorithm)
  - Determine feasible grasping poses for an object

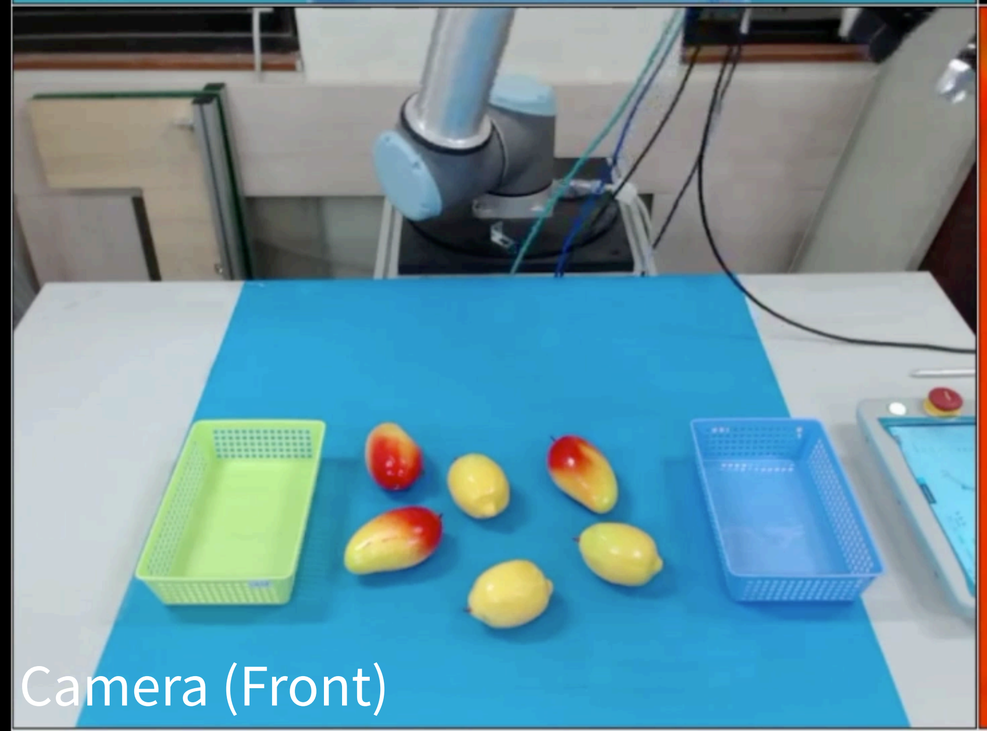
# Learning Objectives-2

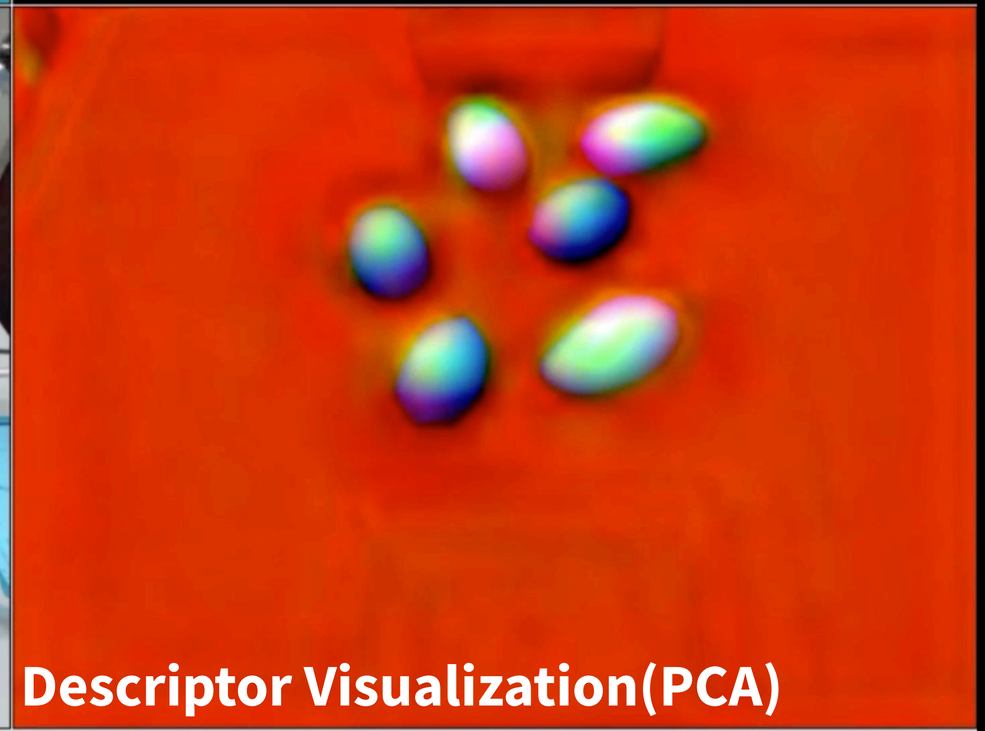
- **Inter-class separation**: semantic segmentation
  - Perform instance segmentation (e.g., Watershed algorithm)
  - Determine feasible grasping poses for an object



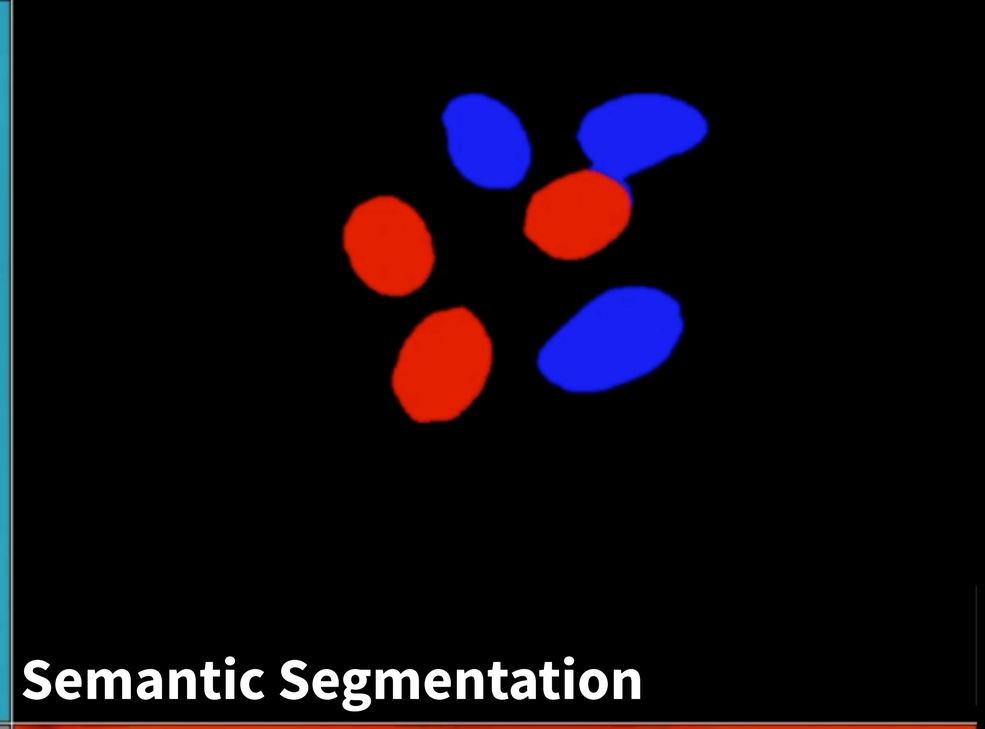

Camera (Arm)

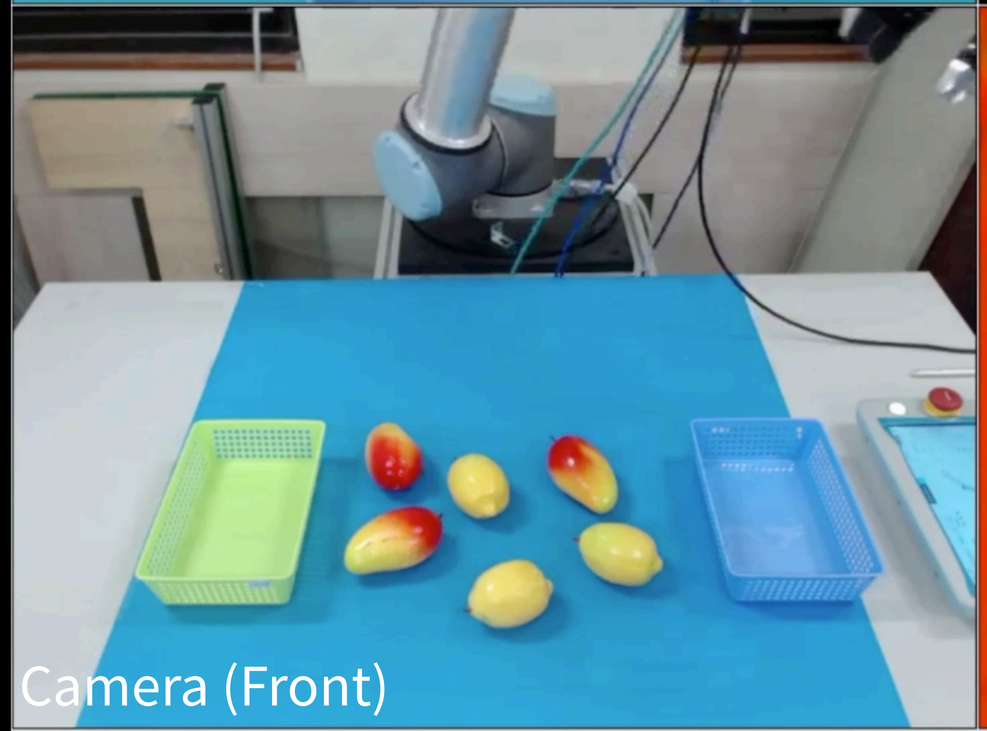

Semantic Segmentation

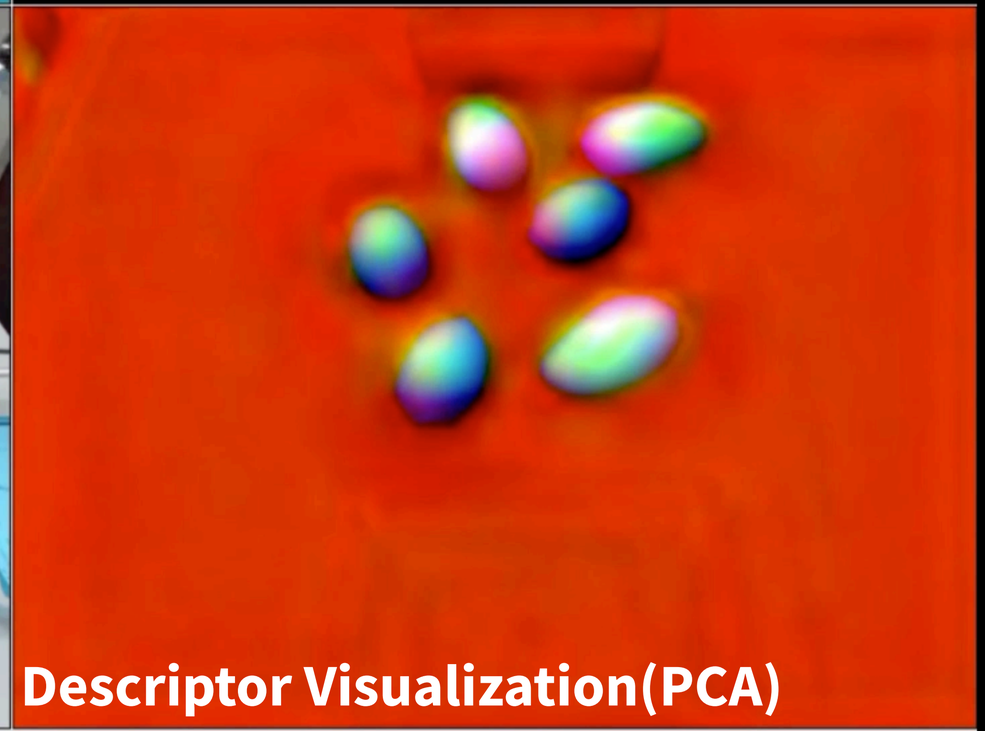

Camera (Front)

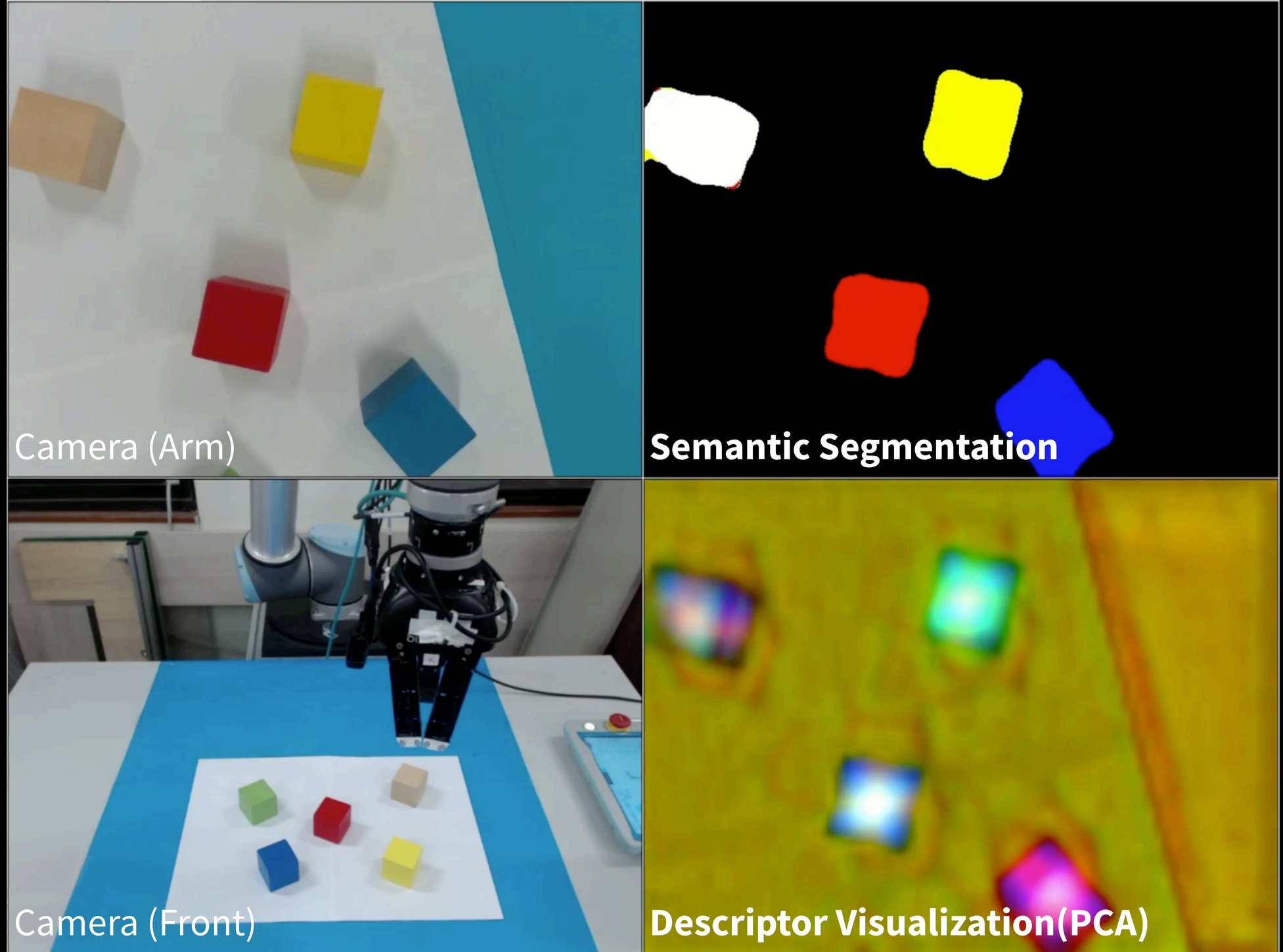

Descriptor Visualization(PCA)

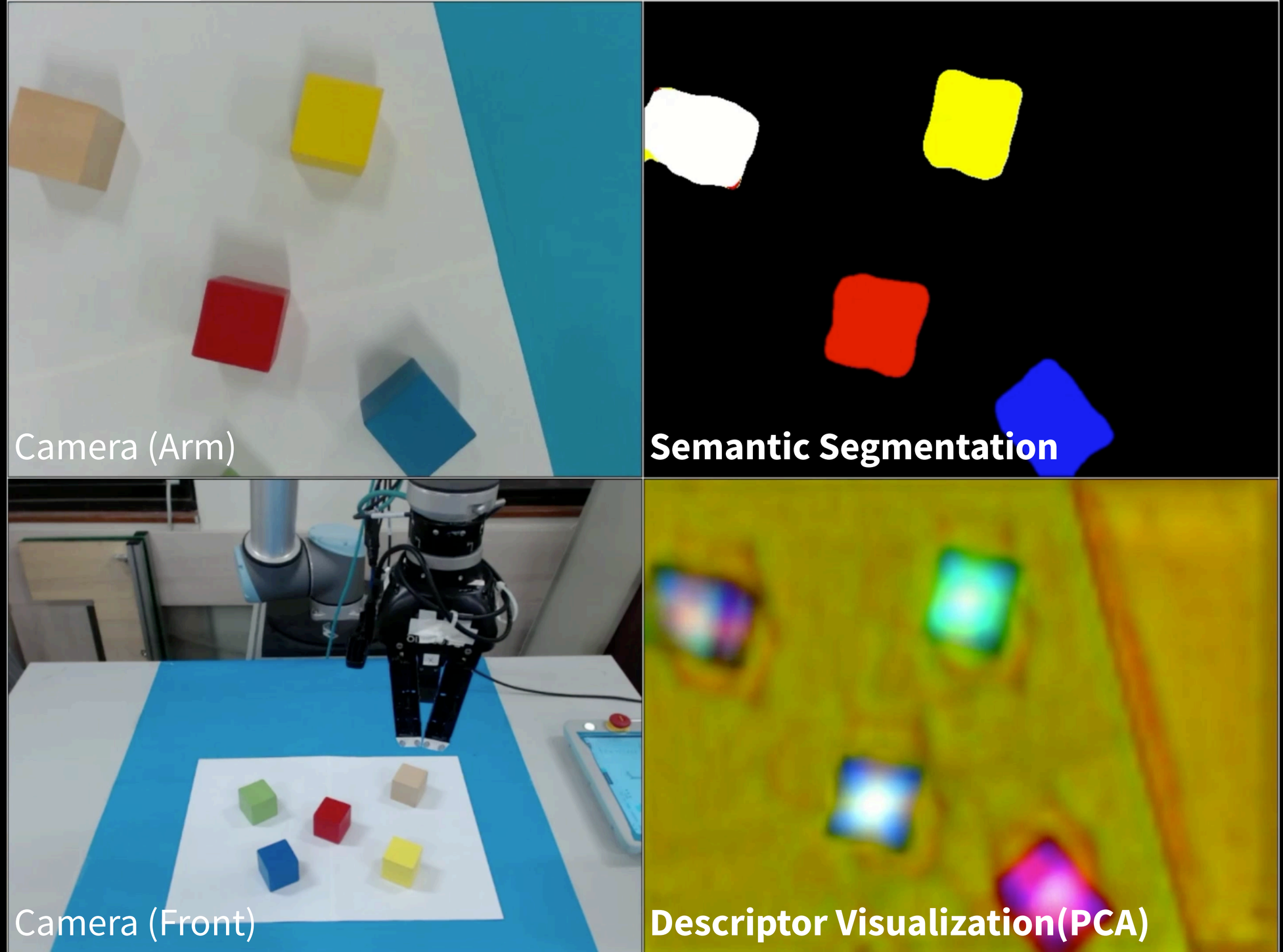

Camera (Arm)


Semantic Segmentation

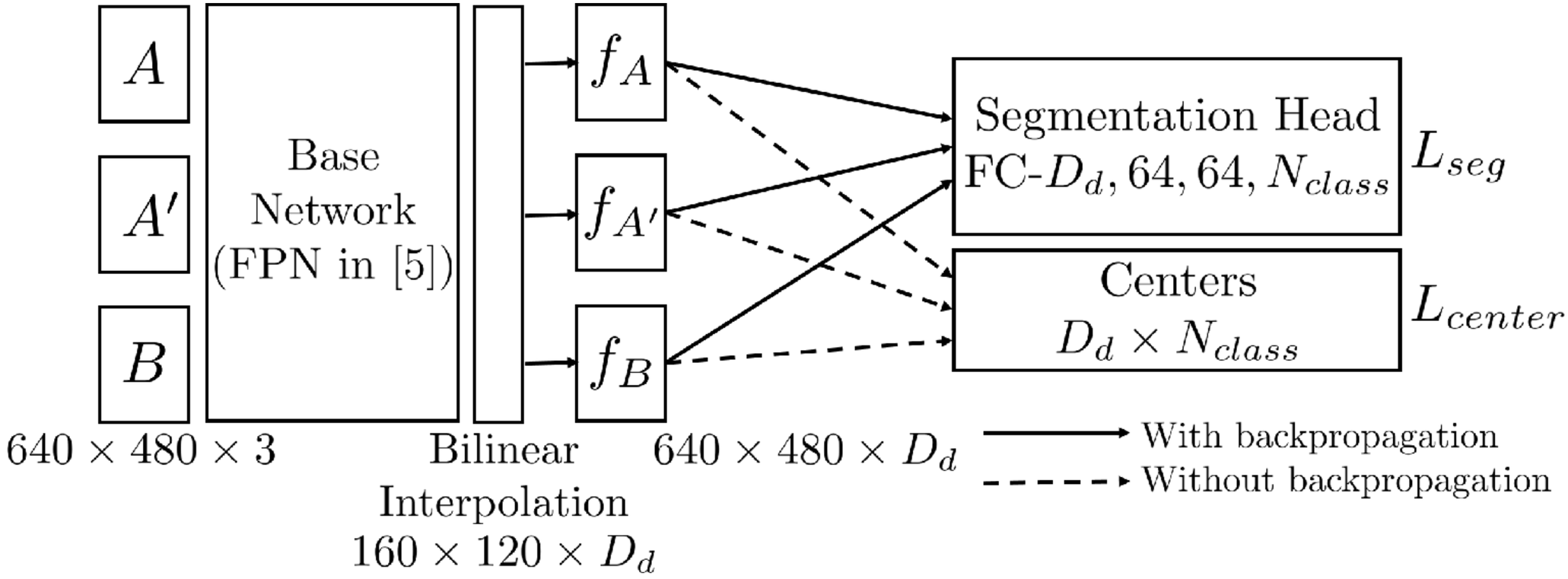

Camera (Front)


Descriptor Visualization(PCA)

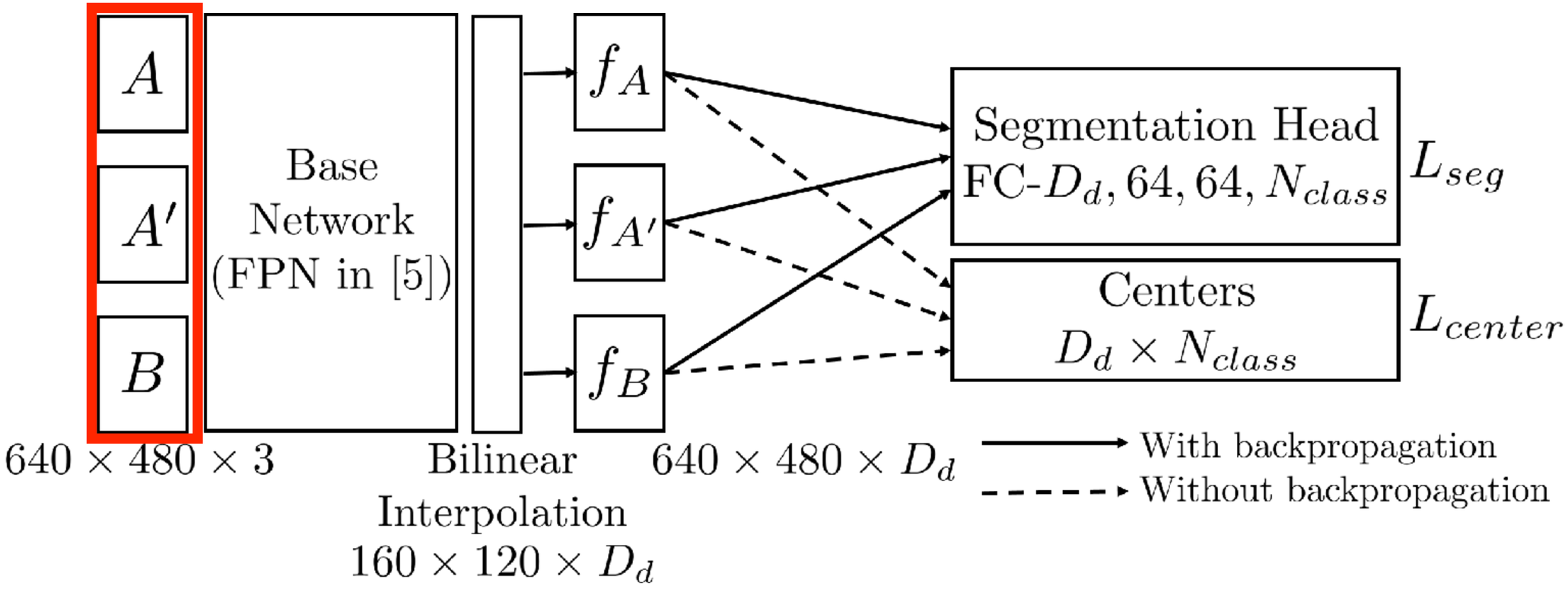

Camera (Arm)
Semantic Segmentation
Camera (Front)
Descriptor Visualization(PCA)


Camera (Arm)
Semantic Segmentation
Camera (Front)
Descriptor Visualization(PCA)

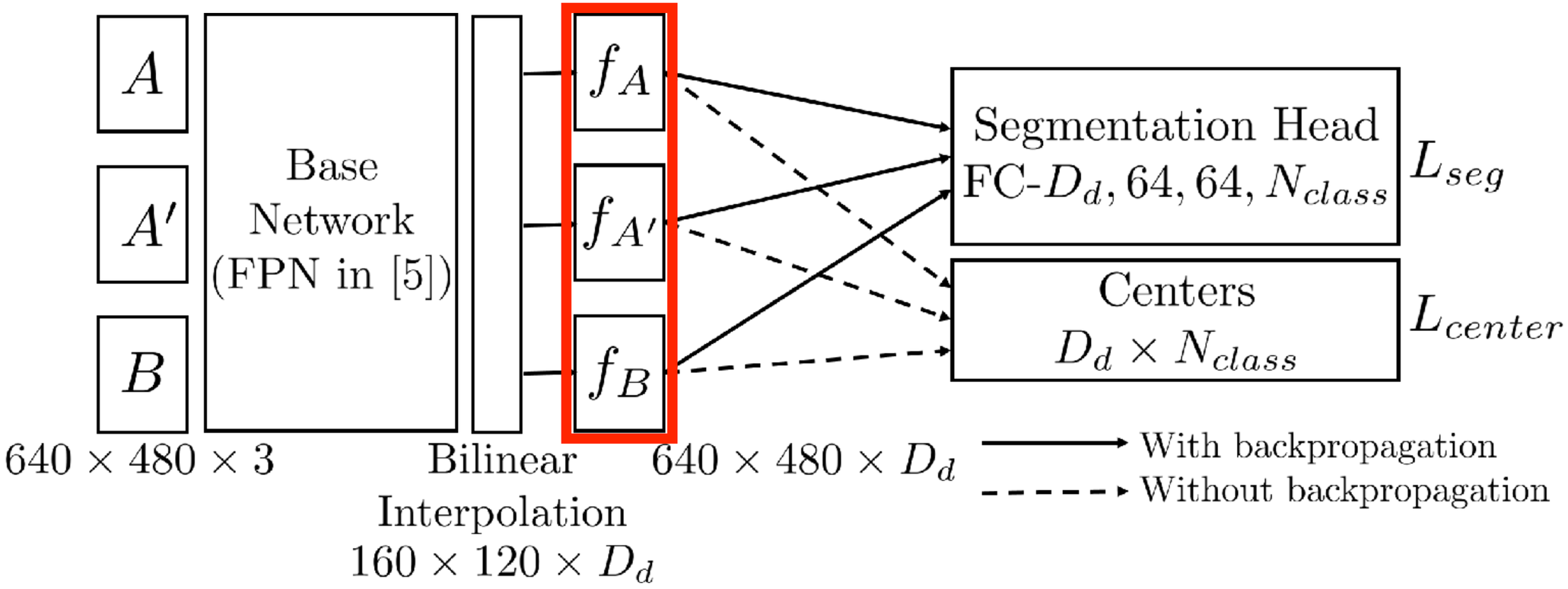
# Multiclass Dense Object Nets - Model



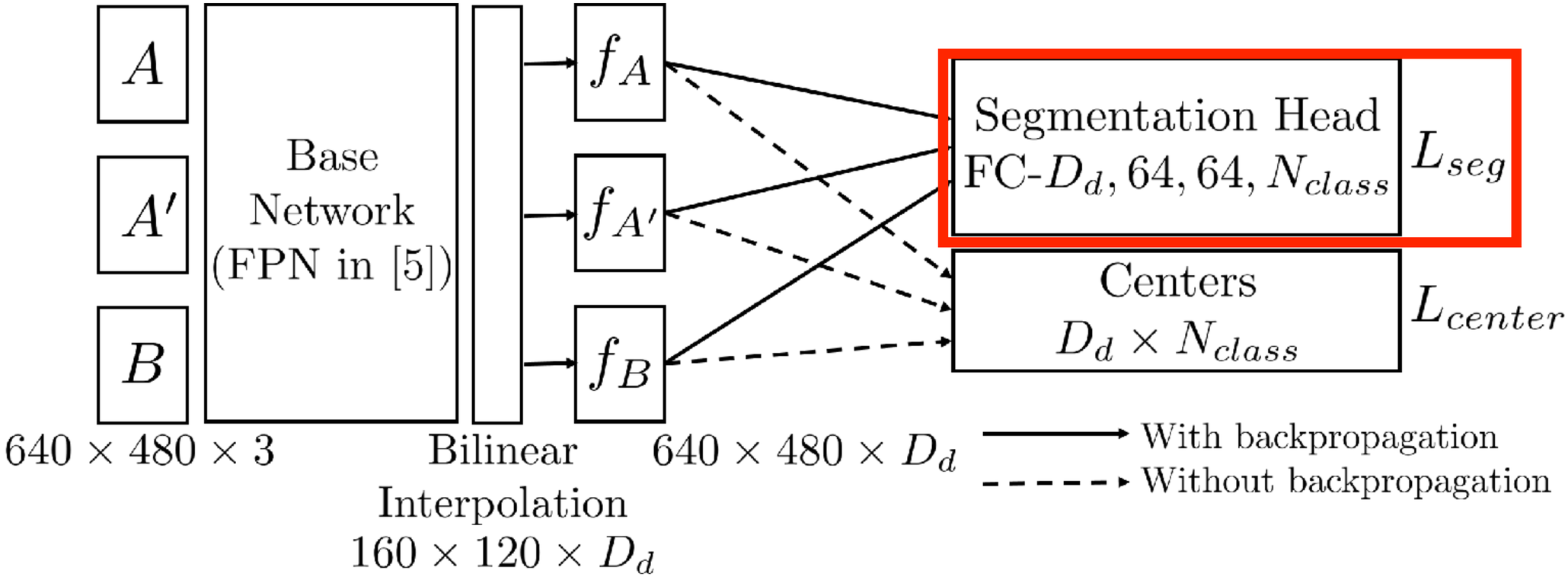
# Multiclass Dense Object Nets - Model



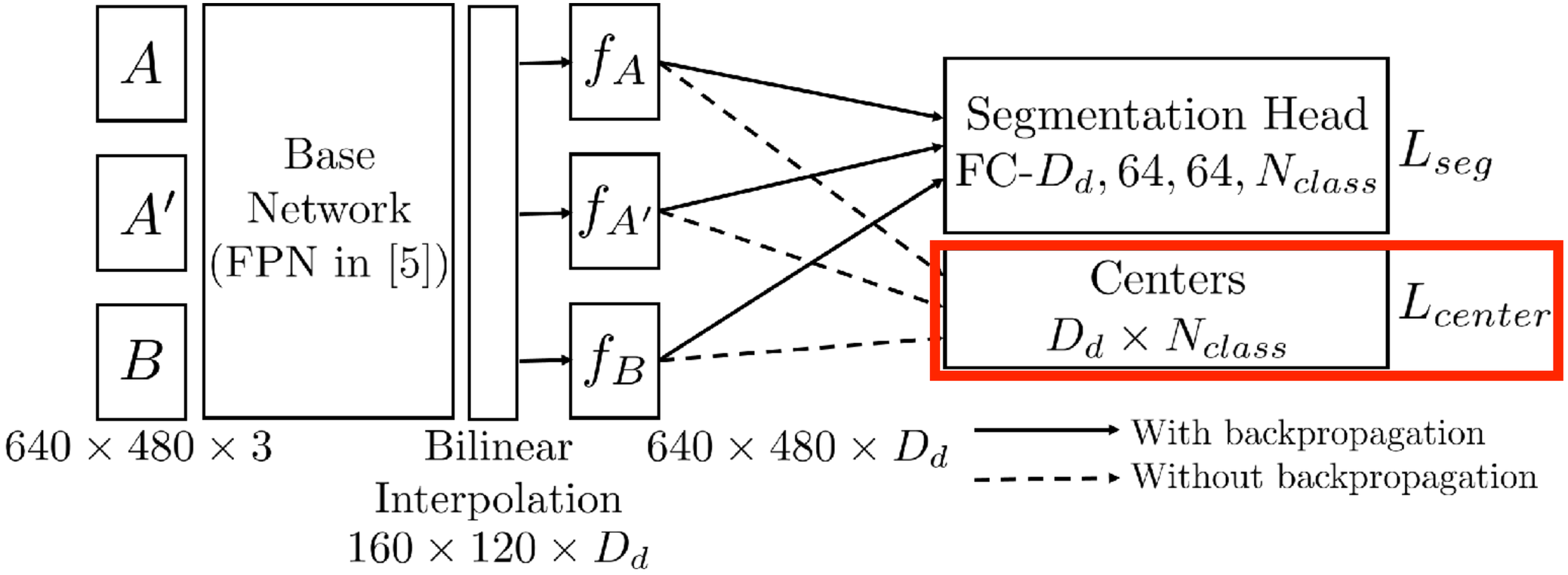
# Multiclass Dense Object Nets - Model



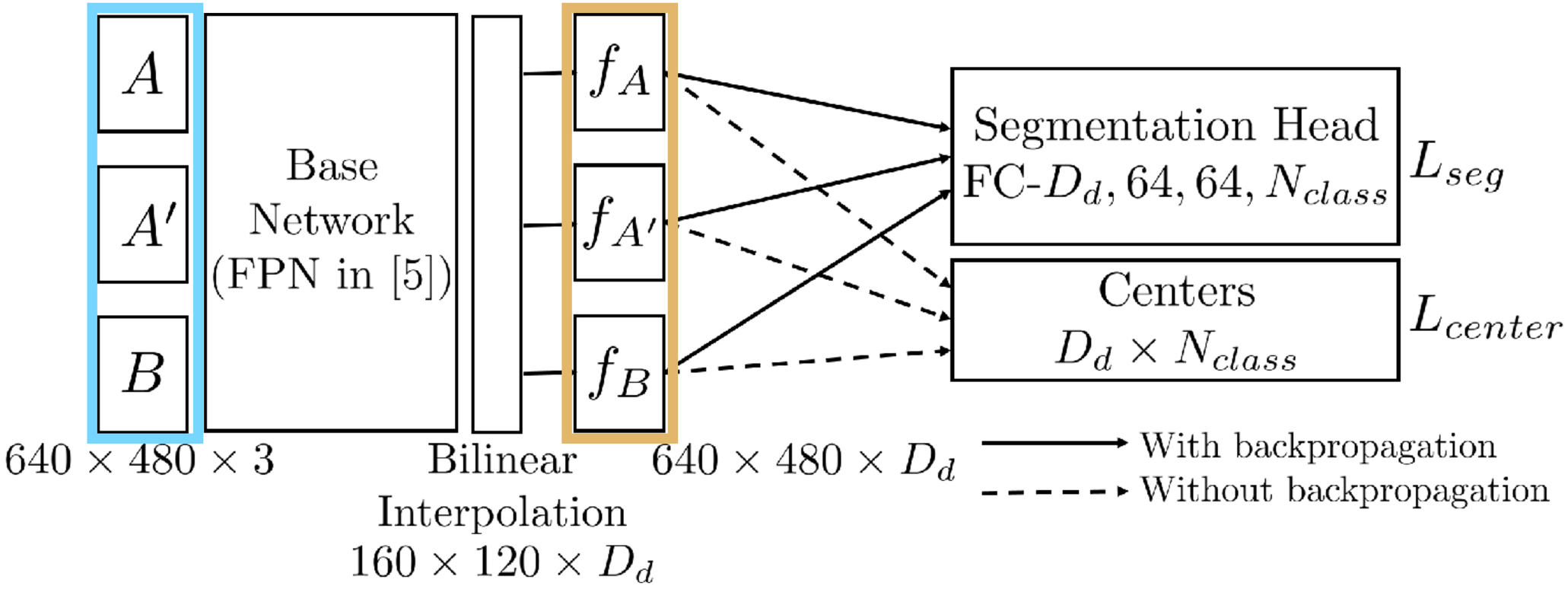
# Multiclass Dense Object Nets - Model



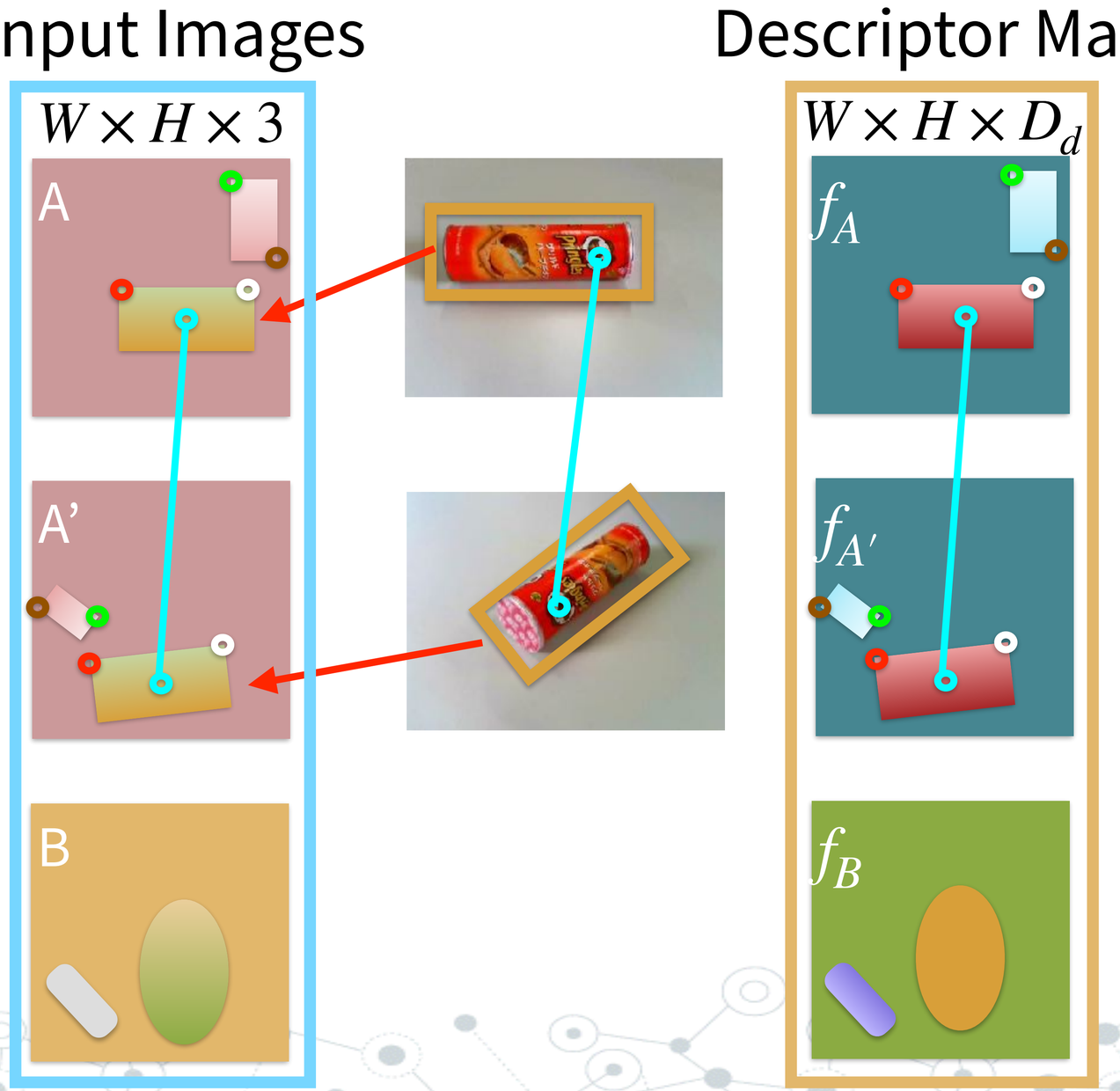
# Multiclass Dense Object Nets - Model



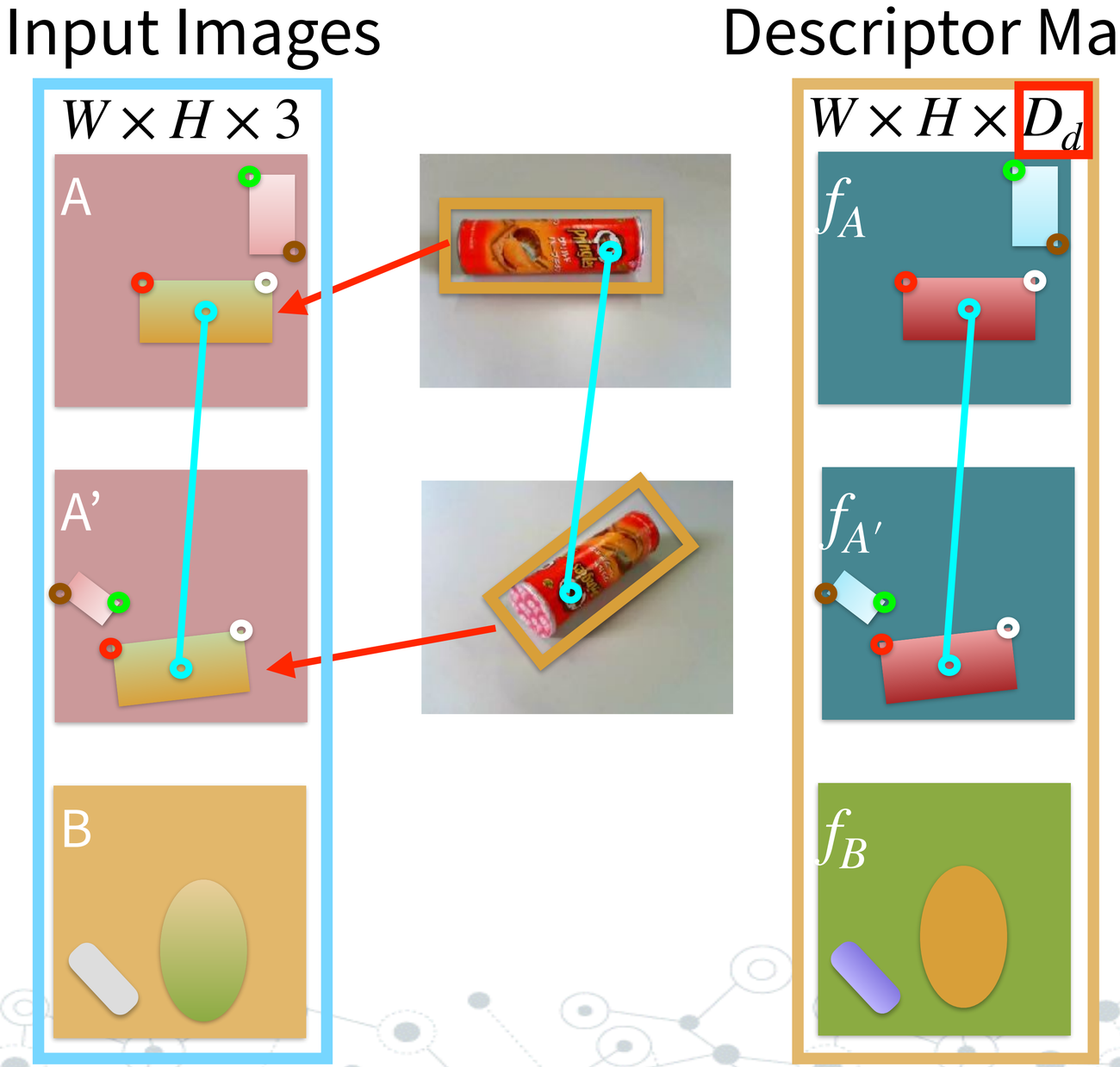
# Multiclass Dense Object Nets - Model



# Data Organization

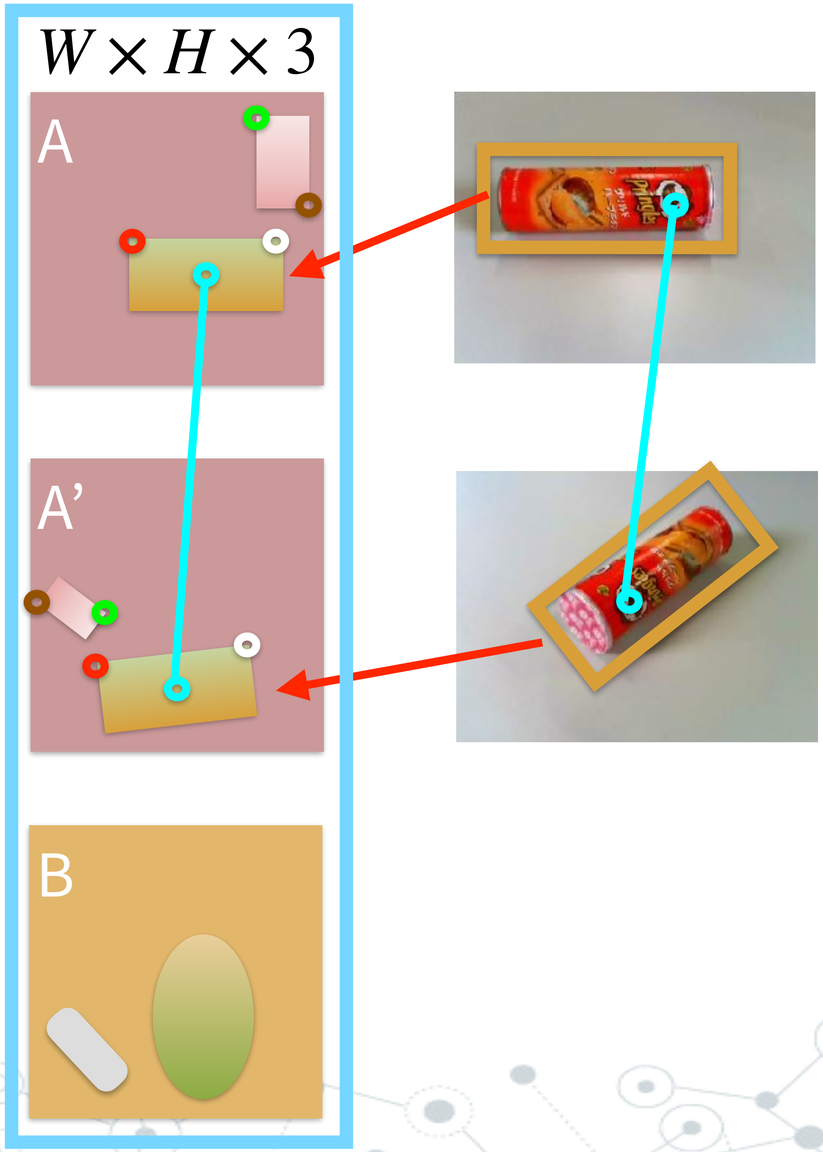


# Data Organization

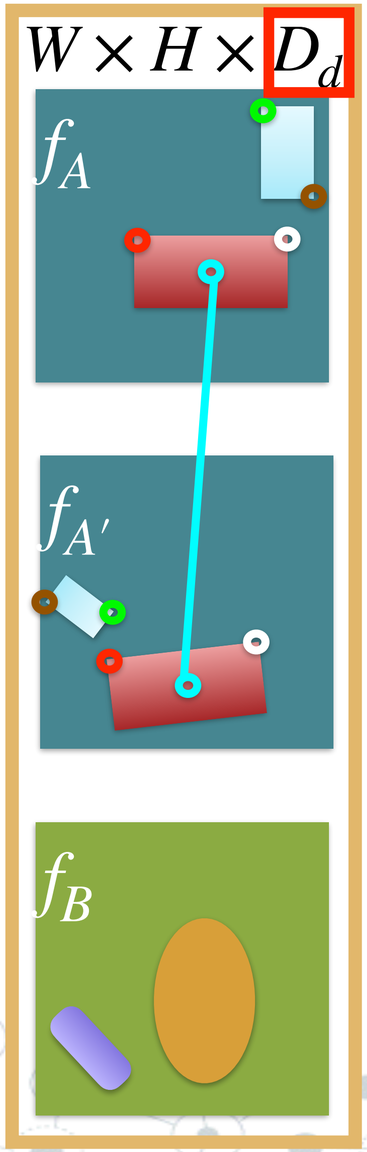


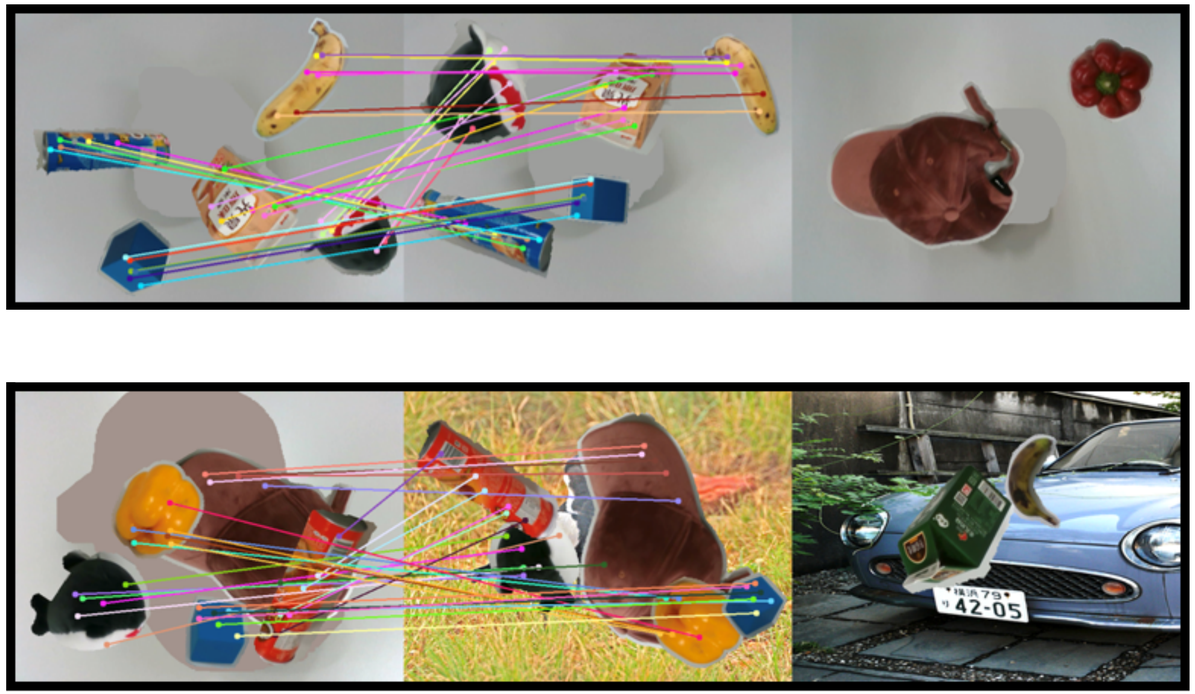
# Data Organization

## Input Images




## Descriptor Maps





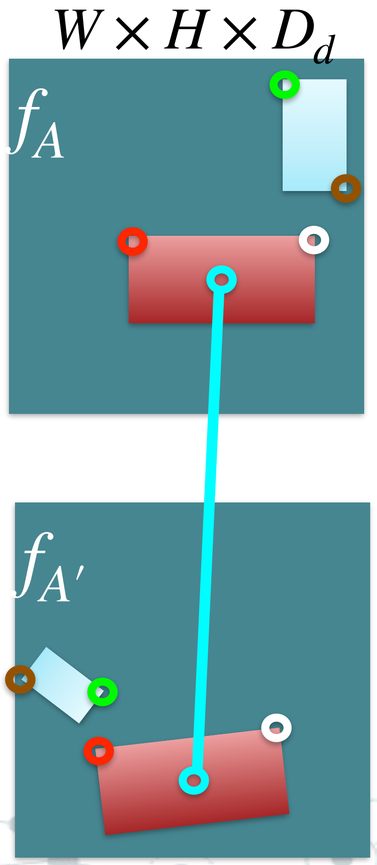
Synthetic image $A$ Synthetic image $A'$ Synthetic image $B$

# Preliminary: Dense Object Nets [1]

# Preliminary: Dense Object Nets [1]

- Contrastive loss for match points

$$D(f_A(\circ), f_{A'}(\circ))^2 = \left\| \begin{bmatrix} x_1 \\ \vdots \\ x_{D_d} \end{bmatrix}_{f_A} - \begin{bmatrix} x_1 \\ \vdots \\ x_{D_d} \end{bmatrix}_{f_{A'}} \right\|_2^2 \doteq$$

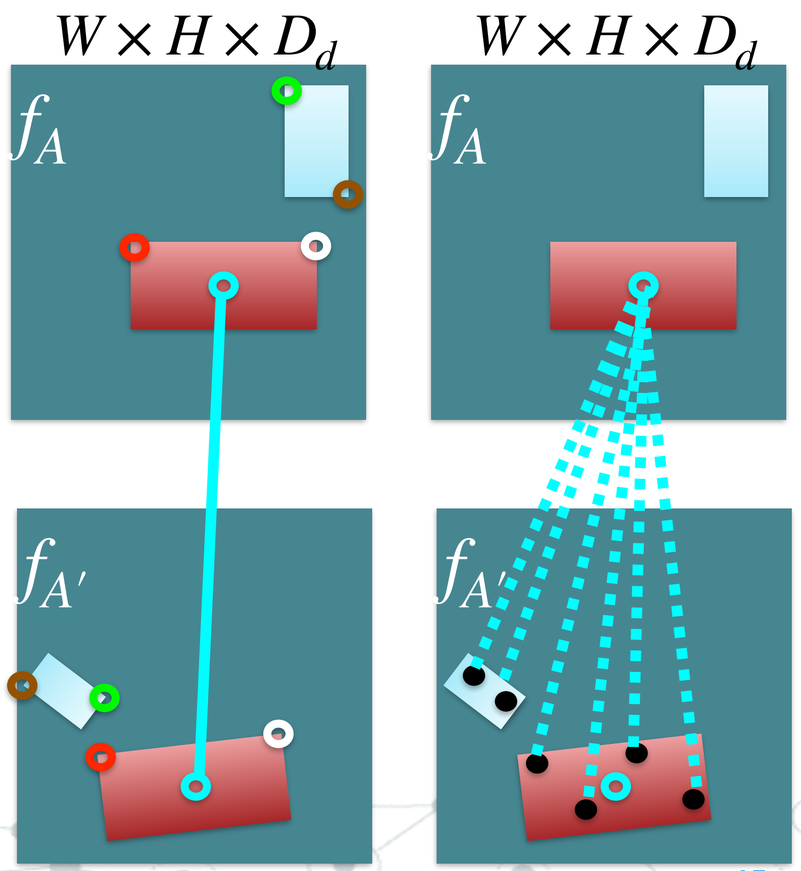


# Preliminary: Dense Object Nets [1]

- Contrastive loss for match points
- Constrastive hard-negative loss for non-match points

$$D(f_A(\circ), f_{A'}(\circ))^2 = \left\| \begin{bmatrix} x_1 \\ \vdots \\ x_{D_d} \end{bmatrix}_{f_A} - \begin{bmatrix} x_1 \\ \vdots \\ x_{D_d} \end{bmatrix}_{f_{A'}} \right\|_2^2 \doteq$$

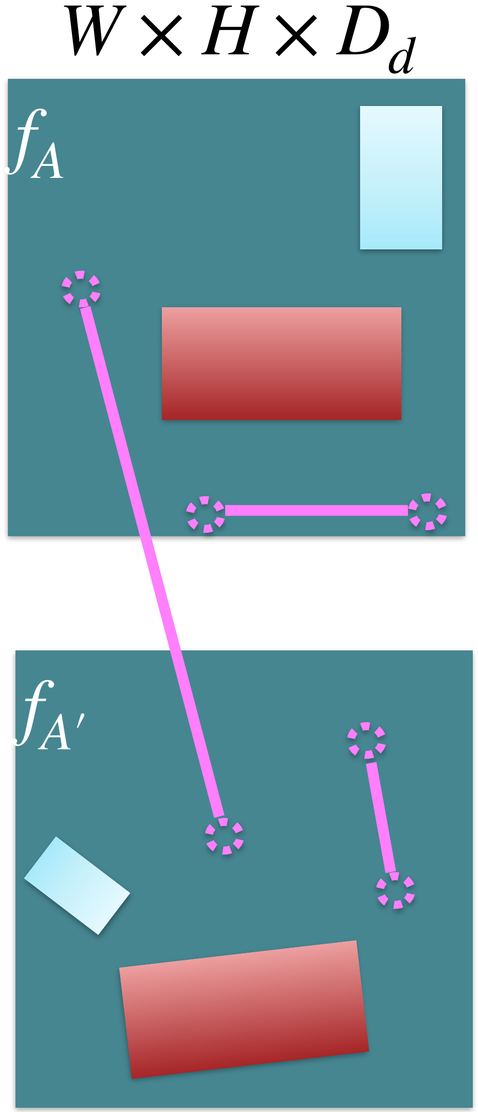
$$\max(0, M_{hn} - D(f_A(\circ), f_{A'}(\bullet)))^2 \doteq$$



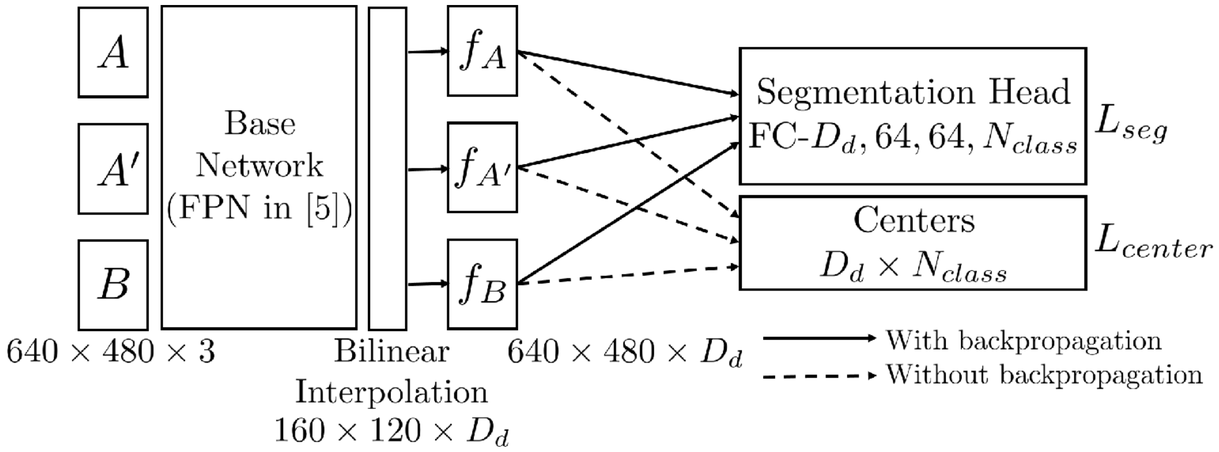
# Background Regularization

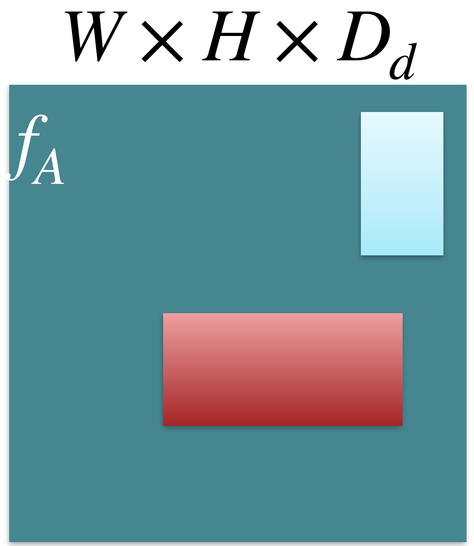
- Contrastive loss for background **match**
  - The background points are regularized to be close to each other within a margin.

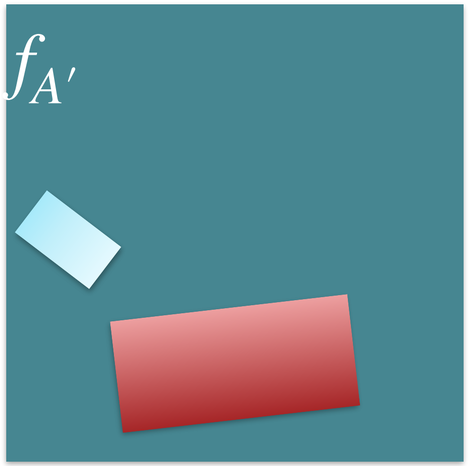
$$\max(0, D(f_A(\ ), f_{A'}(\ )) - M')^2 \doteq$$



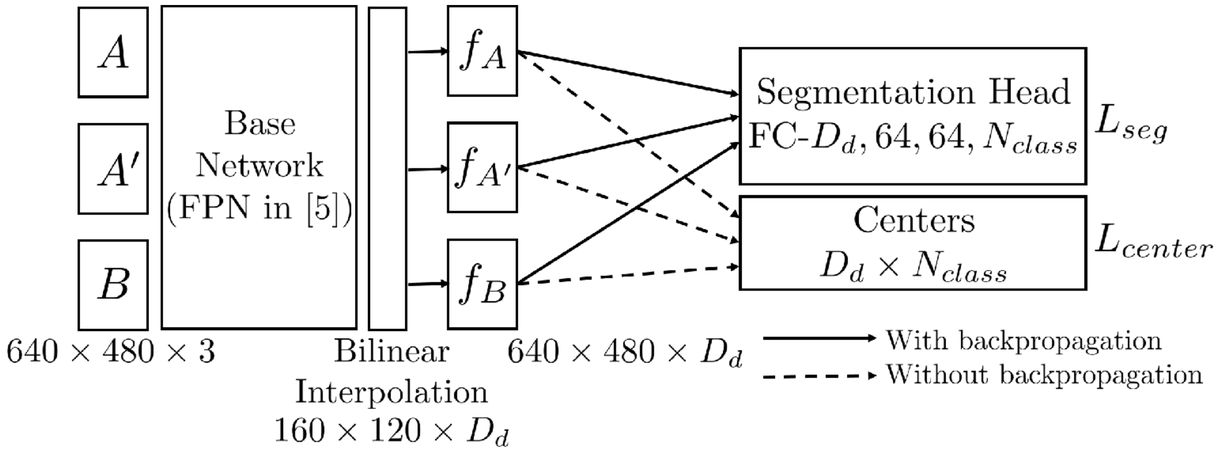
# Inter-class separation

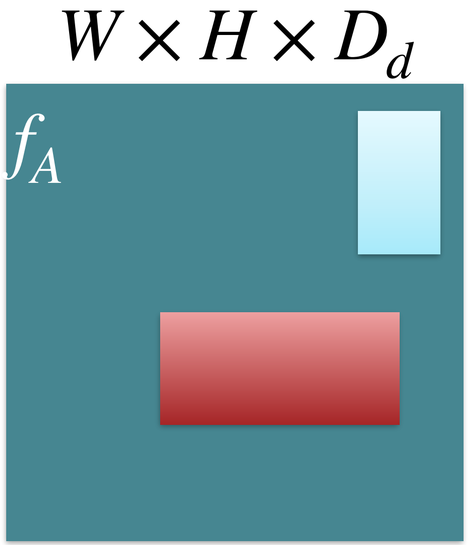


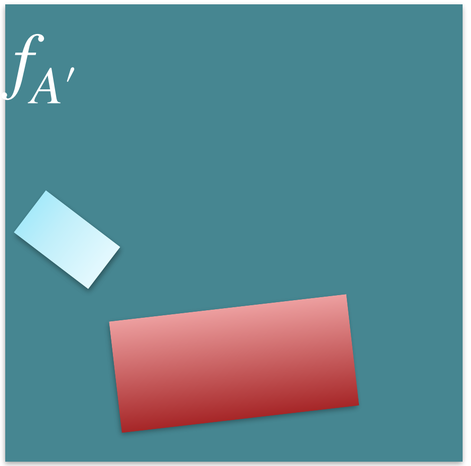



# Inter-class separation



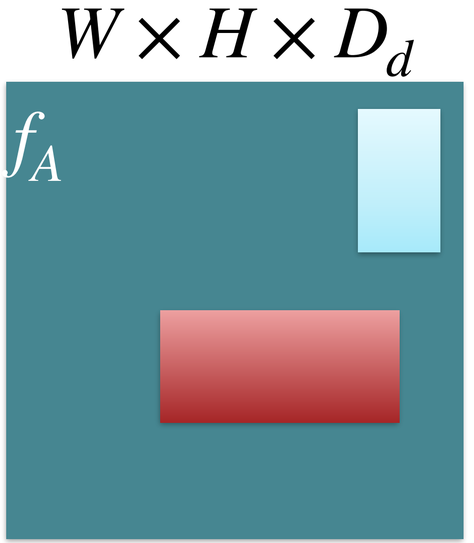



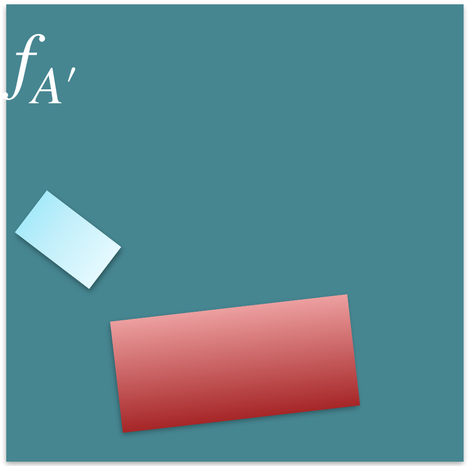



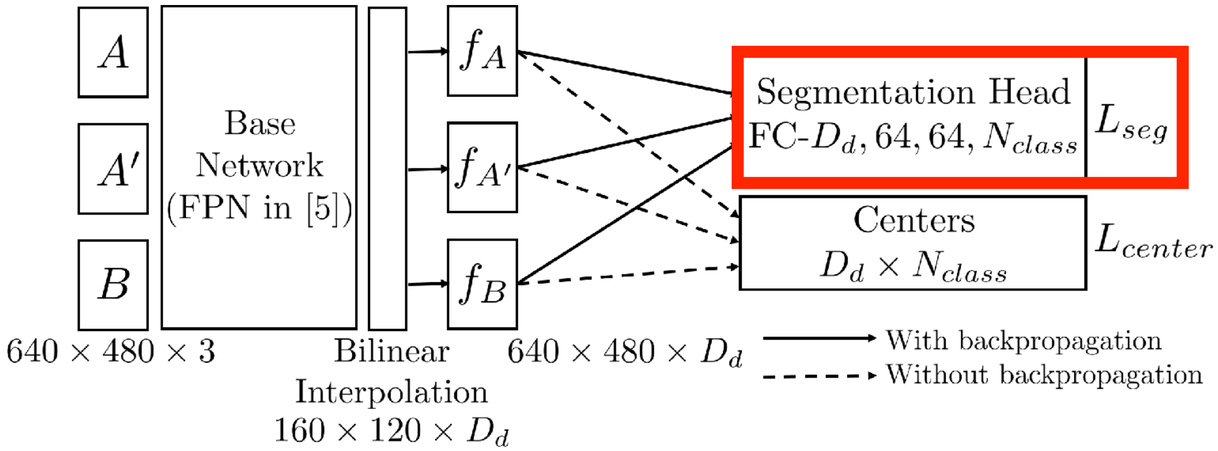

[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. ”A discriminative feature learning approach for deep face recognition”. In ECCV,2016.

# Inter-class separation

- Segmentation loss
  - Softmax cross entropy loss



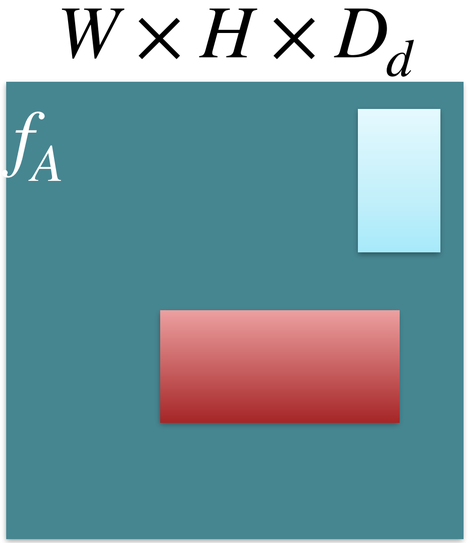



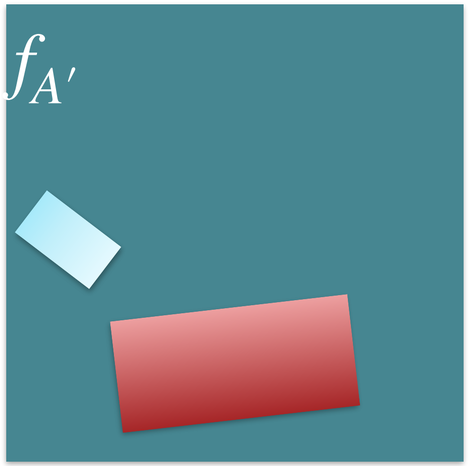



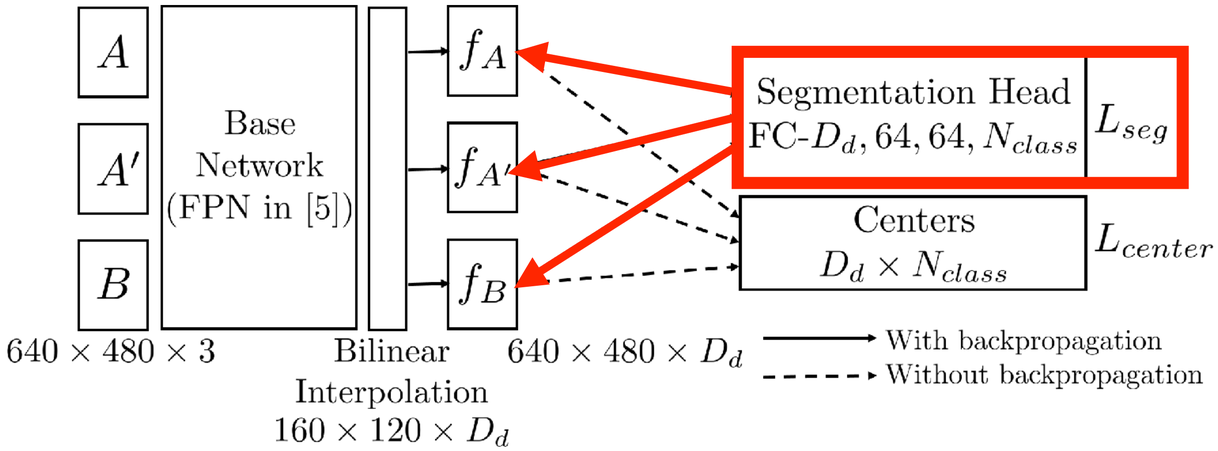

[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. ”A discriminative feature learning approach for deep face recognition”. In ECCV,2016.

# Inter-class separation

- Segmentation loss
  - Softmax cross entropy loss



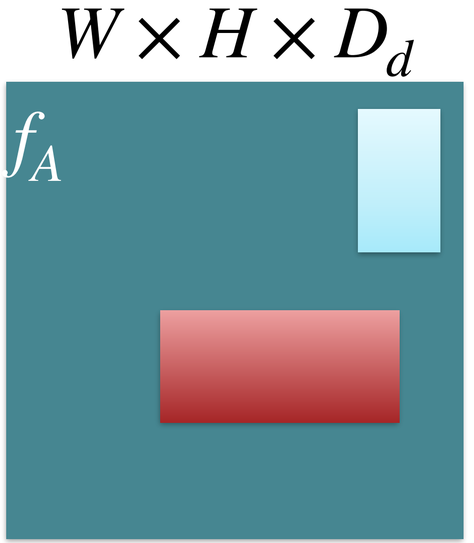



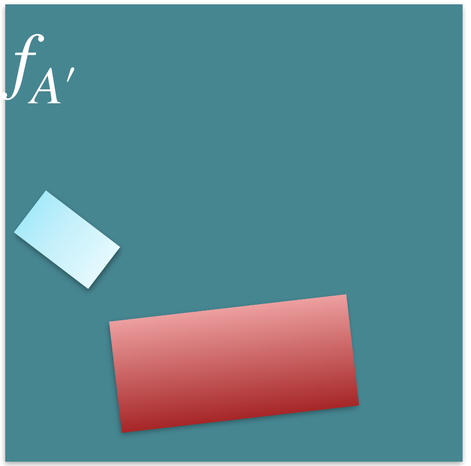



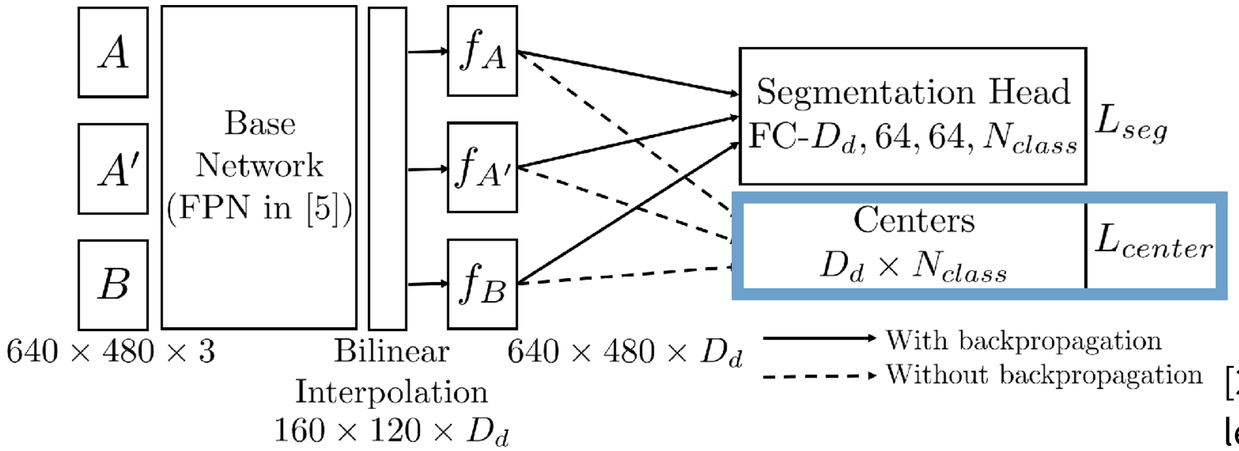

[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. "A discriminative feature learning approach for deep face recognition". In ECCV,2016.

# Inter-class separation

- Segmentation loss
  - Softmax cross entropy loss
- Center loss
  - The centers for triplet center loss



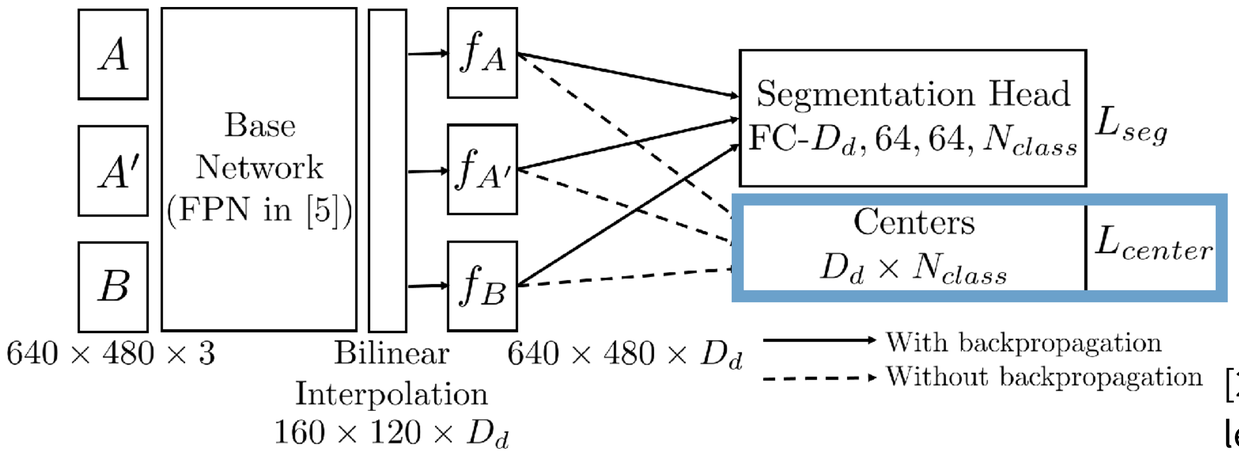



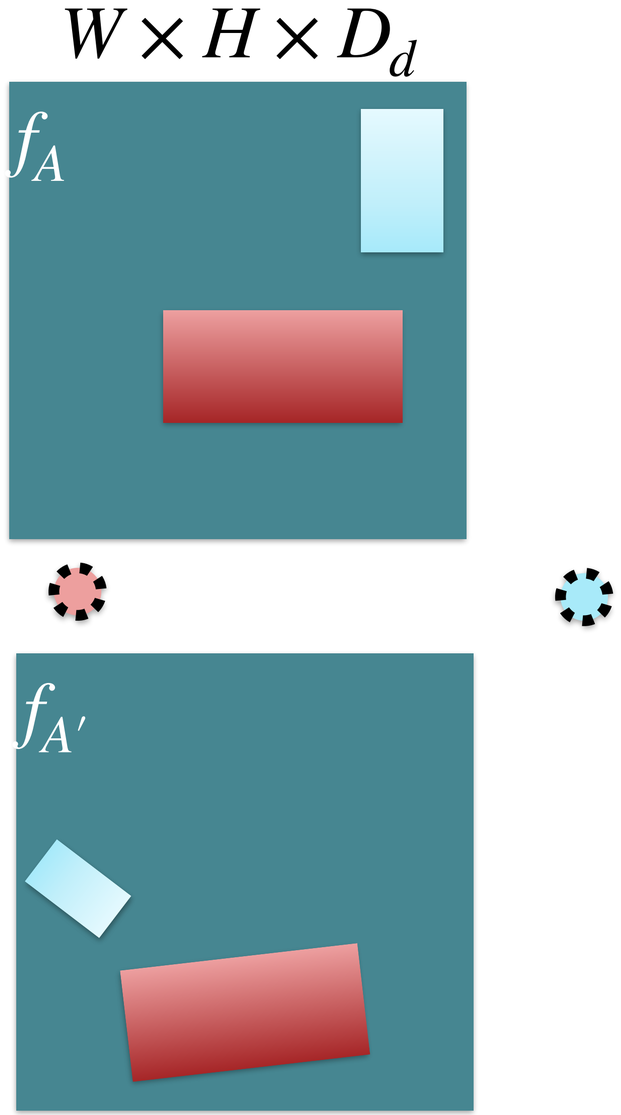




[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. ”A discriminative feature learning approach for deep face recognition”. In ECCV,2016.

# Inter-class separation

- Segmentation loss
  - Softmax cross entropy loss
- Center loss
  - The centers for triplet center loss

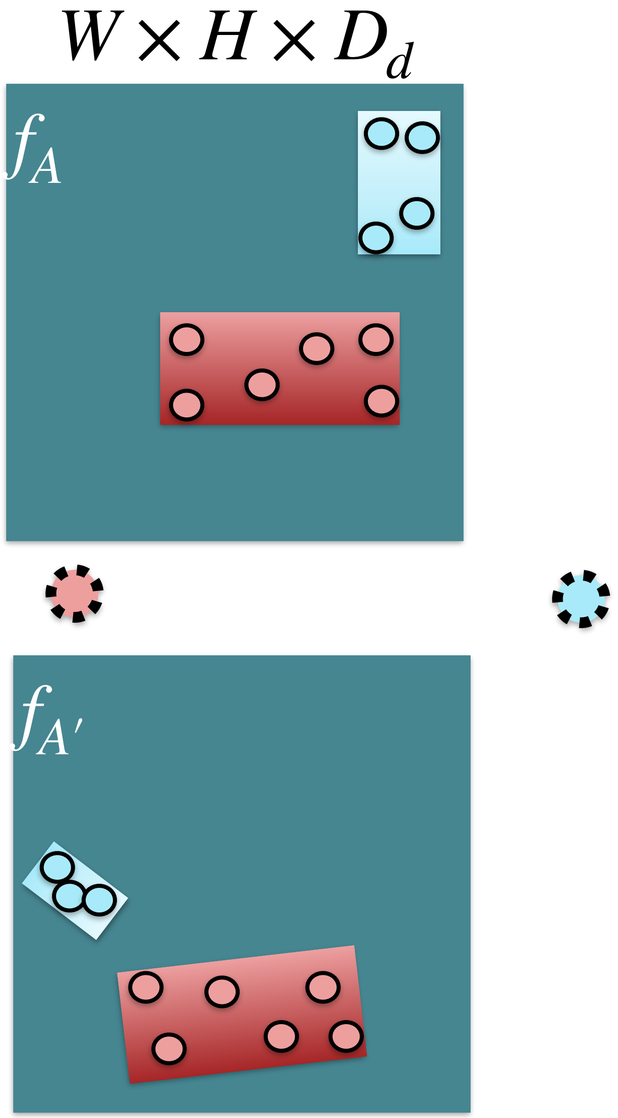


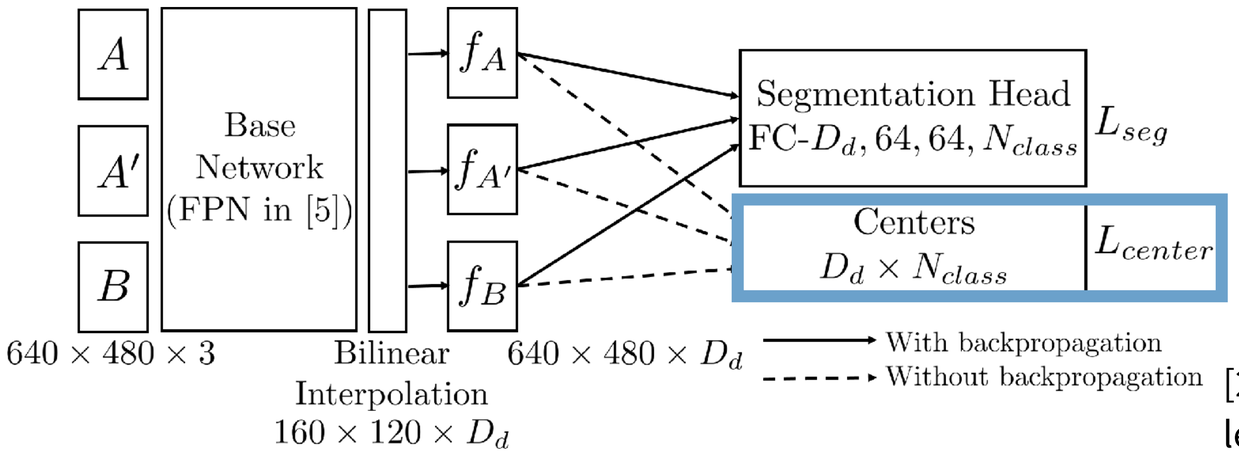




[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. ”A discriminative feature learning approach for deep face recognition”. In ECCV,2016.

# Inter-class separation

- Segmentation loss
  - Softmax cross entropy loss
- Center loss
  - The centers for triplet center loss

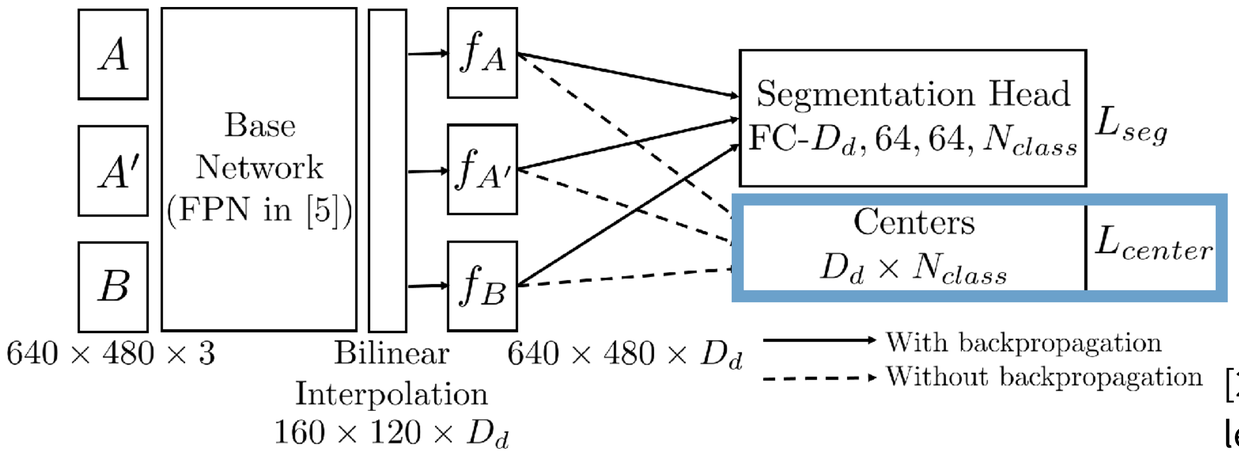


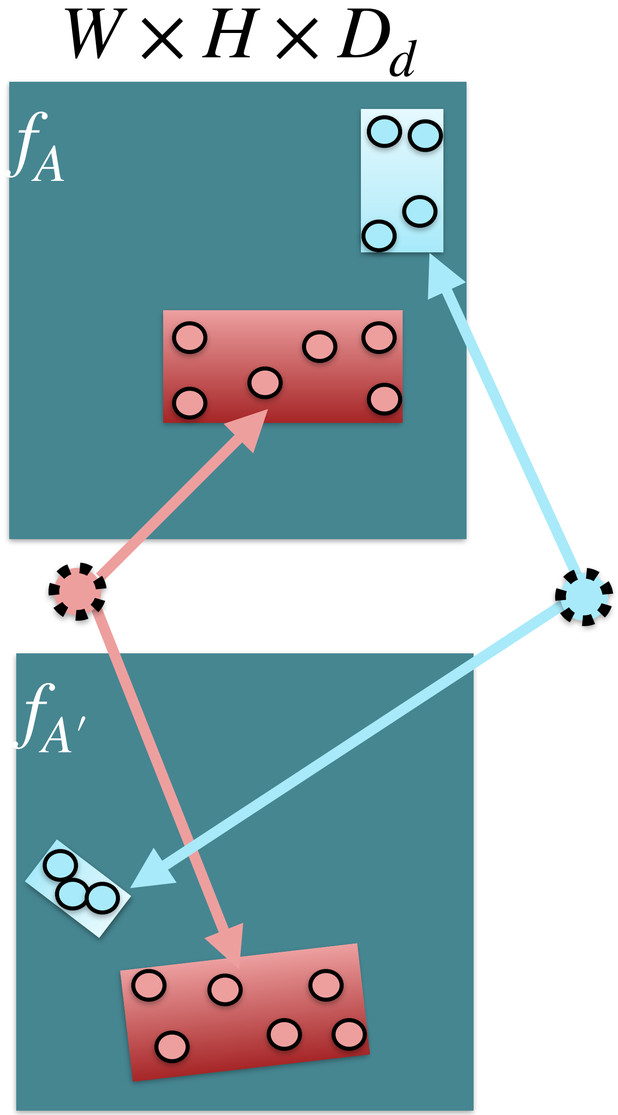




[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. ”A discriminative feature learning approach for deep face recognition”. In ECCV,2016.

# Inter-class separation

- Segmentation loss
  - Softmax cross entropy loss
- Center loss
  - The centers for triplet center loss

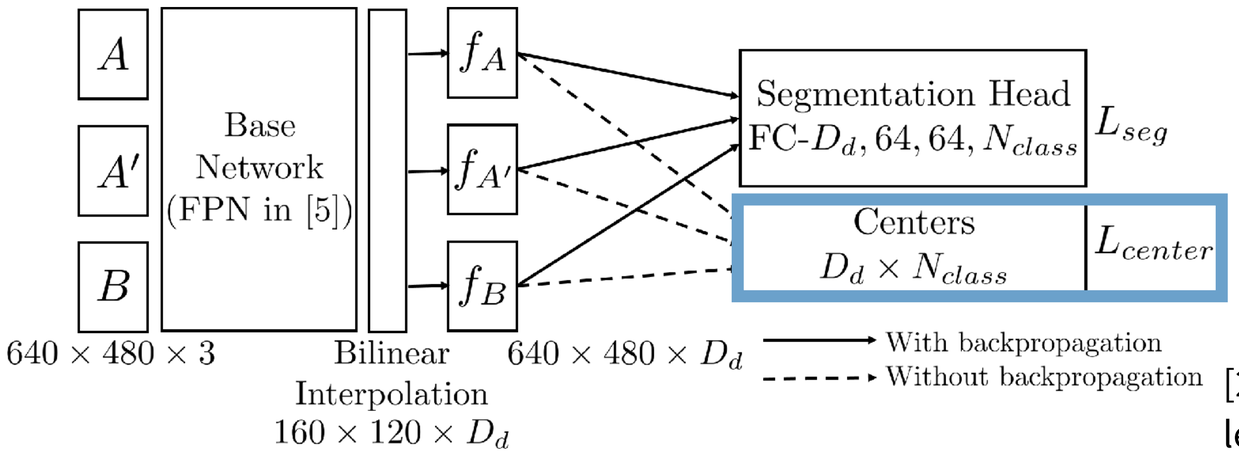


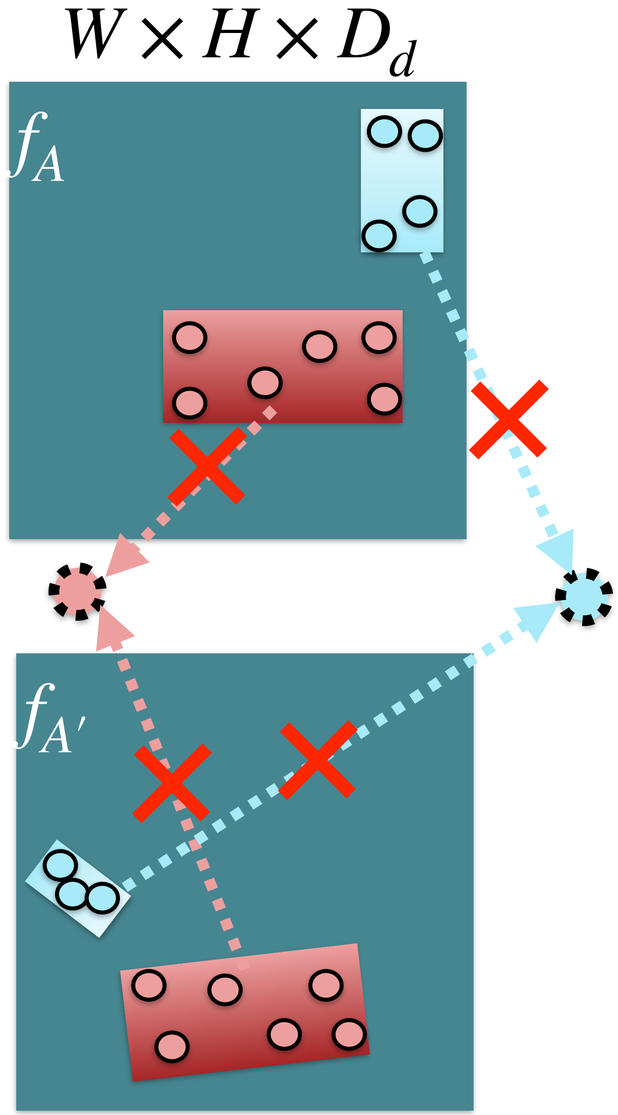




[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. ”A discriminative feature learning approach for deep face recognition”. In ECCV,2016.

# Inter-class separation

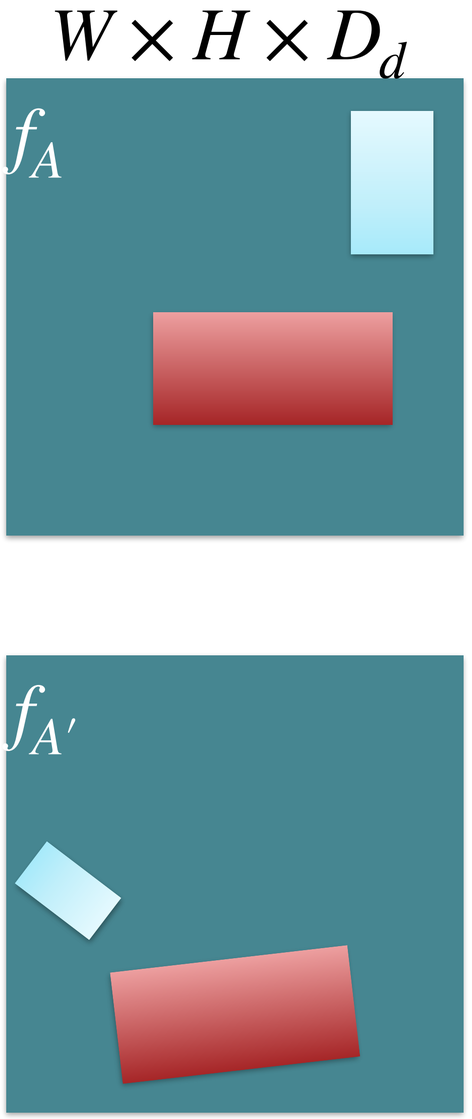
- Segmentation loss
  - Softmax cross entropy loss
- Center loss
  - The centers for triplet center loss




[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. ”A discriminative feature learning approach for deep face recognition”. In ECCV,2016.

# Inter-class separation

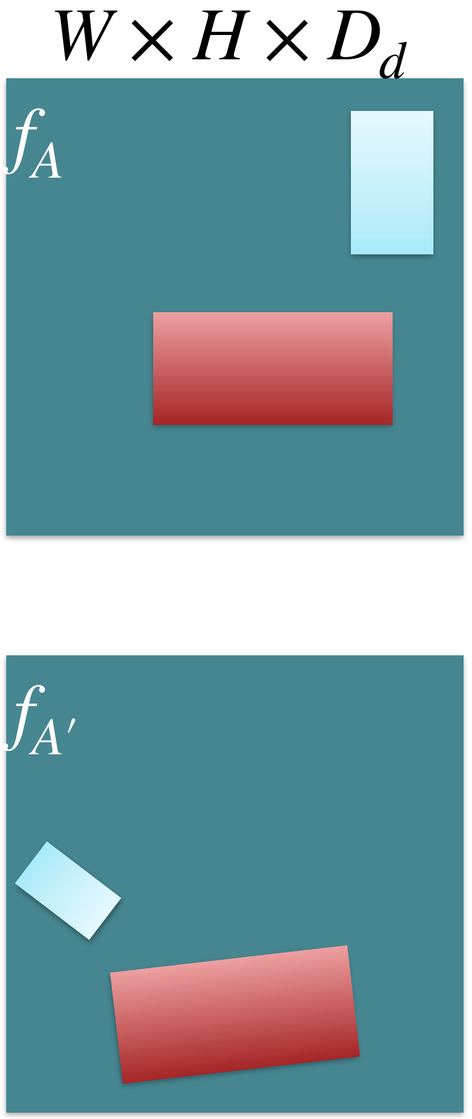



[3] F. Schroff, D. Kalenichenko, and J. Philbin. ”Facenet: A unied embedding for face recognition and clustering”. In CVPR, 2015.

[4] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, ”Triplet-Center Loss for Multi-View 3D Object Retrieval”. In CVPR, 2018.

# Inter-class separation

- Triplet center loss
  - A descriptor in the metric space should have a shorter distance to its class center than any other class center.

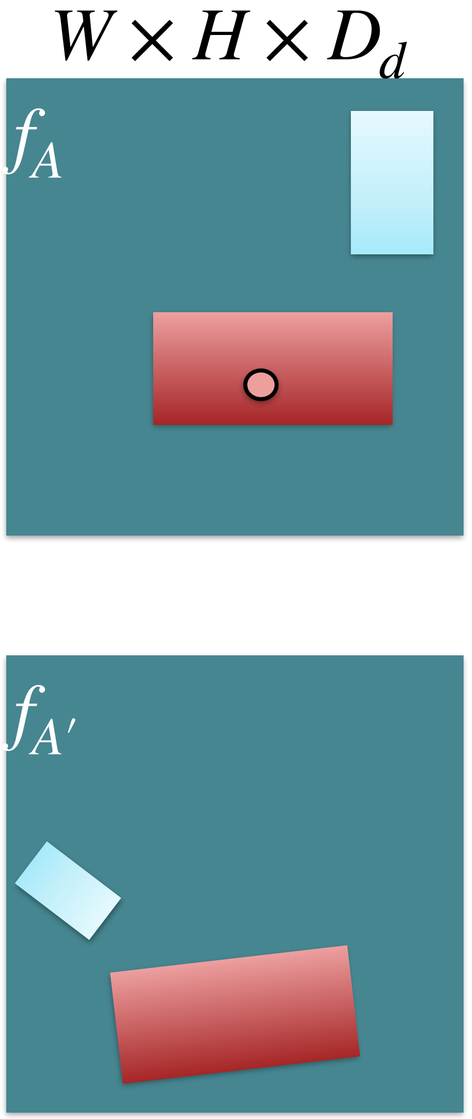



[3] F. Schroff, D. Kalenichenko, and J. Philbin. ”Facenet: A unied embedding for face recognition and clustering”. In CVPR, 2015.

[4] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, ”Triplet-Center Loss for Multi-View 3D Object Retrieval”. In CVPR, 2018.

# Inter-class separation

- Triplet center loss
  - A descriptor in the metric space should have a shorter distance to its class center than any other class center.

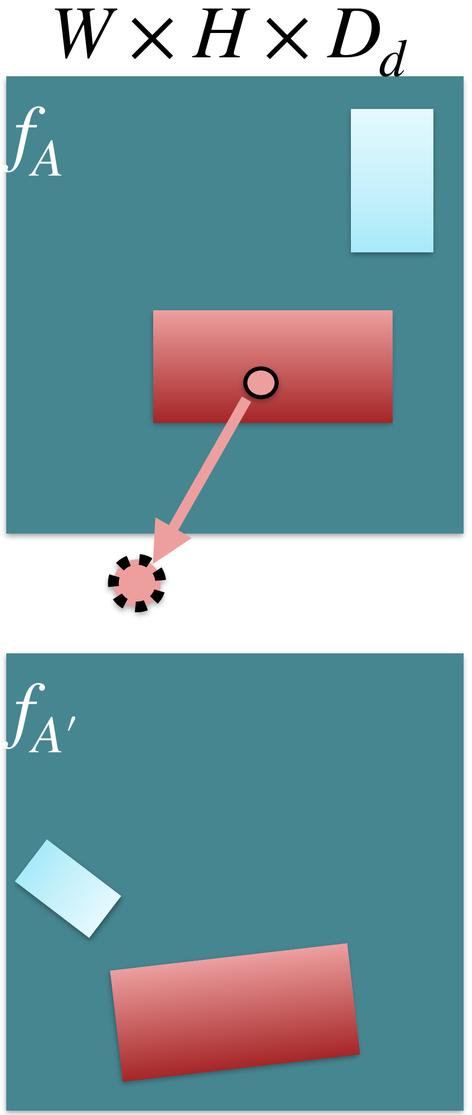



[3] F. Schroff, D. Kalenichenko, and J. Philbin. ”Facenet: A unied embedding for face recognition and clustering”. In CVPR, 2015.

[4] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, ”Triplet-Center Loss for Multi-View 3D Object Retrieval”. In CVPR, 2018.

# Inter-class separation

- Triplet center loss
  - A descriptor in the metric space should have a shorter distance to its class center than any other class center.

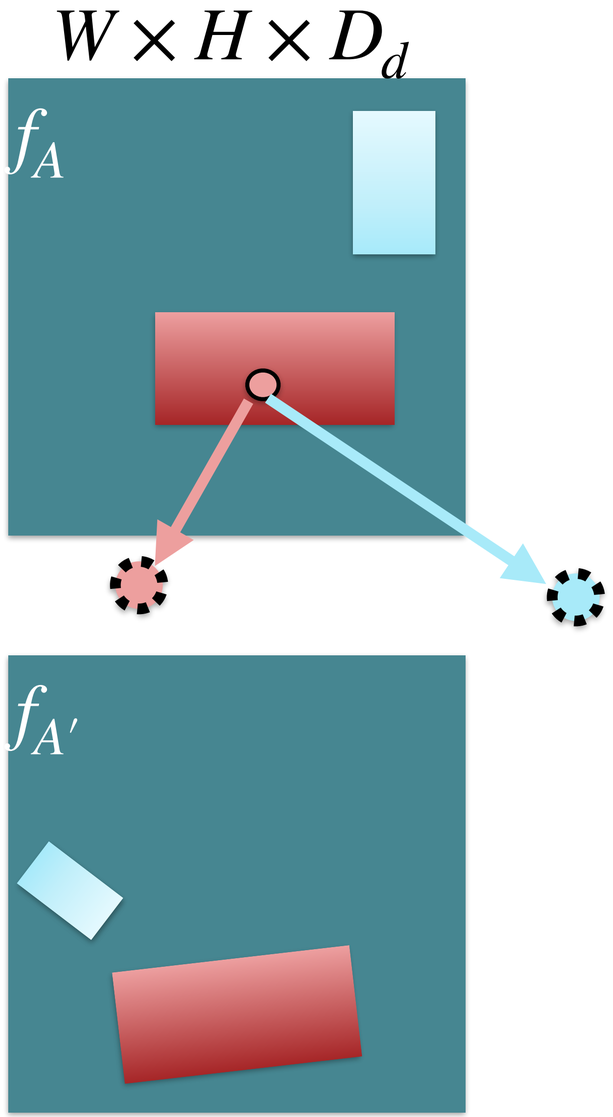



[3] F. Schroff, D. Kalenichenko, and J. Philbin. ”Facenet: A unied embedding for face recognition and clustering”. In CVPR, 2015.

[4] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, ”Triplet-Center Loss for Multi-View 3D Object Retrieval”. In CVPR, 2018.

# Inter-class separation

- Triplet center loss
  - A descriptor in the metric space should have a shorter distance to its class center than any other class center.



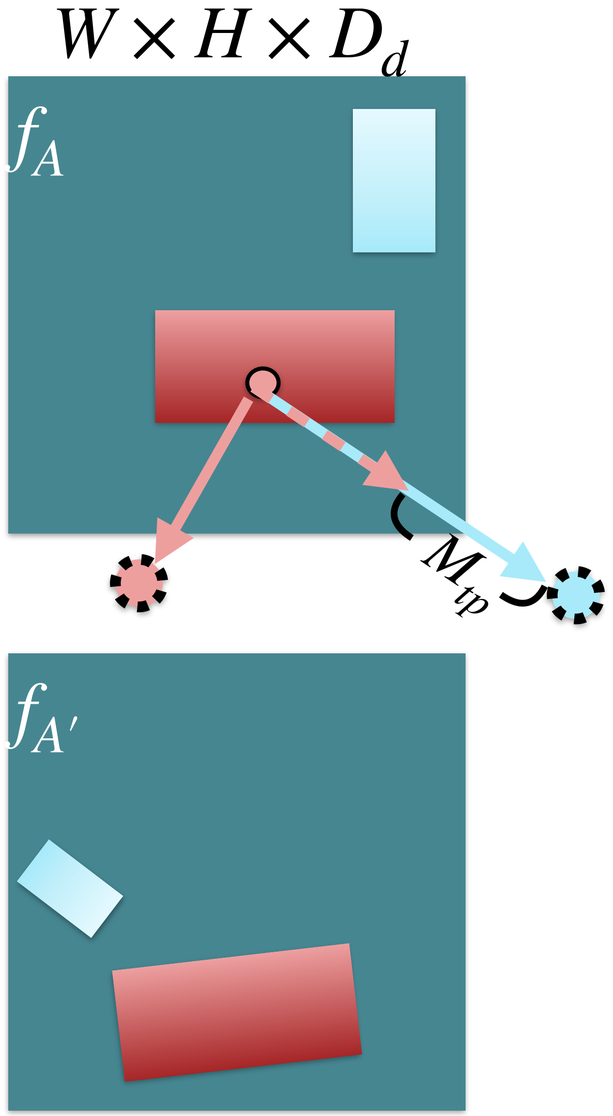

[3] F. Schroff, D. Kalenichenko, and J. Philbin. ”Facenet: A unied embedding for face recognition and clustering”. In CVPR, 2015.

[4] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, ”Triplet-Center Loss for Multi-View 3D Object Retrieval”. In CVPR, 2018.

# Inter-class separation

- Triplet center loss
  - A descriptor in the metric space should have a shorter distance to its class center than any other class center.
  - At least shorter by a margin of $M_{tp}$

$$\max(0, D(f_A(\circ), \circ) - D(f_A(\circ), \circ) + M_{tp})^2$$





[3] F. Schroff, D. Kalenichenko, and J. Philbin. ”Facenet: A unied embedding for face recognition and clustering”. In CVPR, 2015.

[4] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, ”Triplet-Center Loss for Multi-View 3D Object Retrieval”. In CVPR, 2018.

# Experiments

# Experiments

# Experiments

| configurations |
| --- |
| PCK@40 |
| PCK@80 |
| PCK@120 |

# Experiments

| configurations |
| --- |
| PCK@40 |
| PCK@80 |
| PCK@120 |
| Mean Pixel Error |

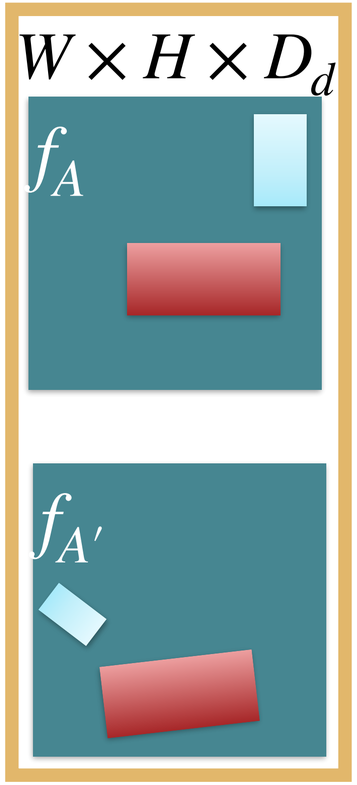
# Experiments

| configurations |
| --- |
| PCK@40 |
| PCK@80 |
| PCK@120 |
| Mean Pixel Error |
| Accuracy |

# Experiments

SYNTHETIC MULTI-OBJECT EVALUATION

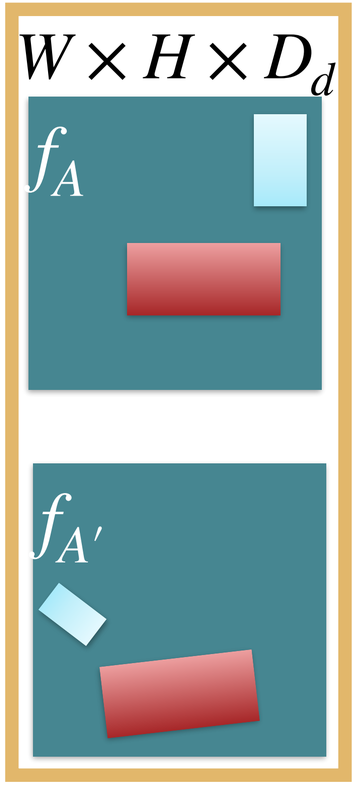
| configurations | *DON Baseline* | M=2 N=5 | M=3 N=4 | M=4 N=3 | M=5 N=2 | M=6 N=1 |
|---|---|---|---|---|---|---|
| PCK@40 | *0.710* | 0.714 | 0.751 | 0.757 | **0.762** | 0.720 |
| PCK@80 | *0.847* | 0.863 | 0.883 | 0.889 | **0.898** | 0.873 |
| PCK@120 | *0.898* | 0.922 | 0.932 | 0.940 | **0.949** | 0.932 |
| Mean Pixel Error | *48.03* | 40.60 | 37.04 | 34.65 | **32.90** | 38.15 |
| Accuracy | - | 0.946 | 0.935 | 0.942 | **0.951** | 0.901 |

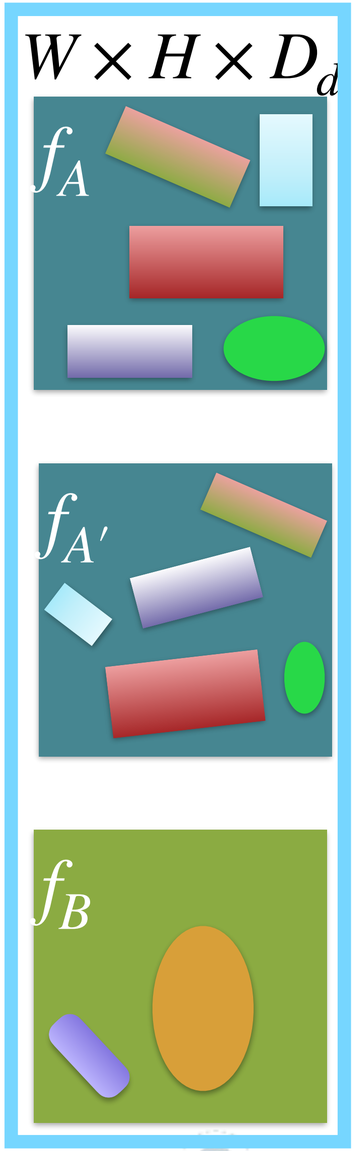


# Experiments

SYNTHETIC MULTI-OBJECT EVALUATION

| configurations | *DON Baseline* | M=2 N=5 | M=3 N=4 | M=4 N=3 | M=5 N=2 | M=6 N=1 |
|---|---|---|---|---|---|---|
| PCK@40 | *0.710* | 0.714 | 0.751 | 0.757 | **0.762** | 0.720 |
| PCK@80 | *0.847* | 0.863 | 0.883 | 0.889 | **0.898** | 0.873 |
| PCK@120 | *0.898* | 0.922 | 0.932 | 0.940 | **0.949** | 0.932 |
| Mean Pixel Error | *48.03* | 40.60 | 37.04 | 34.65 | **32.90** | 38.15 |
| Accuracy | - | 0.946 | 0.935 | 0.942 | **0.951** | 0.901 |

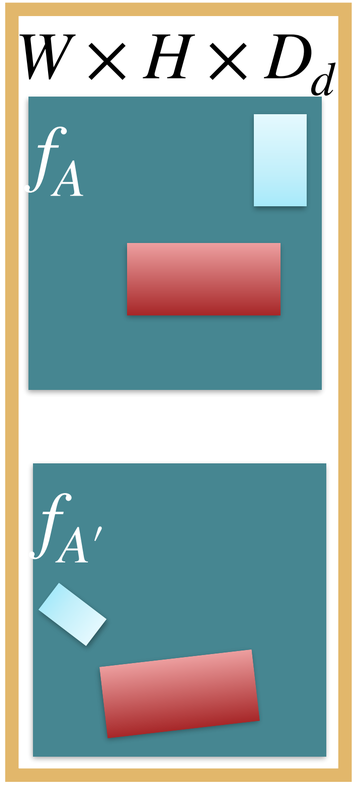


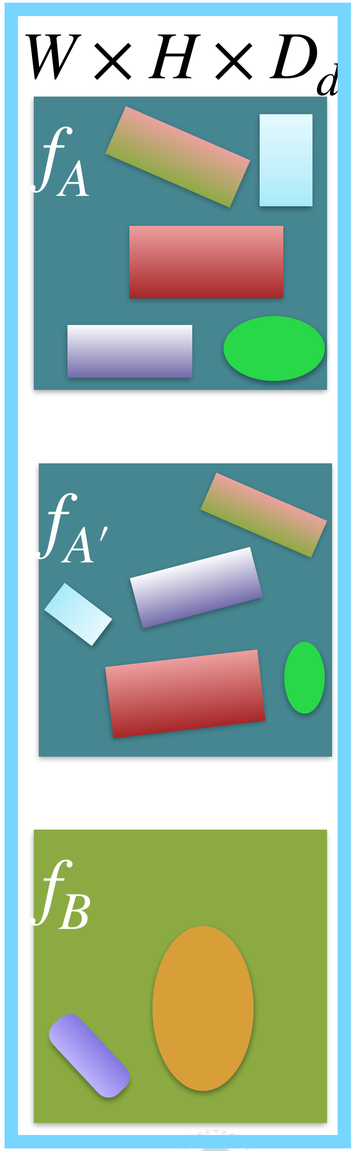



# Experiments

SYNTHETIC MULTI-OBJECT EVALUATION

| configurations | *DON Baseline* | M=2 N=5 | M=3 N=4 | M=4 N=3 | M=5 N=2 → | M=6 N=1 |
|---|---|---|---|---|---|---|
| PCK@40 | *0.710* | 0.714 | 0.751 | 0.757 | **0.762** → | 0.720 |
| PCK@80 | *0.847* | 0.863 | 0.883 | 0.889 | **0.898** → | 0.873 |
| PCK@120 | *0.898* | 0.922 | 0.932 | 0.940 | **0.949** → | 0.932 |
| Mean Pixel Error | *48.03* | 40.60 | 37.04 | 34.65 | **32.90** → | 38.15 |
| Accuracy | - | 0.946 | 0.935 | 0.942 | **0.951** → | 0.901 |

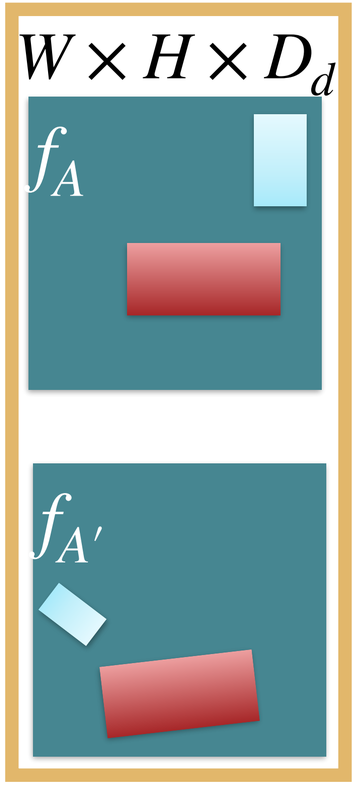


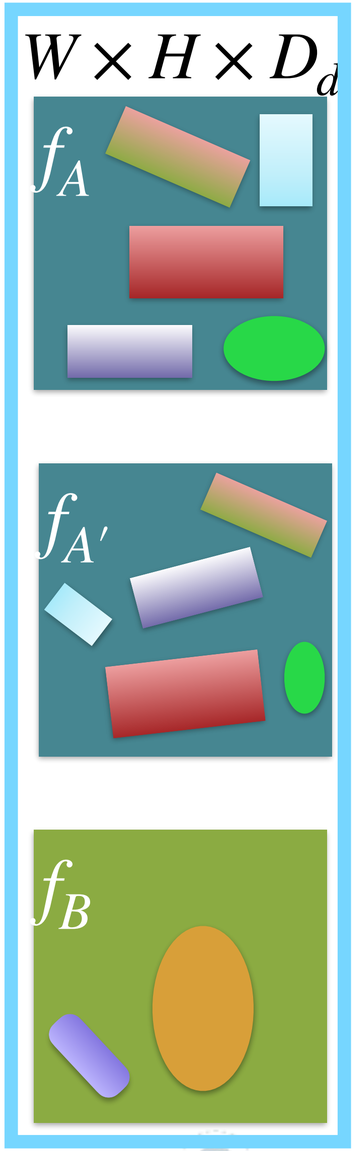



# Experiments

SYNTHETIC MULTI-OBJECT EVALUATION

| configurations | *DON Baseline* | M=2 N=5 | M=3 N=4 | M=4 N=3 | M=5 N=2 | M=6 N=1 |
|---|---|---|---|---|---|---|
| PCK@40 | *0.710* | 0.714 | 0.751 | 0.757 | **0.762** | 0.720 |
| PCK@80 | *0.847* | 0.863 | 0.883 | 0.889 | **0.898** | 0.873 |
| PCK@120 | *0.898* | 0.922 | 0.932 | 0.940 | **0.949** | 0.932 |
| Mean Pixel Error | *48.03* | 40.60 | 37.04 | 34.65 | **32.90** | 38.15 |
| Accuracy | - | 0.946 | 0.935 | 0.942 | **0.951** | 0.901 |



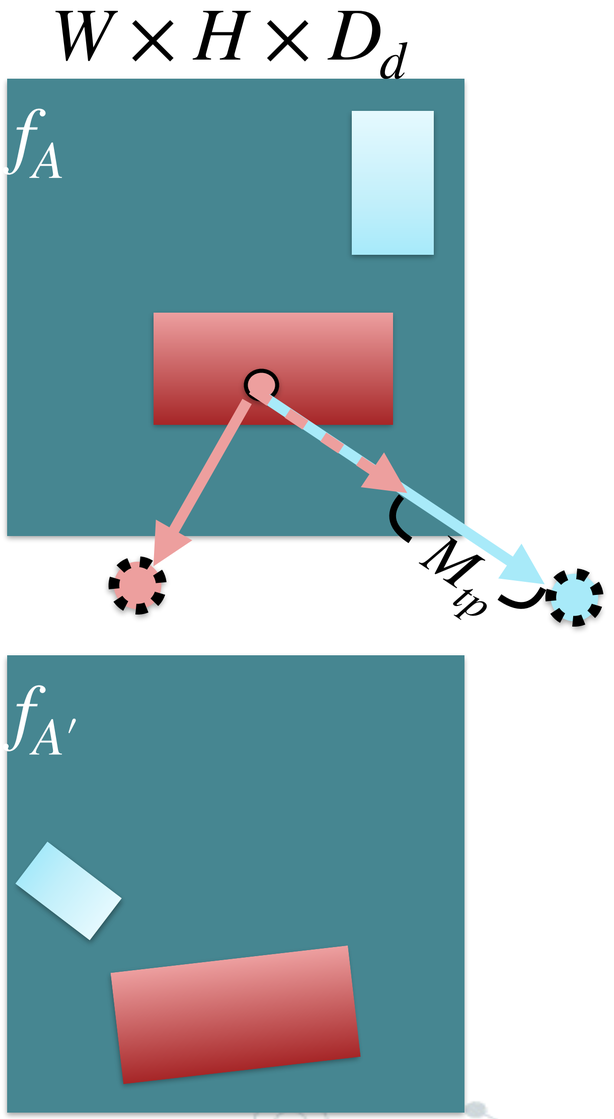
# Experiments

SYNTHETIC MULTI-OBJECT EVALUATION

| configurations | *DON Baseline* | M=2 N=5 | M=3 N=4 | M=4 N=3 | M=5 N=2 | M=6 N=1 |
|---|---|---|---|---|---|---|
| PCK@40 | *0.710* | 0.714 | 0.751 | 0.757 | **0.762** | 0.720 |
| PCK@80 | *0.847* | 0.863 | 0.883 | 0.889 | **0.898** | 0.873 |
| PCK@120 | *0.898* | 0.922 | 0.932 | 0.940 | **0.949** | 0.932 |
| Mean Pixel Error | *48.03* | 40.60 | 37.04 | 34.65 | **32.90** | 38.15 |
| Accuracy | - | 0.946 | 0.935 | 0.942 | **0.951** | 0.901 |

ABLATION ANALYSIS

| configurations | MH 6D | MHS 6D | MHSCT 6D | MH 12D | MHS 12D | MHSCT 12D | MHSCT 6D w/o Bk | MHSCT 12D w/o Bk |
|---|---|---|---|---|---|---|---|---|
| PCK@40 | 0.682 | 0.682 | 0.663 | 0.783 | 0.771 | 0.762 | 0.656 | 0.718 |
| PCK@80 | 0.840 | 0.846 | 0.834 | 0.905 | 0.901 | 0.898 | 0.832 | 0.874 |
| PCK@120 | 0.905 | 0.916 | 0.906 | 0.95 | 0.949 | 0.949 | 0.905 | 0.934 |
| Mean Pixel Error | 46.36 | 43.82 | 45.75 | 32.32 | 32.45 | 32.90 | 47.11 | 38.335 |
| Segmentation Accuracy | - | 0.815 | **0.921** | - | 0.916 | **0.951** | 0.920 | 0.920 |



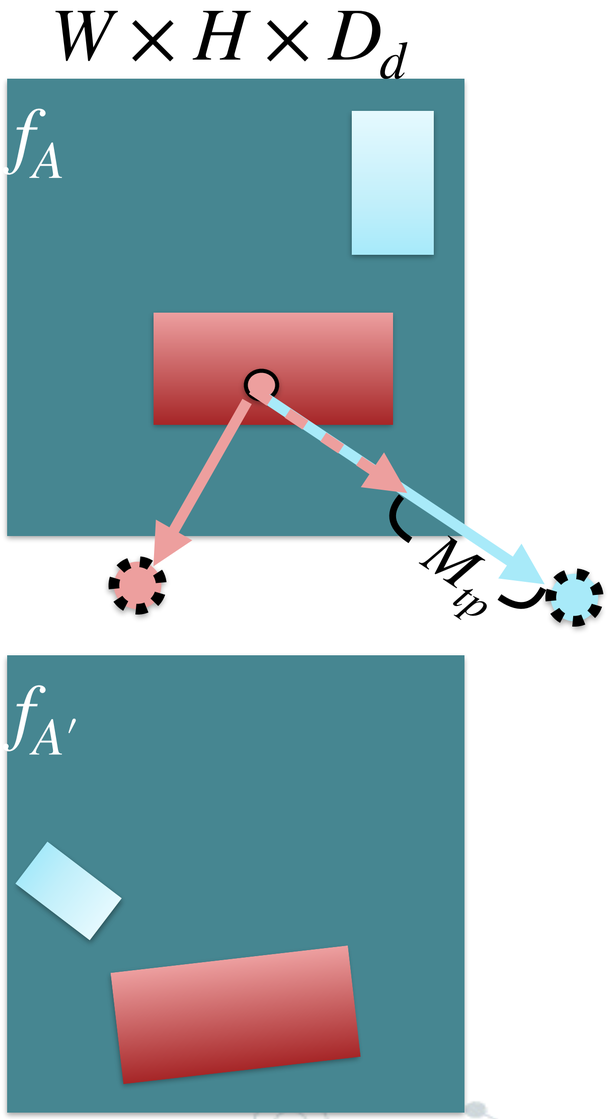
# Experiments

SYNTHETIC MULTI-OBJECT EVALUATION

| configurations | *DON Baseline* | M=2 N=5 | M=3 N=4 | M=4 N=3 | M=5 N=2 | M=6 N=1 |
|---|---|---|---|---|---|---|
| PCK@40 | *0.710* | 0.714 | 0.751 | 0.757 | **0.762** | 0.720 |
| PCK@80 | *0.847* | 0.863 | 0.883 | 0.889 | **0.898** | 0.873 |
| PCK@120 | *0.898* | 0.922 | 0.932 | 0.940 | **0.949** | 0.932 |
| Mean Pixel Error | *48.03* | 40.60 | 37.04 | 34.65 | **32.90** | 38.15 |
| Accuracy | - | 0.946 | 0.935 | 0.942 | **0.951** | 0.901 |

ABLATION ANALYSIS

| configurations | MH 6D | MHS 6D | MHSCT 6D | MH 12D | MHS 12D | MHSCT 12D | MHSCT 6D w/o Bk | MHSCT 12D w/o Bk |
|---|---|---|---|---|---|---|---|---|
| PCK@40 | 0.682 | 0.682 | 0.663 | 0.783 | 0.771 | 0.762 | 0.656 | 0.718 |
| PCK@80 | 0.840 | 0.846 | 0.834 | 0.905 | 0.901 | 0.898 | 0.832 | 0.874 |
| PCK@120 | 0.905 | 0.916 | 0.906 | 0.95 | 0.949 | 0.949 | 0.905 | 0.934 |
| Mean Pixel Error | 46.36 | 43.82 | 45.75 | 32.32 | 32.45 | 32.90 | 47.11 | 38.335 |
| Segmentation Accuracy | - | 0.815 | **0.921** | - | 0.916 | **0.951** | 0.920 | 0.920 |



# Experiments

SYNTHETIC MULTI-OBJECT EVALUATION

| configurations | *DON Baseline* | M=2 N=5 | M=3 N=4 | M=4 N=3 | M=5 N=2 | M=6 N=1 |
|---|---|---|---|---|---|---|
| PCK@40 | *0.710* | 0.714 | 0.751 | 0.757 | **0.762** | 0.720 |
| PCK@80 | *0.847* | 0.863 | 0.883 | 0.889 | **0.898** | 0.873 |
| PCK@120 | *0.898* | 0.922 | 0.932 | 0.940 | **0.949** | 0.932 |
| Mean Pixel Error | *48.03* | 40.60 | 37.04 | 34.65 | **32.90** | 38.15 |
| Accuracy | - | 0.946 | 0.935 | 0.942 | **0.951** | 0.901 |

ABLATION ANALYSIS

| configurations | MH 6D | MHS 6D | MHSCT 6D | MH 12D | MHS 12D | MHSCT 12D | MHSCT 6D w/o Bk | MHSCT 12D w/o Bk |
|---|---|---|---|---|---|---|---|---|
| PCK@40 | 0.682 | 0.682 | 0.663 | 0.783 | 0.771 | 0.762 | 0.656 | 0.718 |
| PCK@80 | 0.840 | 0.846 | 0.834 | 0.905 | 0.901 | 0.898 | 0.832 | 0.874 |
| PCK@120 | 0.905 | 0.916 | 0.906 | 0.95 | 0.949 | 0.949 | 0.905 | 0.934 |
| Mean Pixel Error | 46.36 | 43.82 | 45.75 | 32.32 | 32.45 | 32.90 | 47.11 | 38.335 |
| Segmentation Accuracy | - | 0.815 | **0.921** | - | 0.916 | **0.951** | 0.920 | 0.920 |

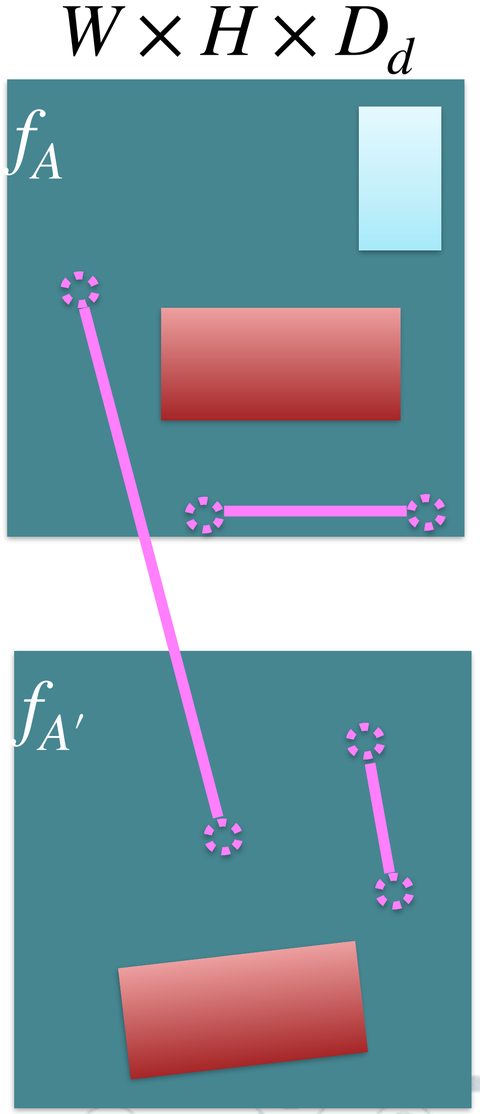
# Experiments

SYNTHETIC MULTI-OBJECT EVALUATION

| configurations | *DON Baseline* | M=2 N=5 | M=3 N=4 | M=4 N=3 | M=5 N=2 | M=6 N=1 |
|---|---|---|---|---|---|---|
| PCK@40 | *0.710* | 0.714 | 0.751 | 0.757 | **0.762** | 0.720 |
| PCK@80 | *0.847* | 0.863 | 0.883 | 0.889 | **0.898** | 0.873 |
| PCK@120 | *0.898* | 0.922 | 0.932 | 0.940 | **0.949** | 0.932 |
| Mean Pixel Error | *48.03* | 40.60 | 37.04 | 34.65 | **32.90** | 38.15 |
| Accuracy | - | 0.946 | 0.935 | 0.942 | **0.951** | 0.901 |

ABLATION ANALYSIS

| configurations | MH 6D | MHS 6D | MHSCT 6D | MH 12D | MHS 12D | MHSCT 12D | MHSCT 6D w/o Bk | MHSCT 12D w/o Bk |
|---|---|---|---|---|---|---|---|---|
| PCK@40 | 0.682 | 0.682 | 0.663 | 0.783 | 0.771 | 0.762 | 0.656 | 0.718 |
| PCK@80 | 0.840 | 0.846 | 0.834 | 0.905 | 0.901 | 0.898 | 0.832 | 0.874 |
| PCK@120 | 0.905 | 0.916 | 0.906 | 0.95 | 0.949 | 0.949 | 0.905 | 0.934 |
| Mean Pixel Error | 46.36 | 43.82 | 45.75 | 32.32 | 32.45 | 32.90 | 47.11 | 38.335 |
| Segmentation Accuracy | - | 0.815 | **0.921** | - | 0.916 | **0.951** | 0.920 | 0.920 |



# Experiments

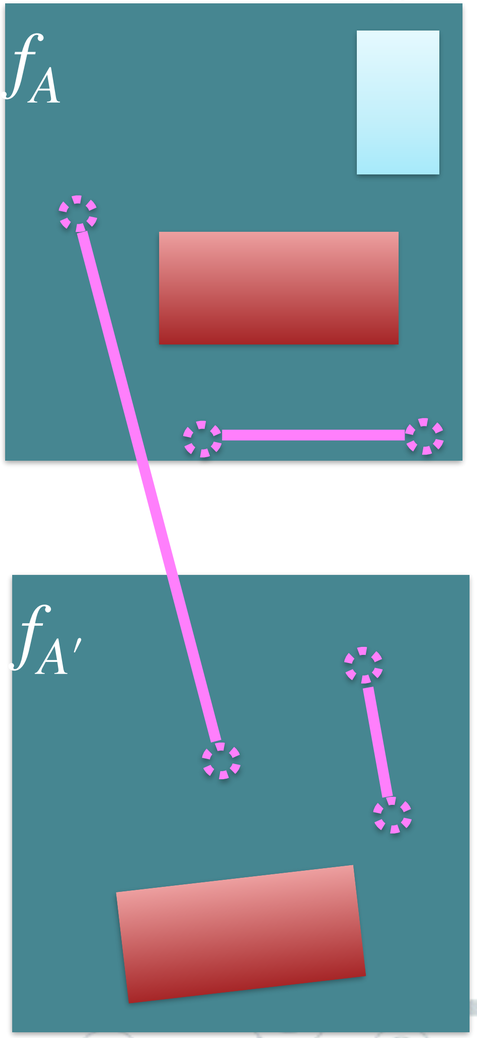
SYNTHETIC MULTI-OBJECT EVALUATION

| configurations | *DON Baseline* | M=2 N=5 | M=3 N=4 | M=4 N=3 | M=5 N=2 | M=6 N=1 |
|---|---|---|---|---|---|---|
| PCK@40 | *0.710* | 0.714 | 0.751 | 0.757 | **0.762** | 0.720 |
| PCK@80 | *0.847* | 0.863 | 0.883 | 0.889 | **0.898** | 0.873 |
| PCK@120 | *0.898* | 0.922 | 0.932 | 0.940 | **0.949** | 0.932 |
| Mean Pixel Error | *48.03* | 40.60 | 37.04 | 34.65 | **32.90** | 38.15 |
| Accuracy | - | 0.946 | 0.935 | 0.942 | **0.951** | 0.901 |

ABLATION ANALYSIS

| configurations | MH 6D | MHS 6D | MHSCT 6D | MH 12D | MHS 12D | MHSCT 12D | MHSCT 6D w/o Bk | MHSCT 12D w/o Bk |
|---|---|---|---|---|---|---|---|---|
| PCK@40 | 0.682 | 0.682 | 0.663 | 0.783 | 0.771 | 0.762 | 0.656 | 0.718 |
| PCK@80 | 0.840 | 0.846 | 0.834 | 0.905 | 0.901 | 0.898 | 0.832 | 0.874 |
| PCK@120 | 0.905 | 0.916 | 0.906 | 0.95 | 0.949 | 0.949 | 0.905 | 0.934 |
| Mean Pixel Error | 46.36 | 43.82 | 45.75 | 32.32 | 32.45 | 32.90 | 47.11 | 38.335 |
| Segmentation Accuracy | - | 0.815 | **0.921** | - | 0.916 | **0.951** | 0.920 | 0.920 |

$W \times H \times D_d$



# Demonstration System

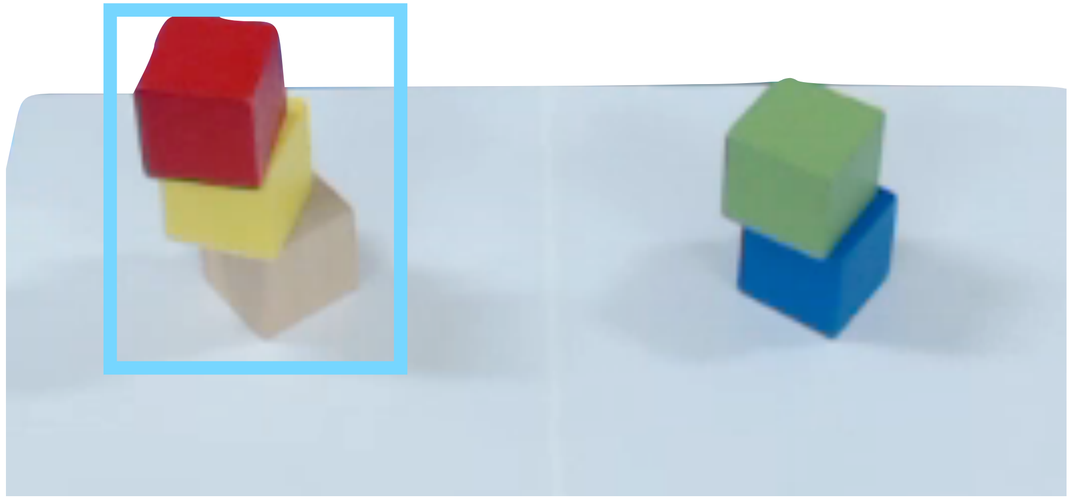
# Demonstration - Two stacks

# Demonstration - Two stacks

- Grasp  and place on 

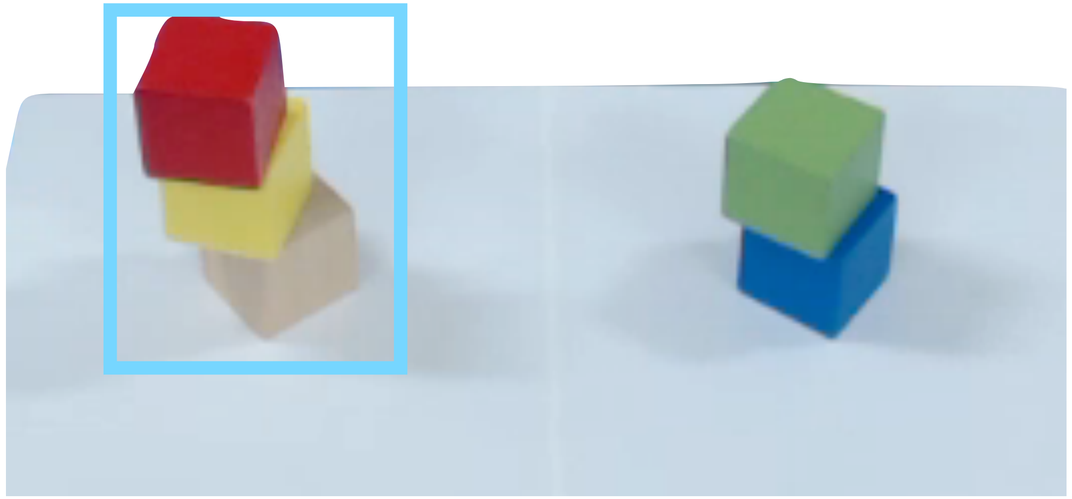
# Demonstration - Two stacks

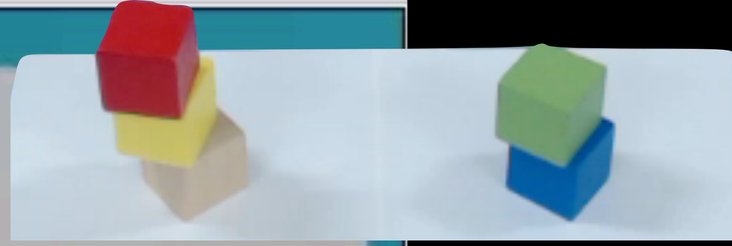
- Grasp  and place on 
- Grasp  and place on 



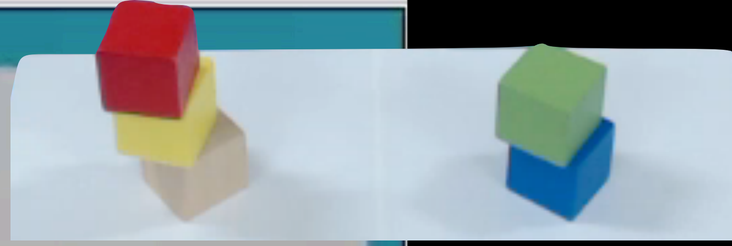
# Demonstration - Two stacks

- Grasp and place on
- Grasp and place on
- Grasp and place on



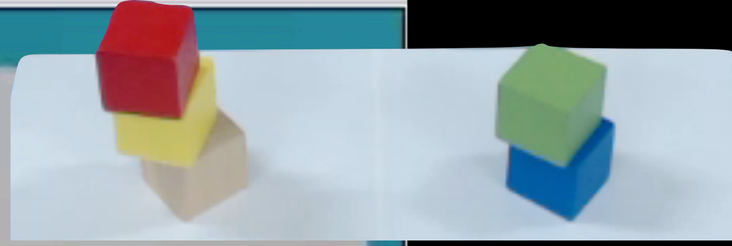


Grasp
Pos-Depth
Grasp Pose
Pose-Desp
Place
Pos-Desp
Add
Delete
Clear All
Show
Save Demo and Rules



Grasp
Pos-Depth
Grasp Pose
Pose-Desp
Place
Pos-Desp
Add
Delete
Clear All
Show
Save Demo and Rules

Camera (Arm)

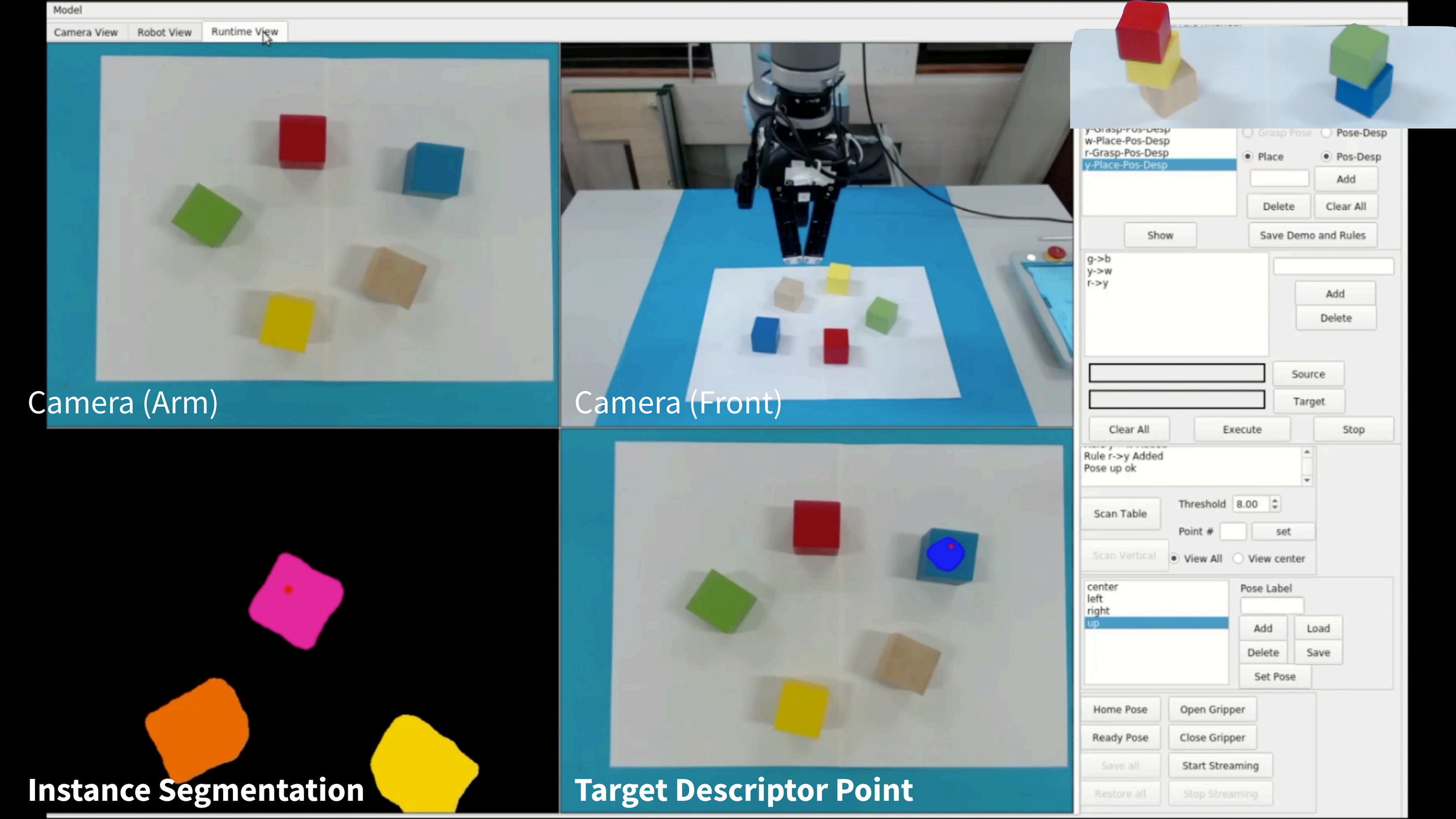


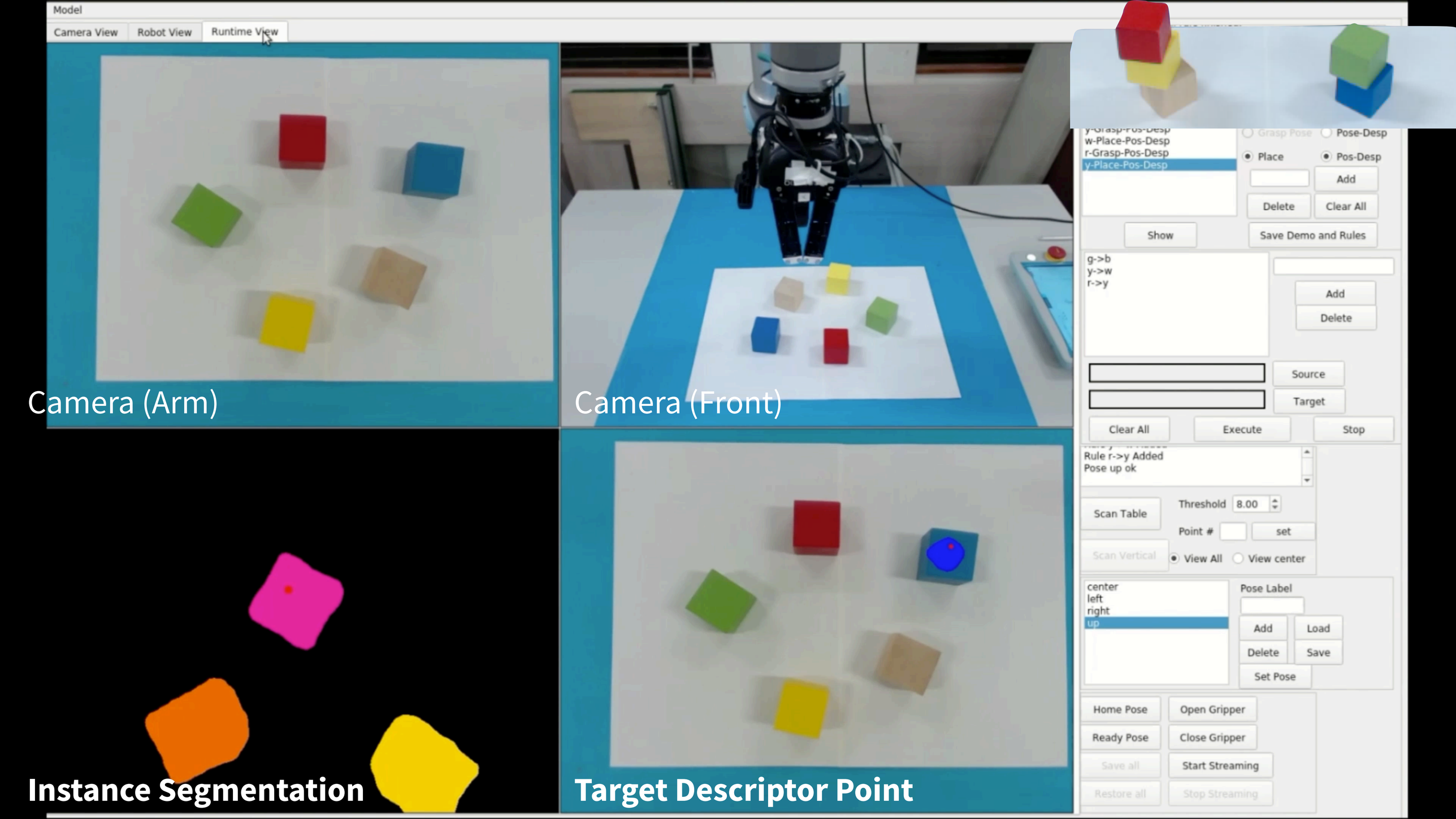
**Semantic Segmentation**

Grasp
Pos-Depth
Grasp Pose
Pose-Desp
Place
Pos-Desp
Add
Delete
Clear All
Show
Save Demo and Rules

Camera (Front)

**Descriptor Visualization(PCA)**


Model
Camera View
Robot View
Runtime View
Camera (Arm)
Camera (Front)
Instance Segmentation
Target Descriptor Point
w-Place-Pos-Desp
r-Grasp-Pos-Desp
y-Place-Pos-Desp
Grasp Pose
Pose-Desp
Place
Pos-Desp
Add
Delete
Clear All
Show
Save Demo and Rules
g->b
y->w
r->y
Add
Delete
Source
Target
Clear All
Execute
Stop
Rule r->y Added
Pose up ok
Scan Table
Threshold
8.00
Point #
set
Scan Vertical
View All
View center
center
left
right
up
Pose Label
Add
Load
Delete
Save
Set Pose
Home Pose
Open Gripper
Ready Pose
Close Gripper
Save all
Start Streaming
Restore all
Stop Streaming


Model
Camera View
Robot View
Runtime View
Camera (Arm)
Camera (Front)
Instance Segmentation
Target Descriptor Point
w-Place-Pos-Desp
r-Grasp-Pos-Desp
y-Place-Pos-Desp
Grasp Pose
Pose-Desp
Place
Pos-Desp
Add
Delete
Clear All
Show
Save Demo and Rules
g->b
y->w
r->y
Add
Delete
Source
Target
Clear All
Execute
Stop
Rule r->y Added
Pose up ok
Scan Table
Threshold
8.00
Point #
set
Scan Vertical
View All
View center
center
left
right
up
Pose Label
Add
Load
Delete
Save
Set Pose
Home Pose
Open Gripper
Ready Pose
Close Gripper
Save all
Start Streaming
Restore all
Stop Streaming

# Conclusion

# Conclusion

- Semantic segmentation objectives for descriptors
  - Learn segmentation and descriptors in the same metric space
  - Balance intra-class variations and inter-class separation

# Conclusion

- Semantic segmentation objectives for descriptors
  - Learn segmentation and descriptors in the same metric space
  - Balance intra-class variations and inter-class separation
- Improvements for multiclass descriptor training
  - Background regularization
  - Multi-patch data augmentation

# Conclusion

- Semantic segmentation objectives for descriptors
  - Learn segmentation and descriptors in the same metric space
  - Balance intra-class variations and inter-class separation
- Improvements for multiclass descriptor training
  - Background regularization
  - Multi-patch data augmentation
- A system for performing a new pick-and-place task by using a single demonstration

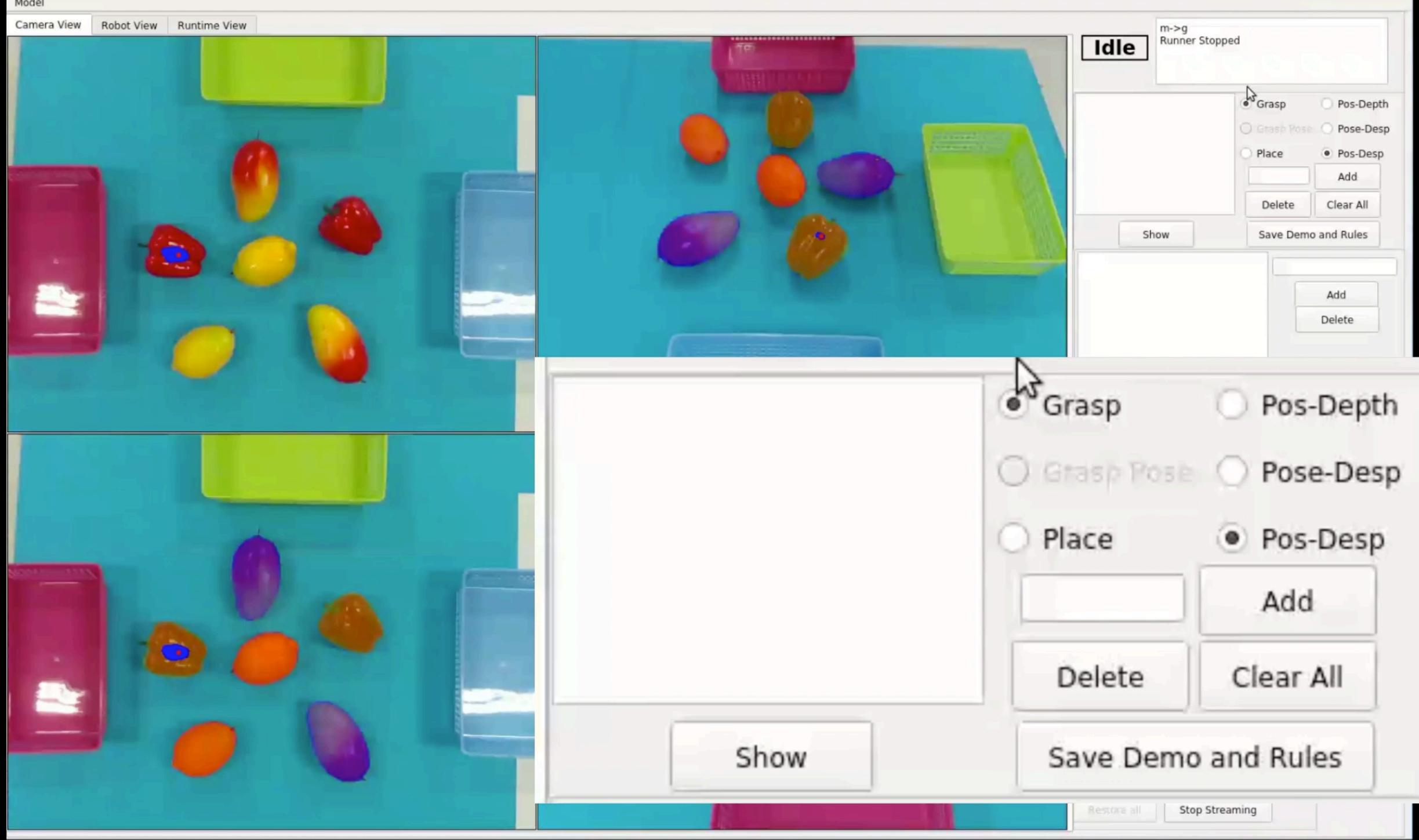
# Backup Slides

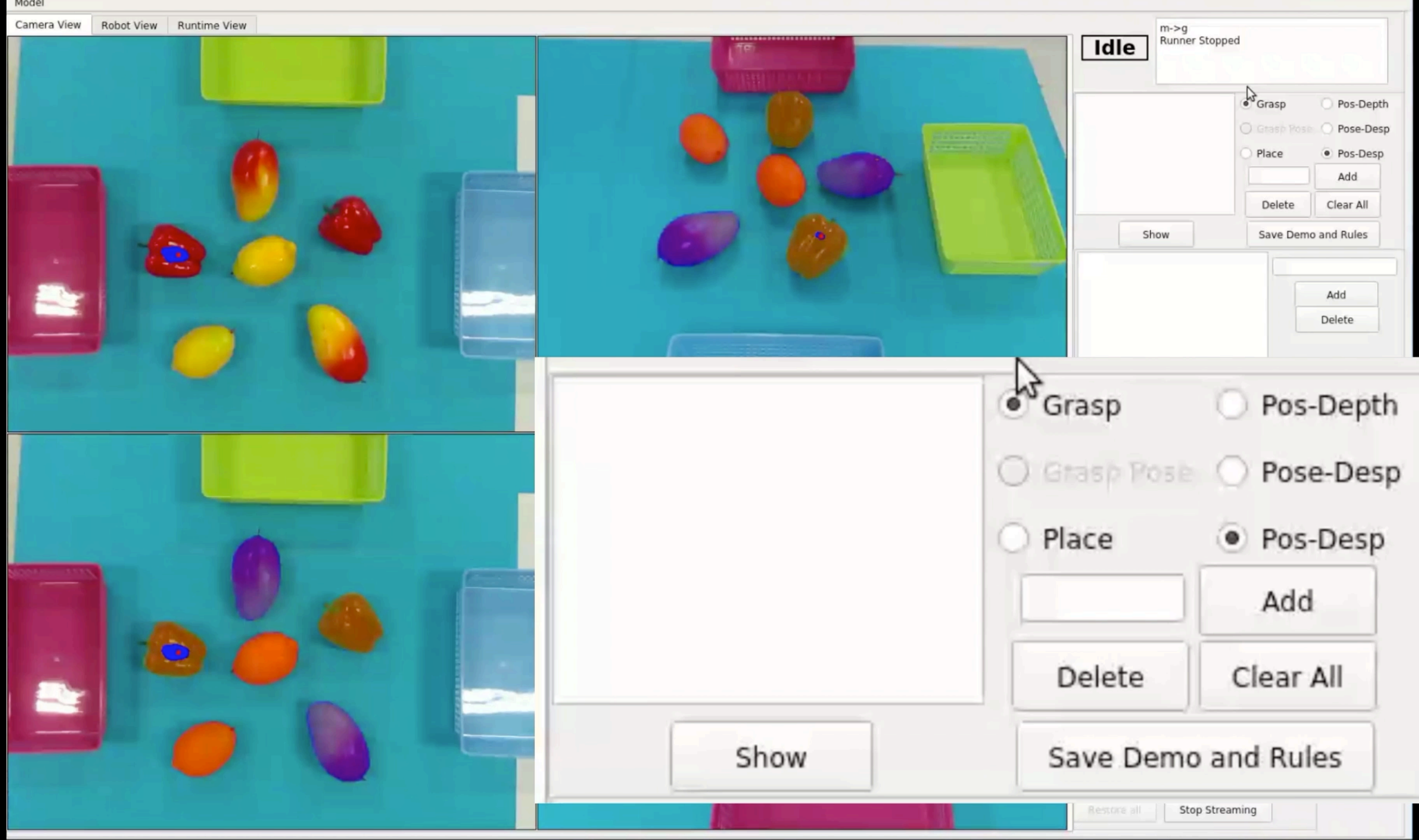
# Pick-and-Place tasks

- Sequential pick and place (Brick stacking)

# Pick-and-Place tasks

- Sequential pick and place (Brick stacking)


Model
Camera View
Robot View
Runtime View
Idle
m->g
Runner Stopped
Grasp
Pos-Depth
Grasp Pose
Pose-Desp
Place
Pos-Desp
Add
Delete
Clear All
Show
Save Demo and Rules
Add
Delete
Grasp
Pos-Depth
Grasp Pose
Pose-Desp
Place
Pos-Desp
Add
Delete
Clear All
Show
Save Demo and Rules
Restore all
Stop Streaming


Model
Camera View
Robot View
Runtime View
Idle
m->g
Runner Stopped
Grasp
Pos-Depth
Grasp Pose
Pose-Desp
Place
Pos-Desp
Add
Delete
Clear All
Show
Save Demo and Rules
Add
Delete
Grasp
Pos-Depth
Grasp Pose
Pose-Desp
Place
Pos-Desp
Add
Delete
Clear All
Show
Save Demo and Rules
Restore all
Stop Streaming

# Limitations

- Intra-class consistency is not guaranteed in learning
- Pose estimation for objects
- Matching performance under severe occlusion
- Viewpoint variations v.s. synthetic patches

# Block Stacking Methods

TABLE I

COMPARISON OF THE METHODS FOR BLOCK STACKING.

| Work | Duan et al. [15] | Nair et al. [7] | Zhu et al. [16] | Tremblay et al. [6] | Xu et al. [12] | Ours |
|---|---|---|---|---|---|---|
| Simulation environment | Yes | Yes | Yes | Yes | Yes | **No** |
| Synthetic demonstration for learning | Yes | **No** | **No** | Yes | Yes | **No** |
| Human demonstration for learning | No | Yes | Yes | **No** | **No** | **No** |
| Real-world evaluation | No | No | **Yes** | **Yes** | **Yes** | **Yes** |
| Task-specific object detector in the real world | - | - | Yes. Transferred auxiliary tasks | Yes. Transferred | Yes. Task-specific APIs | **No. Use dense correspondences across all tasks** |
| Stacking goals can be assigned by human demonstration. | - | - | No. Fixed stacking order | **Yes. Human demonstrations** | No, Synthetic videos as task specifications | **Yes. Object-centric demonstrations** |

# The margins in our experiments

$$M_{bk-nm} = 10$$
$$M_{hn} = M_{tp} = 5$$
$$M_{bk-m} = 2$$