Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

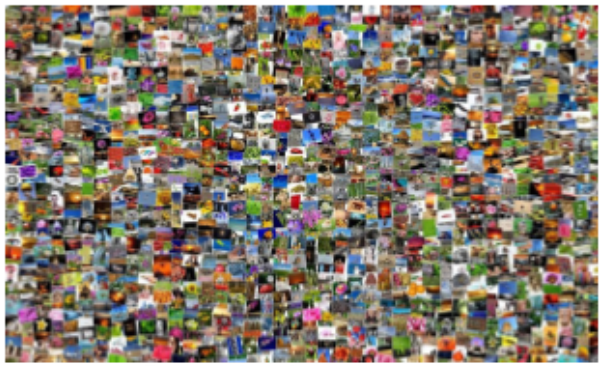
Carina Silberer

IMS, University of Stuttgart

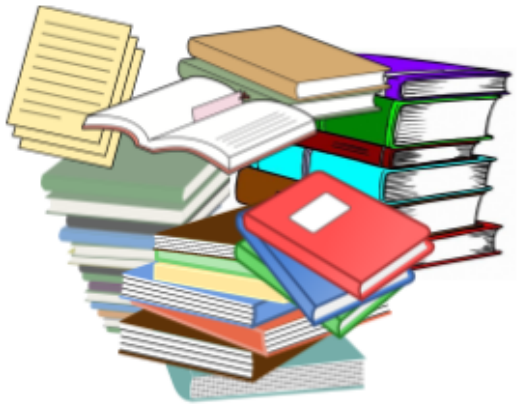
CL Fall School 2024, Passau
16 - 20 September 2024

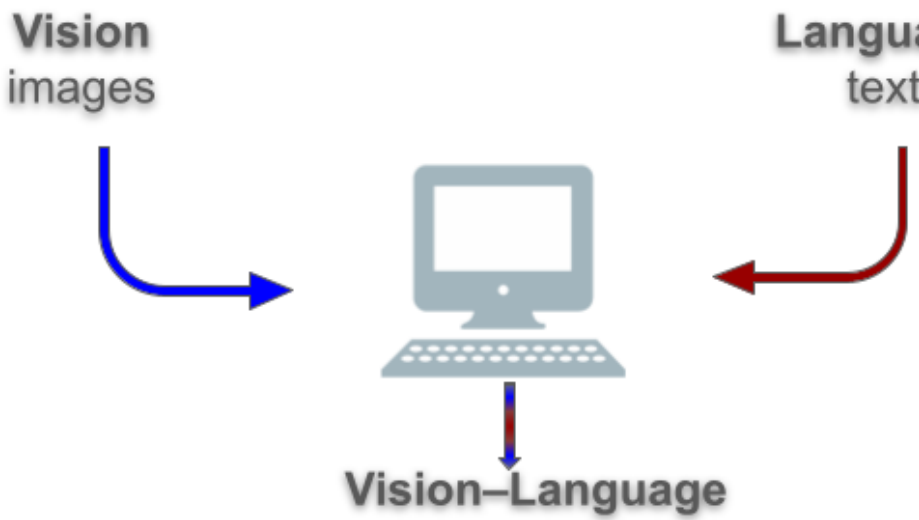
Outline

Multimodal (VL) Representations



https://pixabay.com/photos/photo-collection-pictures-photos-382018





Computer Vision: Transformer-Based Approaches

ViT: Image recognition [Dosovitskiy et al., 2021]

- End-to-end (only self-attention over pixels, no conv layers)
- The input image is split into fixed-size image patches
  e.g., 16×16 or 32×32
- The patches are mapped to vector embeddings
- Positional embeddings and special *classification token* are added (cf. NLP)

$\Rightarrow$ Feed to encoder transformer (see BERT)

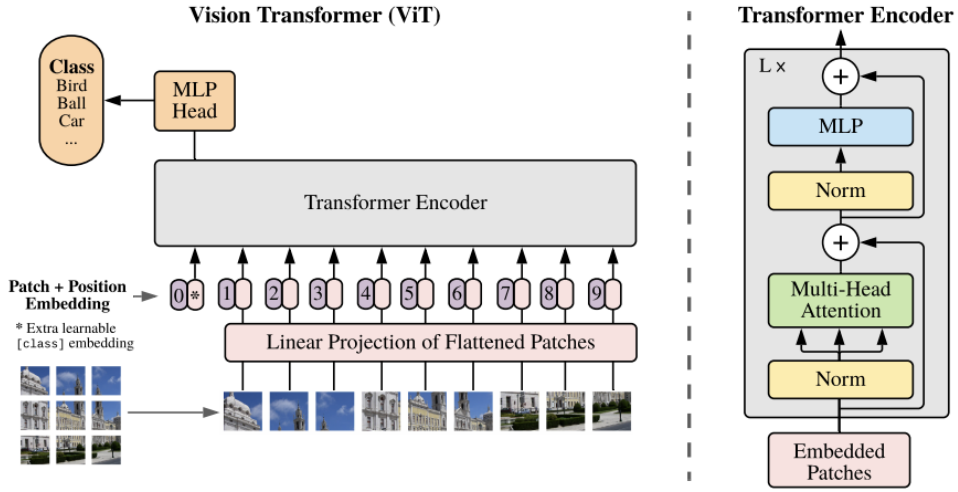


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer

Outline

APPENDIX

Deep NN Architectures for Object Detection

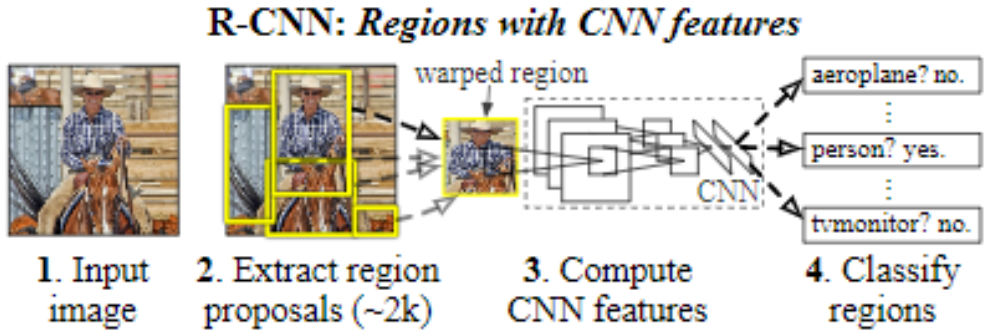
Well Known Architectures

- R-CNN
- Fast R-CNN
- Faster R-CNN
- Mask R-CNN

Blog: `blog.paperspace.com/faster-r-cnn-explained-object-detection/`

Deep NN Architectures for Object Detection

R-CNN: Three stage process

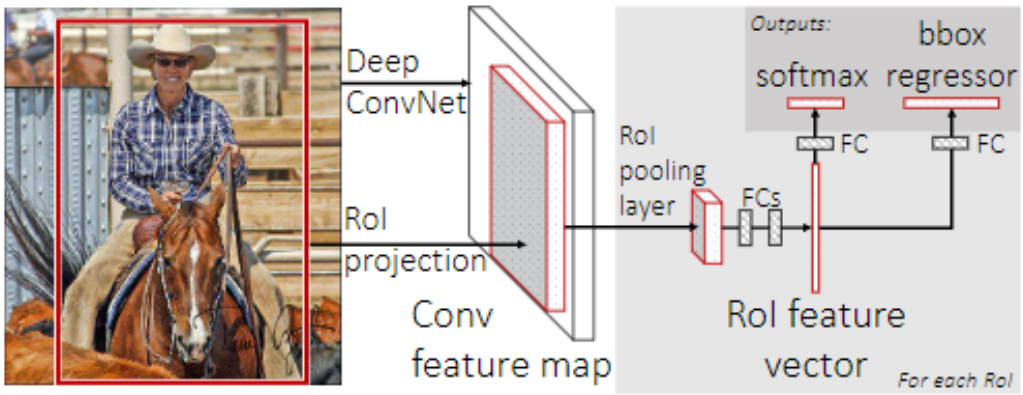


source: *Girshick et al.,2014*

- Extract possible objects using some method -> region proposals
- **Extract features with CNN**
- Use SVM for classification

Deep NN Architectures for Object Detection

Fast R-CNN

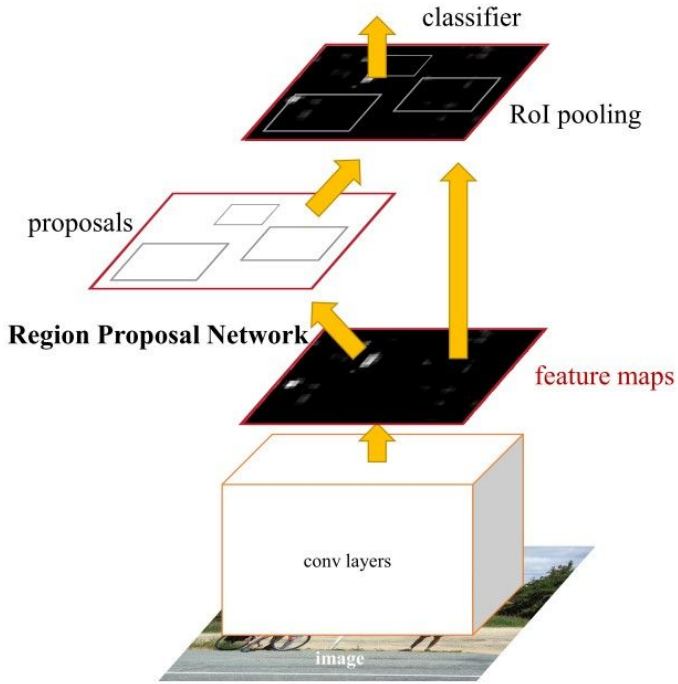


source: *Girshick, 2015*

- Extract features with CNN
- **Convolutional layer computations are shared**
- **Output layers: softmax for classification; FC to predict bounding boxes**

Deep NN Architectures for Object Detection

Faster R-CNN

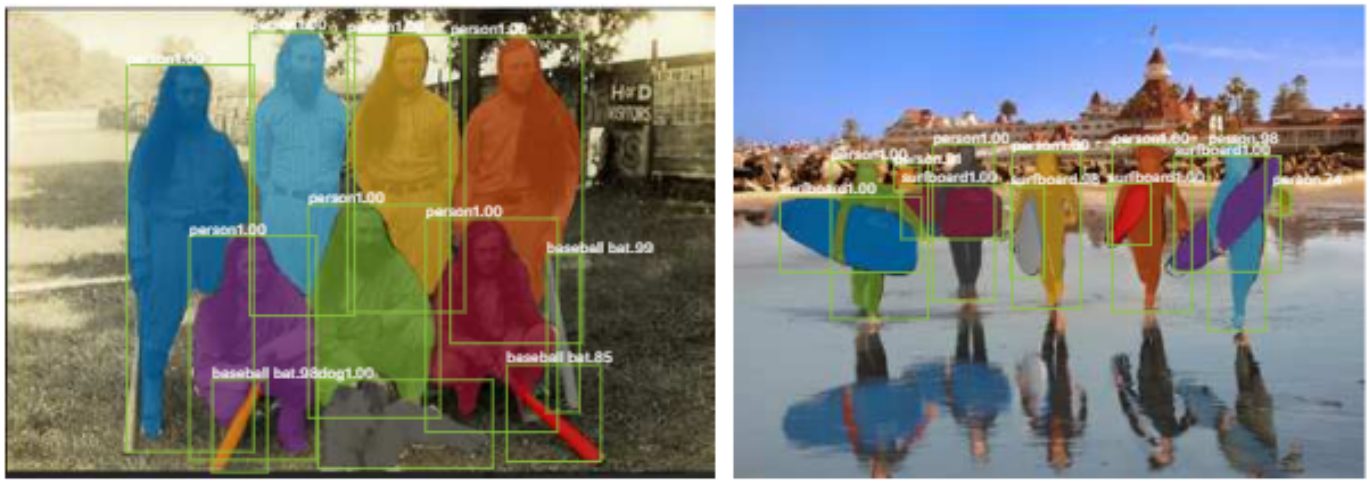


source: *Ren et al., 2015*

- **Generates Region Proposals for Objects**
- Convolutional layer computations are shared
- Output layers: softmax for classification; FC to predict bounding boxes

Deep NN for Instance Segmentation

Mask R-CNN



source: *He et al., 2017*

Outline

Computer Vision Datasets (also L&V datasets)

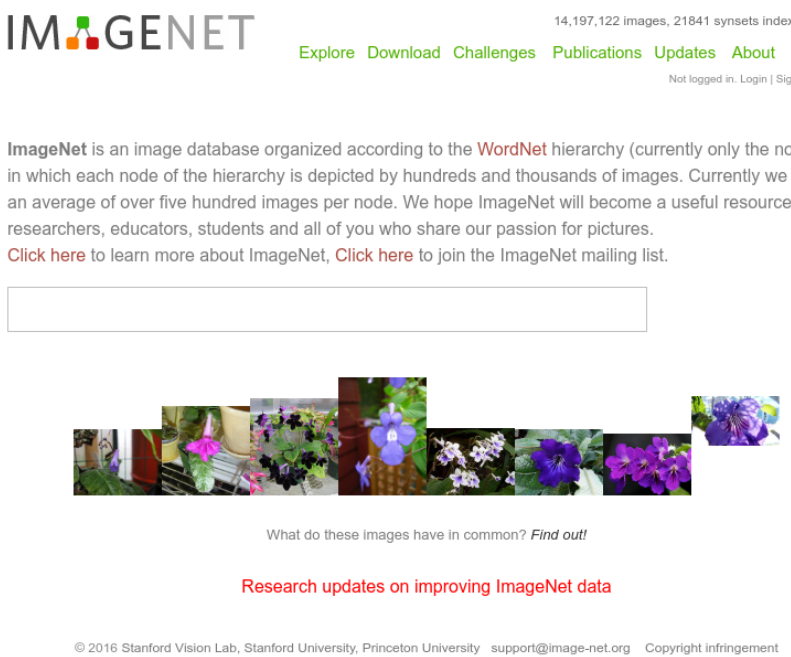
Computer Vision Datasets (also L&V datasets)

- **ImageNet** http://www.image-net.org/
- **Conceptual Captions** [?, ?]
- **MS COCO** [?]
  Human-elicited captions for "Common Objects in Context"
  https://cocodataset.org/
- **Visual Genome** [?]
  Human-elicited descriptions of image regions
  https://visualgenome.org/
- **SBU Captions** [?]
- OpenImages storage.googleapis.com/openimages/web/index.html
- PASCAL http://host.robots.ox.ac.uk/pascal/VOC/
- Scenes (indoor) web.mit.edu/torralba/www/indoor.html
  SUN scenes groups.csail.mit.edu/vision/SUN/
- Attributes
  of celebrities (faces) http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
  of animals https://cvml.ist.ac.at/AwA2
- Pets http://www.robots.ox.ac.uk/~vgg/data/pets

Computer Vision Datasets (also L&V datasets)

ImageNet `http://www.image-net.org/`

- localisation, detection, image classification, segmentation, ...
- Standard dataset for pre-training models



L&V Datasets

Conceptual Captions [?, ?]

- Automatically processed alt-text into clean image captions
- Used for training visual-language transformer models (Week 7)

Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

Conceptual Captions: a worker helps to clear the debris.

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

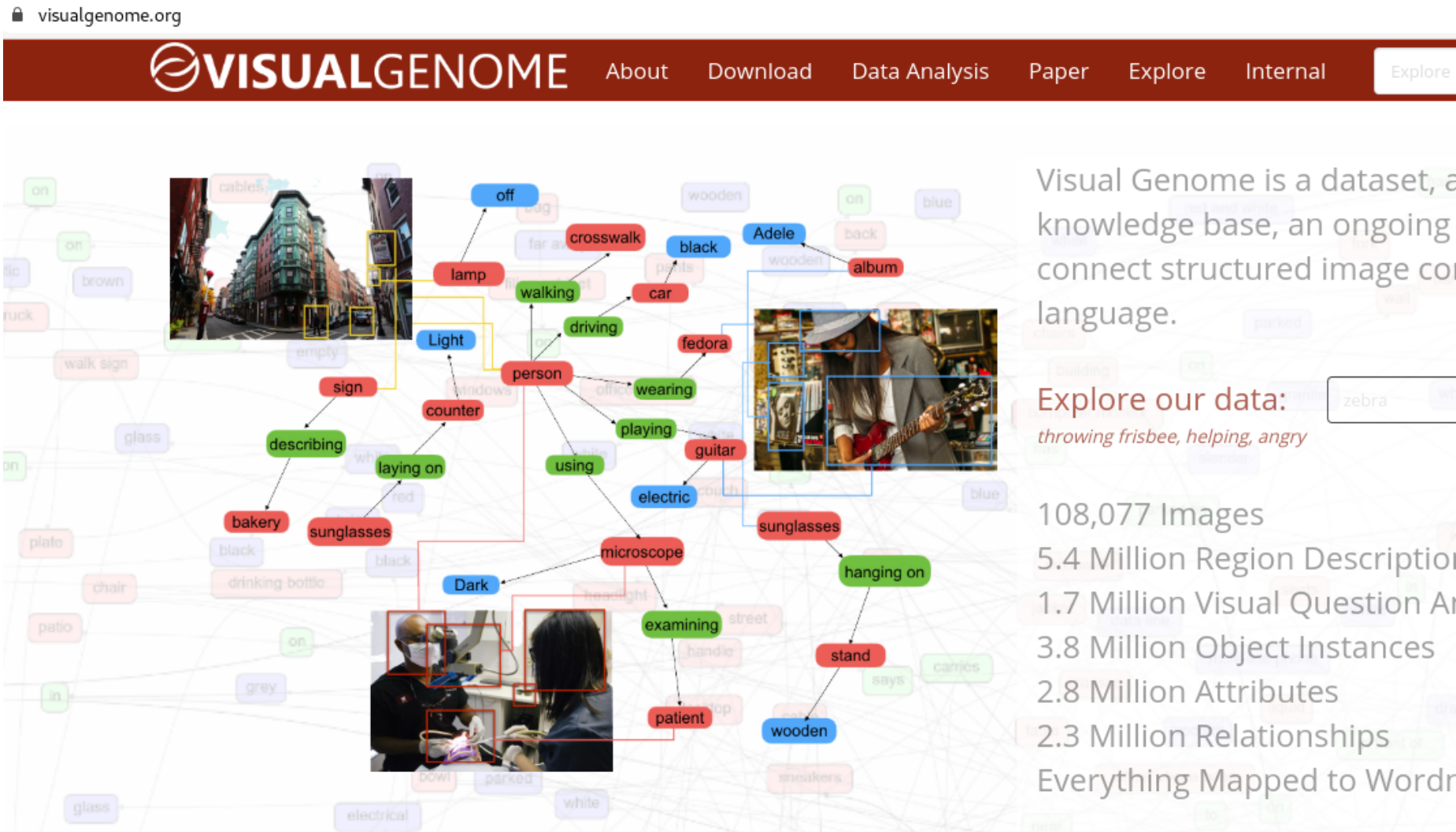
Conceptual Captions: pop artist performs at the festival in a city.

Link to ConceptualCaptions

Computer Vision Datasets (also L&V datasets)

Visual Genome

- region description, Q&A, detection, attributes ...

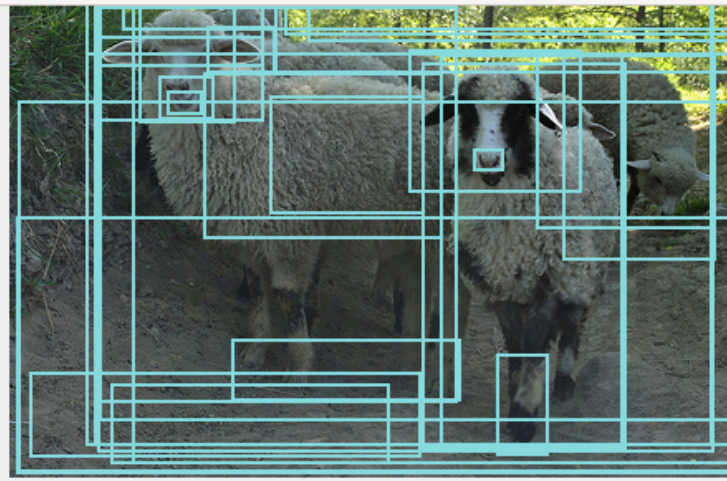


Computer Vision Datasets (also L&V datasets)

Visual Genome

visualgenome.org/VGViz/explore?query=sheep%20eats

a group of white sheep	sheep is white	group OF sheep
a black and white sheep	sheep is black	face OF sheep
the face of a white sheep	face is white	sheep looking down at ground
a sheep looking at the ground	trees is green	trees are in sunlight
green trees in the sunlight	terrain is sandy	sheep are in shade
a group of sheep under the shade	sheep is walking	sheep has ears
sandy terrain	ground is dirt	sheep has wool
a sheep's ear	sheep is looking down	sheep has hooves
		sheep are walking on ground
		trees IN



Question Answers

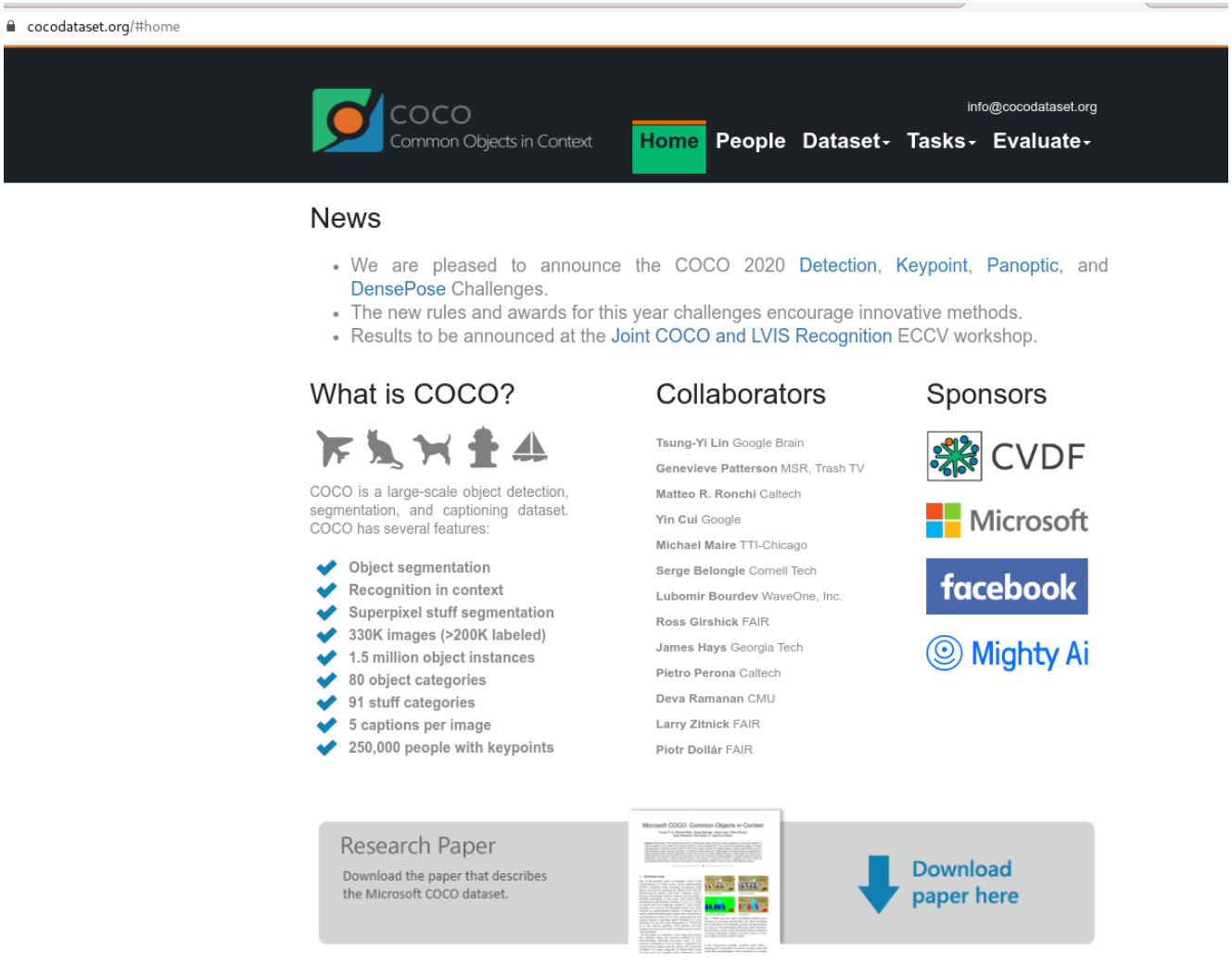
When was the photo taken?	Daytime.
What color are the trees?	Green.
What is in the background?	The trees.
What time of day was this taken?	In the daytime.
What is hanging off the sheep head?	His ears.

Regions Attributes Relationships

Computer Vision Datasets (also L&V datasets)

MS COCO

- localisation, segmentation, detection, image description
- many 3rd-party extensions (see website)



Computer Vision Datasets (also L&V datasets)

MS COCO

cocodataset.org/#explore

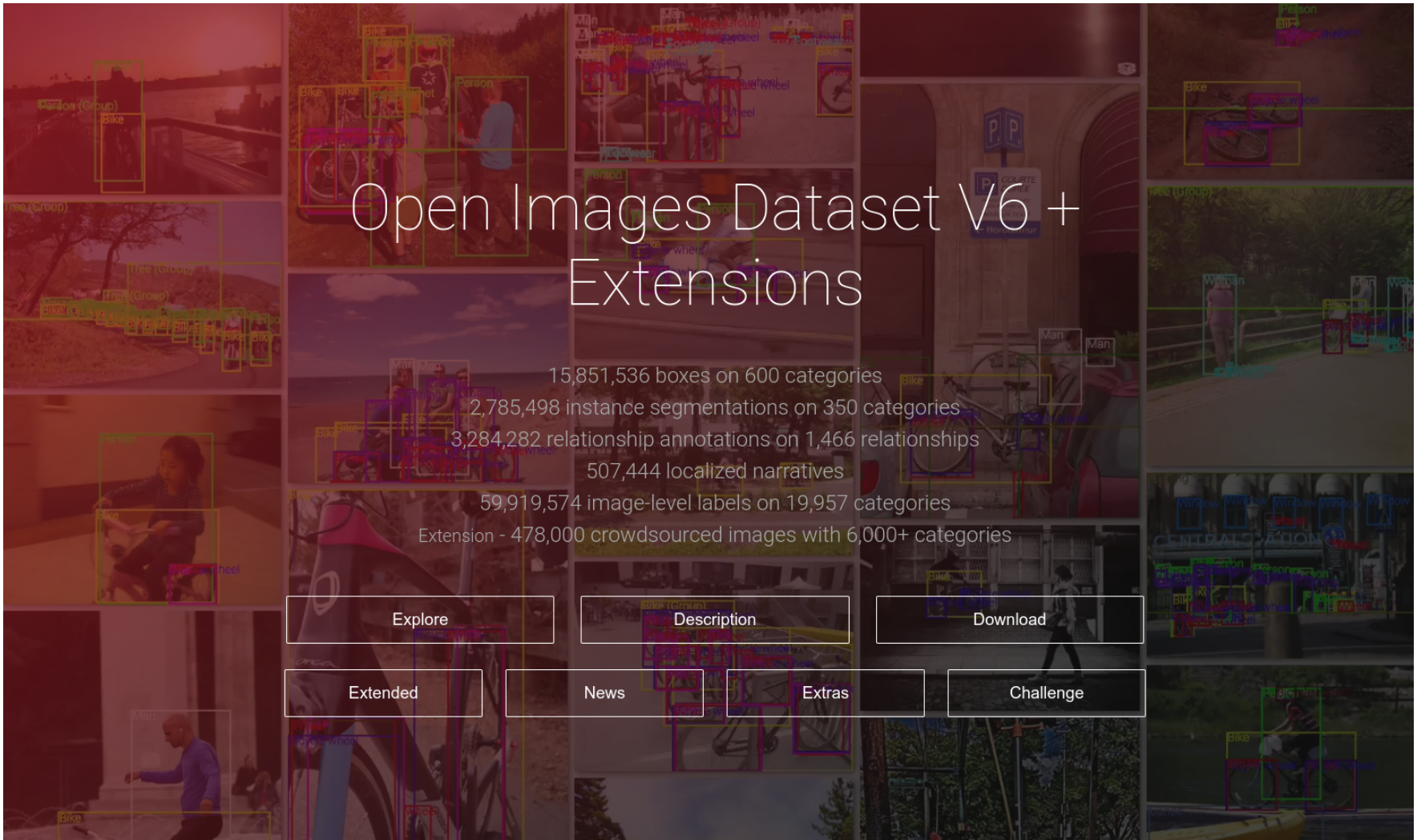
URL sheep

a man and two boys herd 5 sh a truck.
sheep herded together in pen by adults on farm.
the men are herding the sheep into the truck.
there are men hoarding sheep into a crate
a vintage photo of some sheep being herded

Computer Vision Datasets (also L&V datasets)

OpenImages

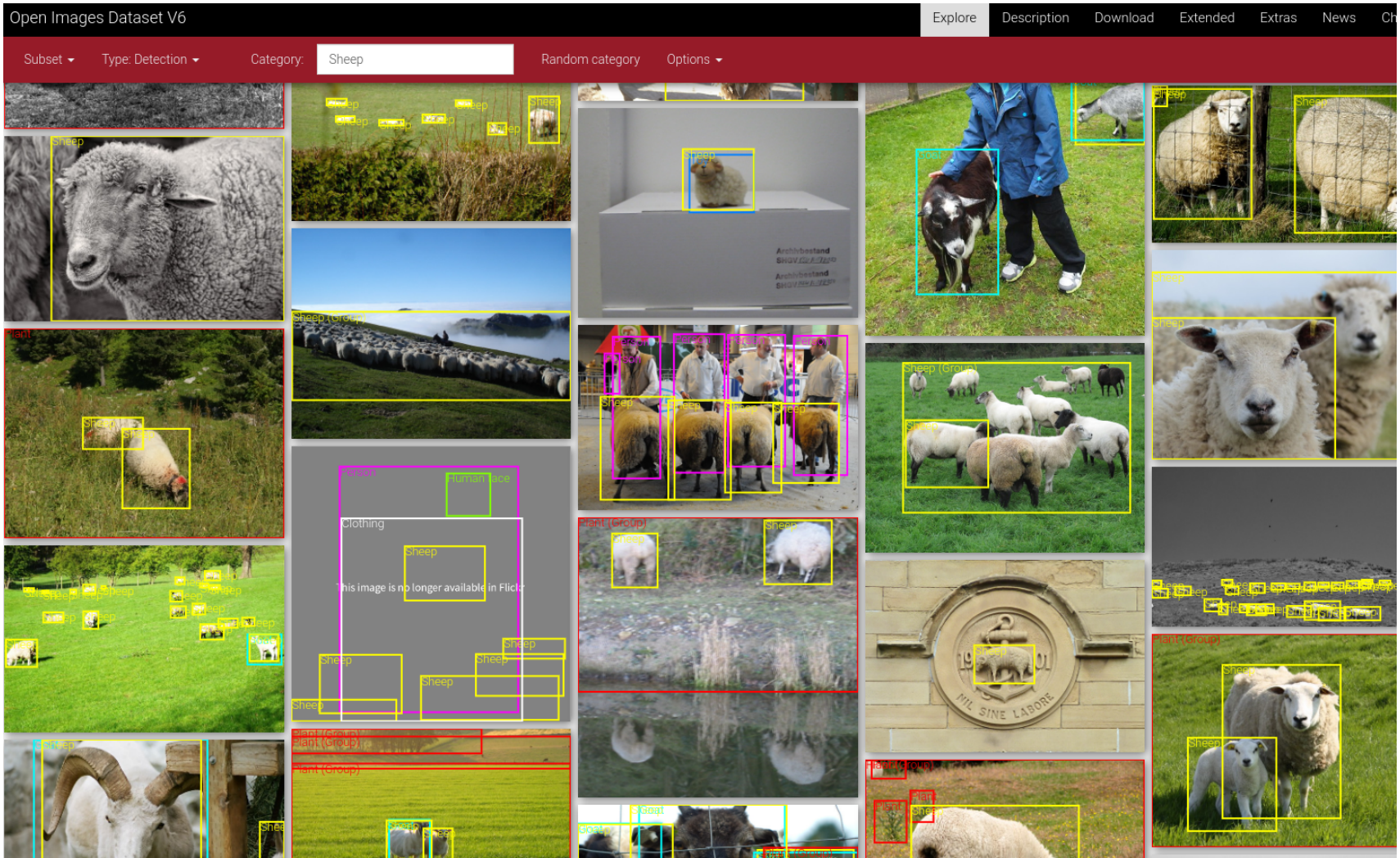
- **detection**, segmentation, relationships, "localised narratives"



Computer Vision Datasets (also L&V datasets)

OpenImages

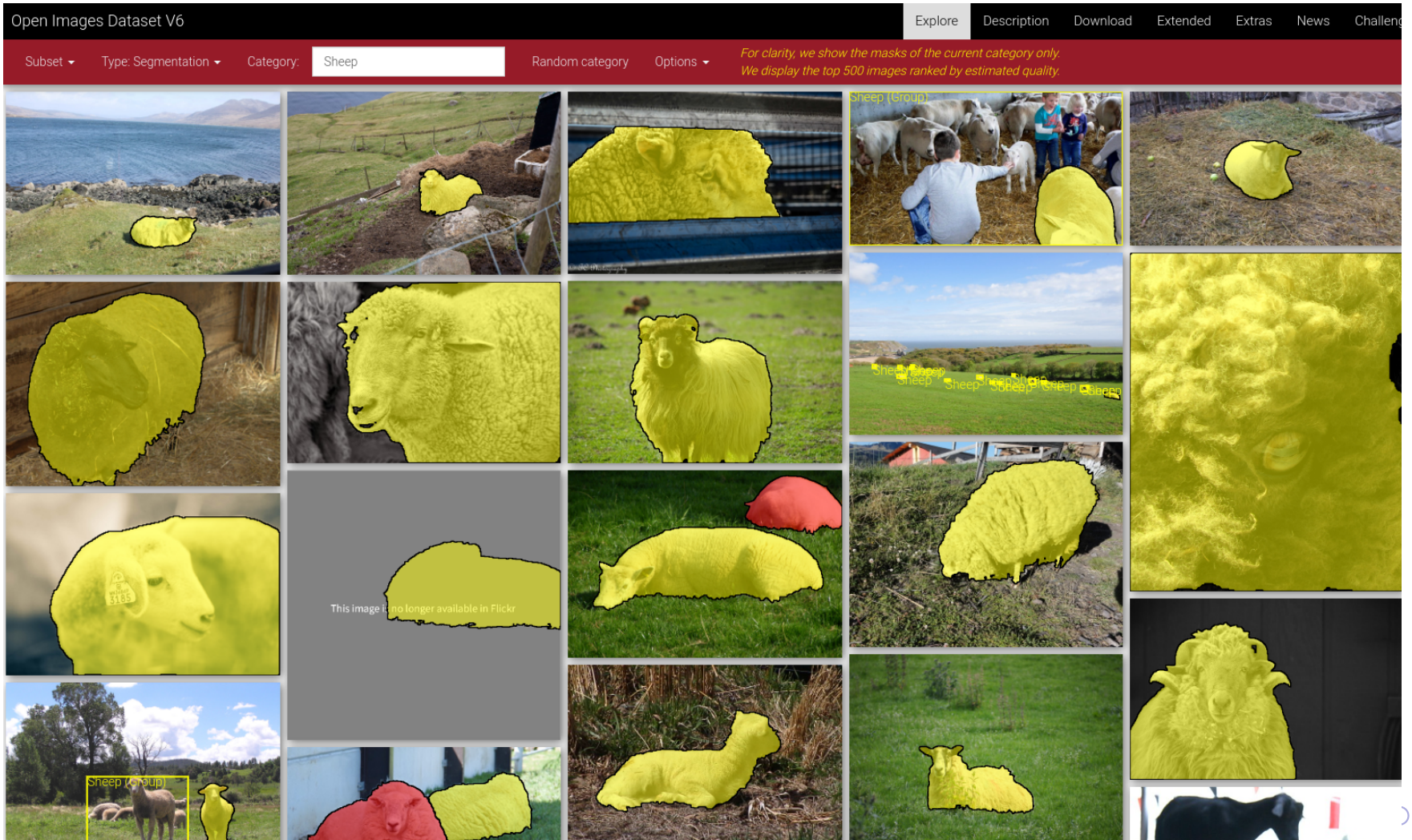
- detection, **segmentation**, relationships, "localised narratives"



Computer Vision Datasets (also L&V datasets)

OpenImages

- detection, segmentation, **relationships**, "localised narratives"



Computer Vision Datasets (also L&V datasets)

OpenImages

- detection, segmentation, relationships, "localised narratives"

References I

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021).
An image is worth 16x16 words: Transformers for image recognition at scale.
In *International Conference on Learning Representations*.