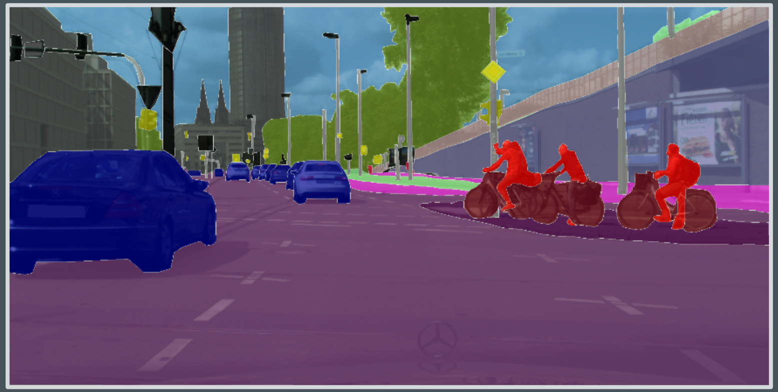
URBAN SCENE SEGMENTATION FOR AUTONOMOUS VEHICLES USING DEEP LEARNING

GROUP: 73

CHUNJHEN LAI / LOUIS LU / TINGSYUAN LIN



From Cityscapes Dataset

BACKGROUND

- Autonomous vehicle: a super computer running on the road.
- A self-driving car has already been involved in five deaths since 2016.
- Perception of objects from image: discern traffic signs, other vehicles, bicycles, and pedestrians.

Computer Vision Mask Semantic Segmentation
(Reference: https://youtu.be/N_g5rO3yj-U)

Lane Detection Using Computer Vision
(Reference: https://youtu.be/fJBHd5S6jgo)

BACKGROUND – CONT.

- Semantic Segmentation: classifies each pixel in image and represents different categories to color.
- Commonly used method: U-Net, SegNet, DeepLab series, FCN, ENet, ICNet, DFN, CCNet , and etc.
- Fully Convolutional Network (FCN): learn mapping in pixel-level prediction , but low resolution.

Figure I: Original Input Image

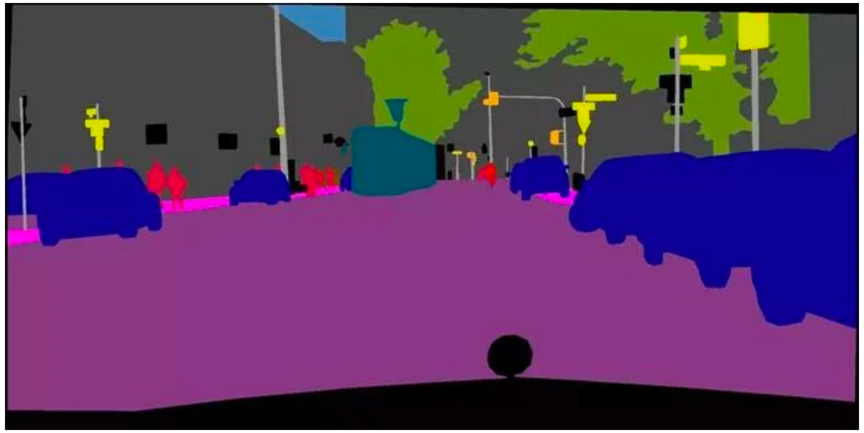
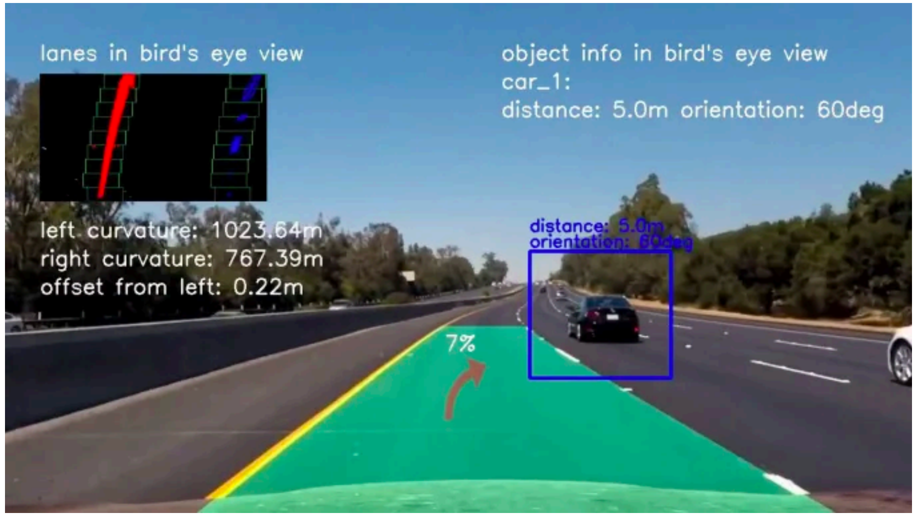
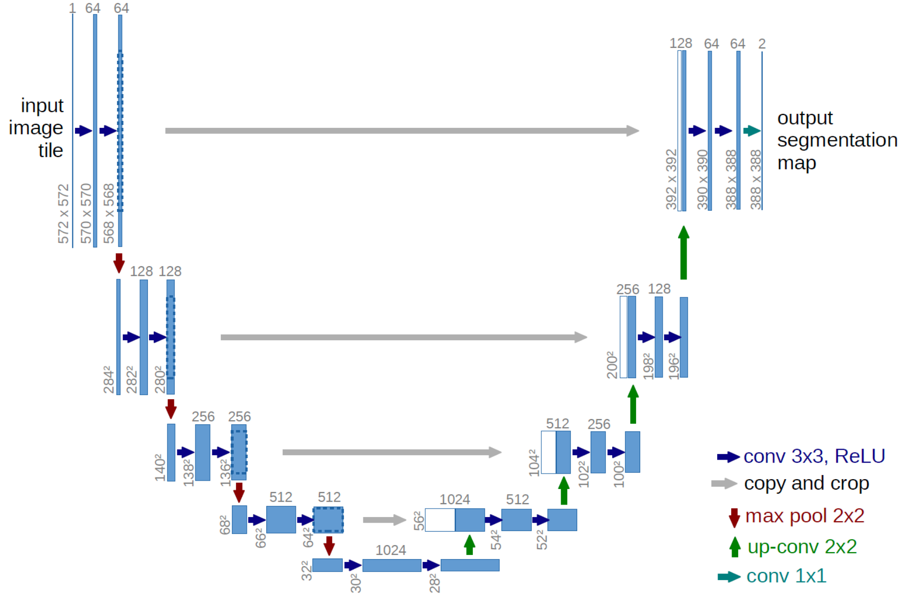


Figure II: Output Segmented Image

DEEP LEARNING CAN HELP SOLVE THIS PROBLEM

- **In deep learning, human gives the rules then neural network learns by itself.**
- Wide coverage and good adaptability on different condition.
- More effective on target classification with limited data sets such as U-Net.

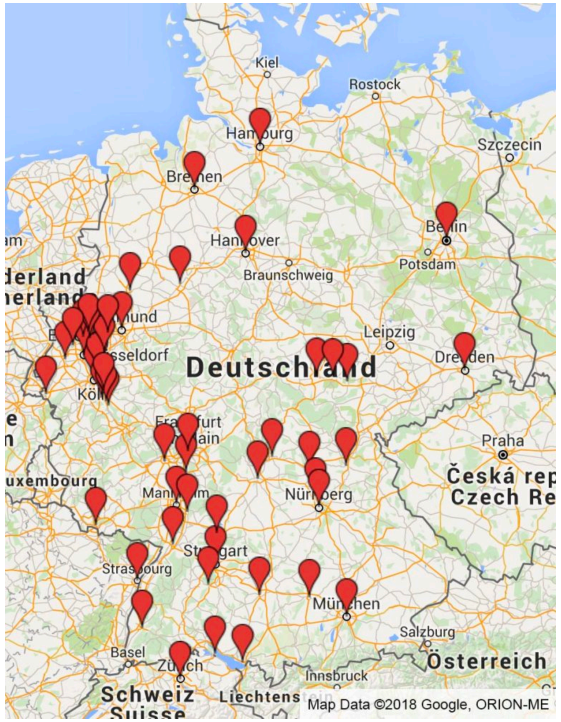




Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation.

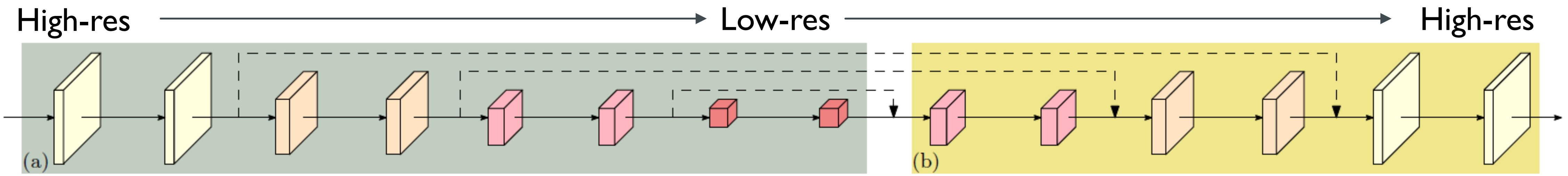
DETAILS ON THE DATASET

- The Cityscapes Dataset: semantic image annotation of urban street scenes.
- Complexity: 30 classes such as humans, cars, road, sky, and etc.
- Diversity:
  - 50 cities in Europe
  - Several months (Spring, Summer, Fall)
  - Weather conditions
  - Large number of dynamic objects
  - Varying scene layout and background
- Volume:
  - 5 000 annotated images with fine annotations
  - 20 000 annotated images with coarse annotations



Contained cities

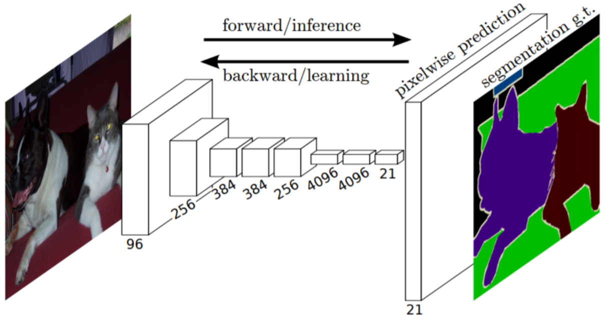
LITERATURE SURVEY

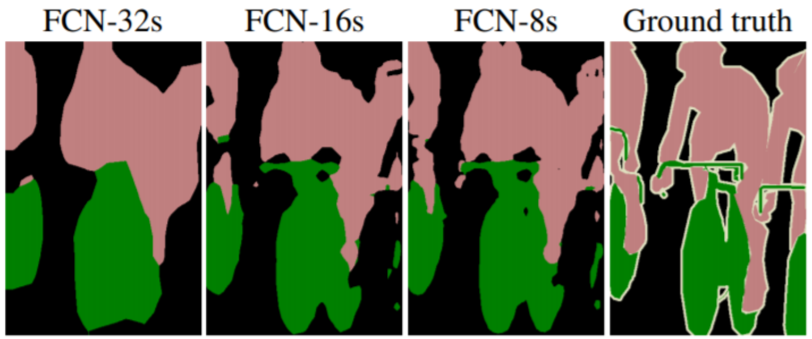


- **(a) Learning low-resolution representations**
  - Compute low-resolution representations by removing the fully-connected layers in a classification network, and estimate their **coarse segmentation maps**
  - FCN, ResNet, VGGNet, and etc.
  - Problems: output resolution

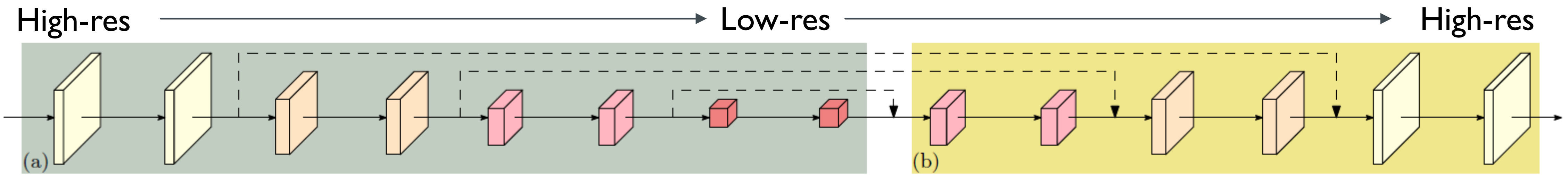
Example of using FCN

Jonathan Long, Evan Shelhamer, Trevor Darrell. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431-3440



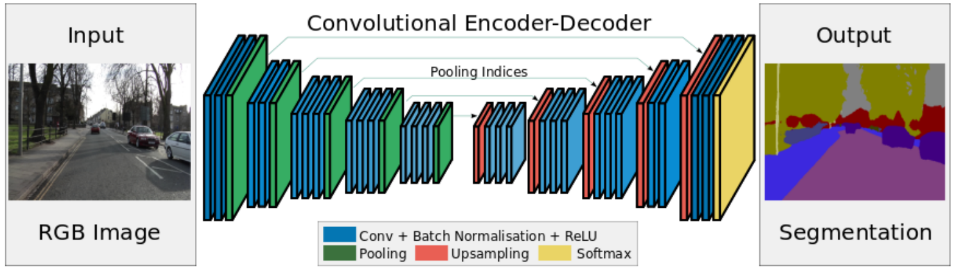


LITERATURE SURVEY – CONT.



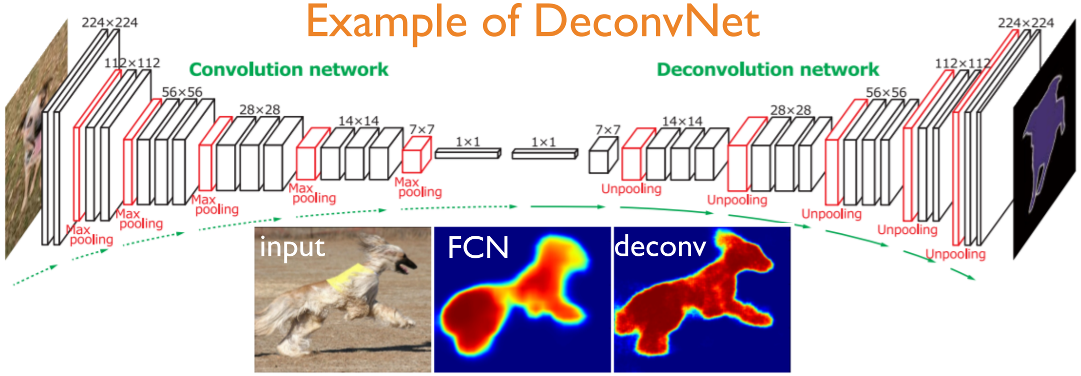
- **(a+b) Recovering high-resolution representations**
  - Use **up-sample** process to gradually **recover** the high-res. representations from the low-res. representations
  - Some models can also **maintain high-resolution** representations
  - DeconvNet, U-Net, SegNet, encoder-decoder, HR-Net and etc.

Example of SegNet



Alex Kendall, Vijay Badrinarayanan and Roberto Cipolla. SegNet

Example of DeconvNet

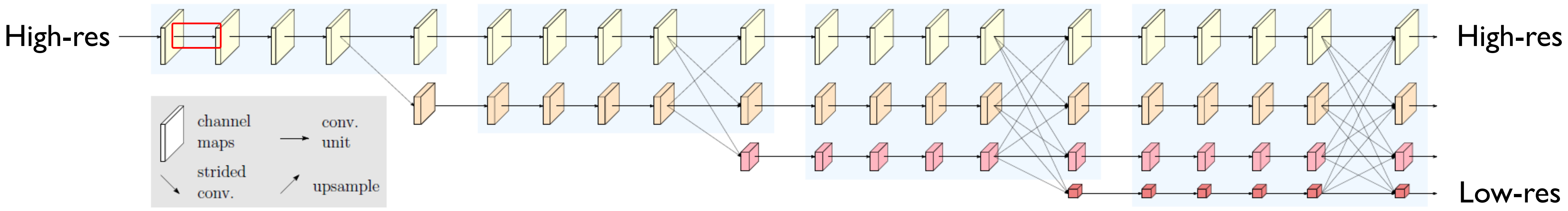


H. Noh, S. Hong and B. Han, "Learning Deconvolution Network for Semantic Segmentation," *2015 IEEE ICCV*, Santiago, 2015, pp. 1520-1528

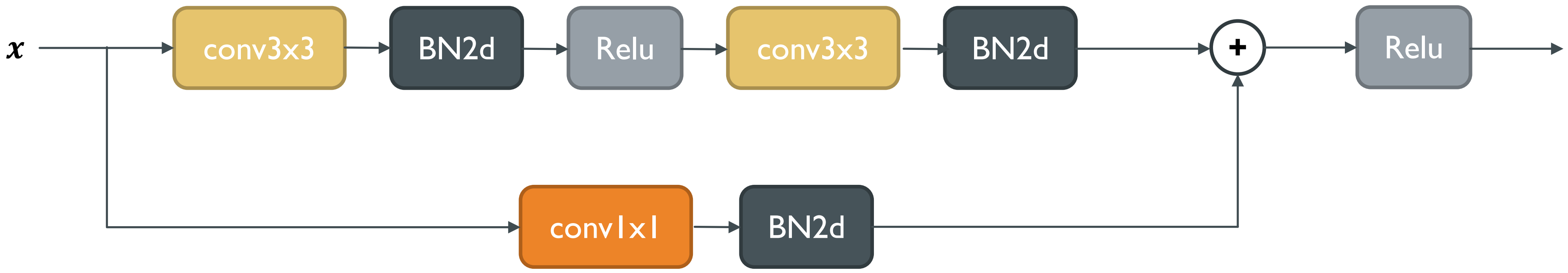
DETAILS ON THE MODEL USED

HRNetV2p

J. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020*



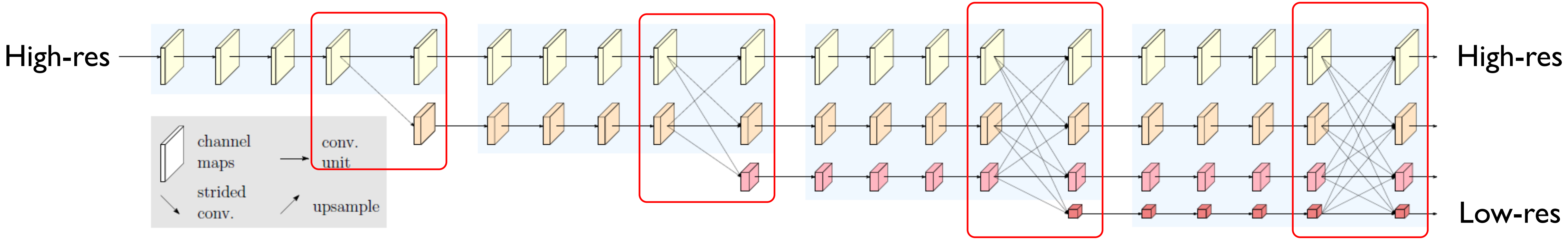
- **Forward propagation block (similar to ResNet block)**



DETAILS ON THE MODEL USED – CONT.

HRNetV2p

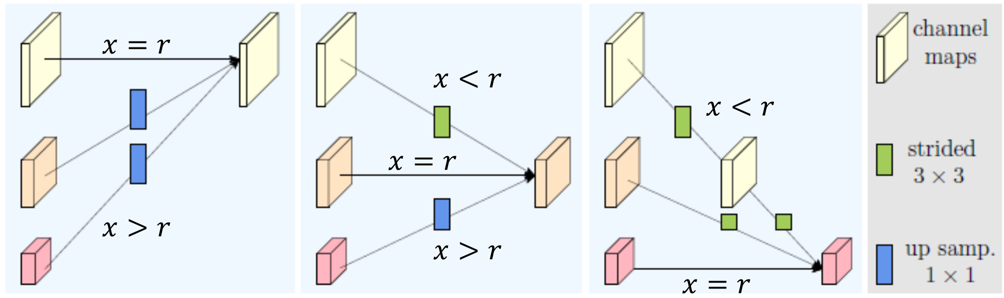
J. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020*



- **Rules of multi-resolution fusion**

$$R_r^{output} = \sum_x f_{xr}\left(R_x^{input}\right), r = 1,2,3$$

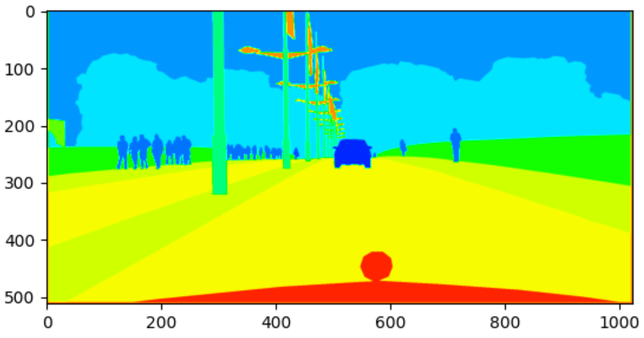
$$f_{xr}(\boldsymbol{R}) = \begin{cases} downsample\ (stride = 2, 3\times3\ conv), & x < r \\ \boldsymbol{R}, x = r \\ upsample\ (1\times1\ conv), & x > r \end{cases}$$

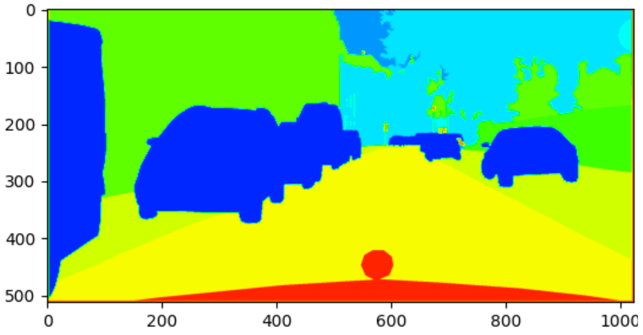


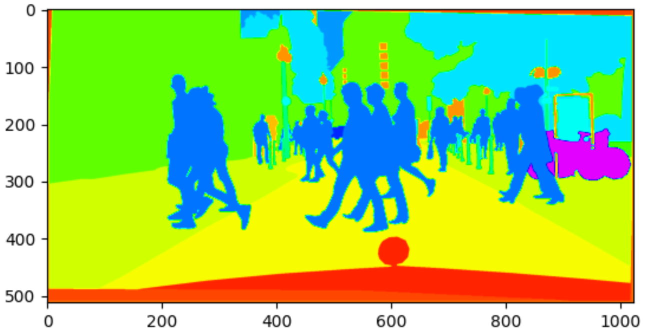
where $\boldsymbol{R}$ is the representations, r is resolution index and $f_{xr}(\cdot)$is the transform function.

RESULTS/OBSERVATIONS

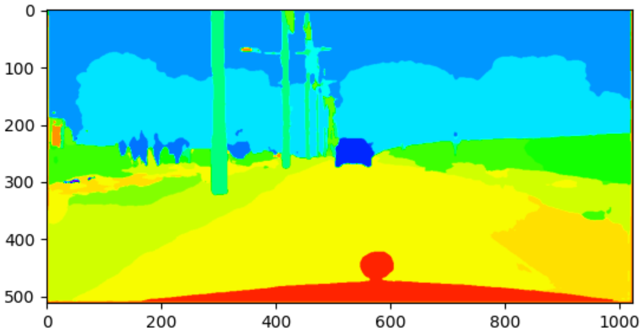
Ground Truth

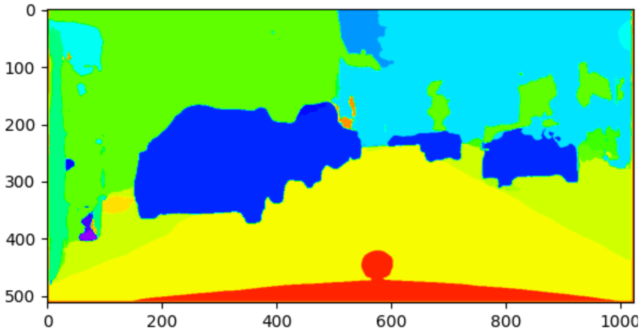


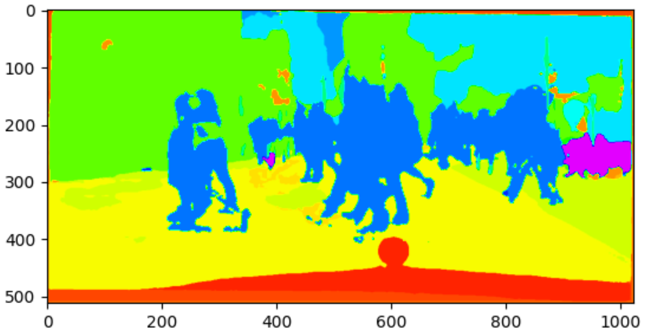




Prediction

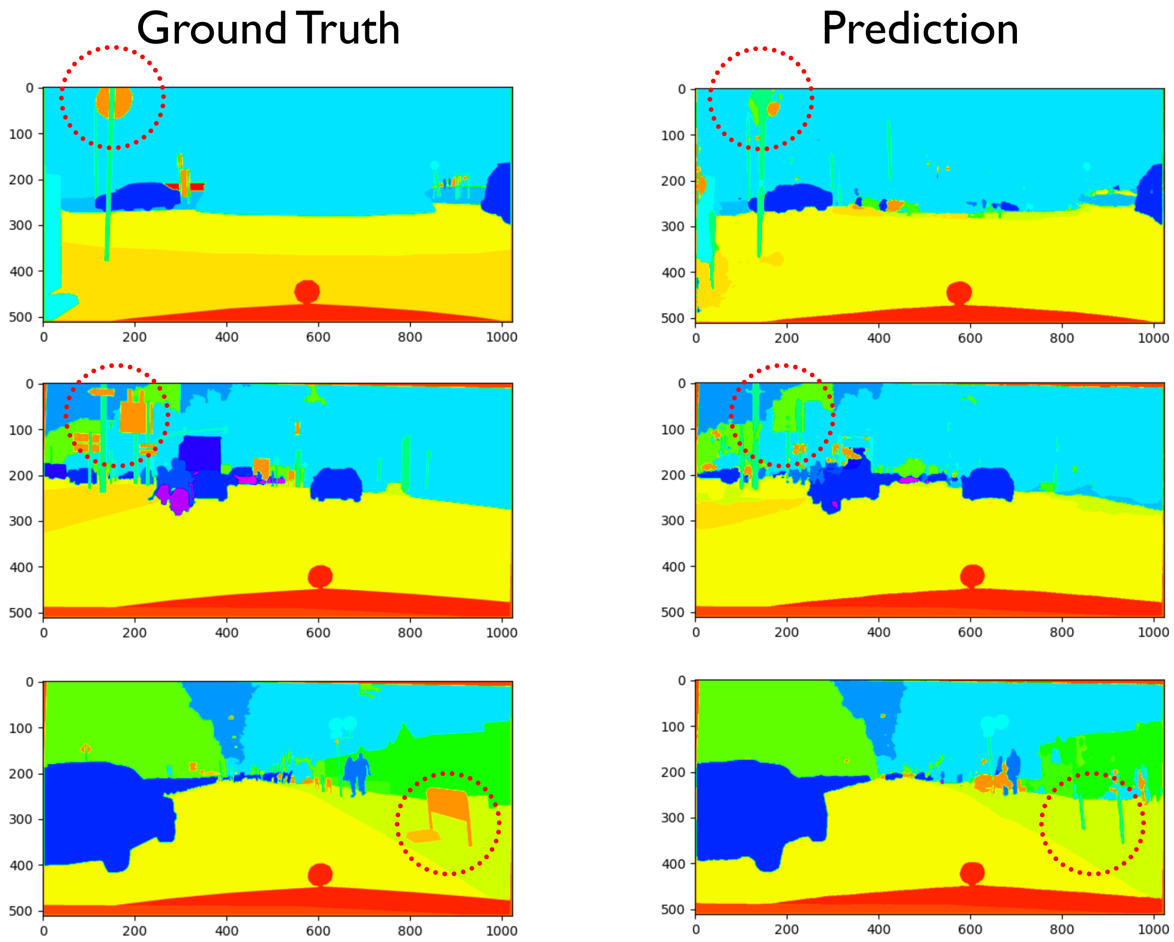






- Our predicted results (false color) can successfully indicate the classes, such as cars, human, constructions and roads, after 100 epochs.

RESULTS/OBSERVATIONS – CONT.



- However, we found something like the traffic signs cannot be predicted very well.
- Reasons:
  - We use cross-entropy as **loss function and the weighs** of all classes are the same
  - Area of traffic signs are small, so they hardly contribute to the loss, while large-area classes like road, sky and constructions make much contribution to the loss

FURTHER ITEMS TO BE COMPLETED

- **Improvement of loss function**
  - Construct **label mapping matrix**
    - not all of the labels are useful; some of them can be ignored (eg. -1 for void)
  - Construct the **class weight tensor**
    - Increase/decrease some of the weight
    - Need several experiments
- **IoU (intersection over union)**
  - Calculate the new confusion matrix

REFERENCES

- Jonathan Long, Evan Shelhamer, Trevor Darrell. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431-3440
- Alex Kendall, Vijay Badrinarayanan and Roberto Cipolla. SegNet (https://mi.eng.cam.ac.uk/projects/segnet/)
- H. Noh, S. Hong and B. Han, "Learning Deconvolution Network for Semantic Segmentation," *2015 IEEE ICCV*, Santiago, 2015, pp. 1520-1528
- J. Wang *et al*., "Deep High-Resolution Representation Learning for Visual Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2020.2983686.