Panoptic Segmentation: Unifying Semantic and Instance Segmentation
Unifying Semantic and Instance Segmentation
Unifying Semantic and Instance Segmentation
Unifying Semantic and Instance Segmentation

Semantic Segmentation

Object Detection
Unifying Semantic and Instance Segmentation

Semantic Segmentation

Object Detection/Seg
Unifying Semantic and Instance Segmentation

Semantic Segmentation

- per-pixel annotation
- simple accuracy measure
- instances indistinguishable

Object Detection/Seg
Unifying Semantic and Instance Segmentation

**Semantic Segmentation**
- per-pixel annotation
- simple accuracy measure
- instances indistinguishable

**Object Detection/Seg**
- each object detected and segmented separately
- “stuff” is not segmented
Unifying Semantic and Instance Segmentation

**Semantic Segmentation**
- per-pixel annotation
- simple accuracy measure
- instances indistinguishable

**Object Detection/Seg**
- each object detected and segmented separately
- “stuff” is not segmented
Unifying Semantic and Instance Segmentation

**Semantic Segmentation**
- per-pixel annotation
- simple accuracy measure
- instances indistinguishable

**Panoptic Segmentation**

**Object Detection/Seg**
- each object detected and segmented separately
- “stuff” is not segmented
Outline

➢ Motivation

➢ Problem Definition

➢ Quality Evaluation

➢ Human Performance

➢ Humans vs Computers

➢ Perspectives
Panoptic Segmentation

For each pixel $i$ predict semantic label $l$ and instance id $z$
Panoptic Segmentation

For each pixel $i$ predict semantic label $l$ and instance id $z$

- no overlaps between segments
Panoptic Segmentation

For each pixel $i$ predict semantic label $l$ and instance id $z$

- no overlaps between segments

- Popular datasets can be used
- We introduce simple, intuitive metric
- Drive novel algorithmic ideas
Popular datasets can be used

For each pixel $i$ predict semantic label $l$ and instance id $z$

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Instance Segmentation</th>
<th>Semantic Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO*</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>ADE20k/Places</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>CityScapes</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Mapillary Vistas</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

*COCO has overlaps (no depth order)*
Outline

- Motivation
- Problem Definition
- Quality Evaluation
- Human Performance
- Humans vs Computers
- Perspectives
Quality Evaluation

Ground Truth

Prediction
Quality Evaluation

Theorem: Matching is unique if overlapping threshold $> 0.5$ IoU and both ground truth and prediction have no overlaps.

Proof sketch:

If $\text{IoU} > 0.5$

then there is no other non overlapping object that has $\text{IoU} > 0.5$. 
Quality Evaluation

Ground Truth

Prediction

\[ TP_1 = \{ (\text{blue box}, \text{blue box}), (\text{red box}, \text{red box}) \} \]

\[ FP_1 = \{ \text{yellow box} \} \]

\[ FN_1 = \{ \text{gray box} \} \]
Quality Evaluation

TP₁ = \{(\text{blue}), (\text{orange})\}, (\text{orange}), (\text{orange}), (\text{orange})\}

FP₁ = \{\text{yellow}\}

FN₁ = \{\text{white}\}

\[
\text{PSQ}_1 = \frac{\text{IoU}(\text{blue}, \text{blue}) + \text{IoU}(\text{orange}, \text{orange})}{|TP₁| + |FP₁| + |FN₁|} \quad = \quad \frac{\sum_{(g,p)\in TP₁} \text{IoU}(g,p)}{|TP₁| + |FP₁| + |FN₁|}
\]
Quality Evaluation

Quality Evaluation

Ground Truth

Prediction

\[
PSQ_{l} = \frac{\sum_{(g,p) \in TP_1} \text{IoU}(g,p)}{|TP_1| + |FP_1| + |FN_1|} = \frac{\sum_{(g,p) \in TP_1} \text{IoU}(g,p)}{|TP_1|} \cdot \frac{|TP_1|}{|TP_1| + |FP_1| + |FN_1|}
\]

Segmentation Quality

Detection Quality
Outline

- Motivation
- Problem Definition
- Quality Evaluation
- Human Performance
- Humans vs Computers
- Perspectives
Panoptic Segmentation Quality (PSQ)

\[
PSQ_1 = \frac{\sum_{(g,p) \in TP_1} \text{IoU}(g,p)}{|TP_1| + |FP_1| + |FN_1|} = \frac{\sum_{(g,p) \in TP_1} \text{IoU}(g,p)}{|TP_1|} \cdot \frac{|TP_1|=1}{|TP_1| + |FP_1| + |FN_1|}
\]

Seg Quality

Det Quality
Panoptic Segmentation Quality (PSQ)

$\text{PSQ}_i = \frac{\sum_{(g,p) \in TP_1} \text{IoU}(g,p)}{|TP_1| + |FP_1| + |FN_1|}$

- **Seg Quality**
  \(\frac{\sum_{(g,p) \in TP_1} \text{IoU}(g,p)}{|TP_1|}\)

- **Det Quality**
  \(\frac{|TP_{i=1}|}{|TP_1| + |FP_1| + |FN_1|}\)

no confidence scores

human performance can be measured
Panoptic Segmentation Quality (PSQ)

\[ PSQ_1 = \frac{\sum_{(g,p) \in TP_1} IoU(g,p)}{|TP_1| + |FP_1| + |FN_1|} = \frac{\sum_{(g,p) \in TP_1} IoU(g,p)}{|TP_1|} \cdot \frac{|TP_1|}{|TP_1| + |FP_1| + |FN_1|} \]

\(\sum_{(g,p) \in TP_1} IoU(g,p)\) \quad Seg Quality

\(\sum_{(g,p) \in TP_1} IoU(g,p)\) \quad Det Quality

CityScapes: 30 images were annotated independently twice.

no confidence scores
human performance
\(\Rightarrow\)
can be measured
Panoptic Segmentation Quality (PSQ)

\[ PSQ_1 = \frac{\sum_{(g,p) \in TP_1} IoU(g,p)}{|TP_1| + |FP_1| + |FN_1|} \]

CityScapes: 30 images were annotated independently twice.

<table>
<thead>
<tr>
<th>class</th>
<th>PSQ</th>
<th>Seg Quality</th>
<th>Det Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>66.6%</td>
<td>87.5%</td>
<td>76.2%</td>
</tr>
<tr>
<td>person</td>
<td>61.8%</td>
<td>80.8%</td>
<td>76.4%</td>
</tr>
<tr>
<td>motorcycle</td>
<td>51.8%</td>
<td>77.8%</td>
<td>66.7%</td>
</tr>
<tr>
<td>pole</td>
<td>46.9%</td>
<td>70.3%</td>
<td>66.7%</td>
</tr>
<tr>
<td>road</td>
<td>98.0%</td>
<td>98.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>traffic sign</td>
<td>67.1%</td>
<td>79.5%</td>
<td>84.4%</td>
</tr>
<tr>
<td>average</td>
<td>62.6%</td>
<td>83.9%</td>
<td>73.43%</td>
</tr>
</tbody>
</table>

All Objects

no confidence scores
human performance can be measured
Panoptic Segmentation Quality (PSQ)

\[
PSQ_1 = \frac{\sum_{(g,p) \in TP_1} \text{IoU}(g,p)}{|TP_1| + |FP_1| + |FN_1|} = \frac{\sum_{(g,p) \in TP_1} \text{IoU}(g,p)}{|TP_1|} \times \frac{|TP_1|}{|TP_1| + |FP_1| + |FN_1|}
\]

Seg Quality
Det Quality

CityScapes: 30 images were annotated independently twice.

<table>
<thead>
<tr>
<th>class</th>
<th>PSQ</th>
<th>Seg Quality</th>
<th>Det Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>89.4%</td>
<td>91.3%</td>
<td>97.9%</td>
</tr>
<tr>
<td>person</td>
<td>82.0%</td>
<td>78.1%</td>
<td>94.1%</td>
</tr>
<tr>
<td>motorcycle</td>
<td>68.8%</td>
<td>79.4%</td>
<td>86.7%</td>
</tr>
<tr>
<td>pole</td>
<td>48.2%</td>
<td>70.3%</td>
<td>68.6%</td>
</tr>
<tr>
<td>road</td>
<td>98.0%</td>
<td>98.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>traffic sign</td>
<td>74.0%</td>
<td>79.5%</td>
<td>93.1%</td>
</tr>
<tr>
<td>average</td>
<td>68.7%</td>
<td>85.1%</td>
<td>80.1%</td>
</tr>
</tbody>
</table>

Objects > 32^2
Human Annotation Flaws

Classification Flaws
Human Annotation Flaws

Segmentation Flaws
Outline

- Motivation
- Problem Definition
- Quality Evaluation
- Human Performance
- Humans vs Computers
- Perspectives
Mask R-CNN + PSPNet Combination Heuristic

Mask R-CNN[1] → instances

PSPNet[2] → semantic scores

panoptic prediction

Mask R-CNN Non-overlapping Instances

Mask R-CNN output
Mask R-CNN filtered
Non-overlapping Instances
Ground Truth
<table>
<thead>
<tr>
<th></th>
<th>PSQ avg.</th>
<th>Seg Quality avg.</th>
<th>Det Quality avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans</td>
<td>62.6%</td>
<td>83.9%</td>
<td>73.43%</td>
</tr>
<tr>
<td>Mask R-CNN + PSPNet</td>
<td>51.7%</td>
<td>81.0%</td>
<td>62.01%</td>
</tr>
</tbody>
</table>
# PSQ – Humans vs Computers

<table>
<thead>
<tr>
<th></th>
<th>PSQ avg.</th>
<th>Seg Quality avg.</th>
<th>Det Quality avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans</td>
<td>62.6%</td>
<td>83.9%</td>
<td>73.43%</td>
</tr>
<tr>
<td>Mask R-CNN + PSPNet</td>
<td>51.7%</td>
<td>81.0%</td>
<td>62.01%</td>
</tr>
</tbody>
</table>

![PSQ Distribution](image)

- **Humans**: Green
- **Heuristic combination of Mask R-CNN and PSPNet**: Blue
## PSQ – Humans vs Computers

<table>
<thead>
<tr>
<th></th>
<th>PSQ avg.</th>
<th>Seg Quality avg.</th>
<th>Det Quality avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Humans</strong></td>
<td>62.6%</td>
<td>83.9%</td>
<td>73.43%</td>
</tr>
<tr>
<td><strong>Mask R-CNN + PSPNet</strong></td>
<td>51.7%</td>
<td>81.0%</td>
<td>62.01%</td>
</tr>
</tbody>
</table>

![IoU (semantic only)](image)

- **Humans**
- **Heuristic combination of Mask R-CNN and PSPNet**
Outline

➤ Motivation

➤ Problem Definition

➤ Quality Evaluation

➤ Human Performance

➤ Humans vs Computers

➤ Perspectives
Why solve it?

**Semantic Segmentation**
- per-pixel annotation
- simple accuracy measure
- instances indistinguishable

**Panoptic Segmentation**

**Object Detection/Seg**
- each object detected and segmented separately
- “stuff” is not segmented
Why solve it?

**Semantic Segmentation**
- per-pixel annotation
- simple accuracy measure
- instances indistinguishable

**Panoptic Segmentation**

**Object Detection/Seg**
- each object detected and segmented separately
- “stuff” is not segmented

FCN 8s, Dilation8, DeepLab, PSPNet, RefineNet, U-Net, etc.

Fast/er R-CNN, DeepMask, SharpMask, Mask R-CNN, FCIS, YOLO, RetinaNet, FPN, etc.
Why solve it?

Mask R-CNN → instances → semantic scores → panoptic prediction

PSPNet → semantic scores → panoptic prediction
Why solve it?
Why solve it?

- instances
- semantic scores
- FPN
- panoptic prediction
Panoptic Segmentation: Future Plans

- Panoptic Segmentation paper on ArXiv
- Efficient evaluation code on GitHub
- Possible competition(s)

Panoptic COCO

Panoptic CityScapes