Exploiting Unlabeled Data in Computer Vision

Xiaohang Zhan
MMLab, The Chinese University of Hong Kong

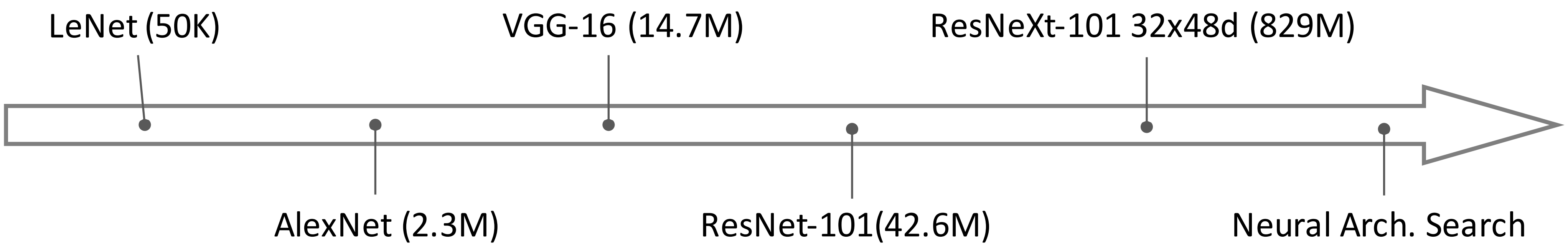
at Tsinghua Shenzhen International Graduate School
Oct. 20, 2020

Outlines

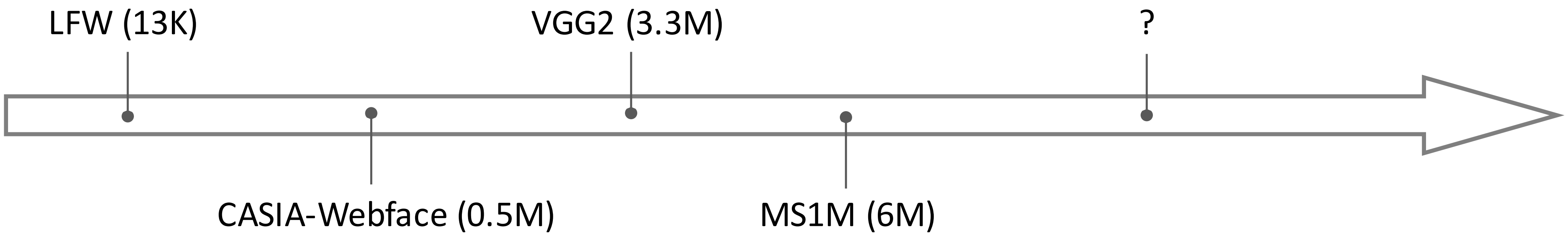
- Why unlabeled data?
- Supervised **face** clustering: a new trend
- Unsupervised representation learning from **object-centric** images
- Self-supervised learning in **scene** understanding

Neural Networks v.s. Labeled Datasets

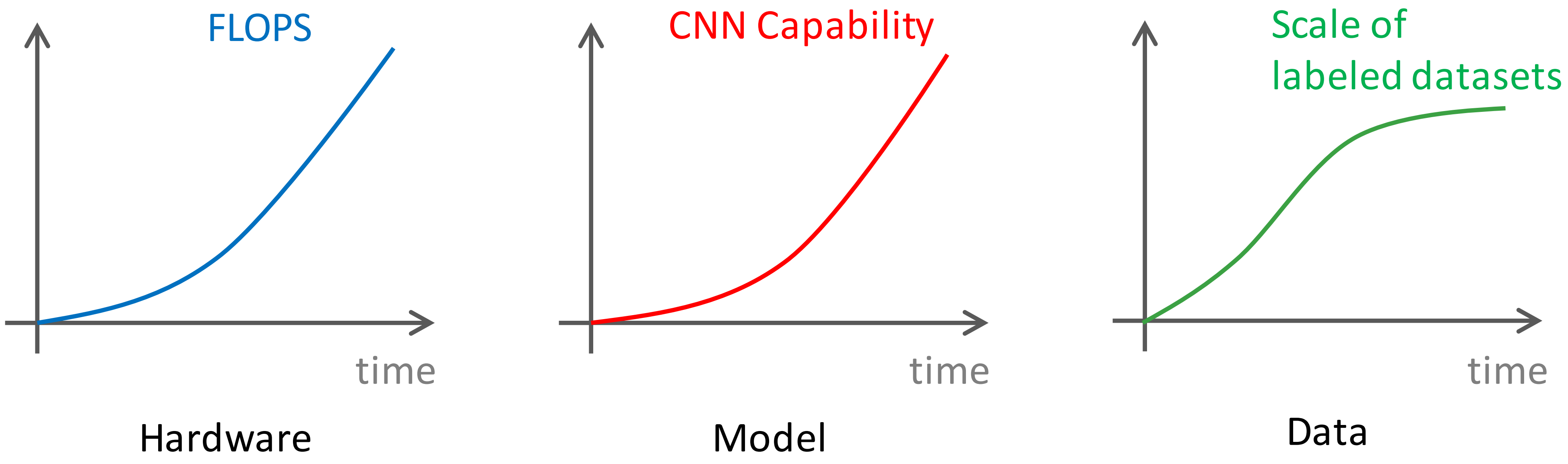
Evolution of CNN architectures (number of parameters in convolution layers)



Evolution of labeled datasets (number of images)



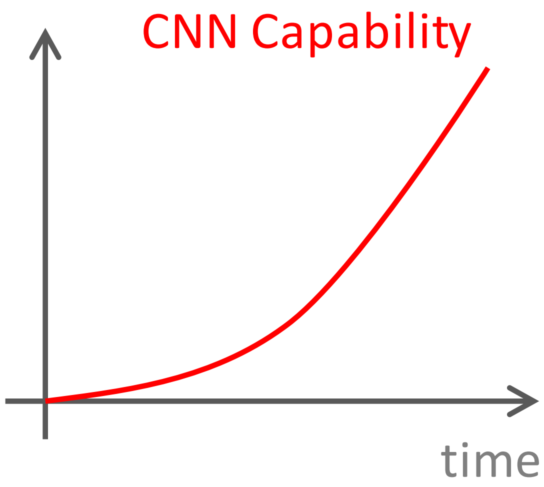
Neural Networks v.s. Labeled Datasets



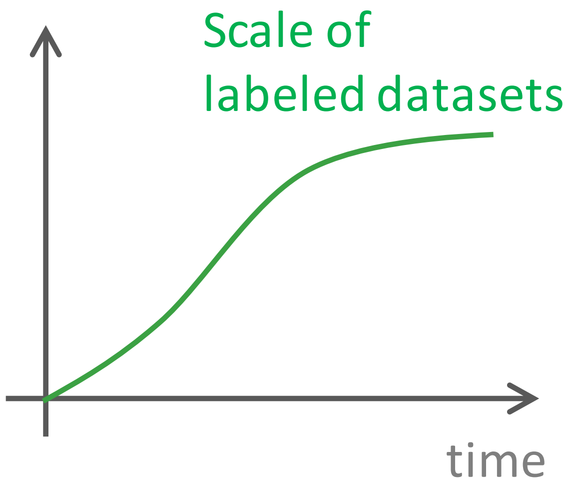
Neural Networks v.s. Labeled Datasets

Issues in labeling datasets:

- High labor cost
- Annotation noise and bias
- Low production speed (several years)



Model



Data

Unlabeled Data in Different Forms

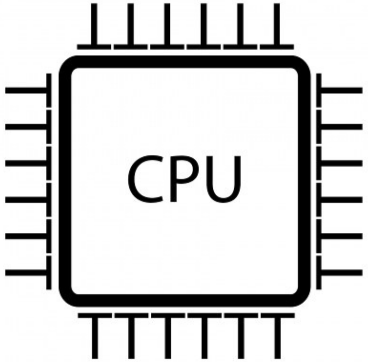

Curated Faces


Object-Centric Images

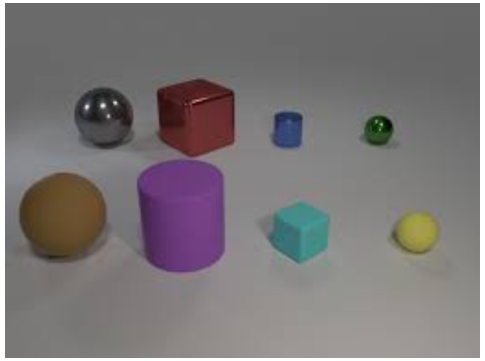

Natural Scenes

Conventional Unsupervised Learning

Small scale datasets



Low efficiency



Limited scenarios

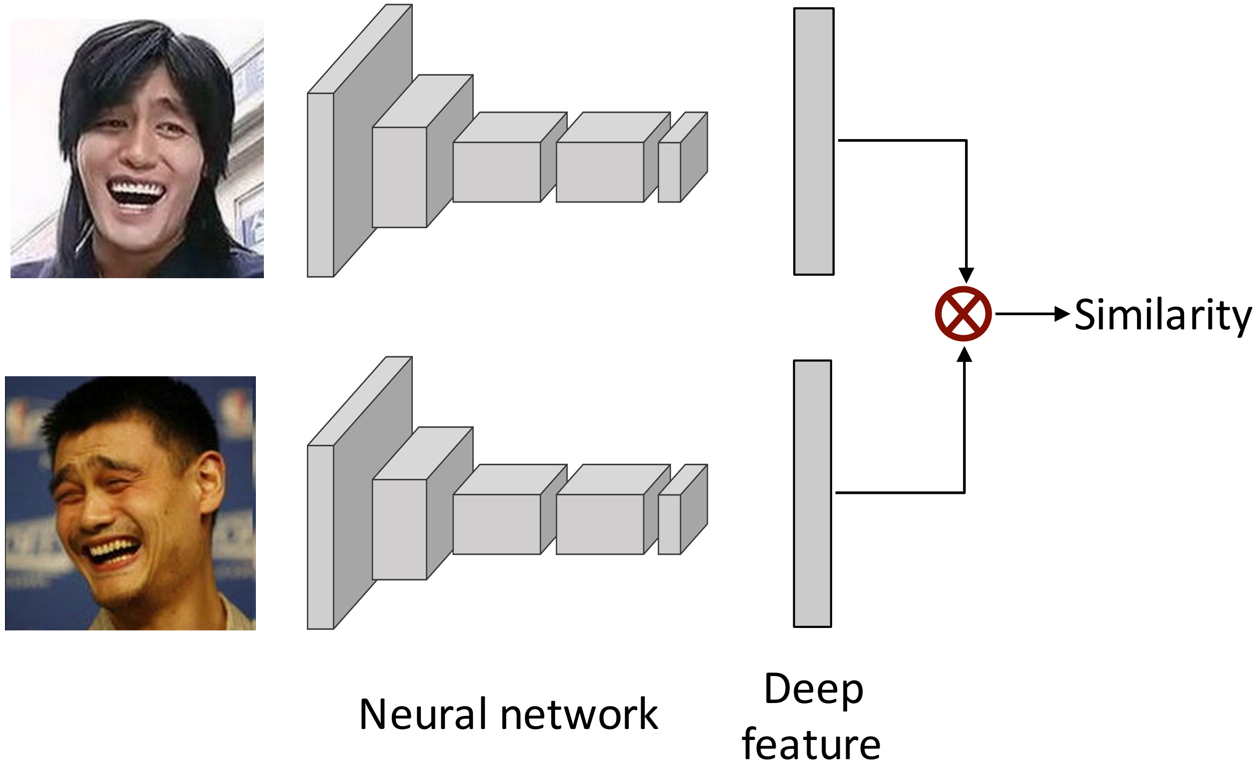
How to better leverage unlabeled data in the era of deep learning?

Outlines

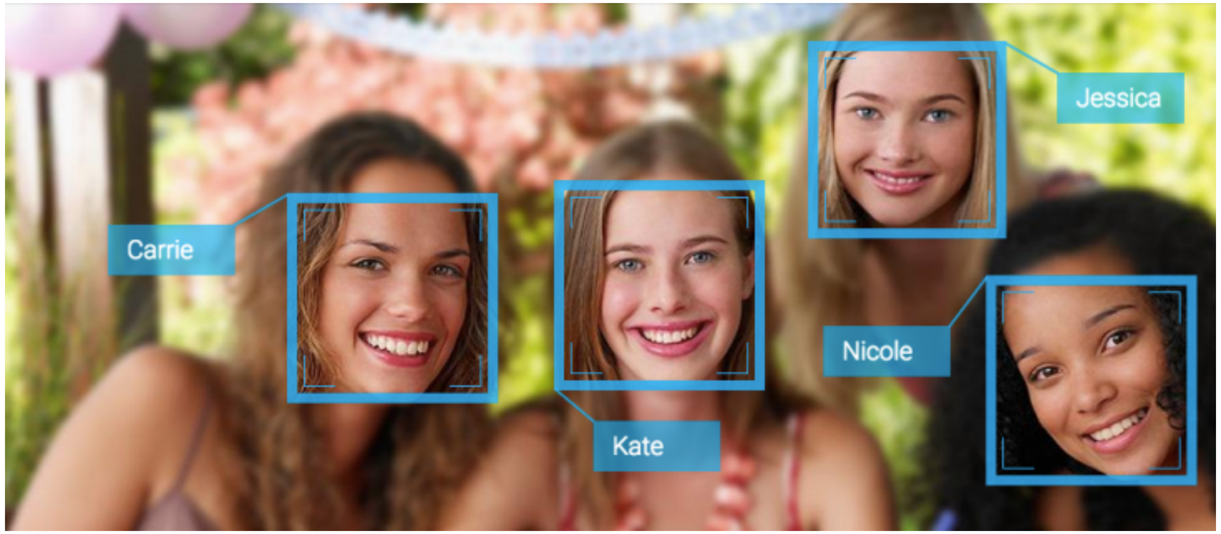
- Why unlabeled data?
- Supervised **face** clustering: a new trend
- Unsupervised representation learning from **object-centric** images
- Self-supervised learning in **scene** understanding

Face Recognition

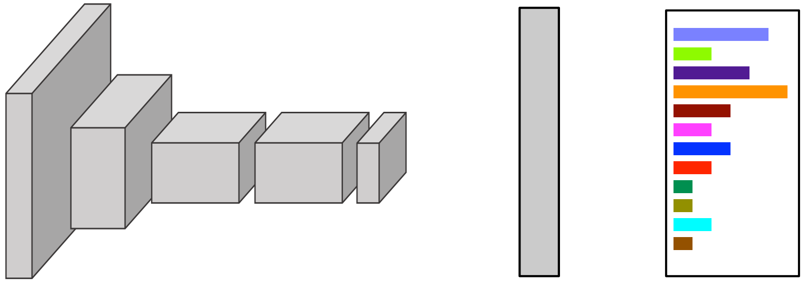
Face recognition in movies



Training of Face Recognition



Labeled data

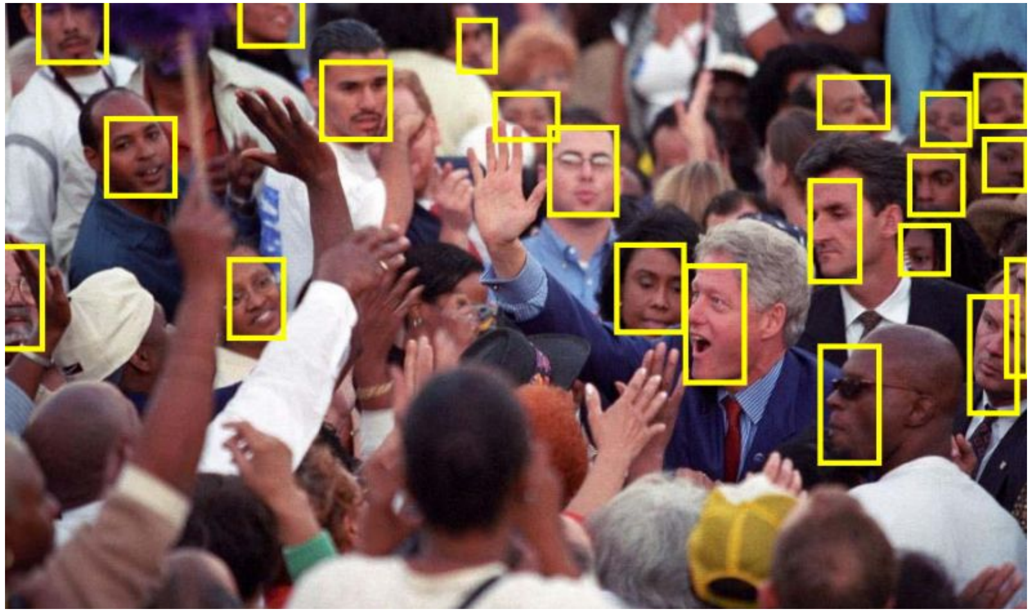


Deep feature

Classifier

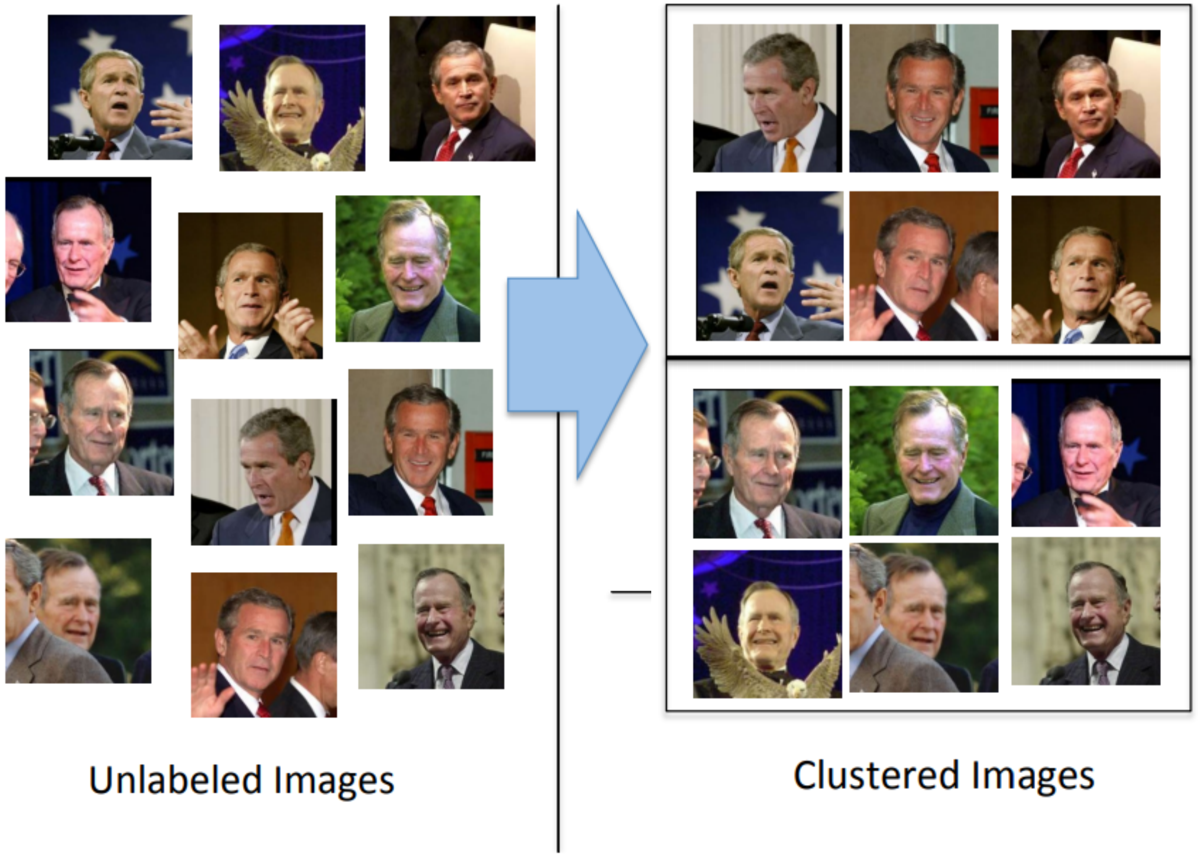
Big Data of Faces

surveillance

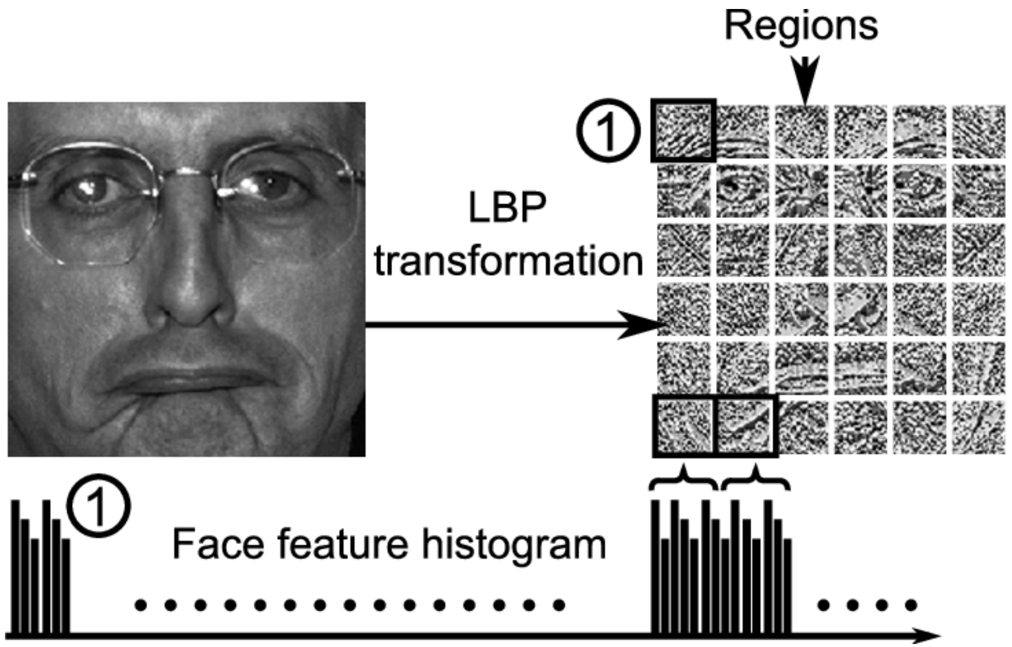


web

Face Clustering

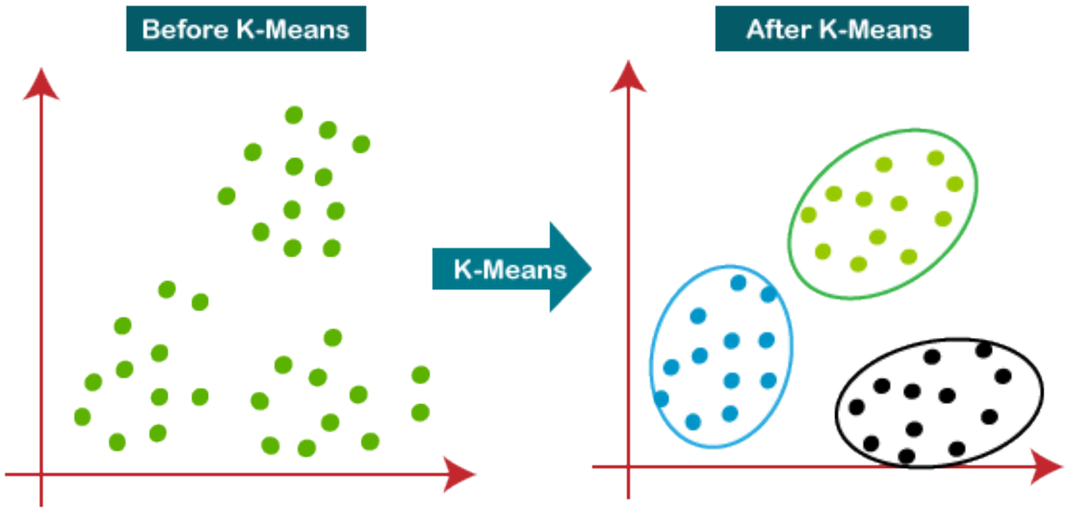


Face Clustering in Ancient Time



LBP features

- Low representability
- Vulnerable
- High dimension



K-Means clustering

- Relying on strong assumptions
- High computational cost

Challenges

1. Unlabeled data collected from unconstrained environments have large variations → hand designed features are unreliable.

occlusion

pose

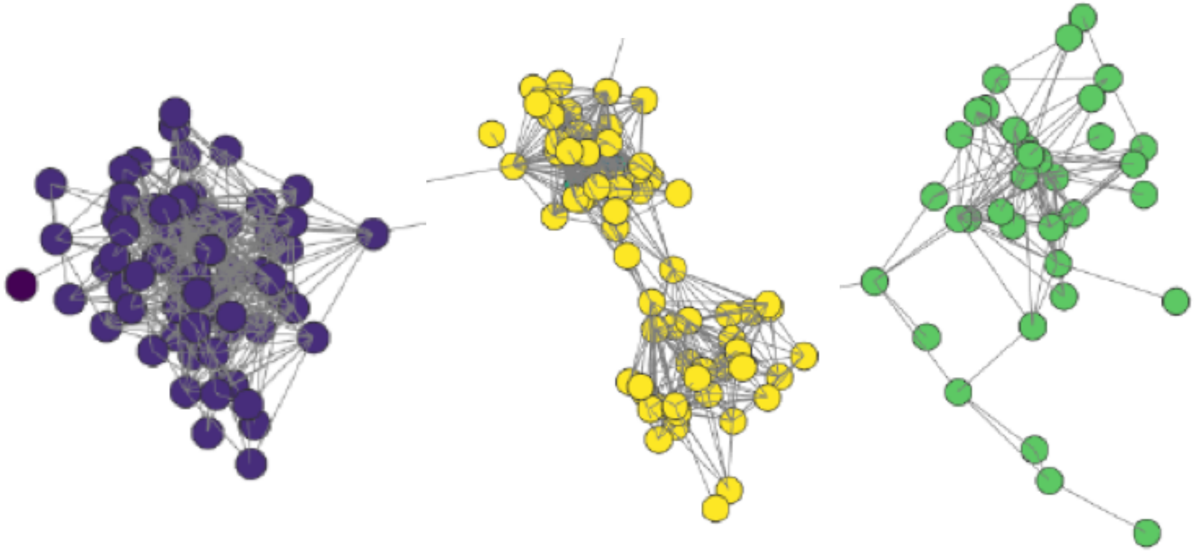
blur

dark

overexposure

Challenges

2. Complicated inner-structures → hard to use priors or assumptions



Sample face clusters in practical

Assumptions:

- KMeans: Samples obey Gaussian distribution.
- Spectral: Clusters' size is balanced.
- DBSCAN: Clusters are dense regions.

Challenges

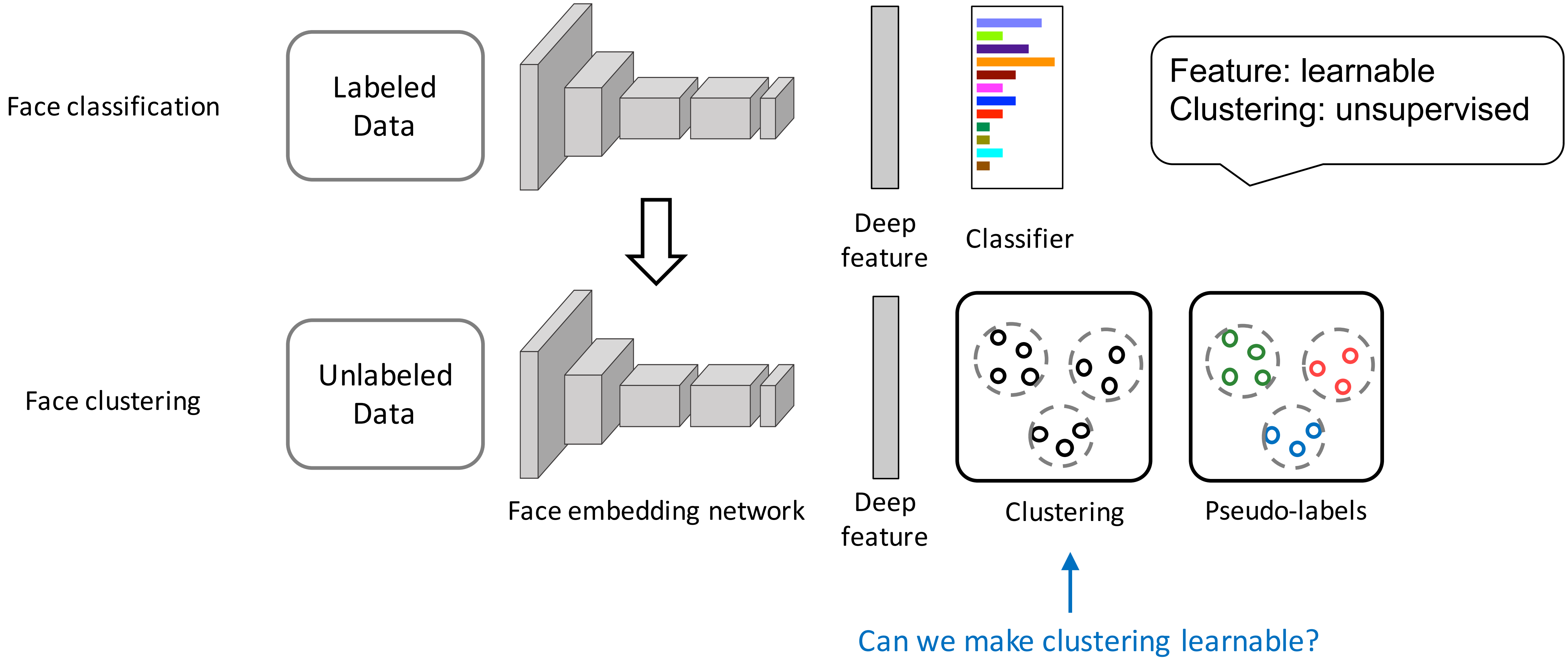
3. Large-scale clustering → computational complexity

Datasets	Images	Identities
VGG2	3.3 M	9 K
MegaFace	4.7 M	672K
MS1M	5.8 M	85 K
Surveillance	Billion-level	Million-level

Computation Complexity:

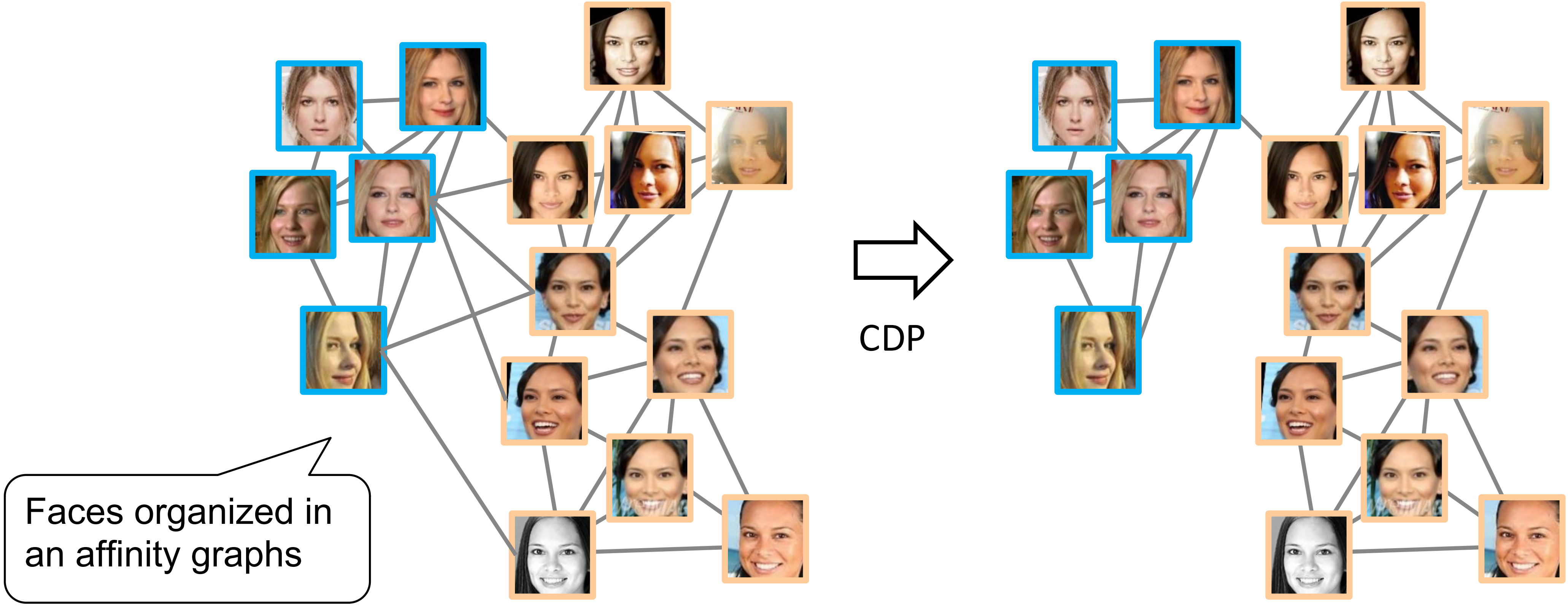
- KMeans: $O(N * Iter * K), K \in N$
- Spectral: $O(N^3)$
- DBSCAN: $O(N^2)$, or $O\big(Nlog(N)\big)$
- HAC: $O(N^3)$

Face Clustering in Deep Learning Era

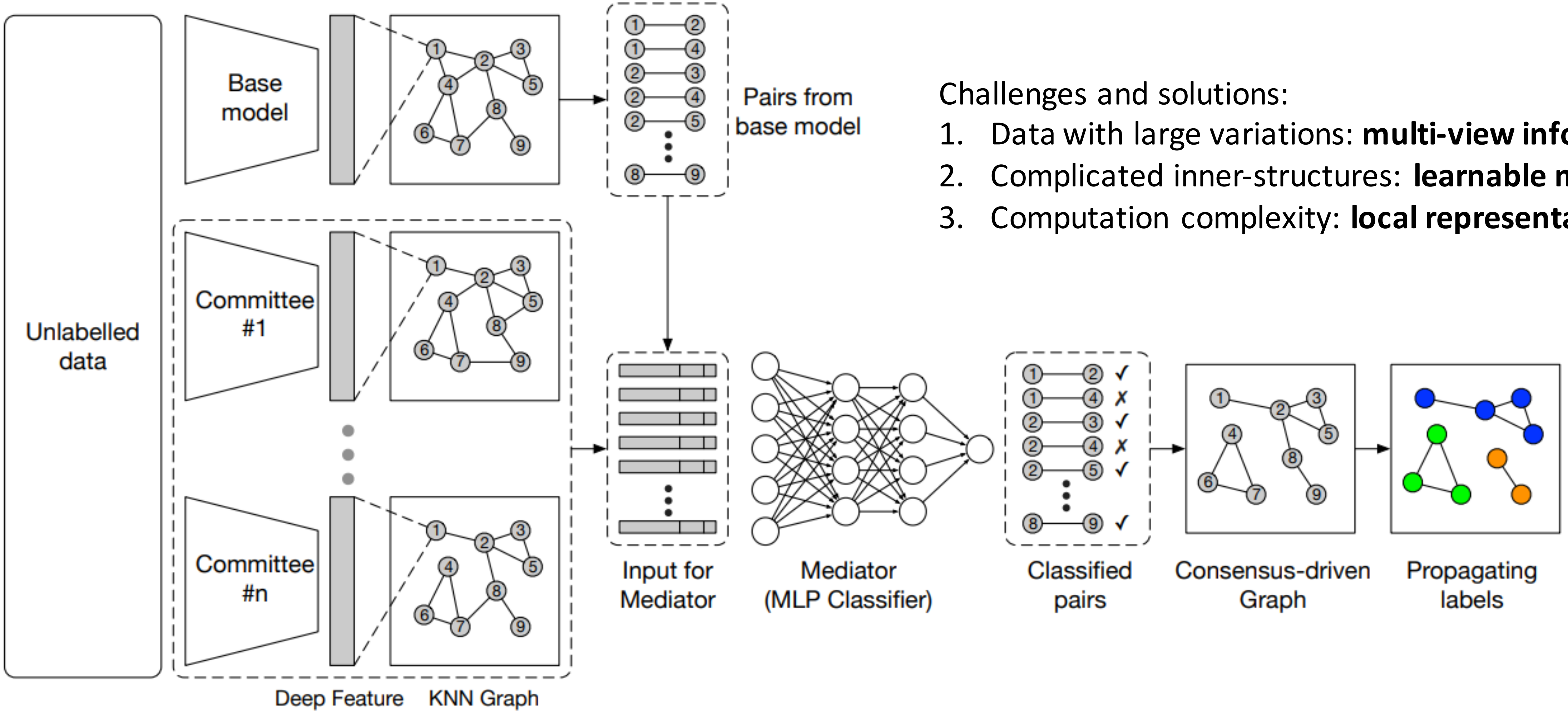


Consensus-Driven Propagation (CDP) [ECCV'18]

- **Objective**: Learning better linkages.



Consensus-Driven Propagation (CDP) [ECCV'18]



Challenges and solutions:

1. Data with large variations: **multi-view information**.
2. Complicated inner-structures: **learnable module**.
3. Computation complexity: **local representations**.

Consensus-Driven Propagation (CDP) [ECCV'18]

Comparison of different methods on MS1M face clustering

methods	F-score (%)			Time
	200K	600K	1.4M	600K
K-Means	83.5	fail	fail	fail
Mini-batch K-Means	88.9	84.0	fail	2266s
HAC	92.6	90.6	fail	61h
FastHAC	69.4	80.9	fail	16h
DBSCAN	79.0	76.2	fail	80h
HDBSCAN	86.1	81.5	fail	48h
CDP (single model)	89.2	86.7	85.2	**85s**
CDP (multi model)	**95.8**	**94.2**	**93.1**	556s

3400x faster

our method

Effectiveness of CDP

Improvements on face recognition in MegaFace through clustering

Data	Performance
9% labeled	61.78%
9% labeled + 91% unlabeled (**HAC**)	62.45%
9% labeled + 91% unlabeled (**CDP**)	78.18%
100% labeled	78.52%

CDP makes unlabeled data as effective as labeled ones.

CDP as a Data Cleaner

CDP cleans out low-quality faces.

From Image to Graph

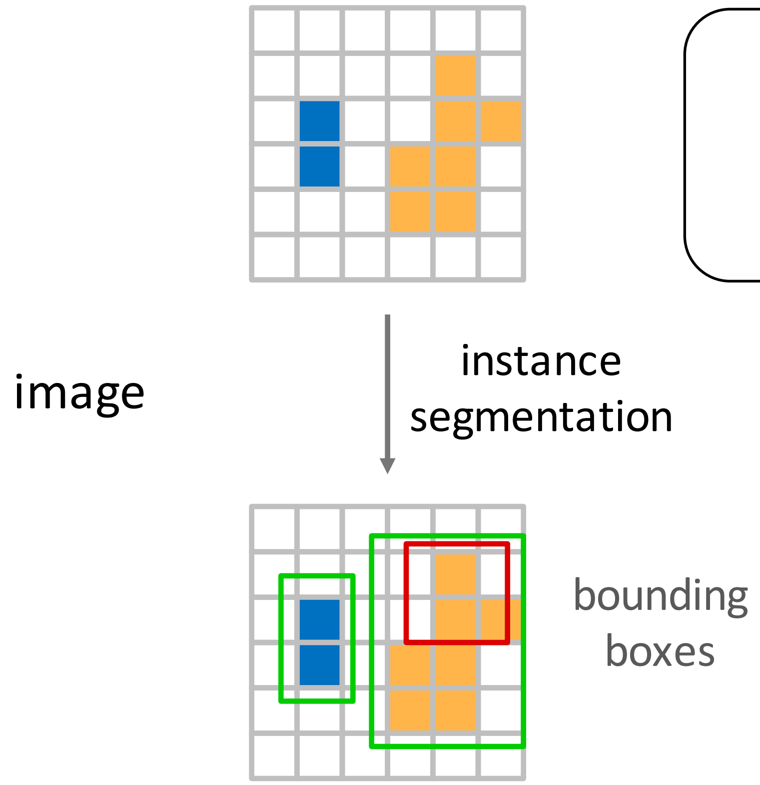
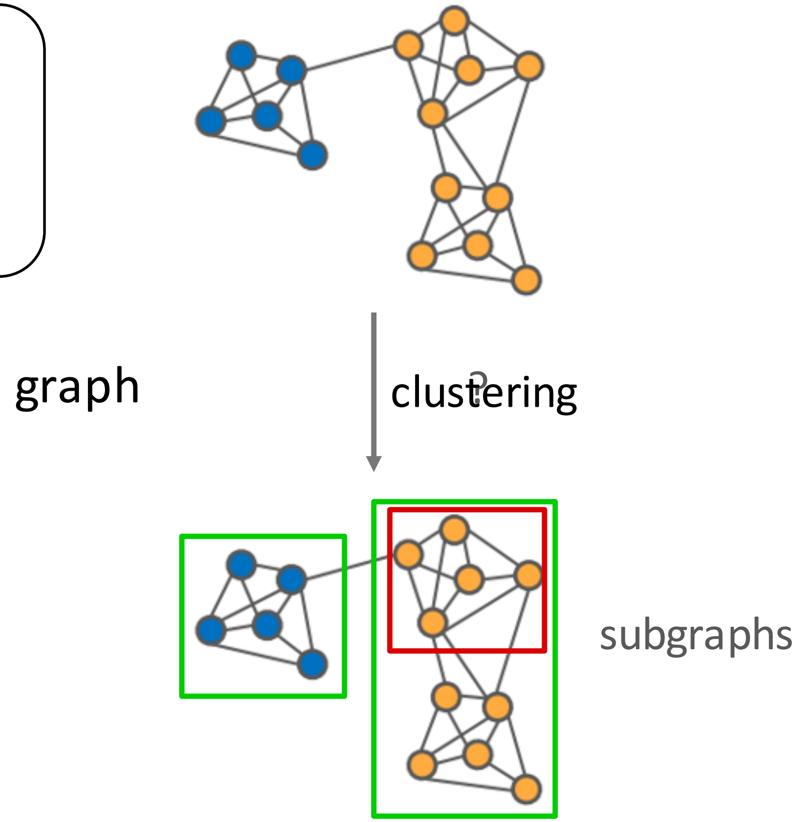
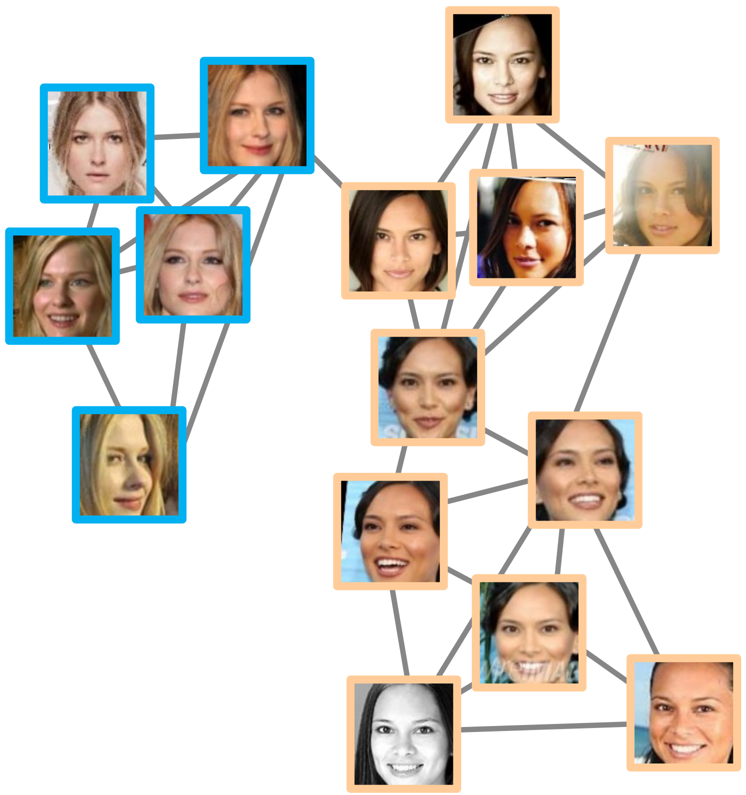
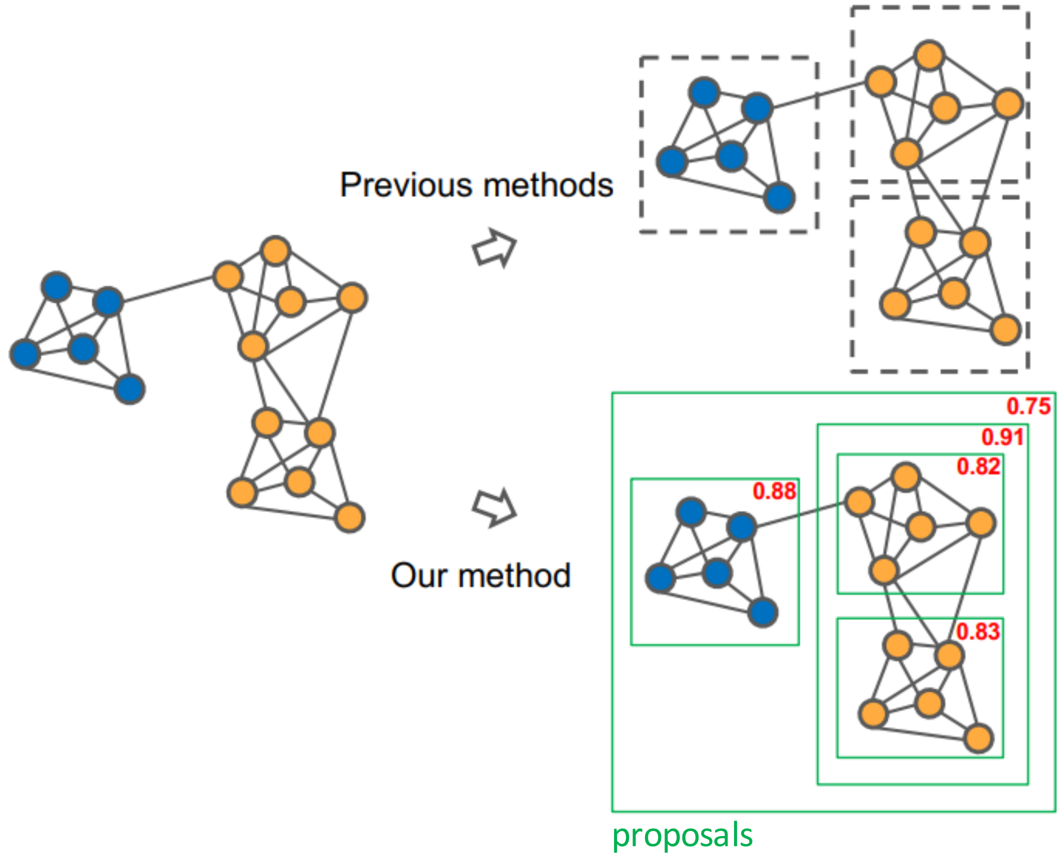


image → graph
pixel → vertex
grid → edge
CNN → GCN

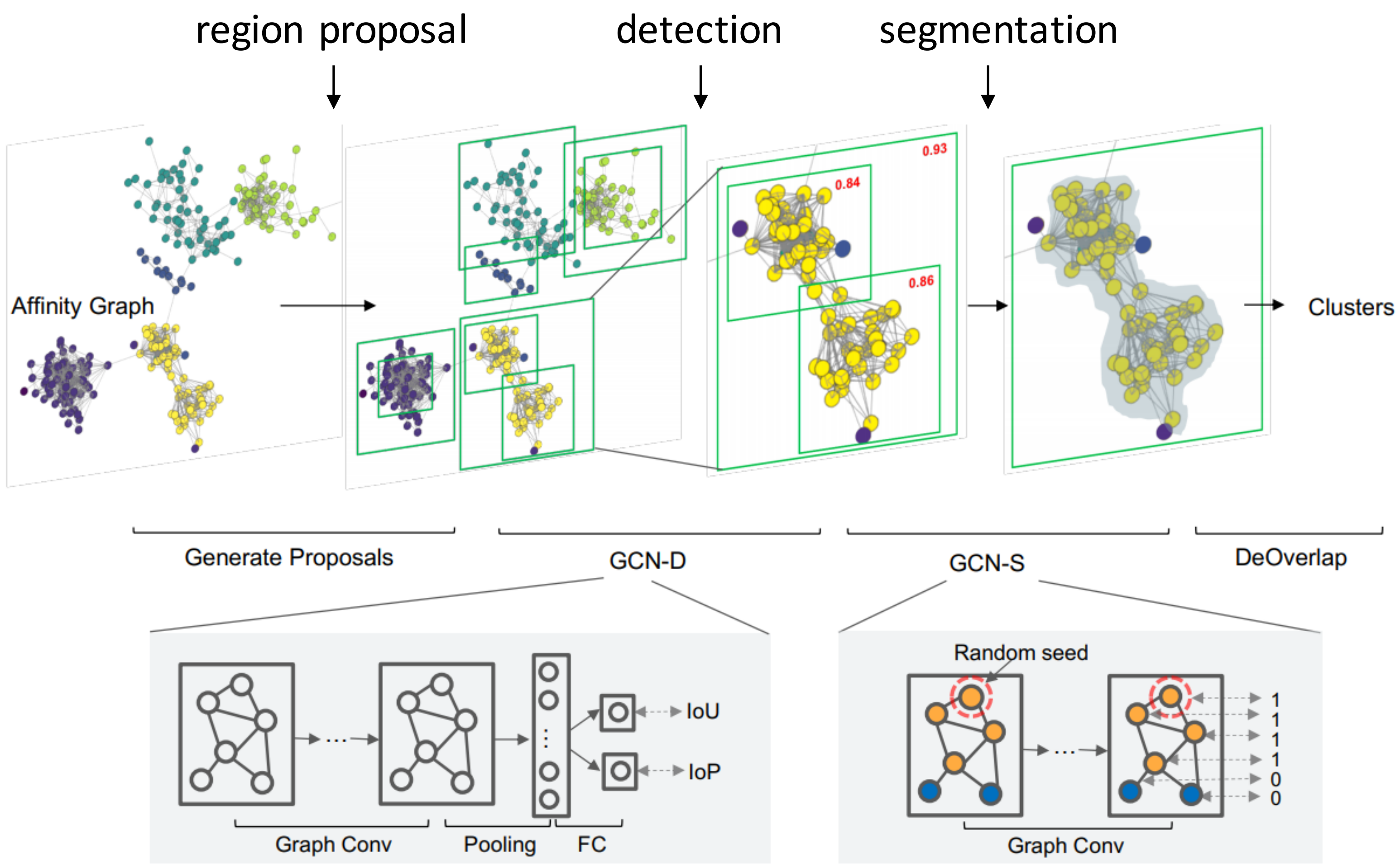


Learning to Cluster Faces on an Affinity Graph [CVPR'19 Oral]





Learning to Cluster Faces on an Affinity Graph [CVPR'19 Oral]



Face clustering as anchor-based detection

Learning to Cluster Faces on an Affinity Graph [CVPR'19 Oral]

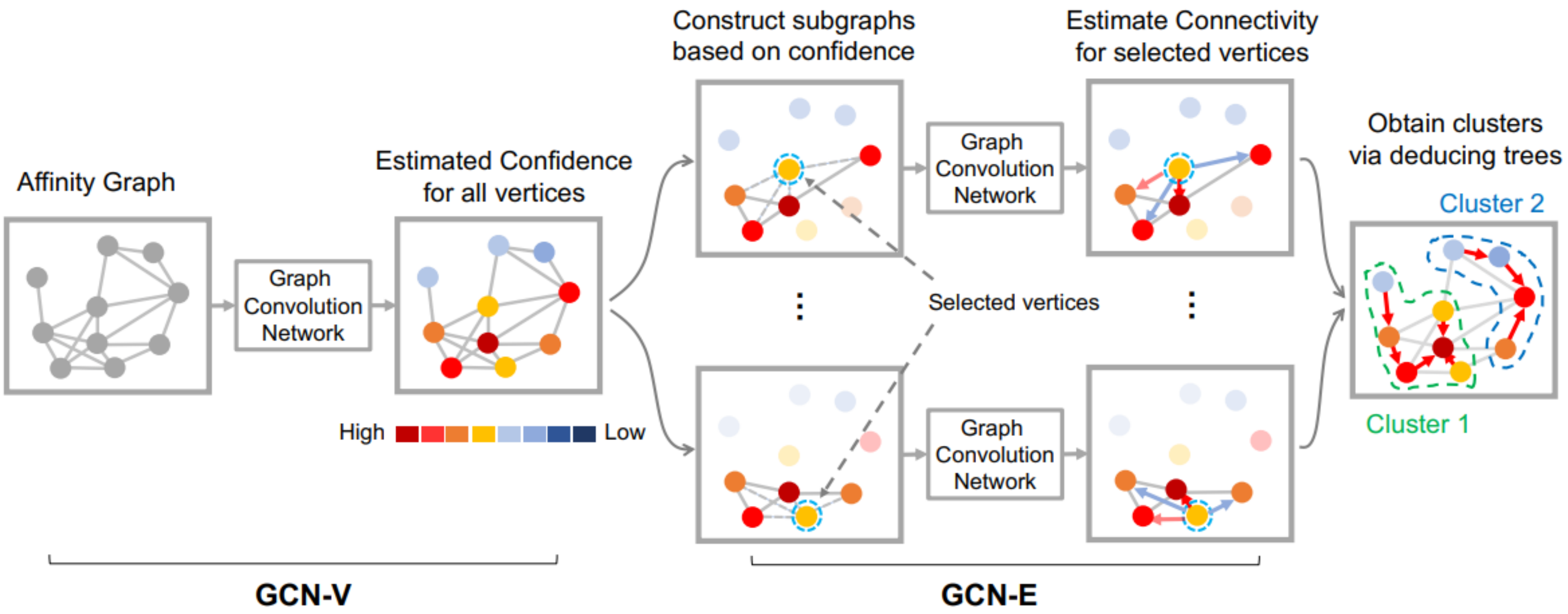
Evaluation on MS1M

Methods	#clusters	Precision	Recall	F-score	Time
K-Means [19]	5000	52.52	70.45	60.18	13h
DBSCAN [4]	352385	72.88	42.46	53.5	**100s**
HAC [24]	117392	66.84	70.01	68.39	18h
Approximate Rank Order [1]	307265	81.1	7.3	13.34	250s
CDP [30]	29658	80.19	70.47	75.01	350s
GCN-D	19879	95.72	76.42	84.99	2000s
GCN-D + GCN-S	19879	98.24	75.93	**85.66**	2200s

Slightly slower

Much stronger

Learning to Cluster Faces via Confidence and Connectivity Estimation [CVPR'20]

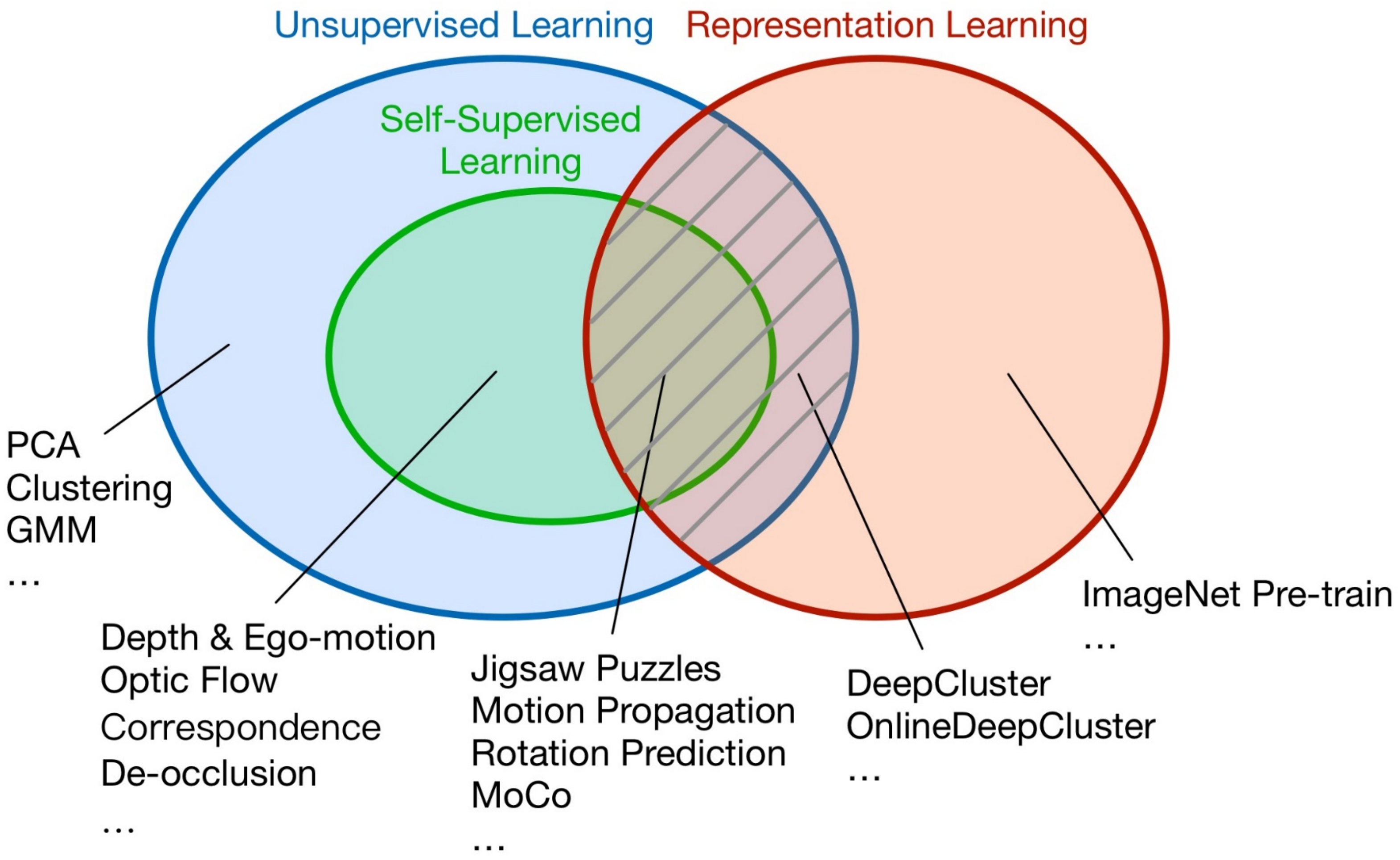


Face clustering as anchor-free detection

Outlines

- Why unlabeled data?
- Supervised **face** clustering: a new trend
- Unsupervised representation learning from **object-centric** images
- Self-supervised learning in **scene** understanding

What is Unsupervised Representation Learning?

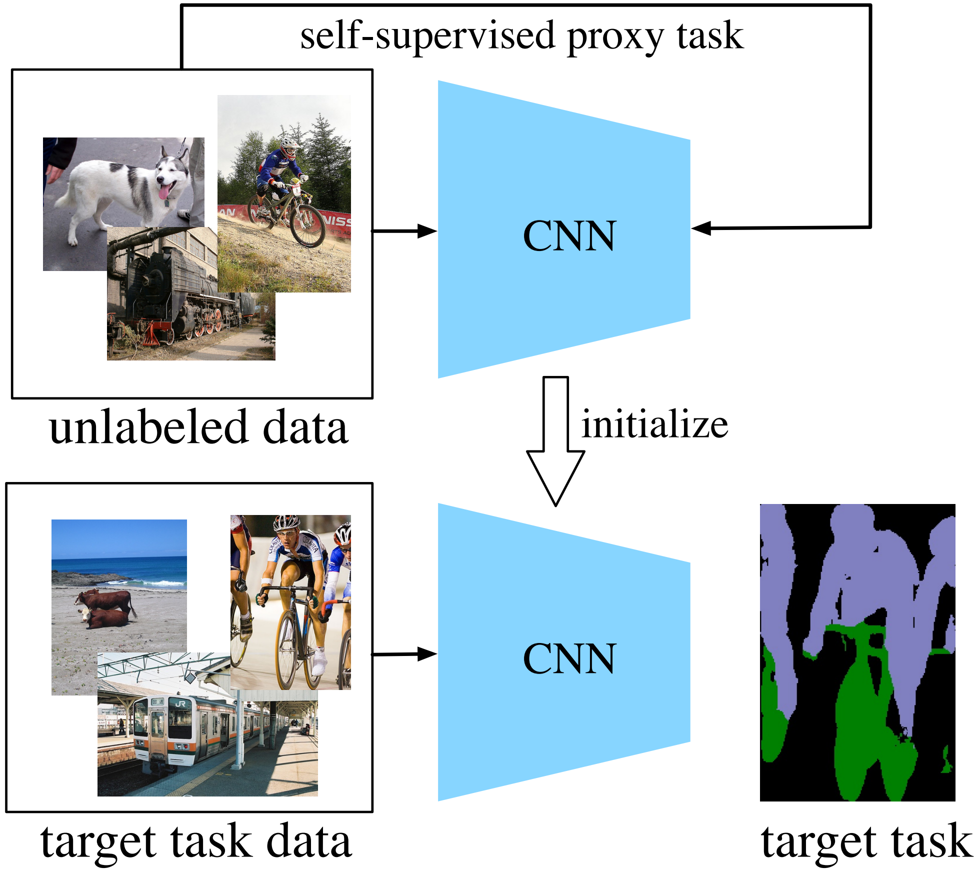


What is Self-Supervised Learning (SSL)?

Self-Supervised → Annotation free

Learning → To learn something new

Does image in-painting belong to self-supervised learning?



A typical pipeline

Self-Supervised Proxy/Pretext Tasks

Image Colorization

Solving Jigsaw Puzzles

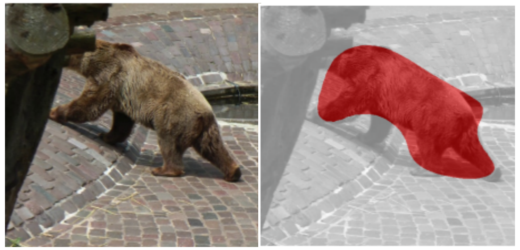
Image In-painting

Rotation Prediction

Instance Discrimination

Counting

Motion prediction



Moving foreground segmentation

Motion propagation

Essence: 1. Prior

- Appearance prior

Image Colorization

Image In-painting

- Physics prior

Rotation Prediction

- Motion tendency prior

Motion prediction
(Fine-tuned for seg: 39.7% mIoU)

- Kinematics prior

Motion propagation
(Fine-tuned for seg: 44.5% mIoU)

Low-entropy priors are more predictable.

Essence: 2. Coherence

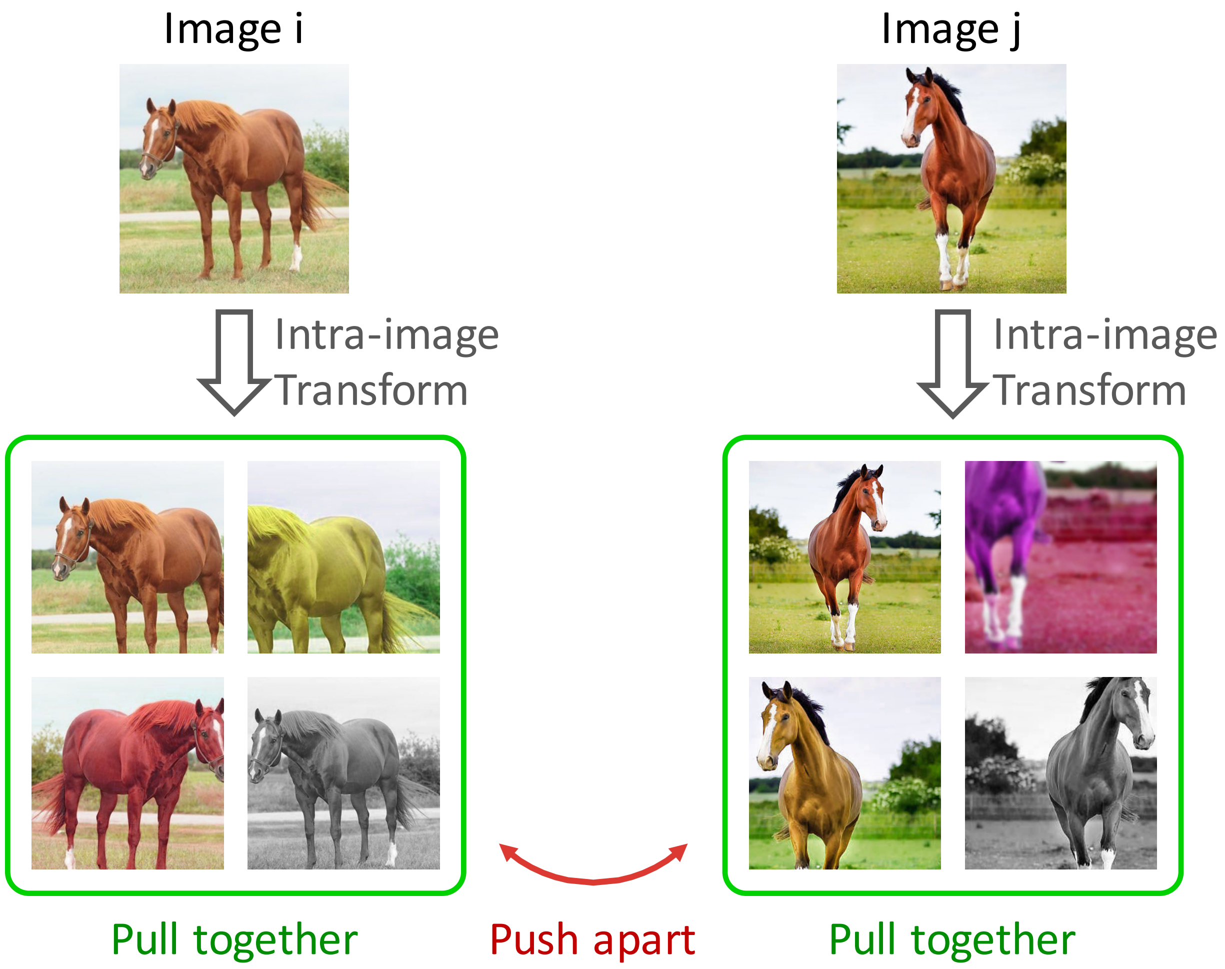
- Spatial coherence

Solving Jigsaw Puzzles

- Temporal coherence

Temporal order verification

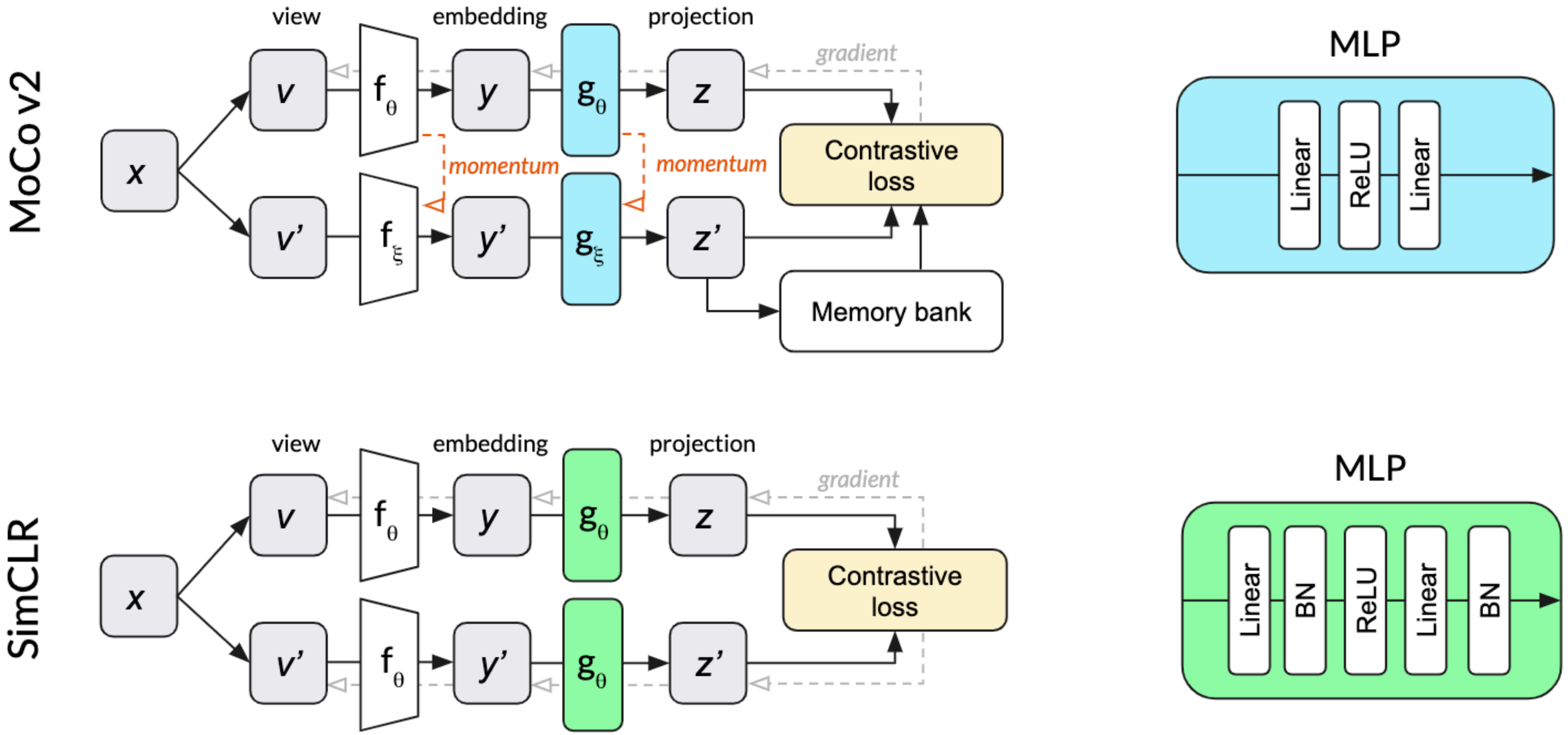
Essence: 3. Structure of Data



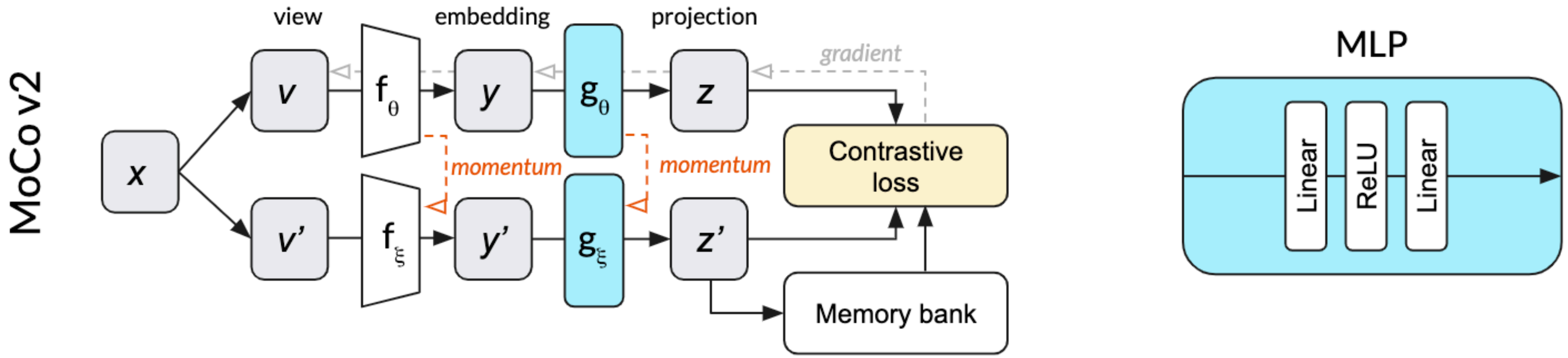
Instance Discrimination
(Contrastive Learning)

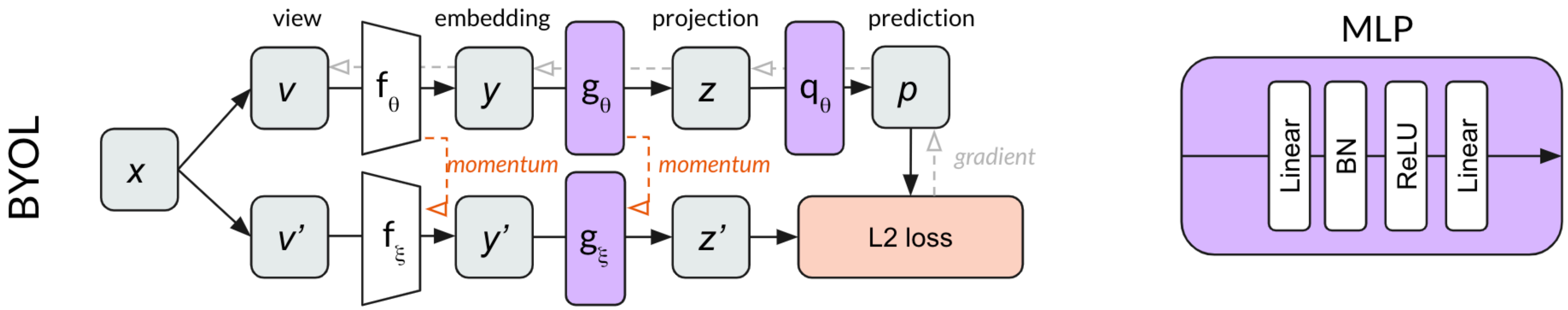
- NIPD
- CPC
- MoCo
- SimCLR
- ...

Typical Contrastive-Based SSL



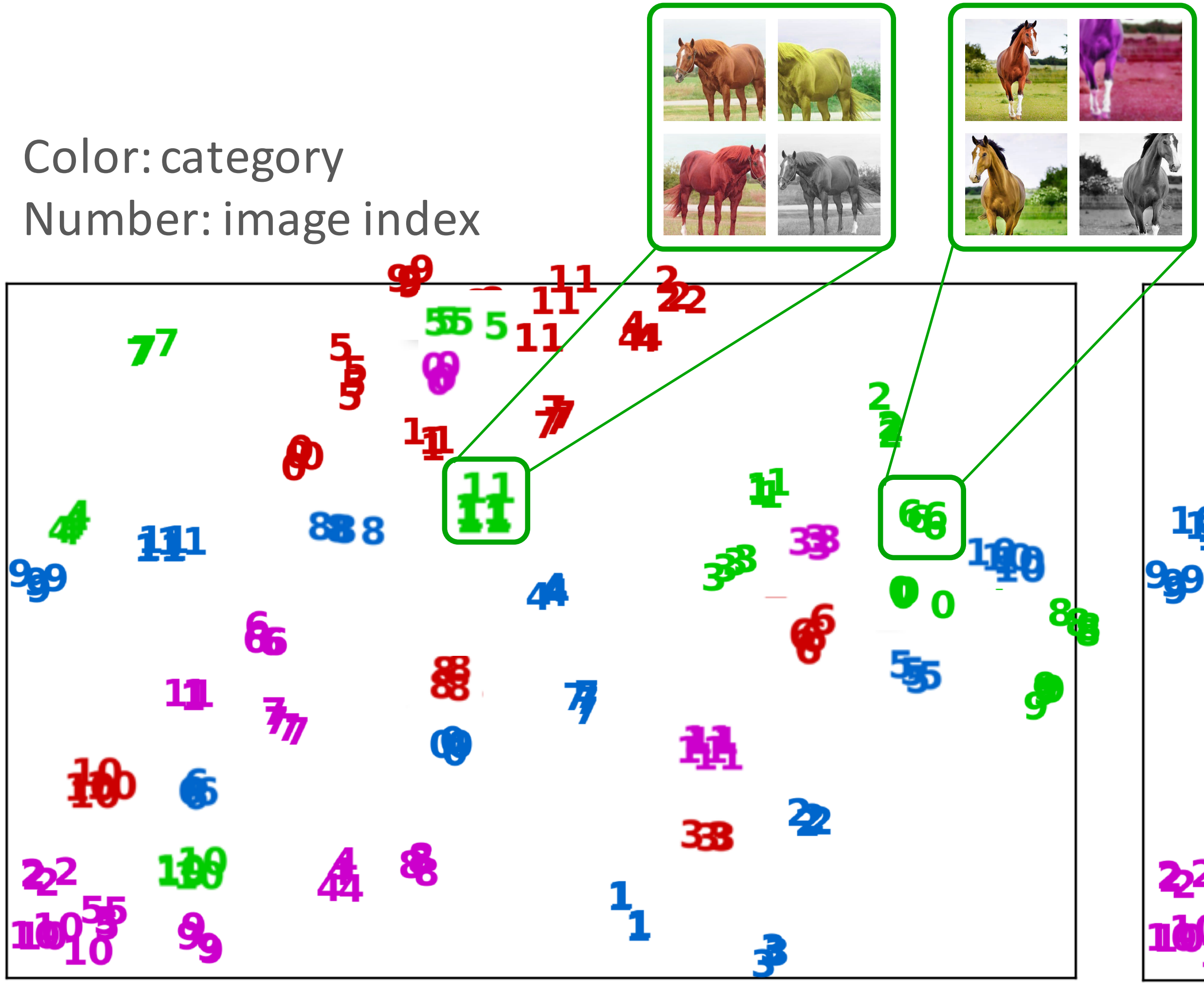
Typical Contrastive-Based SSL





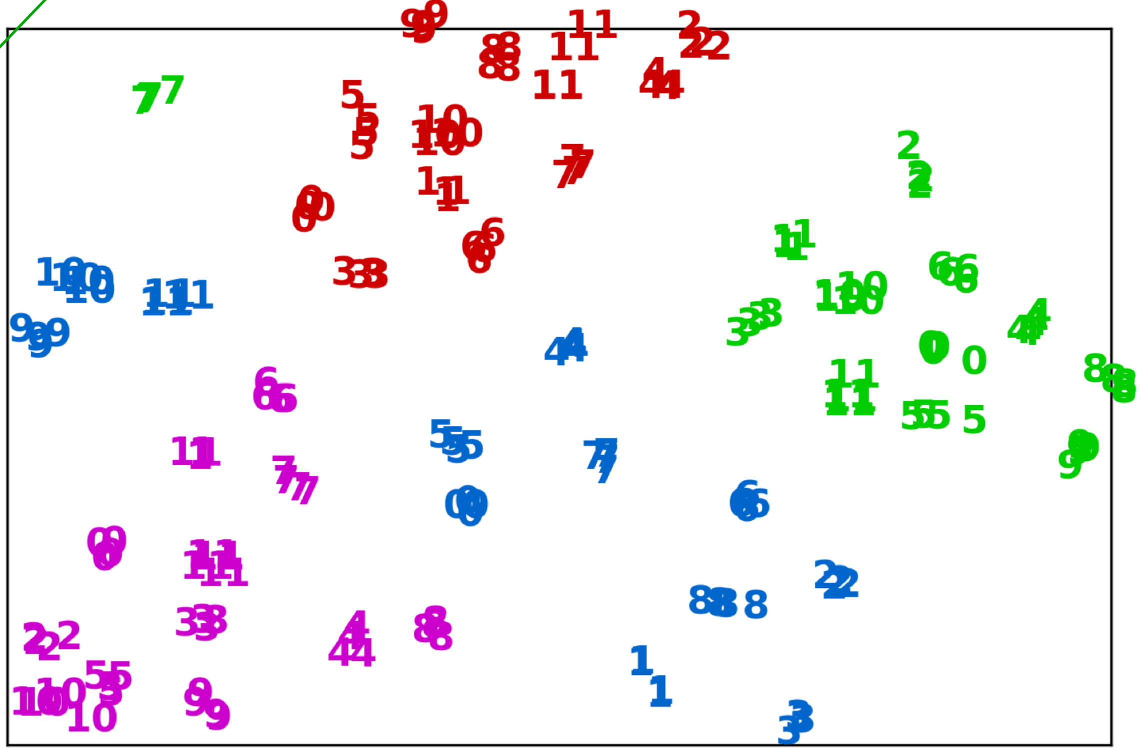
Essence: 3. Structure of Data

Color: category
Number: image index



Optimal solution (suppose)

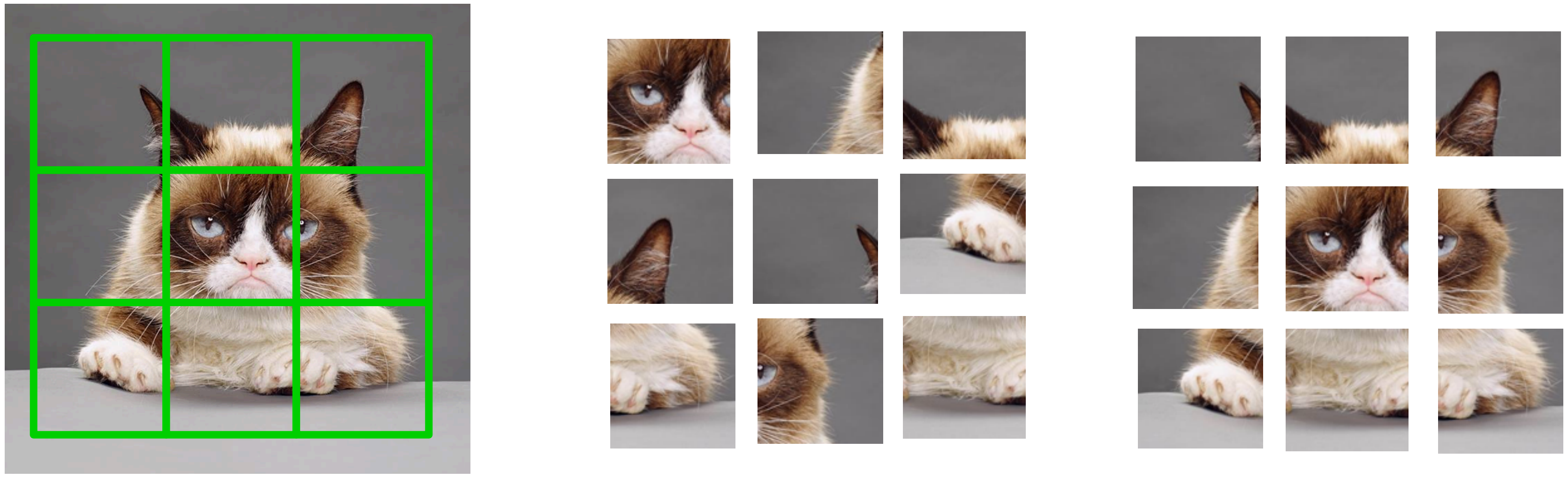
Both are optimal for Instance Discrimination. Why does the final optimized feature space look like the second case?



Optimal solution (actual)

Shortcuts to Avoid

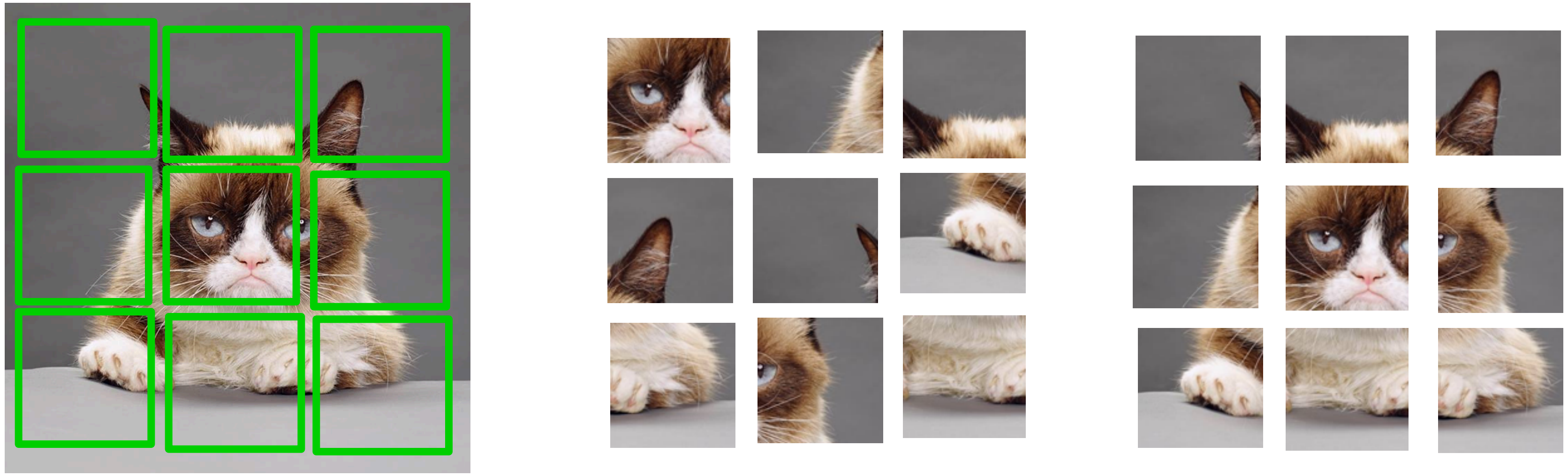
- Continuity



Solving Jigsaw Puzzles

Shortcuts to Avoid

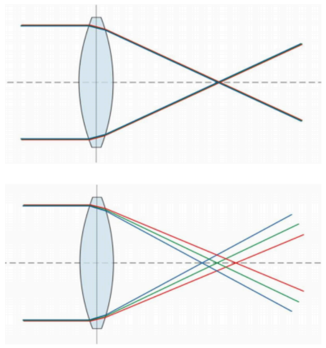
- Solution regarding continuity

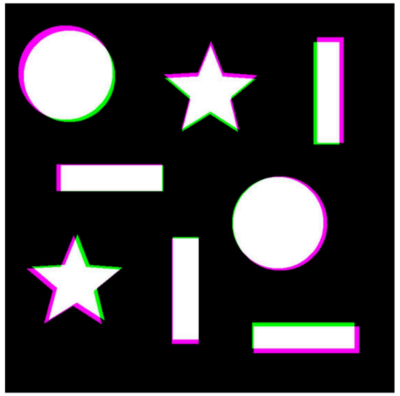


Solving Jigsaw Puzzles

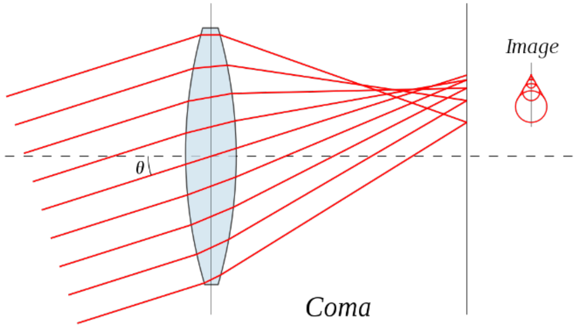
Shortcuts to Avoid

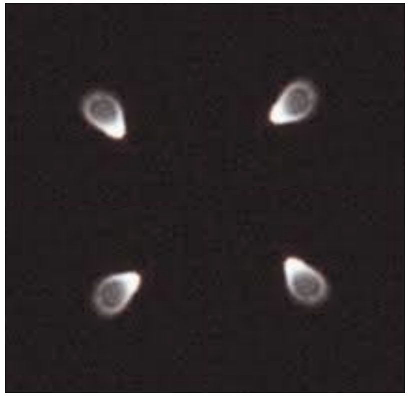
- Chromatic Aberration (色差)



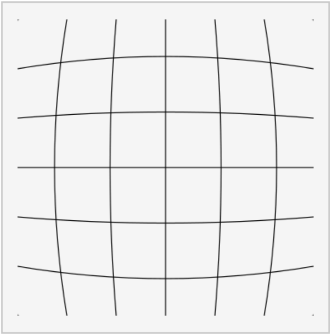


- Coma (彗差)

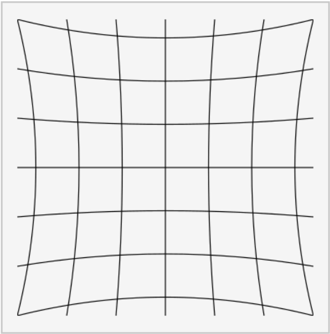




- Distortion (畸变)



Barrel-type



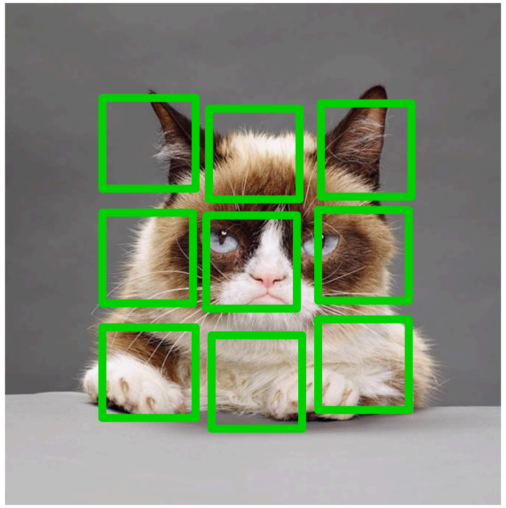
Pincushion-type

- Vignetting (暗角)

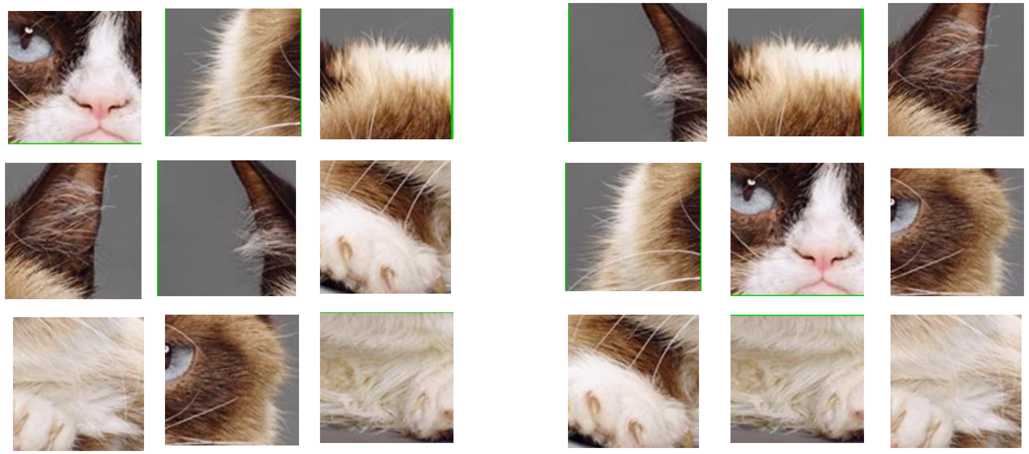
Do not apply heavy vignetting effects in your photos!!!

Shortcuts to Avoid

- Solution regarding aberration



After aberration correction



Solving Jigsaw Puzzles

Ambiguity

- Appearance prior

Image Colorization

Image In-painting

- Physics prior

Rotation Prediction

- Motion tendency prior

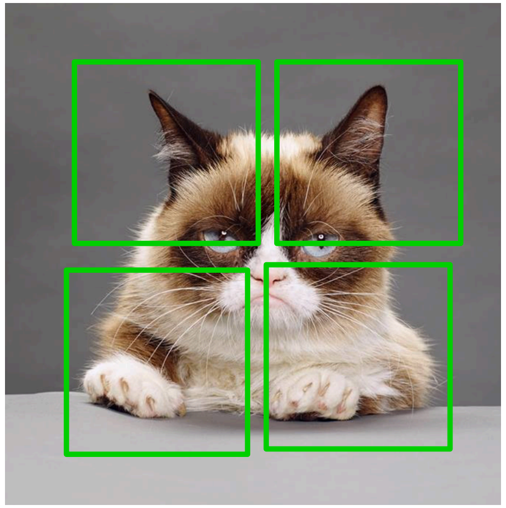
Motion prediction
(Fine-tuned for seg: 39.7% mIoU)

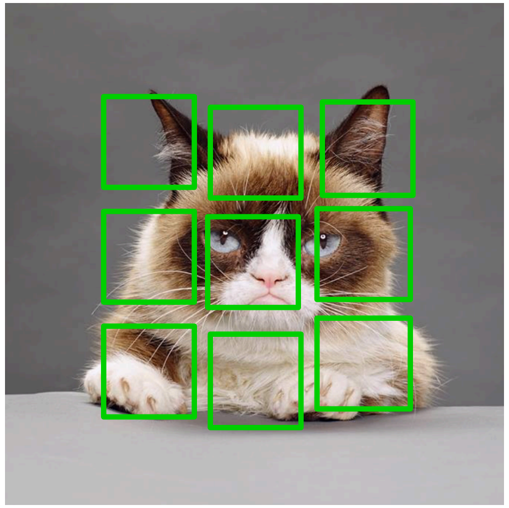
- Kinematics prior

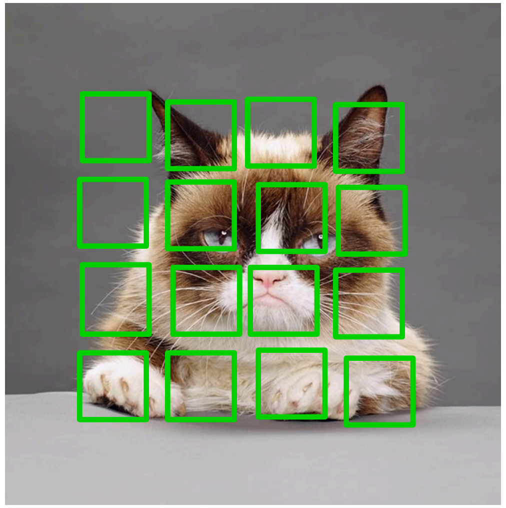
Motion propagation
(Fine-tuned for seg: 44.5% mIoU)

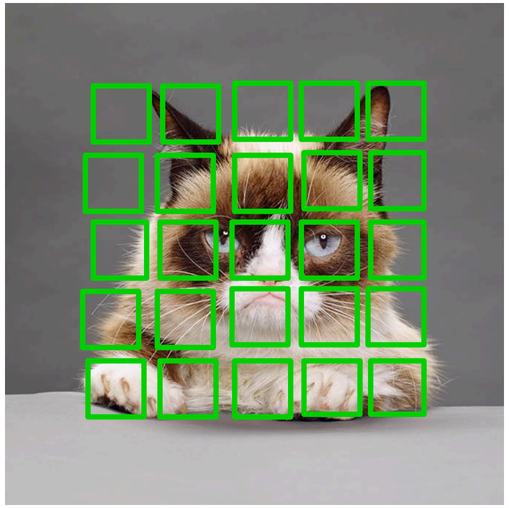
1. Low-entropy priors are less ambiguous.
2. Any other solutions?

Difficulty


Easy mode

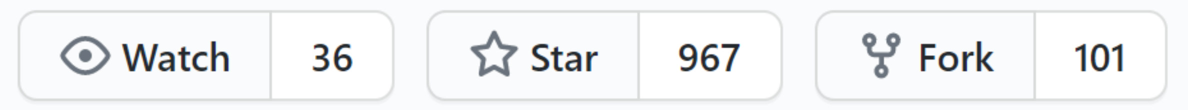

Normal mode


Difficult mode


Hell mode

How to design the difficulty of the task?

open-mmlab/OpenSelfSup



- High-efficiency
  - Distributed & Mixed Precision Training
- Integrity and Extensibility
  - All methods in one framework

Relative Location	Rotation Prediction	Deep Clustering	NPID
ODC	MoCo	SimCLR	BYOL

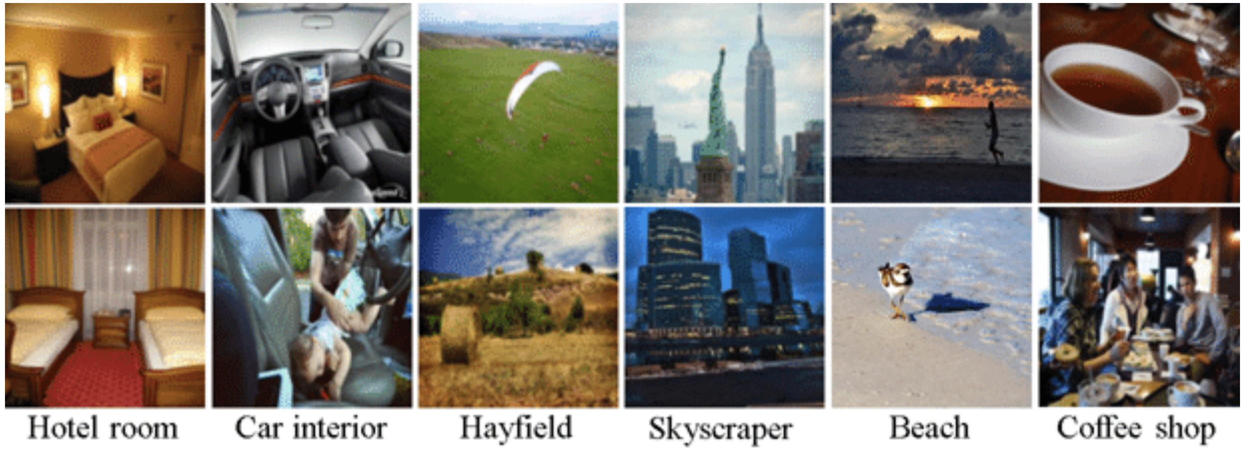
- Fair Comparisons
  - Standardized benchmarks

Linear classification	Semi-supervised classification	SVM & Low-shot SVM	Object detection

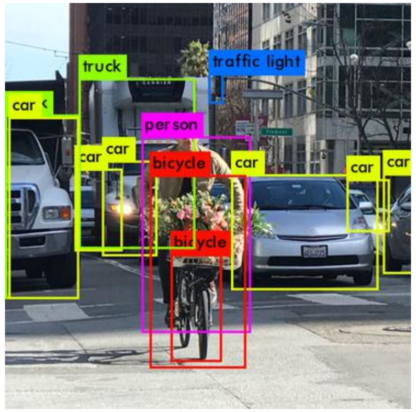
Outlines

- Why unlabeled data?
- Supervised **face** clustering: a new trend
- Unsupervised representation learning from **object-centric** images
- Self-supervised learning in **scene** understanding

Scene Understanding

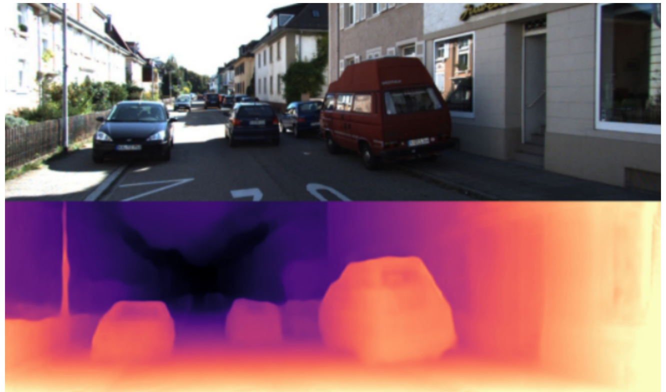


Scene classification



Object detection

Segmentation



Depth estimation

Learning from Motion

1. Motion reflects the kinematic properties or physical structures of objects.
2. Motion is easy to obtain, without manual annotations.

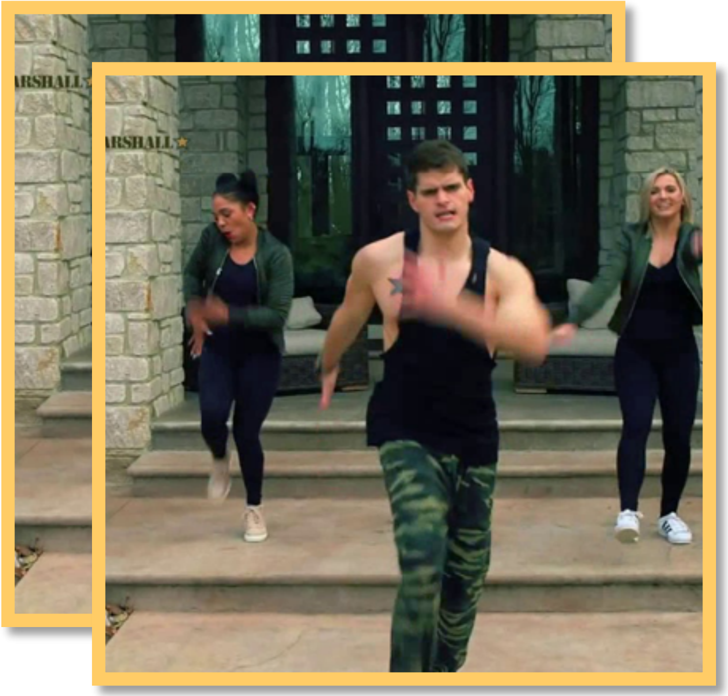


Image pair

Motion (optical flow)

Learning from Motion Tendency Priors

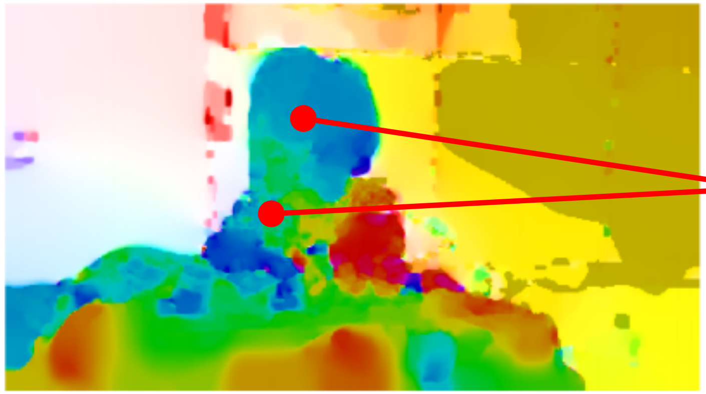
☹ **Motion is ambiguous.**

(a) Input Image (b) Prediction (c) Ground Truth

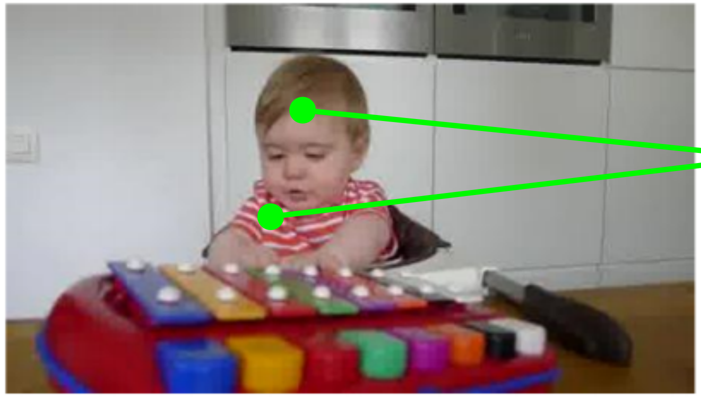
Motion prediction from static images. [1]

[1] Walker J, Gupta A, Hebert M. "Dense optical flow prediction from a static image." *In CVPR*, 2015.

Learning from Motion Consistency Priors



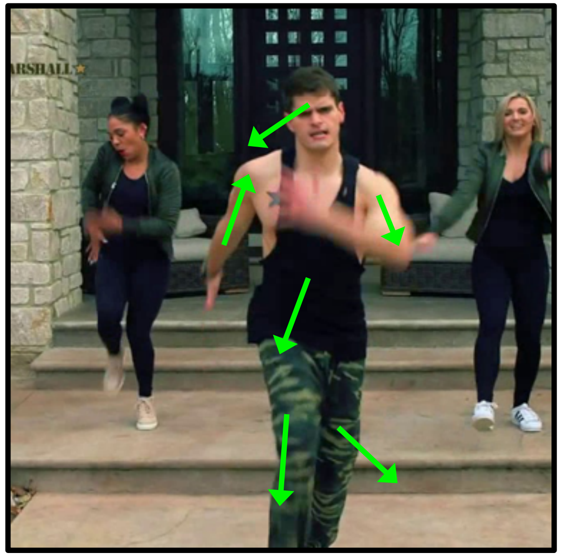
Motion is similar



Pixels on the
same object

Learn from motion consistency [2]

Motion is complicated.



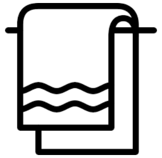
Some objects have high degrees of
freedom, e.g., human.

[2] Mahendran A, Thewlis J, Vedaldi A. Cross pixel optical-flow similarity for self-supervised learning. In ACCV, 2018.

Learning from Kinematic (运动学) Priors

- rigid

- articulated



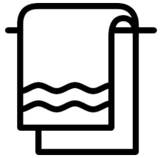
- deformable

Kinematic Priors → affect → Full Motion

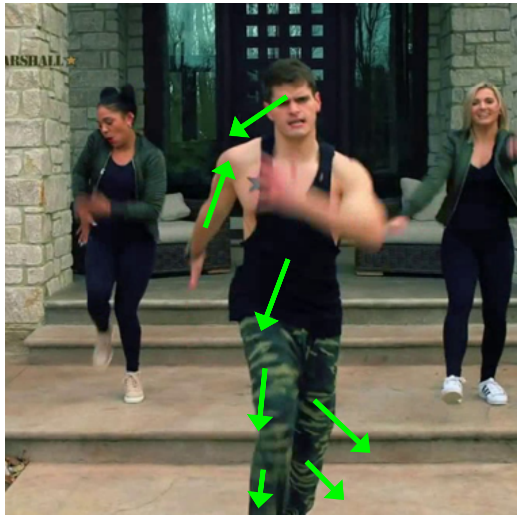
Learning from Kinematic Priors

- rigid

- articulated



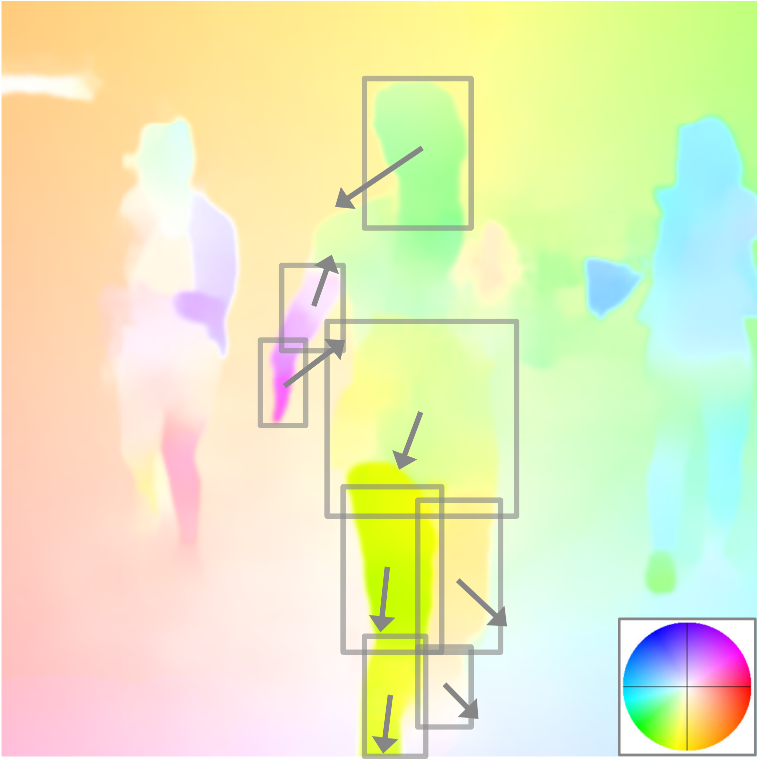
- deformable



Kinematic Priors + Motion Tendency ⇒ Full Motion

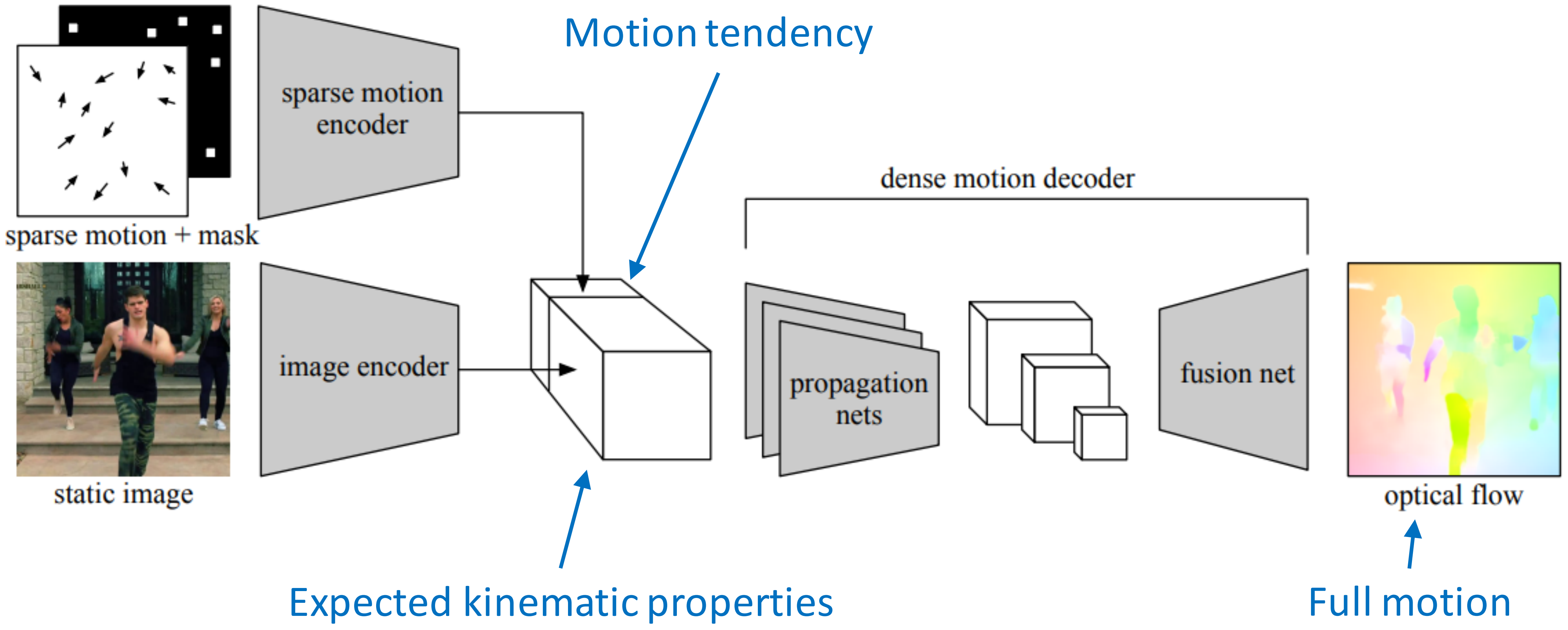
? ✓ ✓

Approximate Motion Tendency

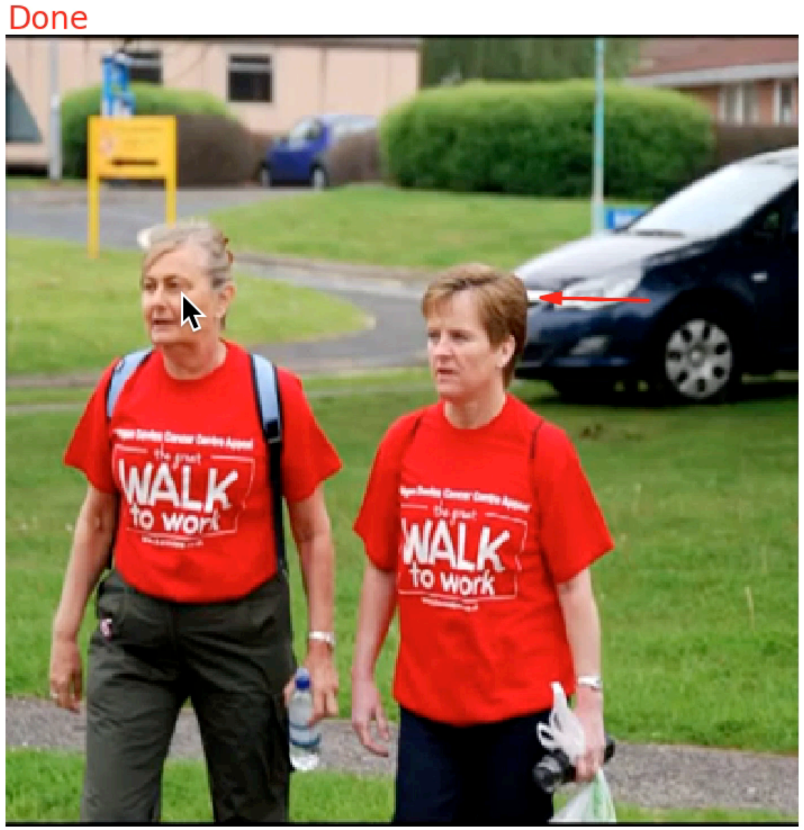


Without the annotation of rigid parts, how to approximate the motion tendency?

Conditional Motion Propagation [CVPR'19]



Conditional Motion Propagation [CVPR'19]



Application: Kinematic-Grounded Video Generation

Application: Kinematic-Grounded Video Generation

Application: Kinematic-Grounded Video Generation

For general objects

Self-Supervised Scene De-occlusion [CVPR'20 Oral]

Real-world scene

Intact objects with invisible parts
+ ordering graph

Background

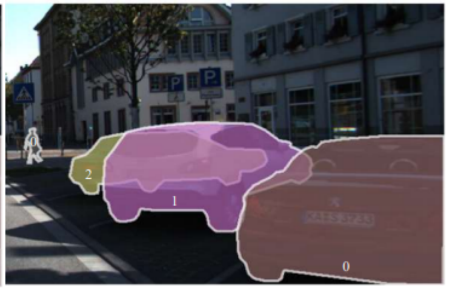
What We Have

- A typical instance segmentation dataset:

RGB image

Modal masks & Category labels

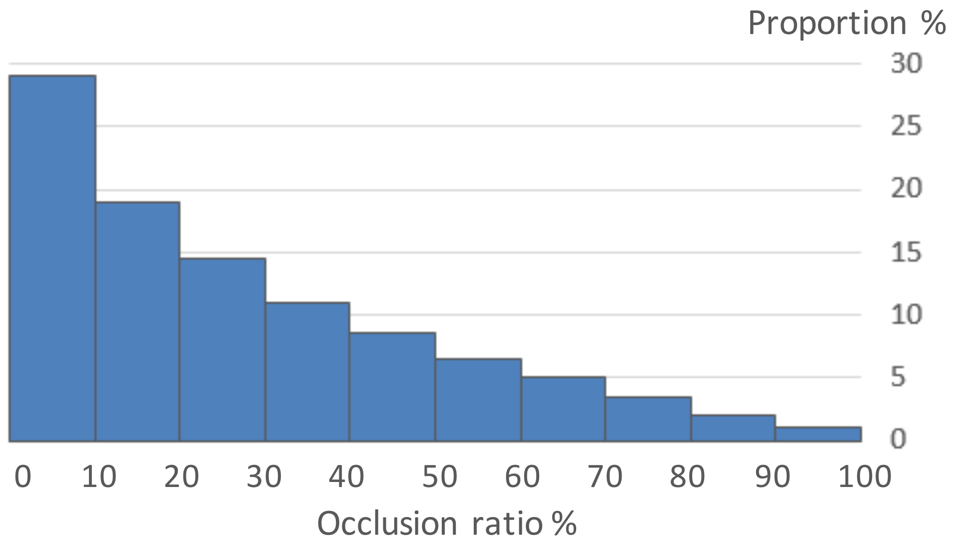
Modal mask

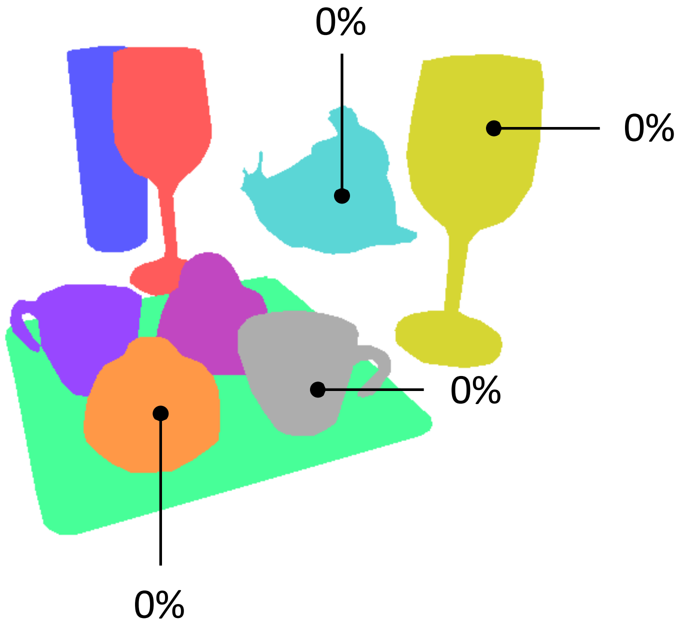


Amodal mask

Modality	Available
Image	√
Modal mask	√
Ordering	×
Amodal Mask	×

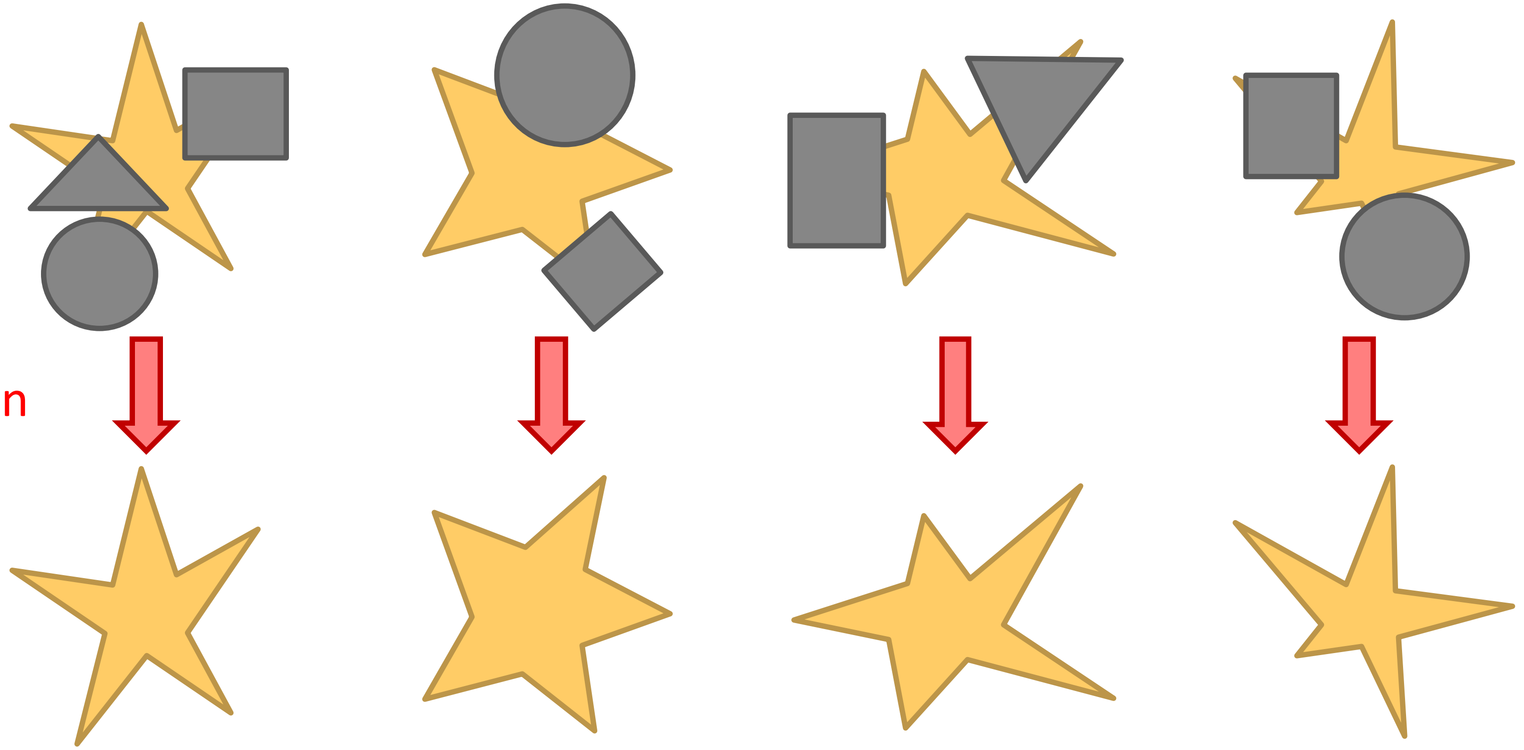
Data Analysis: Occlusion Ratio





Amodal Completion

What's the shape of the yellow objects?

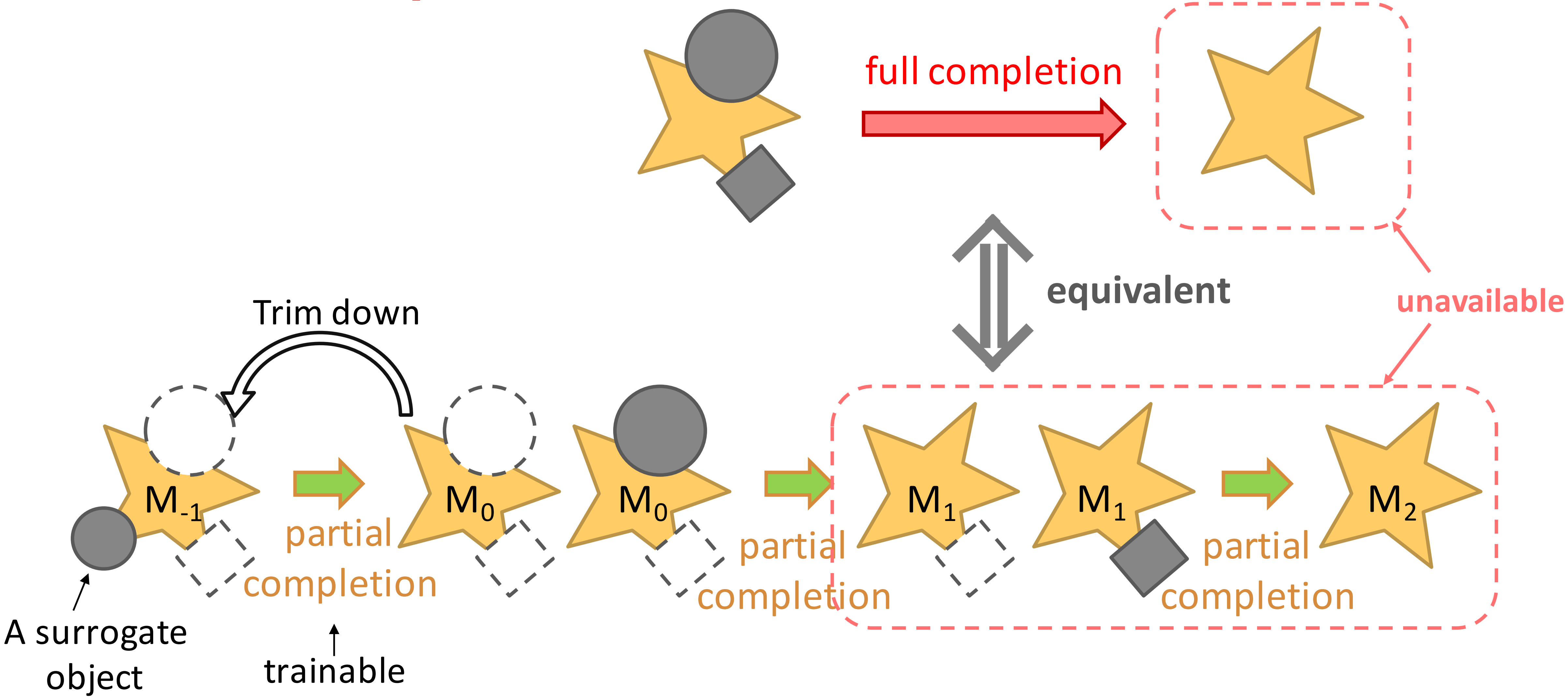


Full completion

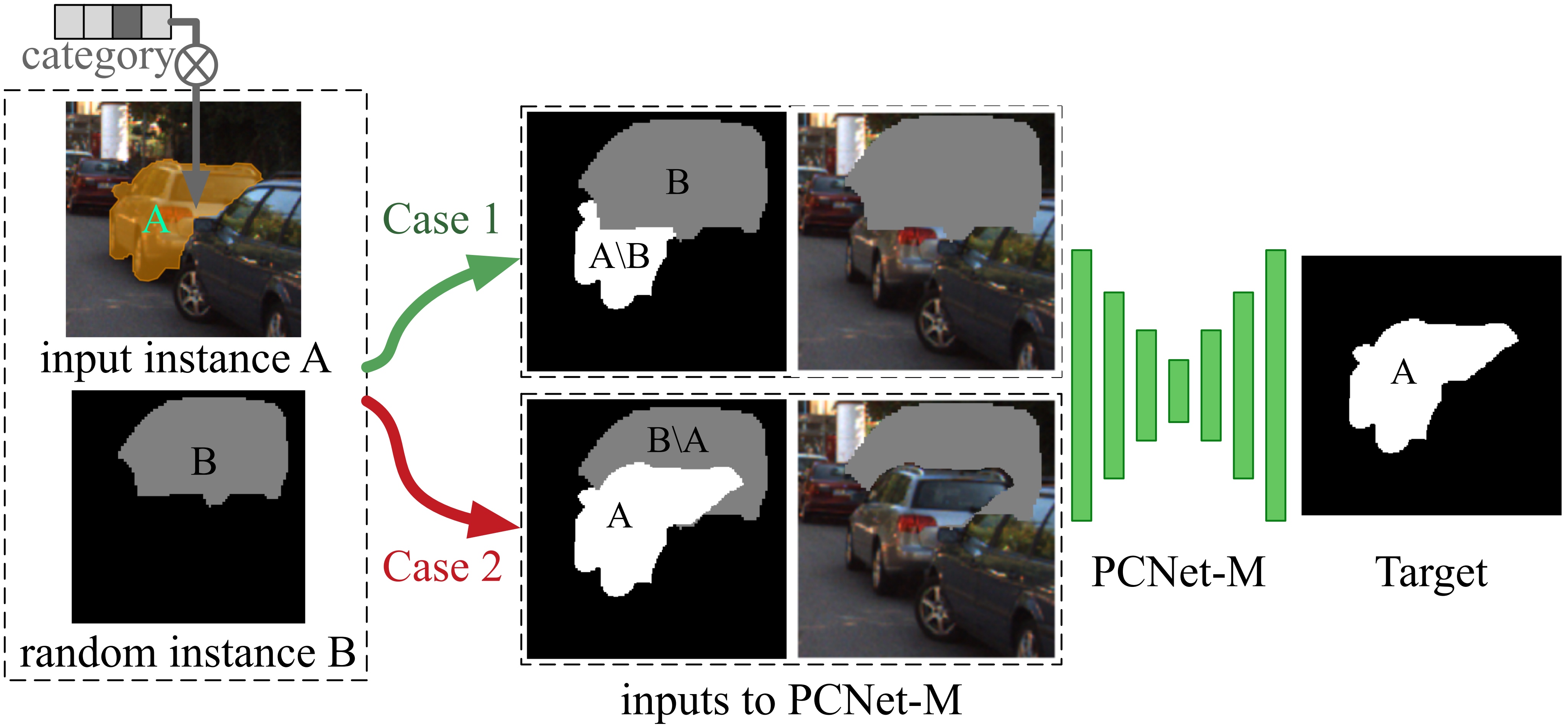
Ground truth amodal masks as supervision

What if we do not have the ground truth?

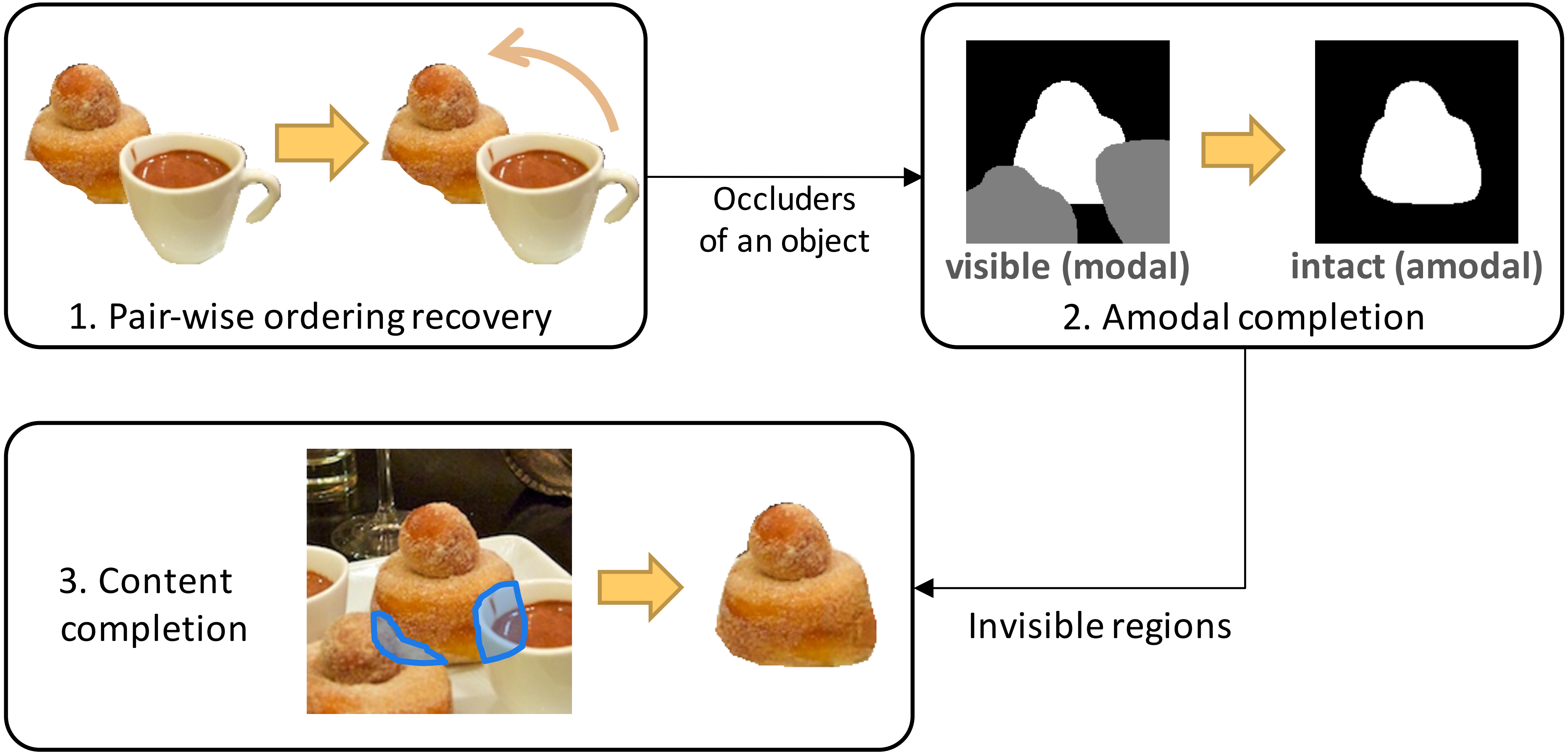
Partial Completion



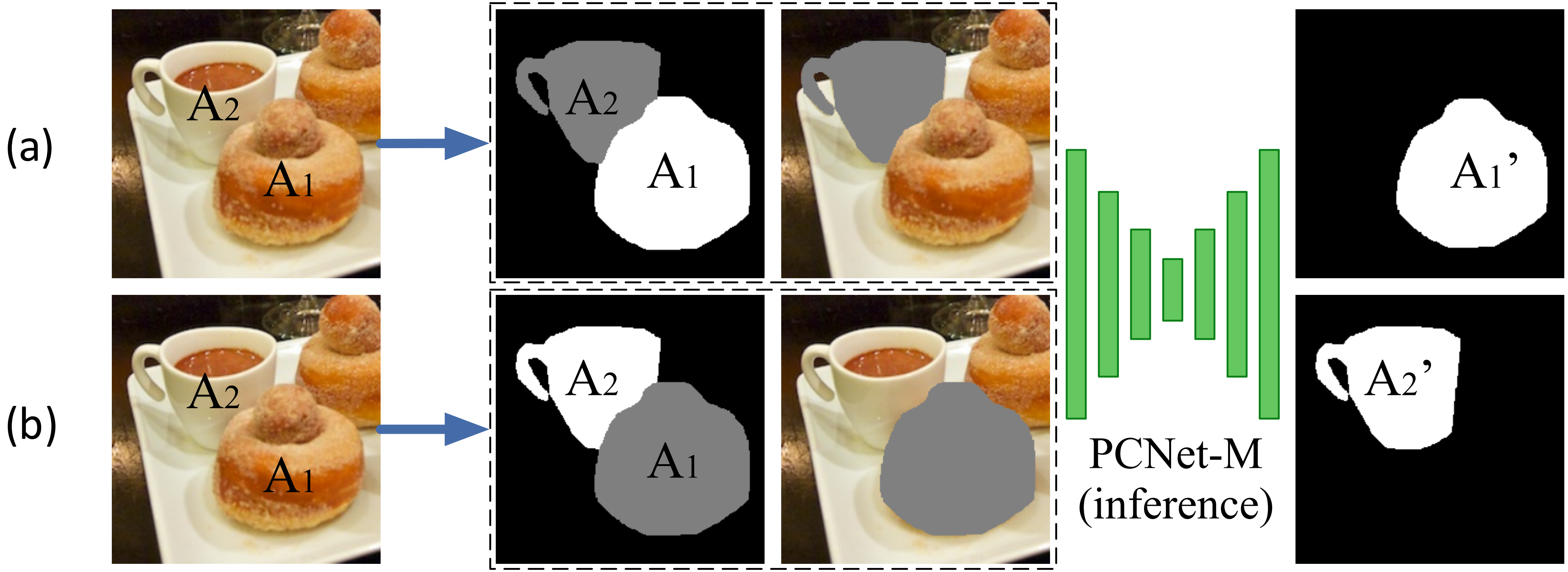
Train Partial Completion Net-Mask (PCNet-M)



Tasks to Solve



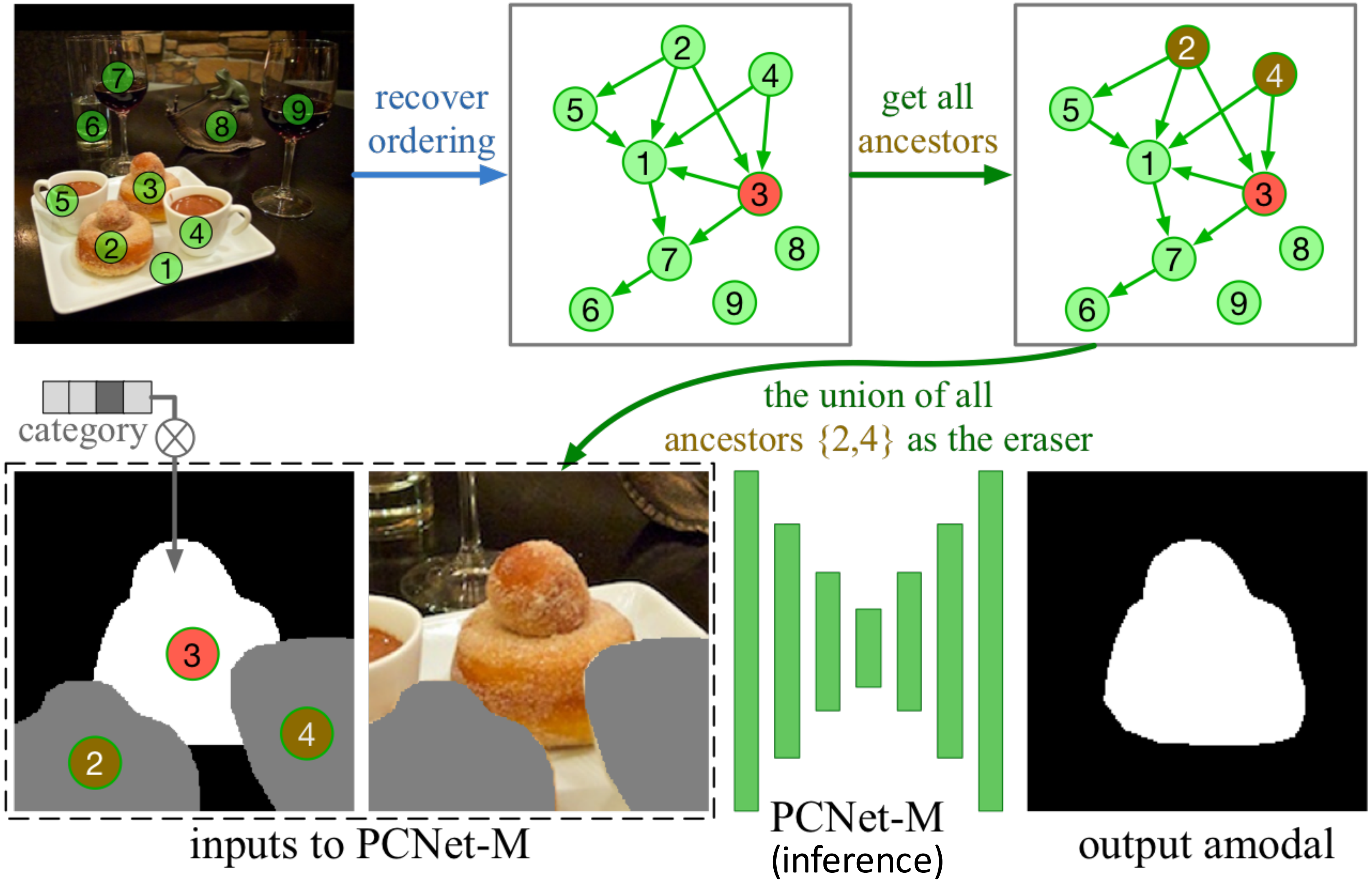
Dual-Completion for Ordering Recovery



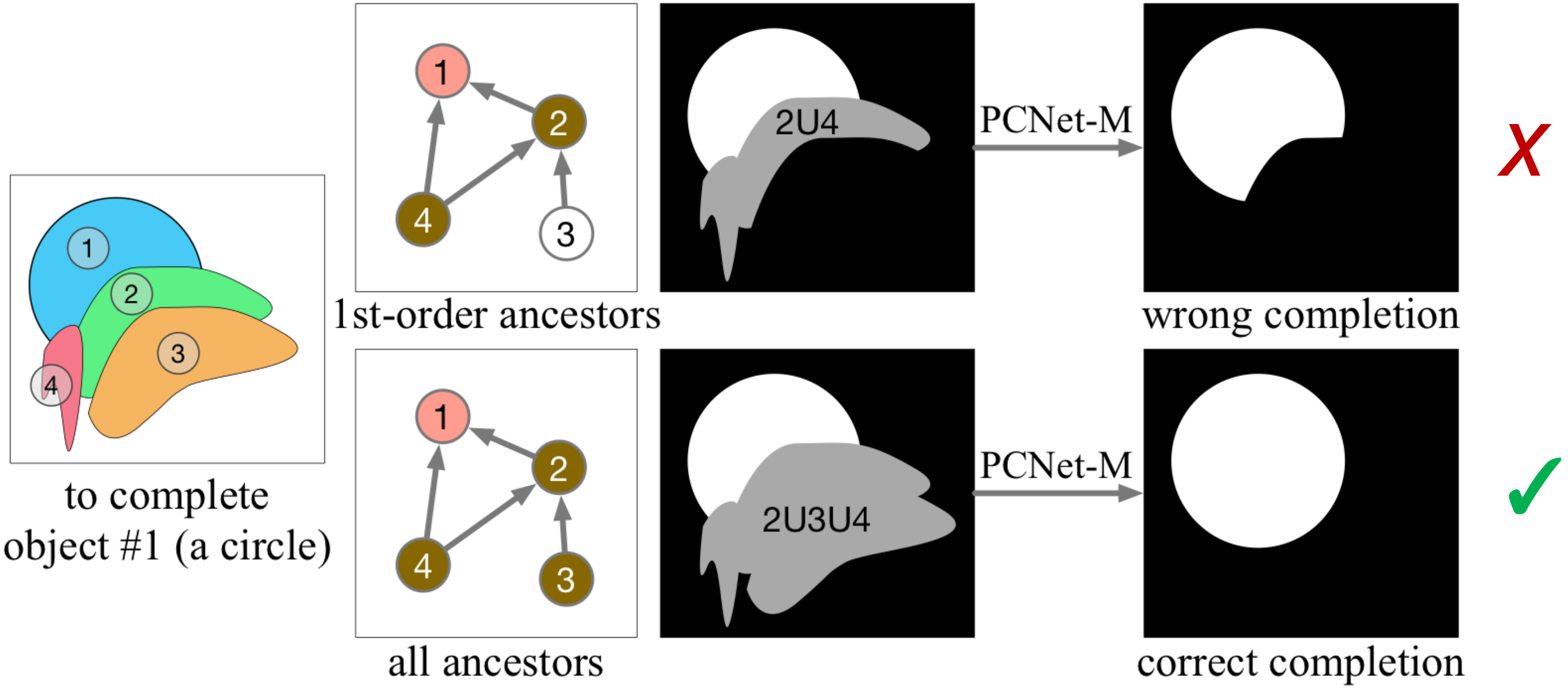
(a) Regarding A1 as the target and A2 as the surrogate occluder, the incremental area of A1: $\Delta A_1' | A_2$
(b) Regarding A2 as the target and A1 as the surrogate occluder, the incremental area of A2: $\Delta A_2' | A_1$

Decision: $\Delta A_1' | A_2 < \Delta A_2' | A_1 \Rightarrow$ A1 is above A2

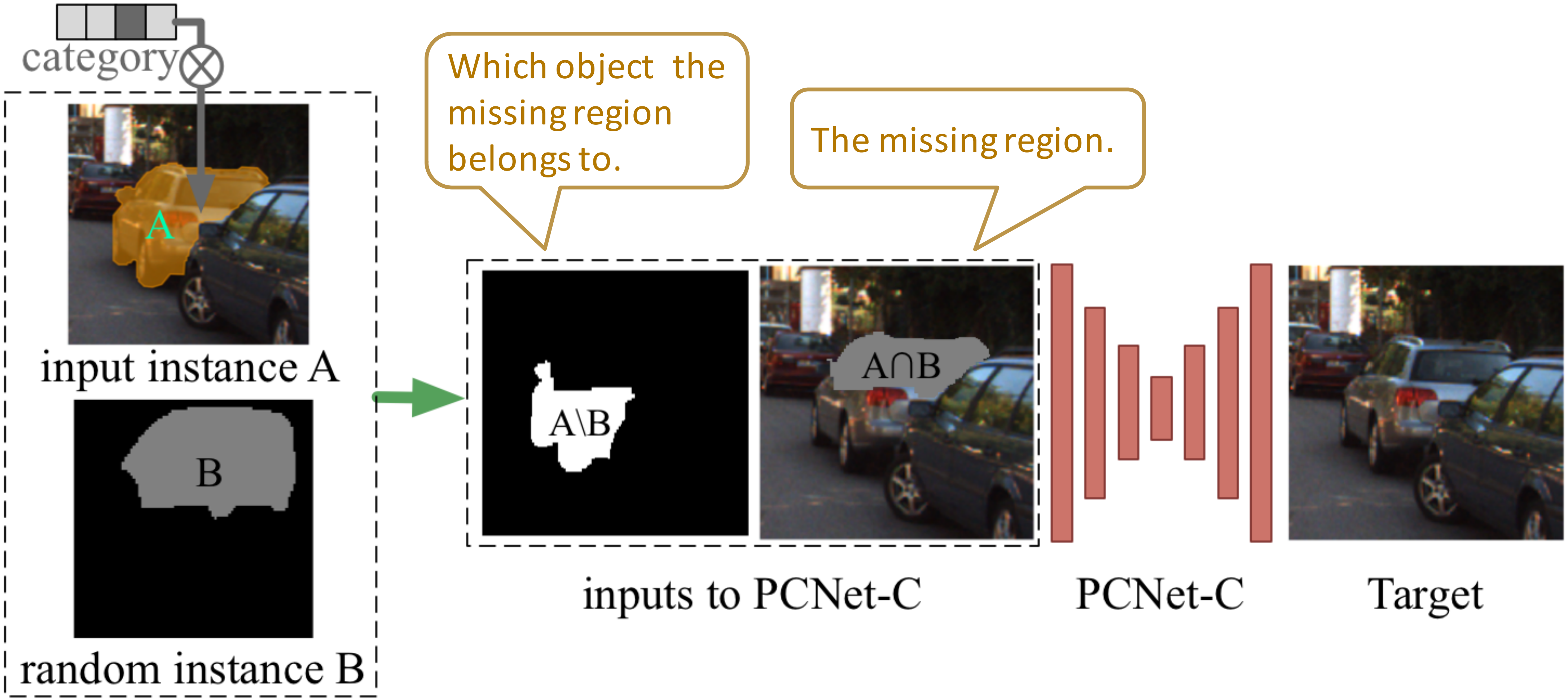
Ordering-Grounded Amodal Completion



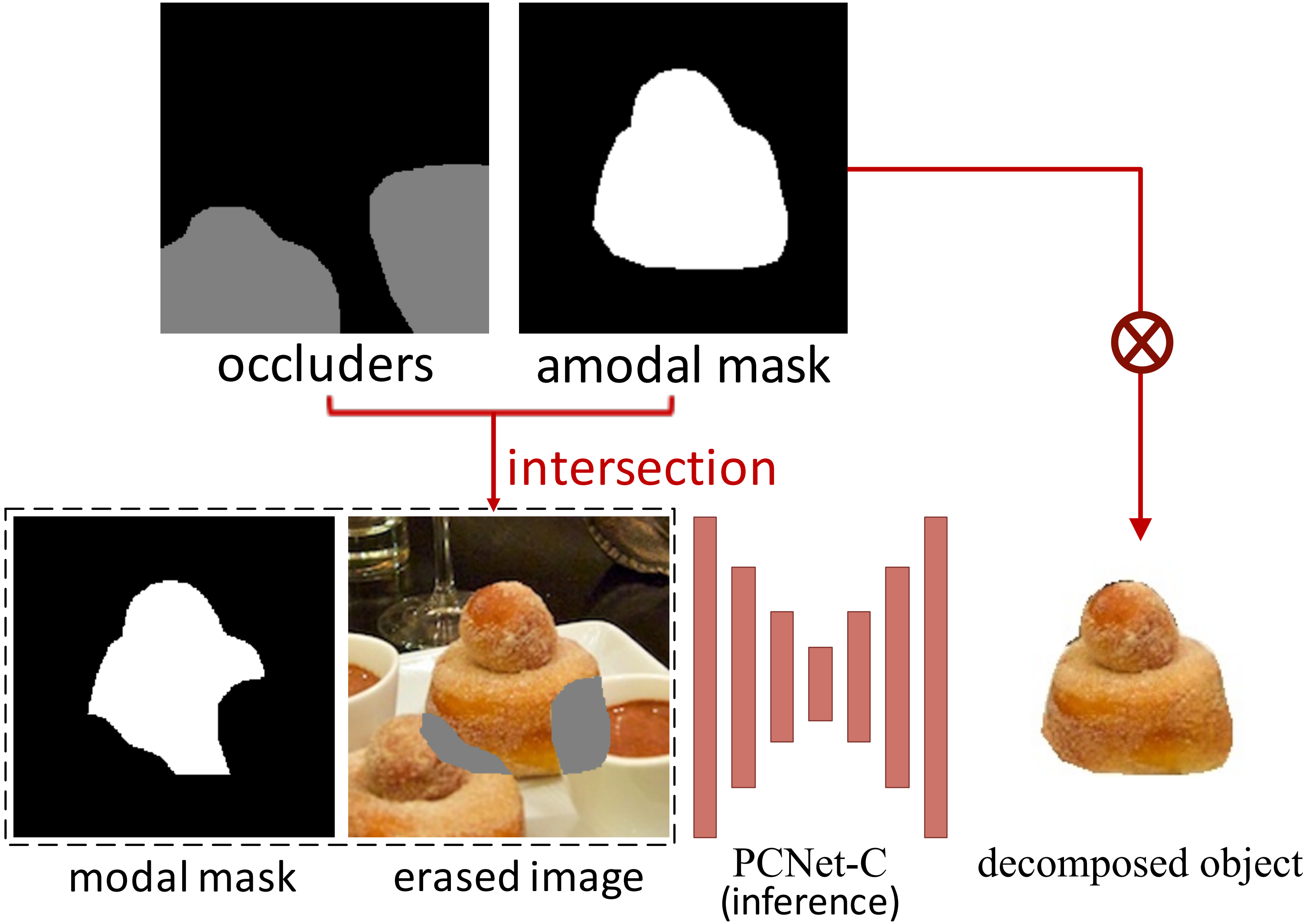
Why All Ancestors?



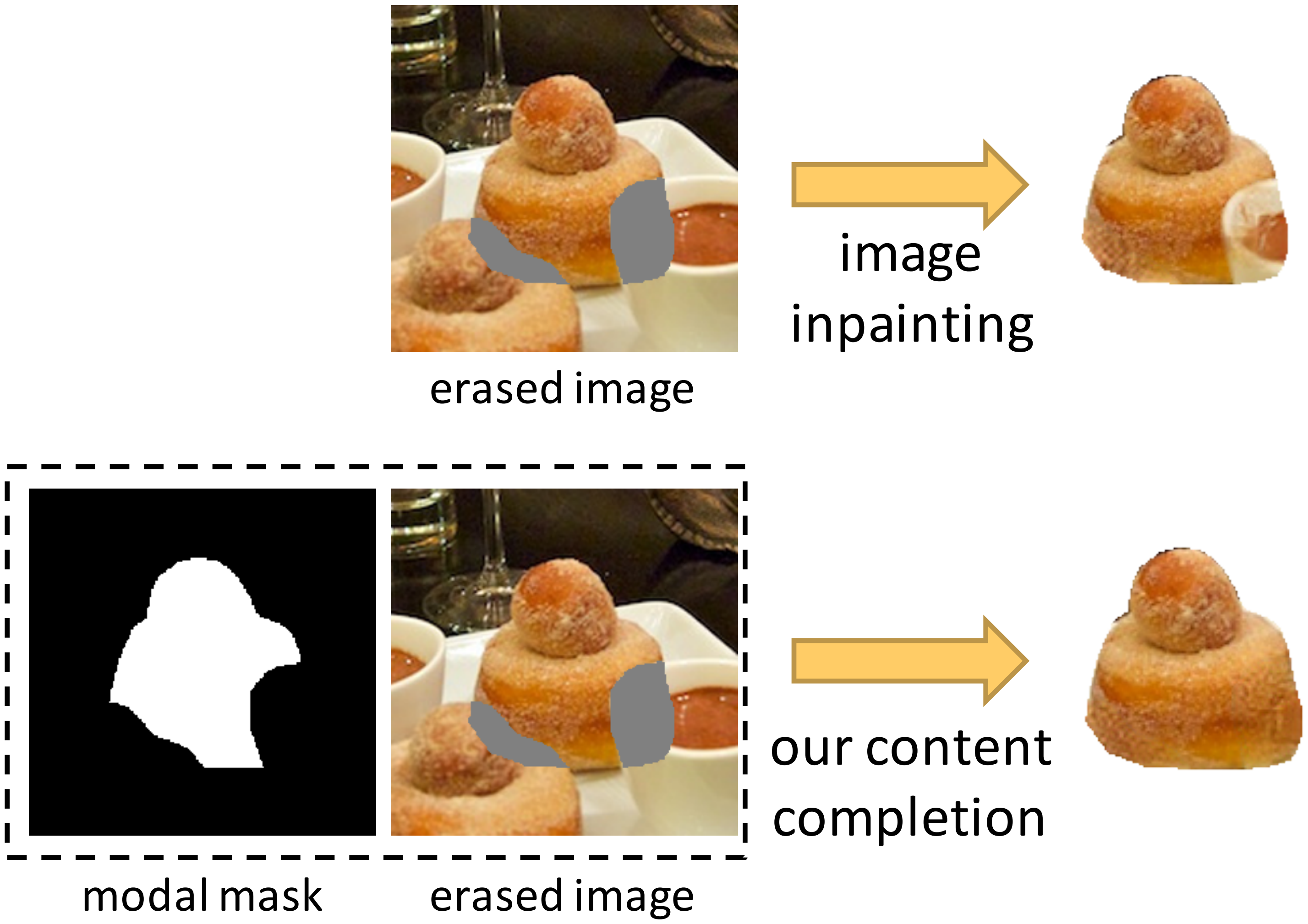
Train Partial Completion Net-Content (PCNet-C)



Amodal-Constrained Content Completion



Compared to Image Inpainting



Scene De-occlusion

Real-world scene

Objects with invisible parts
+ ordering graph

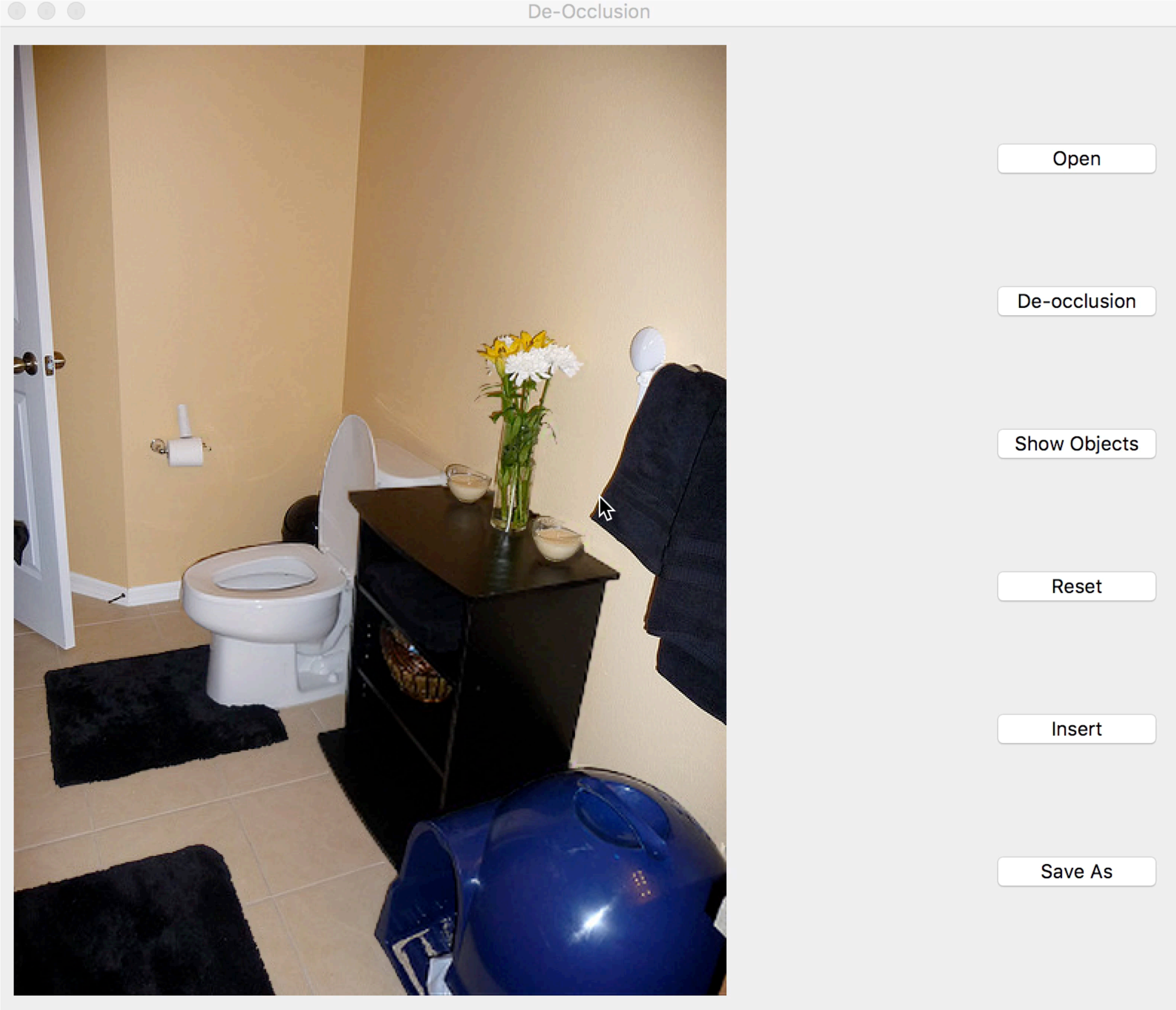
Background

Application: Image Manipulation

delete swap shift reposition

modal-based manipulation

before

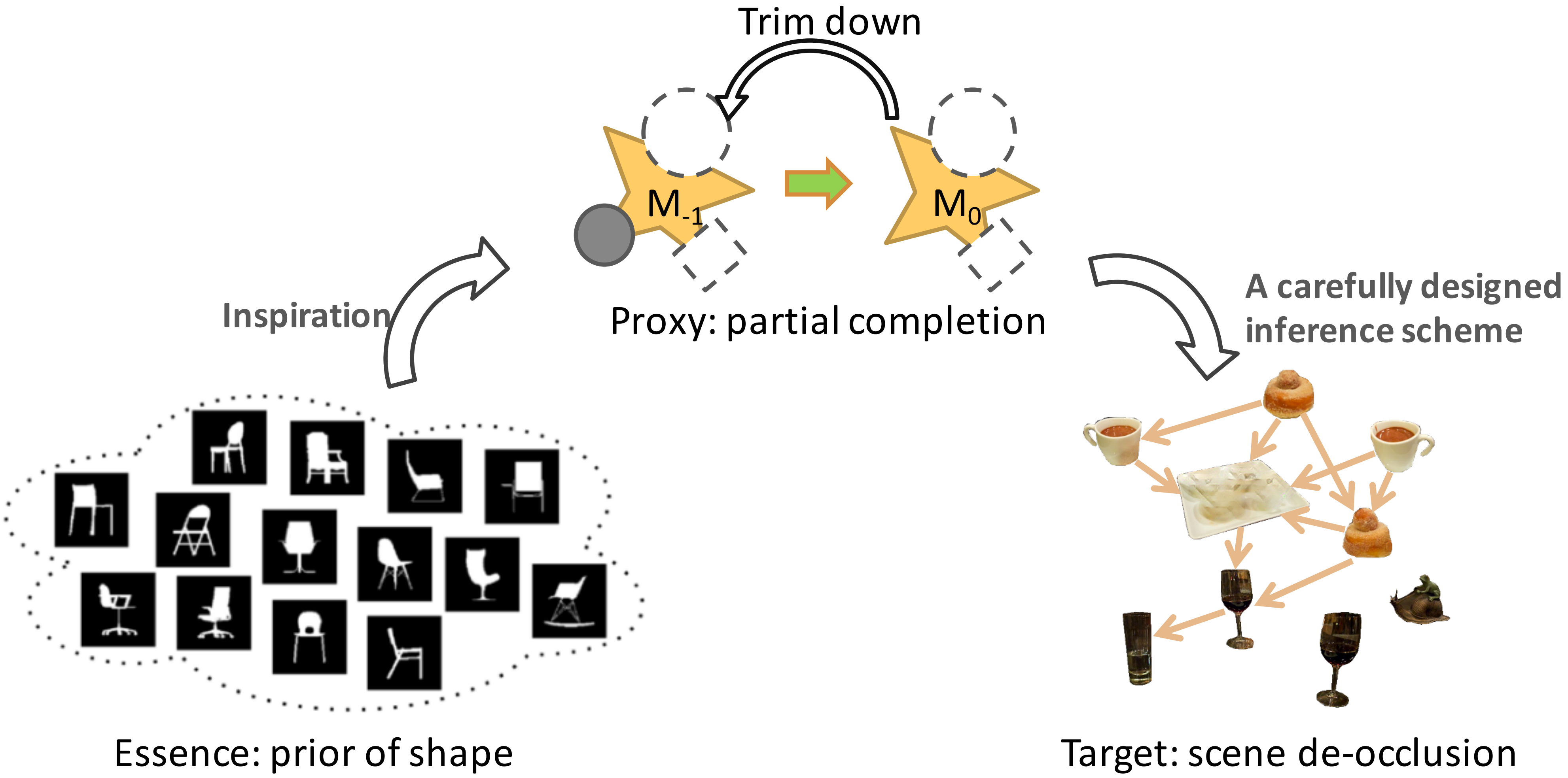

De-Occlusion
Open
De-occlusion
Show Objects
Reset
Insert
Save As

Future Directions with Scene De-occlusion

- Data augmentation / re-composition for instance segmentation.
- Ordering prediction for mask fusion in panoptic segmentation.
- Occlusion-aware augmented reality.

No need for extra annotations!

What's the Intrinsic Methodology?



总结

1. 我们的世界不是像素点的简单组合，而是在严格的物理、数学、化学、生物学、乃至社会学规则/常识下运行的；
2. 我们观测到的数据是这些规则的外在反映，本身就是有结构、可推理、可循因溯果的；
3. 从大量观测数据中归纳反推常识，不一定需要人工标注的显式监督。

Discussion

1. 还有哪些应用场景需要利用无标注数据？
   - 数据量大：自然语言处理、图片视频分类、行人车辆监控、……
   - 标注困难：遥感图像、语义分割、深度估计、……
2. 半监督学习和聚类的区别是？利用人脸无标注数据的时候为何要用聚类，而非半监督？
   - 半监督：类别固定；聚类：类别（identity）是开放集合。
3. 举例graph形式的数据有哪些？用graph相比单点数据的好处是？
   - 社交网络数据、人体骨架、图像中物体关系、电影中人物关系、分子化学结构、……
   - 可表示更多类型的数据、信息更丰富。
4. 无标注图片和视频，除了颜色、纹理、朝向，还可以利用哪些prior来进行自监督学习？
   - 对称性（人脸、动物脸、车辆等）、视频和声音的关联性、……

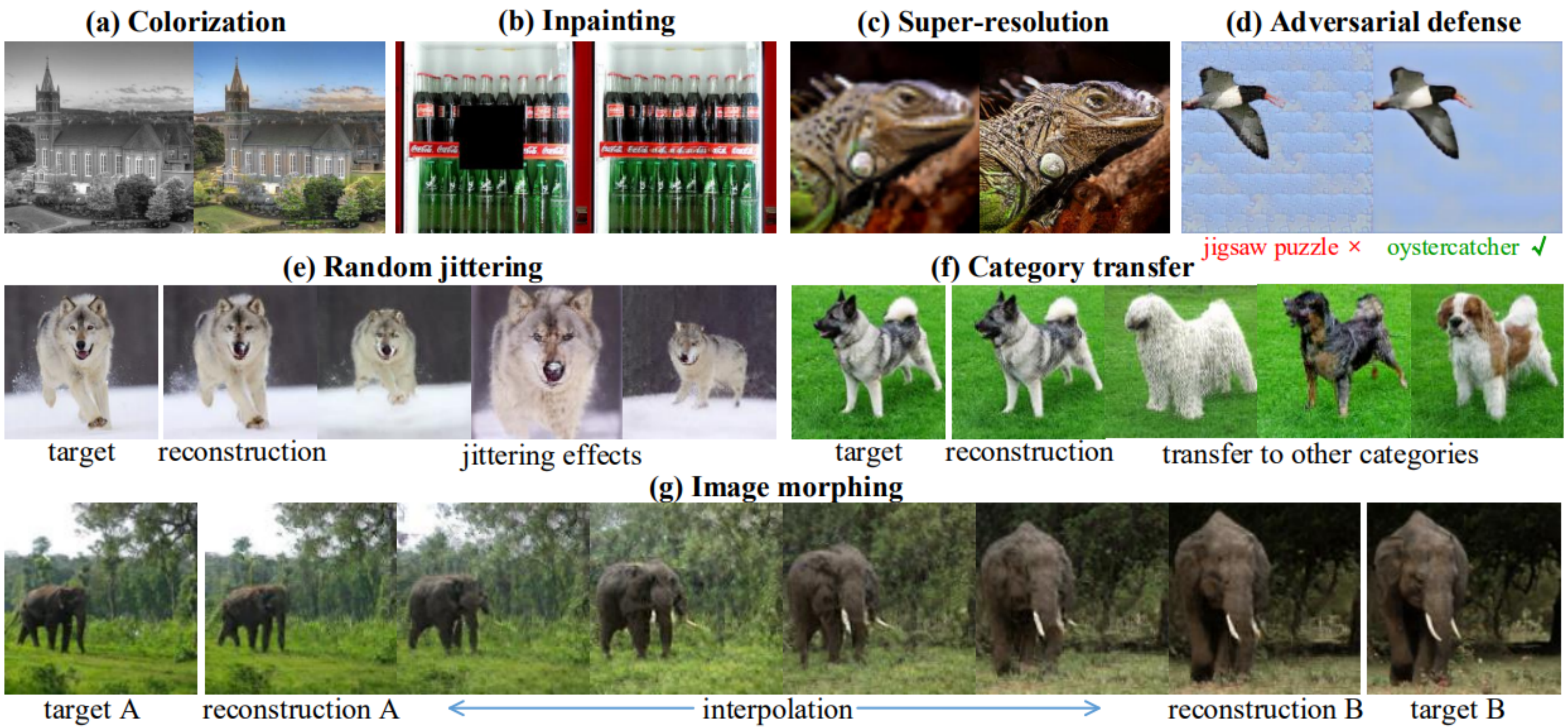
Discussion

Done

利用Conditional motion propagation 还可以设计什么应用？

Application: Interactive Segmentation

Extensions



Deep Generative Prior [ECCV'20 Oral]

More: xiaohangzhan.github.io