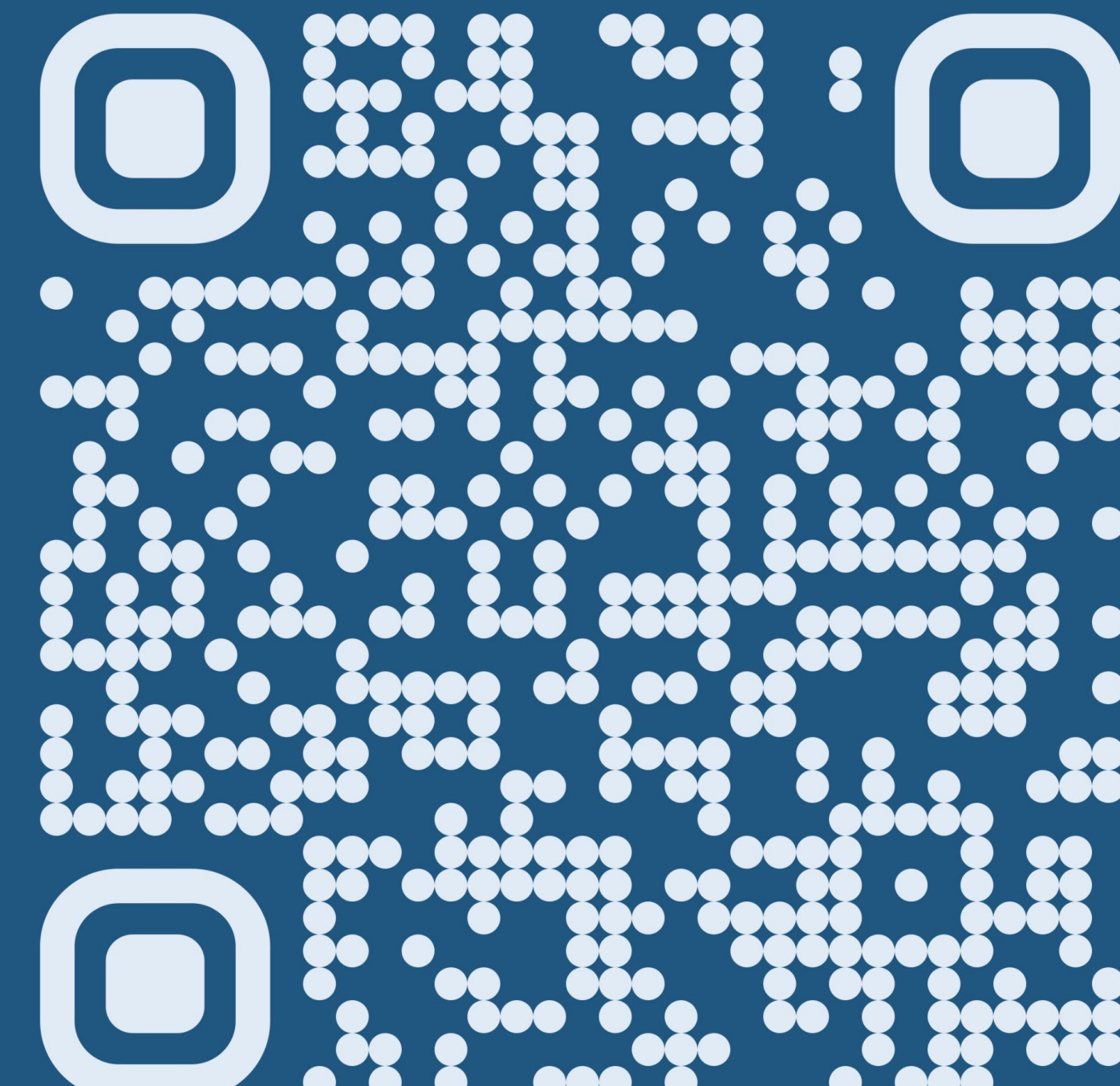
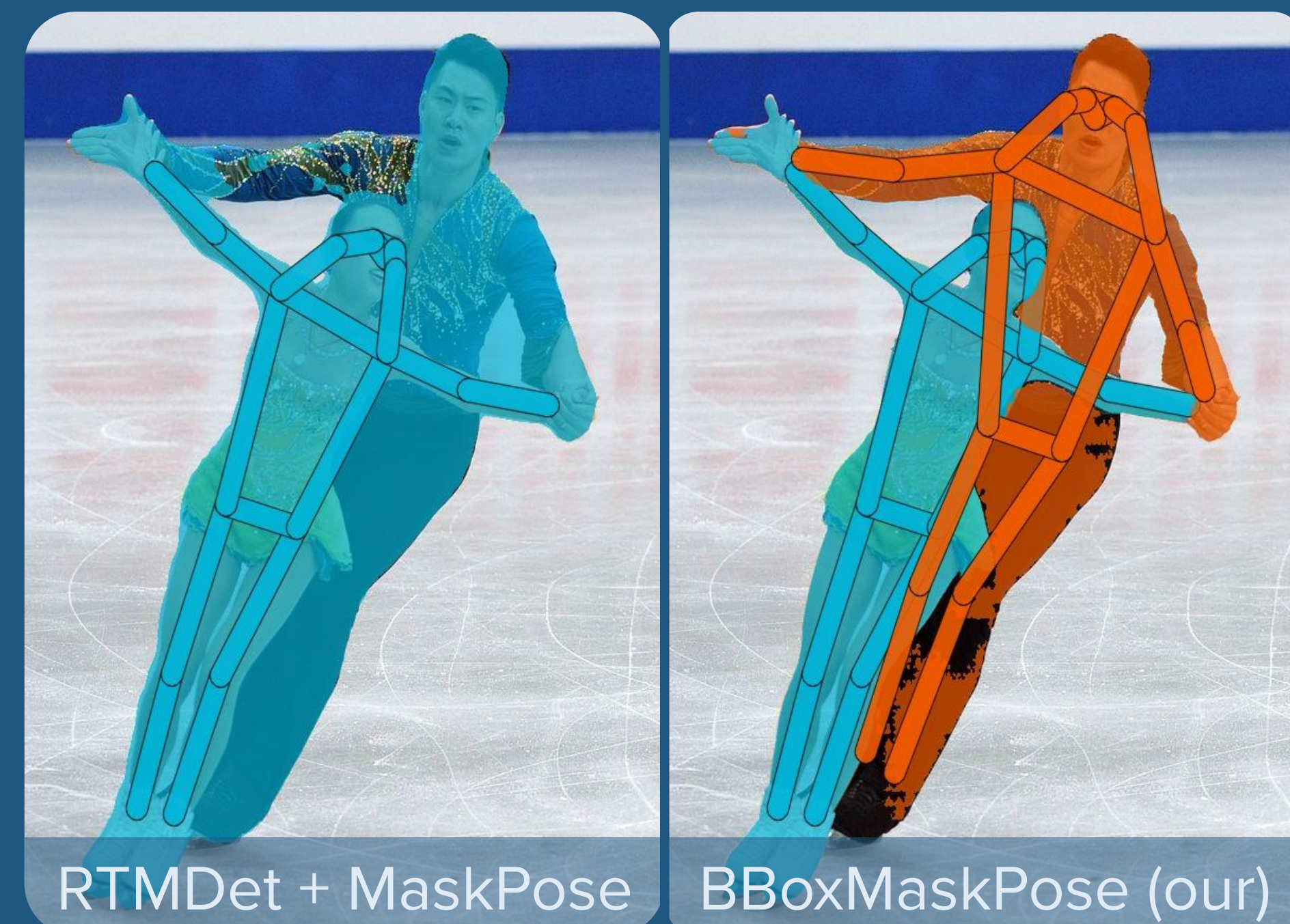


Mutual conditioning of specialists beats human-centered foundational models

Try the Hugging Face demo!



Contributions:

- ↳ New top-down pose estimation model, **MaskPose**, which separates instances better than bbox-based models.
- ↳ A new method, **BBox-Mask-Pose (BMP)**, for detection, segmentation and pose estimation, SOTA on the three tasks, particularly good on multi-body scenes.



Detection, Pose Estimation and Segmentation for Multiple Bodies: Closing the Virtuous Circle

Miroslav Purkrabek and Jiri Matas
Visual Recognition Group, Czech Technical University in Prague



BMP: Detection

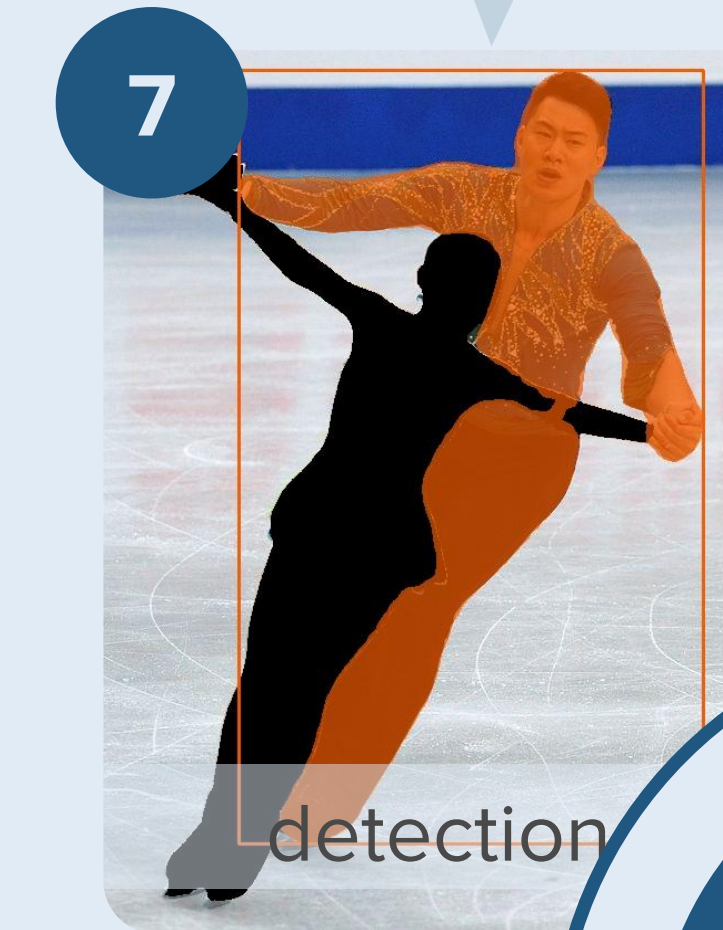
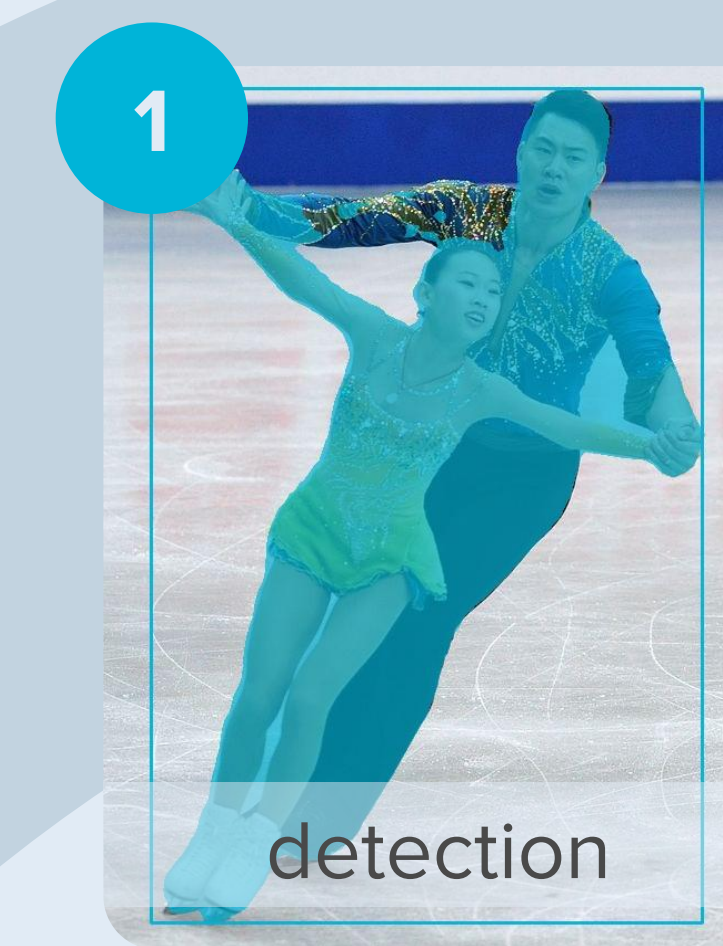
Task: Detect bboxes and masks, *while ignoring already detected instances*

- fine-tune a detector to **ignore masked-out instances** from previous iterations

Finds previously missed instances and/or body parts

Fine-tune on COCO dataset with a new object mask-out augmentation.

Natural loop termination when no instances are found.



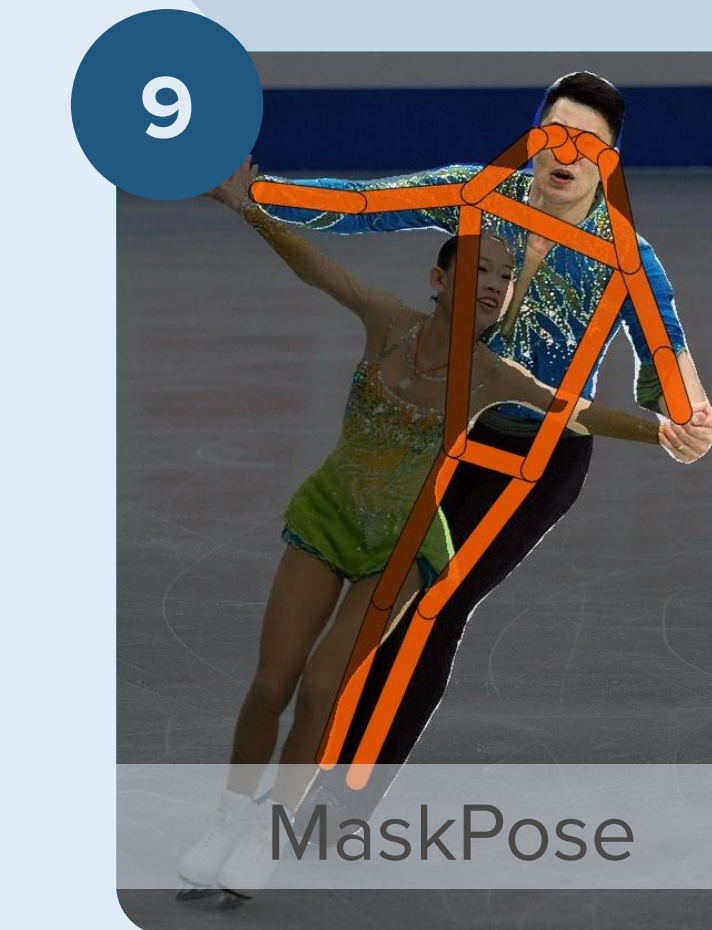
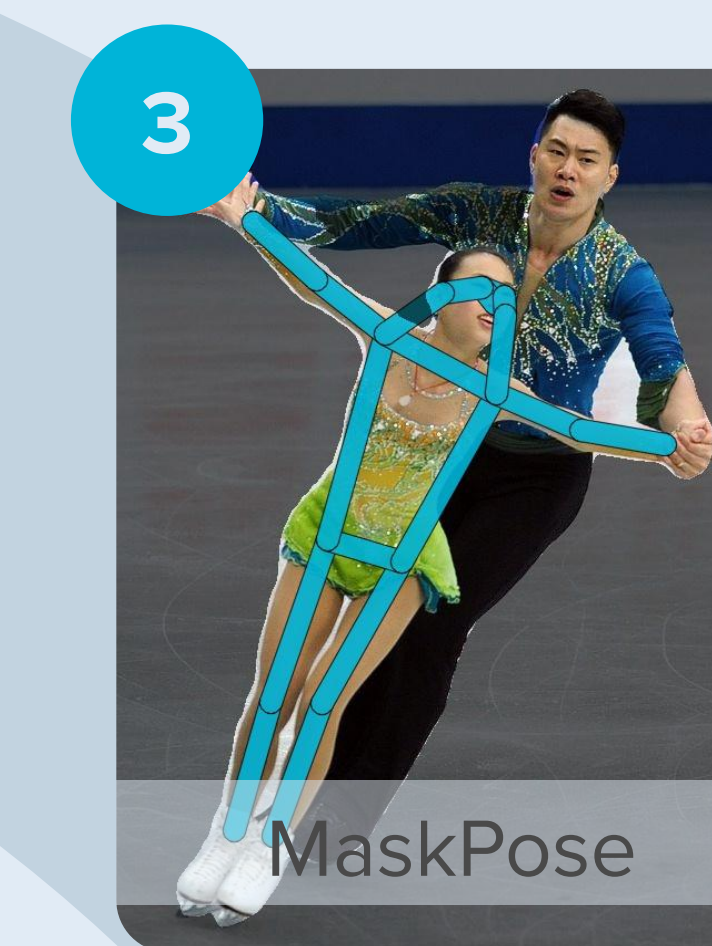
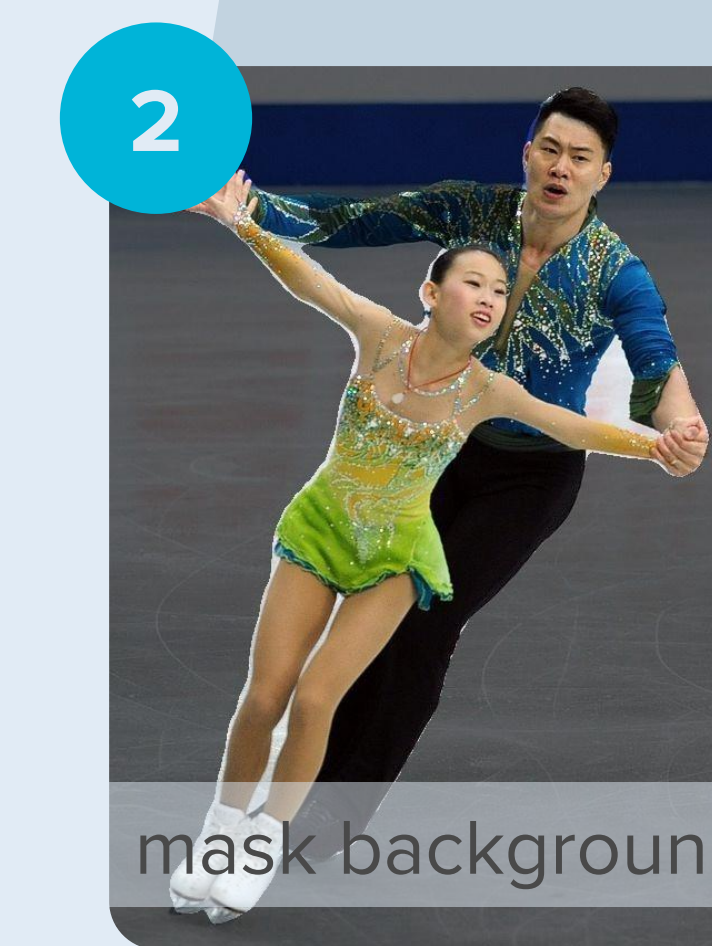
BMP: Pose Estimation

Task: Estimate pose for a given bounding box *and instance segmentation mask*

- **MaskPose**, new pose estimator working on images with **suppressed background**

Pose separates individuals by anatomical constraints

Background masking is more effective than bbox cropping in crowded scenes

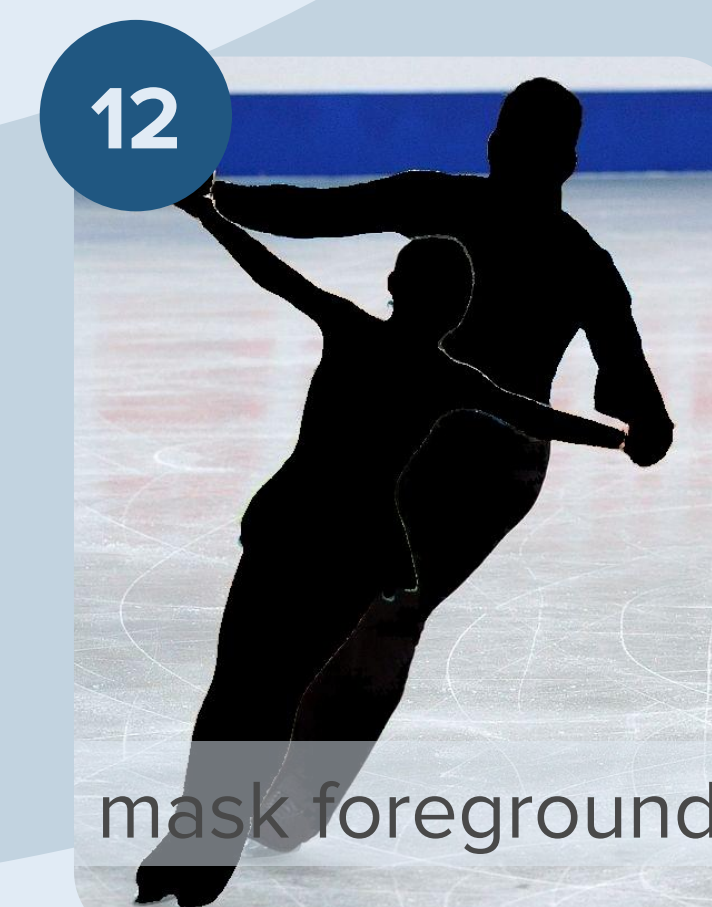
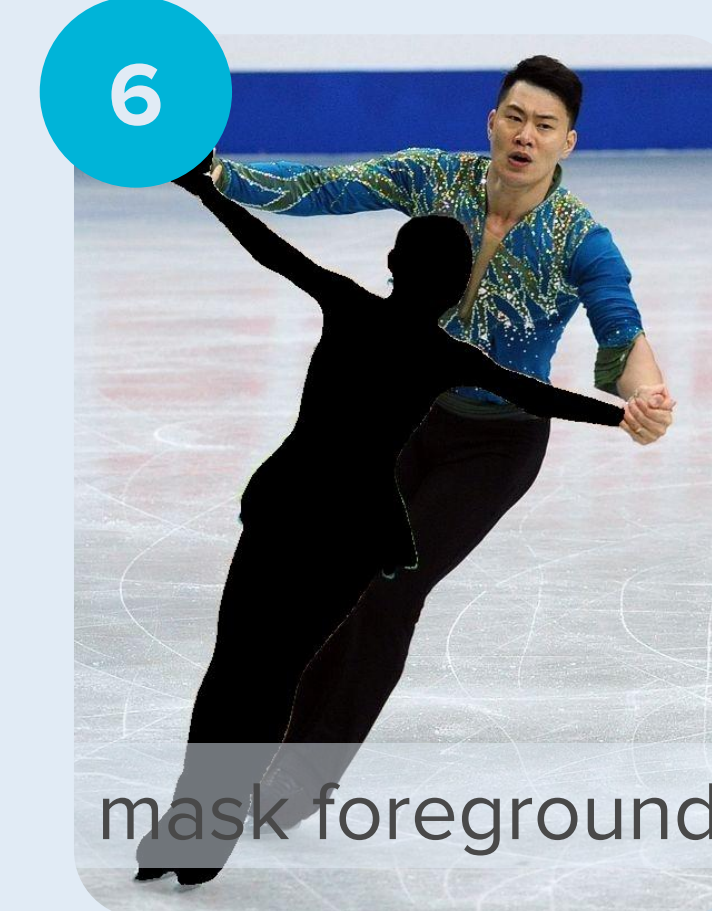
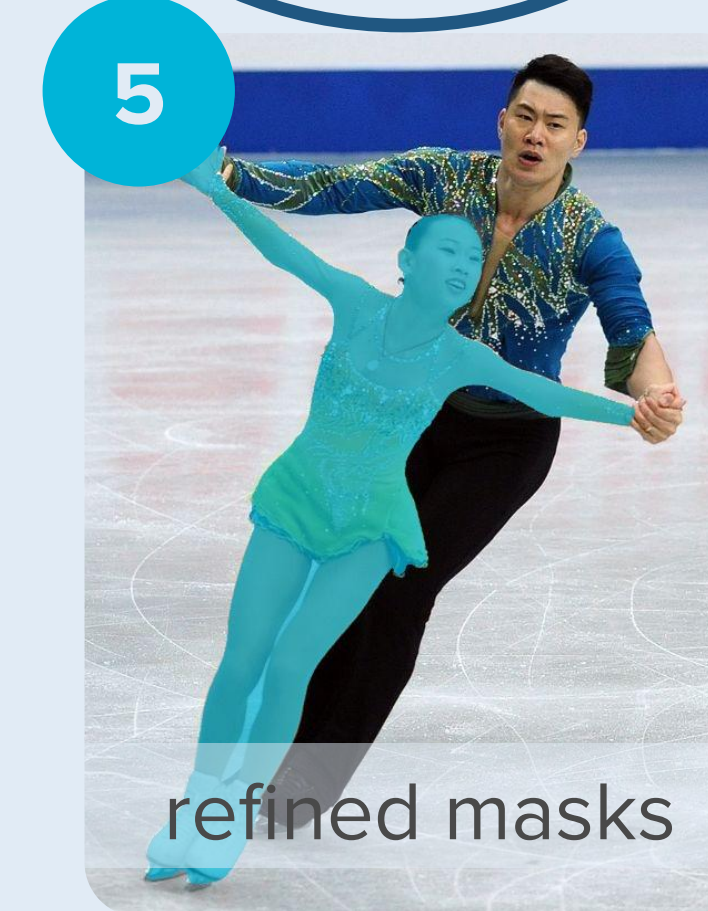
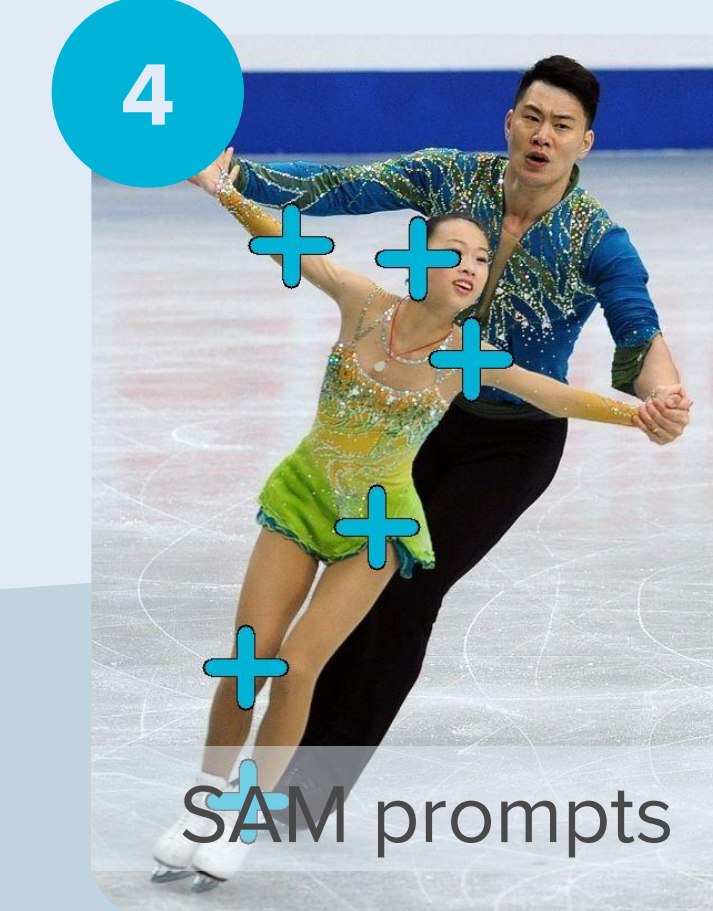


BMP: Mask Refinement

Task: Estimate mask for *a given 2D pose*

- prompt SAM with **selected** keypoints

Offers more detailed segmentation and finer separation of individuals due to prompting with anatomical keypoints.



Relies on prompting with correct prompts; See the extensive ablation study in suppl.



Results & FAQ

Which architecture(s)?

RTMDet-L + MaskPose-B + SAM2-hiera-B+

How fast is it?

Approx **0.5 s/img** per loop (Sapiens: 2.0 s/img)

How many loops?

As long as there are new detections. In our experiments, **2x** was sufficient.

What are the numbers?

COCO: keep-up with SOTA pose: 76.5 mAP

OCHuman: **new SOTA for all three tasks**

det: 30.0 → 35.9 pose: 48.3 → 49.2 segm: 31.4 → 34.0

