Deep Neural Networks II.

Jan Čech

Lecture Outline

1. Deep neural networks for Object detection
2. Deep neural networks for Segmentation
3. “Deeper” insight into the Deep Nets
4. Foundation models

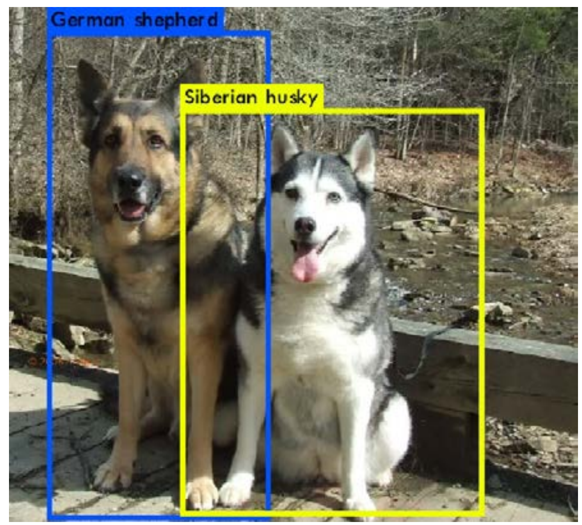
Deep Neural Networks for Object Detection

Convolutional Networks for Object Detection

- What is the object detection?

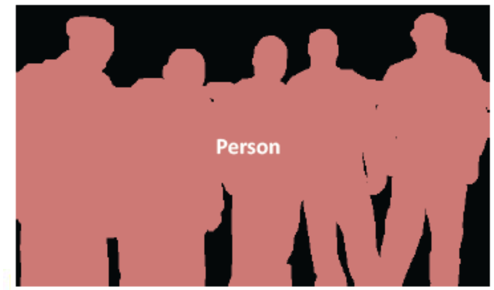
Grocery store

Image recognition
- What?
- holistic



Object detection
- What + Where?
- Bounding boxes

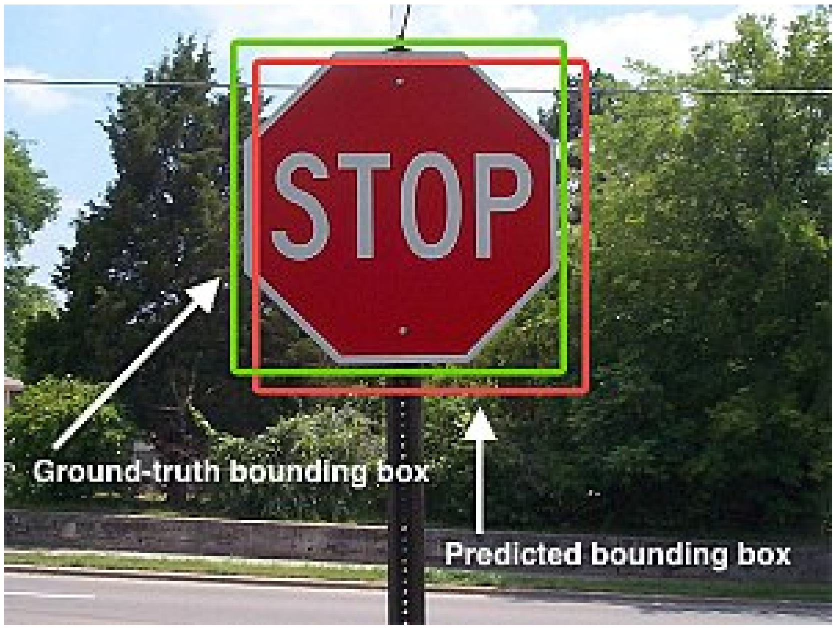
Semantic segmentation
- What + Where?
- Pixel-level accuracy

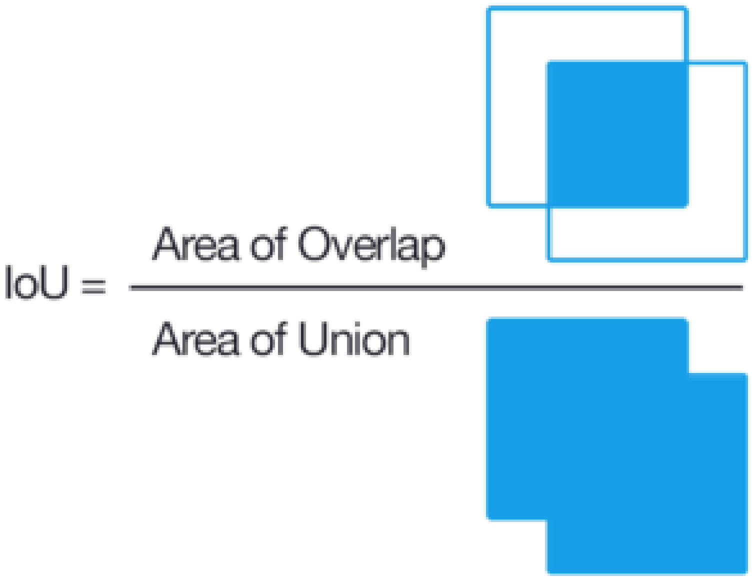


Instance segmentation
- What instance + Where
- Pixel-level accuracy

How to measure detector accuracy?

- Ground-Truth bounding boxes, Detections – predicted bounding boxes
- Intersection over Union (IoU), a.k.a. Jaccard index

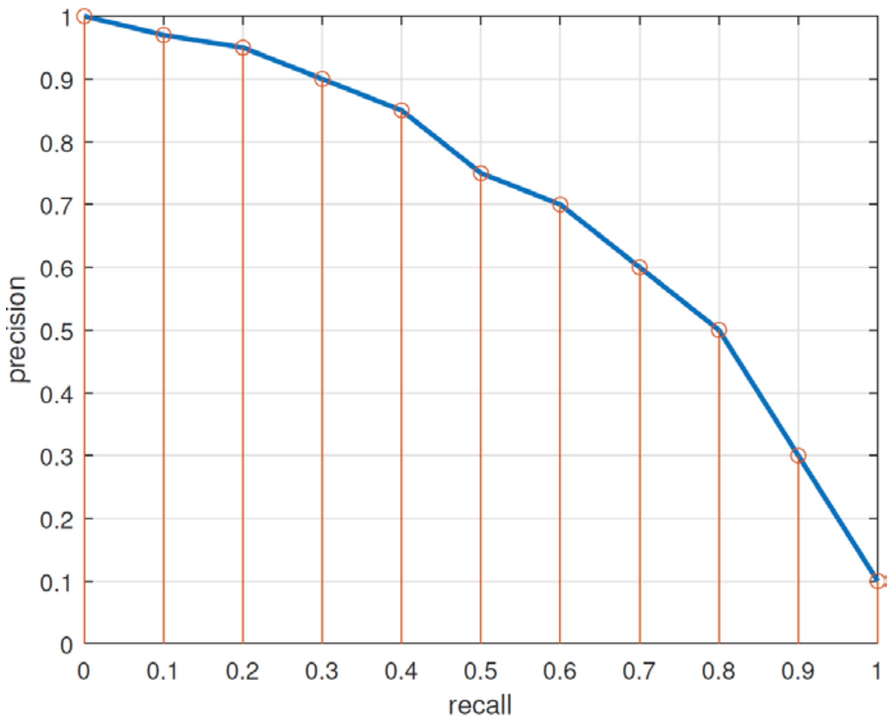


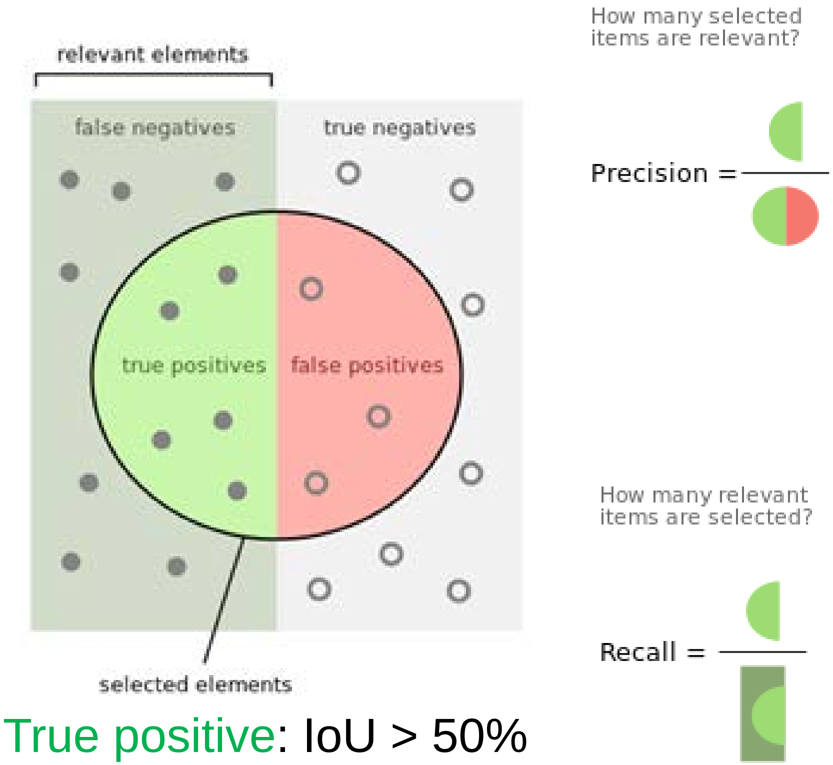


- A detection is correct (= true positive) if it has enough overlap with the ground-truth
  – Typically, IoU > 50%

How to measure detector accuracy?

- Mean Average Precision (mAP)





True positive: IoU > 50%

– Average Precision (Area under the precision-recall curve)

$$\mathrm{AP} = \int_r p(r) dr \approx \frac{1}{N} \sum_i p(r_i)$$

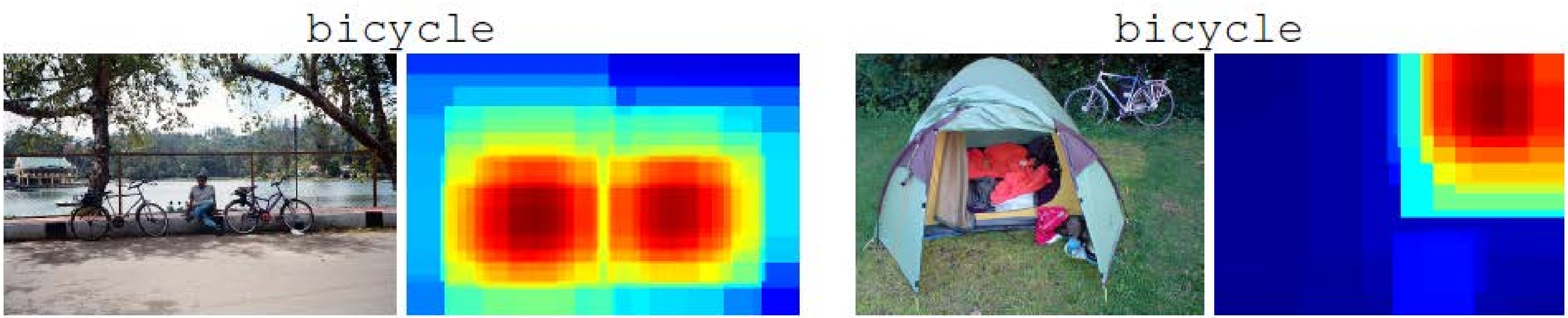
– Mean over all classes

$$\mathrm{mAP} = \frac{1}{C} \sum_c \mathrm{AP}_c$$

Pascal VOC 2007 challenge
(N = 11, r = 0:0.1:1)
(C = 20)
Classes: Person, bird, cat, car, ...

1. Scanning window + CNN

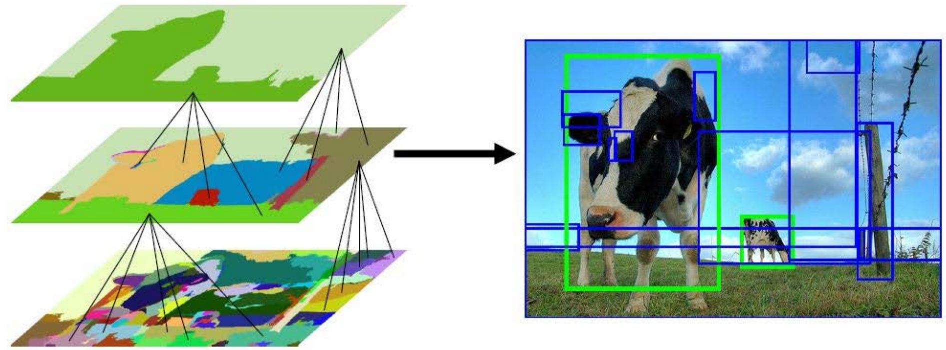
- CNN - Outstanding recognition accuracy of holistic image recognition [Krizhevsky-NIPS-2012]
- A trivial detection extension - exhaustive scanning window
  1. Scan all possible bounding boxes
  2. Crop bounding box, warp to 224x224 (fixed-size input image)
  3. Run CNN
- Works, but
  – prohibitively slow…



Oquab et al. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks, CVPR, 2014.

2. Region proposals + CNN

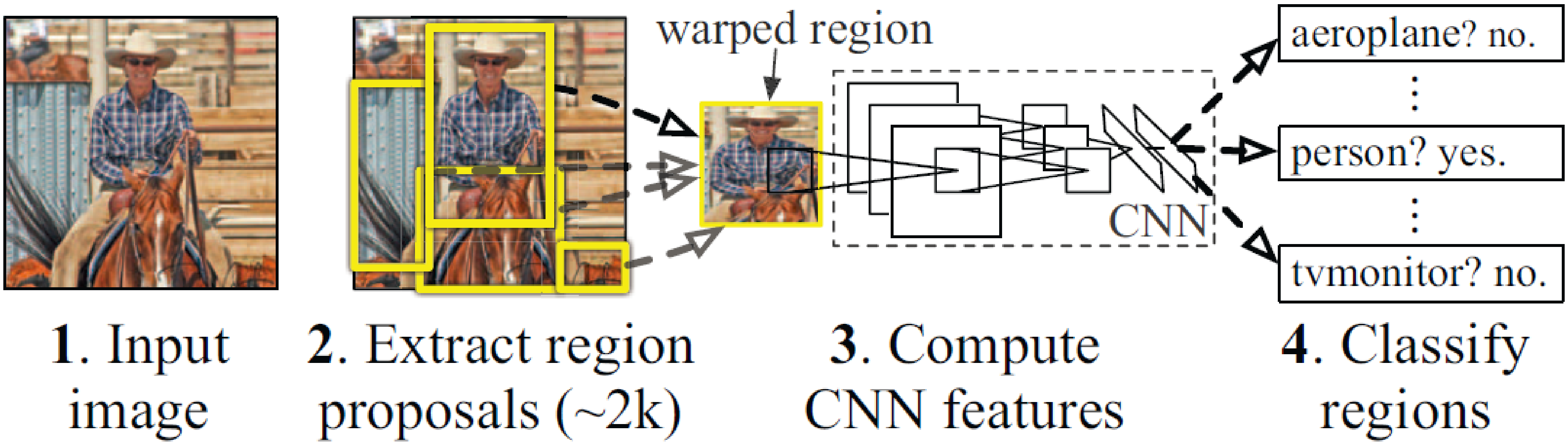
- CNN not evaluated exhaustively, but on regions where objects are likely to be present
- Region proposals (category independent):
  - Selective search [Uijlings-IJCV-2013]



  - Edgeboxes [Zitnick-ECCV-2014]

2. Region proposals + CNN

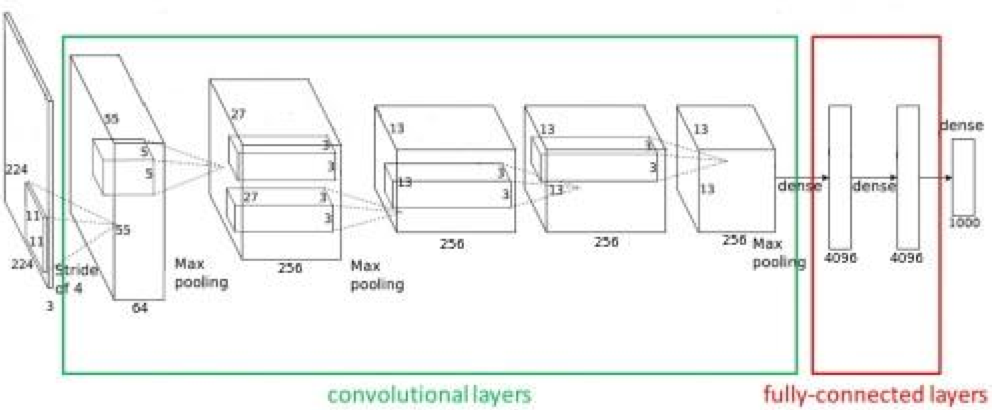
- R-CNN "Regions with CNN feature"
  - Girshick et al. *Rich feature hierarchies for accurate object detection and semantic segmentation*. CVPR 2014.

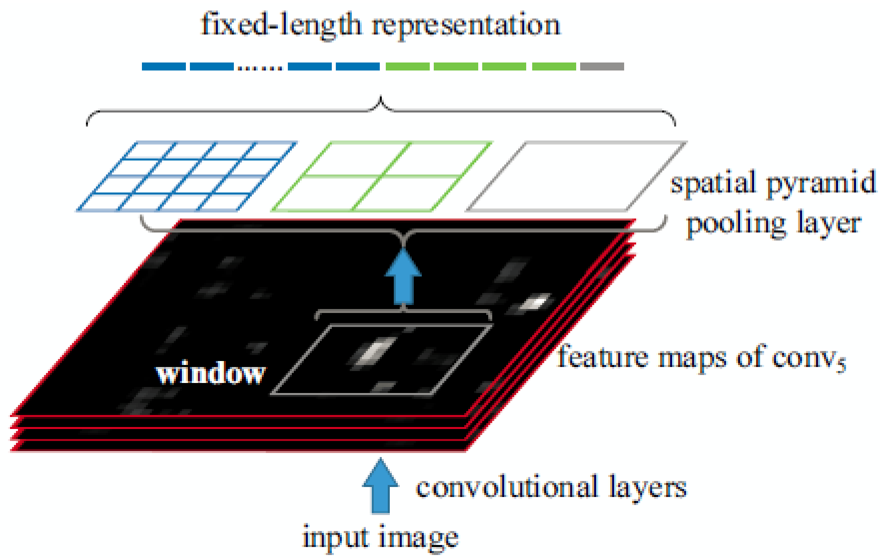


- Highly improved SotA on Pascal VOC 2012 by more than 30% (mAP)
- Still slow
  - For each region: crop + warp + run CNN (~2k)
  - 47 s/image

2. Region proposals + CNN

- Idea (1):
  - Do not run the entire CNN for each ROI, but
    - run convolutional (representation) part once for the entire image and
    - for each ROI pool the features and run fully connected (classification) part
  - He et al. *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recogniton*. ECCV 2014.



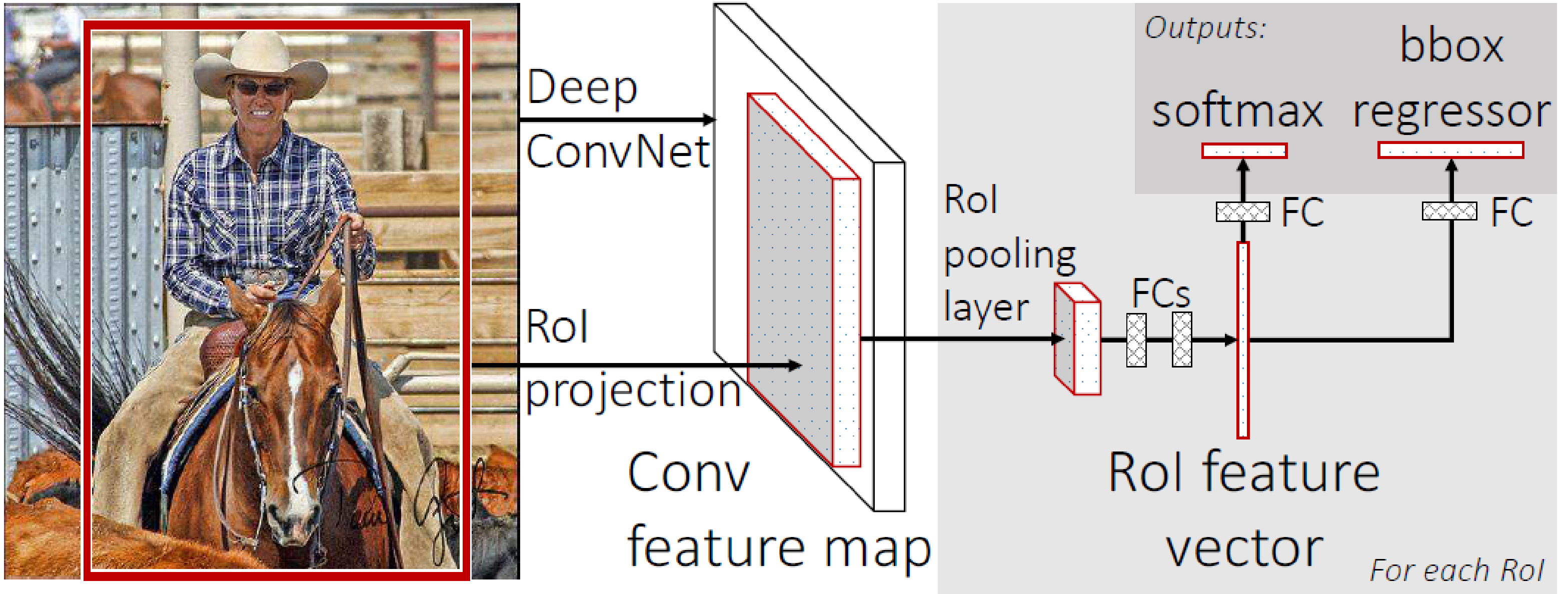


  - Arbitrary size image => fixed-length representation
  - Implemented by max-pooling operations
  - Speeds testing up

2. Region proposals + CNN

- Idea (2):
  – Refine bounding box by regression
  – Multi-task loss: classification + bounding box offset

- Fast R-CNN (= R-CNN + idea 1 + idea 2)
  – Girshick R. Fast R-CNN, ICCV 2015.



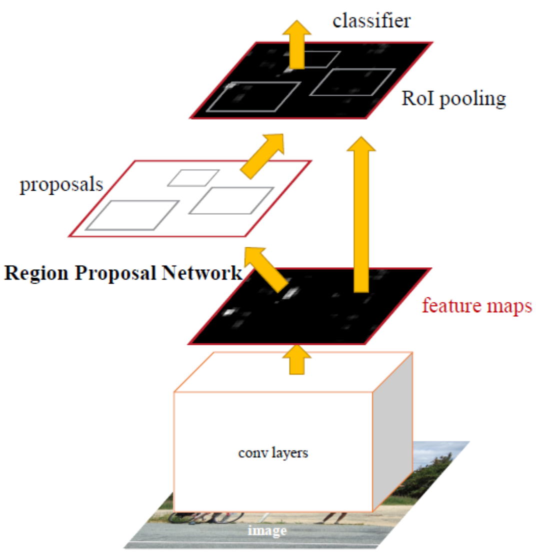
  – End-to-end training
  – Speed up, but proposals still expensive

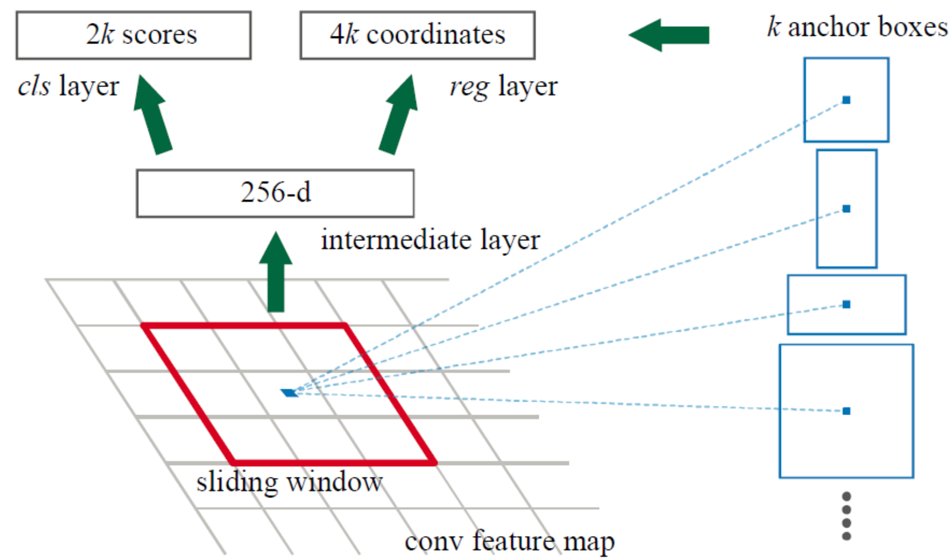
2. Region proposals + CNN

- Idea (3):
  - Implement region proposal mechanism by CNN with shared convolutional features (RPN + fast R-CNN)

⇒ Faster R-CNN

  - Ren et al. Faster R-CNN: *Towards Real-Time Object Detection with Region Proposal Networks*. NIPS 2015.
  - Region proposal network: object/not-object + bb coord. (k-anchor boxes)

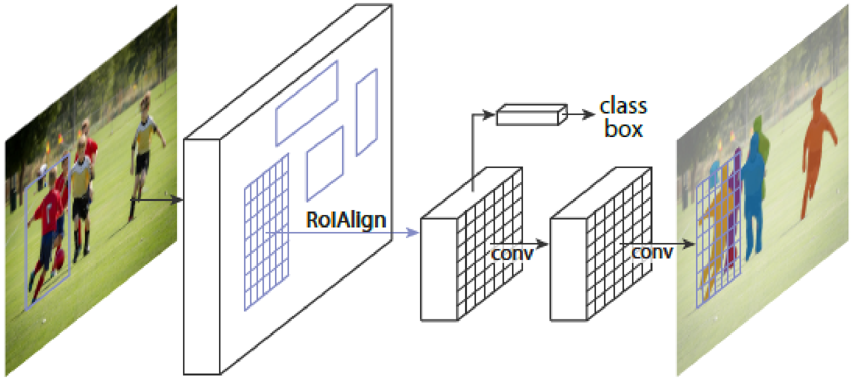


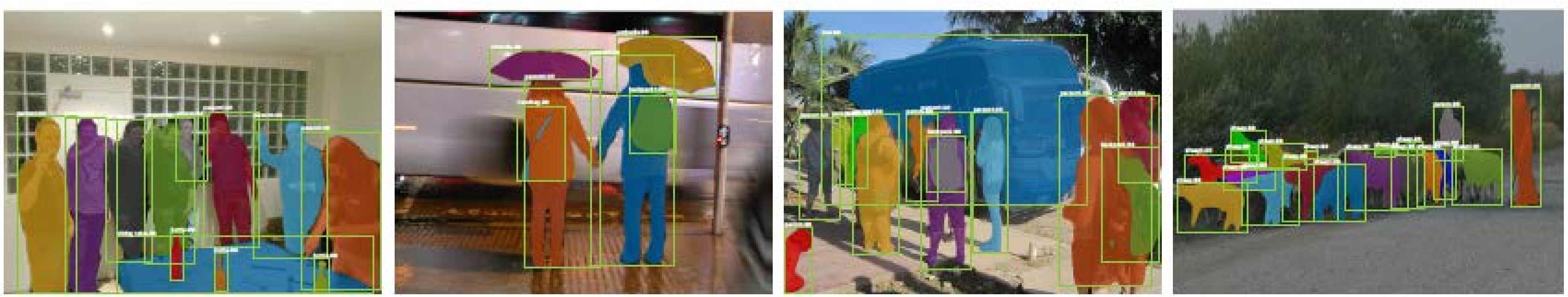


  - Training: simple alternating optimization (RPN, fast R-CNN)
  - Accurate: 73.2% mAP (VOC 2007), Fast: 5 fps

2. Region proposals + CNN + Instance segmentation

- Mask R-CNN
  - He et al., Mask R-CNN. ICCV 2017
  - Faster R-CNN + fully convolutional branch for segmentation
  - ROI alignment
    - Improved pooling with interpolation
  - Running 5 fps





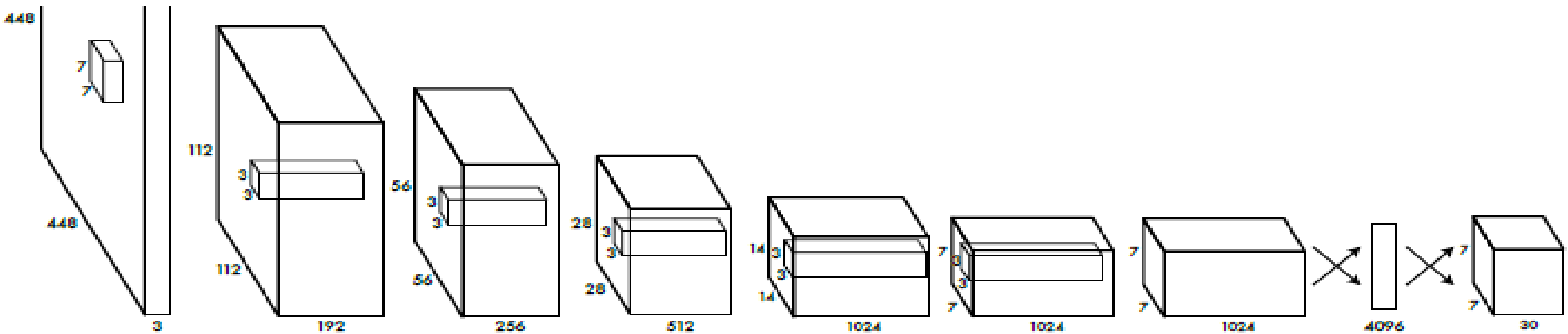
COCO dataset “Common Object in Context” (>200K images, 91 categories)

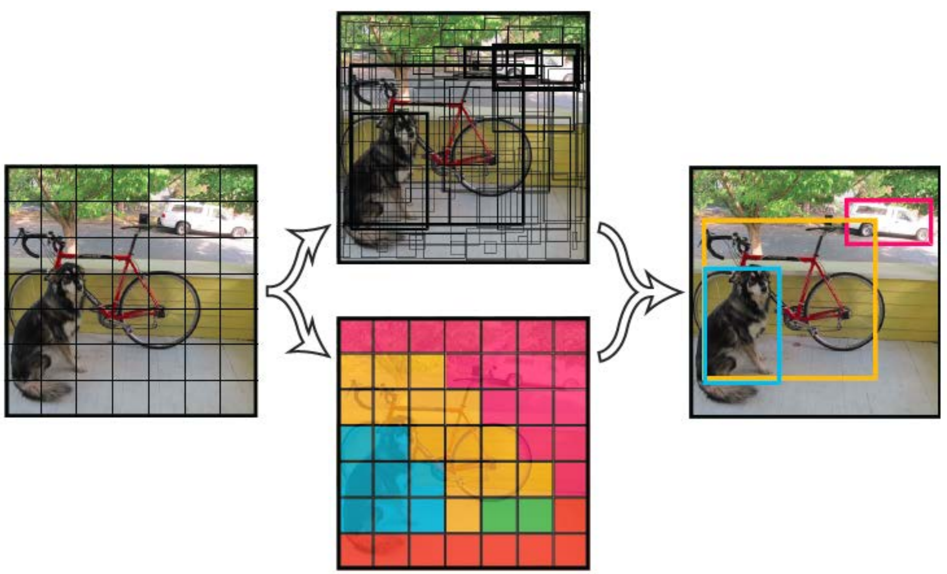
+ keypoint localization (pose estimation)

[video1] [video2]

3. Detection CNN without region proposals

- YOLO “You Only Look Once”
  - Redmond et al. *You Only Look Once: Unified, Real-Time Object Detection*. CVPR 2016.
  - A single net predicts bounding boxes and class probabilities directly from the entire image in a single execution





Output layer:

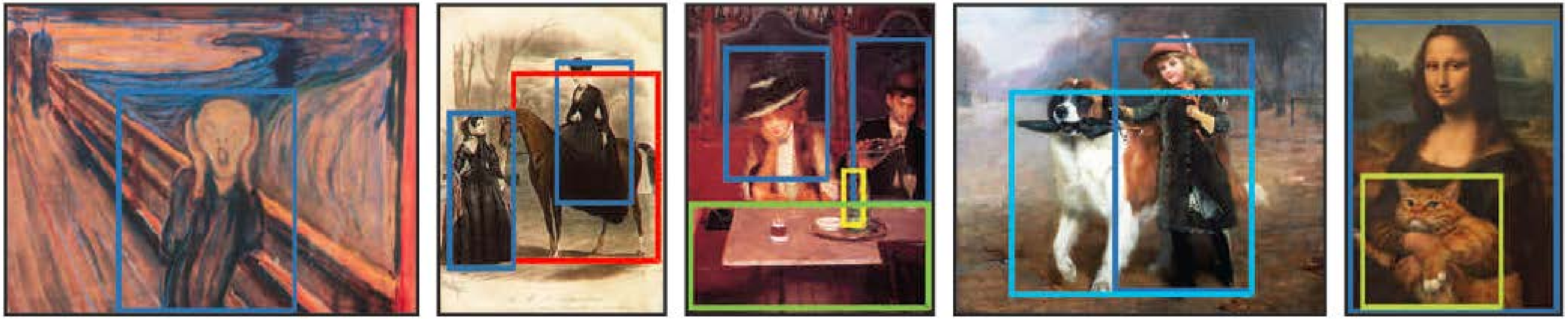
- Tensor 7x7x30

7x7 spatial grid
30=2*5+20

2: number of bboxes per cell
5: (x,y,w,h, overlap score)
20: number of classes

3. Detection CNN without region proposals

- YOLO properties:
  1. Reasons globally
     - Entire image is seen for training and testing, contextual information is preserved (=> less false positives)
  2. Generalization
     - Trained on photos, works on artworks

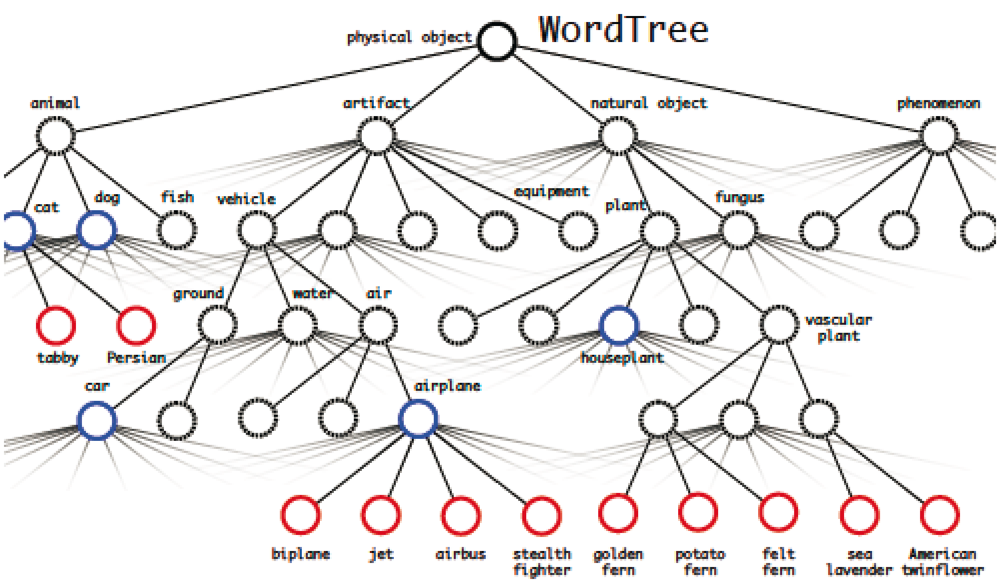


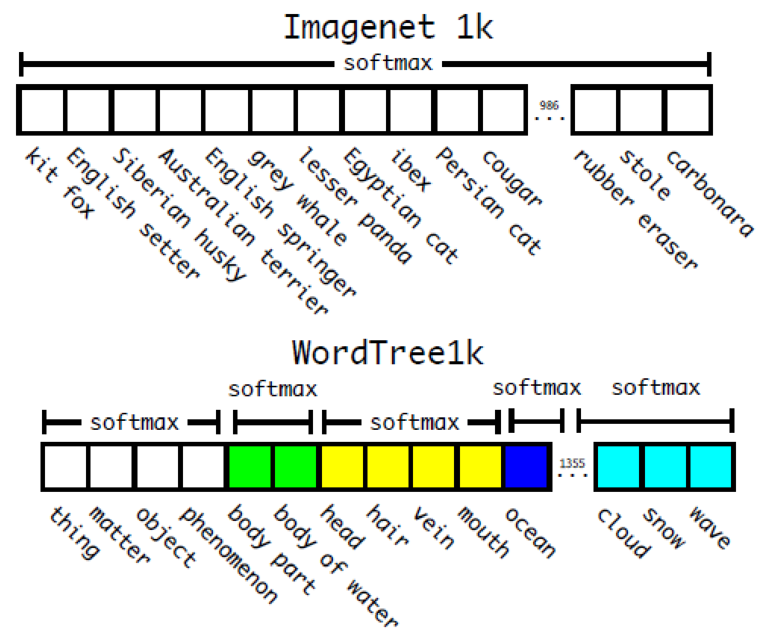
  3. Fast (real-time)

	mAP (VOC 2007)	FPS (GPU Titan X)
YOLO	63.4%	45
fast YOLO	52.7%	150

3. Detection CNN without region proposals

- YOLOv2, YOLO 9000
  - Redmon J., Farhadi A. *YOLO9000: Better, Faster, Stronger*. CVPR 2017
  - Several technical improvements:
    - Batch normalization, Higher resolution input image (448x448), Finer output grid (13x13), Anchor boxes (found by K-means)
  - Hierarchical output labels:

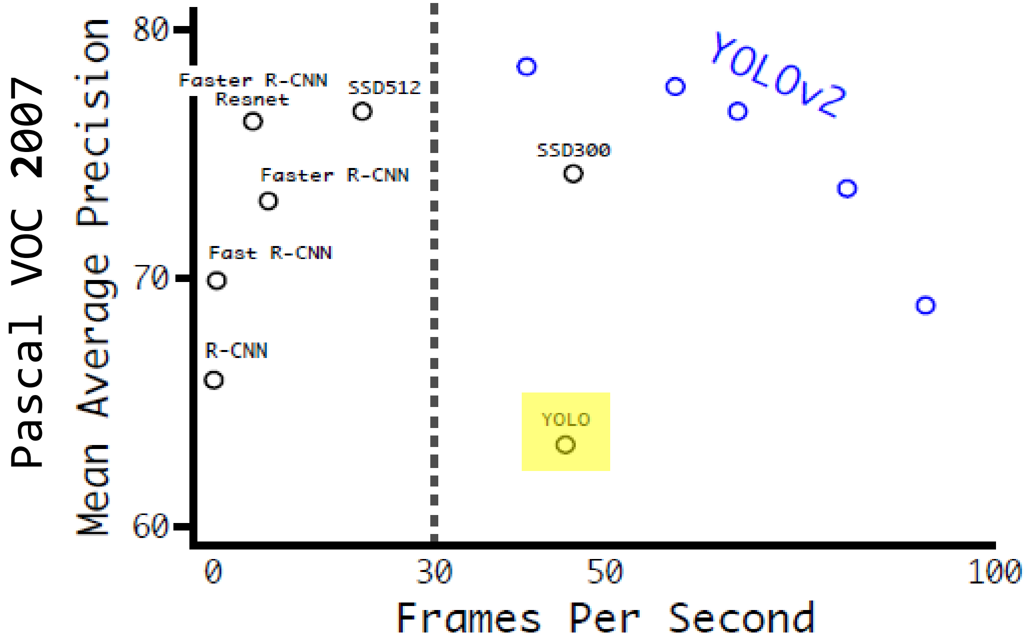




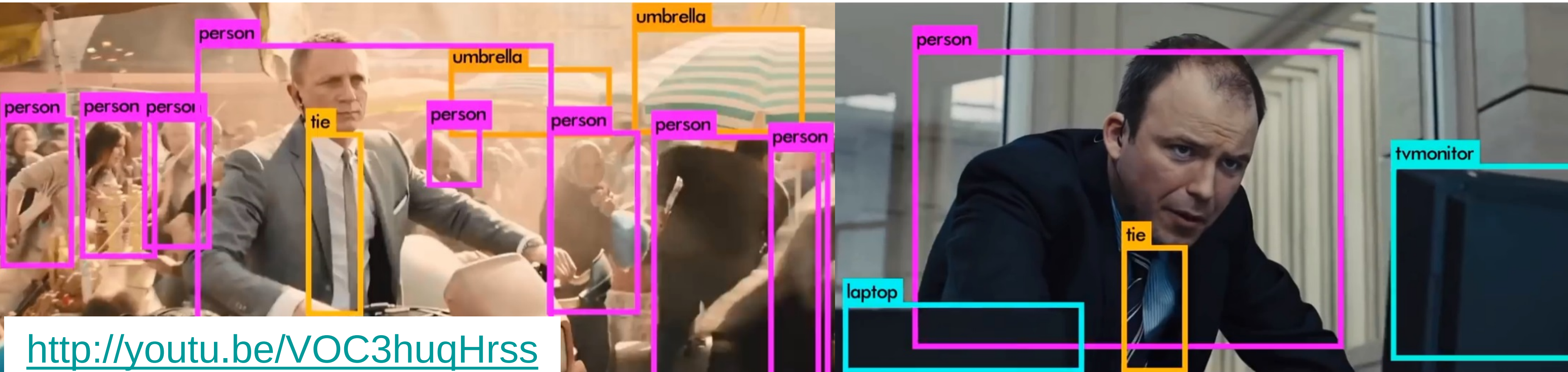
  - Trained on COCO and ImageNET datasets
  - Able to learn from images without bounding box annotation (weak supervision)

3. Detection CNN without region proposals

- YOLOv2, YOLO 9000 summary



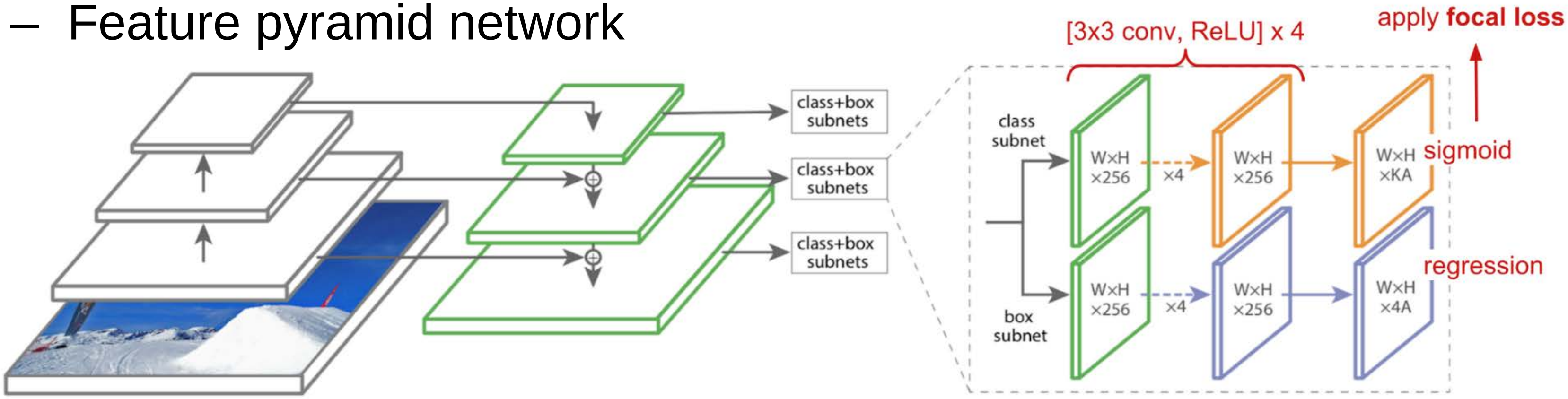
  – The most accurate, the fastest… [video]



http://youtu.be/VOC3huqHrss

3. Detection CNN without region proposals

- RetinaNet (Lin et al., ICCV-2017, IEEE TPAMI 2020)
  - Feature pyramid network



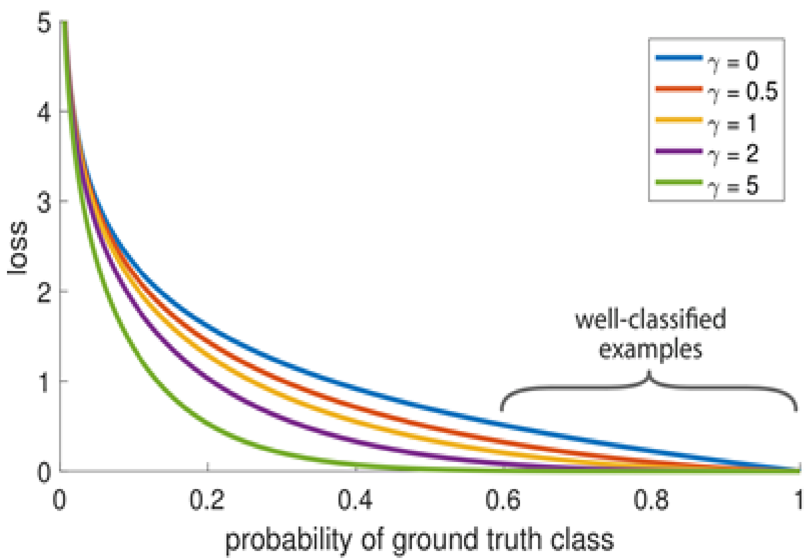
(a) ResNet (b) feature pyramid net (c) class subnet (top) (d) box subnet (bottom)

  - Focal Loss
    - Imbalance between positive and negative (background) classes (1:1000)
    - Assign more weight on hard examples

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise,} \end{cases}$$

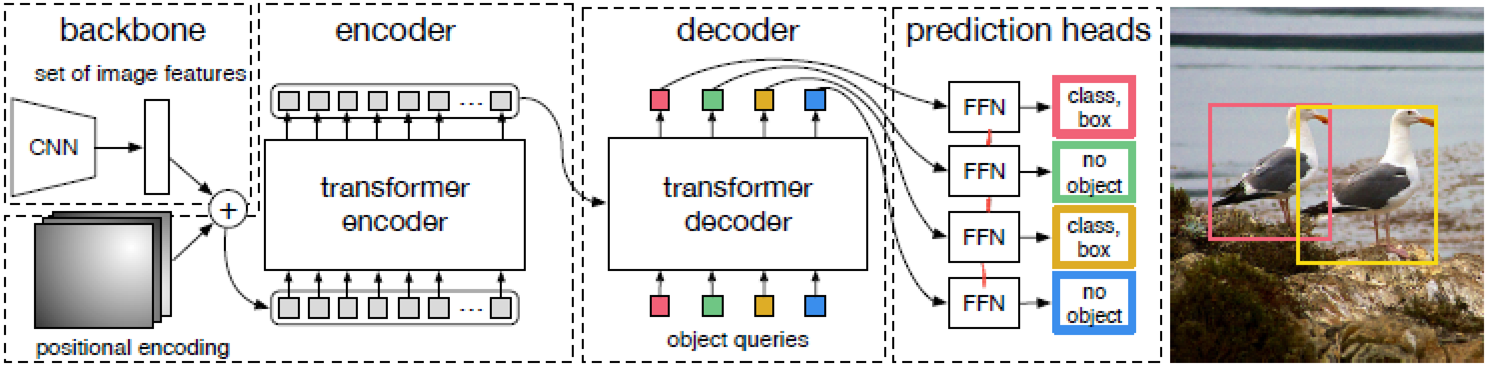
$$\mathrm{CE}(p_t) = -\log(p_t)$$ Cross-entropy loss

$$\mathrm{FL}(p_t) = -(1-p_t)^\gamma \log(p_t)$$ Focal loss



Object Detection with Transformers

- End-to-end Object Detection with Transformers (DETR) [Carion-ECCV-2020]

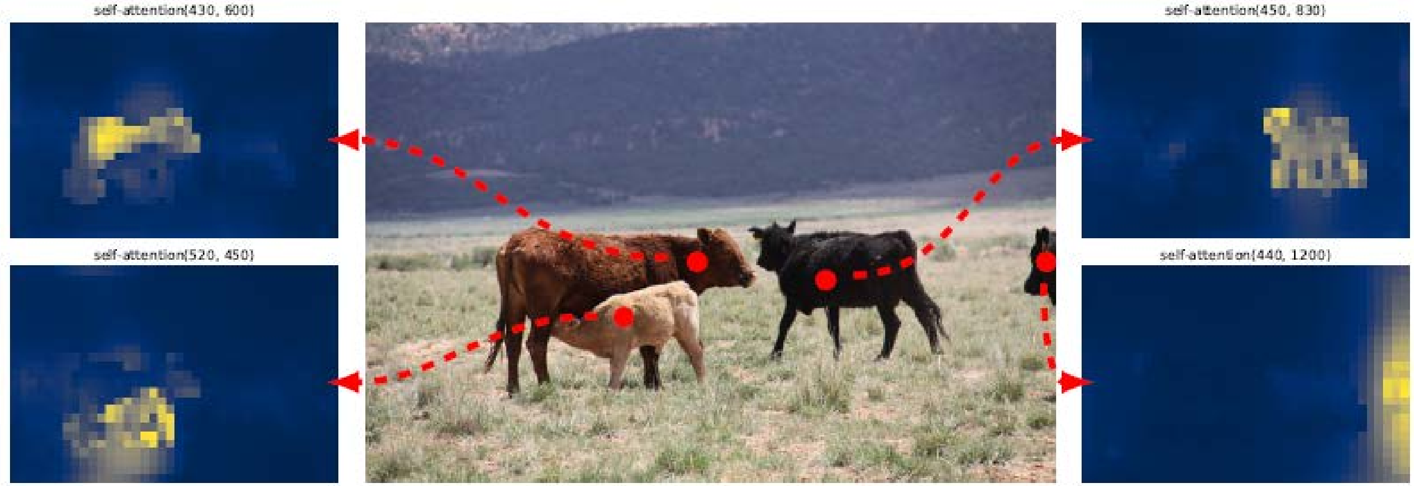


  - CNN as a feature extractor, nowadays image patches instead
  - Transformer encoder – decoder architecture
  - FFN – 3-layer perceptron to predict (bbox + object class/no-object)
  - Bipartite matching between prediction and ground-truth bboxes for training
    - Hungarian algorithm to maximize the matching score
    - Invariant to permutation of predicted objects

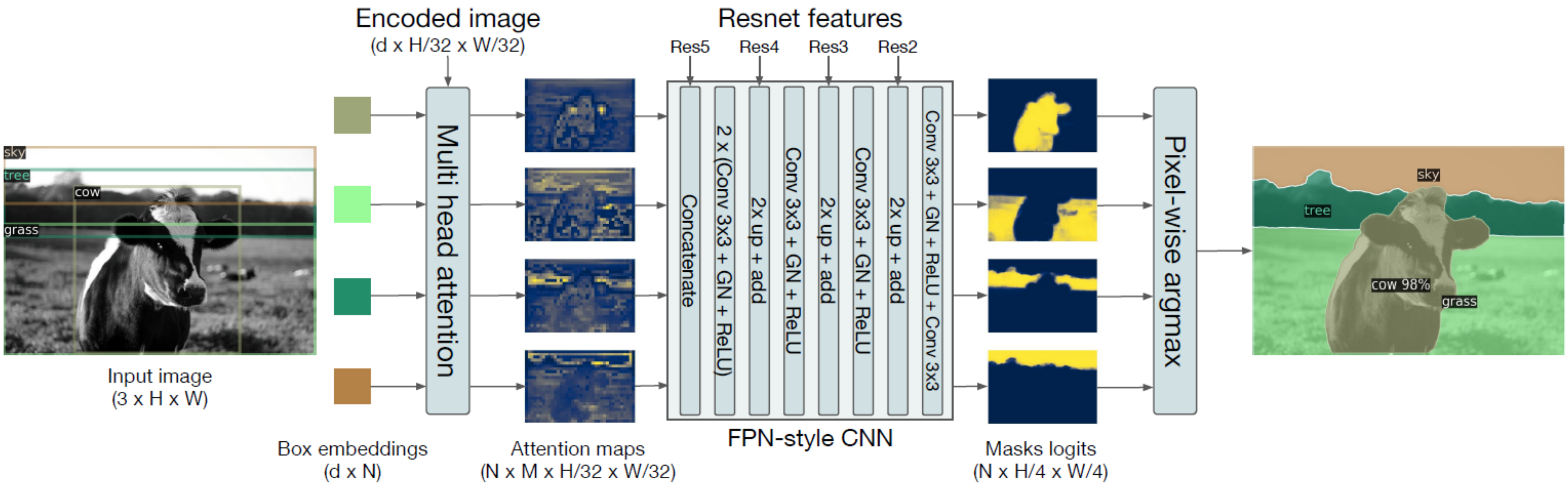
$$\mathcal{L}_{\text{Hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}}(i)) \right]$$

DETR – for segmentation

- Observation: encoder self-attention shows individual instances



- Segmentation head on the attention maps



Detection DNN - summary

1. Exhaustive scanning windows + CNN

2. Region proposals + CNN
    1. R-CNN
    2. Fast R-CNN
    3. Faster R-CNN
    4. Mask R-CNN

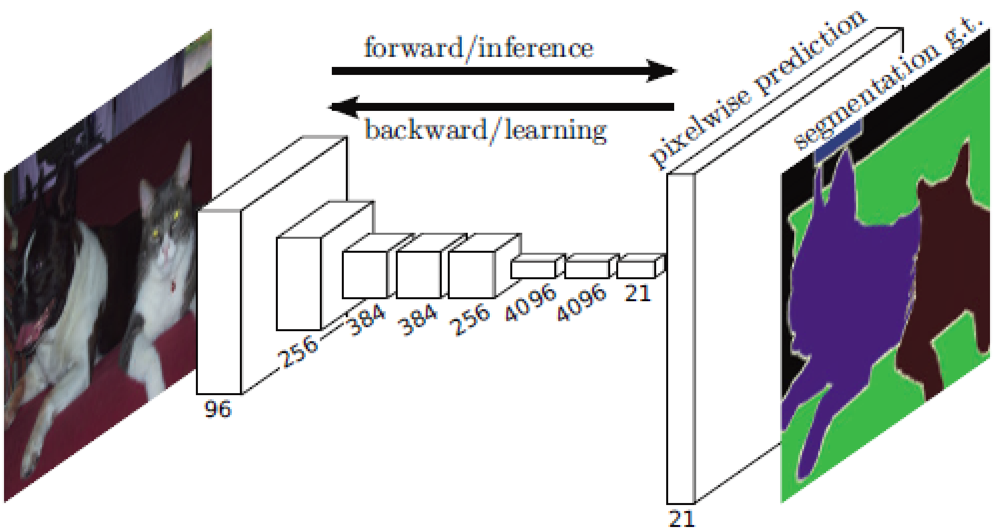
3. CNN/DNN without region proposals
    1. YOLO
    2. YOLO v2, YOLO 9000
    3. RetinaNet
    4. DETR

More recently – (SWIN) transformer backbone + detection/segmentation head

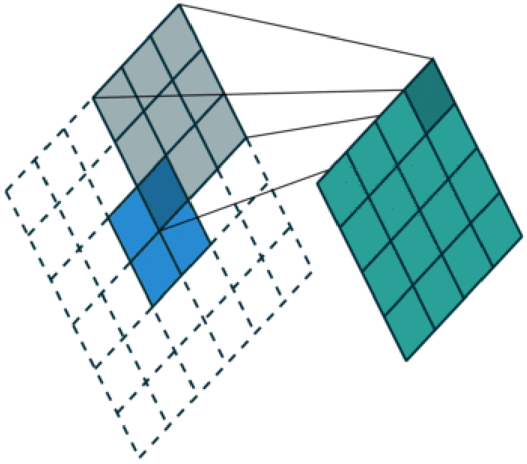
Deep Neural Networks for Semantic Segmentation

Fully Convolutional Net (FCN)

- Shelhammer et al. Fully Convolutional Networks for Semantic Segmentation, TPAMI 2017 (originally CVPR, 2015)

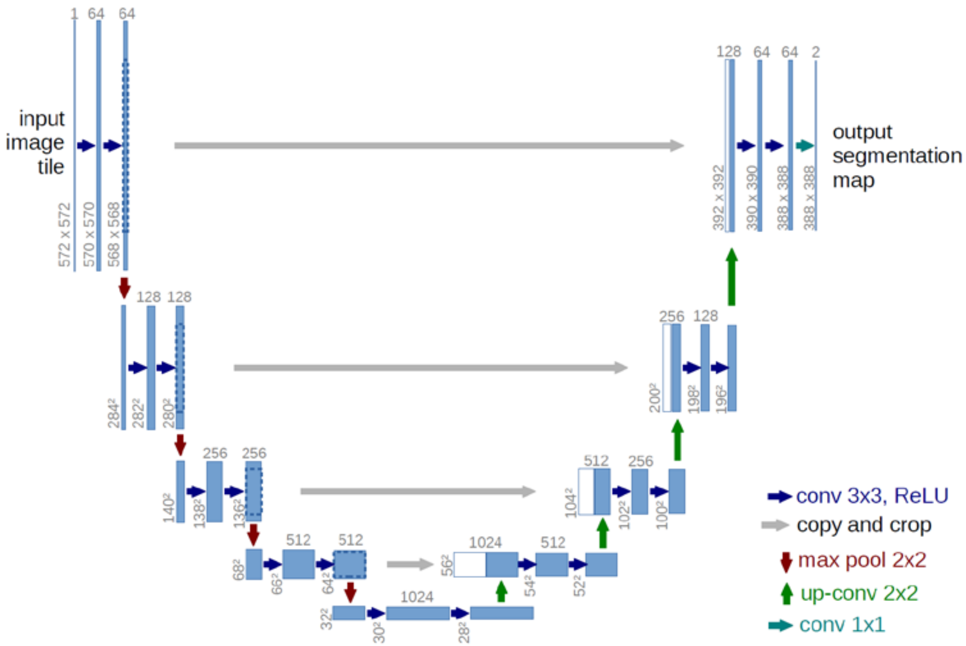


- Fully Convolutional (no fully connected layers)
  - The output size proportional to input size
- Upsampling at the last layer
  - Deconvolution layer (= transposed convolution, fractional-strided convolution)
  - [Dumoulin, Visen, 2018]



U-Net

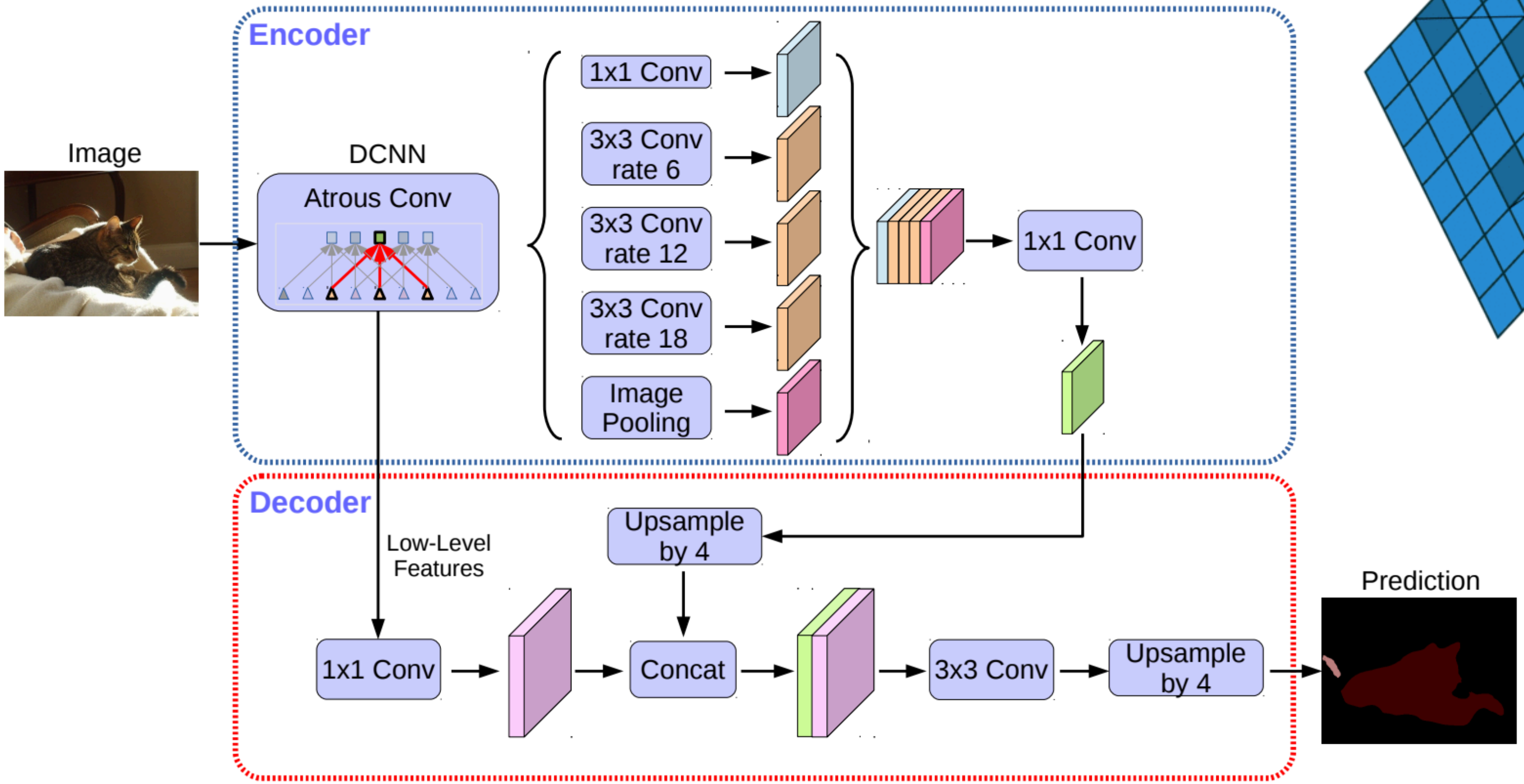
- Ronneberger, et al. [U-Net: Convolutional Networks for Biomedical Image Segmentation, Medical Image Computing and Computer-Assisted Intervention](), 2015

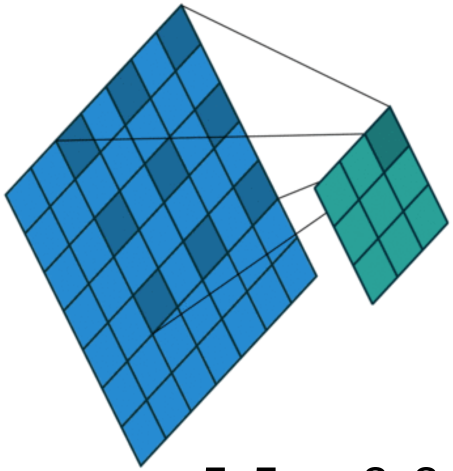


- Bahnik et al., [Visually Assisted Anti-Lock Braking System](). IEEE IV, 2020
  – Surface segmentation

DeepLab v3+

- Chen et al., Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, ECCV 2018.
- Atrous Convolutions (= with "holes", dilated convolutions)
  - Same number of parameters with larger receptive field

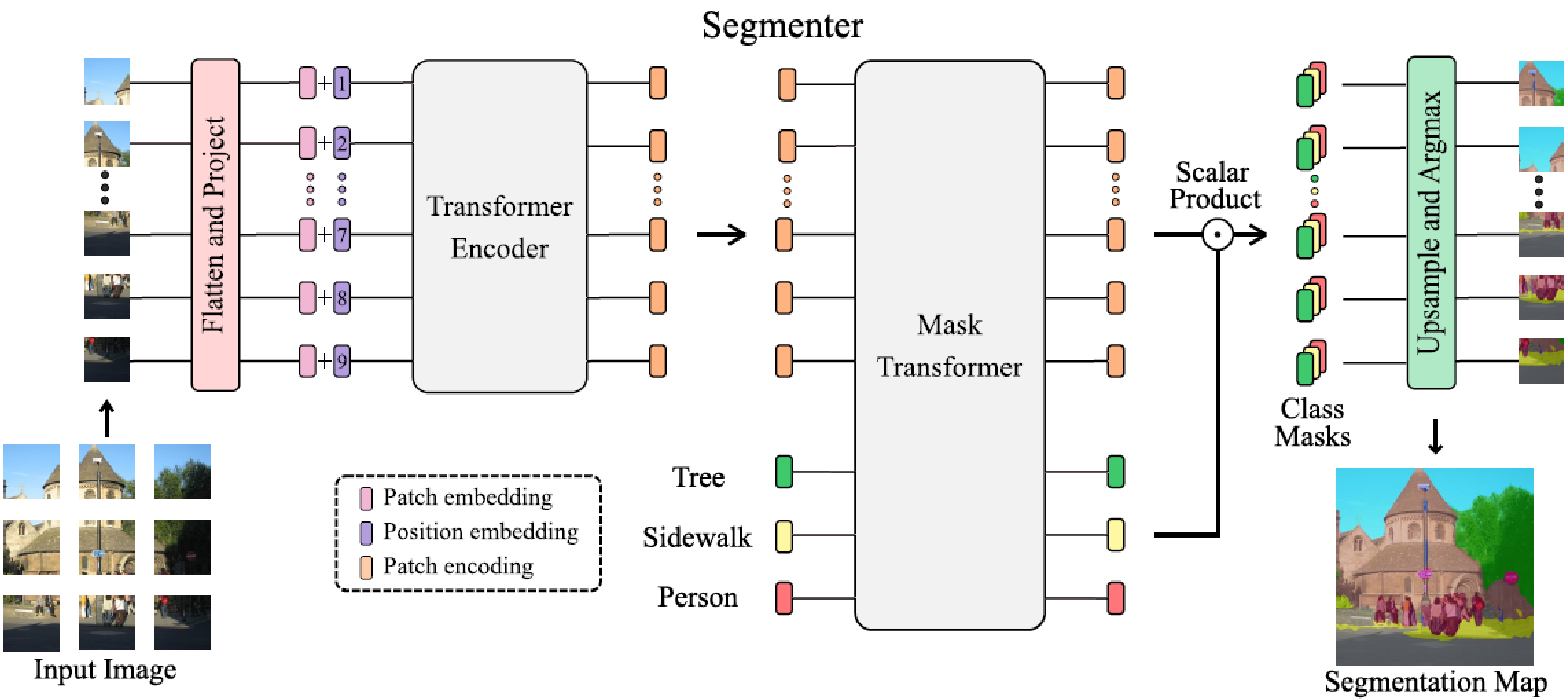




5x5 => 3x3 parameters

Segmentation with Transformers

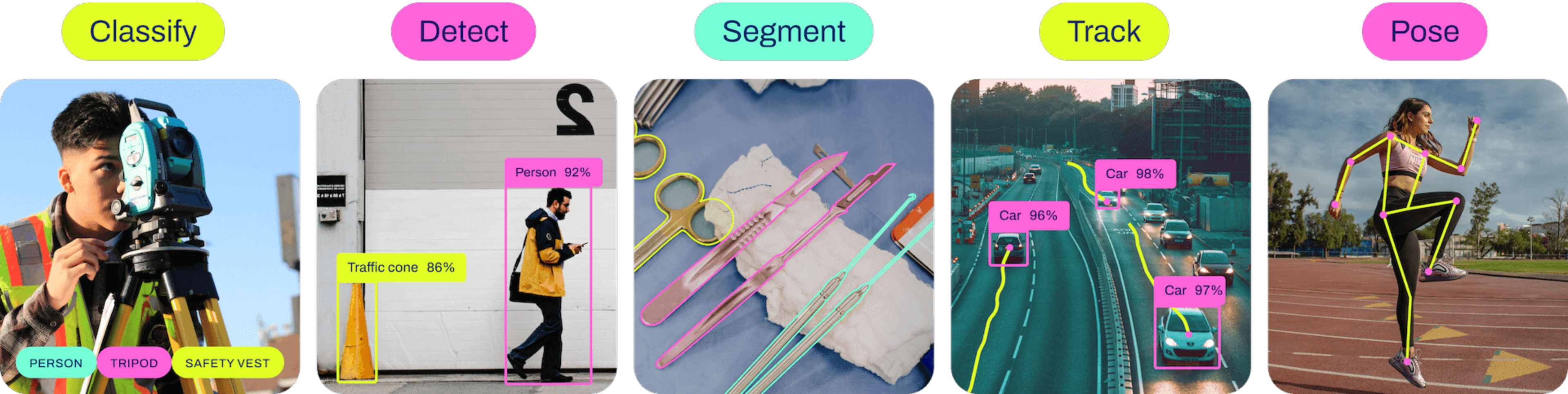
- Segmentation head on top of the transformer features or attention maps
- SEGMENTER [Strudel-ICCV-2021]
  - No convolutions at all



Detection/Segmentation frameworks

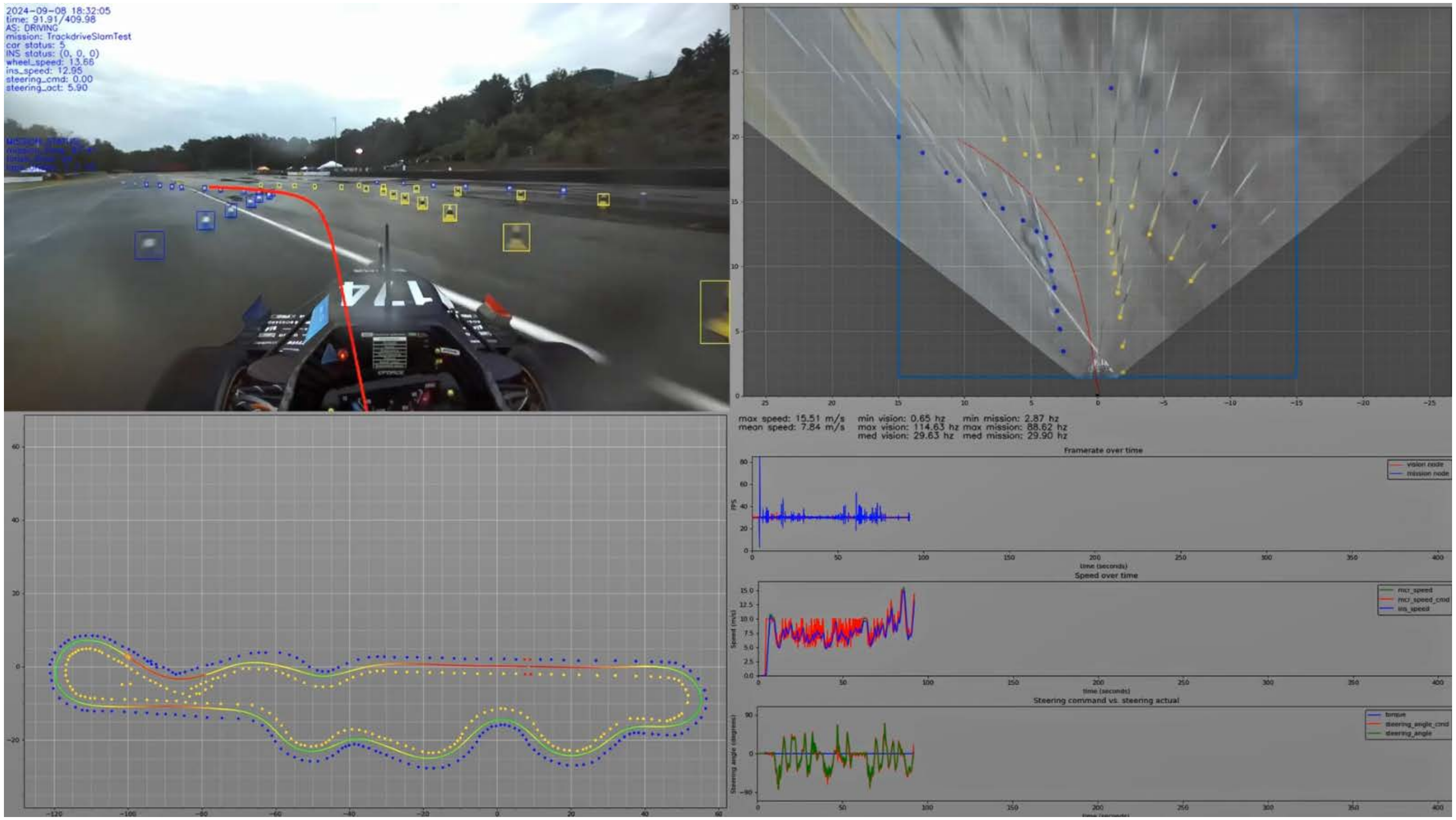
- Detectron2 (Meta, FAIR)
  - Detection, segmentation, keypoints
  - Large model zoo (Faster RCNN, RetinaNet, Mask RCNN, …)

- YOLOv8 (Ultralytics)
  - User-friendly, accurate and fast…



Autonomous student formula (eForce)

- eForce (CTU formula student team)
  - Electric vehicle
  - Acceleration ~ 2.5sec 0-100 km/h
- Driverless disciplines
  - YOLO-type detection of traffic cones



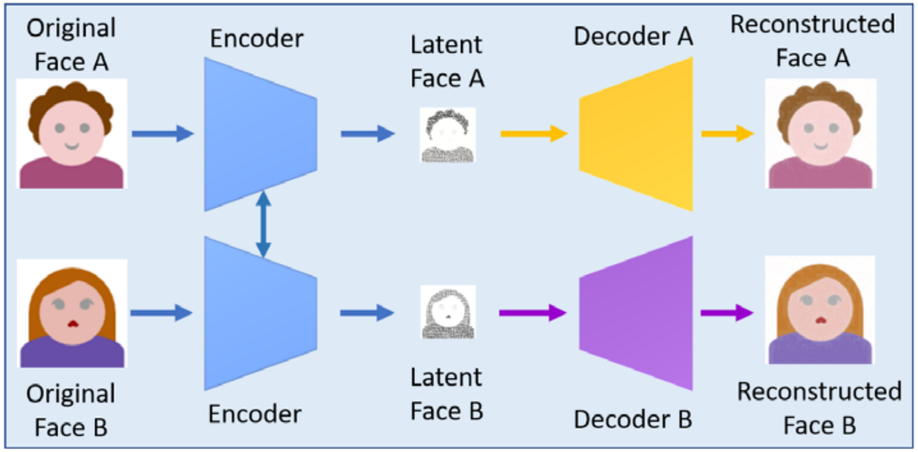
[video]

“Deeper” Insight into the Deep Nets

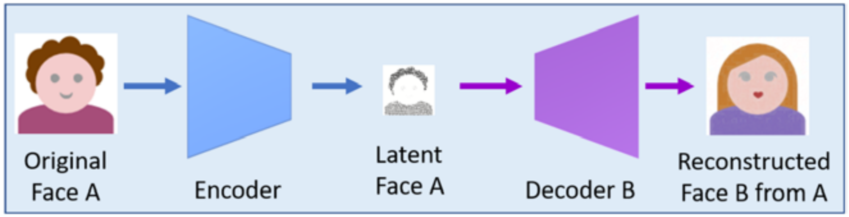
Deep Fake

- Seamless swapping a face in an image/video, e.g. [Nguyen et al., 2020]
- Auto-encoder architecture
  - Single shared encoder (to capture pose / expressions)
  - Two decoders (Source and Target to capture person’s identity)

Training



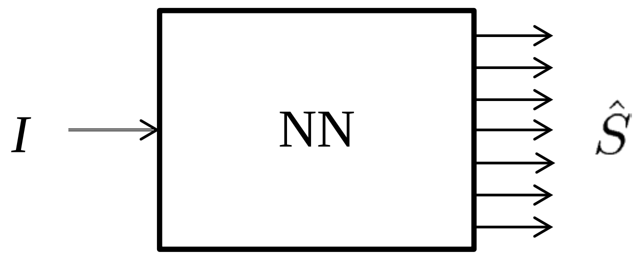
Deployment



- Controversy:
  - fake news, fake porn, …
- Deep fake detection

Deep Network Can Easily Be Fooled

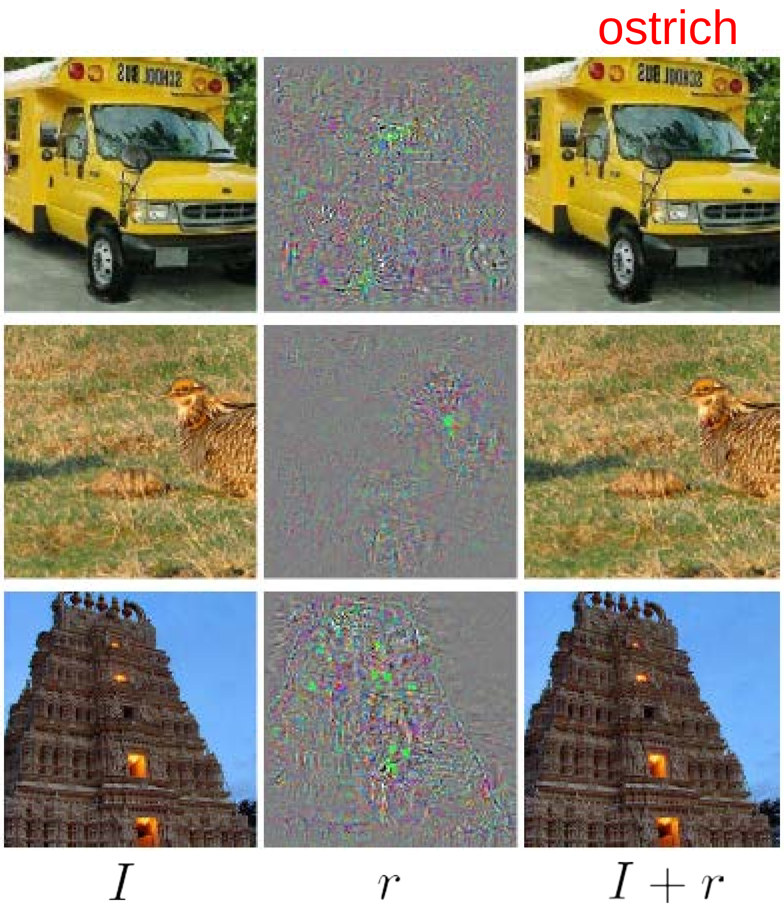
- Szegedy et al. [Intriguing properties of neural networks](). ICLR 2014
  - Small perturbation of the input image changes the output of the trained “well-performing” neural network
  - The perturbation is a non-random image, imperceptible for human



$$\min_r \{ ||\mathrm{NN}(I + r) - S||^2 + \lambda ||r||^2 \}$$

  - Optimum found by gradient descent

$$r^{t+1} = r^t - 2\gamma \Big((\mathrm{NN}(I + r^t) - S) \frac{\partial \mathrm{NN}(I)}{\partial I} + \lambda r^t \Big)$$

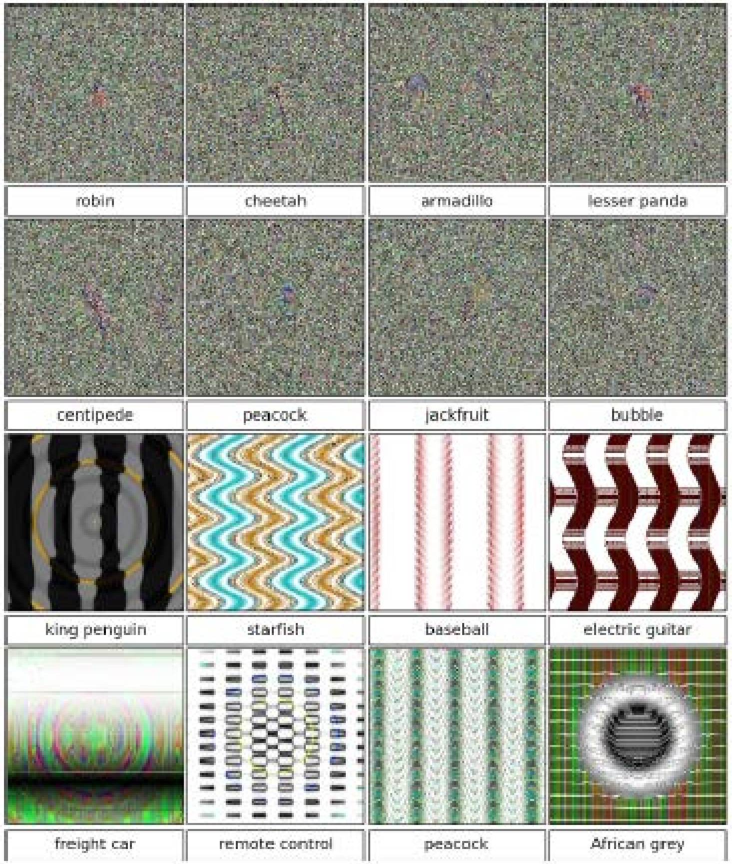


Deep Network Can Easily Be Fooled

- Nguyen et al. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. CVPR 2015.
  - Artificial images that are unrecognizable to humans, producing high output score can be found
  - The optimum images found by evolutionary algorithm
    - Starting from random noise
    - Direct/Indirect encoding

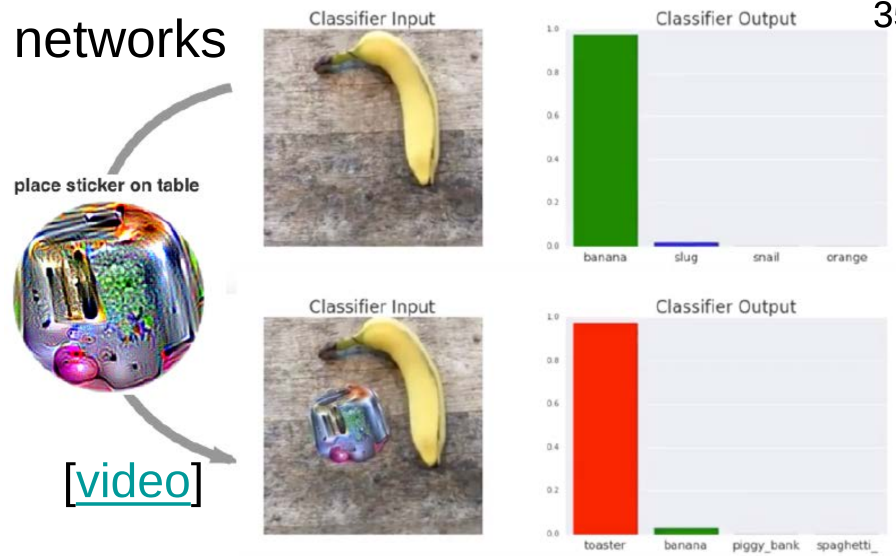
$$\min_I ||\mathrm{NN}(I) - S||^2$$

$\Rightarrow$ The images found do not have the natural image statistics



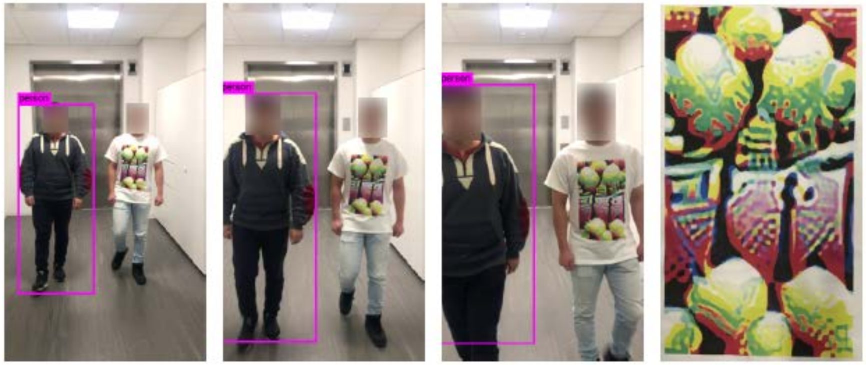
Deep Network Can Easily Be Fooled

- Adversarial physical attacks on neural networks
  - Adversarial sticker [Brown-2018]

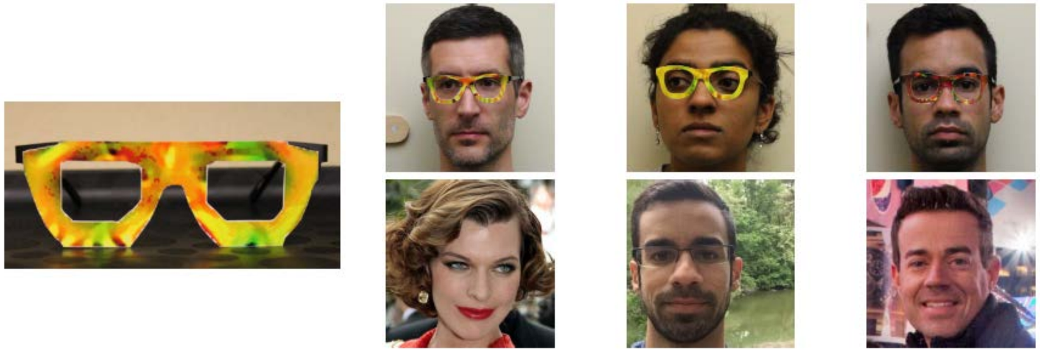


[video]

  - Adversarial T-shirt [Xu-2019]

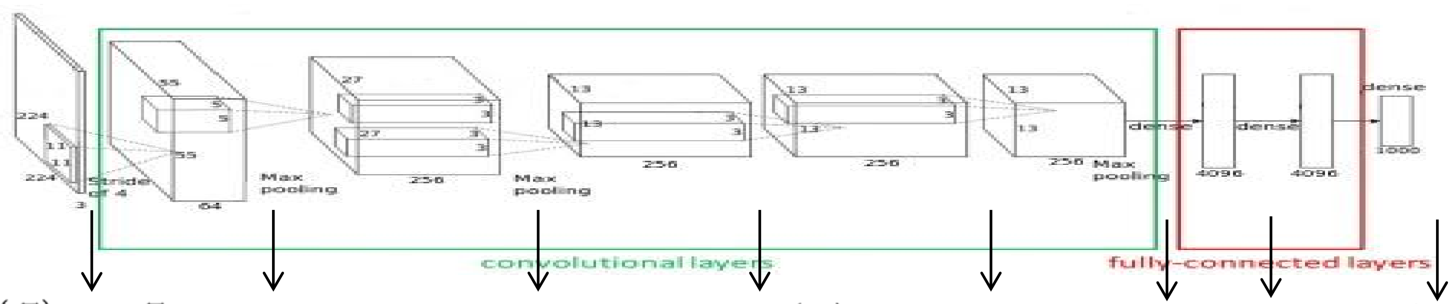


  - Adversarial glasses [Sharif-2016]



Visualization the Deep Nets

- Mahendran A., Vedaldi A. [Understanding Deep Image Representations by Inverting Them](). CVPR 2015.



$\Phi_1(I) = I$ $\Phi_k(I)$ $\Phi_K(I) = S$

- Start from a random Image I
- Best match between features + image regularization (natural image prior)

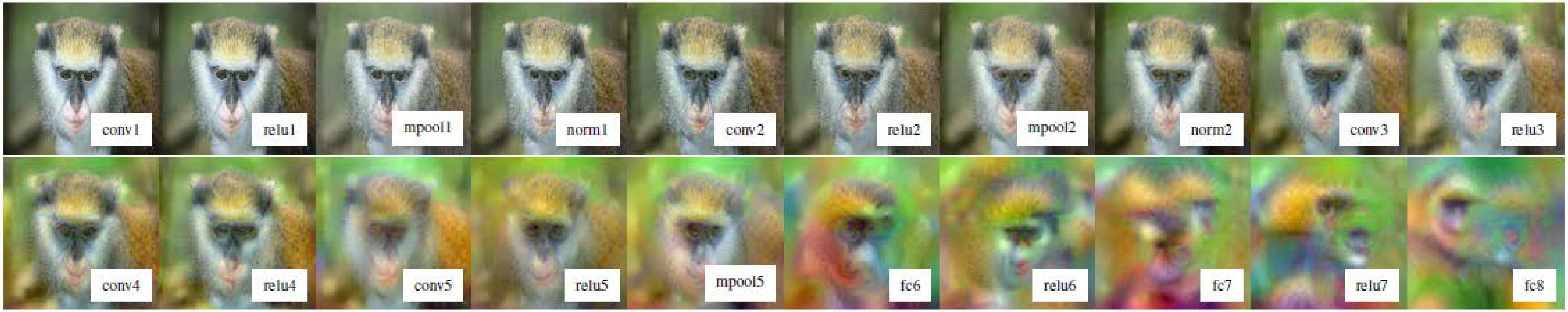
$$\min_I \{ ||\Phi_k(I) - \Phi_k^0||^2 + \lambda R(I) \}$$

- Total Variation regularizer (TV)

$$R(I) = \sum_{x,y} \left(\left(\frac{\partial I(x,y)}{\partial x} \right)^2 + \left(\frac{\partial I(x,y)}{\partial y} \right)^2 \right)^{\frac{\beta}{2}}$$

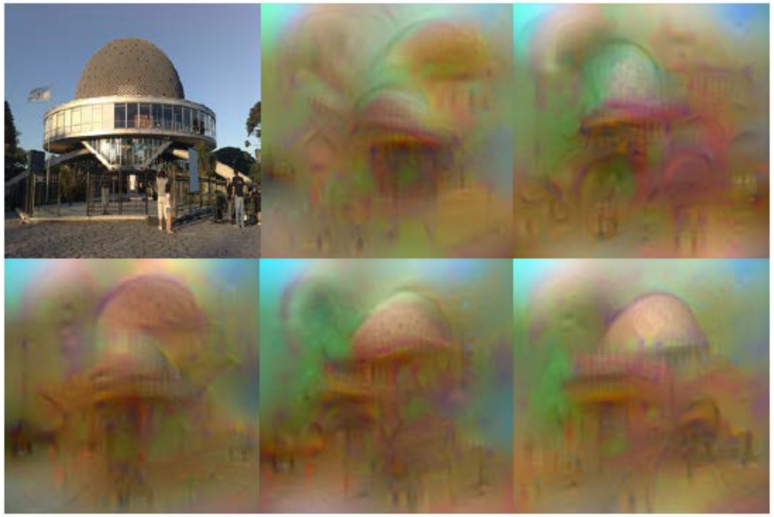
Visualizing the Deep Nets

- CNN reconstruction



  – Gradient descent from random initialization
  – Reconstruction is not unique

  ⇒ All these images are identical for the CNN

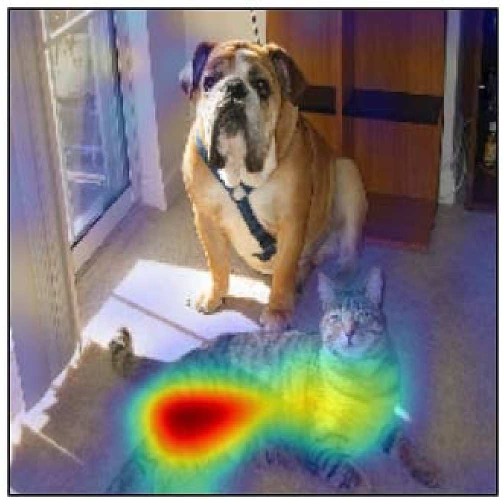


- Similarly, find an image that causes a particular neuron fires (maximally activate)

Verification what the deep net learned

- Deep nets often criticized for a lack of interpretability
- Grad-CAM: Visual Explanations from Deep Networks [Selvaraju-ICCV-2017]
  – GRADient weight Class Activation Mapping
  – Trianed model => Coarse localization map highlighting important regions for a class c

VGG “c=cat”



VGG “c=dog”

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial NN(I)^c}{\partial \Phi_{ij}^k}$$

$$L^c_{\text{Grad-CAM}} = ReLU\left(\sum_k \alpha_k^c \Phi^k\right)$$

$\Phi_{i,j}^k$ …Feature tensor (last convolution layer)

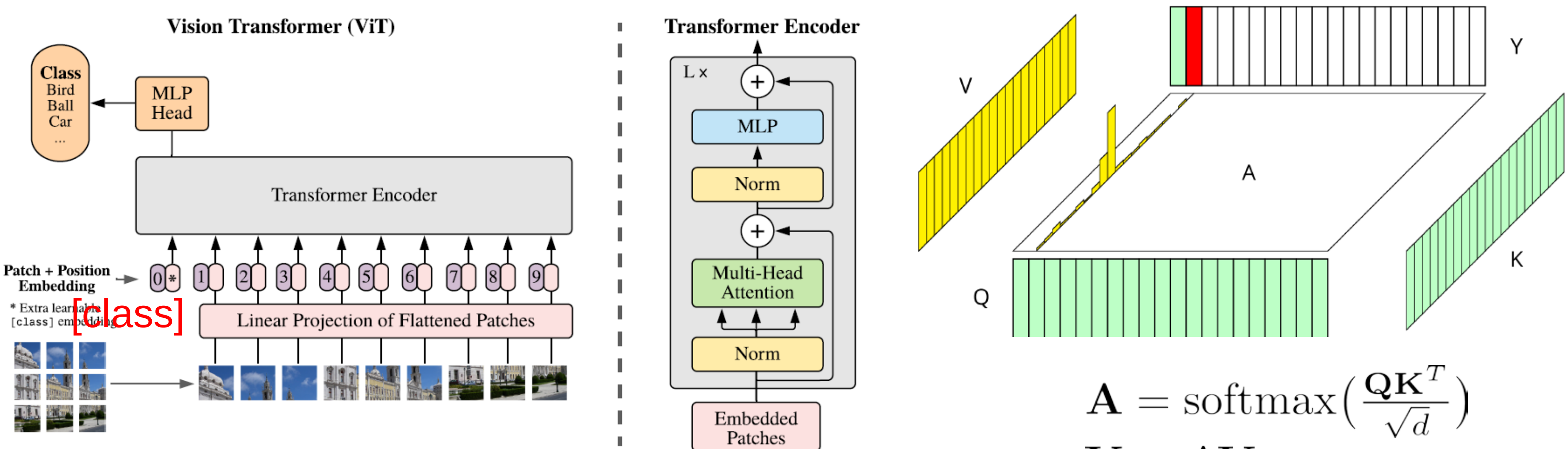
i, j - spans spatial dimensions

k - spans channels

Verification what the deep net learned (2)

- For transformers: Self-Attention exploited

Figure credit: Francois Fleuret



$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d}}\right)$$

$$\mathbf{Y} = \mathbf{AV}$$

- Self-Attention: Query, Key, Value
  - Models long-distance relationships between tokens
  - $\mathbf{A}$ matrix of size *N*x*N*, where *N* is the number of tokens
  - Self-attention map of the [class] token is used (reshaped to image size)
- Multiple heads, multiple layers

(recap)

Verification what the deep net learned (3)

- Attention Roll-out [Abnar-2020]

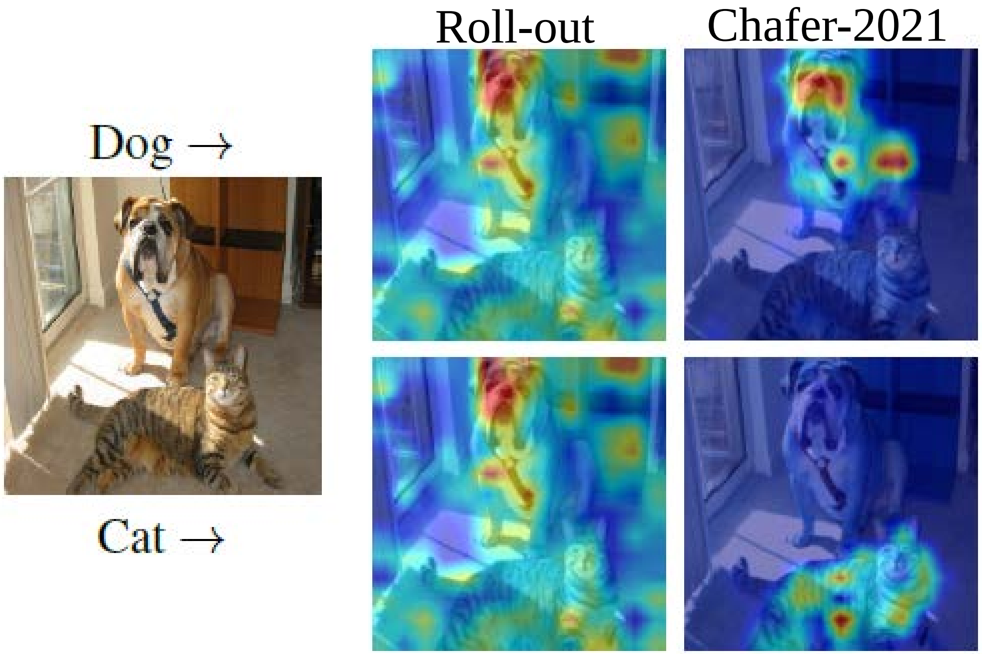
$$\hat{\mathbf{A}}^{(b)} = I + \mathbb{E}_h \mathbf{A}^{(b)}$$

$$\text{rollout} = \hat{\mathbf{A}}^{(1)} \cdot \hat{\mathbf{A}}^{(2)} \cdot \ldots \cdot \hat{\mathbf{A}}^{(B)}$$

- Combination of gradient + attention [Chafer-ECCV-2021]

$$\bar{\mathbf{A}}^{(b)} = I + \mathbb{E}_h (\nabla \mathbf{A}^{(b)} \odot R^{(n_b)})^+$$

$$\mathbf{C} = \bar{\mathbf{A}}^{(1)} \cdot \bar{\mathbf{A}}^{(2)} \cdot \ldots \cdot \bar{\mathbf{A}}^{(B)}$$



Deep Dream

- Manipulate the input image so that response scores are higher for all classes
- Start from an original image
- Regularization with TV prior

$$\max_I \left(||\mathrm{NN}(I)||^2 - R(I) \right)$$

Credit: Eric Wayne

[video]

http://youtu.be/EjiyYtQIEpA

Deep Dream

- Maybe…

Salvador Dalí

Soft Construction with Boiled Beans (1936)

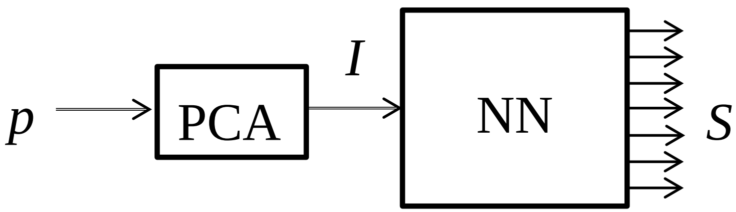
Swans Reflecting Elephants (1937)

Apparition of a Face and Fruit Dish on a Beach (1937)

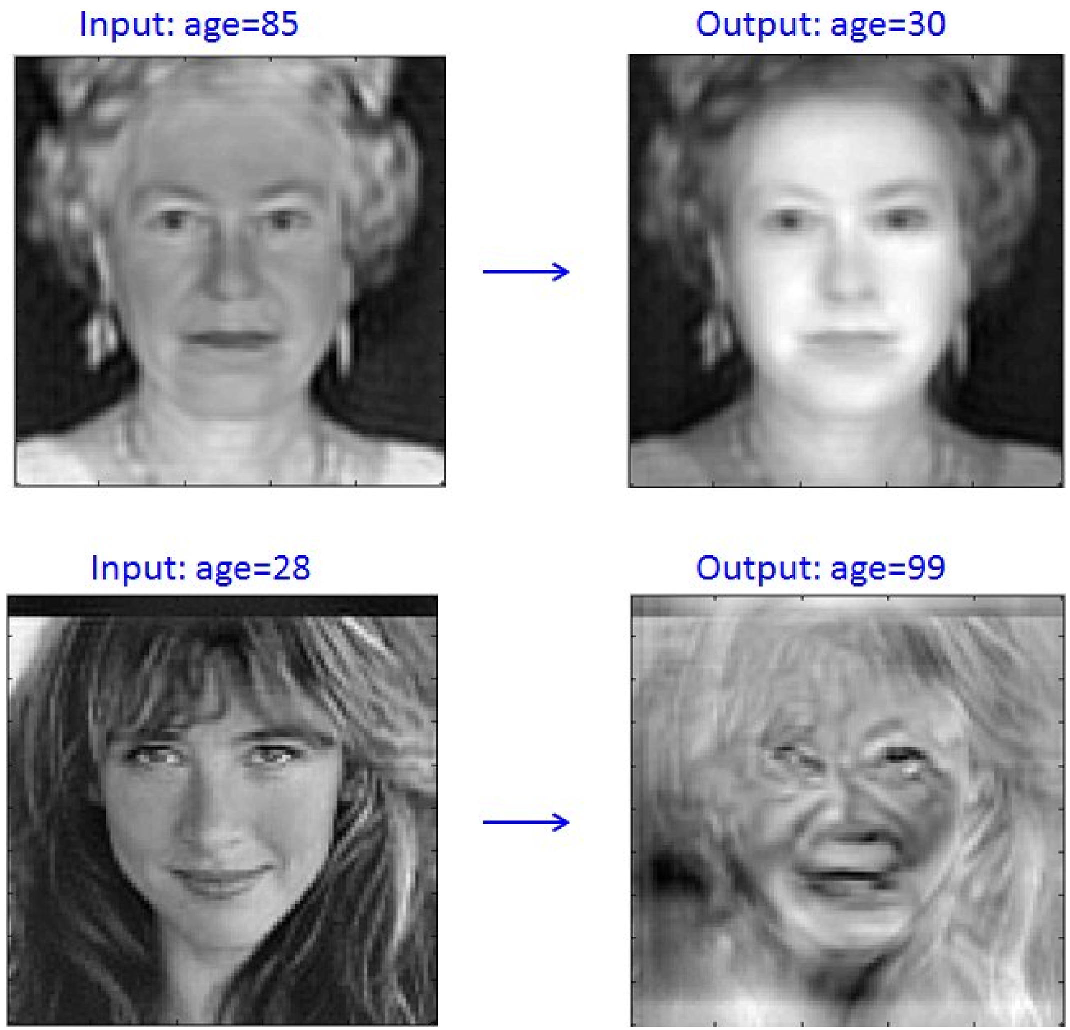
Hieronymus Bosch, Garden of Earthly Delights (~1510), [part]

Deep Aging

- Our network trained for predicting age (gender and landmarks) was used



$$\min_{p} ||\mathrm{NN}(\mathrm{PCA}(p)) - S^t||^2$$



[Čech, J. Unpublished experiment, 2015]

Deep Art – Neural Style

- Gatys et al. *A Neural Algorithm of Artistic Style.* Journal of Vision, 2015.
  - Generate high-quality artistic rendering images from photographs
  - Combines content of the input image with a style of another image

Content image

Style images

Result images

  - More examples at Deepart.io

Deep Art – Neural Style

- Main idea:
  - the style is captured by correlation of lower network layer responses
  - the content is captured by higher level responses

- The optimization problem:

$$\min_I \{\alpha L_{\text{content}}(I_1, I) + \beta L_{\text{style}}(I_2, I)\}$$

$$L_{\text{content}} = \sum_k ||\Phi_k(I) - \Phi_k(I_1)||^2$$

$$L_{\text{style}} = \sum_k w_k ||G(\Phi_k(I)) - G(\Phi_k(I_2))||^2$$

G is a Gram matrix (dot product matrix of vectorized filter responses)

Summary

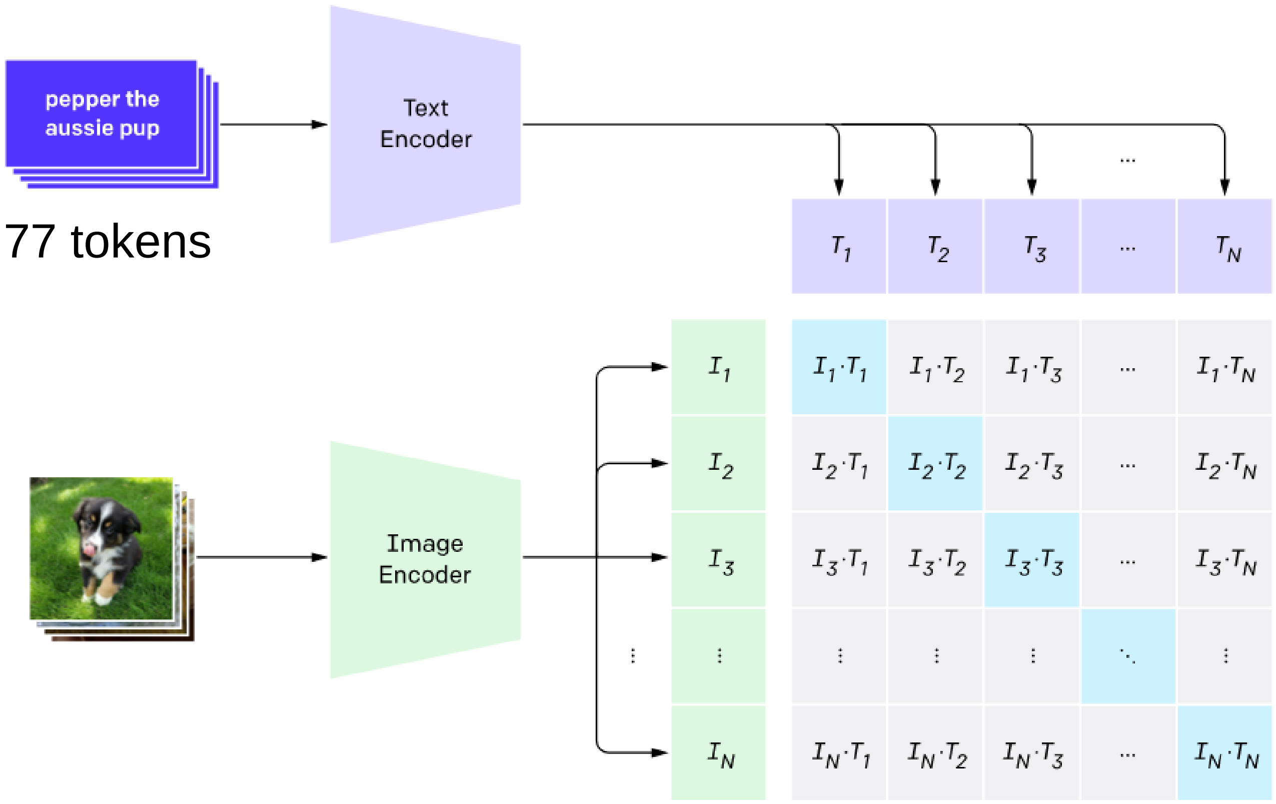
- Deep fake
- Using Network gradient according to the image for various optimization
  - Fooling the net
  - Visualization + Interpretation
  - Dreaming, Hallucination
  - Aging
  - Artistic rendering of photographs

=> Understanding of the trained model

Foundation models

CLIP – Connecting Text and Images

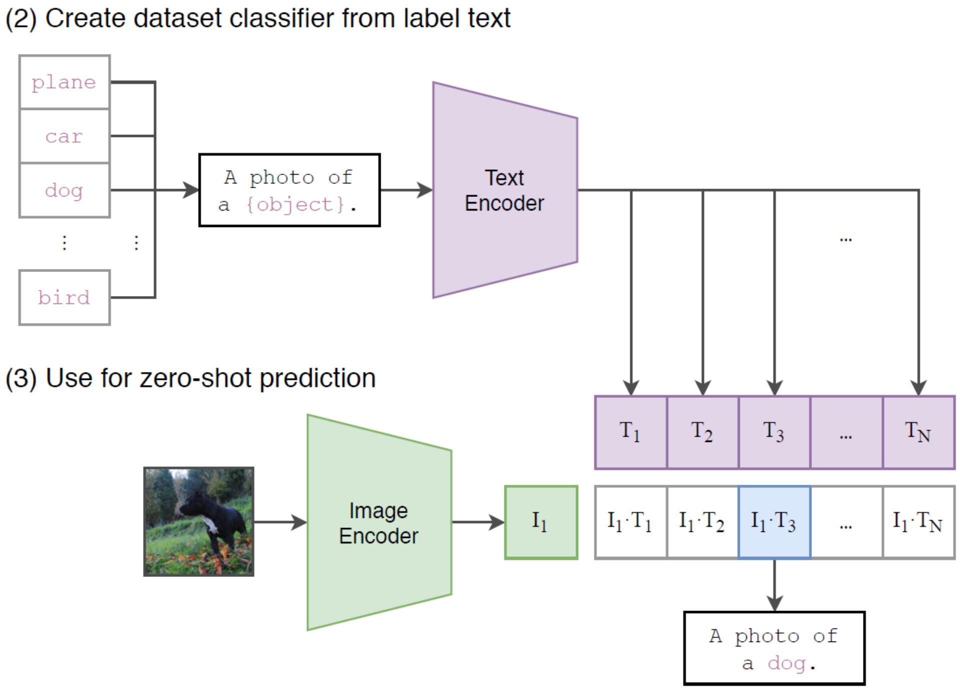
- CLIP [Radford-2021] by OpenAI
  - “*Contrastive Language–Image Pre-training*”
  - Learn joint text-image embedding => Text-image (cosine) similarity
  - Learned from 400M WebImageText (WIT) dataset



CLIP – Connecting Text and Images

- Zero-shot prediction (on par with Resnet on ImageNET benchmark)
  - Loop over ImageNET-classes:

$$\mathit{max}\ \mathrm{CLIP}(\ \mathrm{E_T}(\text{“}\mathit{A\ photo\ of\ a}\ \text{<}\mathit{class}\text{>”}),\ \mathrm{E_I}(I)\ )$$



⇒ 76.2% top-1 accuracy on ImageNET

- Trained model publicly available
- Alternative model: ALIGN [Jia-ICML-2021] (by Google), but not public
  - A Large scale ImaGe and Noisy-text embedding

DINO – self-supervised vision transformer

- DINO (self-Distillation with NO labels) [Caron-ICCV-2021] by Meta

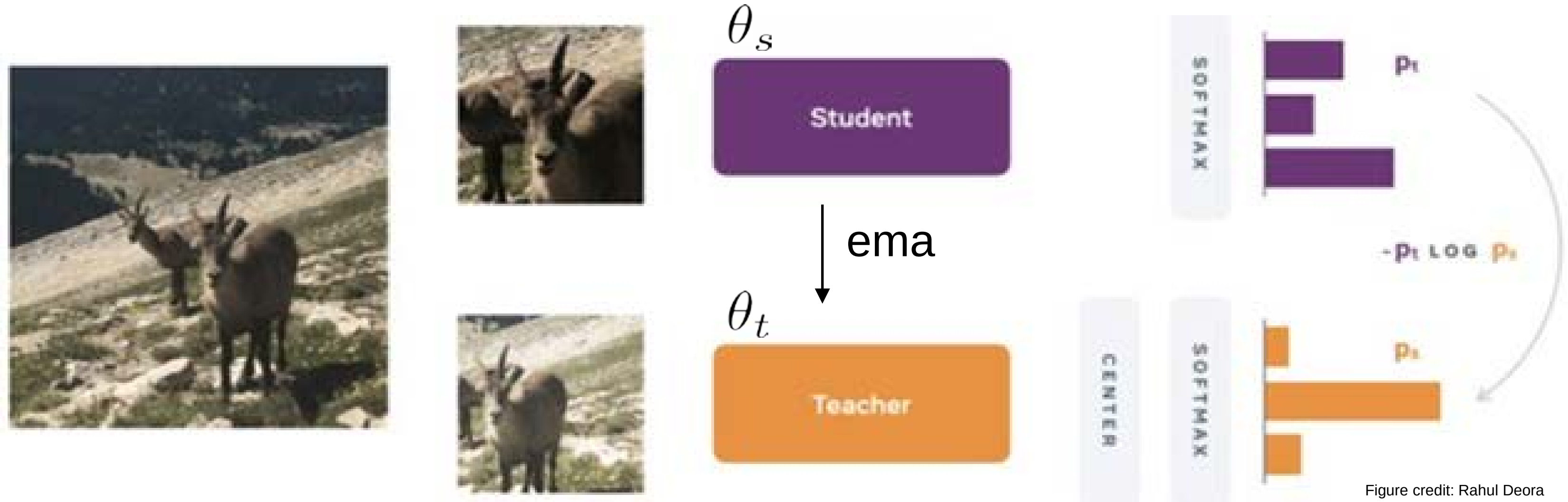


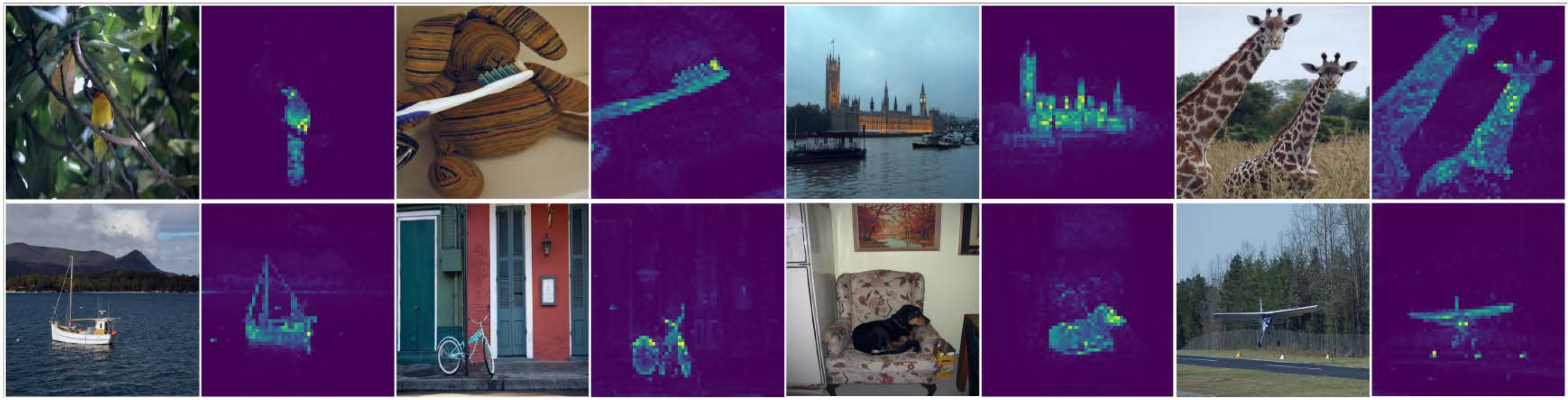
Figure credit: Rahul Deora

- No labels, random crops of the same image
- Student – Teacher training
  - Student and teacher nets of the same architecture
  - Student updated by Cross-entropy loss $\min_{\theta_s} H(P_t(x), P_s(x))$
  - Teacher's weights are exponentially moving average of the student

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$$

DINO – self-supervised vision transformer

- Model learns class-specific features without label supervision



Self-attention of the [CLASS] token on the heads of last hidden layer [video]

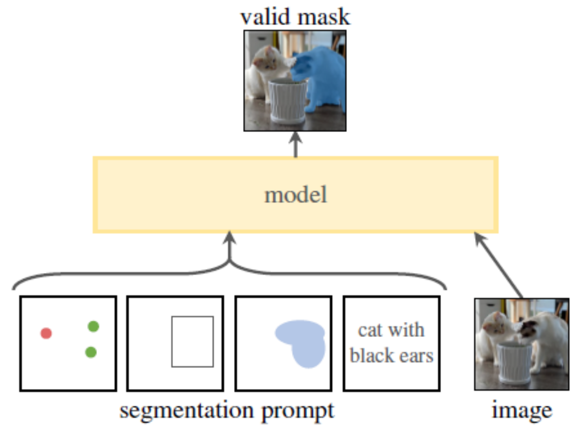
- Universal representation for downstream tasks
  - k-NN/linear classifier on the features 78.3/80.1% top-1 accuracy on ImageNET
  - Transfer learning (fine-tuning on other datasets)
  - Image retrieval
  - Segmentation
  - …

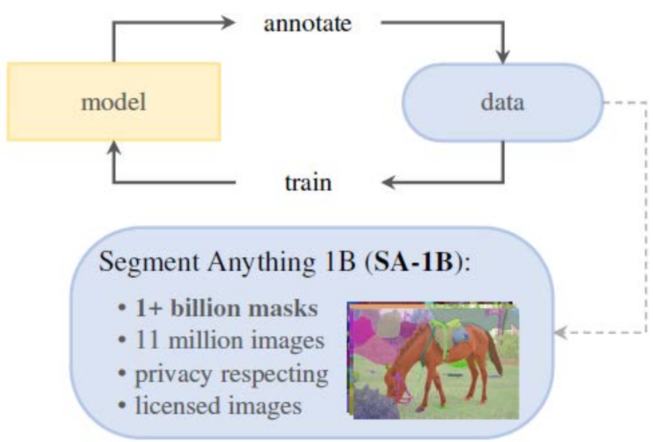
Segment Anything

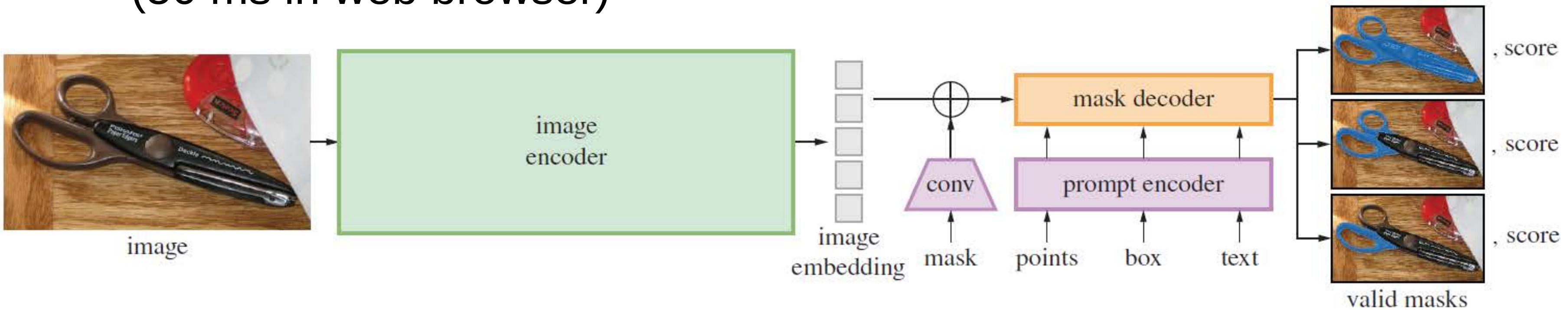
- Segment Anything Model (SAM) [Kirillov-ICCV-2023] by Meta
- Promptable segmentation
- Human in the loop training (11M images, 1B masks)
  - 3 stages (assisted-manual 120k, semi-automatic 180k, fully-automatic 11M)
- Handles natural ambiguity by providing multiple solutions (3)
- Lightweight prompt encoder and mask decoder

  ⇒ Interactive segmentation

  (50 ms in web browser)

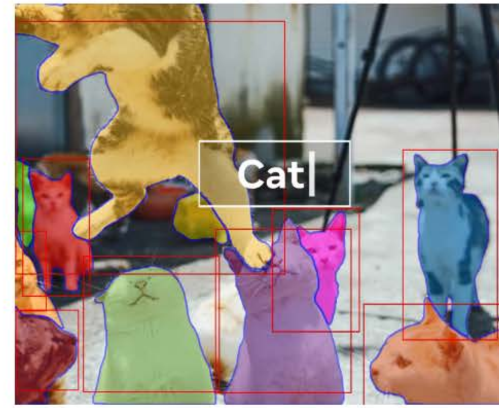




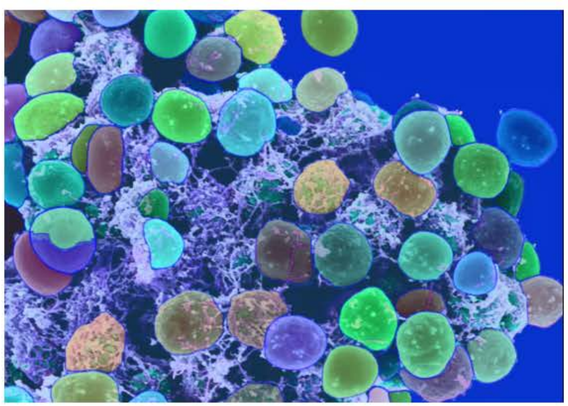


Segment Anything

- Qualitative results – various prompts



- Outstanding zero-shot capabilities



[project-page / demo]

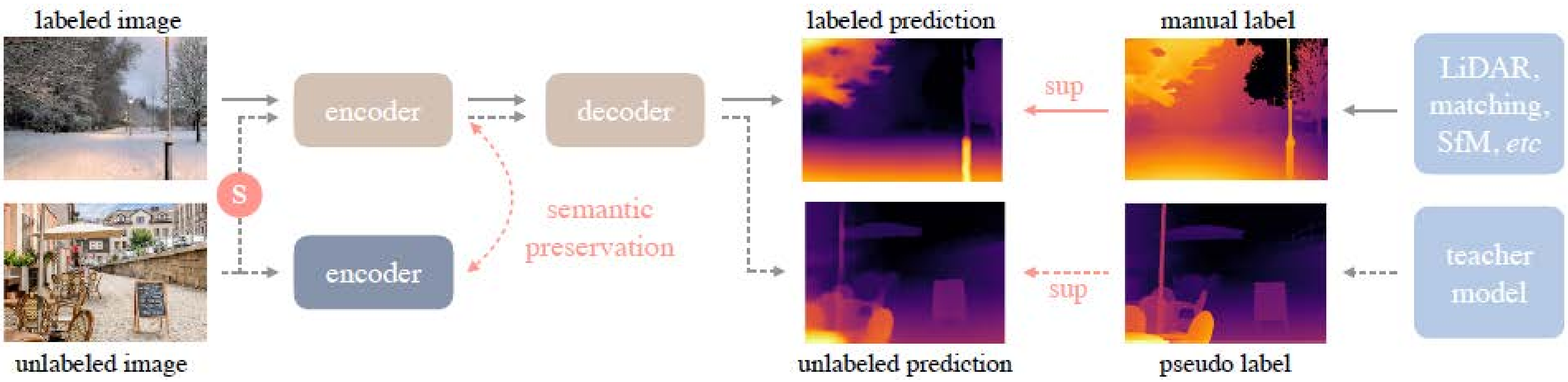
Depth Anything

- Large Monodepth model [Yang-CVPR-2024] by TikTok
- Trained from 1.5M of depth labeled images + 62M of unlabeled images
  - Semi-Supervised Learning (SSL):
    - Teacher – trained from labeled,
    - Student – trained from labeled + pseudo-labeled (from the Teacher)
  - Normalizing depth (inverse depth, 0-1 range)
  - Strong data augmentation (color jitter, blur, geometry - CutMix)

$$u_{ab} = u_a \odot M + u_b \odot (1 - M)$$

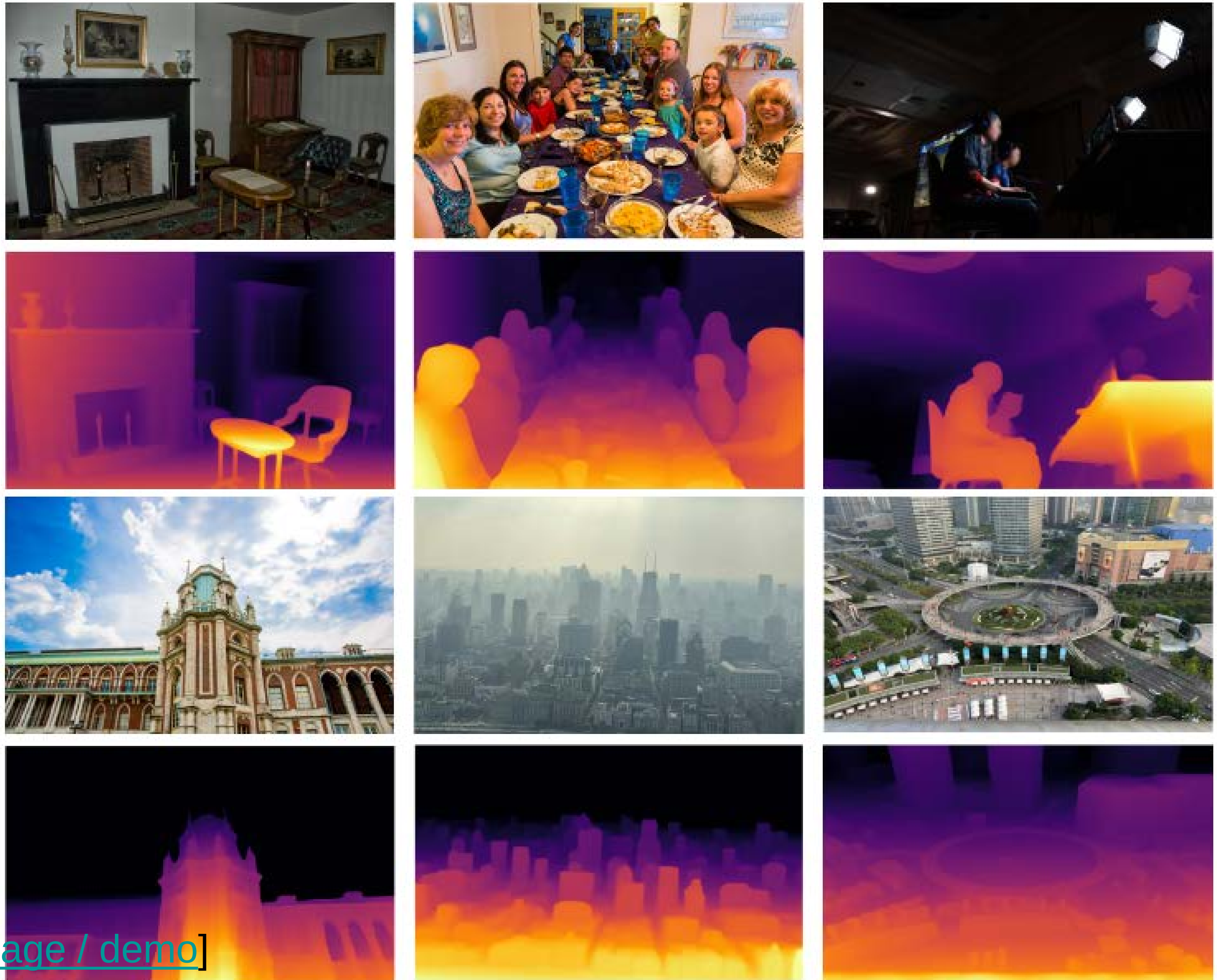
$$\mathcal{L}_u^M = \rho\big(S(u_{ab}) \odot M, T(u_a) \odot M\big),$$
$$\mathcal{L}_u^{1-M} = \rho\big(S(u_{ab}) \odot (1 - M), T(u_b) \odot (1 - M)\big)$$

  - Semantic preservation (alignment with DINO features)



Depth Anything

- Qualitative results



[project-page / demo]

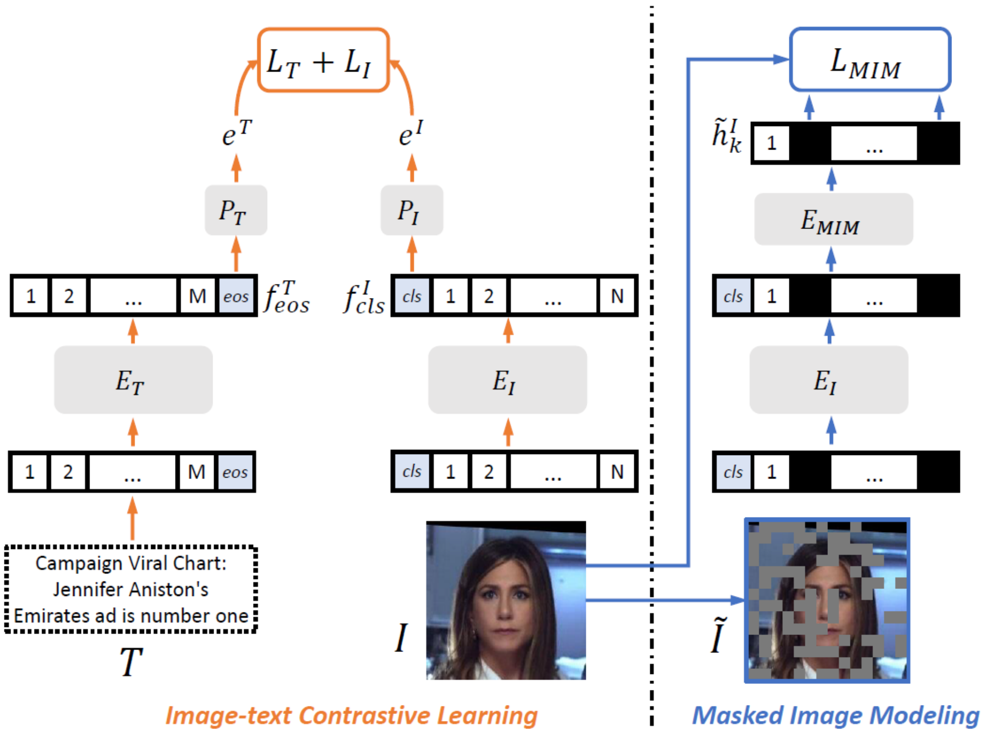
FARL – FAcial Representation Learning

- FARL [Zheng-CVPR-2022] by Microsoft
- Universal representation for face images
- Trained from 20M LAION-Face dataset
- Combines text-image contrastive learning and masked image modeling

$$L_I = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(e_i^I e_i^T/\sigma)}{\sum_{j=1}^{B}\exp(e_i^I e_j^T/\sigma)},$$

$$L_T = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(e_i^T e_i^I/\sigma)}{\sum_{j=1}^{B}\exp(e_i^T e_j^I/\sigma)},$$

$$L_{MIM} = -\sum_{k\in\mathcal{M}}\log p\left(q_\phi^k(I)|\tilde{I}\right)$$



- "CLIP for faces", many downstream tasks (segmentation, landmarks, age)

[project page]

Conclusions

- No doubt that the paradigm has shifted
- Turbulent period
  - The research is extremely accelerated, many novel approaches
  - New results are still astonishing
- Large foundation models appear and are usually publicly available