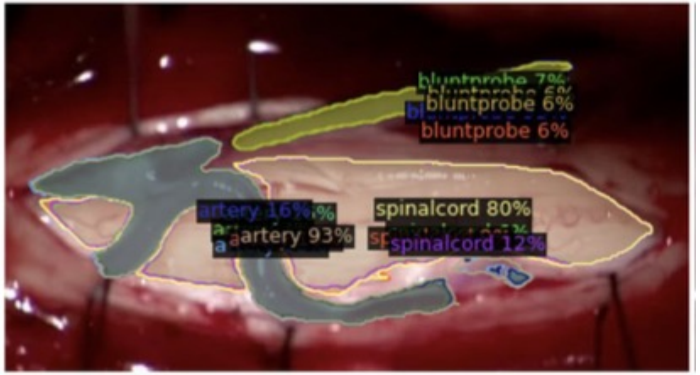
# Beyond mAP: Towards better evaluation of instance segmentation

Rohit Jena[1] Lukas Zhornyak[1]* Nehal Doiphode[1]* Pratik Chaudhari[1] Vivek Buch[2] James Gee[1] Jianbo Shi[1]

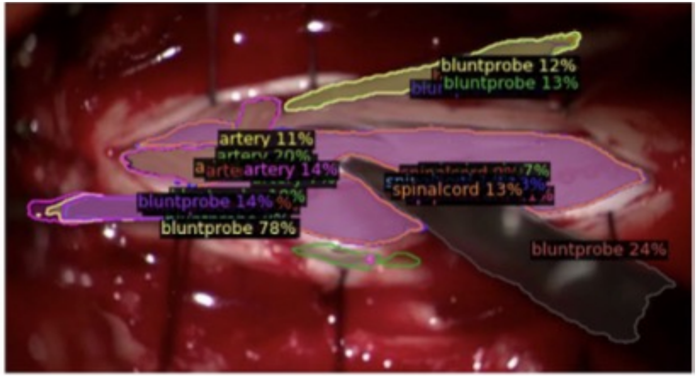
[1]University of Pennsylvania, [2]Stanford University

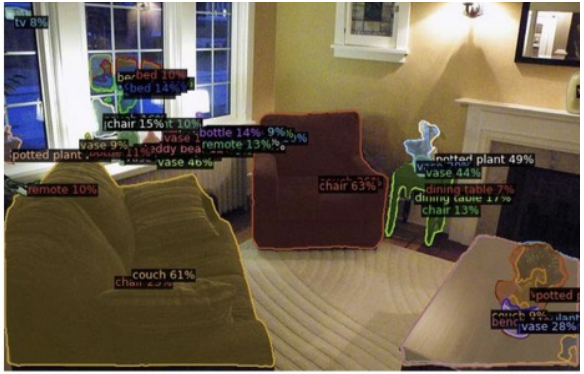
Session WED-AM-293

# Motivation


bluntprobe 6%
bluntprobe 6%
artery 16%
artery 93%
spinalcord 80%
spinalcord 12%

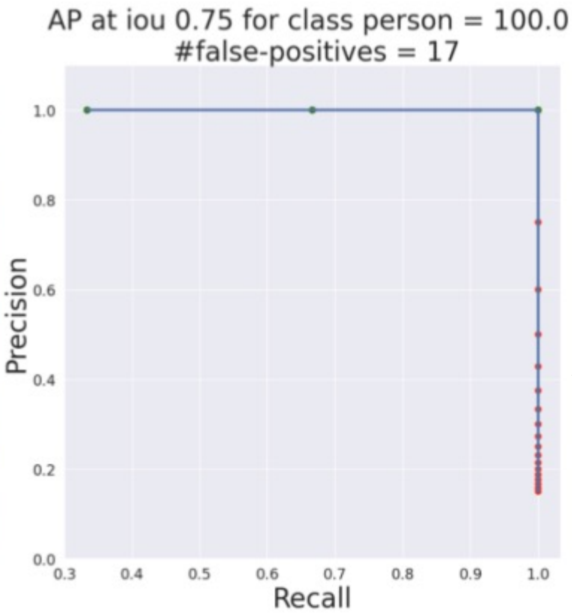

giraffe 8%
giraffe 68%
giraffe 8%
giraffe 61%
giraffe 16%
giraffe 9%
giraffe 76%
giraffe 11%


bluntprobe 12%
bluntprobe 13%
artery 11%
artery 14%
spinalcord 13%
bluntprobe 14%
bluntprobe 78%
bluntprobe 24%


tv 8%
bed 14%
chair 15%
bottle 14%
remote 13%
vase 9%
vase 46%
remote 10%
chair 63%
potted plant 49%
vase 44%
dining table 7%
chair 13%
couch 61%
vase 28%

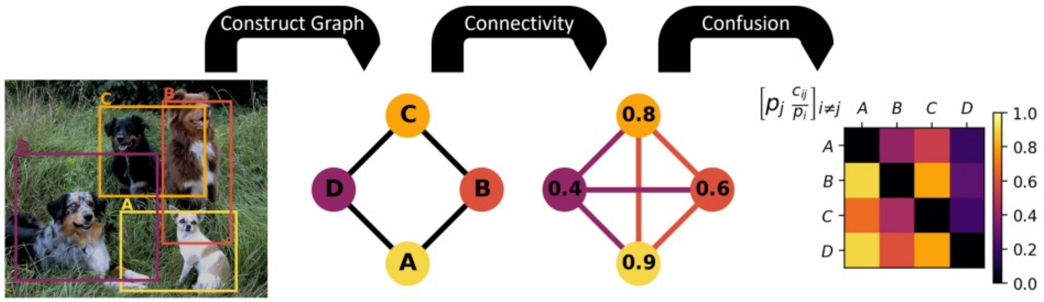
# Motivation

person 9%
person 11%
person 5%
person 10%
person 5%

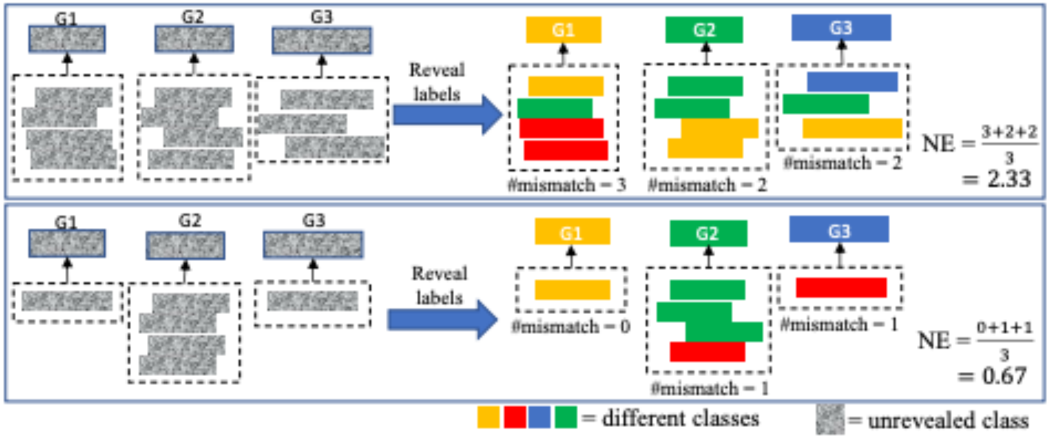



AP at iou 0.75 for class person = 100.0
#false-positives = 17
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

# Hedged predictions




Measures spatial hedging




Measures categorical hedging

# Mitigating hedging

Inspiration from bottom-up methods.

Semantic Sorting: re-rank instances based on semantic masks.

Semantic NMS: Remove instances that do not have "occupancy" from semantic mask.

**Algorithm 1:** Pseudocode for semantic sorting and NMS, given instances $D_k$ with category $c_k$ and confidence $\tau_k$, threshold $thr$, semantic masks $M$

**Data:** $\{D_k, c_k, \tau_k\}_{k=1 \dots N}$, $\{M_c\}_{c=1 \dots C}$

**Result:** Boolean array $keep$ indicating preservation of instances

**for** $k = 1 \dots N$ **do**

  $pr \leftarrow \mathrm{precision}(D_k, M_{c_k})$;

  $iou \leftarrow \mathrm{IoU}(D_k, M_{c_k})$;

  $\tau_k \leftarrow \tau_k + pr + (1 - iou)$;

**end**

$(D, c, \tau) = \mathrm{sort}(D, c, \tau)$;  // sort by decreasing $\tau$

**for** $k = 1 \dots N$ **do**

  $overlap \leftarrow \mathrm{precision}(D_k, M_{c_k})$;

  **if** $overlap \geq thr$ **then**

    $keep[k] = True$;

    $M_{c_k} = M_{c_k} \backslash D_k$

  **else**

    $keep[k] = False$

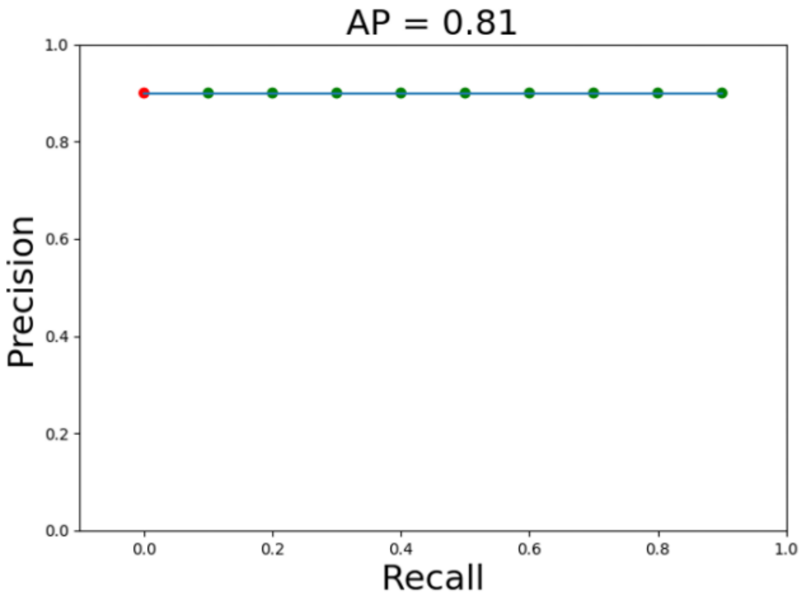
  **end**

**end**

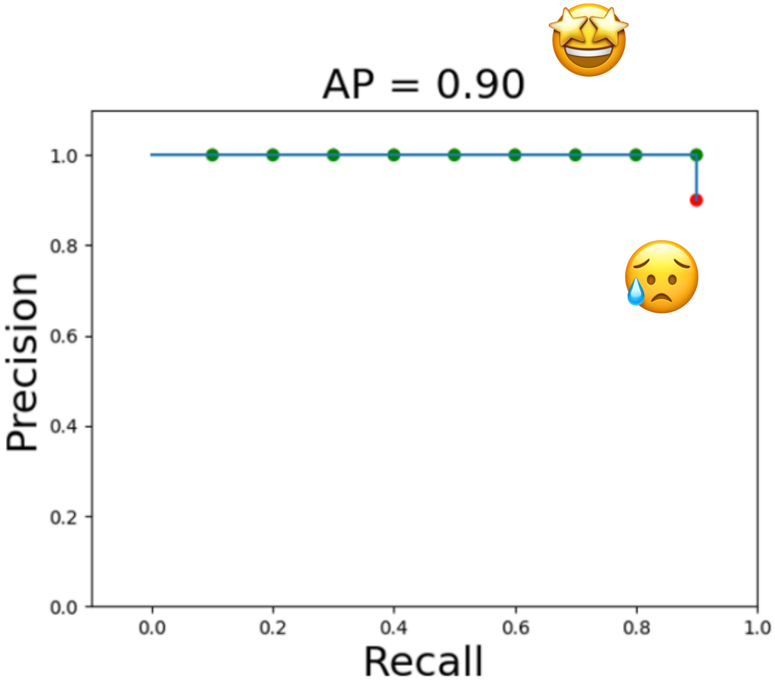
# Qualitative results

# Let’s dive deeper!

# A toy example

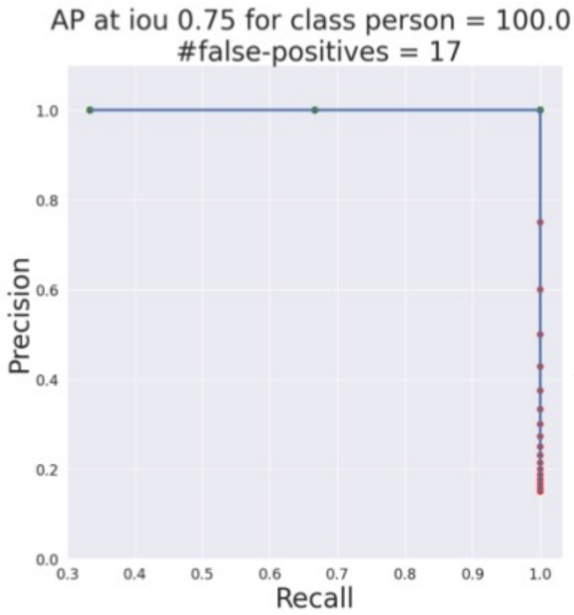



(a)

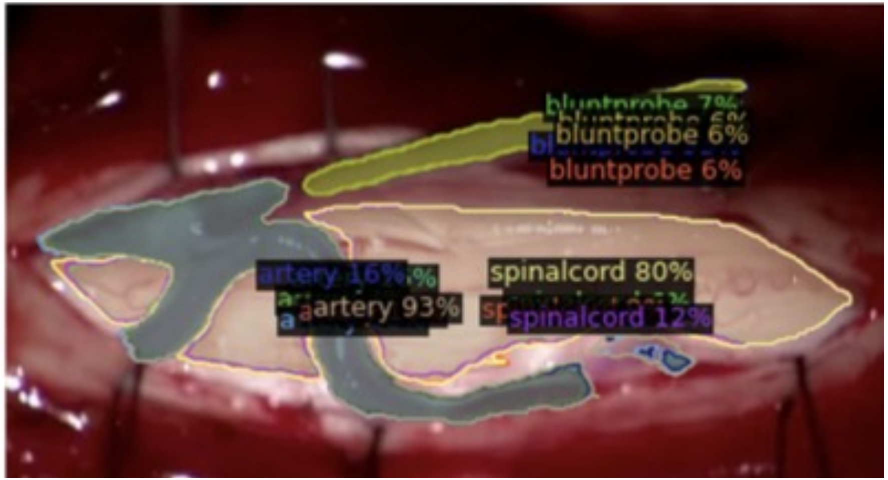


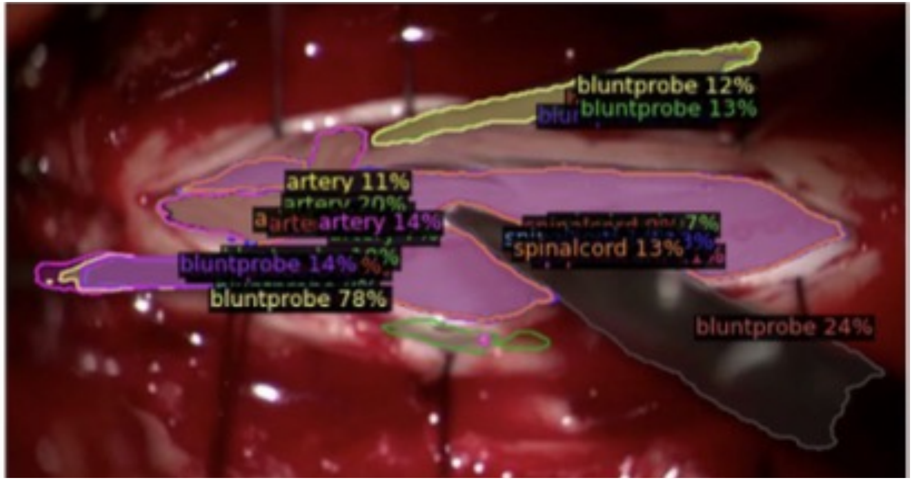

(b)

# Defining hedging

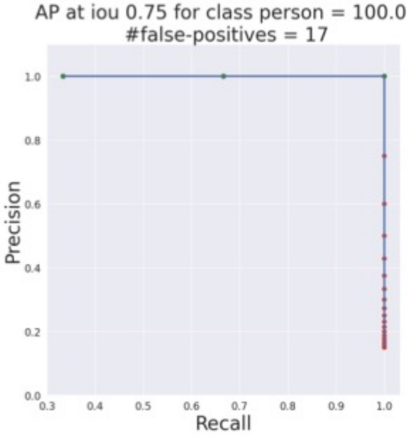




# Motivation

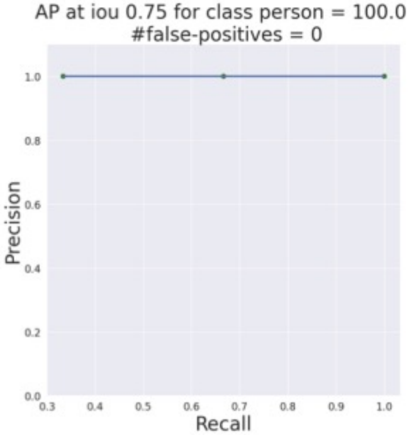

bluntprobe 6%
bluntprobe 6%
artery 16%
artery 93%
spinalcord 80%
spinalcord 12%


bluntprobe 12%
bluntprobe 13%
artery 11%
artery 14%
spinalcord 13%
bluntprobe 14%
bluntprobe 78%
bluntprobe 24%

# A closer look

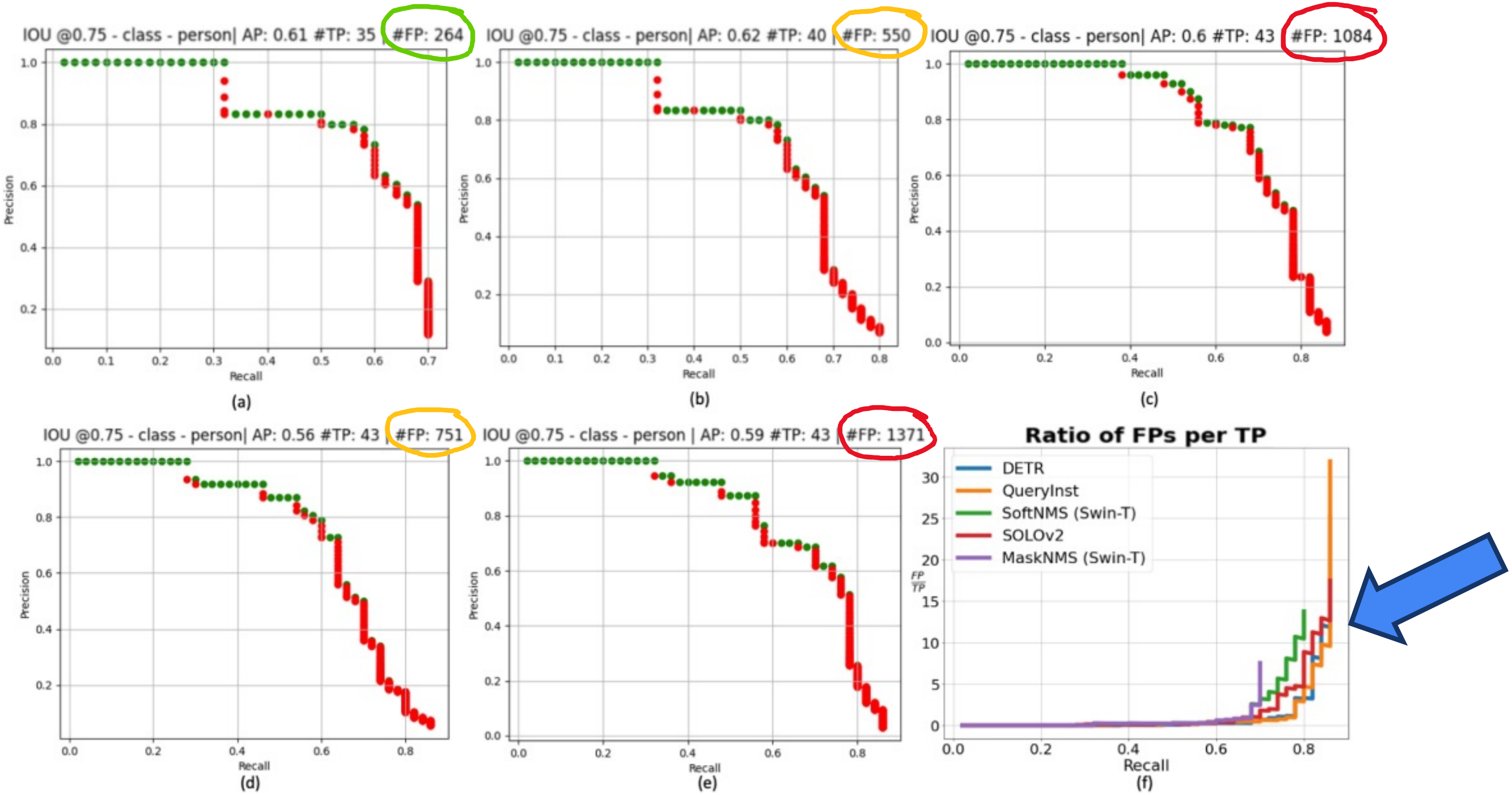





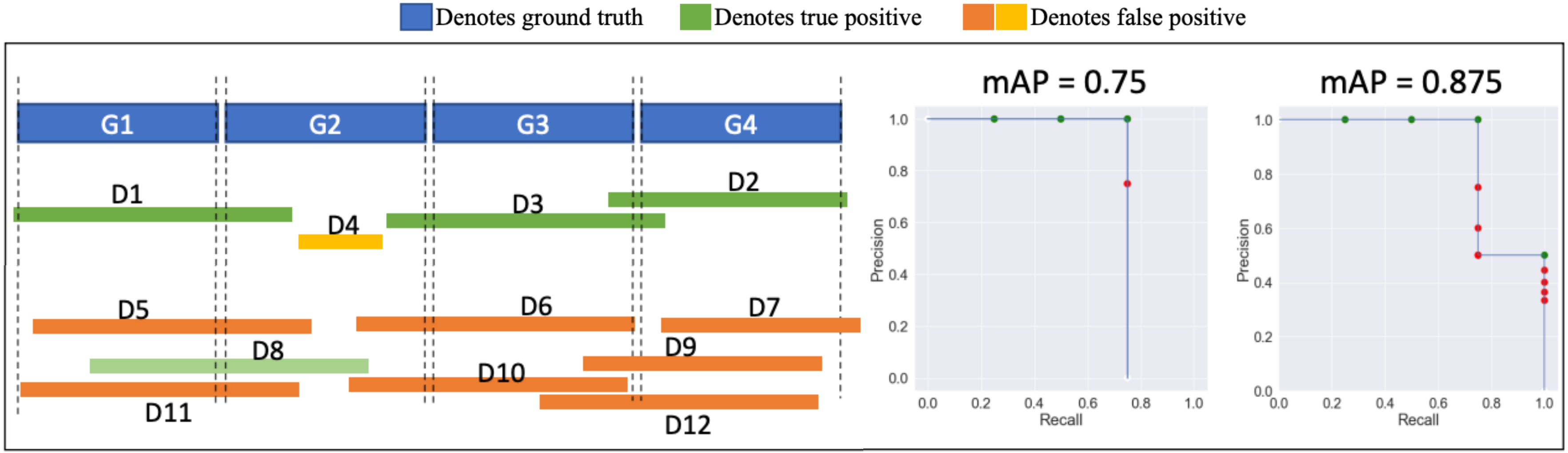



How to distinguish this?

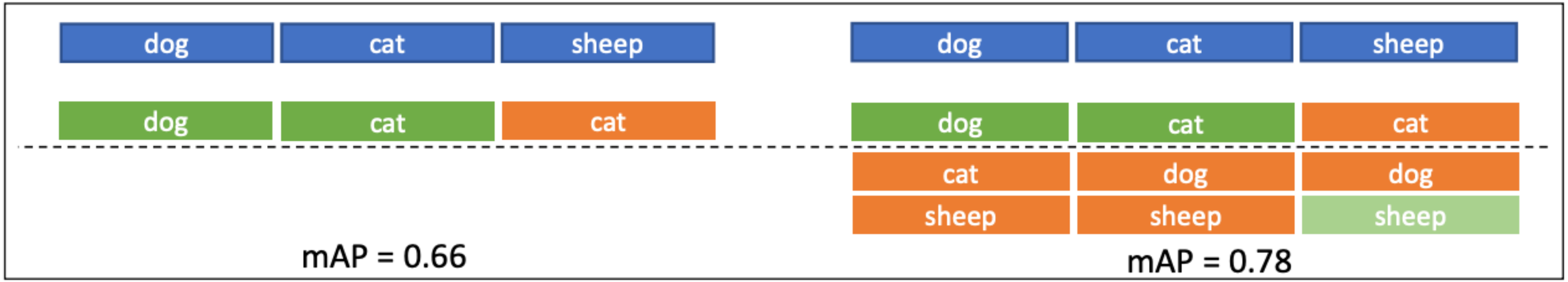
# Shouldn’t NMS be clearing this up?

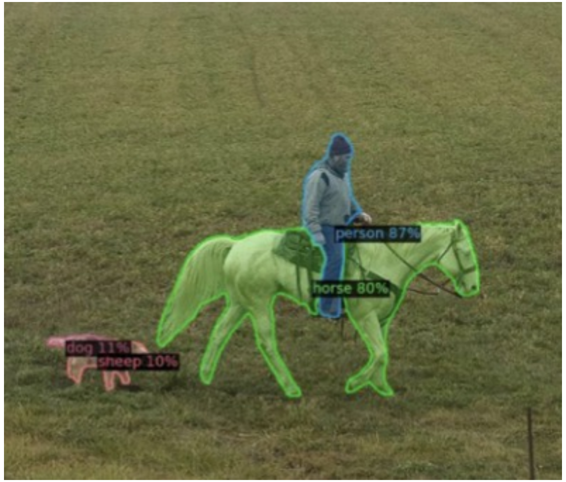


# Spatial hedging



# Categorical hedging

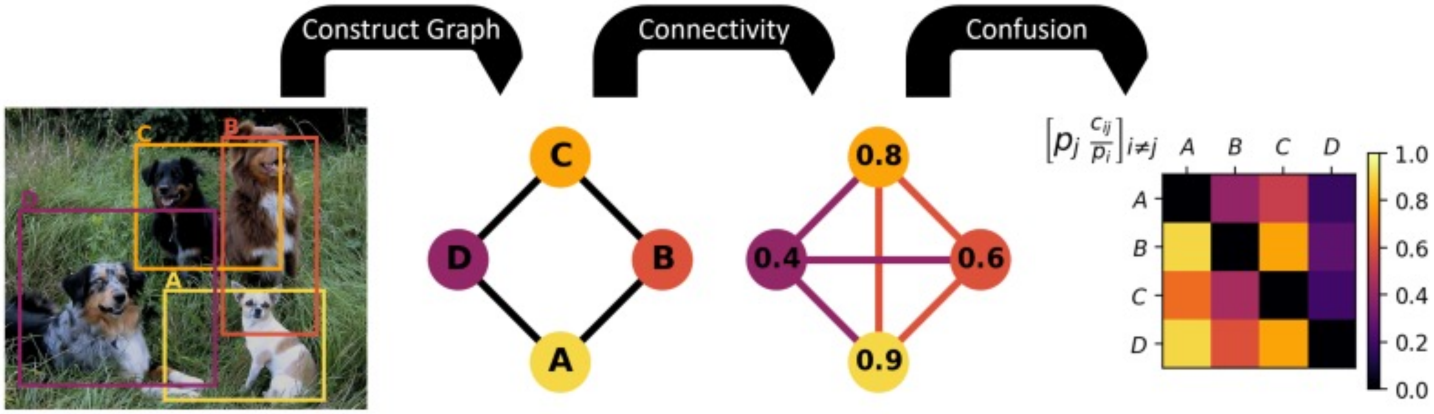


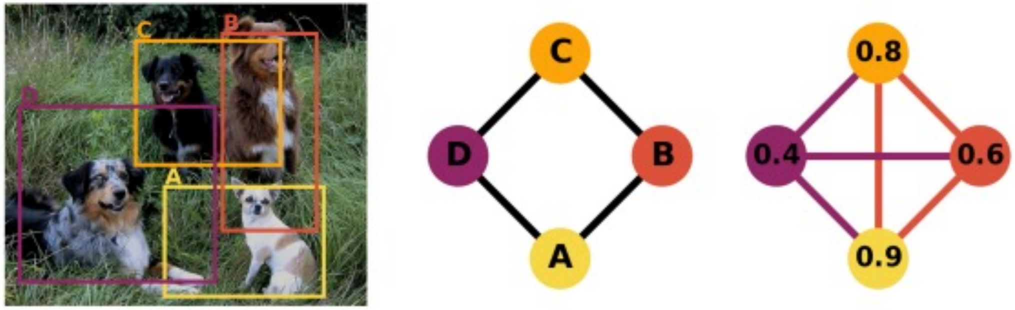

# Quantifying hedging

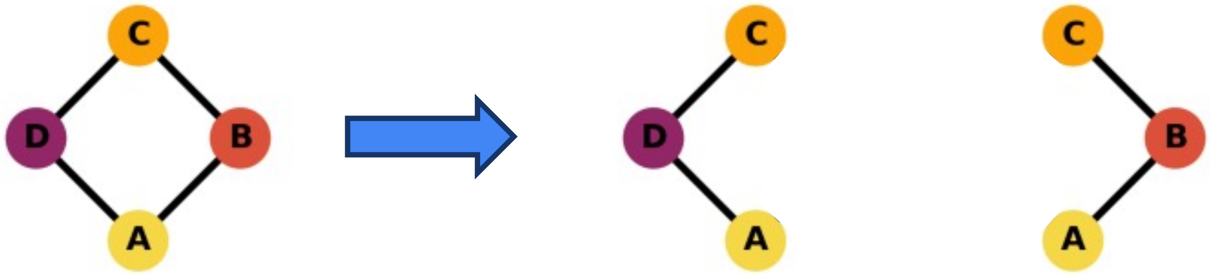
# Duplicate confusion (DC)

What is the average overlap between any two instances?

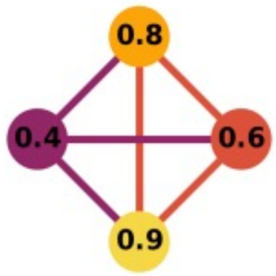


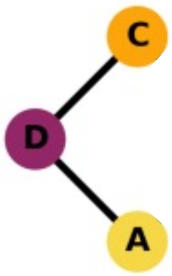
# Duplicate confusion (DC)



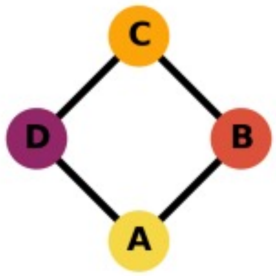



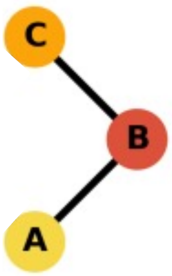
# Duplicate confusion (DC)




$$\min_{k \in \pi_1} \tau_k = 0.4$$

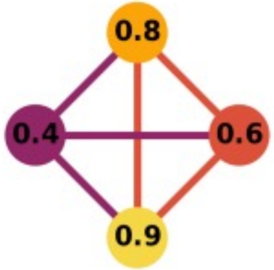



$$\min_{k \in \pi_2} \tau_k = 0.6$$

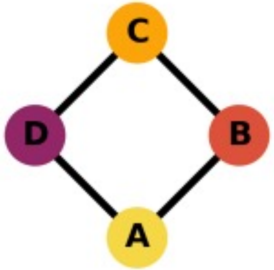
$$c_{AC} = \max_{\pi} \min_{k \in \pi} \tau_k = 0.6$$

# Duplicate confusion (DC)



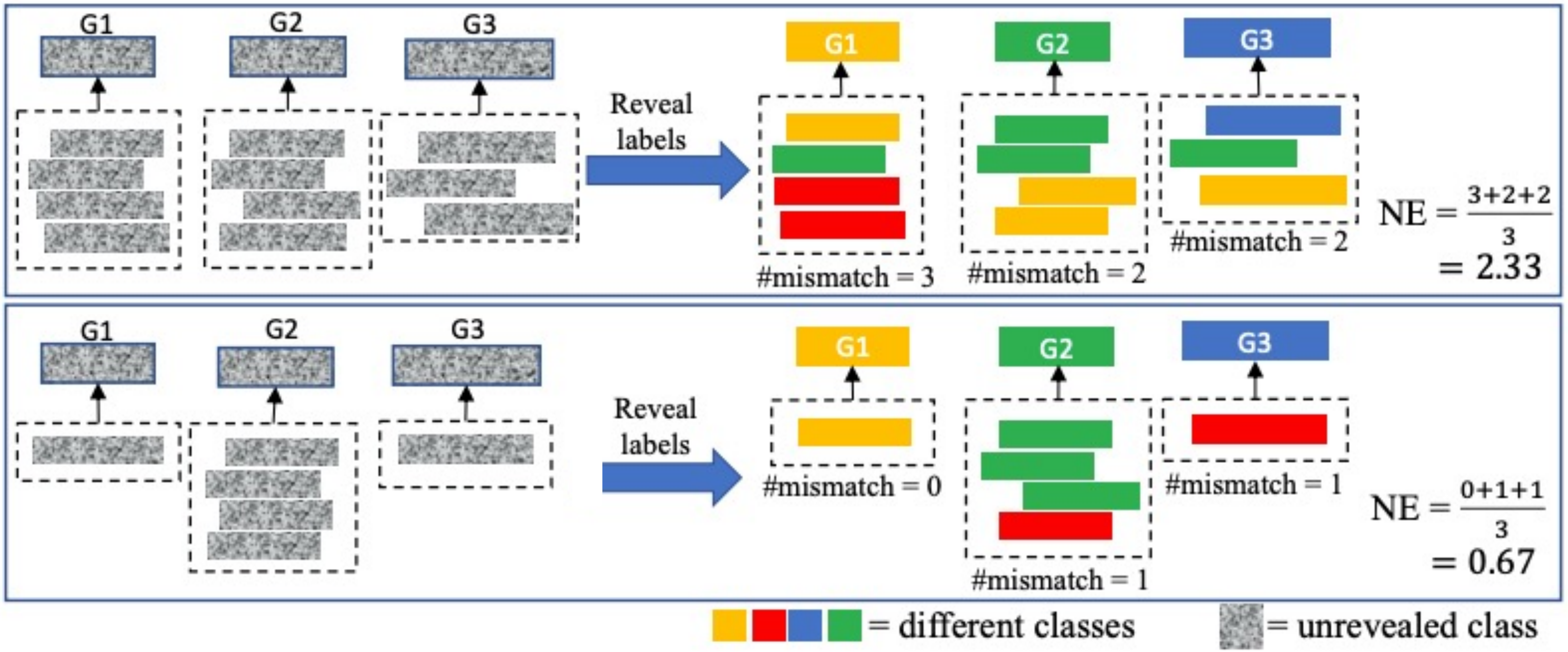

$$c_{ij} = \max_{\pi \in T_{ij}} \min_{k \in \pi} \tau_k$$

$$\sum_{j \neq i} \tau_j c_{ij}$$

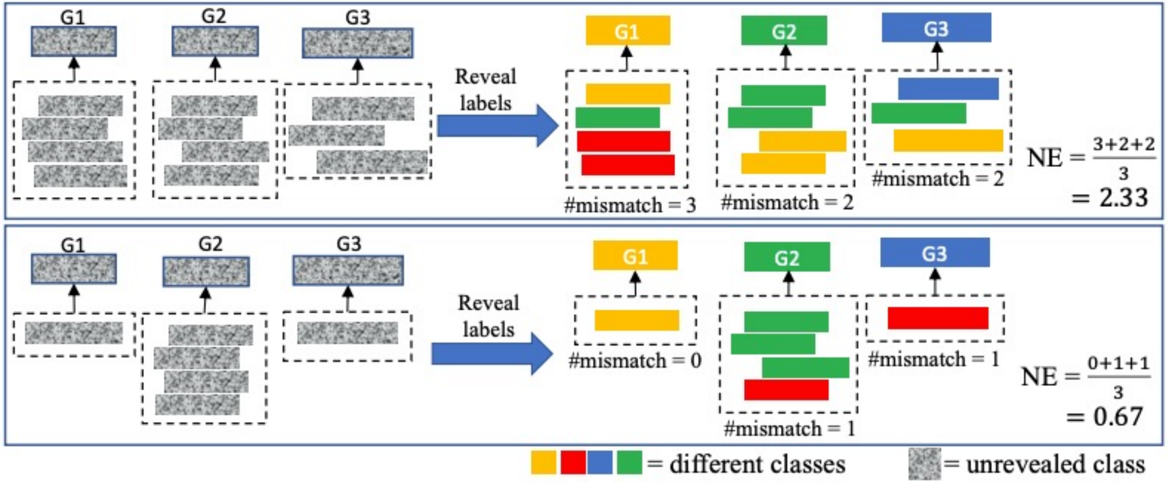



$$\mathrm{DC}_{tv} = \frac{1}{m} \sum_{i}^{m} \sum_{j \neq i} \frac{\tau_j c_{ij}}{\tau_i}$$

# Naming Error (NE)



# Naming Error (NE)




$$g(D_j) = \begin{cases} \arg\max_i \mathrm{IoU}(D_j, G_i) & , \max_i \mathrm{IoU}(D_j, G_i) \geq 0.5 \\ -1 & , \text{otherwise} \end{cases}$$

$$\mathrm{NE} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j:g(D_j)=i} \mathbb{I}\left[l_{D_j} \neq l_{G_i}\right]$$

# Other metrics

IEEE TRANSACTION OF PATTERN ANALYSIS AND MACHINE INTELLIGENCE 

## One Metric to Measure them All: Localisation Recall Precision (LRP) for Evaluating Visual Detection Tasks

Kemal Oksuz†, Baris Can Cam, Sinan Kalkan‡, and Emre Akbas‡

**Abstract**—Despite being widely used as a performance measure for visual detection tasks, Average Precision (AP) is limited in (i) reflecting localisation quality, (ii) interpretability and (iii) robustness to the design choices regarding its computation, and its applicability to outputs without confidence scores. Panoptic Quality (PQ), a measure proposed for evaluating panoptic segmentation (Kirillov et al., 2019), does not suffer from these limitations but is limited to panoptic segmentation. In this paper, we propose Localisation Recall Precision (LRP) Error as the average matching error of a visual detector computed based on both its localisation and classification qualities for a given confidence score threshold. LRP Error, initially proposed only for object detection by Oksuz et al. (2018), does not suffer from the aforementioned limitations and is applicable to all visual detection tasks. We also introduce Optimal LRP (oLRP) Error as the minimum LRP Error obtained over confidence scores to evaluate visual detectors and obtain optimal thresholds for deployment. We provide a detailed comparative analysis of LRP Error with AP and PQ, and use nearly 100 state-of-the-art visual detectors from seven visual detection tasks (i.e. object detection, keypoint detection, instance segmentation, panoptic segmentation, visual relationship detection, zero-shot detection and generalised zero-shot detection) using ten datasets to empirically show that LRP Error provides richer and more discriminative information than its counterparts. Code available at: https://github.com/kemaloksuz/LRP-Error.

**Index Terms**—Localisation Recall Precision Average Precision Panoptic Quality Object Detection Keypoint Detection Instance Segmentation Panoptic Segmentation Performance Metric Threshold.



### 1 Introduction

Many vision applications require identifying objects and object-related information from images. Such identification can be performed at different levels of detail, which are addressed by different detection tasks such as "object detection" for identifying labels of objects and boxes bounding them, "keypoint detection" for finding keypoints on objects, "instance segmentation" for identifying the classes of objects and localising them with masks, and "panoptic segmentation" for classifying both background classes and objects by providing detection ids and labels of pixels in an image. Accurately evaluating performances of these methods is crucial for developing better solutions.

#### 1.1 Important features for a performance measure

To facilitate our analysis, we define three important features for performance measures of visual detection methods:

**Completeness.** Arguably, three most important performance aspects that an evaluation measure should take into account in a visual detection task are false positive (FP) rate, false negative (FN) rate and localisation error. We call a performance measure "complete" if it precisely takes into account all three quantities.

strengths and weaknesses of the detector being evaluated. To provide such insight, the evaluation measure should ideally comprise interpretable components.

**Practicality.** Any issue that arises during practical use of a performance measure diminishes its practicality. This could be, for example, any discrepancy between the well-defined theoretical description of the evaluation measure and its actual application in practice, or any shortcoming that limits the applicability of the measure to certain cases.

#### 1.2 Overview of Average Precision and Its Limitations

Today "average precision" (AP) is the de facto standard for evaluating performance on many visual detection tasks and competitions [1], [2], [3], [4], [5], [6], [7]. Computing AP for a class involves a set of detection results with confidence scores and a set of ground-truth items (e.g. bounding boxes in the case of object detection). First, detections are matched to ground-truth items (GT) based on a predefined spatial overlap criterion such as Intersection over Union (IoU)[1] being larger than 0.50. Each GT can only match one detection and if there are multiple detections that satisfy the overlap criterion, the one with the highest confidence score is matched. A detection that is matched to a GT is

Good at counting FPs, FNs.



## A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation

F. Perazzi[1,2] J. Pont-Tuset[1] B. McWilliams[2] L. Van Gool[1] M. Gross[1,2] A. Sorkine-Hornung[2]
[1]ETH Zurich [2]Disney Research

### Abstract

*Over the years, datasets and benchmarks have proven their fundamental importance in computer vision research, enabling targeted progress and objective comparisons in many fields. At the same time, legacy datasets may impend the evolution of a field due to saturated algorithm performance and the lack of contemporary, high quality data. In this work we present a new benchmark dataset and evaluation methodology for the area of* video object *segmentation. The dataset, named DAVIS (Densely Annotated VIdeo Segmentation), consists of fifty high quality, Full HD video sequences, spanning multiple occurrences of common video object segmentation challenges such as occlusions, motion-blur and appearance changes. Each video is accompanied by densely annotated, pixel-accurate and per-frame ground truth segmentation. In addition, we provide a comprehensive analysis of several state-of-the-art segmentation approaches using three complementary metrics that measure the spatial extent of the segmentation, the accuracy of the silhouette contours and the temporal coherence. The results uncover strengths and weaknesses of current approaches, opening up promising directions for future works.*

Figure 1: Sample sequences from our dataset, with ground truth segmentation masks overlayed. Please refer to the supplemental material for the complete dataset.

### 1. Introduction

Video object segmentation is a binary labeling problem aiming to separate foreground object(s) from the background region of a video. A pixel-accurate, spatio-temporal bipartition of the video is instrumental to several applications including, among others, action recognition, object tracking, video summarization, and rotoscoping for video editing. Despite remarkable progress in recent years, video object segmentation still remains a challenging problem and most existing approaches still exhibit too severe limitations in terms of quality and efficiency to be applicable in practical applications, *e.g.* for processing large datasets, or video

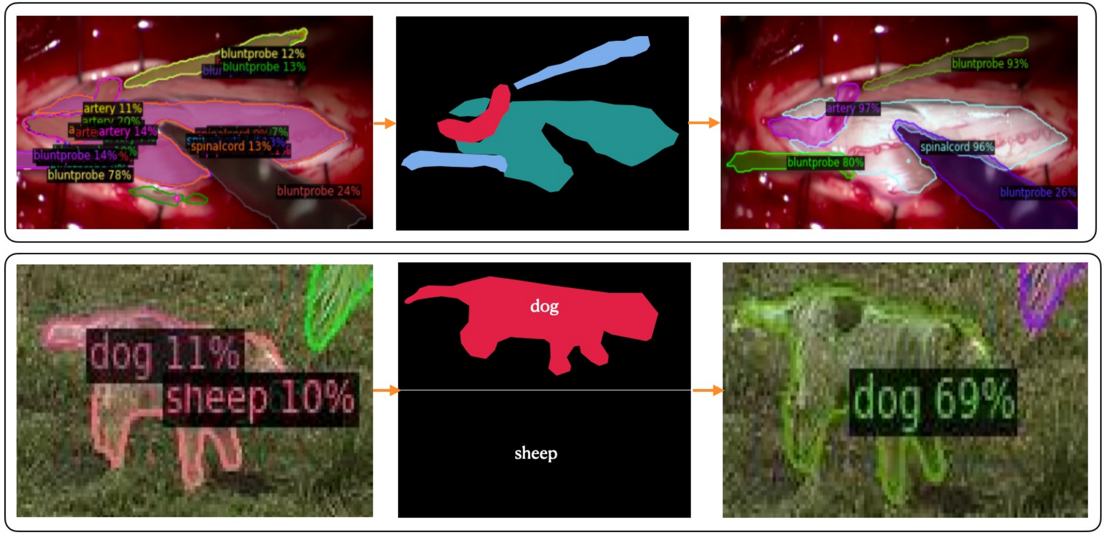
and object recognition, which have experienced remarkable progress in the recent years. A key factor bootstrapping this progress has been the availability of large scale datasets and benchmarks [12, 26, 29, 42]. This is in stark contrast to video object segmentation. While several datasets exists for various different video segmentation tasks [1, 4, 5, 15, 20, 21, 25, 38, 41, 44, 46, 47], none of them targets the specific task of video *object* segmentation.

To date, the most widely adopted dataset is that of [47], which, however, was originally proposed for joint segmentation and tracking and only contains six low-resolution video sequences, which are not representative anymore for the image quality and resolution encountered in today's video processing applications. As a consequence, evaluations performed on such datasets are likely to be overfitted, without reliable indicators regarding the differences between individual video segmentation approaches, and the real performance on unseen, more contemporary data becomes difficult to determine [6]. Despite the effort of some authors to augment their evaluation with additional datasets,

Evaluates mask quality.

# Mitigating hedging

# Semantic Sorting and NMS




**Algorithm 1:** Pseudocode for semantic sorting and NMS, given instances $D_k$ with category $c_k$ and confidence $\tau_k$, threshold $thr$, semantic masks $M$

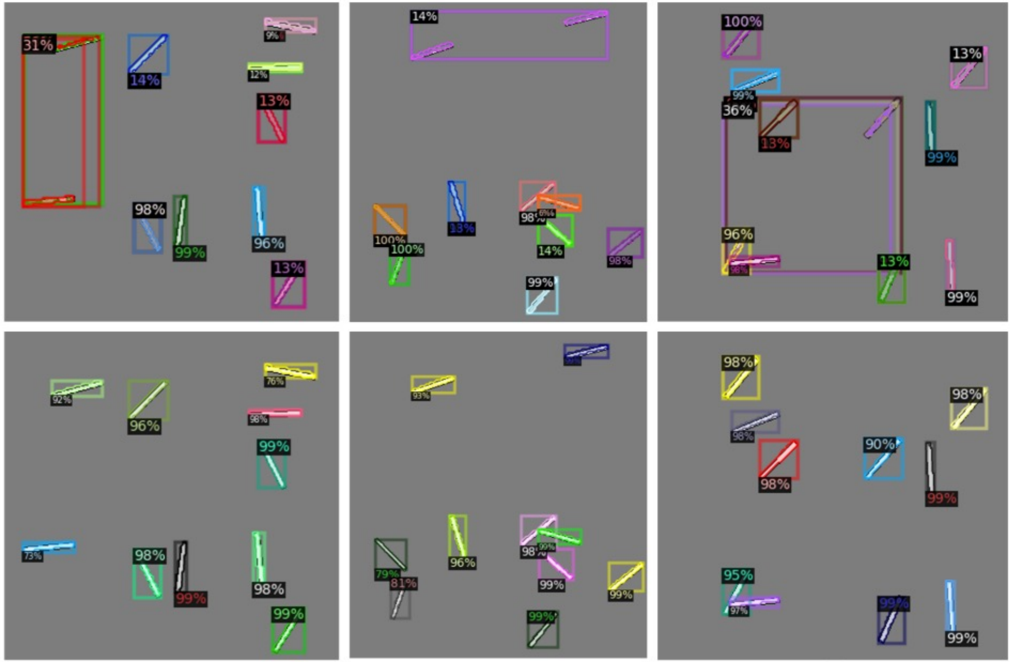
**Data:** $\{D_k, c_k, \tau_k\}_{k=1...N}$, $\{M_c\}_{c=1...C}$
**Result:** Boolean array *keep* indicating preservation of instances

**for** $k = 1 \ldots N$ **do**
  $pr \leftarrow \text{precision}(D_k, M_{c_k})$;
  $iou \leftarrow \text{IoU}(D_k, M_{c_k})$;
  $\tau_k \leftarrow \tau_k + pr + (1 - iou)$;
**end**
$(D, c, \tau) = \text{sort}(D, c, \tau)$; // sort by decreasing $\tau$
**for** $k = 1 \ldots N$ **do**
  $overlap \leftarrow \text{precision}(D_k, M_{c_k})$;
  **if** $overlap \geq thr$ **then**
    $keep[k] = True$;
    $M_{c_k} = M_{c_k} \backslash D_k$
  **else**
    $keep[k] = False$
  **end**
**end**

Time complexity = O(n)

# Experiments

Toy experiment (isolate the spatial hedging problem)



(a) (b) (c)

| Model | CoordConv | $AP_{50}$ | $F1_{0.5}$ | LRP | $LRP_{Loc}$ |
|---|---|---|---|---|---|
| SOLOv2 | ✗ | 96.87 | 0.47 | 79.65 | 16.55 |
| SOLOv2 | ✓ | 96.90 | 0.46 | 79.87 | 16.06 |
| Ours | ✗ | **98.01** | **0.99** | 33.46 | 15.87 |
| Ours | ✓ | **98.01** | **0.99** | **33.37** | **15.75** |

# Experiments

Performance on COCO dataset (Ours = SOLOv2 + Semantic NMS and Sorting)

| Method | Spatial hedging | | | Mask quality | | Category hedging | AP↑ | LRP↓ |
|---|---|---|---|---|---|---|---|---|
| | DC↓ | $\text{LRP}_{\text{FP}}$↓ | F1↑ | b-IoU↑ | $\text{LRP}_{\text{Loc}}$↓ | NE↓ | | |
| **ResNet-50-FPN** | | | | | | | | |
| Mask-RCNN | 76.1 | **80.3** | **38.4** | 49.6 | 20.6 | **0.63** | 37.2 | **88.4** |
| SOLOv2 | 64.1 | 90.4 | 20.8 | 49.8 | 20.6 | 1.13 | **37.6** | 94.4 |
| HTC | 62.3 | 93.9 | 23.3 | 49.9 | **20.4** | 2.19 | 37.4 | 96.3 |
| QueryInst (100 queries) | **14.9** | 95.1 | 17.1 | 16.9 | 20.6 | 2.78 | **37.5** | 97.1 |
| CondInst | 144.1 | 88.1 | 30.7 | **50.2** | 20.5 | 1.35 | 37.4 | 92.9 |
| Ours | **2.0** | **78.1** | **43.3** | **50.5** | **20.1** | **0.94** | 34.7 | **87.6** |
| **ResNet-101-FPN** | | | | | | | | |
| Mask-RCNN | 62.6 | **77.5** | **41.7** | 50.4 | 20.0 | **0.56** | 38.6 | **86.6** |
| SOLOv2 | 63.1 | 89.5 | 21.6 | 50.8 | 20.0 | 1.05 | 39.0 | 93.7 |
| HTC | 48.3 | 92.7 | 26.4 | **51.1** | 20.0 | 1.98 | **39.6** | 95.5 |
| QueryInst (100 queries) | **10.9** | 94.7 | 19.9 | 17.0 | **19.6** | 2.64 | **41.0** | 96.7 |
| CondInst | 126.2 | 86.1 | 33.5 | 50.9 | 20.2 | 1.17 | 38.5 | 91.6 |
| Ours | **1.9** | **70.6** | **45.9** | **51.4** | **19.2** | **0.57** | 37.4 | **83.4** |

Indicates best result   Indicates second best result

# Experiments

Ablation of different NMS techniques.

| Method | NMS | Spatial hedging | | | Mask quality | | Category hedging | AP↑ | LRP↓ |
|---|---|---|---|---|---|---|---|---|---|
| | | DC↓ | $LRP_{FP}$↓ | F1↑ | b-IoU↑ | $LRP_{Loc}$↓ | NE↓ | | |
| SOLOv2 | Matrix | 55.6 | 92.61 | 18.46 | 43.0 | 22.43 | 1.93 | 26.34 | 95.88 |
| SOLOv2 | Mask | **16.0** | 88.52 | 29.82 | 42.9 | 22.13 | 1.56 | 26.16 | 93.54 |
| Ours | Matrix | 63.5 | 91.99 | 17.87 | 44.1 | 22.44 | 1.68 | **28.15** | 95.56 |
| Ours | Mask | 17.4 | **86.76** | **30.82** | **44.3** | **22.12** | **1.33** | **27.94** | **92.60** |
| Ours | Semantic | **2.3** | **79.29** | **36.05** | **44.7** | **21.84** | **0.98** | 26.37 | **89.25** |

Indicates best result     Indicates second best result

# Results


zebra 11%
zebra 41%
zebra 74%
zebra 76%
zebra 15%
zebra 85%
zebra 29%
zebra 73%
zebra 75%
zebra 83%
sheep 18%
sheep 11%
sheep 10%
sheep 81%
sheep 19%
sheep 78%
sheep 17%
sheep 16%
sheep 9%
sheep 78%
sheep 58%
sheep 17%
sheep 80%
sheep 82%
sheep 77%
sheep 66%
sheep 25%

# Results


bench 9%
bench 50%
person 70%
person 71%
bench 50%
clock 70%
tv 67%
chair 15%
bed 61%
chair 58%
chair 65%
couch 23%
oven 25%
bed 17%
clock 72%
tv 67%
bench 10%
chair 74%
oven 21%
bed 61%
couch 14%

# Summary

**mAP:**
•penalizes high confidence FPs ✅
•doesnt penalize trailing low-confidence FPs 📉
- can reward “accidental TPs” → promotes hedging 😩

**Need to capture and quantify this behavior!**

**DC:** Confidence-weighted overlap of the network outputs 🤔
**NE**: interclass labelling confusion 🙈
**F1, LRP**: counting metrics (FPs, FNs) 1️⃣2️⃣3️⃣

Proposed Semantic NMS+Sorting provides a great tradeoff! ✨

Questions? 🙋‍♀️🙋‍♂️