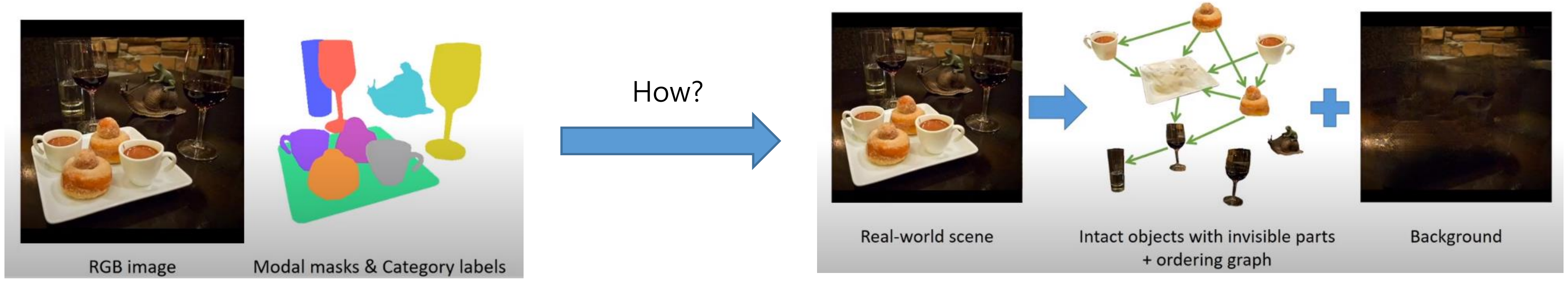
Self-Supervised Scene De-occlusion

CVPR20 oral

https://xiaohangzhan.github.io/projects/deocclusion

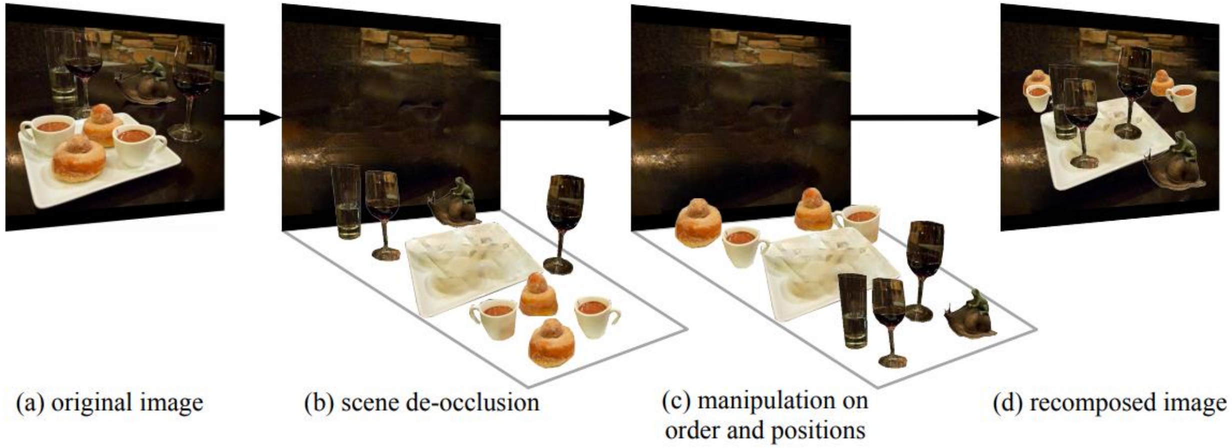
Scene de-occlusion

- Decomposes a real world scene into intact objects with invisible parts and the background



Instance segmentation dataset

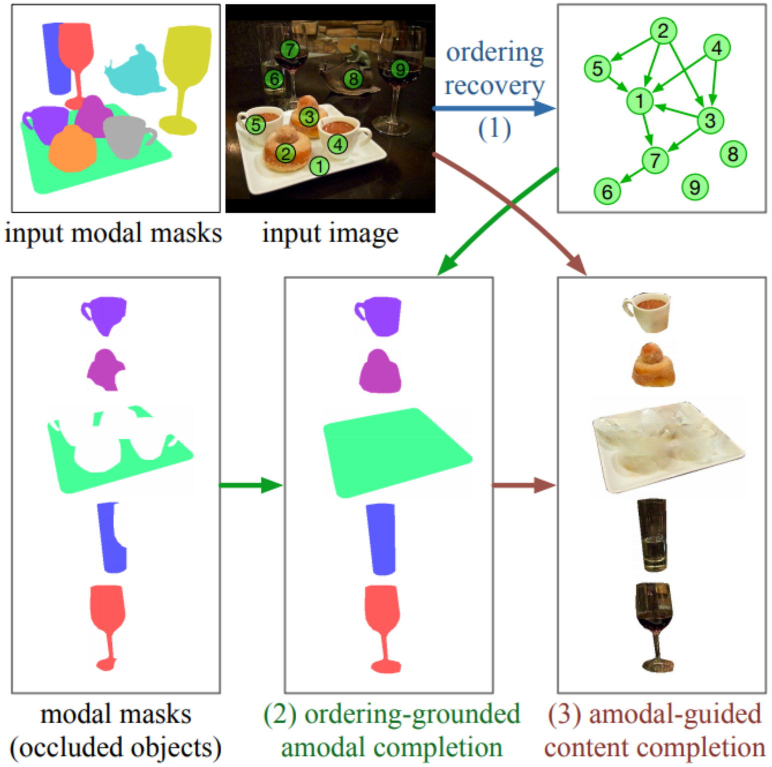
Scene de-occlusion



Application to scene recomposition

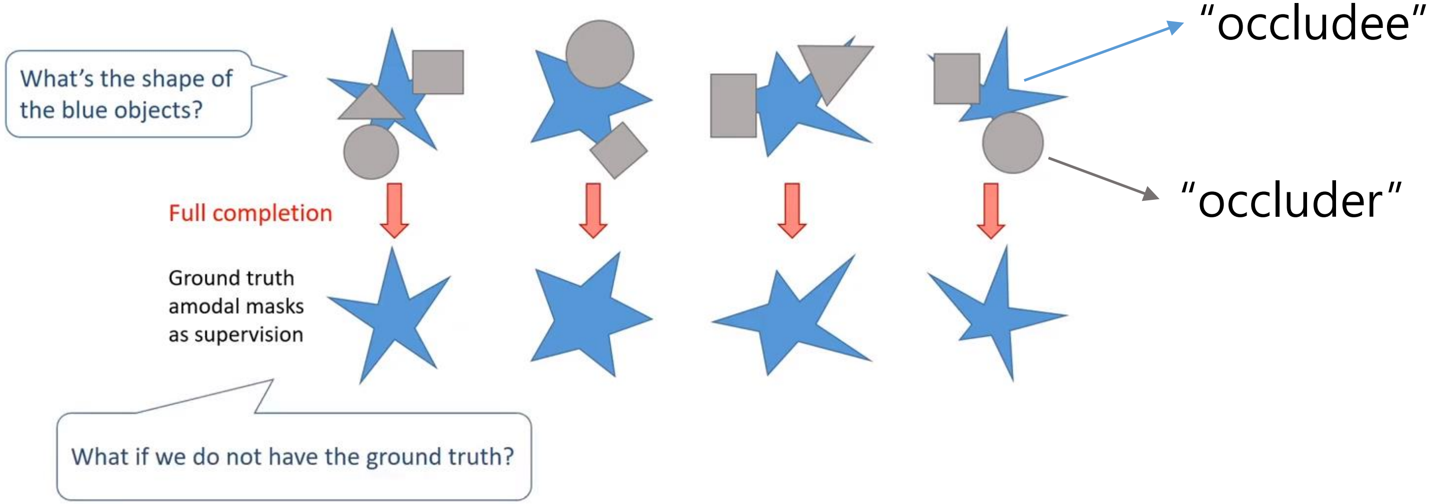
Framework

- Given a scene and its corresponding modal masks of objects as inputs,
  - 1) Recover the underlying occlusion ordering
  - 2) Complete the invisible parts of occluded objects
    - Amodal completion / Content completion

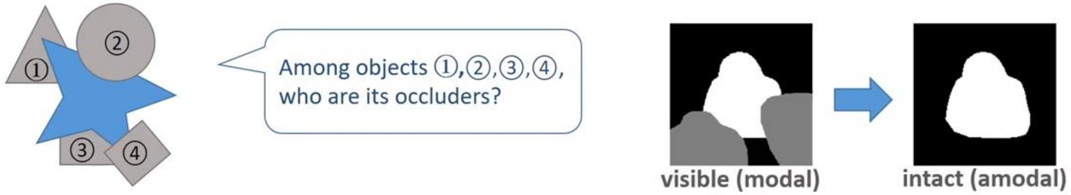


Progressive inference scheme

Amodal completion



Human can easily predict the original shape of occluded objects while it is challenging for machines.



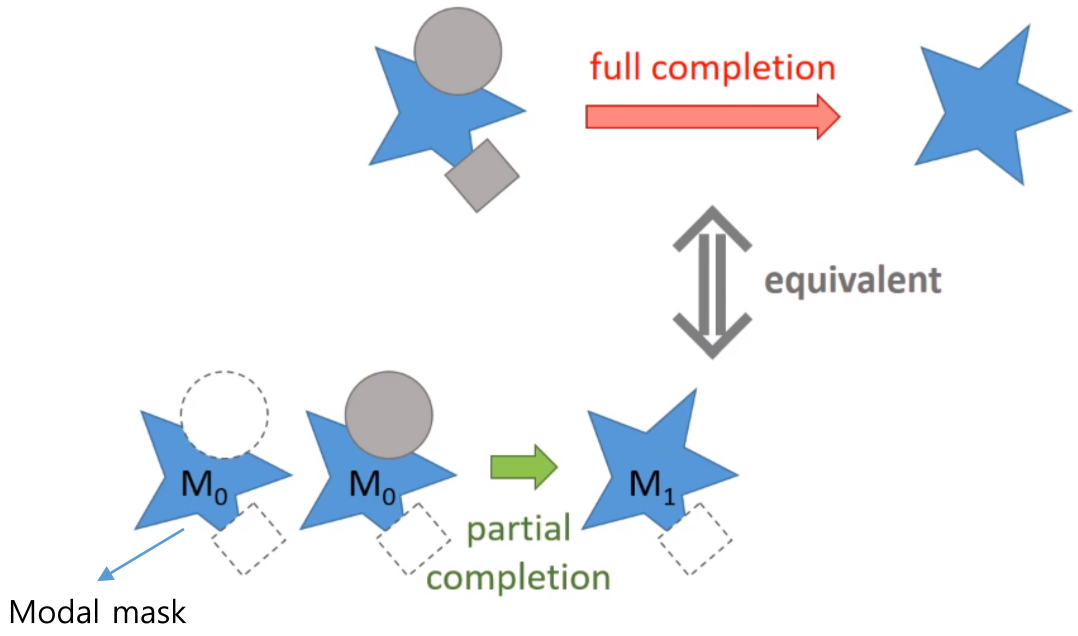
Getting Ground truth of **occlusion orderings** and **amodal masks** is laborious and costly.

A novel self-supervised framework without manual annotations in this work.

Approach

Self-supervised Partial Completion

- Amodal completion can be broken down into a sequence of partial completions.
  - Complete an occluded object progressively with one occluder involved at a time.

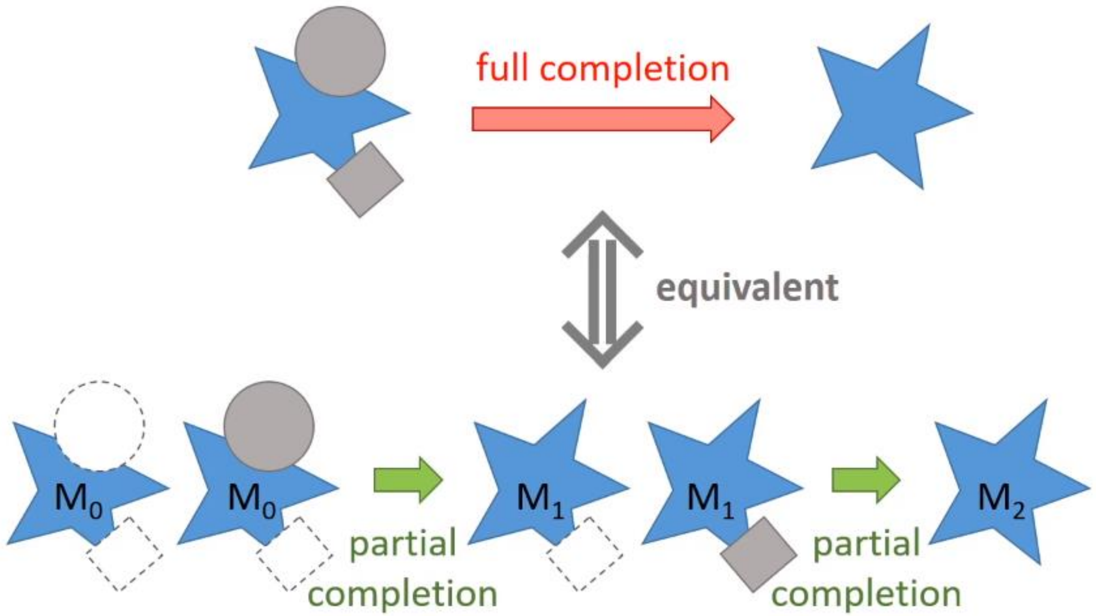


Consider one of the occluders, the gray circle at first

Perform partial completion to complete the occluded part

Self-supervised Partial Completion

- Amodal completion can be broken down into a sequence of partial completions.
  - Complete an occluded object progressively with one occluder involved at a time.

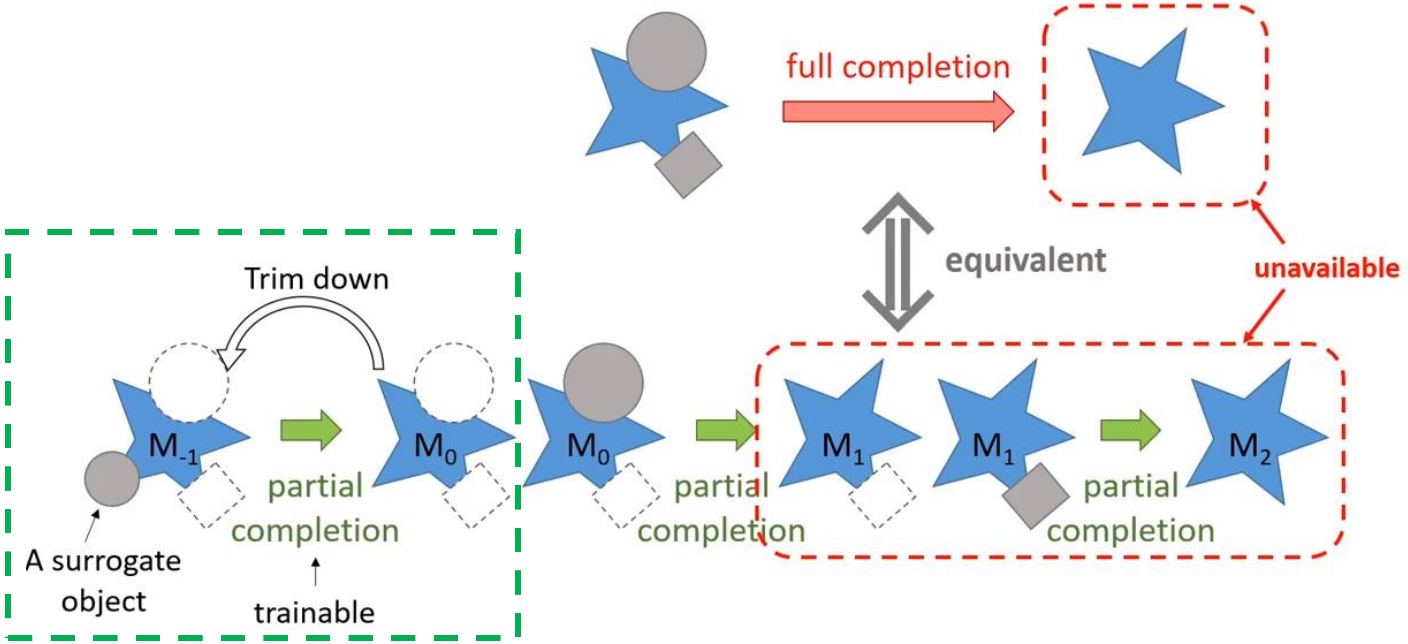


Consider the other occluder, the gray rectangle

Perform partial completion again

Self-supervised Partial Completion

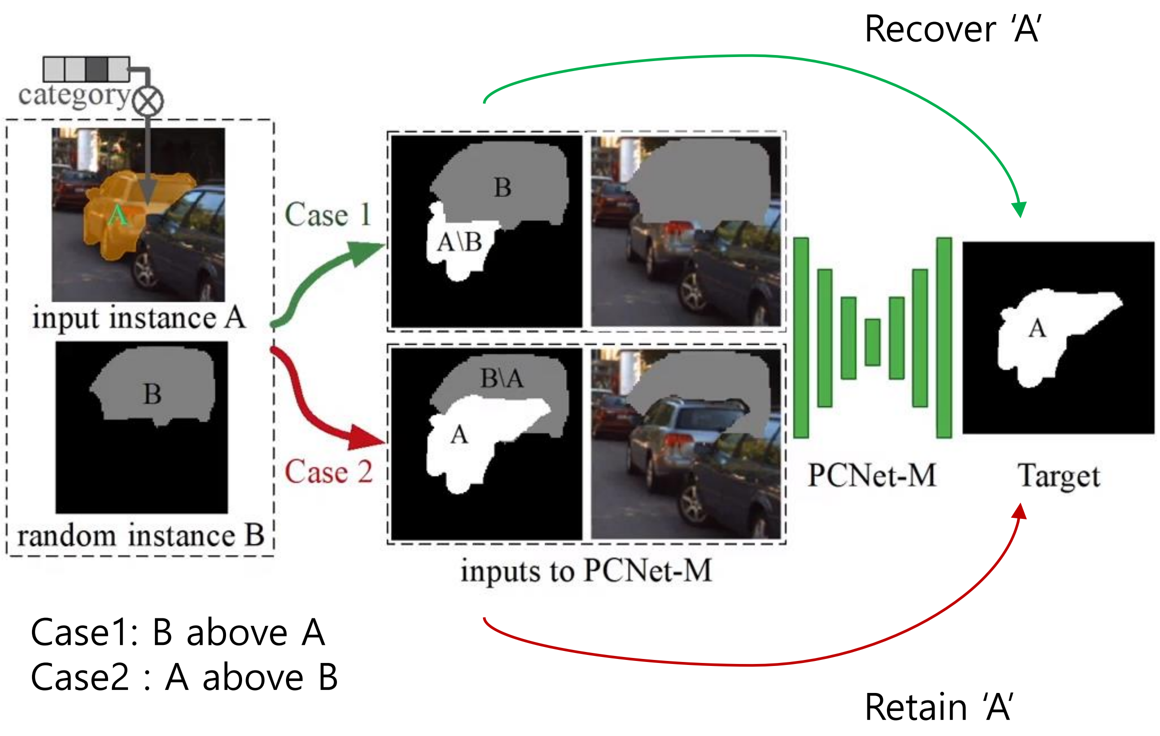
- Ground-truth M_1,M_2 to train the partial completion process is unavailable.
- Self-supervised learning strategy
  - 1) Trim down M_0 with a other random modal mask chosen from dataset to obtain M_{-1}
  - 2) Perform partial completion on M_{-1} to recover M_0



Trimming down and recovering strategy for self-supervised partial completion

PCNet-M for Mask Completion

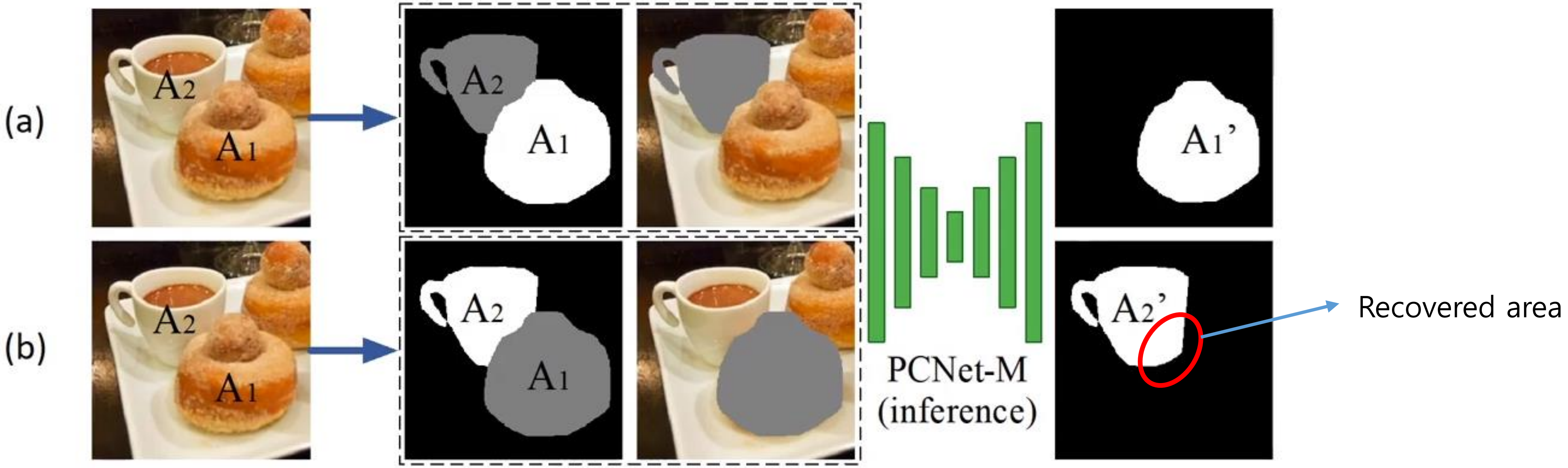
- PCNet-M is trained to partially recover the invisible mask of the occludee corresponding to an occluder.
- Two different training input cases for
  - 1) partial completion
  - 2) regularization to prevent over-completion
- PCNet-M learns to determine whether to complete or retain, and recover the occluded shape correctly.



Case1: B above A
Case2 : A above B

Ordering Recovery via Dual-Completion

- Infer the order between A_1 and A_2 by comparing their incremental area of predicted mask.
- Obtain ordering graph by performing dual-completion for all neighboring pairs.

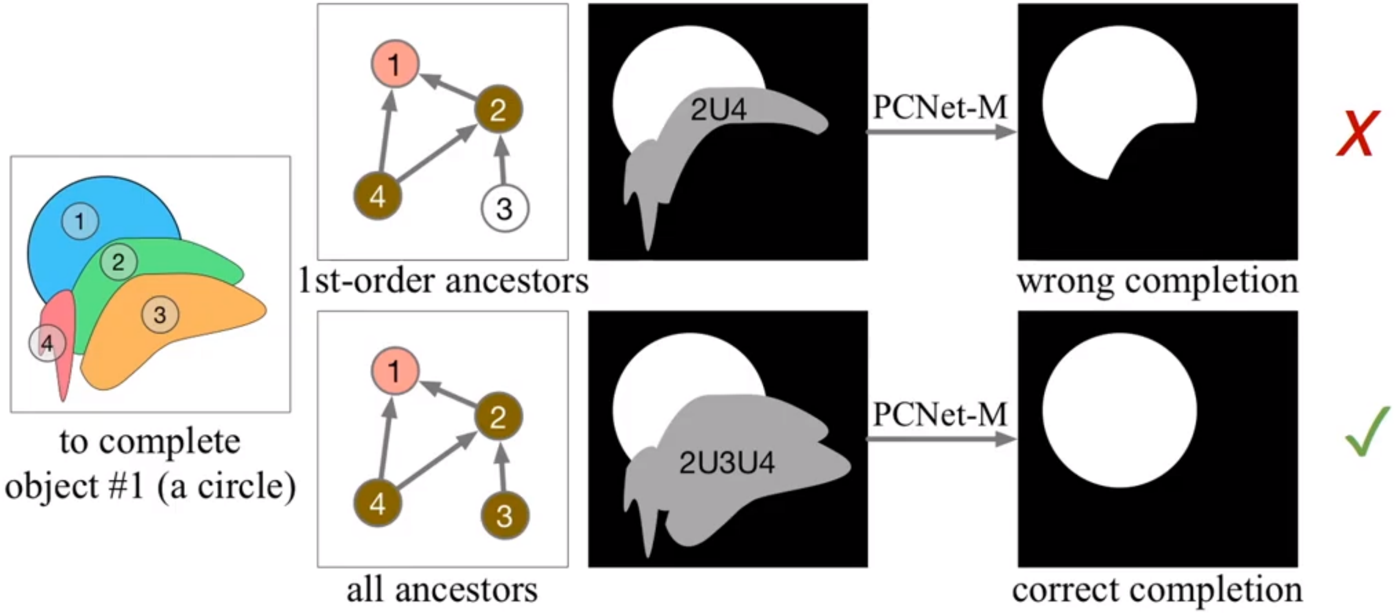


(a) Regarding A1 as the target and A2 as the surrogate occluder, the incremental area of A1: $\Delta A_1'|A_2$
(b) Regarding A2 as the target and A1 as the surrogate occluder, the incremental area of A2: $\Delta A_2'|A_1$

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

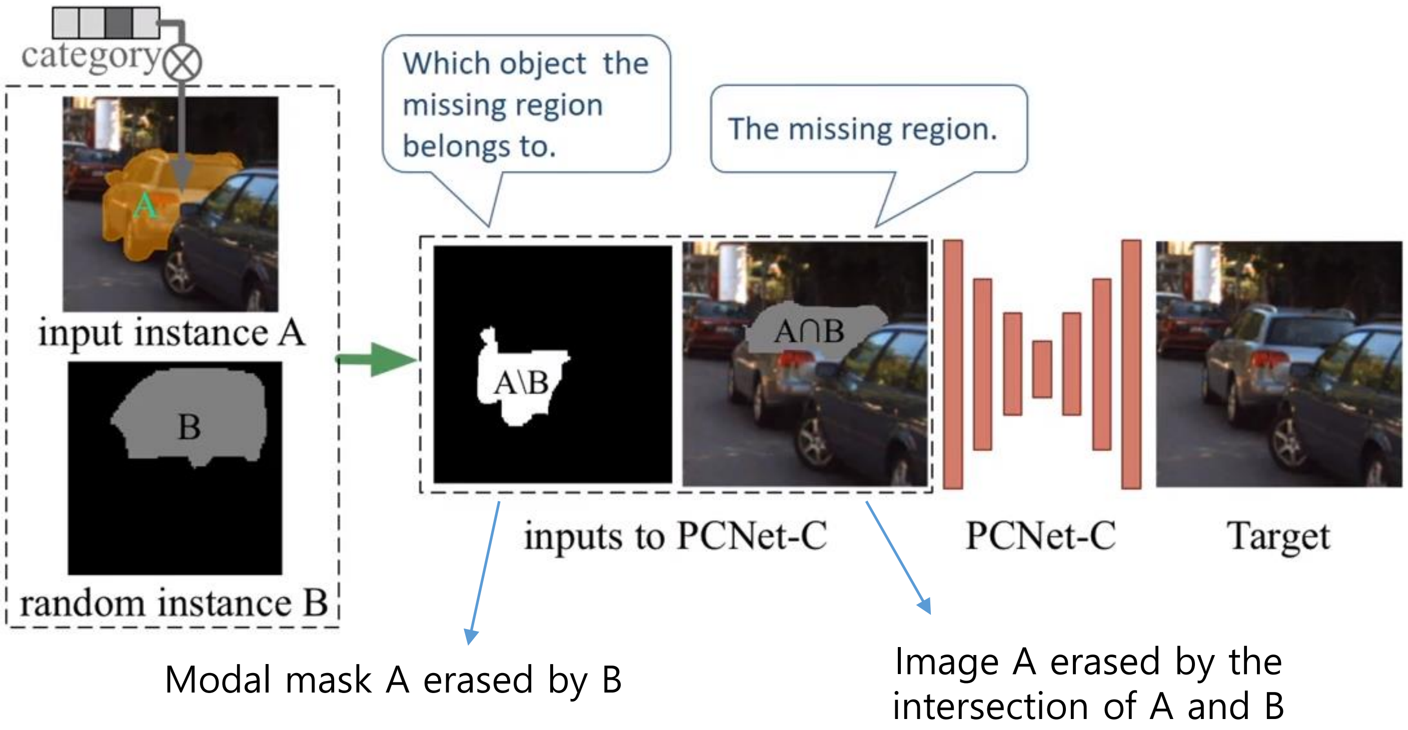
Ordering-Grounded Amodal Completion

- Find all ancestors of occludee in the graph as occluders via Breadth-First Search including higher-order!
  - May **indirectly occlude** the target instance.
  - Trained PCNet-M can perform amodal completion **in one step** conditioned on the union of all ancestors' model masks.



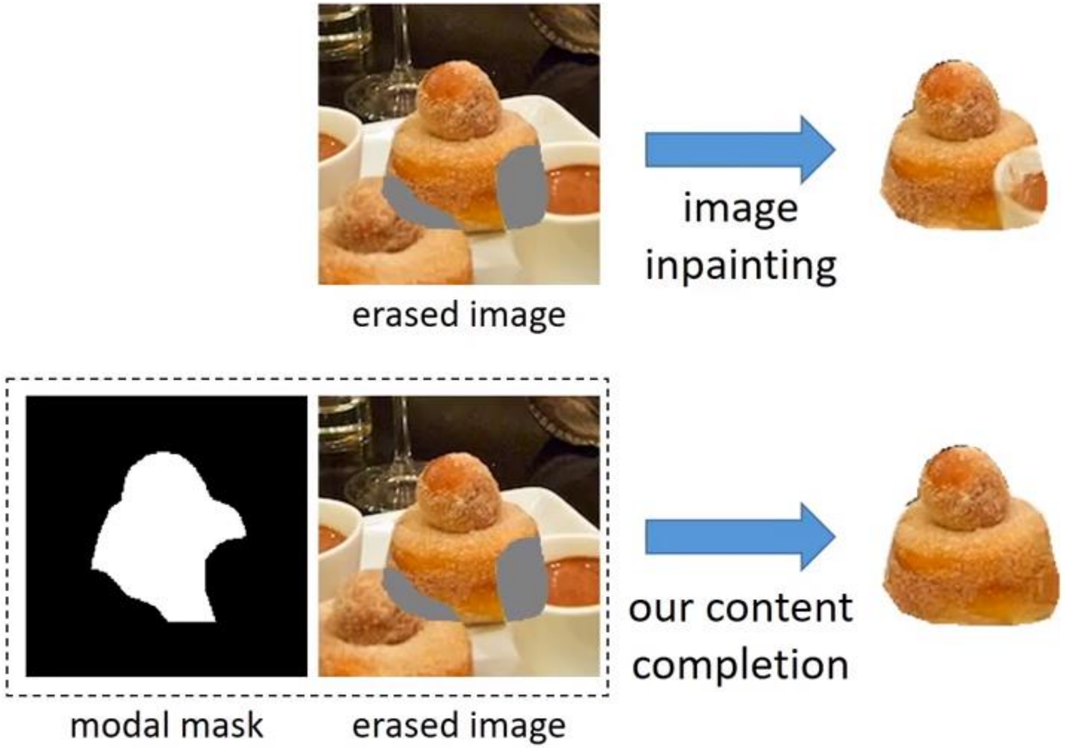
PCNet-C for Content Completion

- PCNet-C is trained to partially complete an object with RGB content given its occluders.

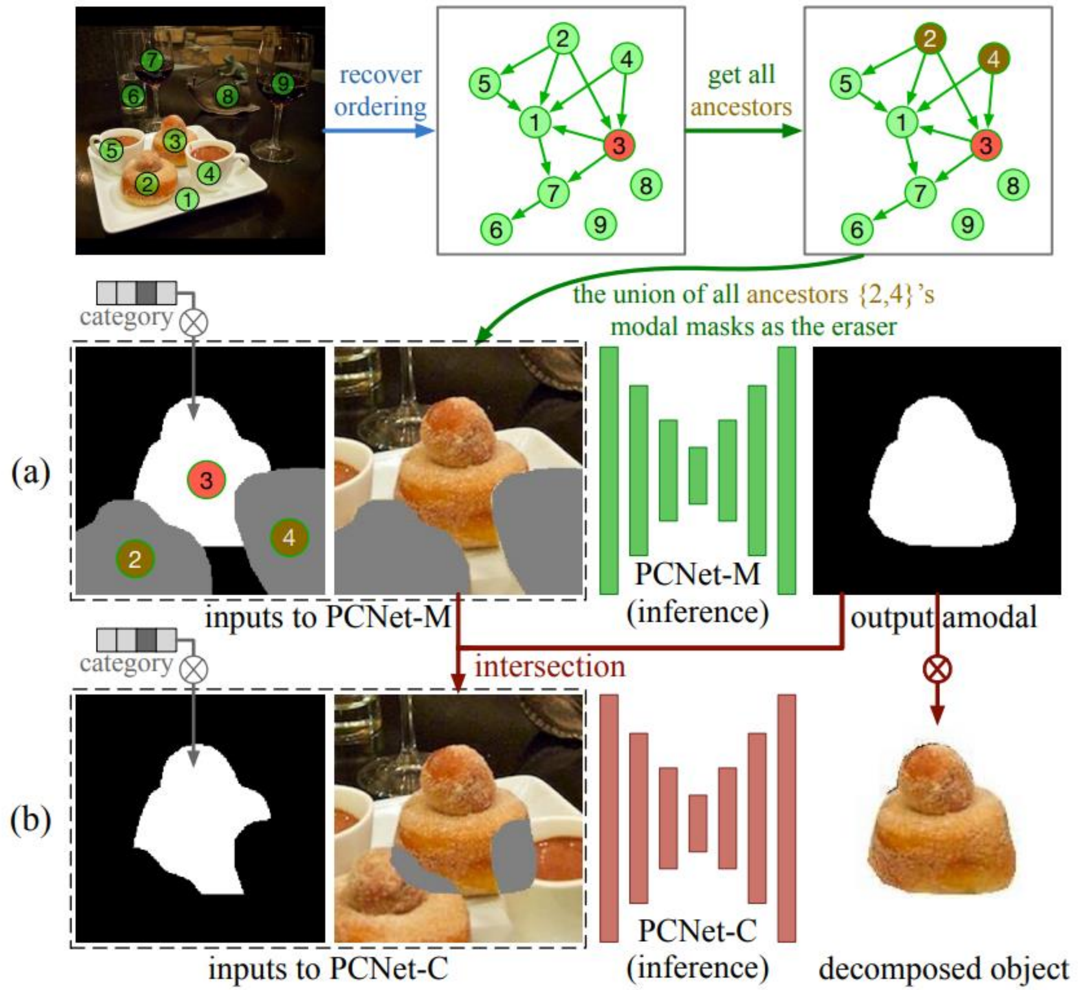


Difference between PCNet-C and image inpainting

- Image inpainting does not require the modal mask of the target object.
  - Not care about which object the missing region belongs to
- Content completion **cannot be simply replaced** by standard image inpainting!



Progressive inference scheme



Experiments

Ordering Recovery

method	gt order (train)	COCOA	KINS
Supervised			
OrderNet^{M} [17]	✔	81.7	87.5
OrderNet^{M+I} [17]	✔	88.3	94.1
Unsupervised			
Area	✘	62.4	77.4
Y-axis	✘	58.7	81.9
Convex	✘	76.0	76.3
Ours	✘	**87.1**	**92.5**

Pair-wise ordering accuracy on occluded instance pairs.

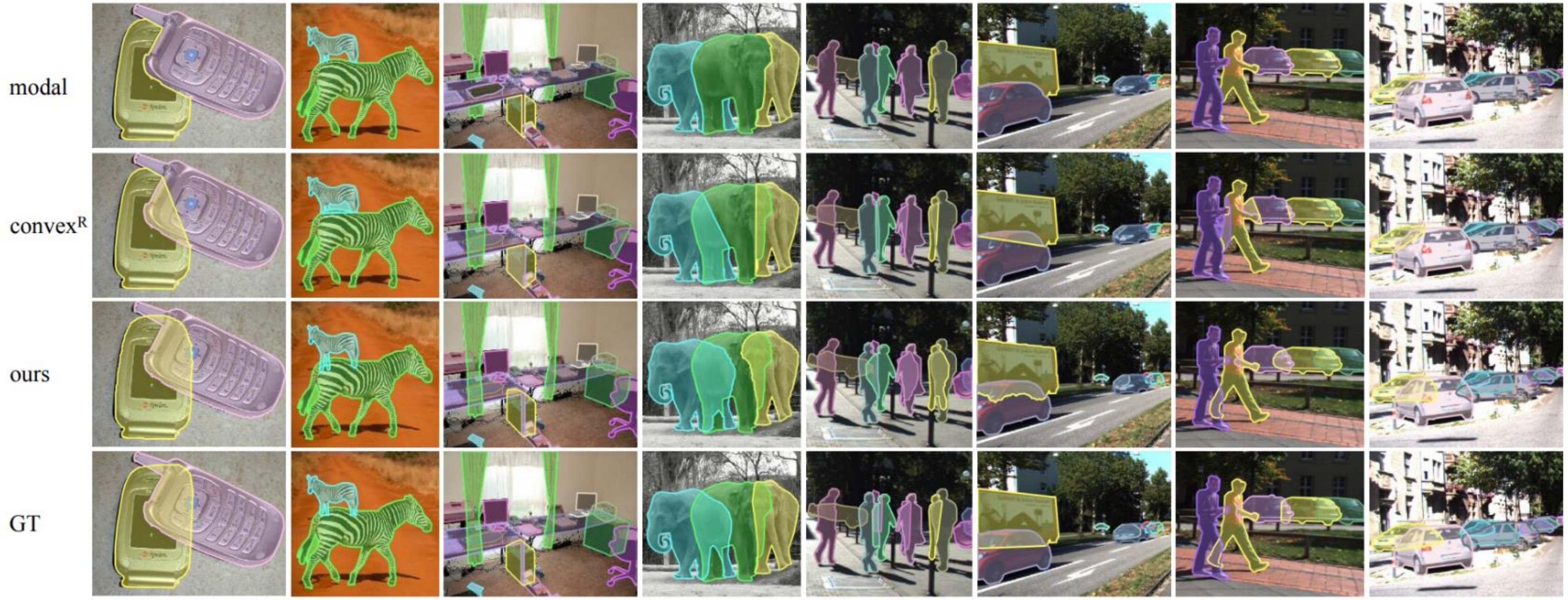
Comparable accuracy to the supervised counterparts.

Amodal Completion

- Ordering-grounded (OG) method shows better performance than Non-ordering-grounded (NOG).
  - Importance of ordering in amodal completion.

method	amodal (train)	COCOA %mIoU	KINS %mIoU
Supervised	✔	82.53	94.81
Raw	✘	65.47	87.03
Convex^R	✘	74.43	90.75
Ours (NOG)	✘	76.91	93.42
Ours (OG)	✘	**81.35**	**94.76**

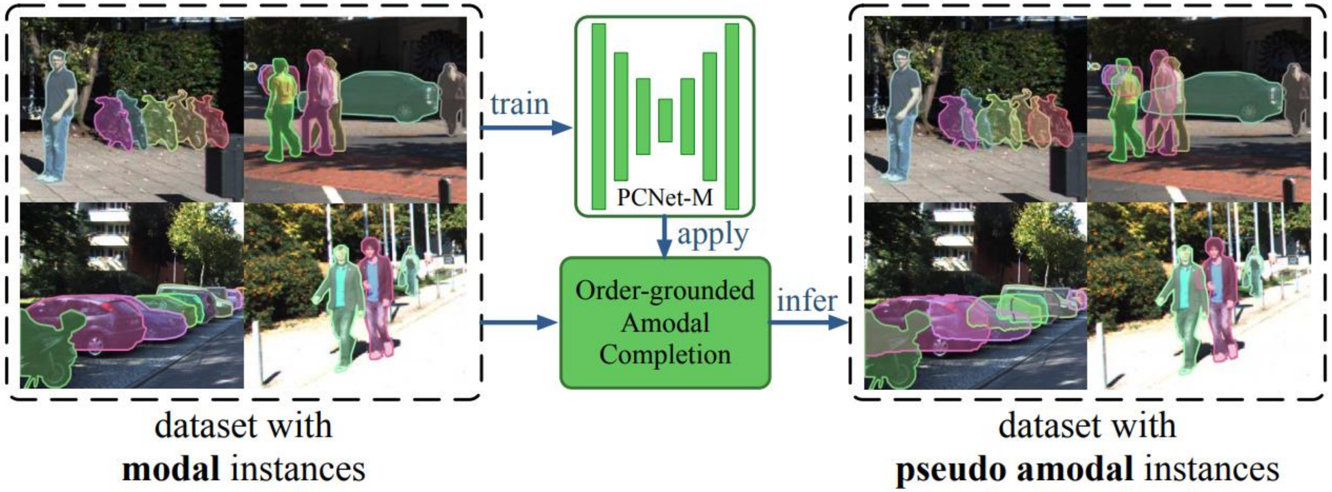
Intersection Over Union on predicted amodal masks



More natural than GT, especially yellow objects

Label Conversation for Amodal instance segmentation

- Convert modal annotations into *pseudo amodal annotations*
- Amodal instance segmentation dataset without manual amodal annotations



Using manual annotations

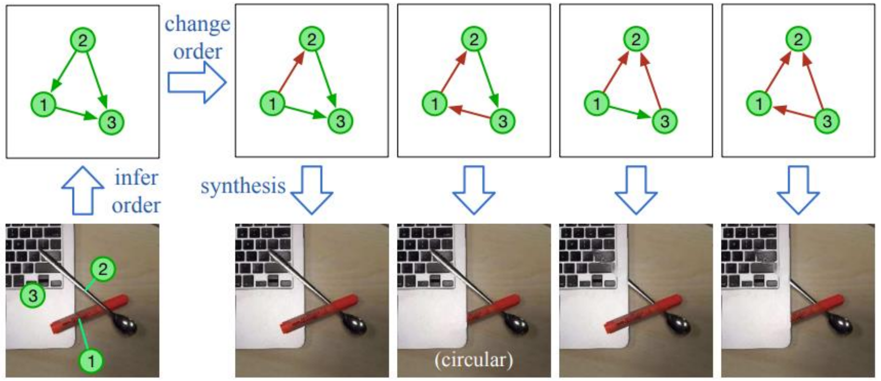
Ann. source	modal (train)	amodal (train)	%mAP
GT [17]	✘	✔	29.3
Raw	✔	✘	22.7
Convex	✔	✘	22.2
Convex^{R}	✔	✘	25.9
Ours	✔	✘	**29.3**

Using our pseudo annotations

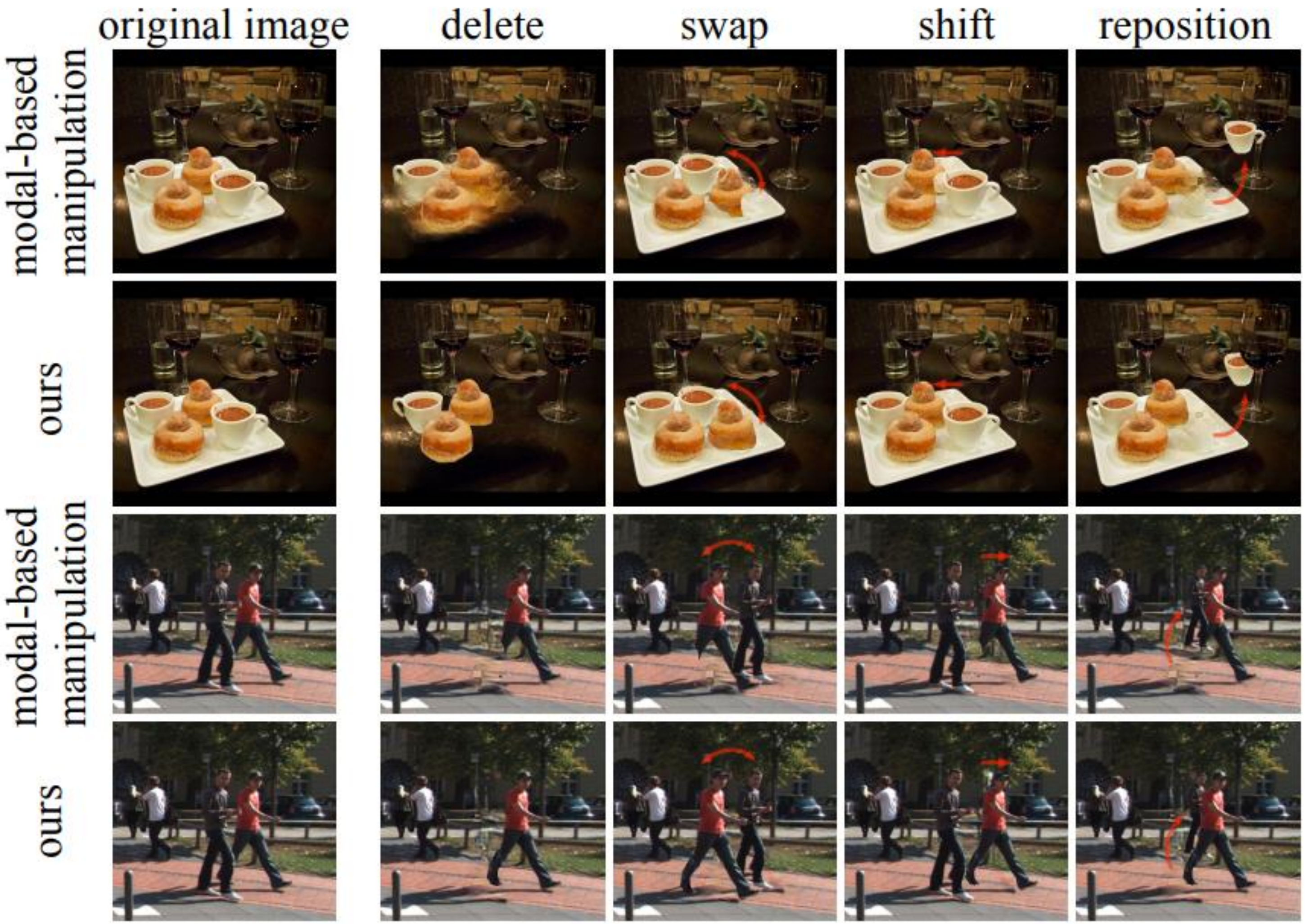
Maybe in the future, we do not need to annotate amodal masks anymore!

Application on Scene Manipulation

- Controlling order and position is possible with an occlusion ordering graph.



Scene synthesis by chaning the ordering graph



Discussion

- For mutual occlusions, the ordering graph cannot be defined, therefore fine-grained boundary-level de-occlusion is required.