Scalability of Deep Learning

Dr Yifan Liu
Invited Lecture
yf856@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

About me

- Research interests:
  - Dense prediction tasks
  - Efficient model training
  - Self-supervise/unsupervised training
  - Robust models in the wild

Code

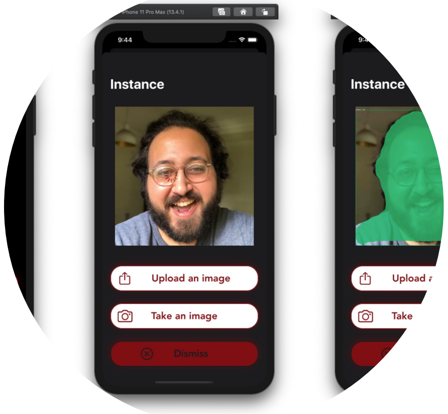
Homepage

Publication

Content

- The power of large model
  - Increased model size
  - Increased labeled training dataset
  - Multimodality
- Efficient model training
  - Knowledge distillation
  - Network pruning/ Quantization

Deploying highly efficient, compact models on edge devices (e.g., AIoT)

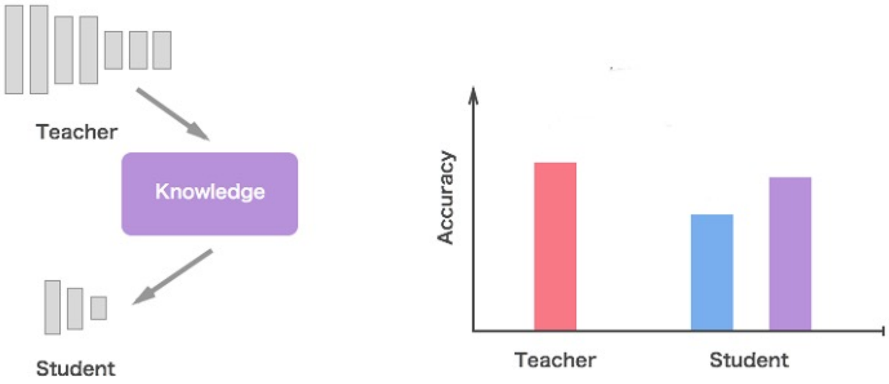


Content

- The power of large model
  - Increased model size
  - Increased labeled training dataset
  - Multimodality
- Efficient model training
  - **Knowledge distillation**
  - Network pruning/ Quantization

Knowledge Distillation

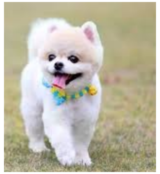
- Knowledge distillation for classification
  - Geoffrey Hinton, (2015)
  - Soften output
  - Compact model (student) learns from large models (teacher)



UNIVERSITY OF CAMBRIDGE

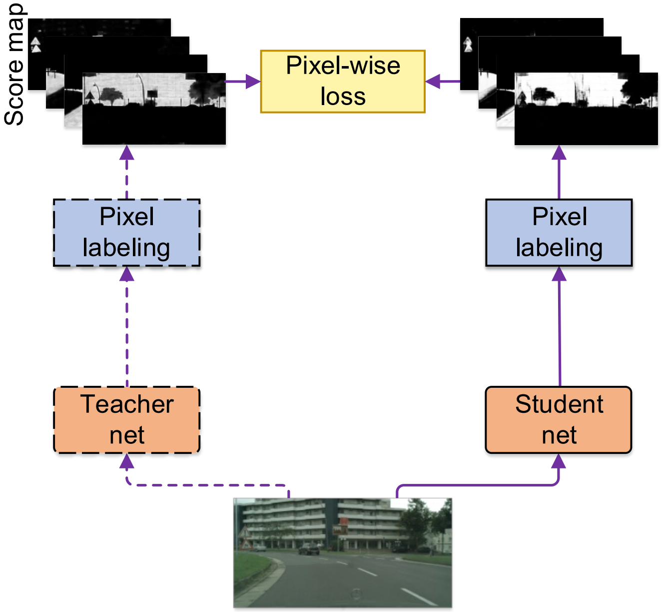
Knowledge Distillation

- Knowledge distillation for classification
  - Geoffrey Hinton, (2015)
  - Soften output
  - Compact model (student) learns from large models (teacher)



Hard label:	1	0	0
Soft target	0.8	0.19	0.01

Knowledge distillation for semantic segmentation



Baseline: applying KD to each pixel on the logits

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

 Structured Knowledge Distillation for Dense Prediction, **Liu, Yifan** and Shu, Changyong and Wang, Jingdong and Shen, Chunhua, TPAMI, 2020

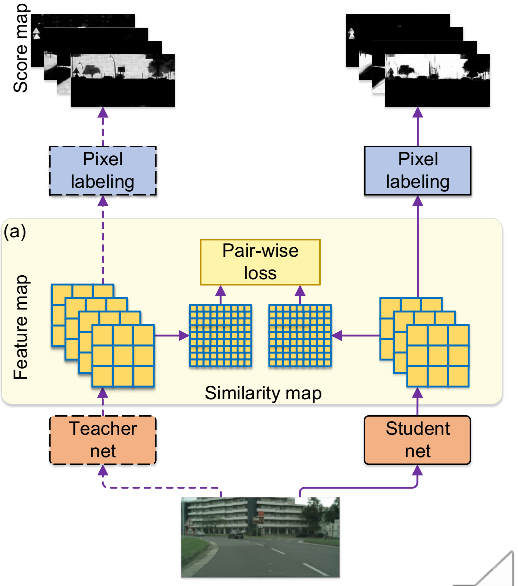
Structural Knowledge Distillation

- Ours: Knowledge distillation considering structural correlations

Idea1: Learn from correlations among spatial locations

✓ **Pair-wise**

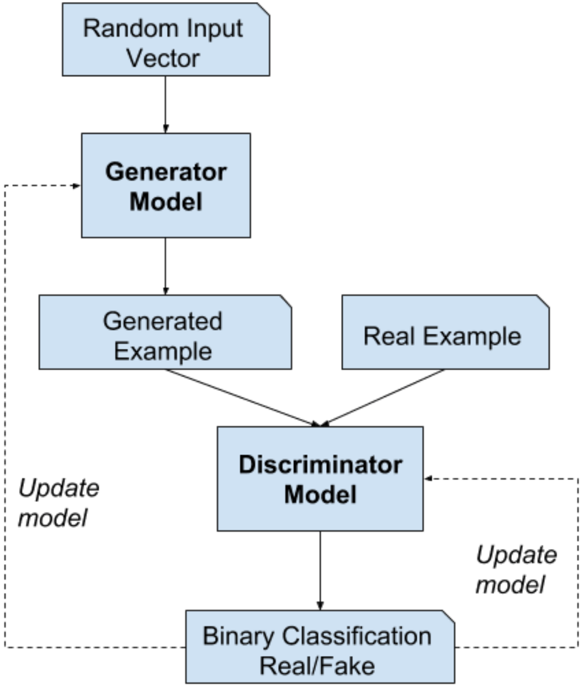
✓ Holistic



Structured knowledge distillation for semantic segmentation, **Liu, Yifan** and Chen, Ke and Liu, Chris and Qin, Zengchang and Luo, Zhenbo and Wang, Jingdong, 2019 ,CVPR

Structured Knowledge Distillation for Dense Prediction, **Liu, Yifan** and Shu, Changyong and Wang, Jingdong and Shen, Chunhua, TPAMI, 2020

External: GAN



Generator: try to generate fake distributions which is similar to the real ones, to fool the discriminator

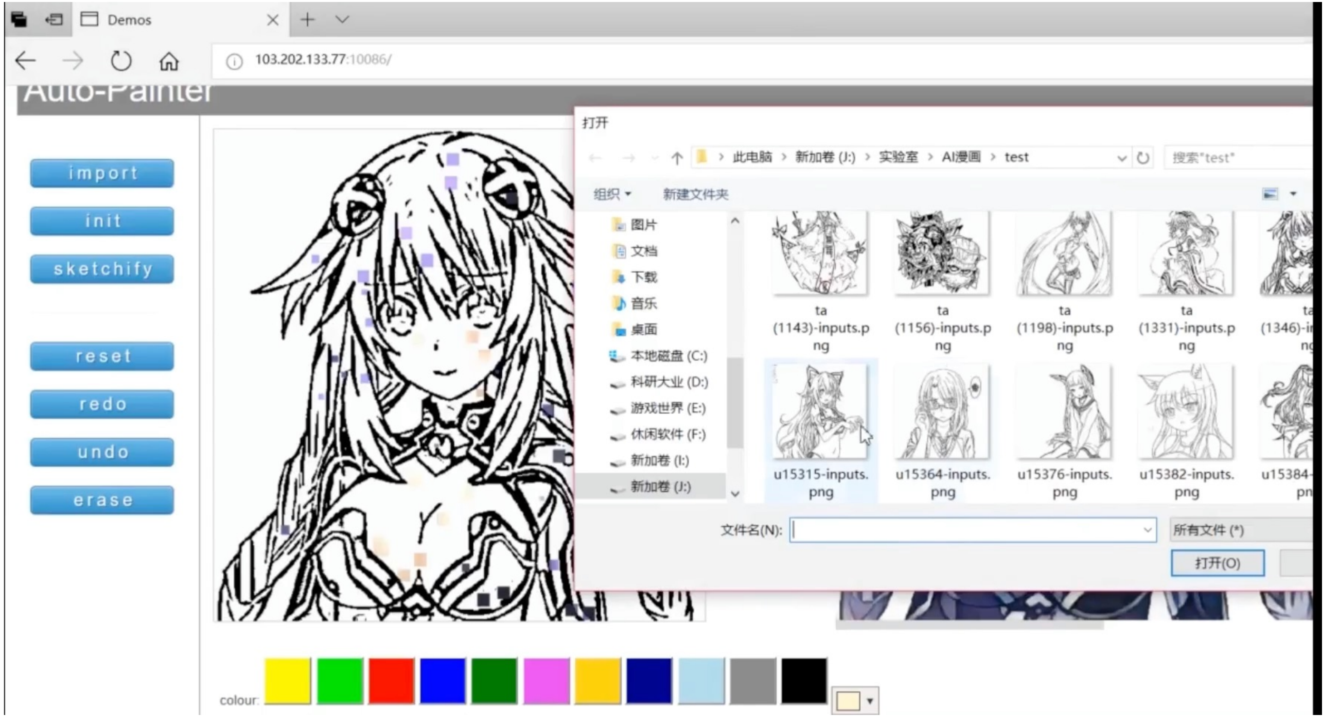
Discriminator : try distinguish between the real distribution and the fake distribution

First proposed in image generation tasks

External: GAN

Analyzing and Improving the Image Quality of StyleGAN, Tero Karras et al. 2019

External: GAN

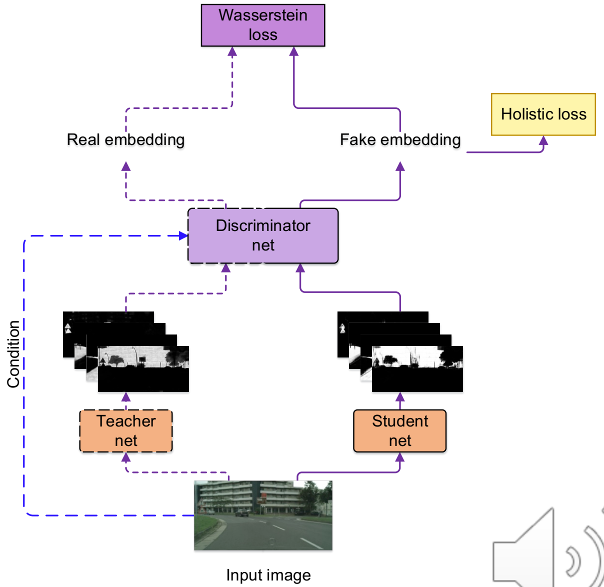


Structural Knowledge Distillation

- Ours: Knowledge distillation considering structural correlations

Idea1: Learn from correlations among spatial locations

- ✓ Pair-wise
- ✓ **Holistic**



Structured Knowledge Distillation for Dense Prediction, **Liu, Yifan** and Shu, Changyong and Wang, Jingdong and Shen, Chunhua, TPAMI, 2020

Spatial distillation

- Mimic

  Minimize the L2 similarity among features

  A 1 x 1 convolution is employed to align the channel of the feature
- Attention transfer

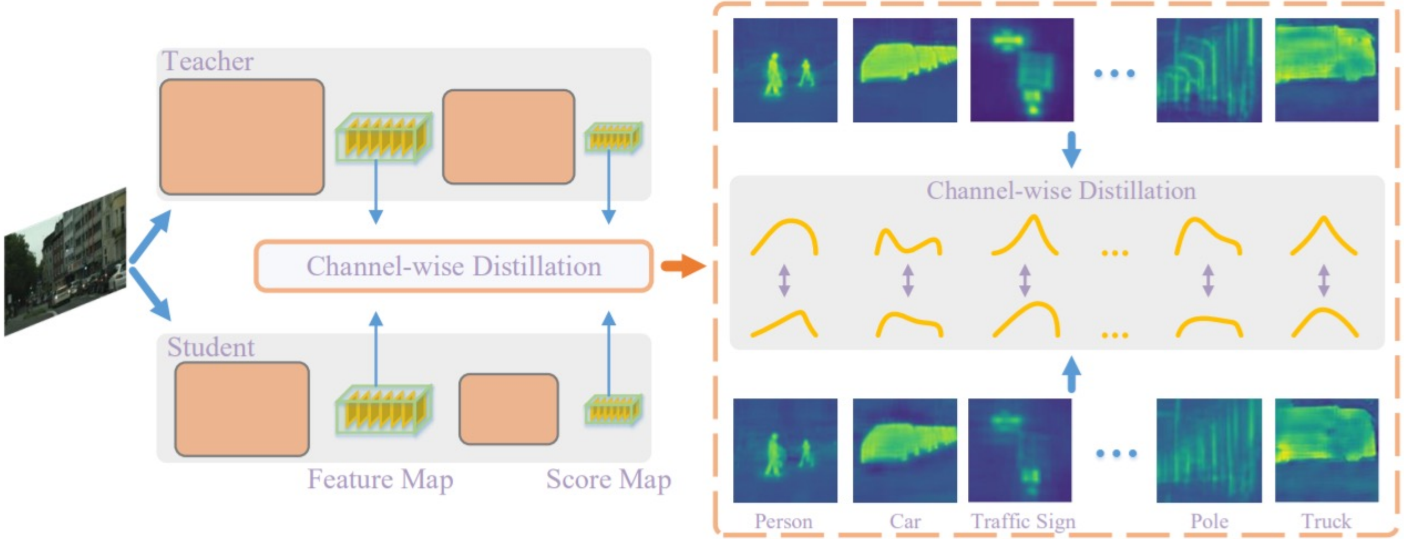
  Get an attention map with one channel from the feature map.

  Merging all the channels into one channel.

UNIVERSITY OF CAMBRIDGE

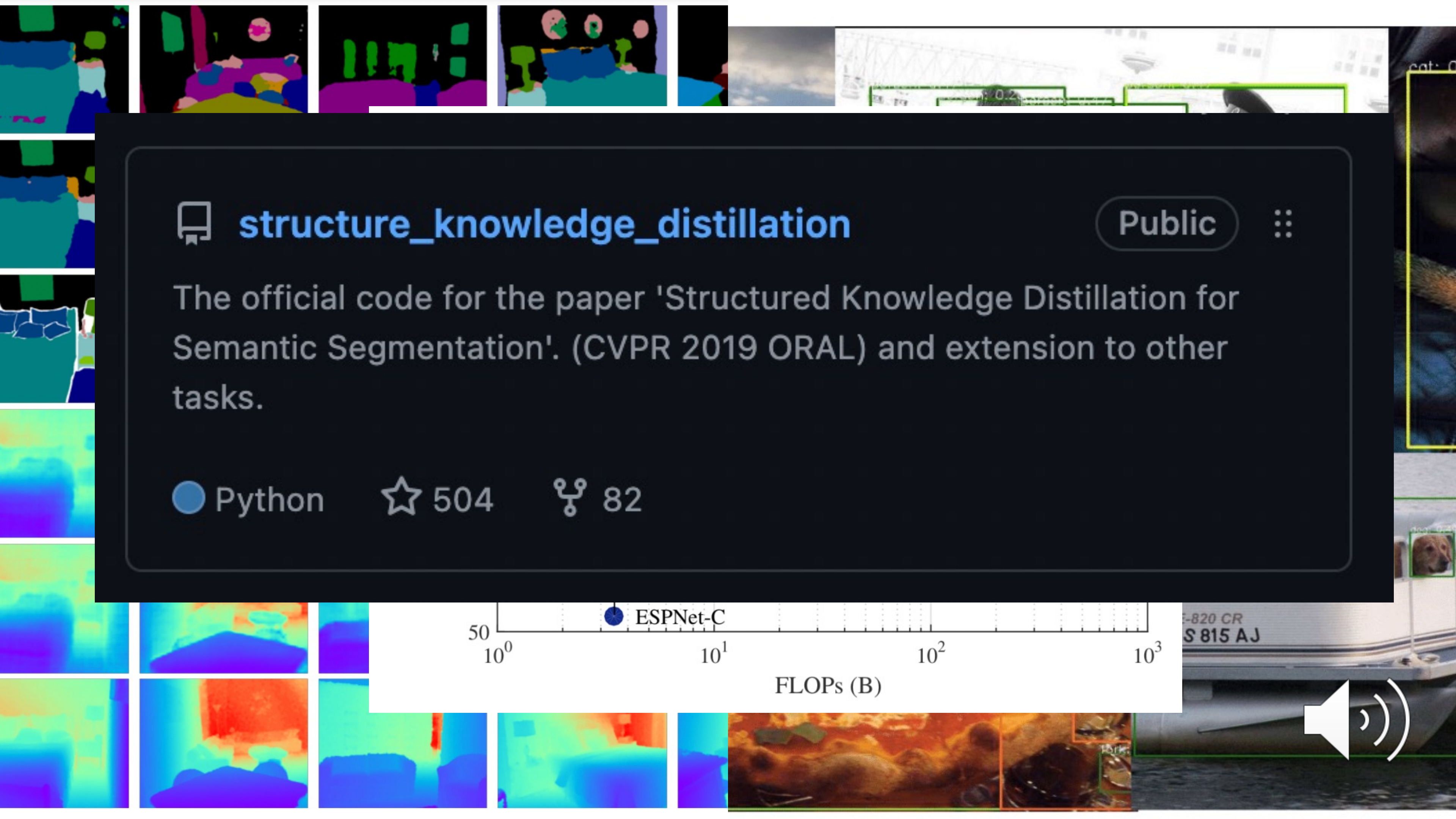
Channel-wise Distillation

- Ours: Knowledge distillation considering the information in the channels.



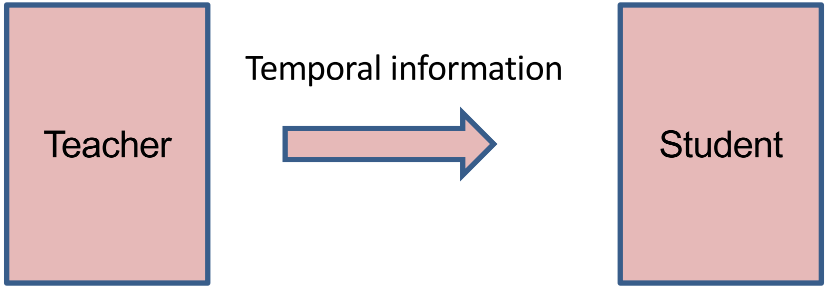
UNIVERSITY OF CAMBRIDGE

Channel-wise Knowledge Distillation for Dense Prediction, **Yifan Liu** *, Changyong Shu*, Jianfei Gao, Zheng Yan, Chunhua Shen, ICCV, 2021

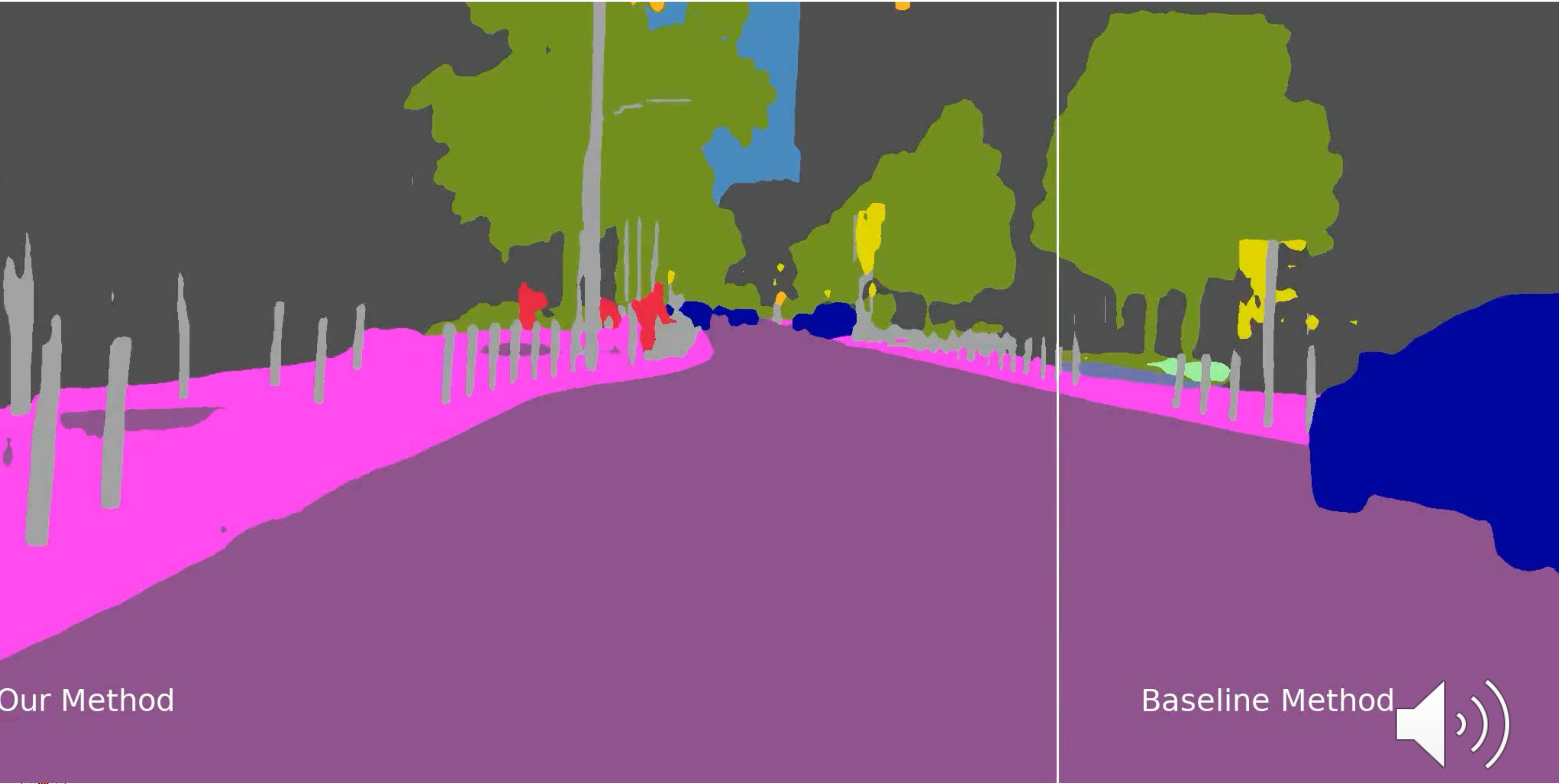

structure_knowledge_distillation
Public
The official code for the paper 'Structured Knowledge Distillation for Semantic Segmentation'. (CVPR 2019 ORAL) and extension to other tasks.
Python
504
82
ESPNet-C
50
10^0
10^1
10^2
10^3
FLOPs (B)
S 815 AJ

Knowledge distillation on video frames

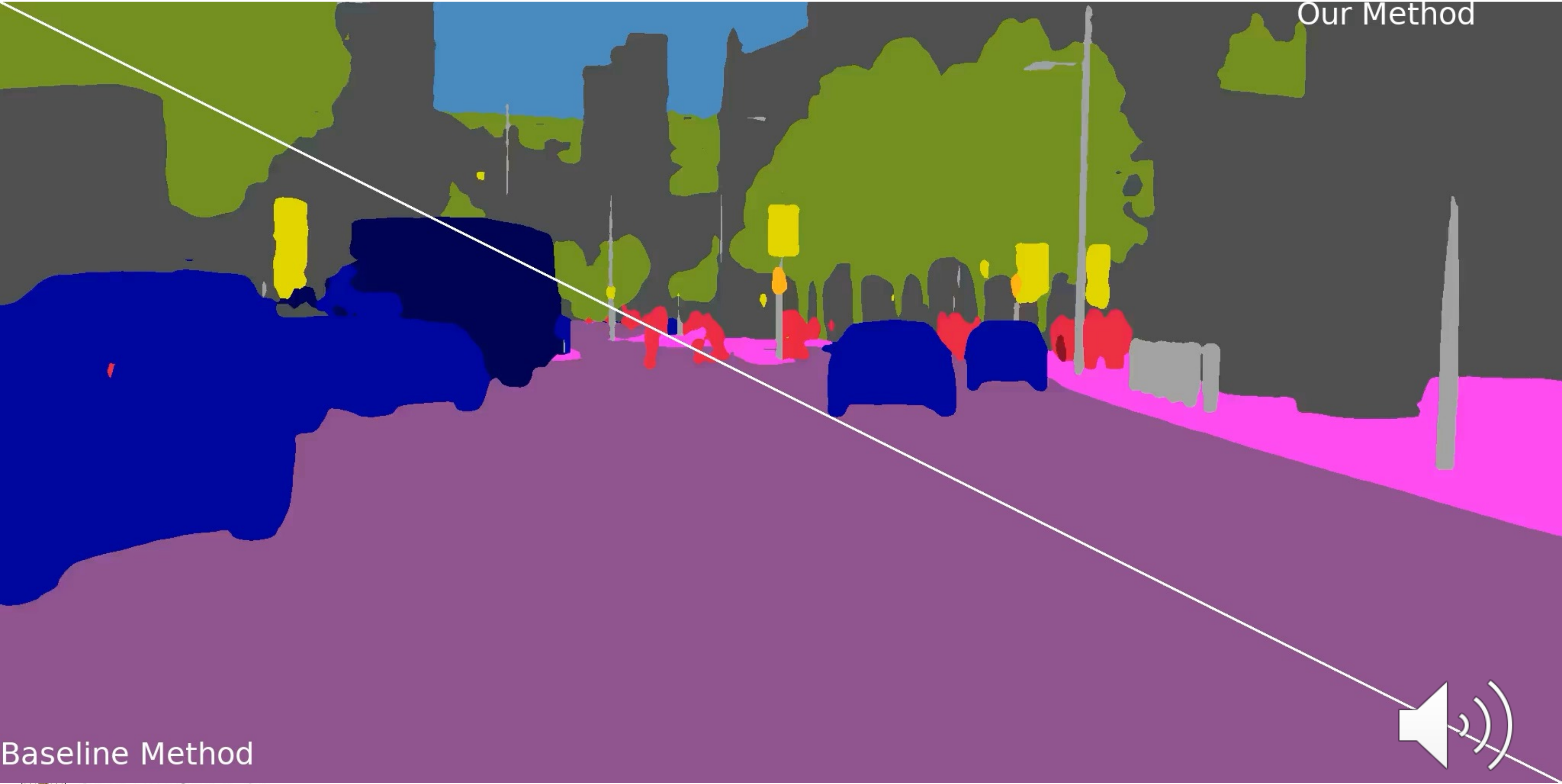
- Core idea:
  - Considering the correlations among frames during training, and **inference on single frames**:
    - Learning from a large **temporally consistent model**
    - Learning the correlations from a large **optical flow model**



Efficient semantic video segmentation with per-frame inference, **Liu, Yifan et al**, ECCV, 2020



CAMBRIDGE



CAMBRIDGE

Content

- The power of large model
  - Increased model size
  - Increased labeled training dataset
  - Multimodality
- Efficient model training
  - Knowledge distillation
  - **Network pruning/ Quantization**

Pruning Happens in Human Brain

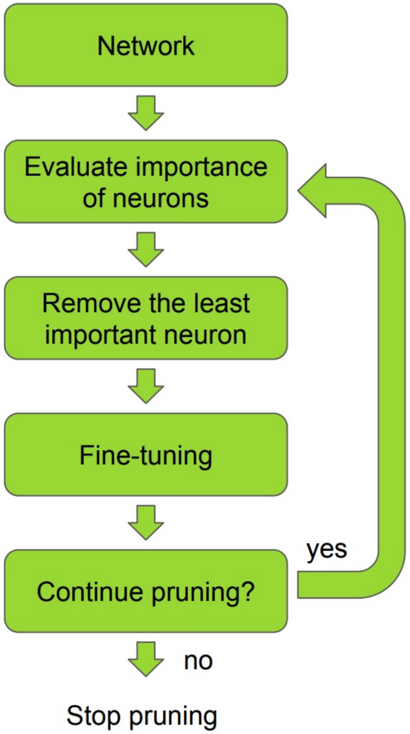
50 Trillion Synapses → 1000 Trillion Synapses → 500 Trillion Synapses

Newborn

1 year old

Adult

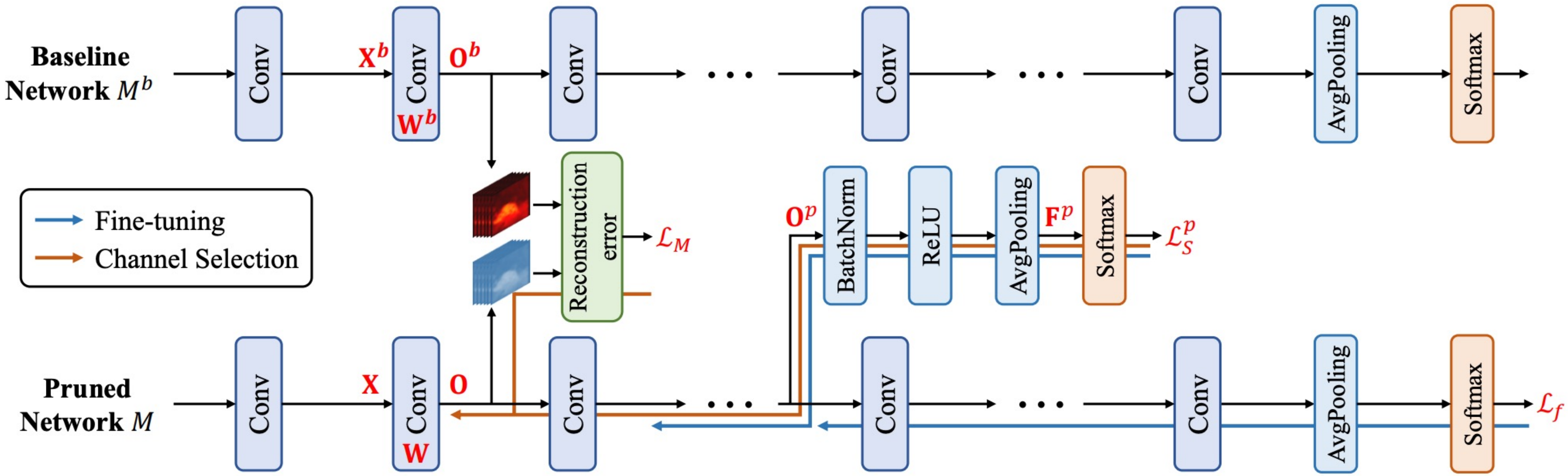
Pruning



1. Prune weights
   - setting individual parameters to zero and making the network sparse.

2. Remove entire nodes from the network
   - make the network architecture itself smaller, while aiming to keep the accuracy of the initial larger network

Channel pruning



UNIVERSITY OF CAMBRIDGE

Network Quantization

- 0.33323134411 → float point → range(0,1)
- 01011010→ int8 →range(-127,128)
- First, we normalize the weight of the network into the range (-127,128)

Network Quantization

- If the output of the network is (X1,X2),
- For a weight x, we can use

new_w = round((X2-X1)/255*x)

UNIVERSITY OF CAMBRIDGE

Network Quantization

- 1. Training
- 2. Quantization
- 3. Retraining

Network Quantization

Challenges:

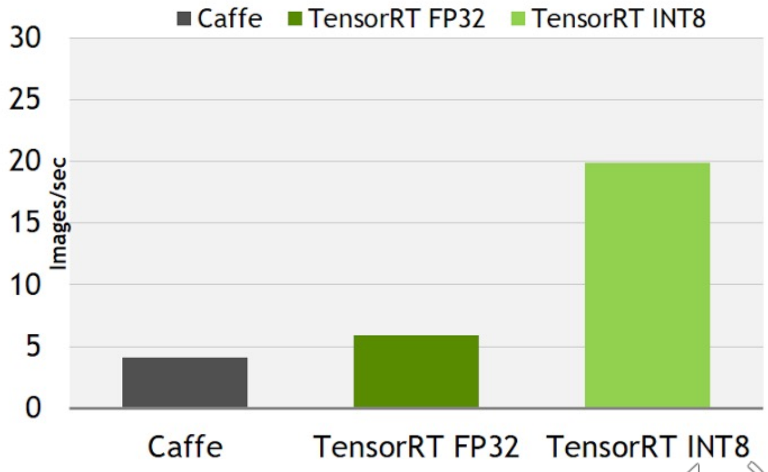
- Non-differentiable quantization functions (e.g., round, sign).
- Quantized structure needs to be re-designed.
- Large gap between theory and reality.

Network Quantization

How efficient?

NVIDIA INT8:   >3x speedup vs. 32-bit

	CAFFE	TENSORRT FP32	TENSORRT INT8
Runtime (ms)	242	170	50
Images/sec	4	6	20
Class IoU	48.4	48.4	48.1
Category IoU	76.9	76.9	76.8



Batch Size = 1, Input/Output Resolution = 512 x 1024

CAMBRIDGE

Summary

- Large model:
  - Powerful
  - Expensive
  - Inefficient

Summary

- Small model:
  - Hard to train
    - Knowledge distillation: Improve the performance
    - Pruning: change the model structure to reduce the size
    - Quantization: keep the structure and change the type of the weights of the network

Thanks!

UNIVERSITY OF CAMBRIDGE