Paths towards Open World Generalization

Thomas Brox

Computer Vision Group

University of Freiburg

The problem of out-of-distribution data

CityScapes

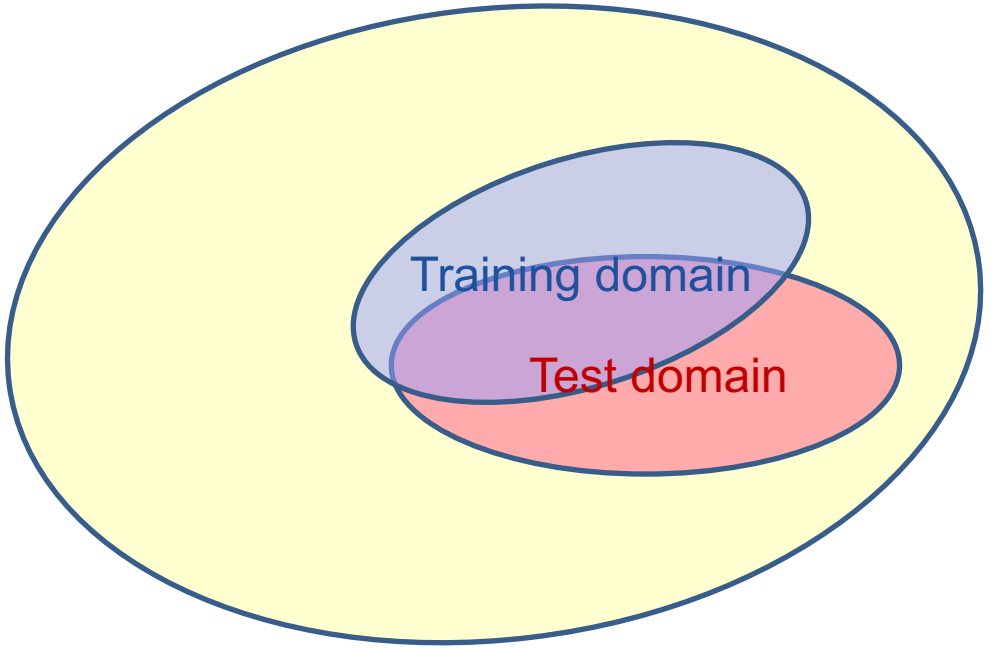
Examples from the DAWN dataset

Distribution shifts

- Other examples:
  - Training on European roads, testing on US roads
  - Training on rendered images, testing on natural images
  - New camera hardware
  - Adding new classes

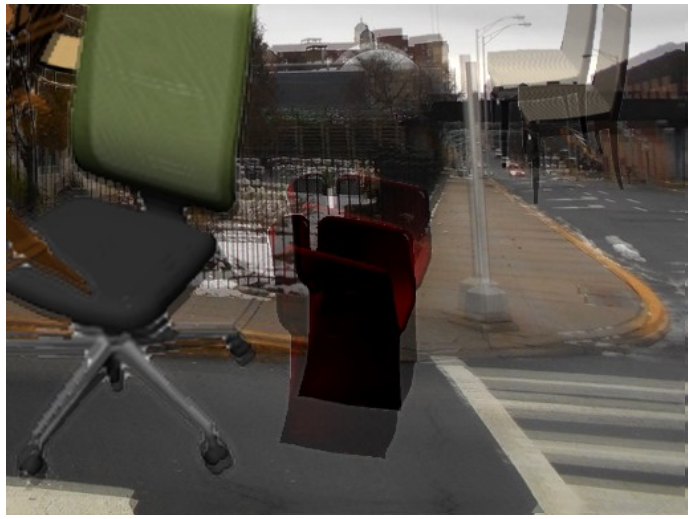
- Approaches:
  - Adaptation (various types of supervision)
  - Generalization

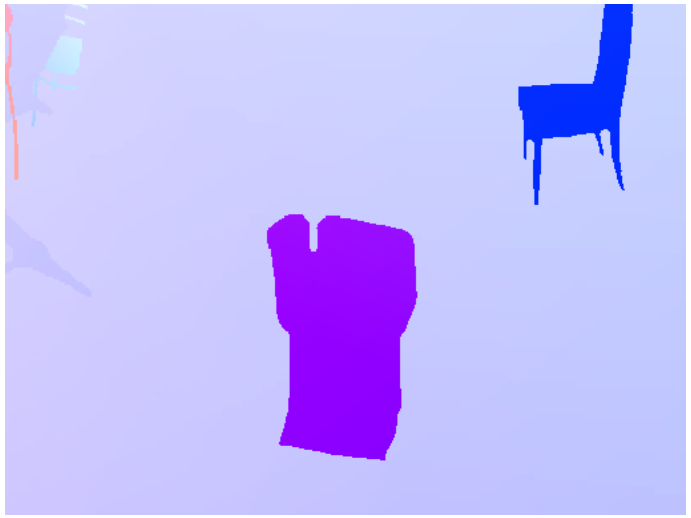
- Self-supervised learning promising
  - easier to expand the distribution
  - learning of short cuts less likely

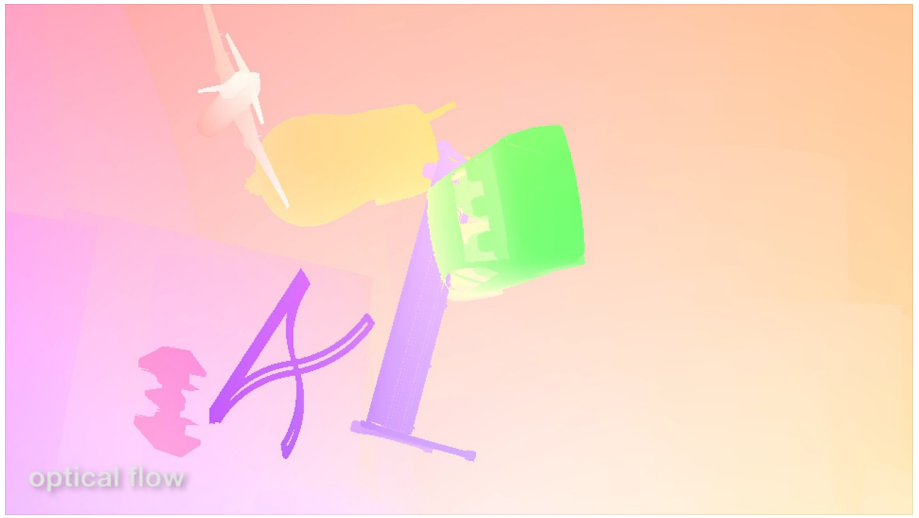


Open world generalization for optical flow

Synthetic training data



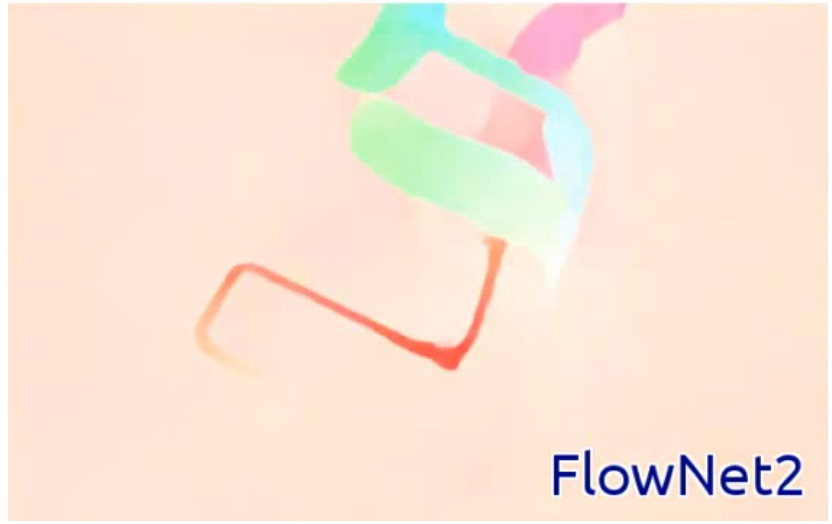




FlyingChairs

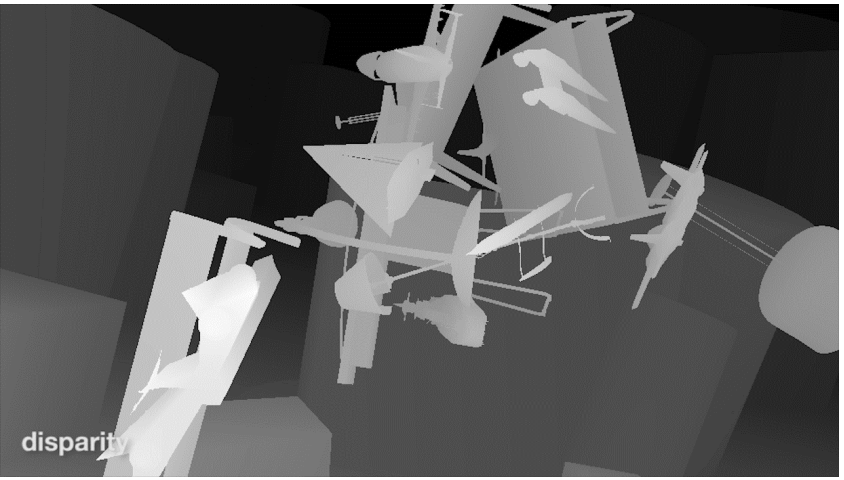
FlyingThings3D

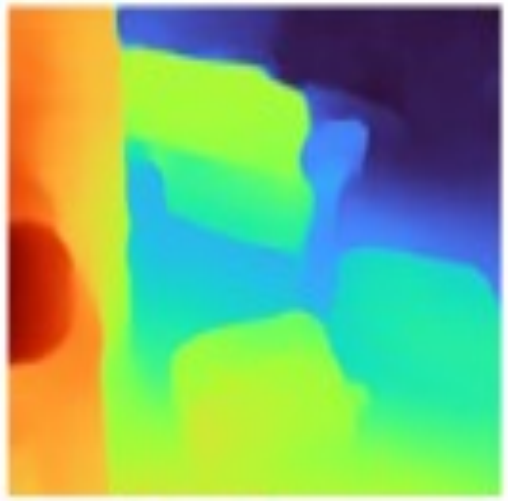
Test data

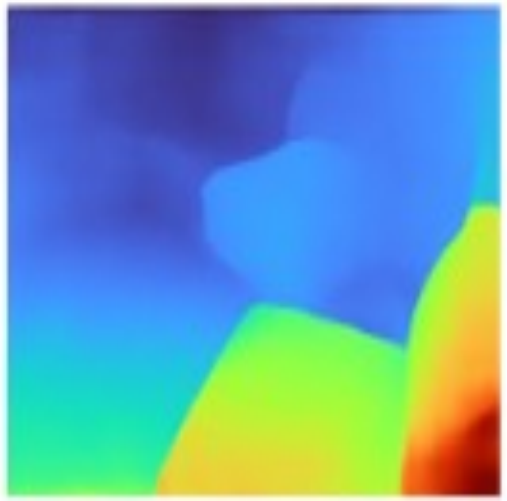


Dosovitskiy et al. 2015, Ilg et al. 2017

Open world generalization for stereo



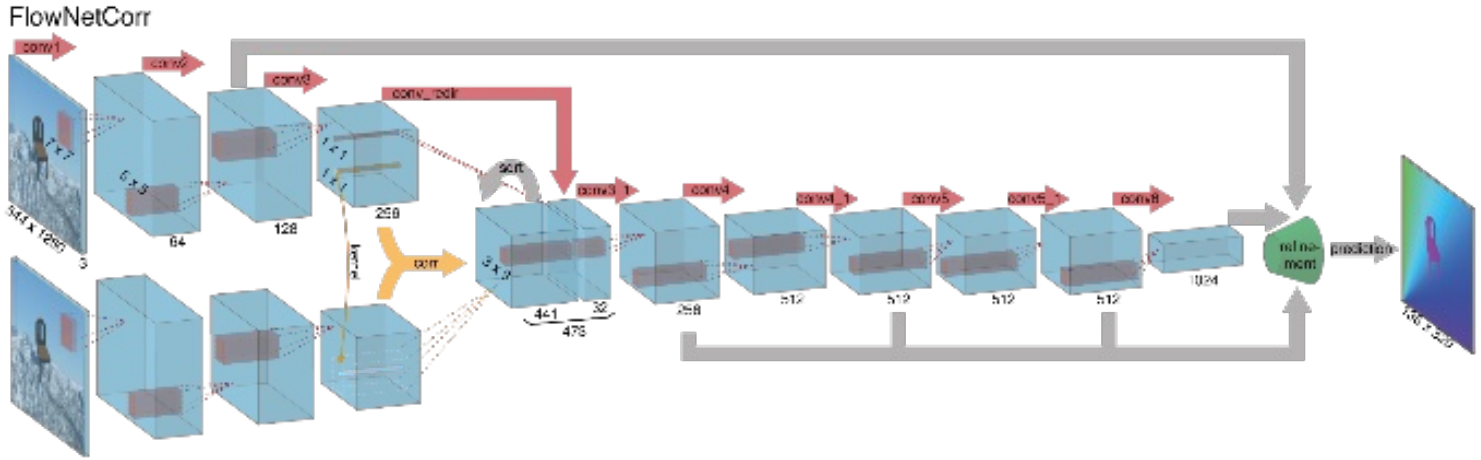




Mayer et al. 2016, Schröppel et al. 2022

Why are recognition tasks different?

- Optical flow and stereo are correspondence tasks



→ features determined based on how well they serve matching
→ only priors for resolving ambiguities are domain specific

- Recognition tasks are “remembering” tasks
  → features are learned to discriminate training samples
  → they are not necessarily descriptive (short cuts likely)

Instance matching: invariance to transformations

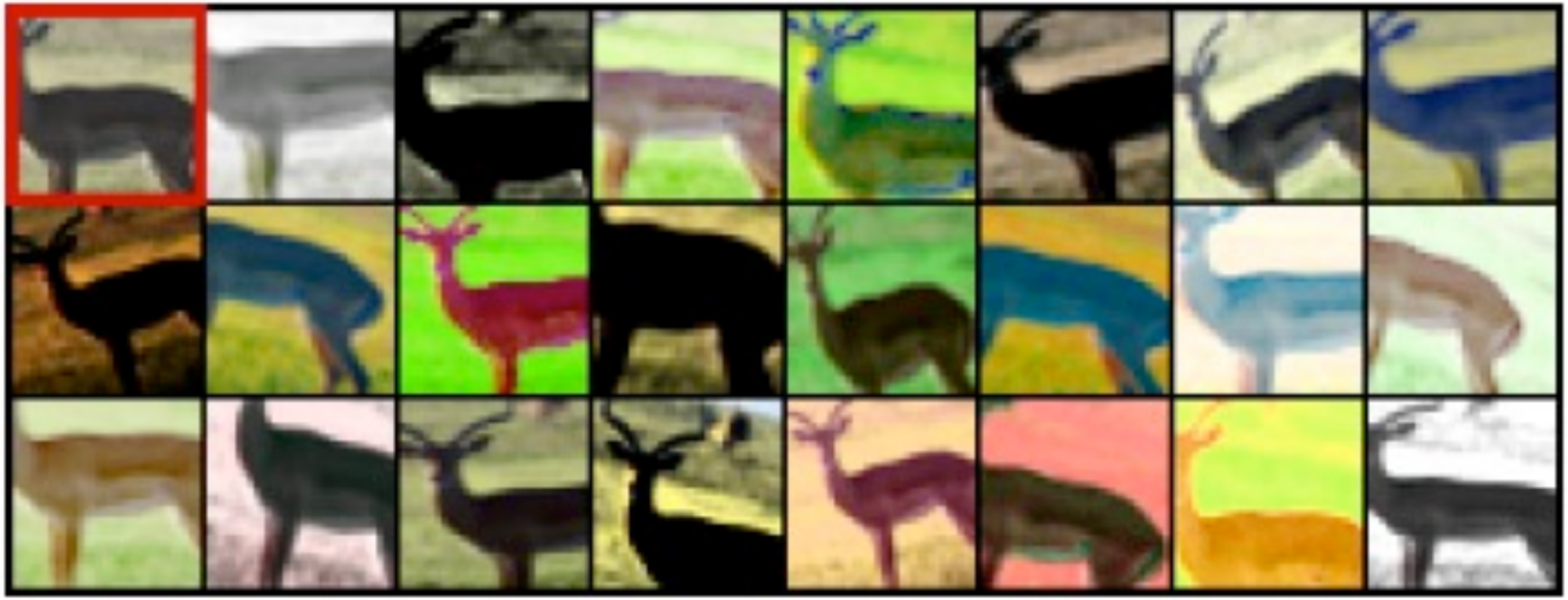
- Recognition at the instance level is closer to a correspondence task

- Variation w.r.t.
  - Pose, camera parameters
  - Lighting
  - Background
  - Occlusion

- Much of this can be simulated (approximately) by data augmentation
  - self-supervised feature learning with contrastive losses
  - learn feature embedding that contracts all instance variations

Exemplar-CNN

- Train CNN to discriminate **surrogate classes** defined by data augmentation



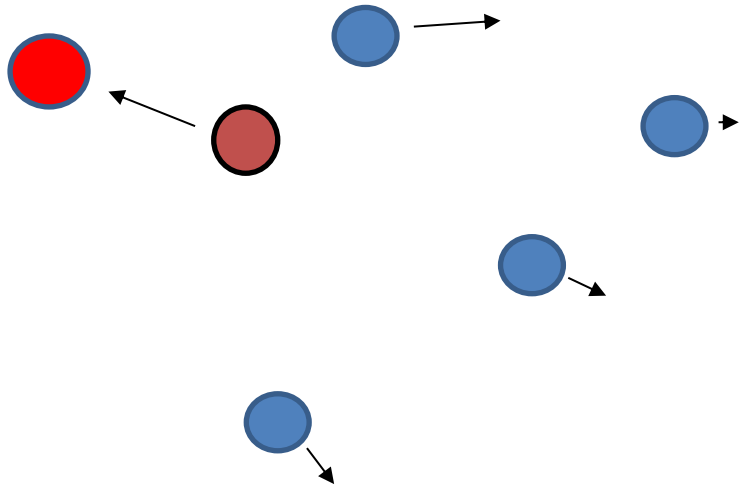
- Yielded good features to match instances

Dosovitskiy et al. 2014

Contrastive learning on synthetic transformations

- Contrastive losses based on positive pairs, for example:

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \; [k \neq i] \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$



- Typically established via synthetic data augmentation

→ Contrastive learning fosters instance matching

Why is this good for recognition in general?

Classification as instance matching?

- How different are instances really?



a) Much variation is covered by instance transformations

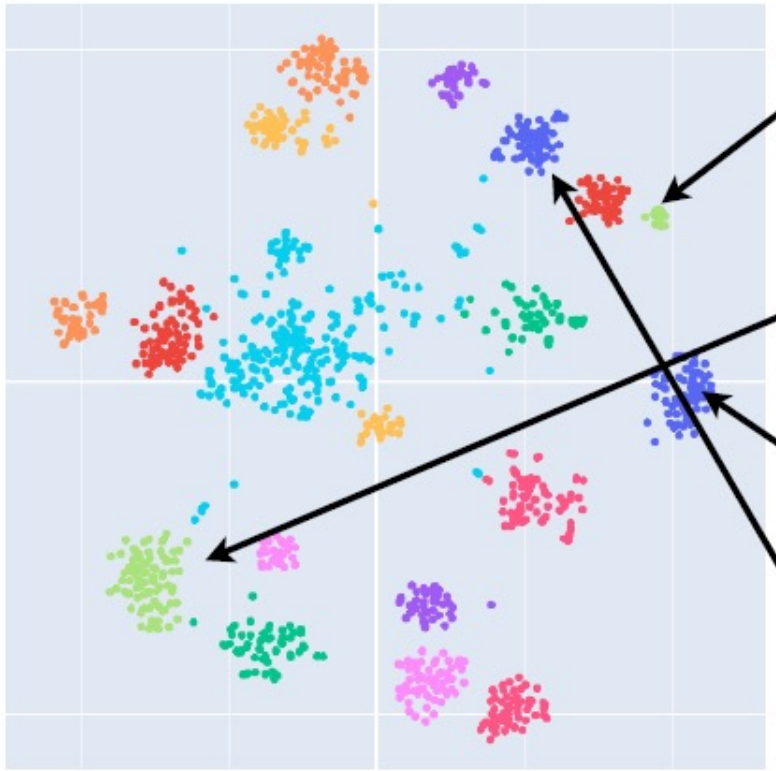
b) Larger differences often covered via transitivity

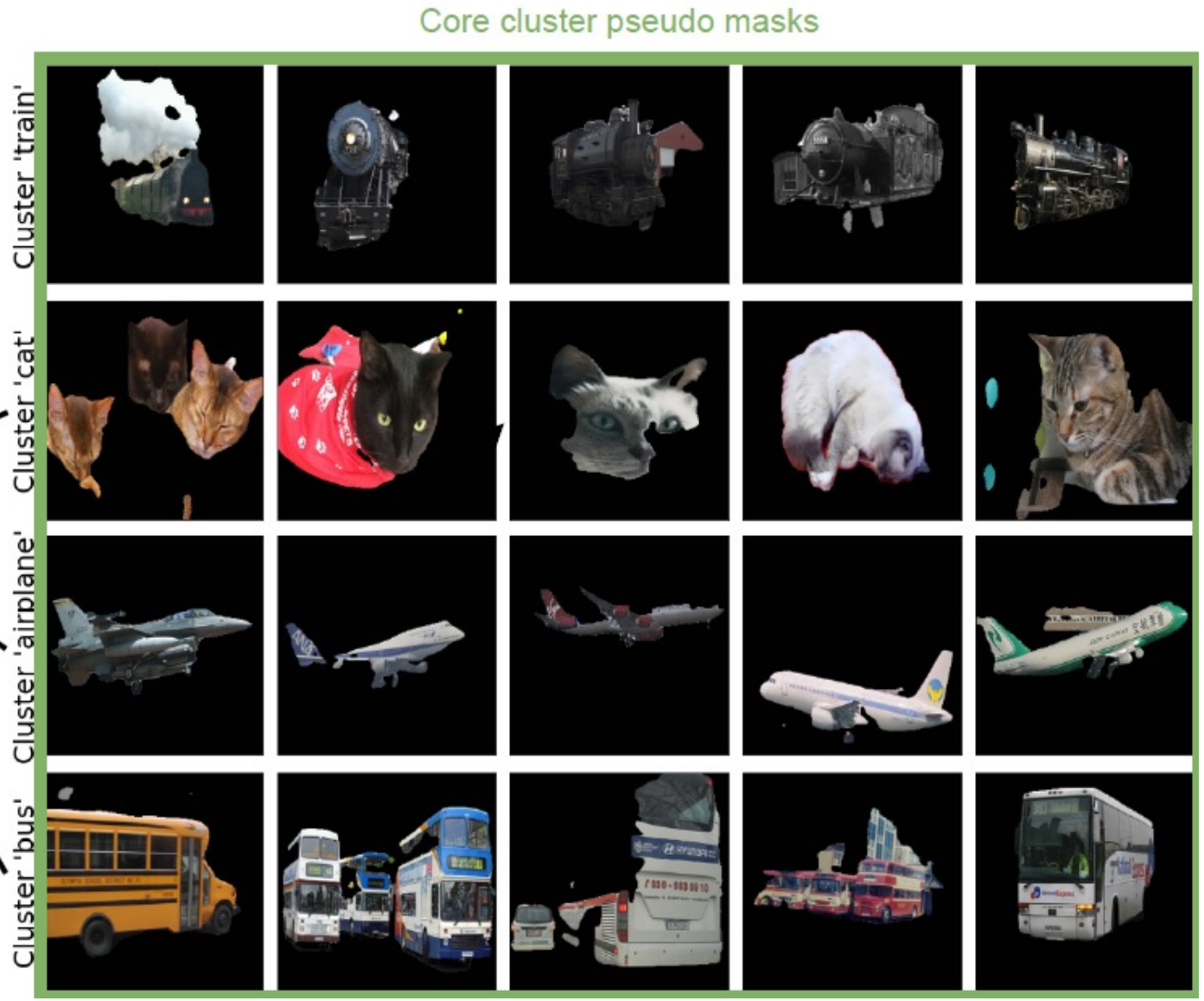
- Contrastive learning respects transitivity (though hard to control)

→ Explains why self-supervised learning works at class level

(e.g. Caron et al. 2021, Zadaianchuk et al. 2023)

Unsupervised semantic segmentation





Zadaianchuk et al. 2023

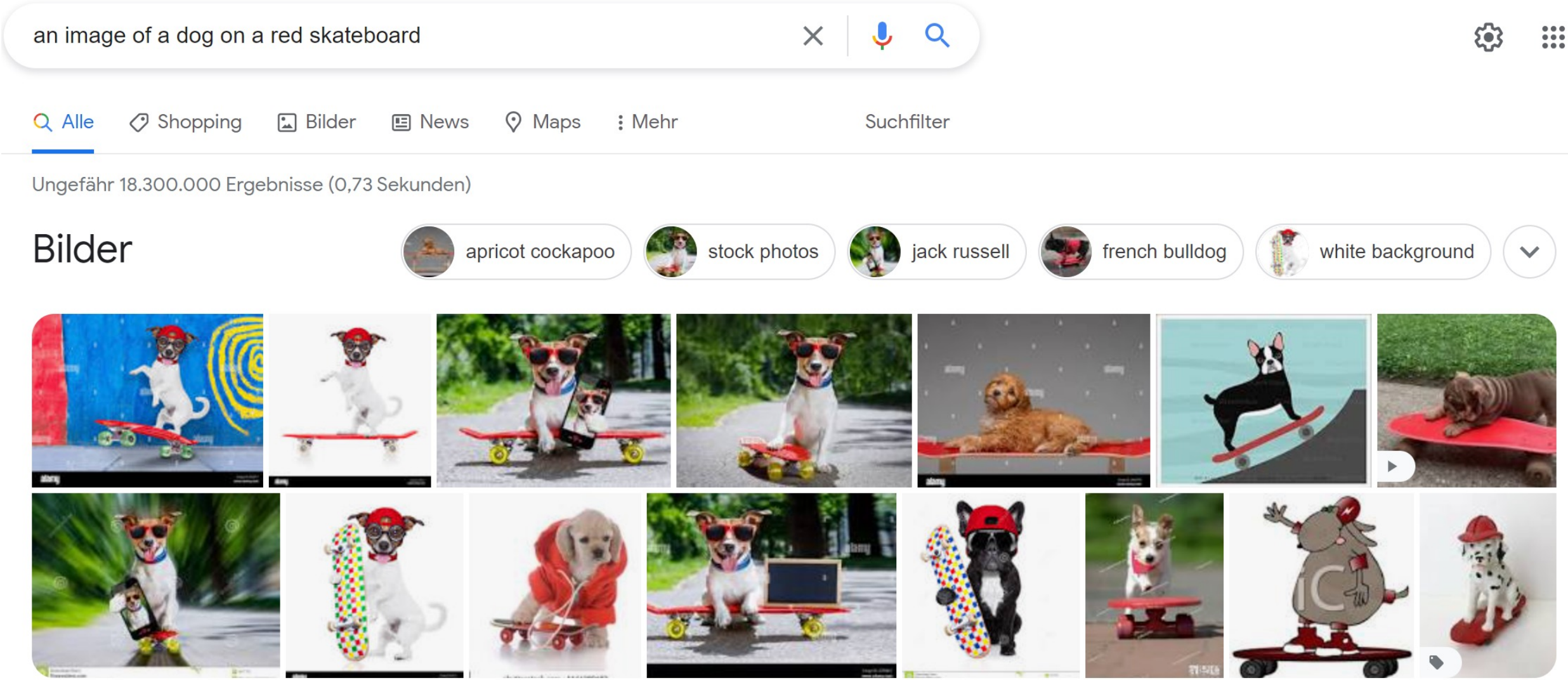
Semantic labels for free: image-text learning

- Image-text pairing is more natural and more powerful than labeling

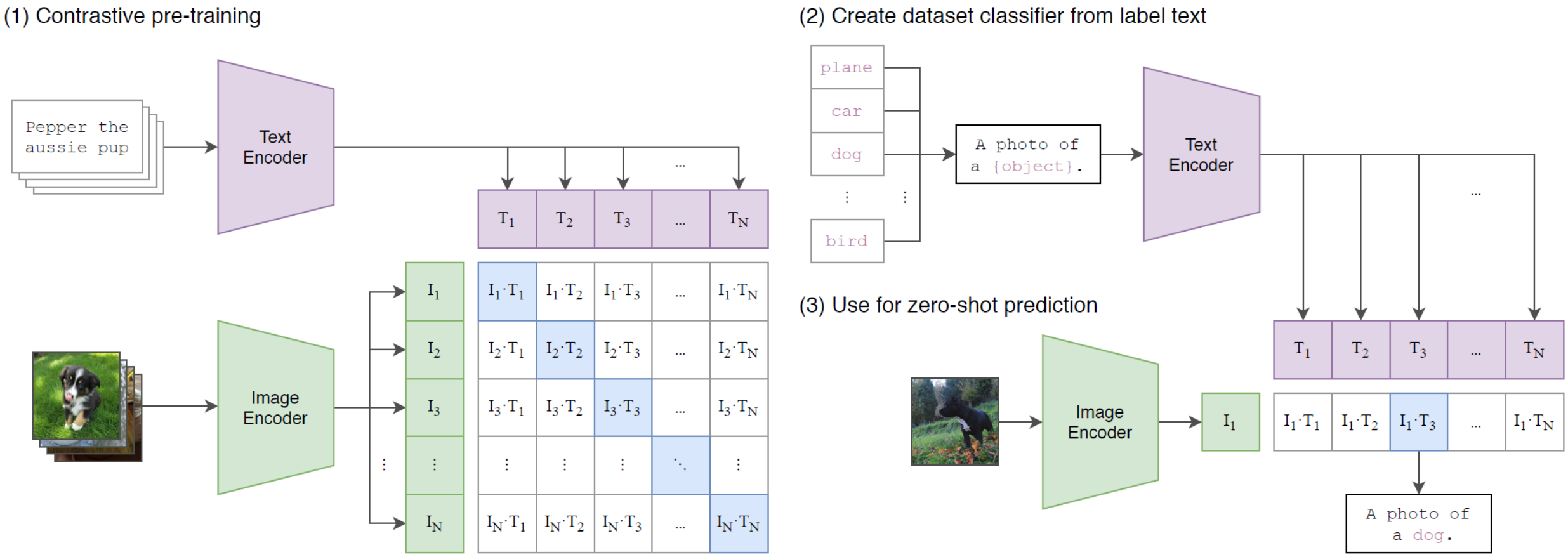
- Often these pairs already exist (in large numbers)
  - Image captions or tags in internet photos
  - Video subtitles, video descriptions
  - Speech recognition in videos

- Data for driving scenarios is scarce

Example for paired image and text



Building image-text embeddings

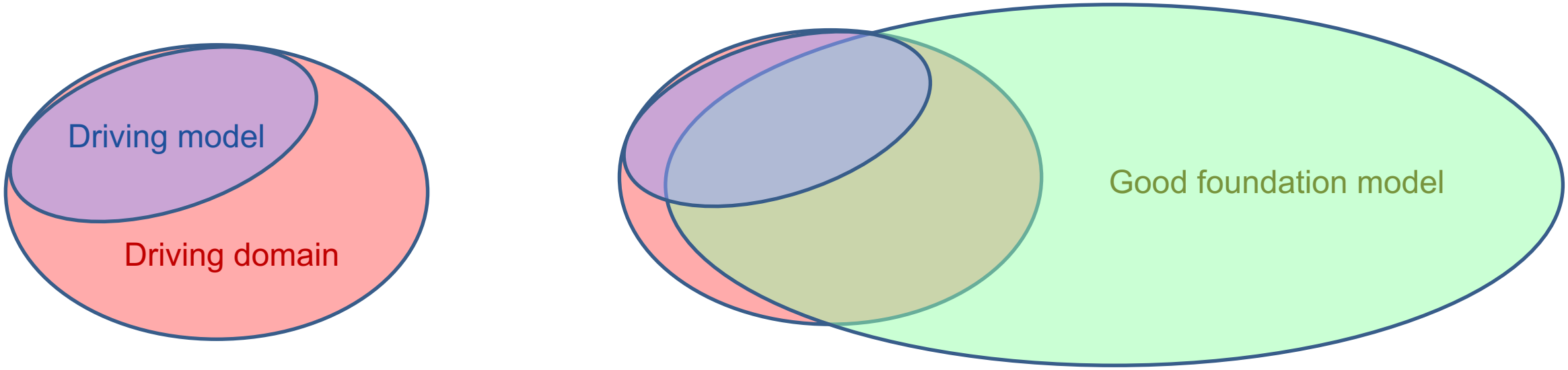


Trained on 400 million image-text pairs obtained from the internet via 500.000 word queries

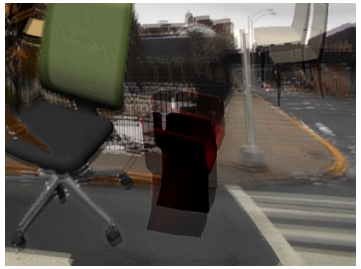
CLIP, Radford et al. 2021

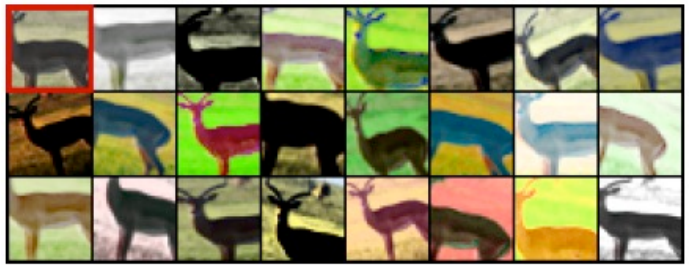
Foundation models and automatic driving: a love story?

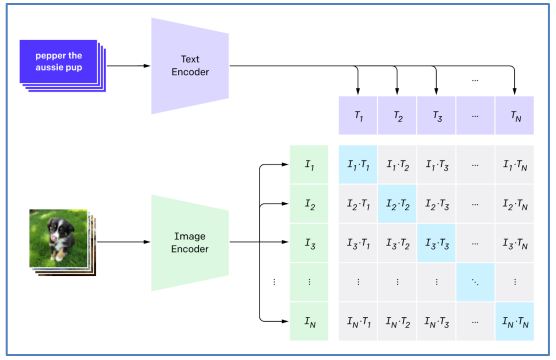
- CLIP yields a “world” model that applies to many downstream tasks
- Out-of-distribution problem stays: no driving data on the web
- Concepts of a good foundation model could transfer (e.g. fog or snow)



Summary







- Self-supervised learning enables open world generalization
- Powerful learning cue comes from instance matching
- Can lead to class-level embeddings via transitivity
- Image-text pairing yields high-level learning cues