A Holistic View of Perception in Intel. Vehicles

Part II: Deep Learning for Perception

IEEE IV 2023
IEEE INTELLIGENT VEHICLES SYMPOSIUM

OLIVES @GeorgiaTech

Georgia Tech

Objectives

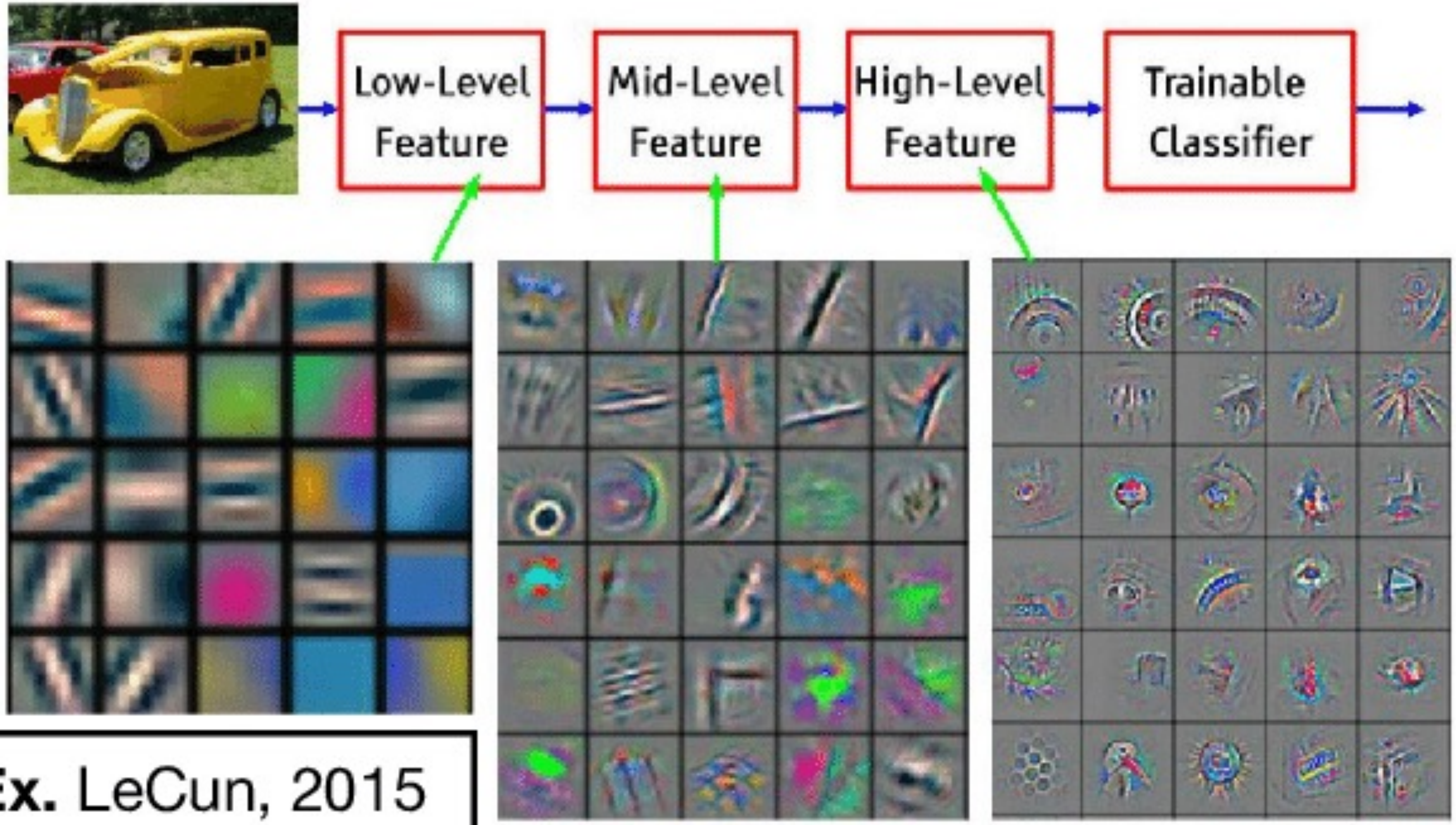
Objectives in Part II

- Discuss myths surrounding deep learning
- Brief history of deep learning
- Review deep learning models for vision
- Deep learning extensions into sensor domain
- Transfer Learning and foundation models
- Self-supervised learning
- Case study: Self-supervised learning for fisheye images

Deep Learning
Meme to start off with

People's expectation of AI and Deep Learning

Deep Learning
Model Decomposition



Ex. LeCun, 2015

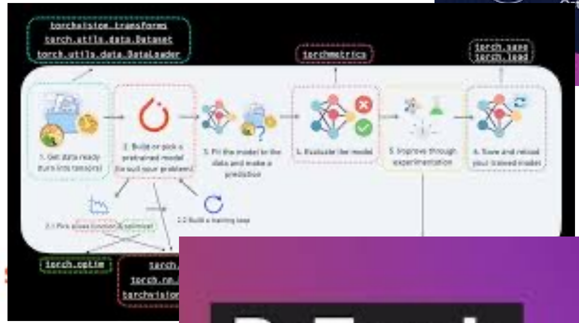
"Deep learning is hard to train"

Convolution Layers

nn.Conv1d	Applies a 1D convolution over an input signal composed of several input planes.
nn.Conv2d	Applies a 2D convolution over an input signal composed of several input planes.
nn.Conv3d	Applies a 3D convolution over an input signal composed of several input planes.
nn.ConvTranspose1d	Applies a 1D transposed convolution operator over an input image composed of several input planes.
nn.ConvTranspose2d	Applies a 2D transposed convolution operator over an

109,392 repository results

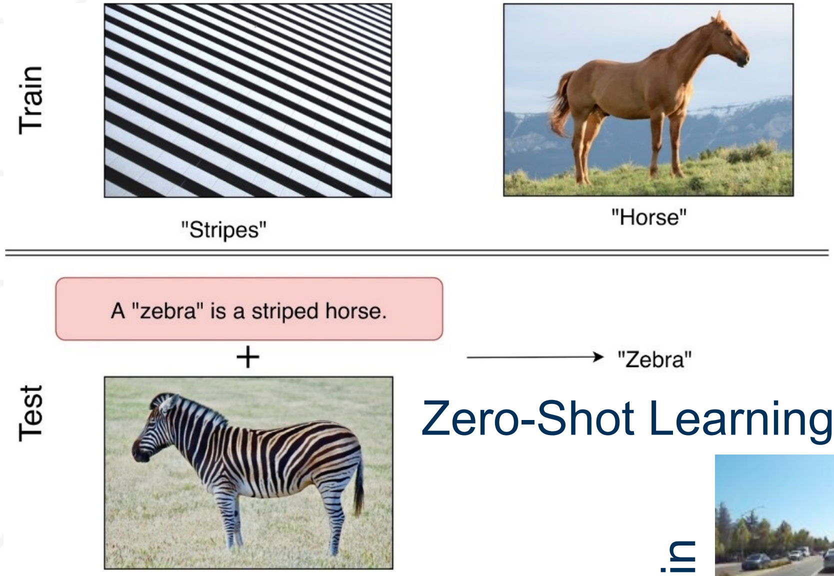
- Containers
- Convolution Layers
- Pooling layers
- Padding Layers
- Non-linear Activations (weighted
- Non-linear Activations (other)
- Normalization Layers
- Recurrent Layers
- Transformer Layers
- Linear Layers



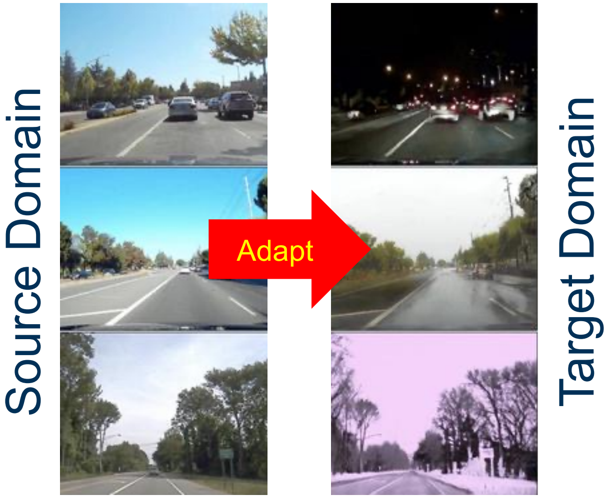
Deep Learning

Some Common Myths about Deep Learning

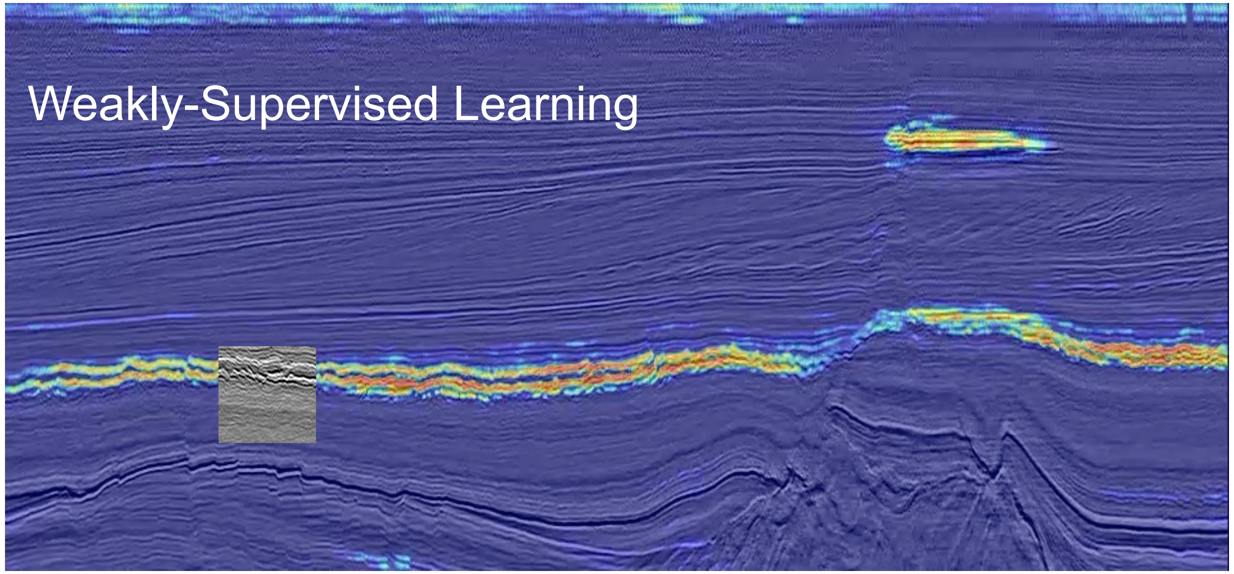
"Deep learning requires lots of data"



Zero-Shot Learning

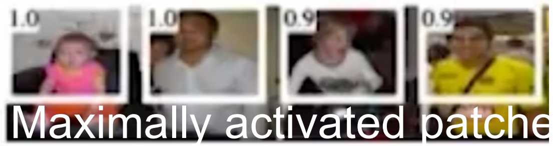


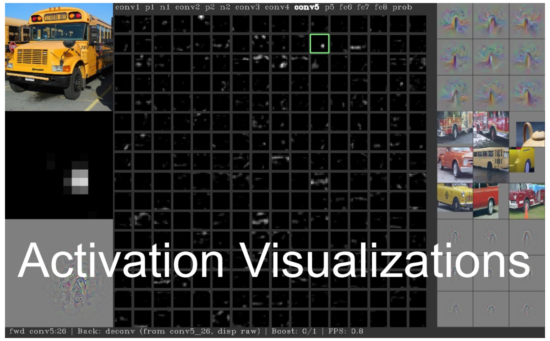
Domain Adaptation

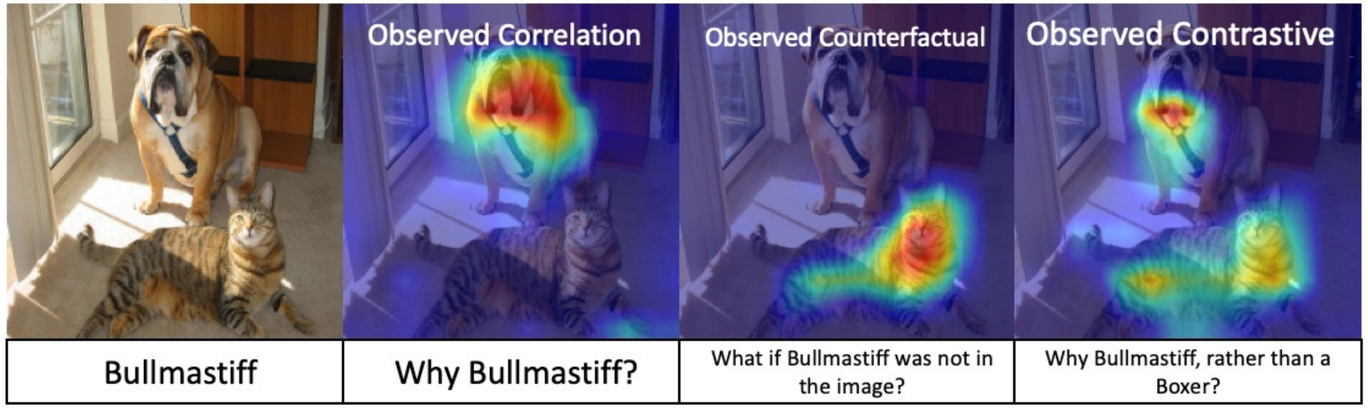


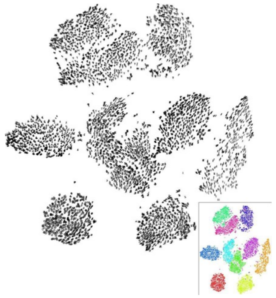
Some Common Myths about Deep Learning

"Deep learning has poor interpretability"

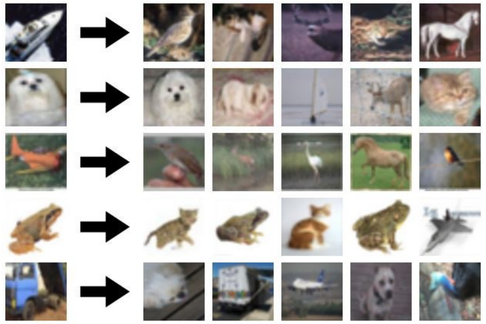






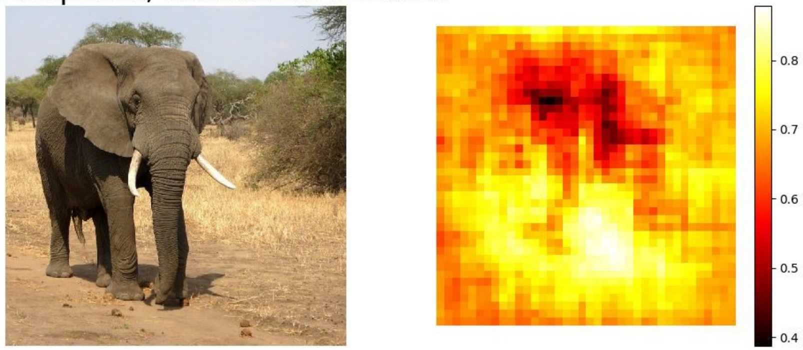


Dimensionality Reduction



Nearest Neighbor

African elephant, Loxodonta africana

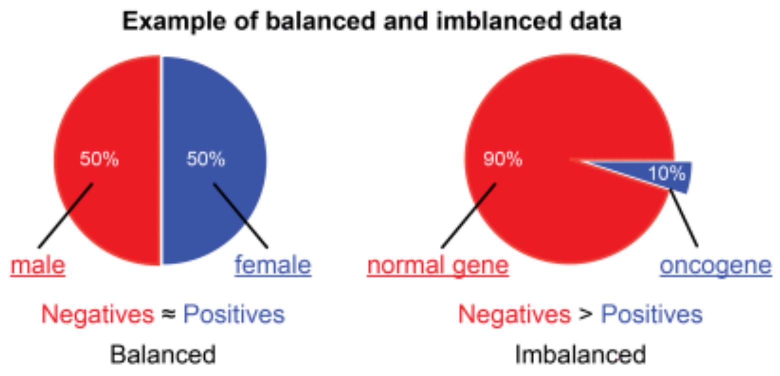


Saliency via occlusion

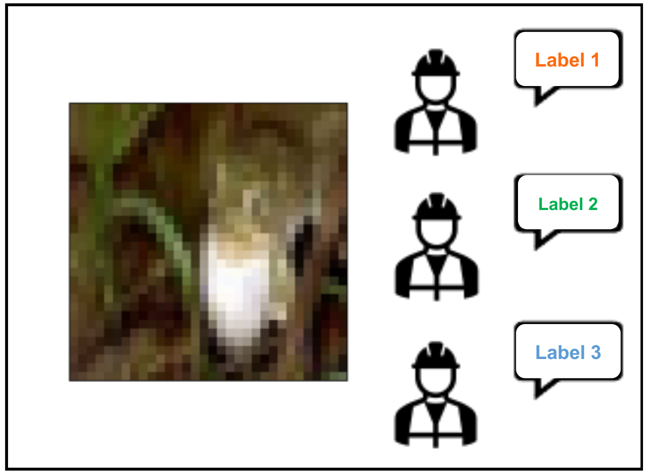
Deep Learning

Some Common Myths about Deep Learning

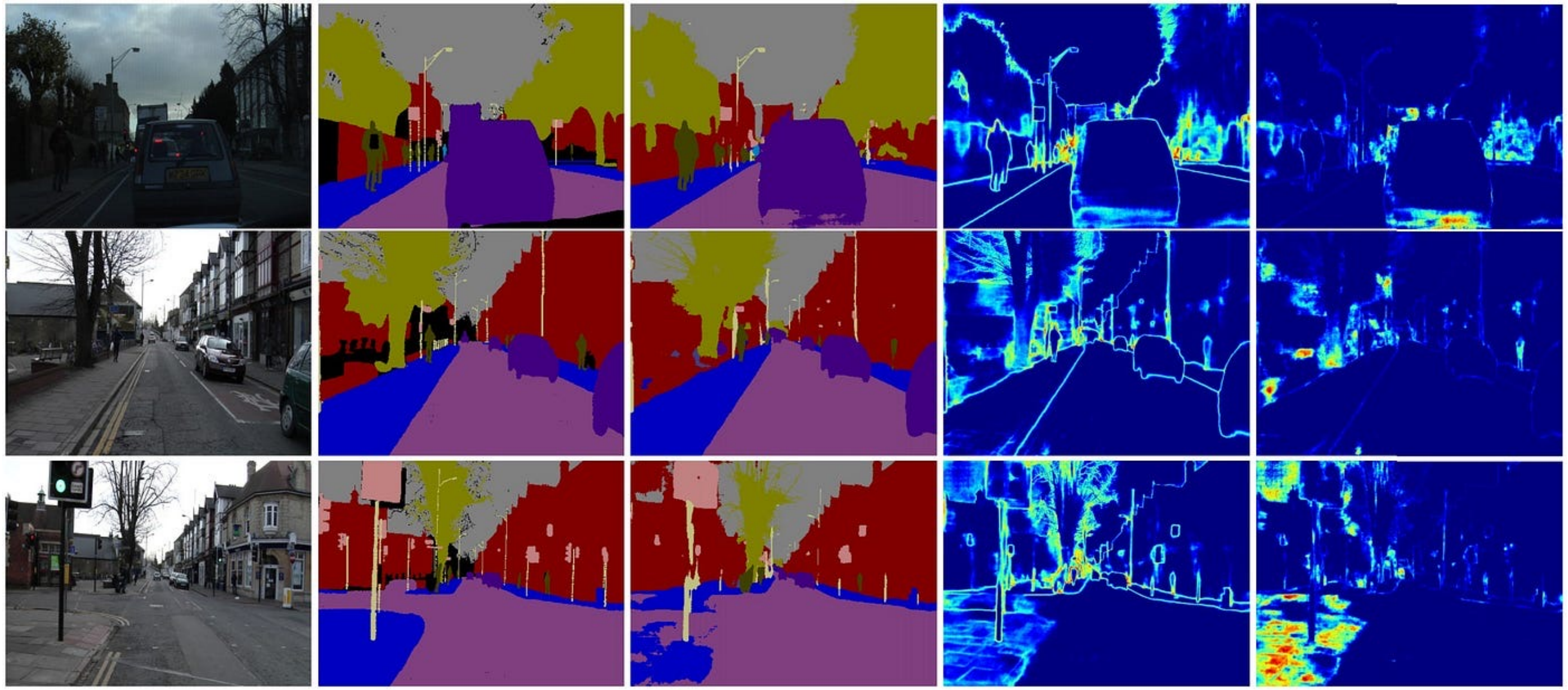
"More the data, better the model"



Data imbalance issues



Human labeling issues



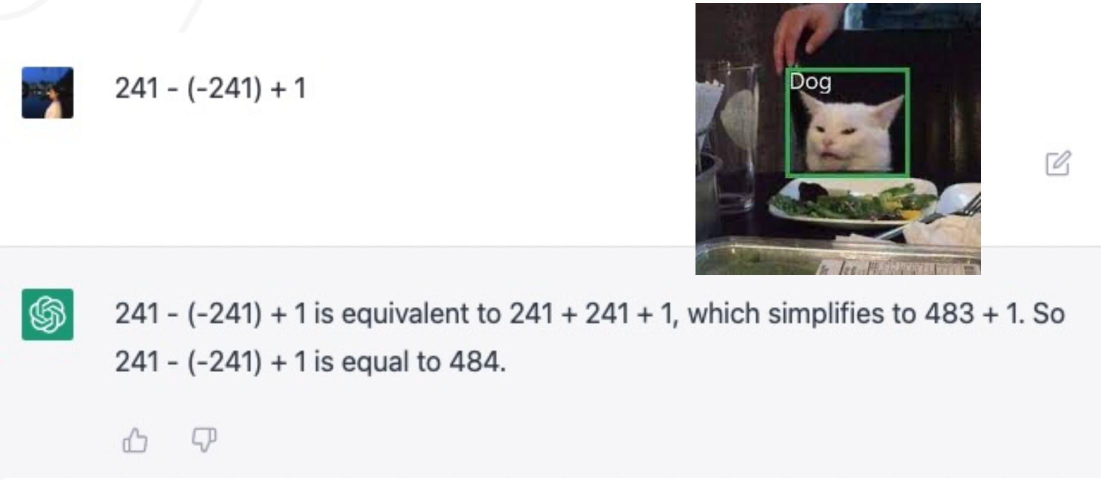
(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

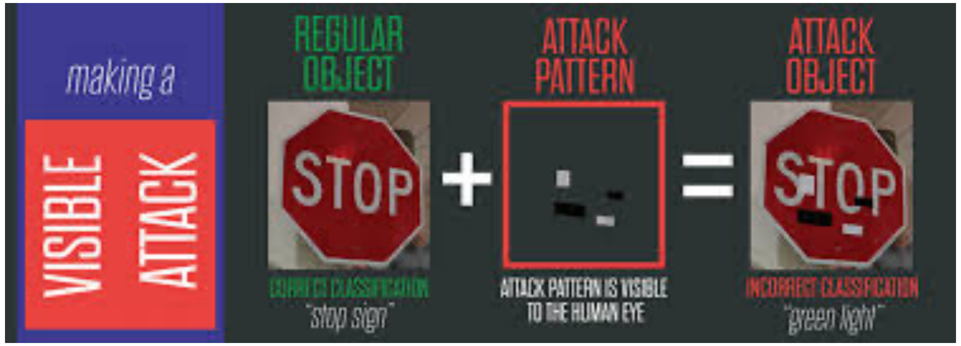
Dataset uncertainties

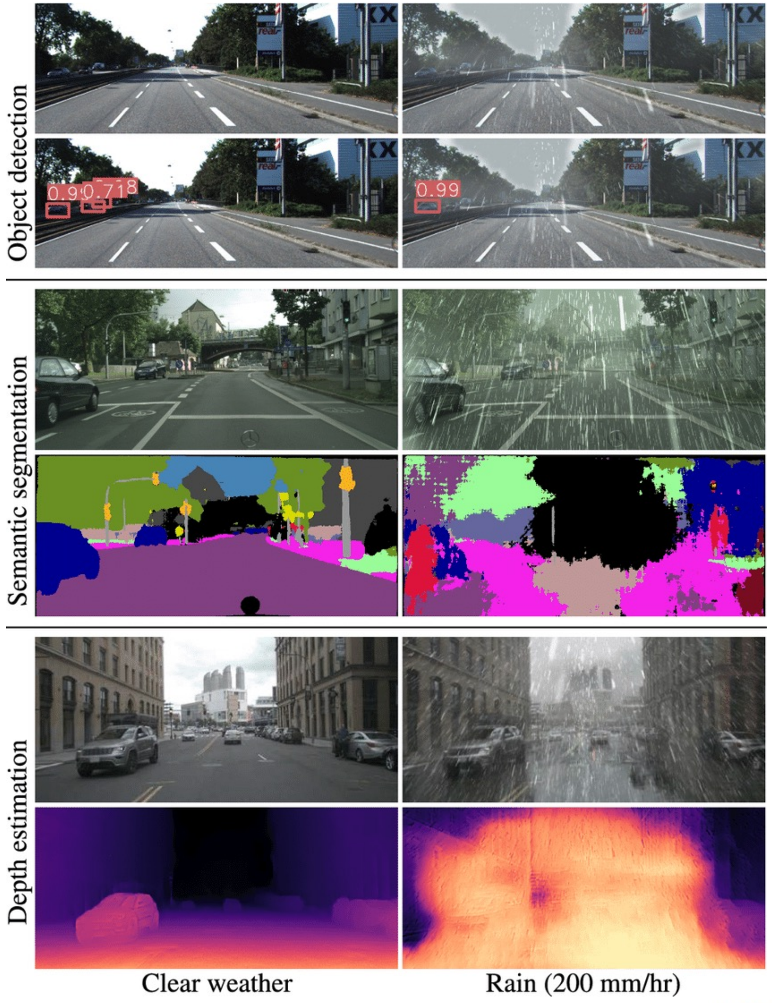
Deep Learning

Some Common Myths about Deep Learning

"Deep learning is State-of-the-Art in every field"







IEEE IV 2023 IEEE INTELLIGENT VEHICLES SYMPOSIUM

[Tutorial] | [Ghassan AlRegib and Mohit Prabhushankar] | [June 4, 2023]

OLIVES @GeorgiaTech

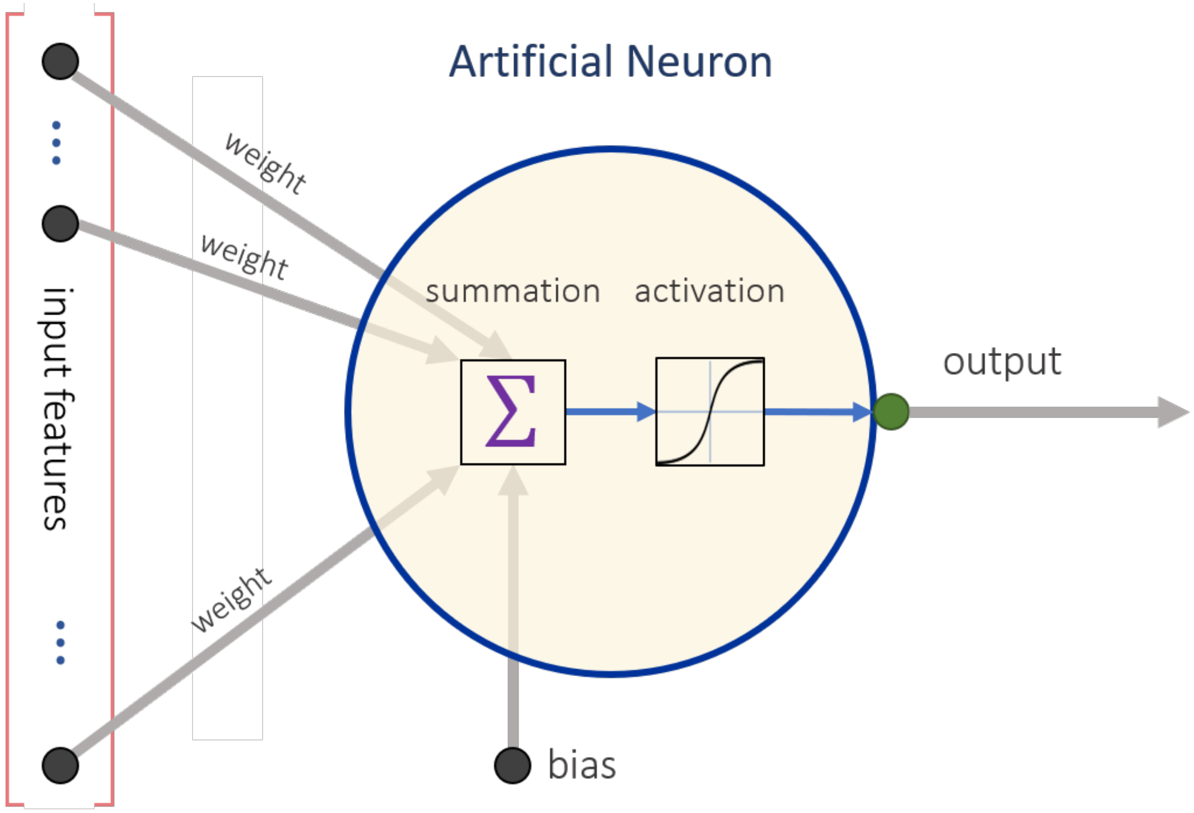
Georgia Tech

Deep Learning
The Building Block

The underlying computational unit is the artificial neuron

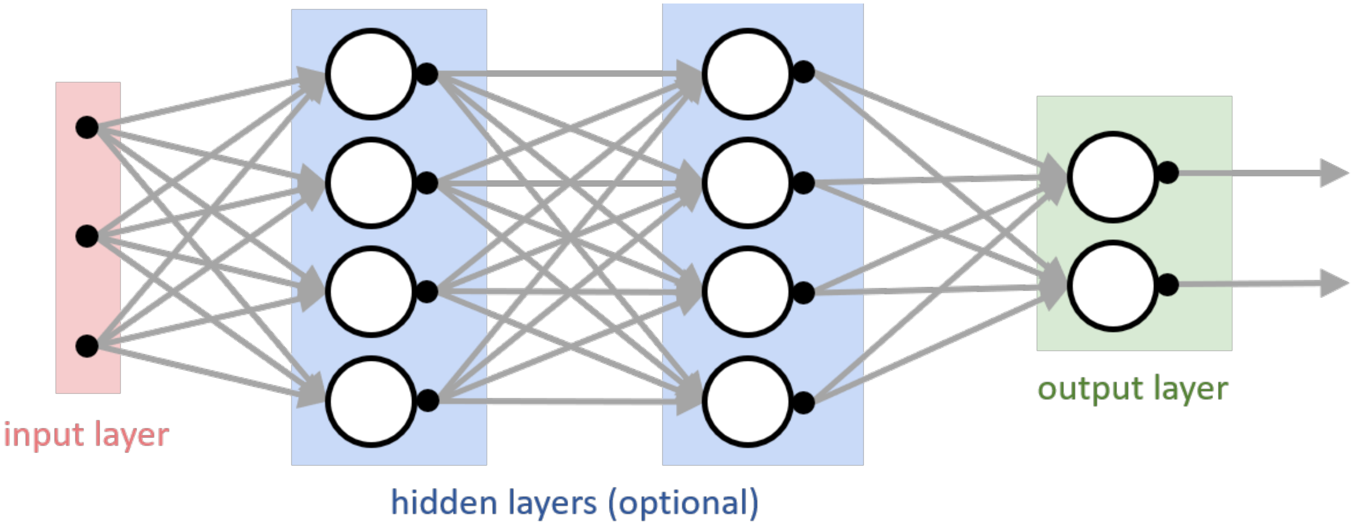
Artificial neurons consist of:

- A single **output**
- Multiple **inputs**
- Input **weights**
- A **bias** input
- An **activation function**



Deep Learning

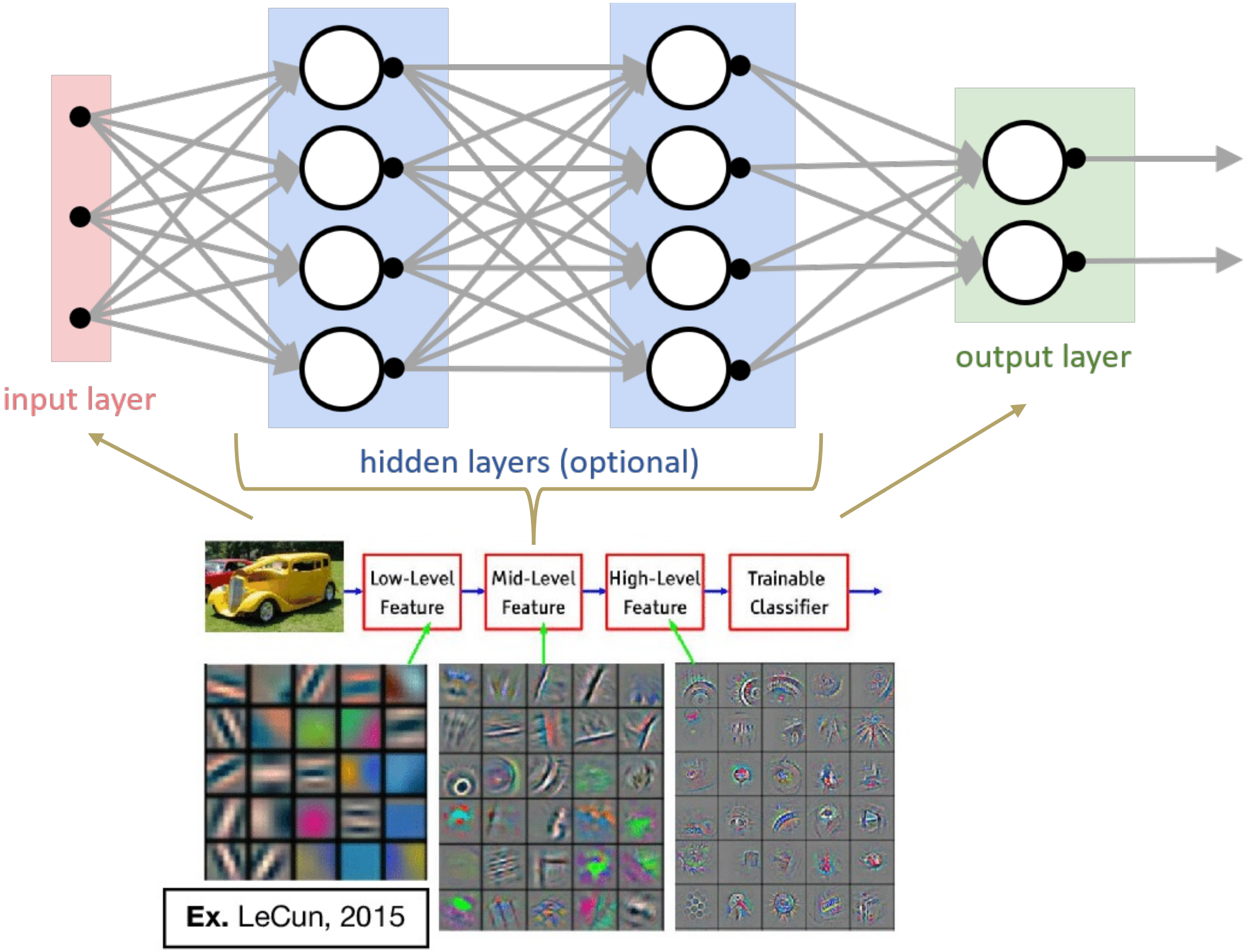
Artificial Neural Networks



Typically, a neuron is part of a network organized in **layers**:

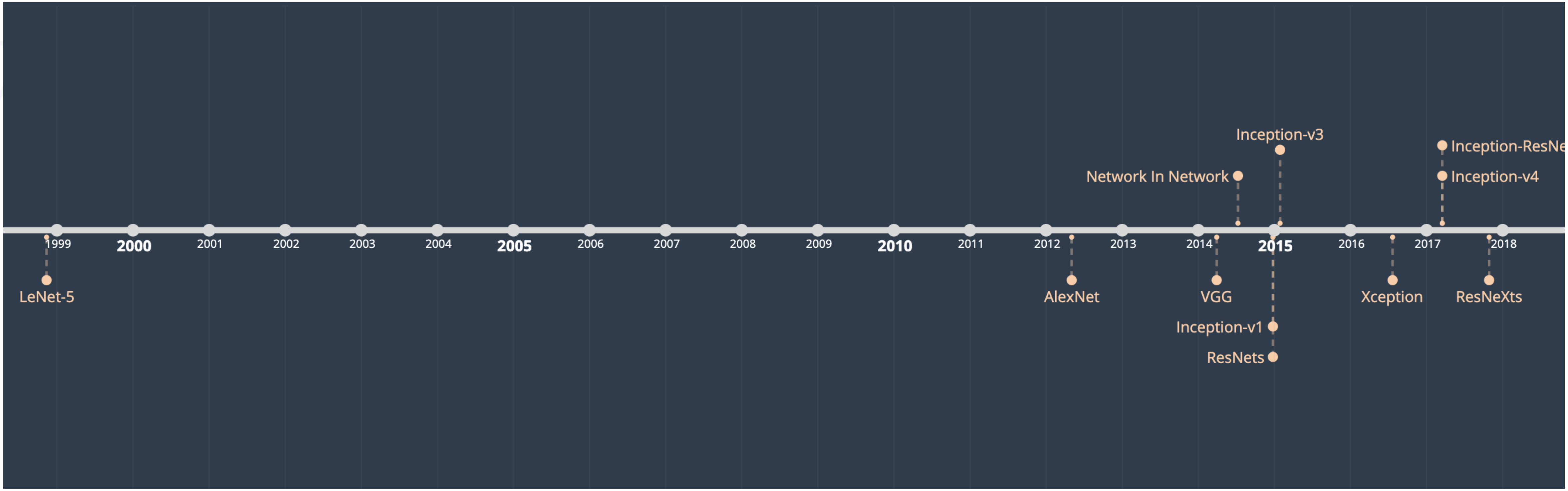
- An **input** layer (Layer 0)
- An **output** layer (Layer K)
- Zero or more **hidden** (middle) layers (Layers $1 \dots K-1$)

Deep Learning
Convolutional Neural Networks

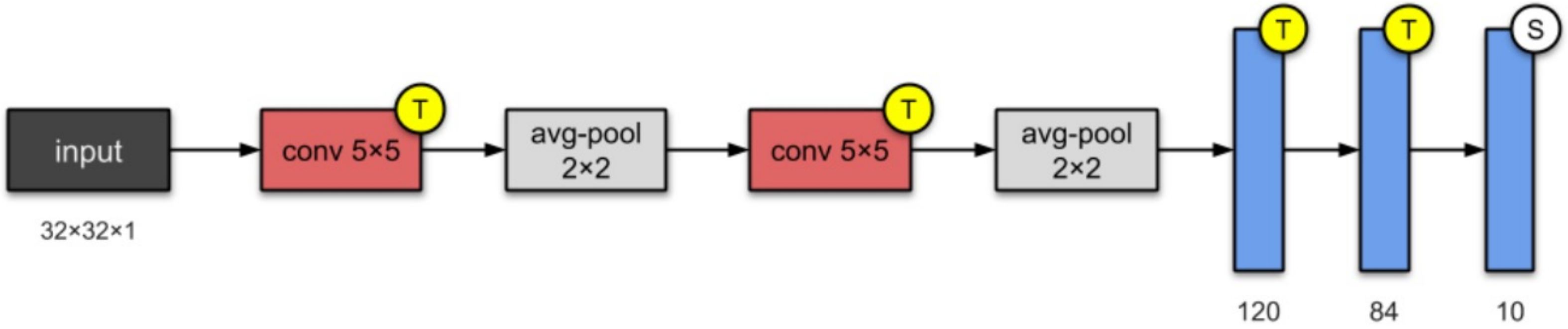


Deep Learning
Evolution of CNN Architectures

- LeNet
- AlexNet
- VGG
- GoogLeNet (Inception-V1)
- ResNet



CNN Architectures
LeNet5 (1998)



Novelty:

- Reduced number of learnable parameters and learned from raw pixels automatically
- The 1st popular CNN that became the “standard” template of CNNs
  - Stacking convolutional, activation, pooling layers
  - Ending with fully connected layers
- Good results on small datasets
  - Top-5 error rate on MNIST is 0.95%

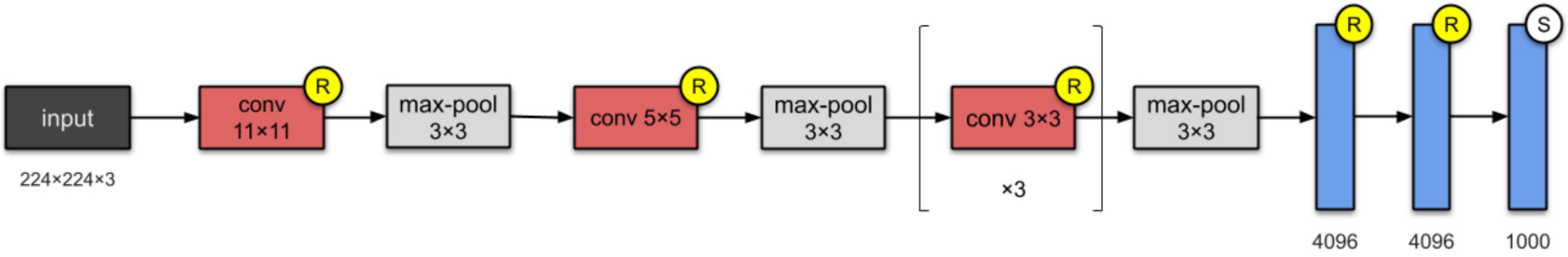
https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d

Long Gap (1998 – 2012)

- Working to improve **computational power**
  - Existing accelerators were not yet sufficiently powerful to make deep multichannel, multilayer CNNs with a large number of parameters.
- Existing **datasets** were relatively **small**
  - Limited storage capacity of computers
- **Tricks** for **neural network training** were not established yet
  - Parameter initialization
  - Variants of stochastic gradient descent
  - Non-squashing activation functions
  - Effective regularization techniques

https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d

CNN Architectures
AlexNet (2011)

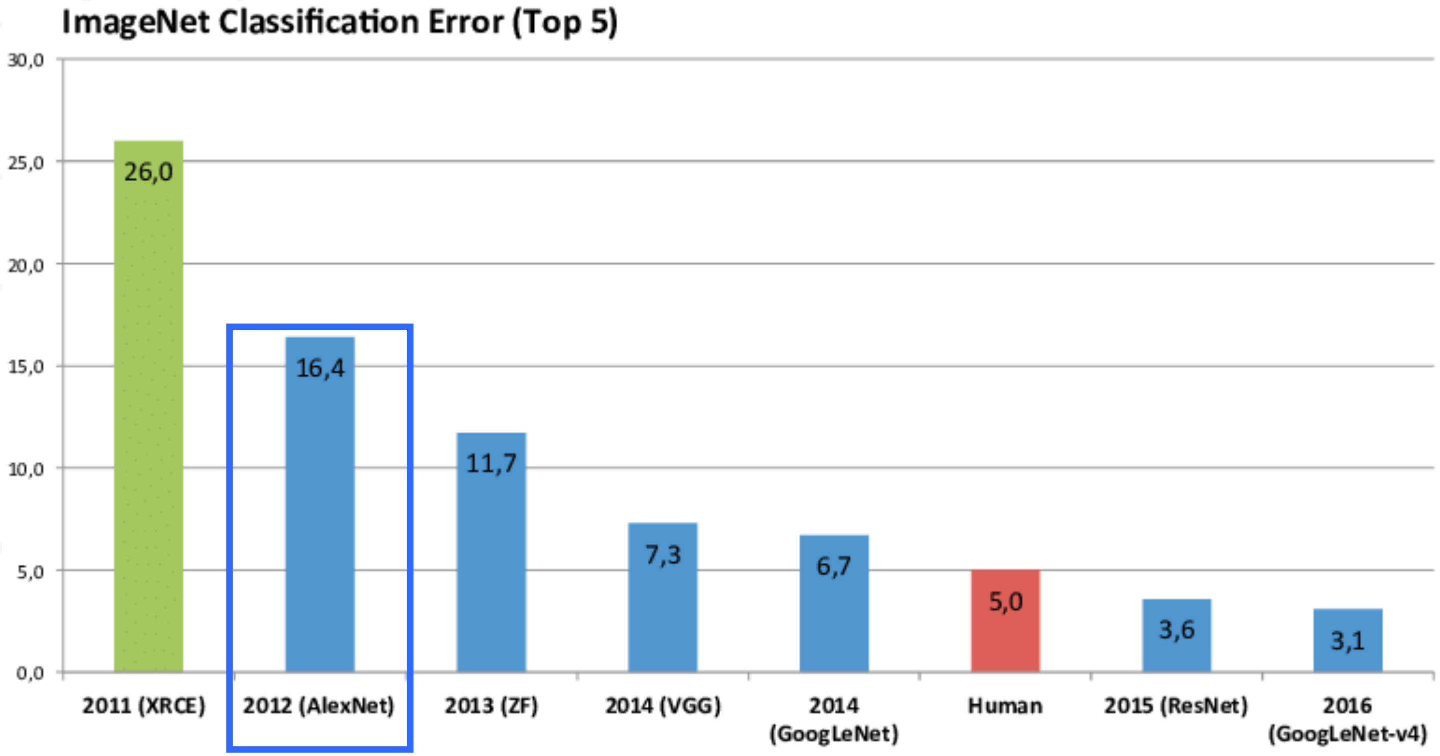


Novelty:

- First to implement Rectified Linear Units (ReLUs) as activation, solving the vanishing gradient problem
- Applied dropout regularization to fully connected layer to control complexity
- Deep CNN that runs on GPU hardware
- Deeper and wider than LeNet
- More robust than LeNet (data augmentation)
- **Won ImageNet Challenge and significantly outperformed traditional methods**

https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d

AlexNet (2012)



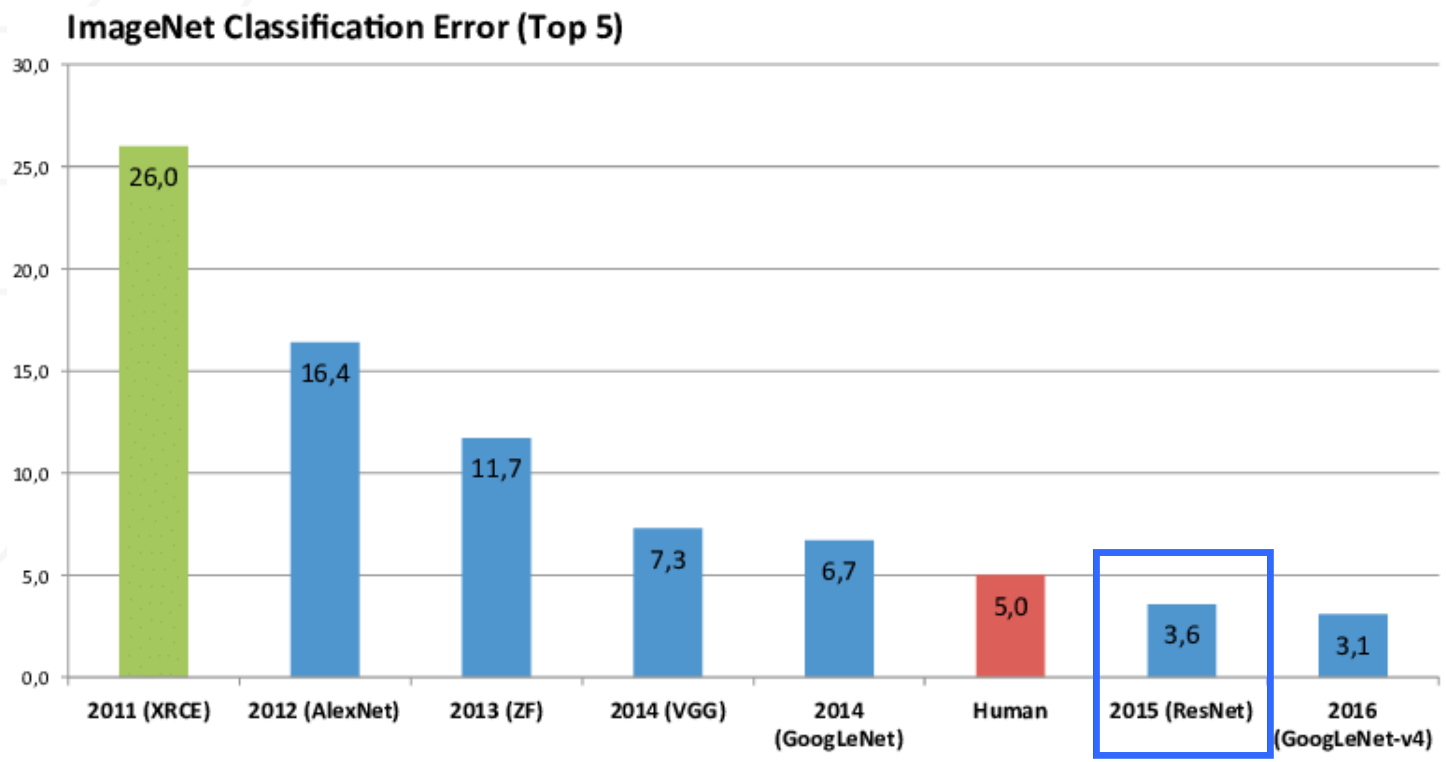
Imagenet:
1000 classes, 1.2M training images, 150K for testing

16.4% top 5 error in ILSVRC 2012

Figure Credit: Zitzewitz, Gustav. "Survey of neural networks in autonomous driving." (2017)

Krizhevsky et al. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012)

ResNet (2015)



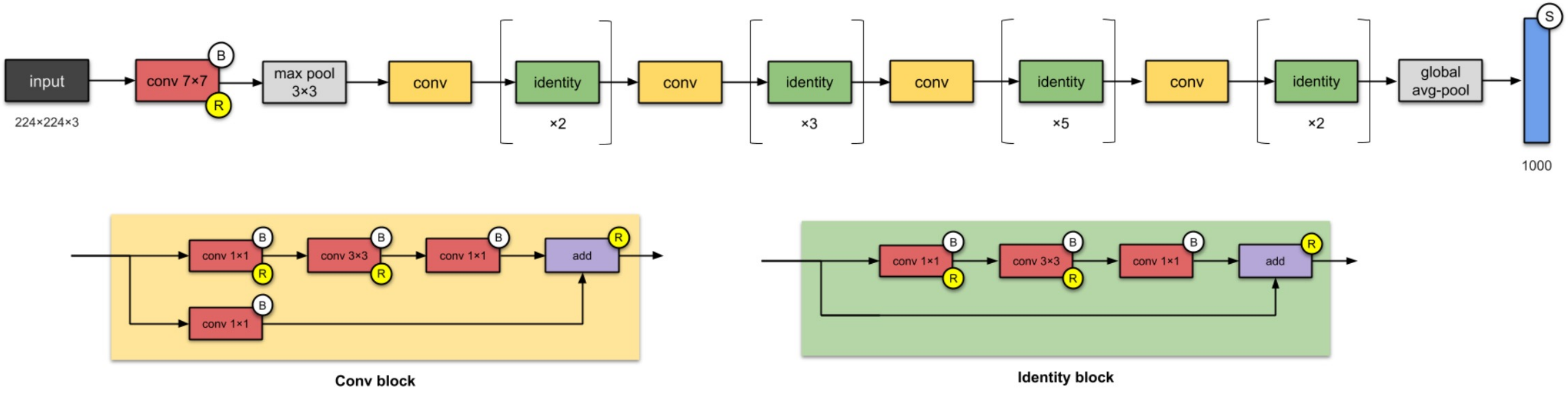
Imagenet:
1000 classes, 1.2M training images, 150K for testing

~3.6% top 5 error in ILSVRC 2015, lower than human recognition error!

Figure Credit: Zitzewitz, Gustav. "Survey of neural networks in autonomous driving." (2017)

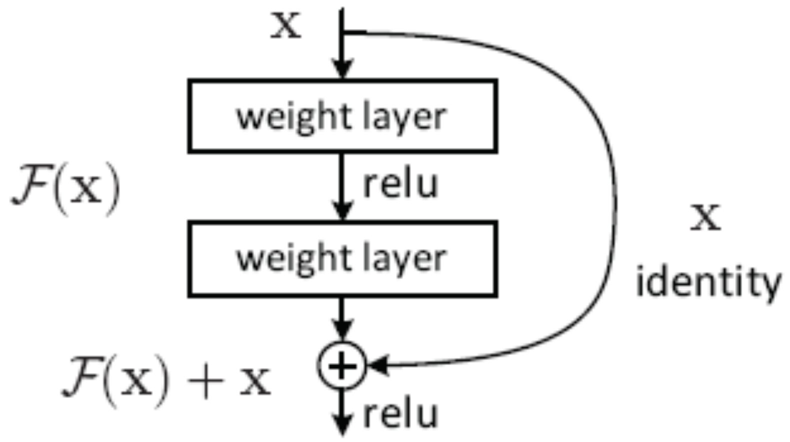
He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

CNN Architectures
ResNet (2015)



Novelty:

- Introduced residual learning (Residual blocks)
  - Shortcut connections with identity mapping
- Popularized skip connections
- 20 and 8 times deeper than AlexNet and VGG, respectively with less computational complexity and without compromising generalization power



He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Object Detection Architectures
YOLO (2016 - Ongoing)

All previous object detection techniques required multiple stages of detection

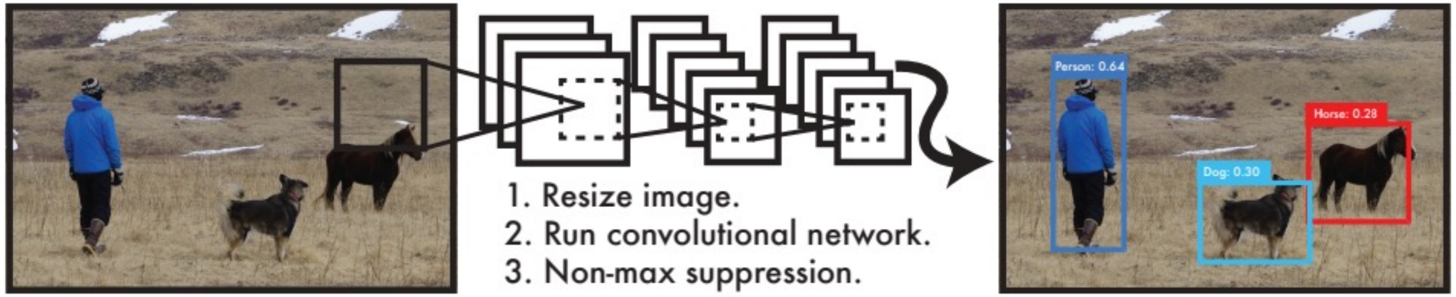


Figure 1: The YOLO Detection System. Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to 448 × 448, (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence.

Novelty:

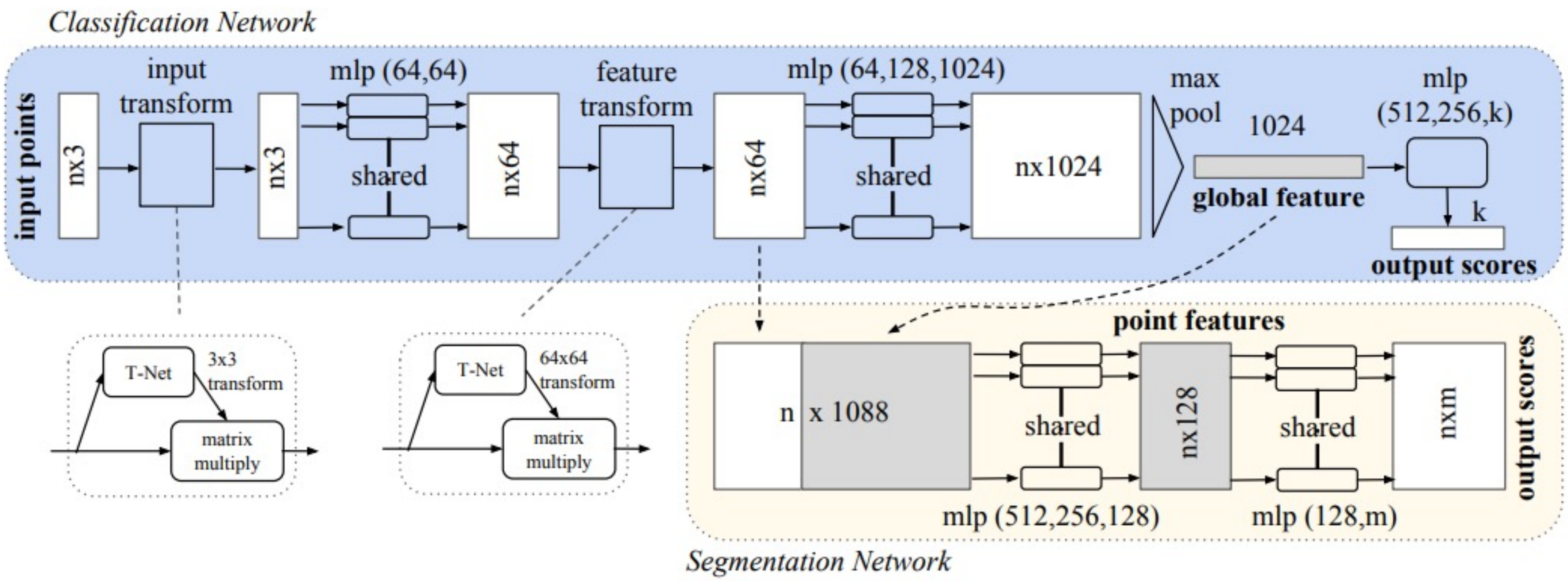
- Object detection is reformulated as a regression problem from image space to bounding-box coordinate space
- Single stage object detectors
  - Feature extraction, detection, classification performed in one go
- Contextual information is encoded within each prediction

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Deep Learning for LIDAR data

PointNet (2017)

The challenge in utilizing LIDAR data is the volume of point cloud data and the permutation of their processing

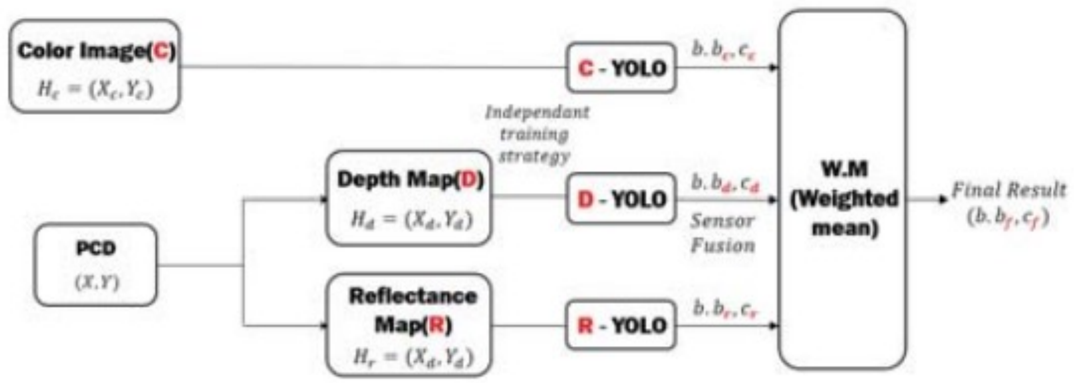


- Performed classification and segmentation on *n* points of LIDAR data. Input *nx3* refers to n points with $\{x, y, z\}$ coordinate dimensions
- Used RNNs to overcome the permutation issues within LIDAR data

Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Deep Learning for Sensor Fusion

Vision and LIDAR



YOLO Framework is used to independently extract features from cameras and LIDAR sensors and fused to detect missed boxes

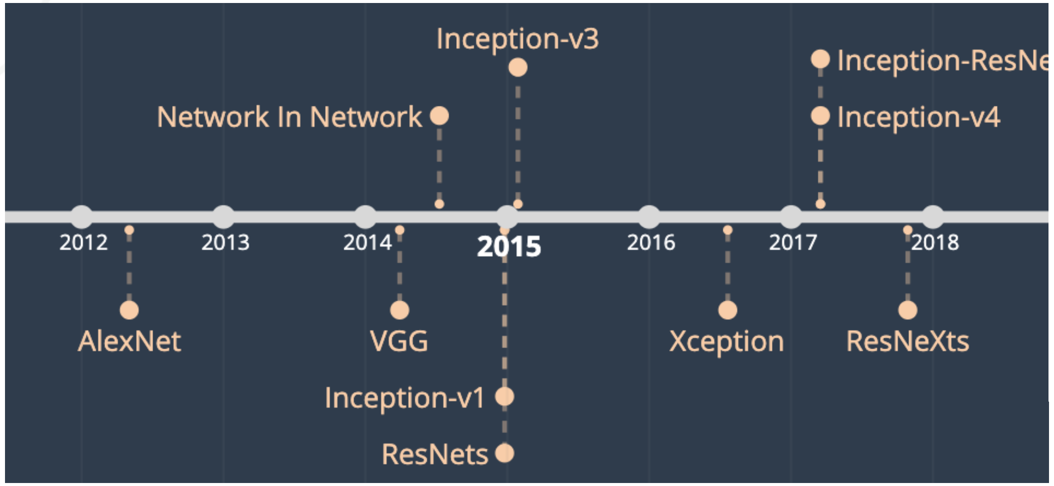
This is *'late fusion'*, in the sense that each sensor modality is independently evaluated

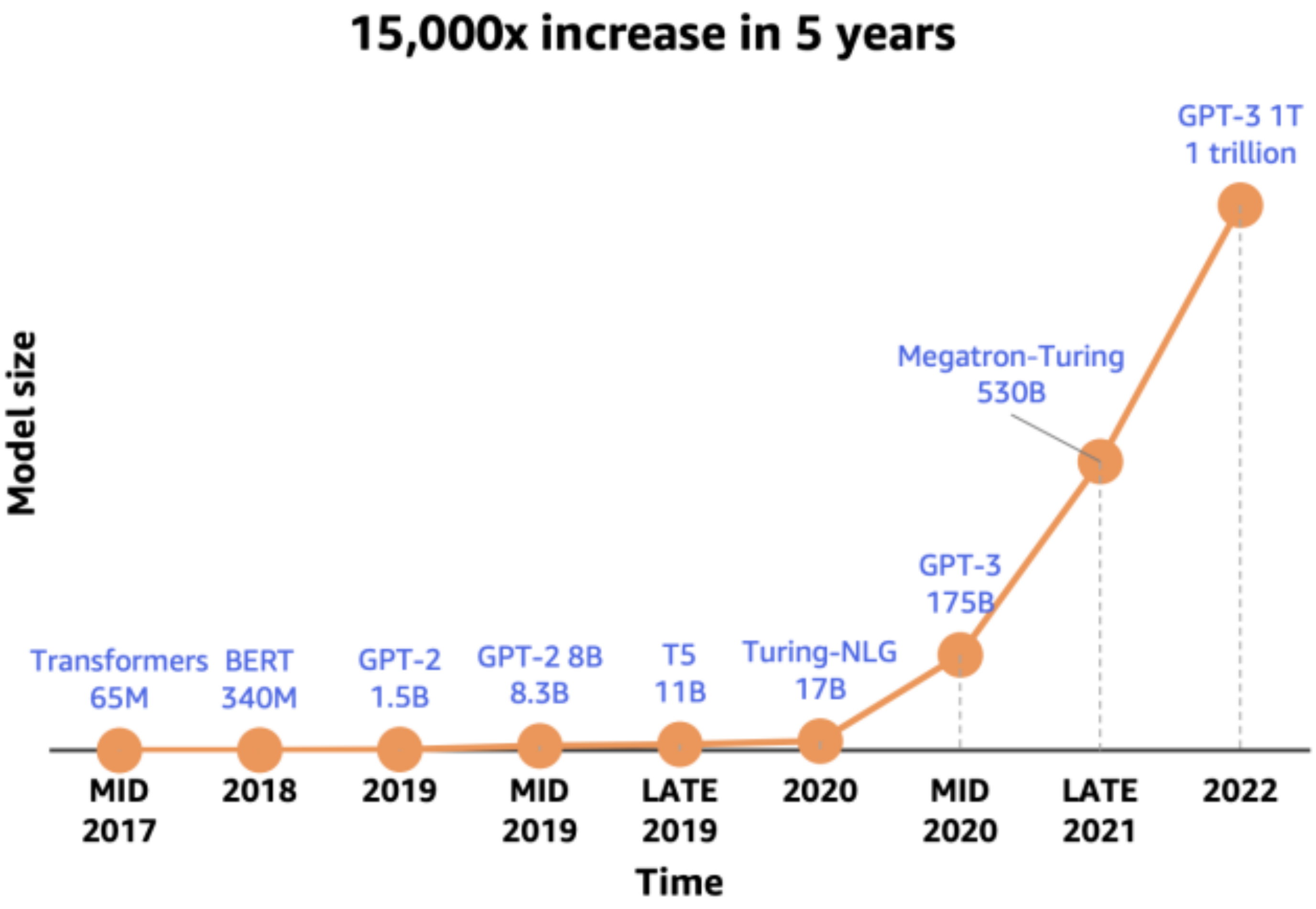
Kim, J., Kim, J., & Cho, J. (2019, December). An advanced object classification strategy using YOLO through camera and LiDAR sensor fusion. In *2019 13th International Conference on Signal Processing and Communication Systems (ICSPCS)* (pp. 1-5). IEEE.

Deep Deep Deep … Deep Deep Learning

Recent Advancements

The number of parameters in models has increased exponentially

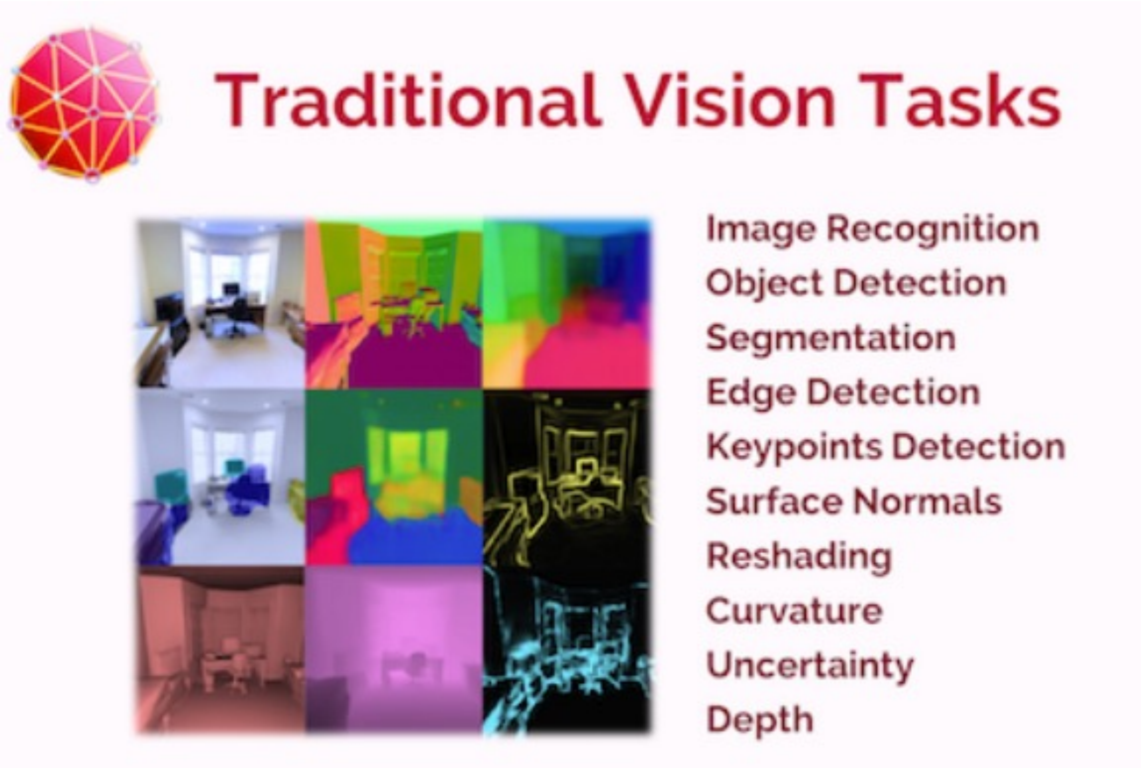




Deep Deep Deep … Deep Deep Learning

Motivation

Underlying features among different vision tasks are similar



This similarity leads to Transfer Learning

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

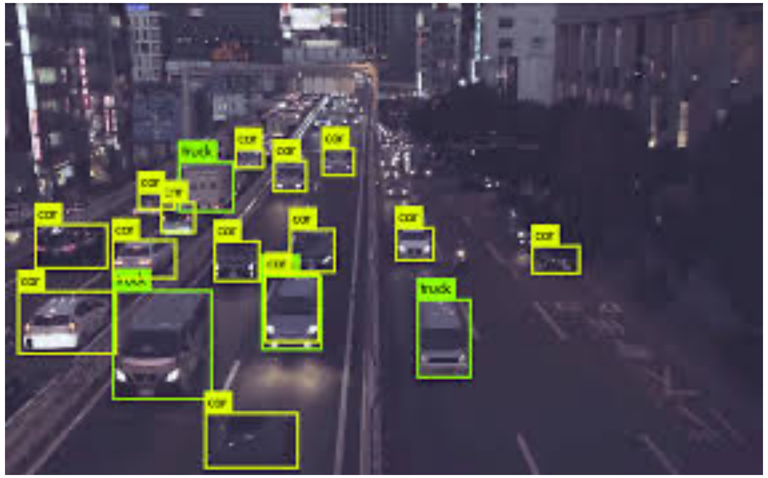
Transfer Learning
What is Transfer Learning?

- Deep networks tend to **learn common representations** for various tasks in their earlier layers
- Can be exploited **to transfer representations from networks trained on large datasets** on one task (i.e., Image Classification on ImageNet) called the *source* to a different task called the *target* task
- Usually done by **taking large pretrained network** and then **finetuning last layer** (with all other layers frozen) on target dataset
- **Pre-trained frozen backbone** acts as a **feature extractor** while **finetuned last layer** acts to project the representations into the **decision boundary for the target task**
- Utility depends on how closely related the source and target datasets and/or tasks are

Transfer Learning

Foundation Models

Source: https://gluon-cv.mxnet.io/



Source: https://www.move-lab.com/blog/tracking-things-in-object-detection-videos

Pretraining

Foundation Model

Finetuning

Source: https://www.saagie.com/blog/object-detection-part1/

Foundation Models

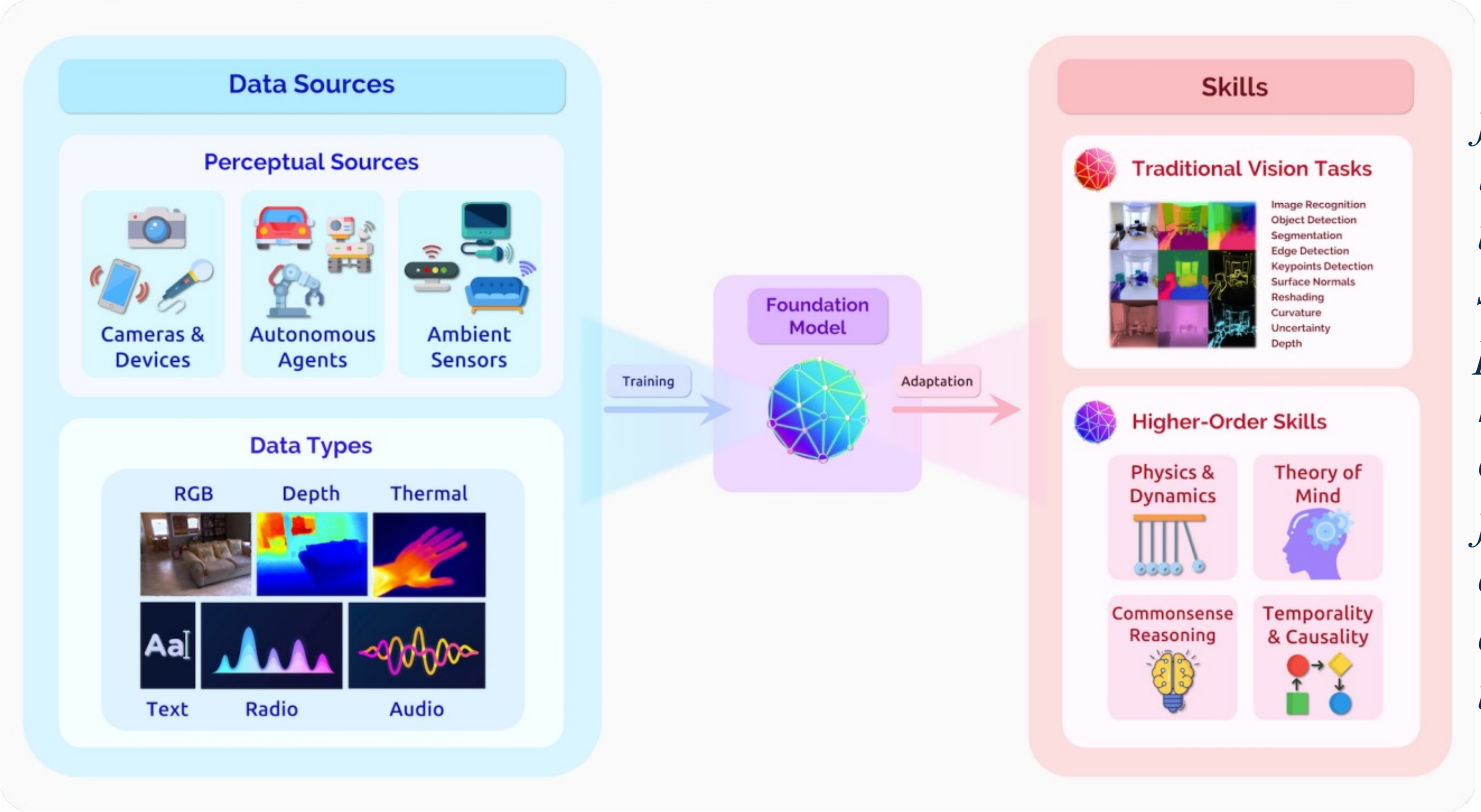
Origin of the term Foundation Models

- **Foundation models** are like any other deep network that have employed **transfer learning**, except **at *scale***
- ***Scale*** brings about ***emergent* properties** that are common between tasks
- **Before 2019:** Base architectures that powered multiple neural networks were **ResNets, VGG** etc.
- **Since 2019: BERT, DALL-E, GPT, Flamingo**
- Changes since 2019: **Transformer architectures and Self-Supervision**

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Foundation Models

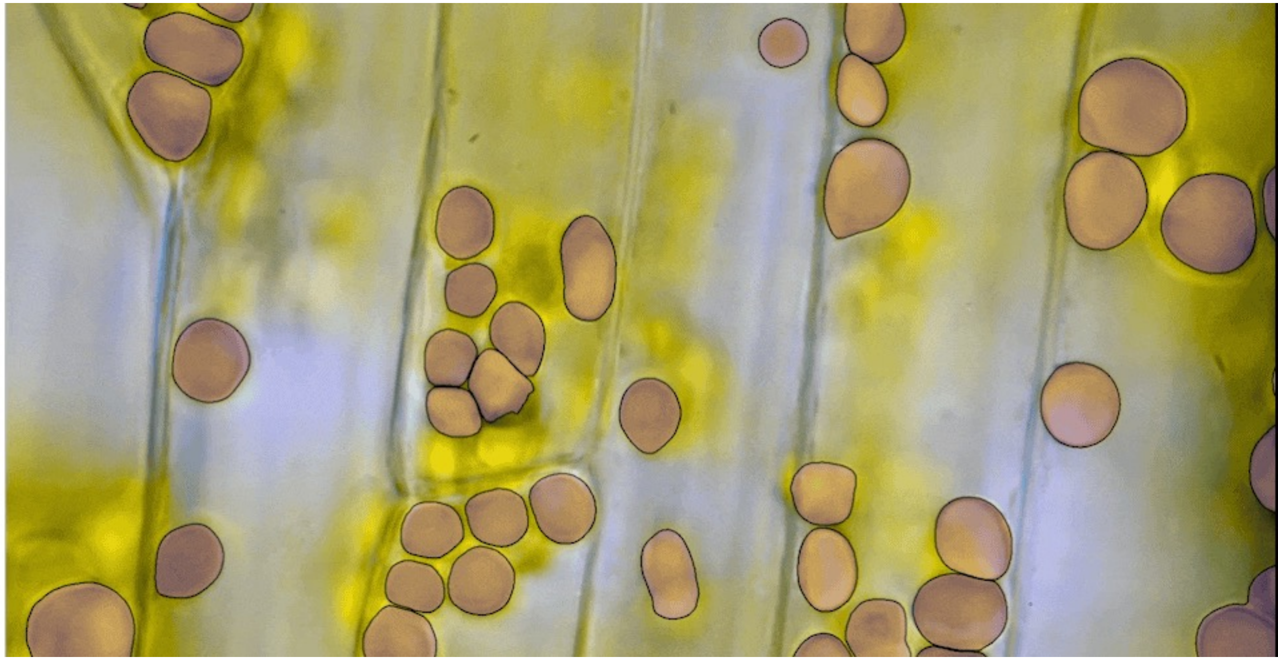
Origin of the term Foundation Models



*'By harnessing **self-supervision at scale**, foundation models for vision have the potential to **distill raw, multimodal sensory information into visual knowledge**, which may effectively support traditional **perception tasks** and possibly enable new progress on challenging higher-order skills like **temporal and commonsense reasoning** These inputs can come from a **diverse range of data sources** and application domains, suggesting promise for applications in **healthcare and embodied, interactive perception settings**'*

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Foundation Models
Segment Anything Model



Segment Anything Model (SAM) released by Meta on April 5, 2023 was trained on Segment Anything 1 Billion dataset with 1.1 billion high-quality segmentation masks from 11 million images

Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao et al. "Segment anything." *arXiv preprint arXiv:2304.02643* (2023).

Foundation Models
Segment Anything Model

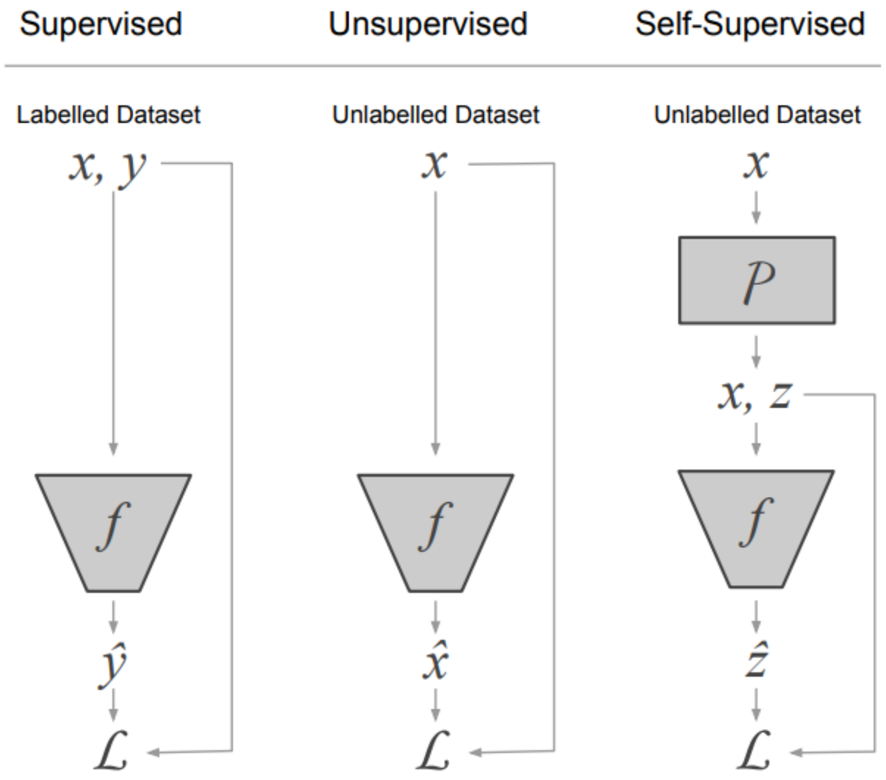
Cityscapes dataset semantic segmentation annotation took ~90 mins per image

Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao et al. "Segment anything." *arXiv preprint arXiv:2304.02643* (2023).

Foundation models are trained via Self-Supervision

Self-Supervision:

- Type of unsupervised learning
- Primary difference is the introduction of a **“pre-text task.”**
- The pre-text task generates pseudo-labels that are used to train a network.



Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2021). Self-Supervised Representation Learning: Introduction, Advances and Challenges. *arXiv preprint arXiv:2110.09327*.

Self-Supervision
Overall Training Process

1. **Identify Labeled and Unlabeled Data**

Unlabeled Data
$(x_1 \dots x_N)$

Labeled Data
$(x_1 \dots x_M), (y_1 \dots y_M)$

2. **Generate pseudo-labels with some pre-text task P**

Unlabeled Data
$(x_1 \dots x_N)$

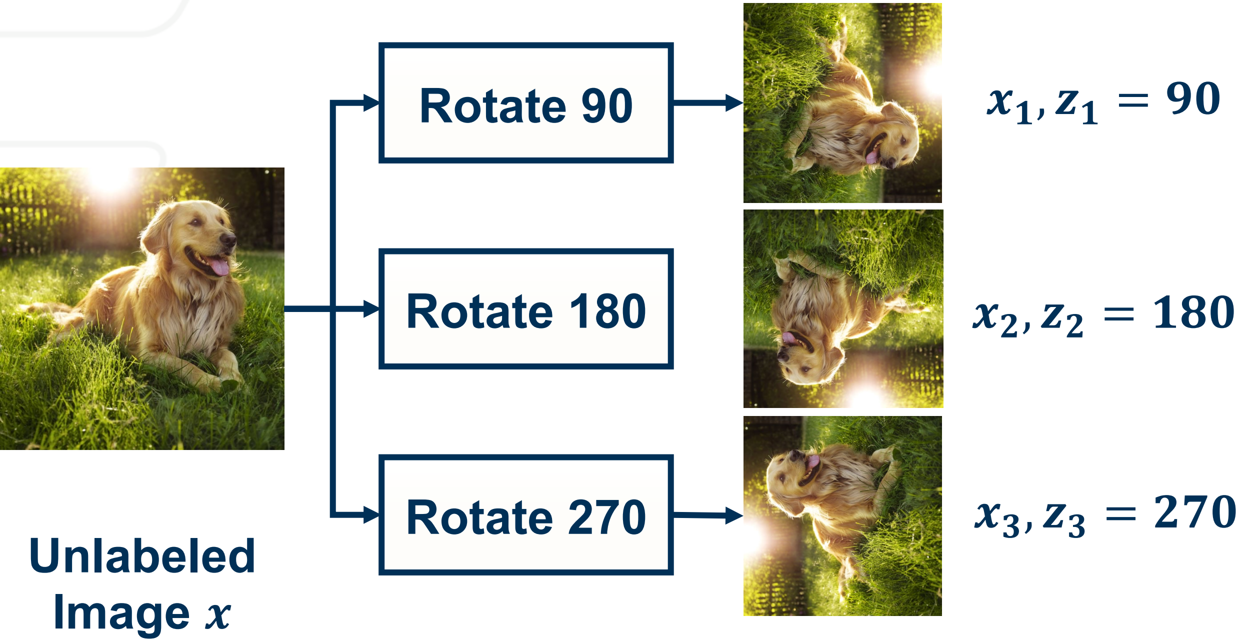
$\Rightarrow$ P $\Rightarrow$

Pseudo - Labeled Data
$(x_1 \dots x_N), (z_1, \dots, z_M)$

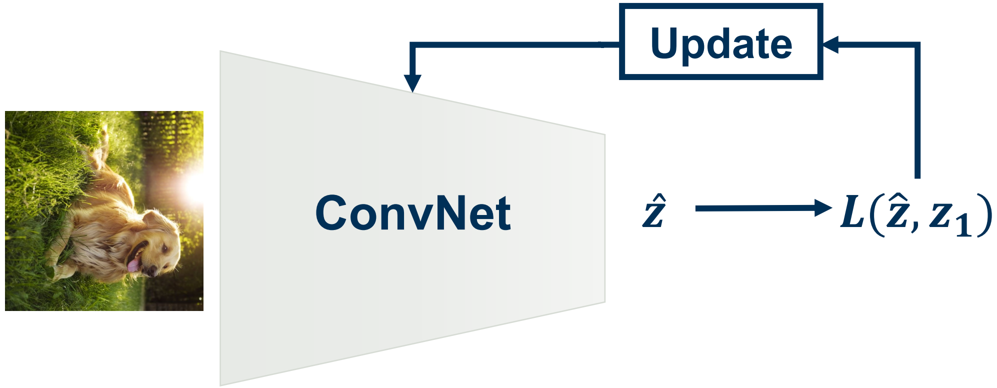
Self-Supervision

Example Training Process

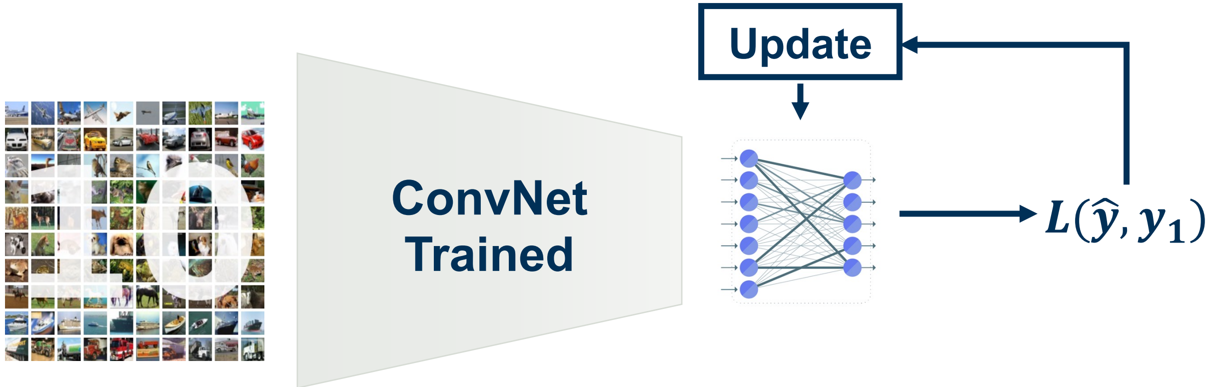
Step 1: Generate pseudo-labels via image rotations



Step 2: Network learns to predict angle image is rotated



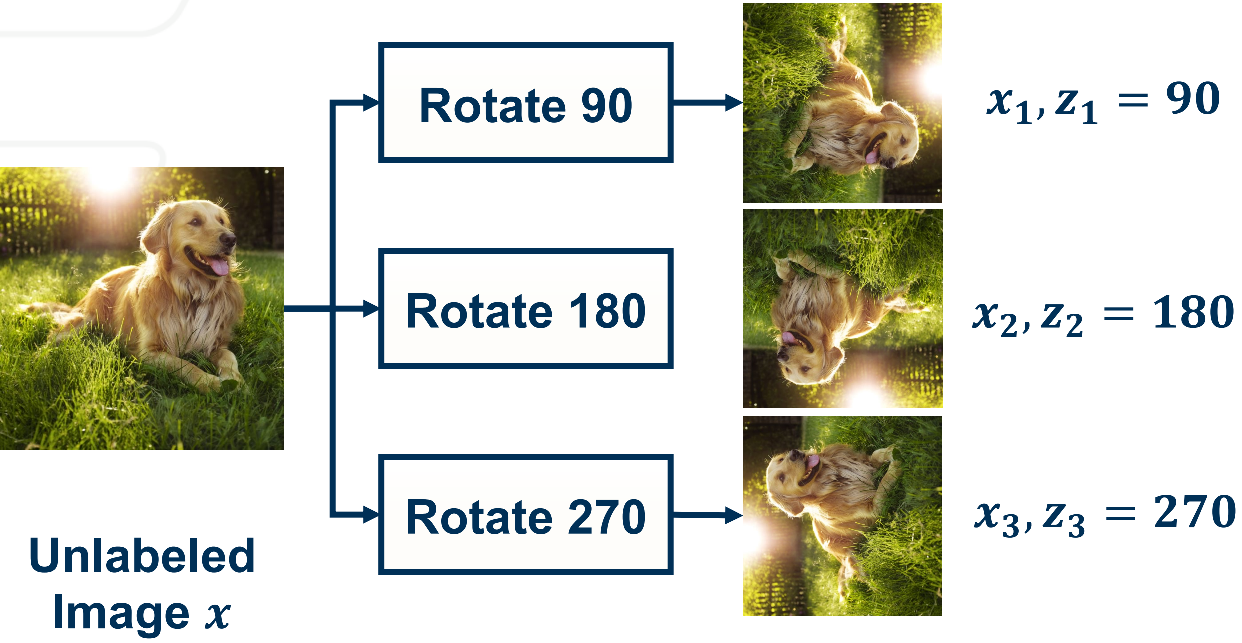
Step 3: Attach linear layer and train to classify labels (y) on labeled dataset



Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Self-Supervision
Motivation

Step 1: Generate pseudo-labels via image rotations



Learning pre-text task will allow network to learn relevant features without needing explicit labels!

Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Self-Supervision
Types of Pre-text Tasks

Differences in self-supervision are based on the type of pre-text task that is defined

Transformation Prediction

- Pre-text task performs some transformation on data and tasks model with trying to learn nature of transformation.

Masked Prediction

- Pre-text task removes some part of the data and the model is tasked with trying to predict what was removed.

Deep Clustering

- Identify clusters of features and iteratively assign pseudo-labels to train model.

Contrastive Learning

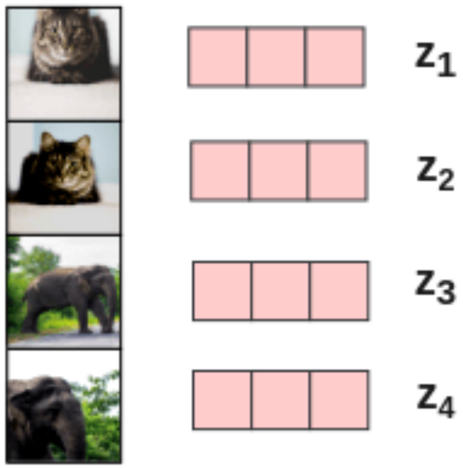
- Pre-text task identifies positive and negative pairs of data and the model is tasked with learning similarities to discriminate between positive and negatives.

Contrastive Learning
Sim-CLR Framework

The Pseudo-labels are used to create positive-negative pairs within each batch

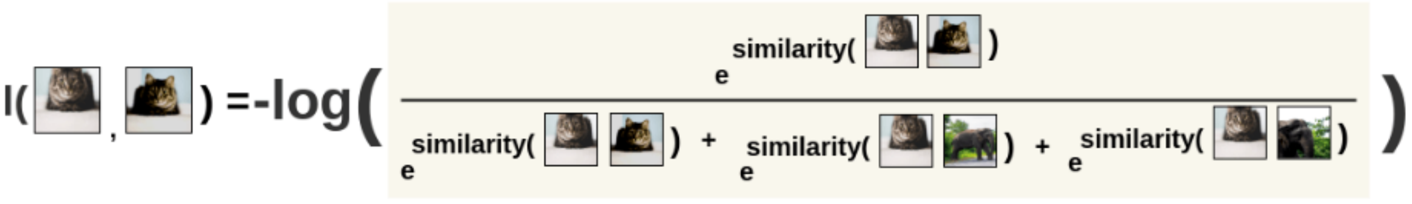
Calculated Embeddings

Batch Augmented Images



Contrastive loss on embeddings

$$l(i,j) = -log\frac{exp(s_{i,j})}{\sum_{k=1}^{2N} l_{[k!=i]}exp(s_{i,k})}$$

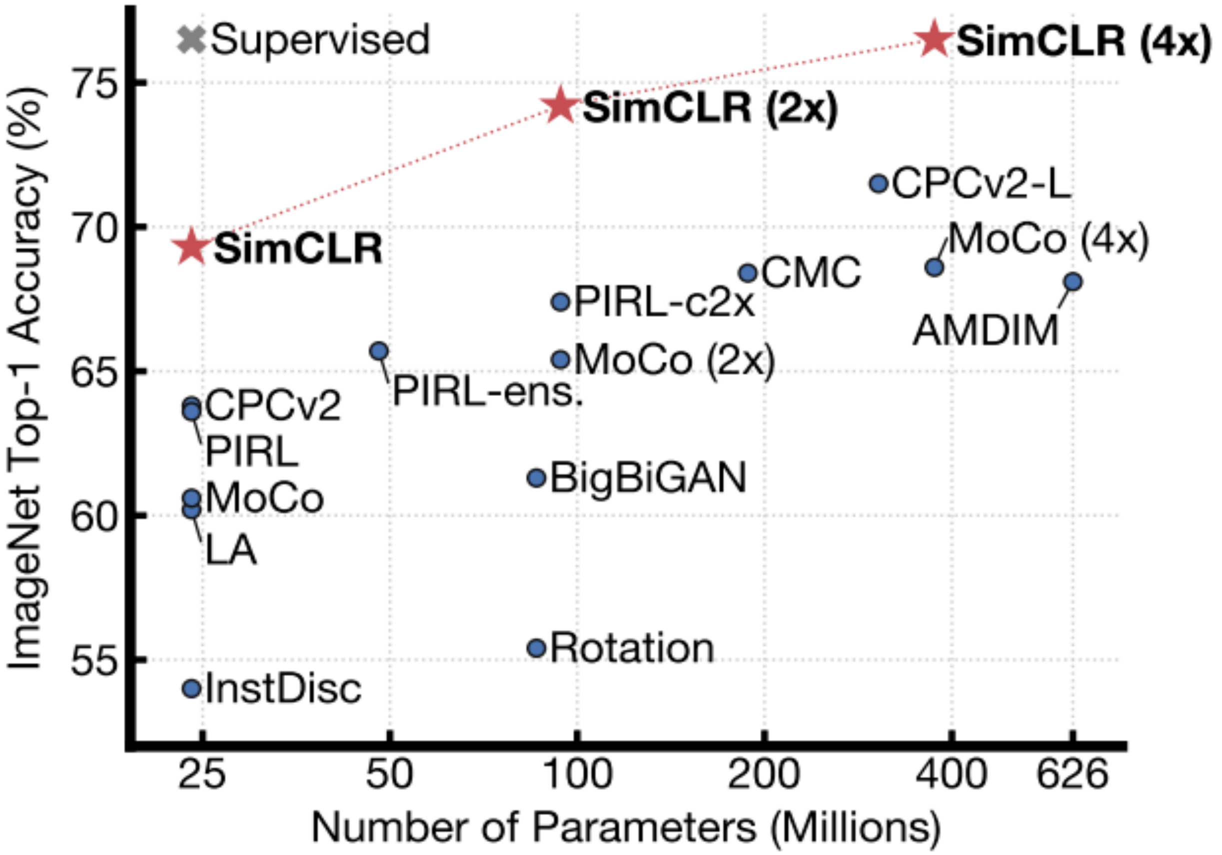


Note: The positive pairs are only the augmentations and negative pairs are all other images in the batch

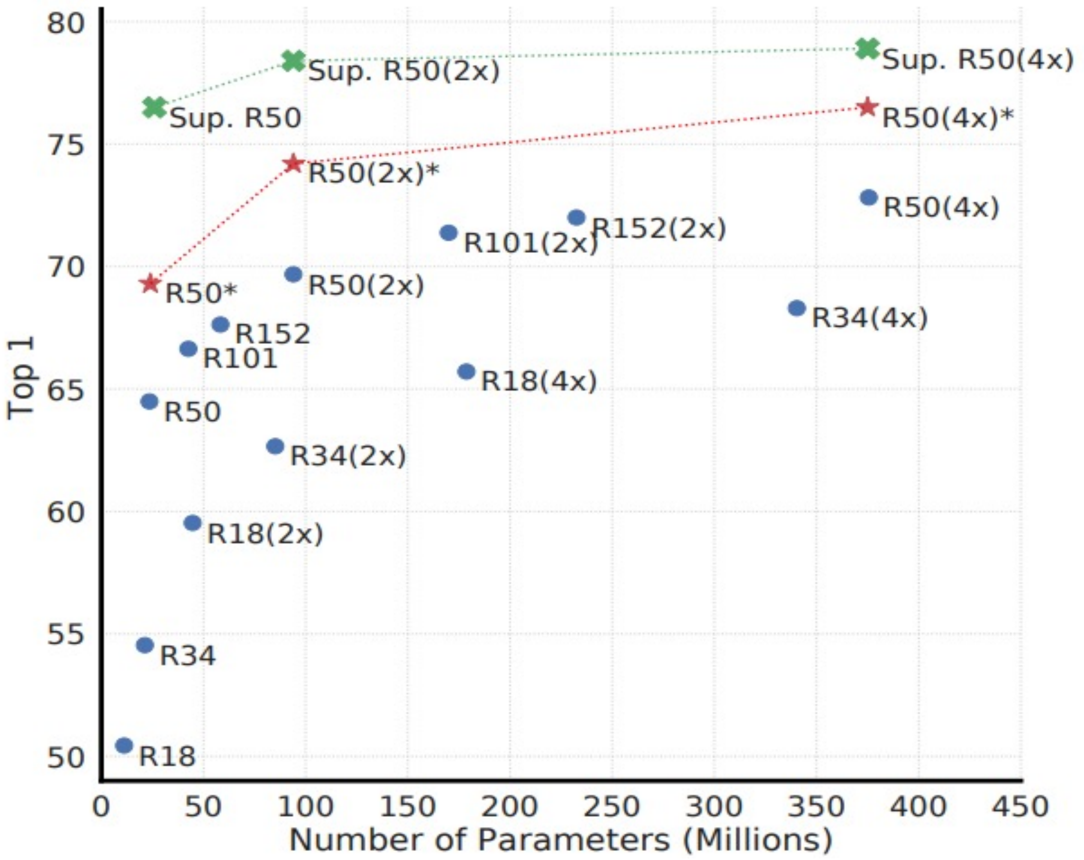
Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).

Contrastive Learning
Contrastive Learning vs Supervised Learning

Performance vs Models



Performance vs Parameters



Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).

Contrastive Learning
Contrastive Learning other than SIM-CLR

What differentiates other Contrastive Learning methods from Sim-CLR?

The way that similar pairs (positives) and dissimilar pairs (negatives) are generated.

Paper	Short description	Topics of contribution
Becker and Hinton [8]	Maximise MI between two views	Foundation
Bromley et al. [11]	Siamese network in metric learning setting	Foundation
Chopra, Hadsell, and LeCun [20]	Learn similarity metric with contrastive pair loss	Energy-based loss, Application
Hadsell, Chopra, and LeCun [39]	Learn invariant representation from pair loss	Energy-based loss, Application
Weinberger, Blitzer, and Saul [108]	Learn distance metric with triplet loss	Energy-based loss
Collobert and Weston [21]	Learn language model with triplet loss	Application
Chechik et al. [15]	Learn image retrieval model with triplet loss	Application
Noise Contrastive Estimation [38]	Introduce NCE, a general methods to learn unnormalised probabilistic model	Probabilistic loss
Mnih and Teh [71]	Learn language model with NCE-based loss	Application
Mikolov et al. [68]	Learn word embedding with Negative Sampling (NEG), a modified version of NCE	Probabilistic loss, Application
Wang et al. [105]	Learn fine-grained image similarity using deep network and triplet loss	Application
Wang and Gupta [107]	Use video's sequential coherence to learn unsupervised video representation	Similarity, Application
Lifted-structure loss [75]	Extend triplet loss to multiple positive and negative pairs per query	Energy-based loss
N-pair loss [92]	Proposed non-parametric classification loss with multiple negative pairs per query	Probabilistic loss
Wu et al. [109]	Focus on the quality of negative samples through a distance-weighted margin loss	Similarity, Energy-based loss
Hermans, Beyer, and Leibe [45]	State the important of mining hard samples in triplet loss	Similarity
Wu et al. [110]	Self-supervised representation with instance discrimination Memory bank to holds keys for next epoch	Application Encoder
CPC [77]	Mutual Information with the contrastive loss Define similarity with past-future context-instance relationship	Mutual Information loss Similarity
DIM [46]	Evaluate multiple mutual information bound for the contrastive loss Global-local context-instance relationship	Mutual Information Loss Similarity
MoCo [43]	Use momentum encoder to store features to memory queue	Encoder
SimCLR [16]	Simplify and demonstrate large empirical improvement in instance discrimination task Focus on the use of separate heads	Application Transform heads
BYOL [34]	Learning similarity without negative samples	Loss

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).

IEEE Open Journal of Signal Processing

Exploiting the Distortion-Semantic Interaction in Fisheye Data

Kiran Kokilepersaud, PhD Student

Mohit Prabhushankar, PhD Postdoc

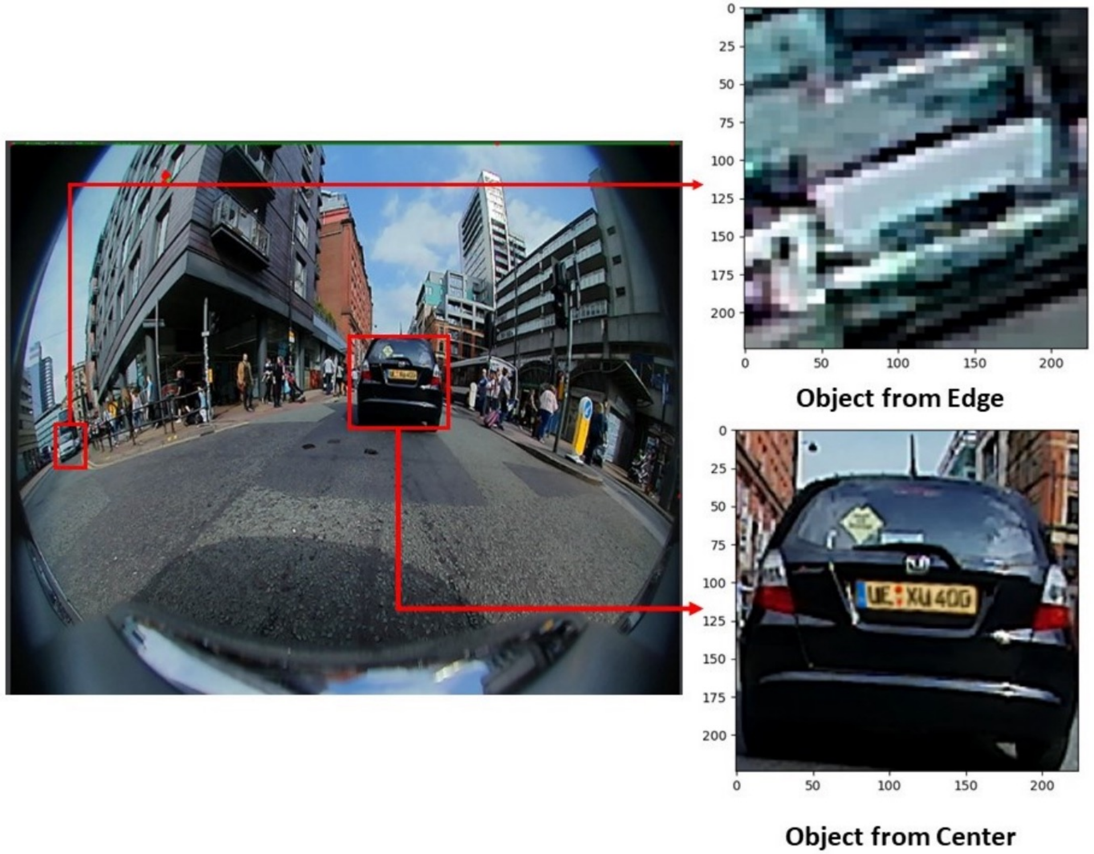
Ghassan AlRegib, PhD Professor

SCAN ME

Contrastive Learning for Fisheye Images

Positive-negative pairs in Fisheye Images

Intuition: Regions within a fisheye image are their own class. Hence, any object within them are positives



Intuition for Loss 1:

All objects from the edge (be it a car, bike, pedestrian) are positives and objects from the centre (be it a car, bike, pedestrian) are negatives

Intuition for Loss 1:

All objects from labeled car (be it in the center or the edge) are positives and all other objects (be it in the center or the edge) are negatives

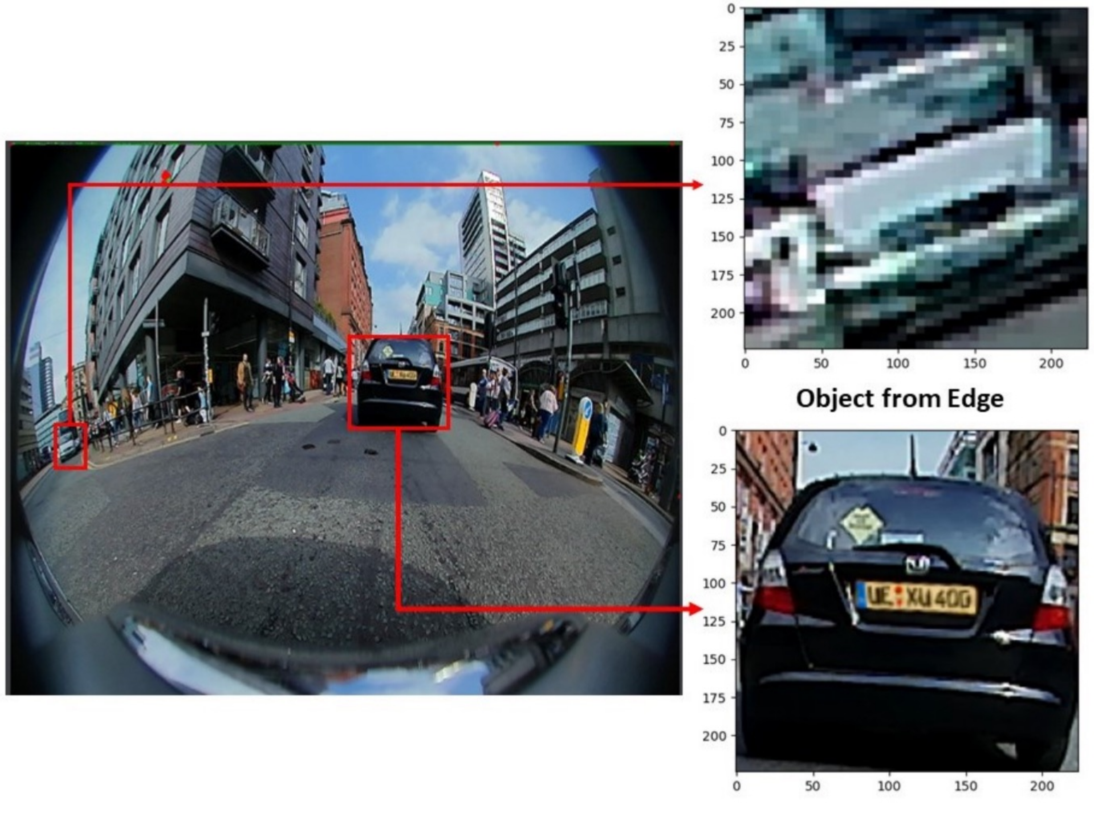
Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).

Contrastive Learning for Fisheye Images
Positive-negative pairs in Fisheye Images

Exploiting the Distortion-Semantic Interaction in Fisheye Data

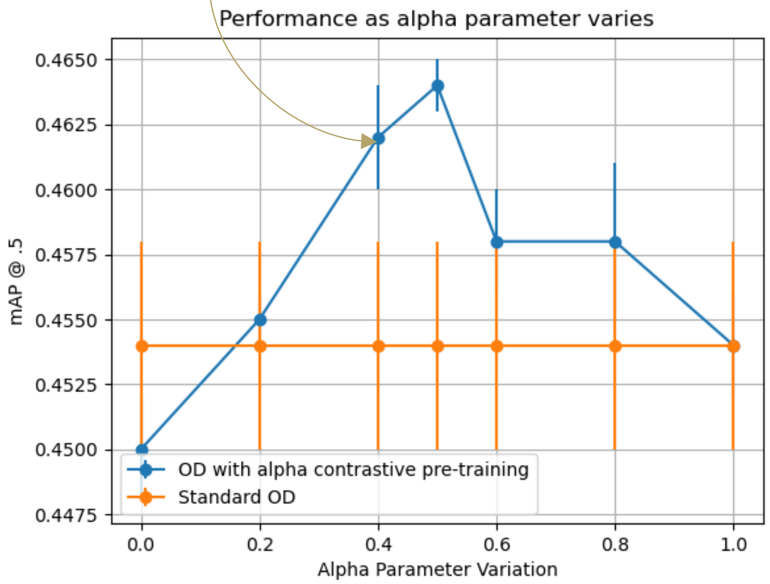
SCAN ME

Intuition: Regions within a fisheye image are their own class. Hence, any object within them are positives



$$\alpha L_{class} + (1-\alpha) L_{RegionClass}$$

α controls the level of unsupervised contrastive learning



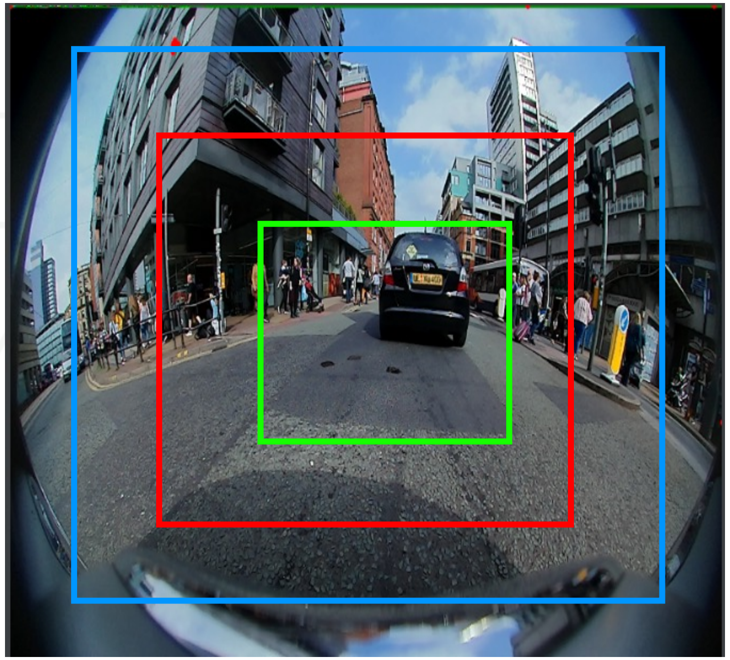
Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).

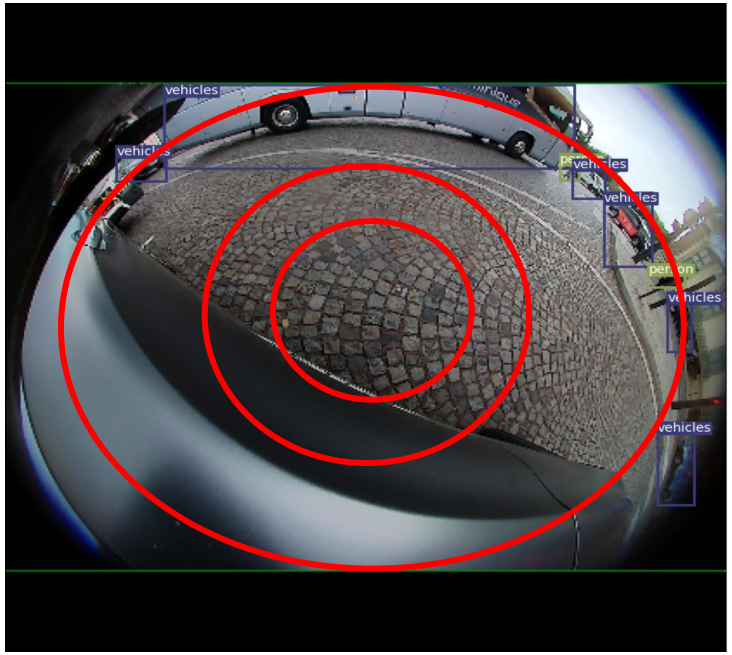
Contrastive Learning for Fisheye Images
Positive-negative pairs in Fisheye Images

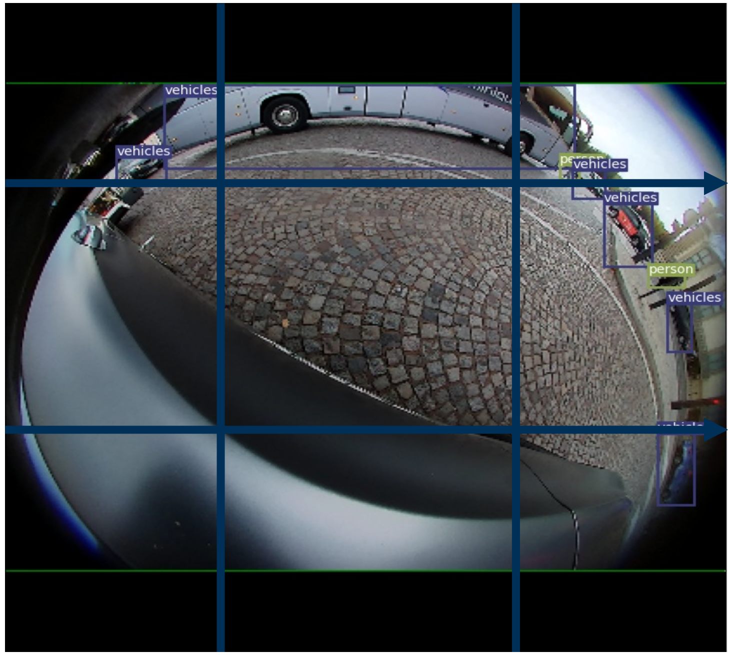
Exploiting the Distortion-Semantic Interaction in Fisheye Data

SCAN ME

Are there alternative ways of partitioning the regions?







Defining the positive-negative pairs is application dependent

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).

Objectives

Takeaways from Part II

- Part I: Challenges in Perception and Autonomy
- **Part II: Deep Learning for Perception**
  - Transfer Learning and training at scale are essential for foundation model development
  - Self-supervised Learning provides a framework for large scale learning on unannotated data
- Part III: Existing Deep Learning solutions to Challenges in Perception
- Part IV: Remaining Challenges and Future Directions