UberNet: Training a 'Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory.

Summary:

Introduces a convolutional neural network (CNN) that jointly handles low-, mid-, and high-level vision tasks in a unified architecture that is trained end-to-end and simultaneously addresses:

a. boundary detection
b. normal estimation
c. saliency estimation
d. semantic segmentation
e. human part segmentation
f. semantic boundary detection
g. region proposal generation
h. object detection

∴ There is no free lunch, and the performance measures of the different tasks act like communicating vessels.

(= Improving normal estimation comes at the cost of decreasing performance in remaining tasks.)

<UberNet Architecture>

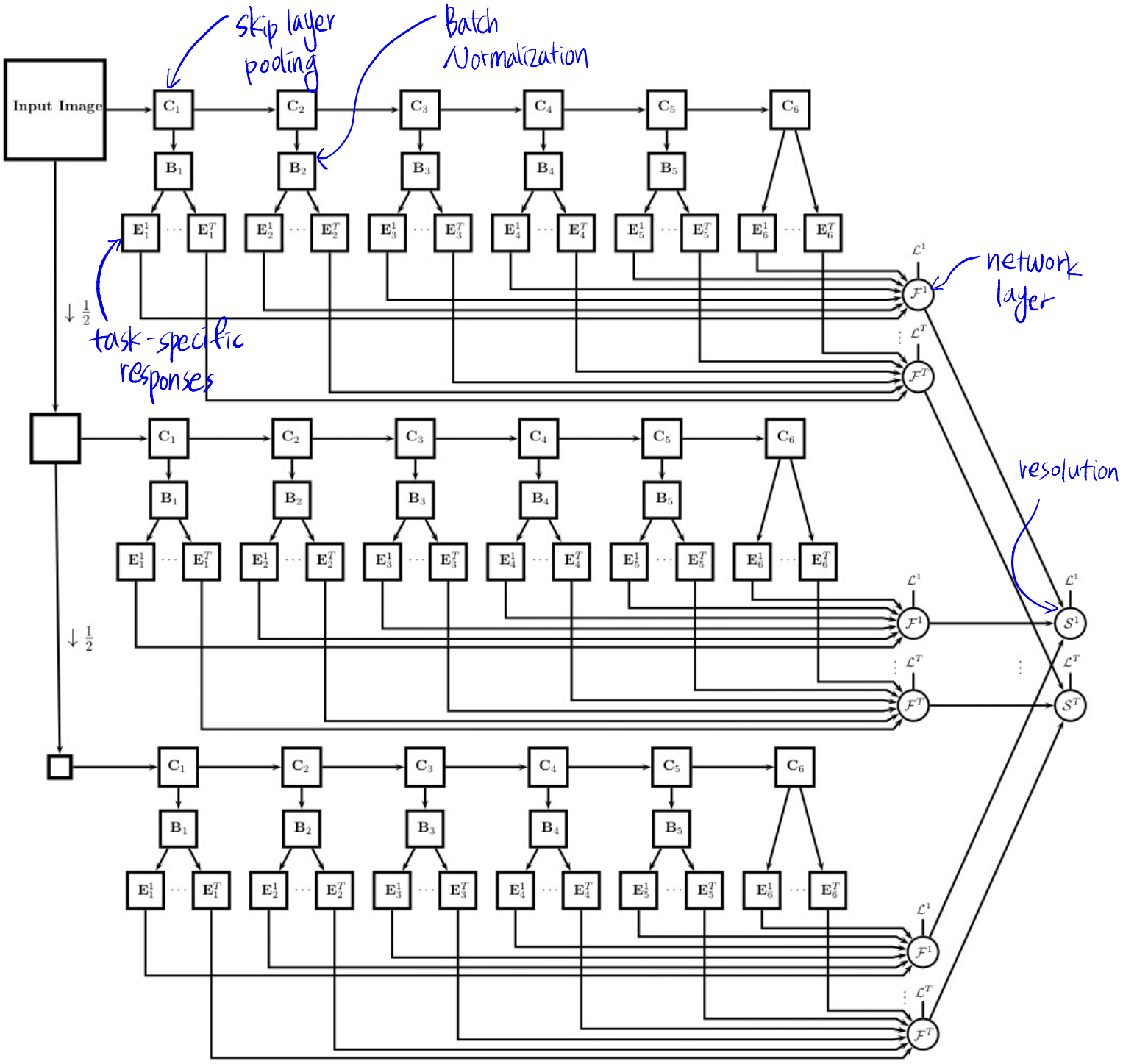
: A minimumal number of additional, task-specific layers on top of a common CNN trunk that is based on the VGG network

- Skip Layers
  - Combine the top-layer neurons with the activations of intermediate neurons to form the network output
- Skip-layer normalization
  - batch normalization (applied except for the last layer)
- Cumulative task-specific operations

$$S_t = W_t^T f = \sum_{k=1}^{6} W_{t,k}^T f_k$$

↑ computing intermediate results per layer

- Fusion layers
  - Instead of simply adding the scores (sum-fusion), one can accelerate training by concatenating the score maps and learning a linear function that operates on top of the concatenated score maps.
- Atrous convolution
  - convolution with holes which allows us to control the spatial resolution of the output layer.
- Multi-resolution CNN
  - an image pyramid and pass scaled versions of the same image through CNNs with shared weights. → can deal with the scale variability of image patterns.


skip layer pooling
Batch Normalization
Input Image
network layer
task-specific responses
resolution