UPerNet

Unified Perceptual Parsing Network

Lukas Göbl, Peer Schäfer, Lukas Scheib

03.07.2025

Computer Vision Journal Club
University of Cologne

Paper

Unified Perceptual Parsing for Scene Understanding

Tete Xiao[1]*, Yingcheng Liu[1]*, Bolei Zhou[2]*, Yuning Jiang[3], Jian Sun[4]

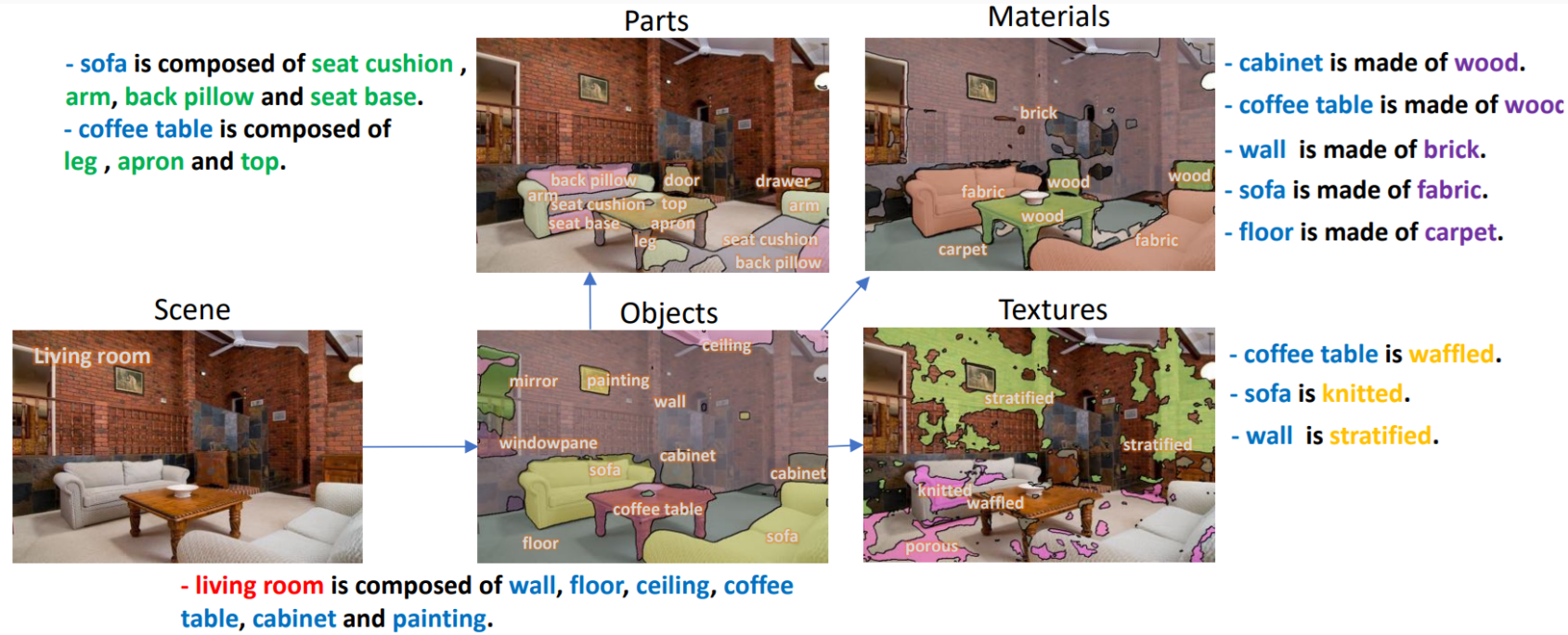
[1]Peking University [2] MIT CSAIL [3] Bytedance Inc. [4] Megvii Inc.
* indicates equal contribution.
{jasonhsiao97, liuyingcheng}@pku.edu.cn,
bzhou@csail.mit.edu, jiangyuning@bytedance.com,
sunjian@megvii.com

[1] Xiao et al. (2018)

What is Unified Perceptual Parsing?

- Humans perceive a scene by identifying:
  - Scene type (e.g., kitchen)
  - Objects (e.g., stove, table)
  - Objects parts (e.g., table leg, stove knob)
  - Materials (metal, wood)
  - Textures (smooth, rough)
- **Unified Perceptual Parsing** aims to do all this with a single model.

What is Unified Perceptual Parsing?



[1] Xiao et al. (2018), Figure 1

Why Is This Challenging?

- Perceptual parsing combines very different tasks:
  - Scene classification (high-level understanding)
  - Object and part segmentation (mid-level)
  - Material and texture recognition (low-level, detailed)
- These tasks vary in:
  - The type of information they require
  - Spatial scale and detail level
- **Challenge:** One model must handle all levels — from big picture to fine details.

Previous Work

- **Scene Classification:** CNNs like ResNet, trained on datasets like Places
- **Semantic Segmentation:** FCN, DeepLab (object-level)
- **Part Segmentation:** Pascal-Parts
- **Material / Texture:** OpenSurfaces, DTD

Before This Paper:

- Each task was handled by a separate model
- No feature sharing between tasks
- Training and inference were inefficient and fragmented

Main Contribution of This Paper

Objective: Develop a single unified network for parsing multiple semantic levels from one image.

- Create **Broden+** – unified dataset from various sources
- Propose **UPerNet** – a multi-task network on top of FPN
- Define a **joint benchmark** for perceptual parsing

One image in, multiple perceptual layers out.

Why a new Dataset?

No single existing dataset had all the annotations needed for unified perceptual parsing.

Limitations of existing datasets:

- Each focuses on **only one or two tasks** (e.g., objects OR textures)
- Different label sets, formats, and definitions
- No shared vocabulary across datasets (e.g. 'car' in one dataset might be 'vehicle' in another)

Manual merging needed to create a unified dataset that supports all tasks.

How Was Broden+ Merged?

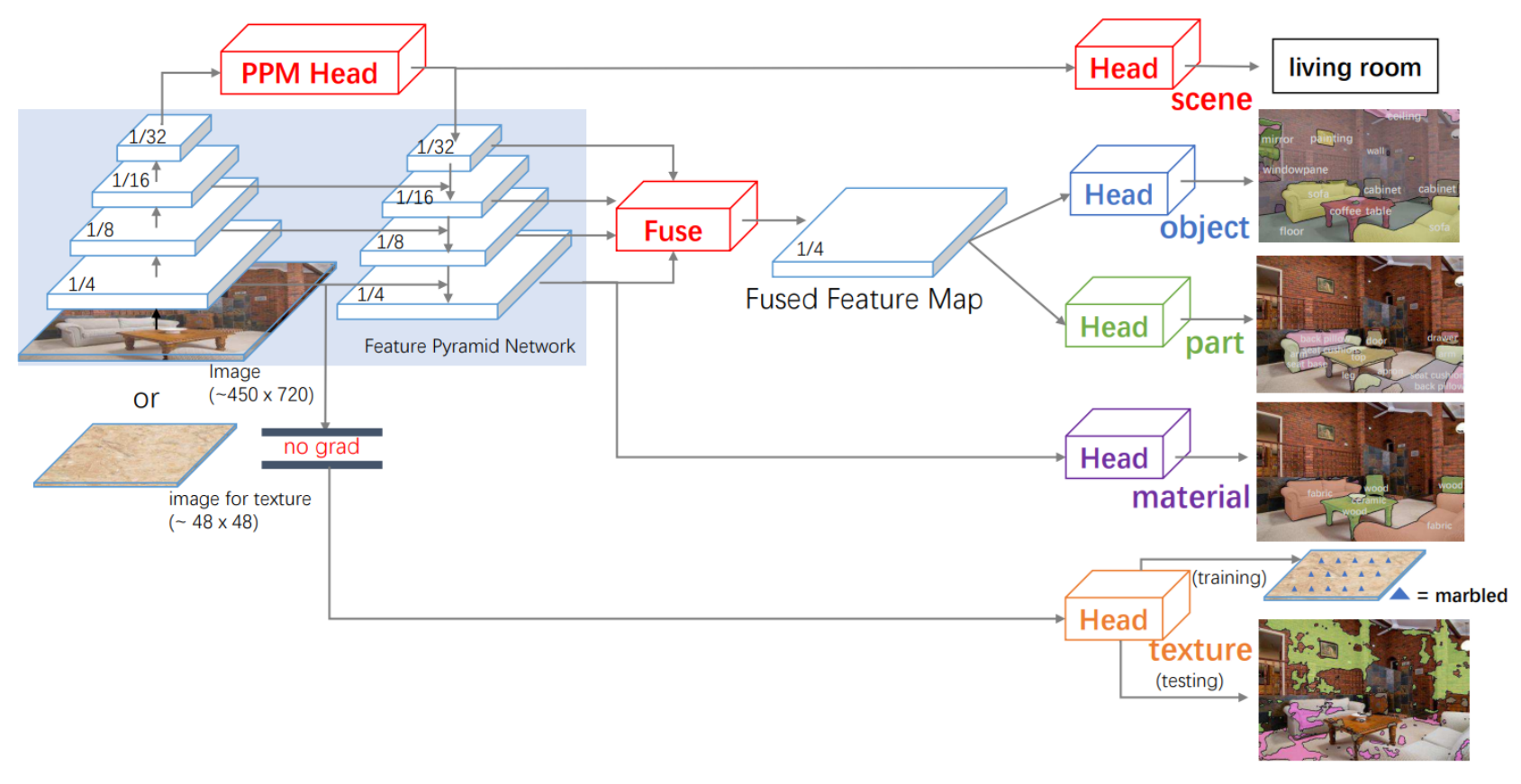
Merging Criteria:

- All annotations are **mapped to a shared vocabulary**
- Images are labeled with **as many tasks as possible**
- Only high-quality, pixel-aligned annotations are kept
- Some rare classes were **merged into broader categories** to reduce noise and improve consistency (e.g 'stone' and 'concrete' are merged into 'stone')

What is Broden+

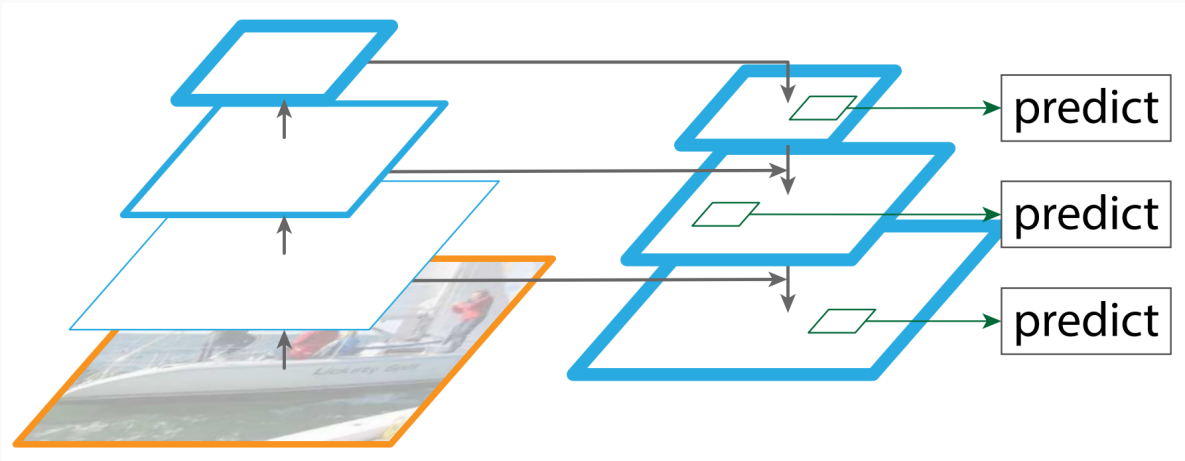
- **Broden+** is a large-scale, unified dataset built by merging:
  - ADE20K: Scene and object segmentation
  - Pascal-Context: Part segmentation
  - OpenSurfaces: Material labels
  - DTD (Describable Textures Dataset): Texture attributes
- Provides a **diverse set of annotations** per image:
  - Scene labels, object masks, parts, textures, materials

UPerNet Architecture Overview



[1] Xiao et al. (2018), Figure 4 (edited)

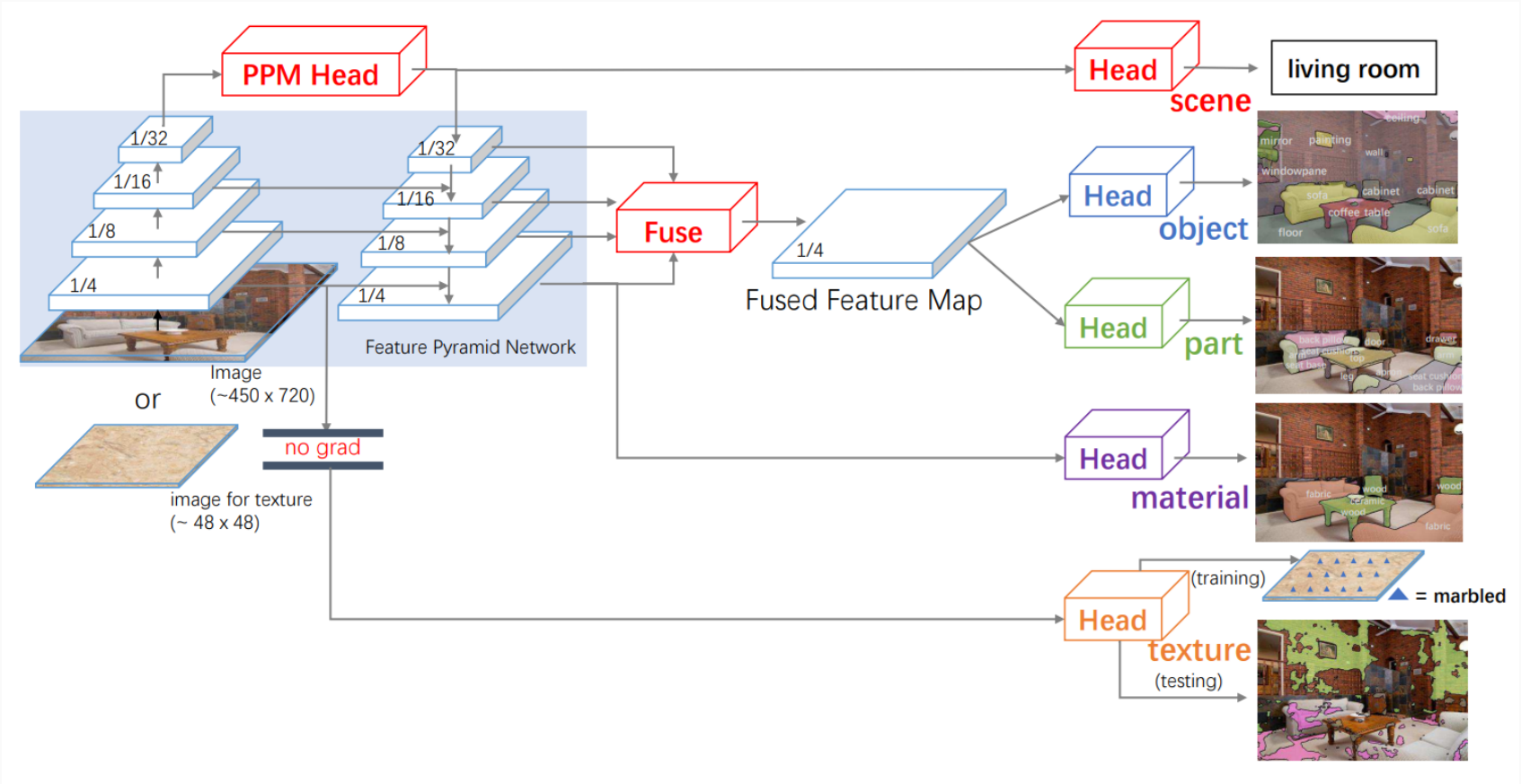
Feature Pyramid Networks (FPN)



- **Bottom-up pathway:** Computes feature maps at different scales using a convolutional NN
- **Top-down pathway:** Upsamples high-level feature maps back to original size (nearest neighbour upsampling)
- **Lateral connections:** Merges feature maps of **Top-down pathway** with **Bottom-up pathway** from same level in hierarchy

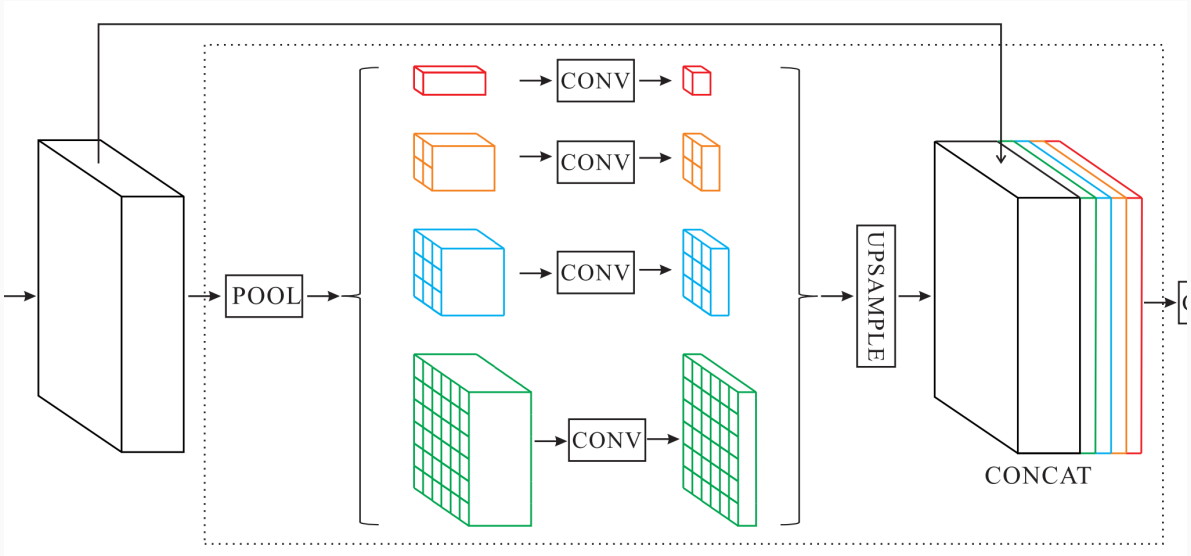
[2] Lin et al. (2017), Figure 1 (edited)

UPerNet Architecture Overview



[1] Xiao et al. (2018), Figure 4 (edited)

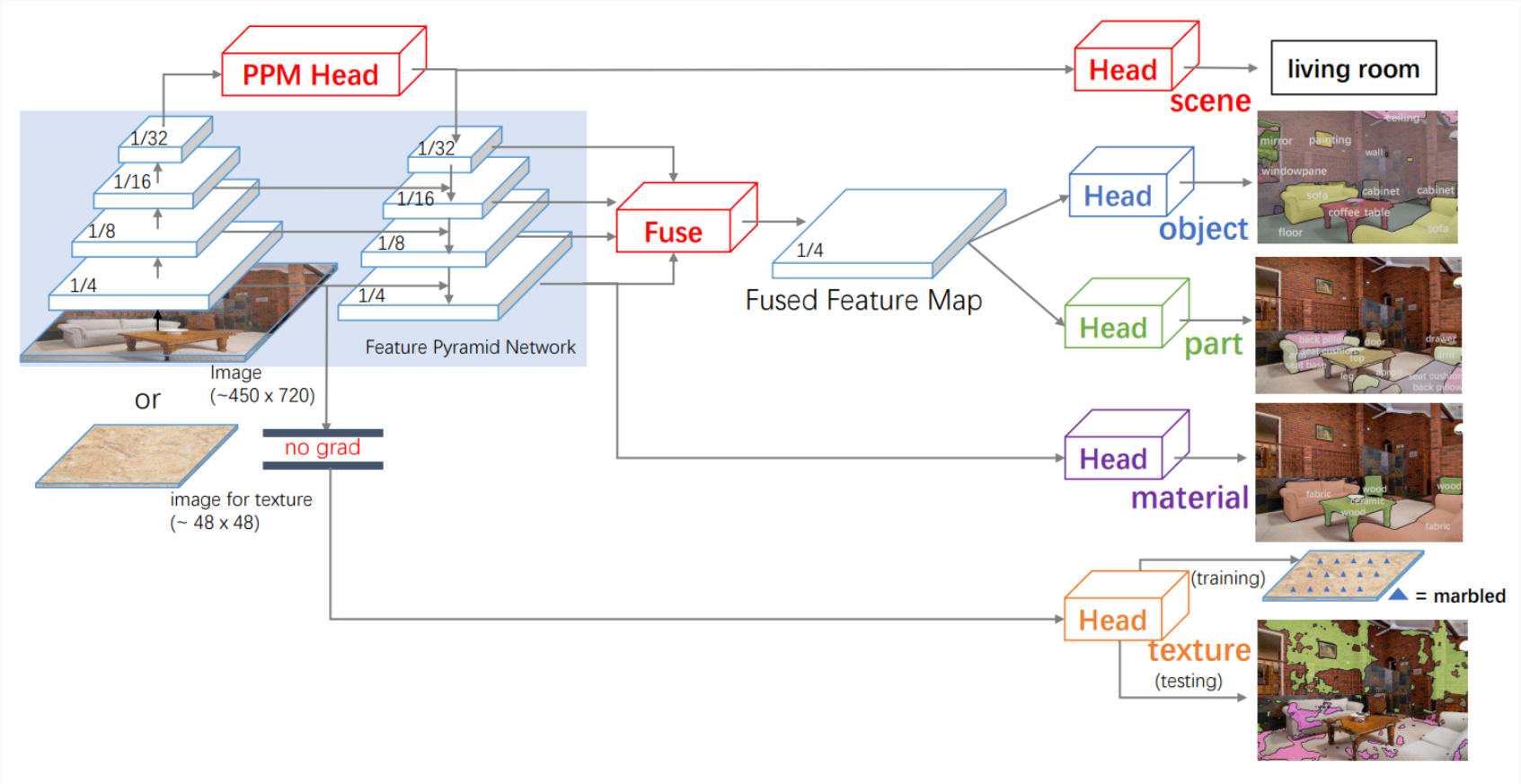
Pyramid Pooling Module



- Different pooling operations produce different sized feature maps
- Reduce dimensionality with a 1x1 convolution
- Upsample via interpolation to original size
- Concatenate results with original feature map
- PPM is used to address limited receptive field of deep CNNs

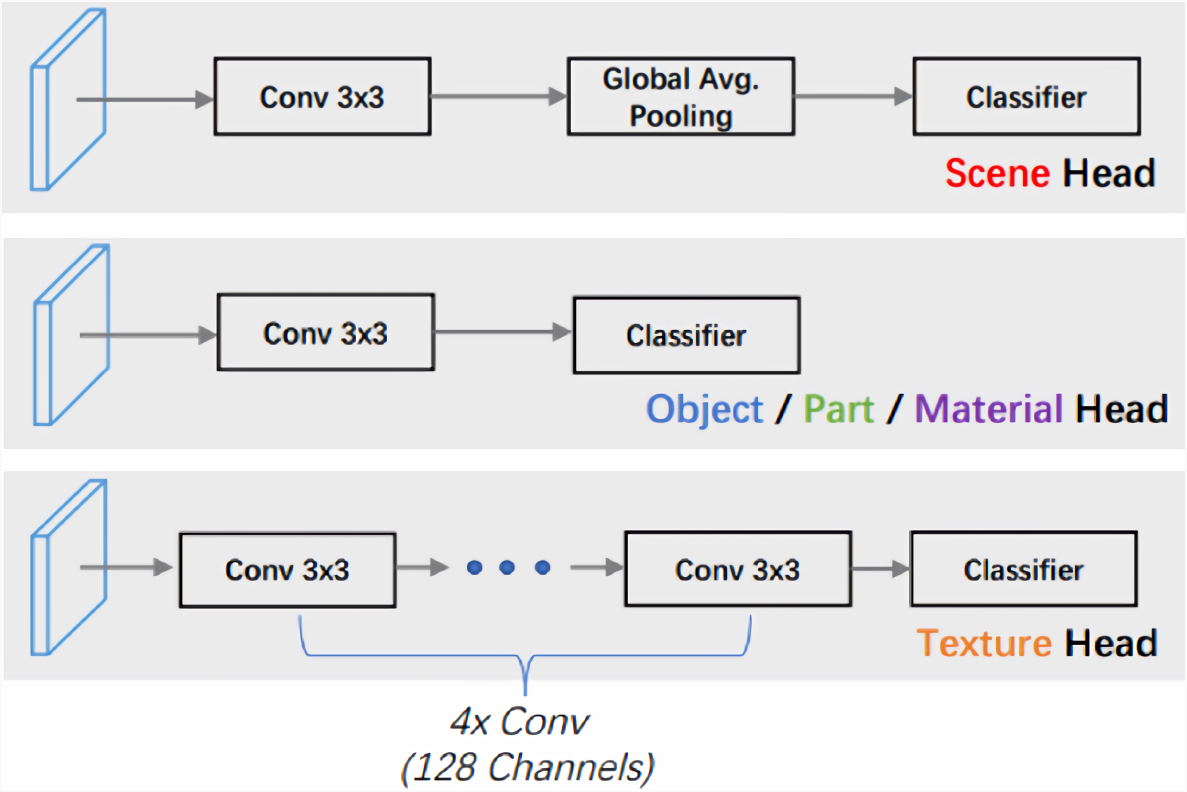
[3] Zhao et al. (2017), Figure 3 (edited)

UPerNet Architecture Overview



[1] Xiao et al. (2018), Figure 4 (edited)

Classification Heads



[1] Xiao et al. (2018), Figure 4 (edited)

Experiment 1: Standard Semantic Segmentation

Comparison with state-of-the-art methods for semantic segmentation of objects on the ADE20K dataset:

Method	Mean IoU(%)	Pixel Acc.(%)	Overall(%)	Time(hr)
FCN [11]	29.39	71.32	50.36	-
SegNet [42]	21.64	71.00	46.32	-
DilatedNet [14]	32.31	73.55	52.93	-
CascadeNet [2]	34.90	74.52	54.71	-
RefineNet (Res-152) [15]	40.70	-	-	-
DilatedNet*† (Res-50) [16]	34.28	76.35	55.32	53.9
PSPNet† (Res-50) [16]	**41.68**	**80.04**	**60.86**	61.1
FPN (/16)	34.46	76.04	55.25	18.1
FPN (/8)	34.99	76.54	55.77	20.2
FPN (/4)	35.26	76.52	55.89	21.2
FPN+PPM (/4)	40.13	79.61	59.87	27.8
FPN+PPM+Fusion (/4)	**41.22**	**79.98**	**60.60**	38.7

[1] Xiao et al. (2018), Table 2

Experiment 2: Unified Perceptual Parsing

Results of UPP on the Broden+ dataset:

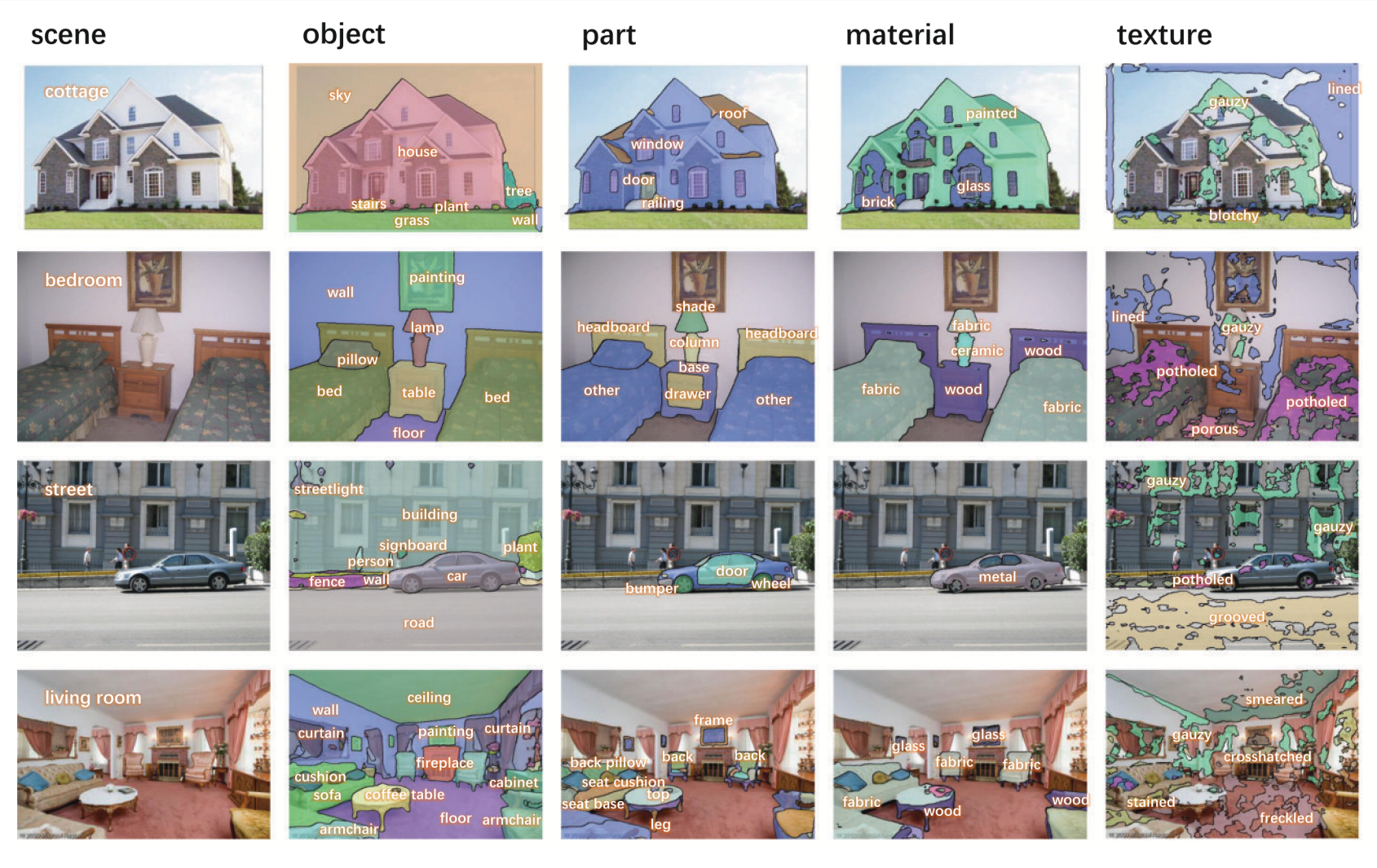
Training Data					Object		Part		Scene	Material		Texture
+O	+P	+S	+M	+T	mI.	P.A.	mI.(bg)	P.A.	T-1	mI.	P.A.	T-1
✓					24.72	78.03	-	-	-	-	-	-
			✓		-	-	-	-	-	52.78	84.32	-
✓	✓				23.92	77.48	30.21	48.30	-	-	-	-
✓	✓	✓			23.83	77.23	30.10	48.34	71.35	-	-	-
✓	✓	✓	✓		23.36	77.09	28.75	46.92	70.87	54.19	84.45	-
✓	✓	✓	✓	✓	23.36	77.09	28.75	46.92	70.87	54.19	84.45	35.10

- Joint training across 5 perceptual levels
- Performance remains stable for most tasks and even improves for material
- Texture prediction requires fine-tuning due to different data distribution

[1] Xiao et al. (2018), Table 3

Experiment 2: Unified Perceptual Parsing

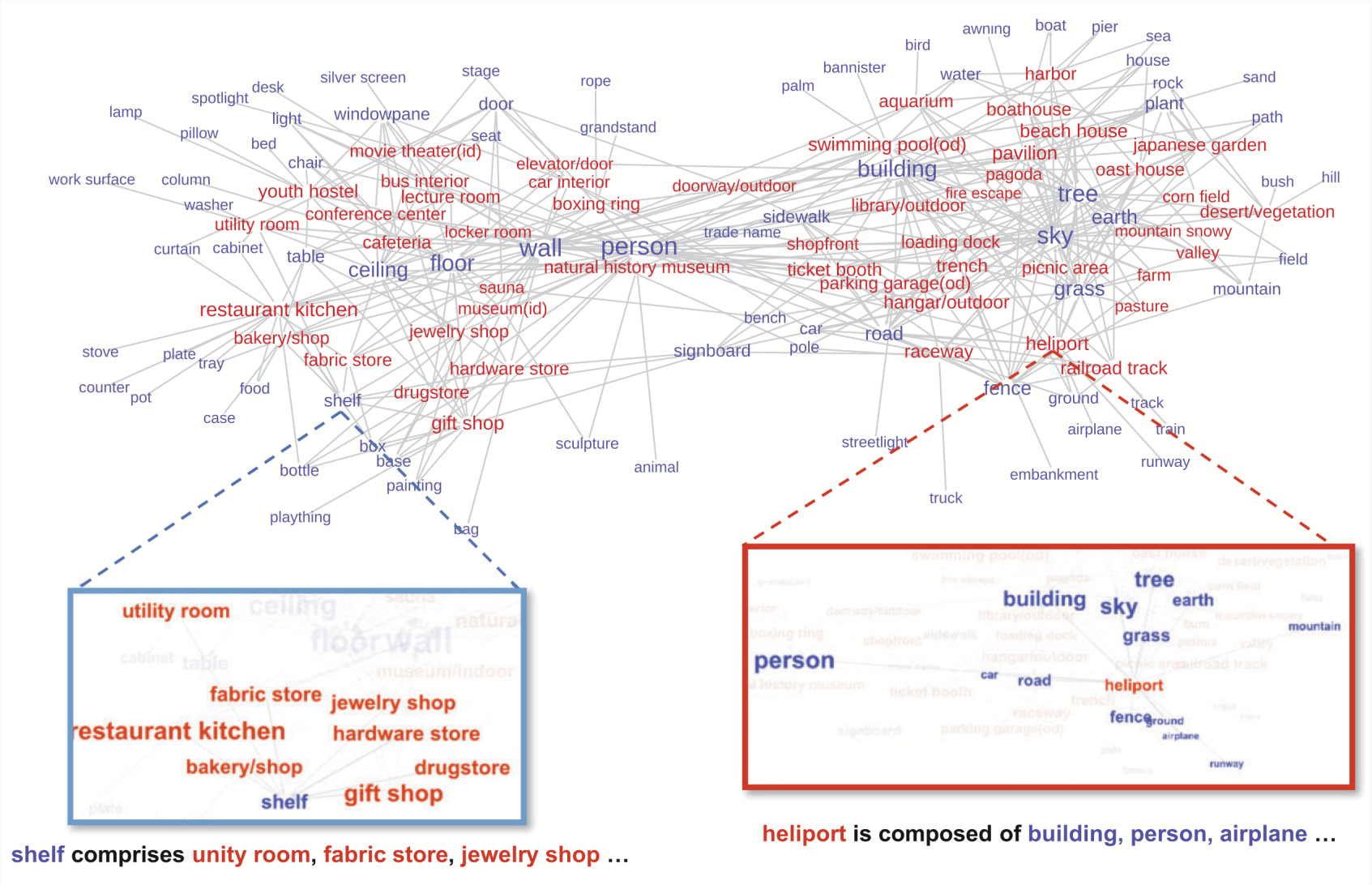
Some qualitative results:



[1] Xiao et al. (2018), Figure 5

Visual Knowledge Discovery

Visualization of scene-object relations after clustering:



shelf comprises **unity room**, **fabric store**, **jewelry shop** ...

heliport is composed of **building**, **person**, **airplane** ...

[1] Xiao et al. (2018), Figure 6a

Visual Knowledge Discovery

Scene-object Relations

garage (indoor) is composed of floor, wall, ceiling, car, door, person, building, windowpane, box, and signboard.
glacier is composed of mountain, sky, earth, tree, snow, rock, water, and person.
laundromat is composed of wall, floor, washer, ceiling, door, cabinet, person, table and signboard.

Object-material Relations

toilet is made of ceramic (65%) and plastic (35%).
microwave is made of glass (55%), and metal (45%).
sidewalk is made of tile (65%), stone (18%), and wood (17%).

Part-material Relations

coffee table top is made of wood (69%) and glass (31%).
bed headboard is made of wood (77%) and fabric (23%).
tv monitor screen is made of glass (100%).

Material-texture Relations

brick is stratified (42%), stained (34%) and crosshatched (24%) .
stone is stained (43%), potholed (31%) and matted (26%) .
mirror is gauzy (54%), crosshatched (26%) and grooved (20%) .

[1] Xiao et al. (2018), Table 4

Conclusion and Perspectives

- UPerNet performs competitively on semantic segmentation with less training time
- The network can handle heterogeneous annotations and tasks across multiple perception levels
- Allows for discovery of compositional visual knowledge (e.g. scene-object-material relations)
- Future directions:
  - Improving texture parsing and handling more complex real-world images
  - Potential for downstream applications in robotics, AR and reasoning

Open Questions

Literature

[1] T. Xiao et al., "Unified Perceptual Parsing for Scene Understanding" in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part V*, 2018, pp. 432–448, doi: 10.1007/978-3-030-01228-1_26.

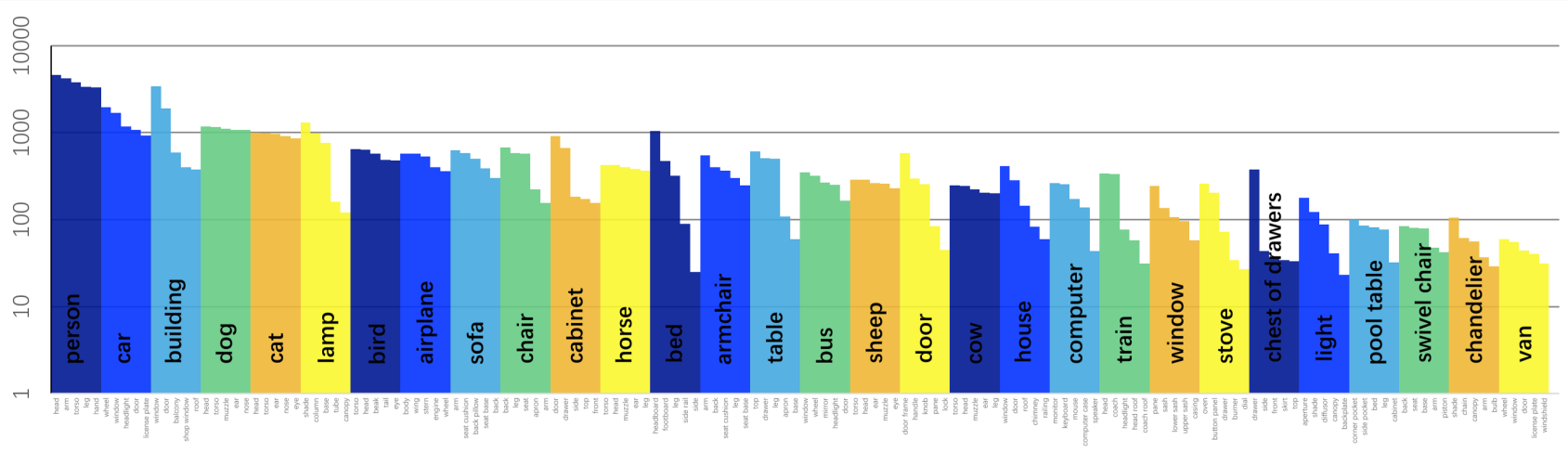
[2] T.-Y. Lin et al., "Feature Pyramid Networks for Object Detection" in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[3] H. Zhao et al., "Pyramid Scene Parsing Network" in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.

Backup

How Was Broden+ Merged?

Category	Classes	Sources	Eval. Metrics
scene	365	ADE [2]	top-1 acc.
object	335	ADE [2], Pascal-Context[27]	mIoU & pixel acc.
object w/ part	77	ADE [2], Pascal-Context[27]	-
part	152	ADE [2], Pascal-Part [28]	mIoU (bg) & pixel acc.
material	26	OpenSurfaces [6]	mIoU & pixel acc.
texture	47	DTD [4]	top-1 acc.



[1] Xiao et al. (2018), Table 1 + Figure 2b

How Is Performance Measured?

Task	Metric
Scene Classification	Top-1 Accuracy
Object Segmentation	Mean IoU (Intersection over Union)
Part Segmentation	Mean IoU
Material Prediction	Pixel Accuracy
Texture Prediction	Image-level Classification Accuracy

- Evaluated on the unified benchmark created from **Broden+**.
- **Multi-task setting:** All tasks evaluated using a shared backbone and features.

Goal: Assess how well the unified model performs across diverse visual tasks.