UberNet

Kamil Bladoszewski

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

Plan of presentation

1. Introduction
2. Tasks
3. Architecture
4. Multi-Task Training
5. Low memory usage
6. Results
7. Bibliography

Introduction

The *swiss knife* of Convolutional Neural Networks

Abstract

In this work we introduce 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. Such a universal network can act like a 'swiss knife' for vision tasks; we call this architecture an UberNet to indicate its overarching nature.

Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

The *swiss knife* of Convolutional Neural Networks

Source: https://assets.victorinox.com/mediahub/33849/560Wx490H/SAK_1_6795__S2.jpg

The *swiss knife* of Convolutional Neural Networks

Source: https://odditymall.com/includes/content/giant-swiss-army-knife-thumb.jpg

The *swiss knife* of Convolutional Neural Networks

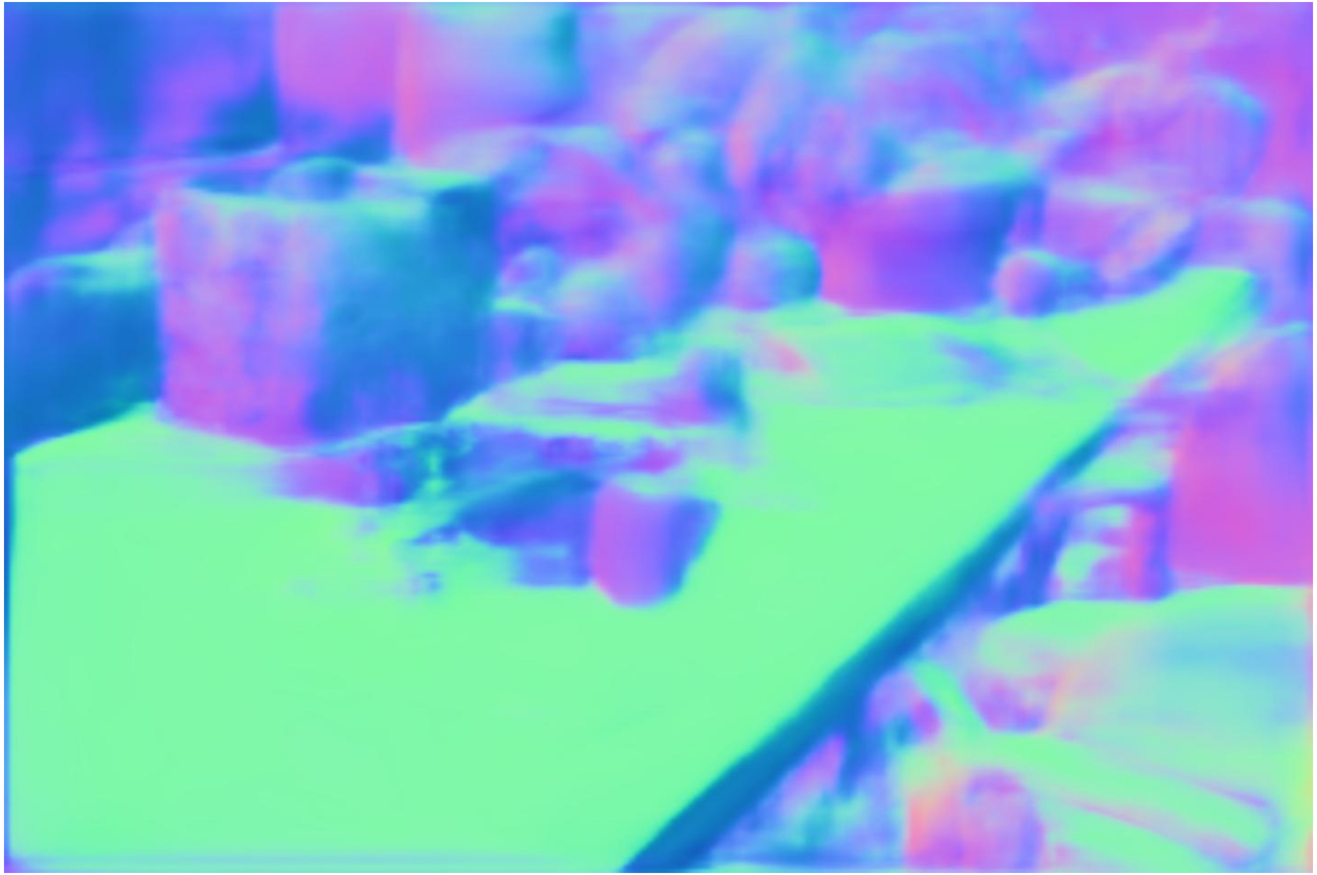
Source: https://assets.victorinox.com/mediahub/37364/640Wx560H/SAK_1_3603__S1.jpg

Tasks

Boundary Detection

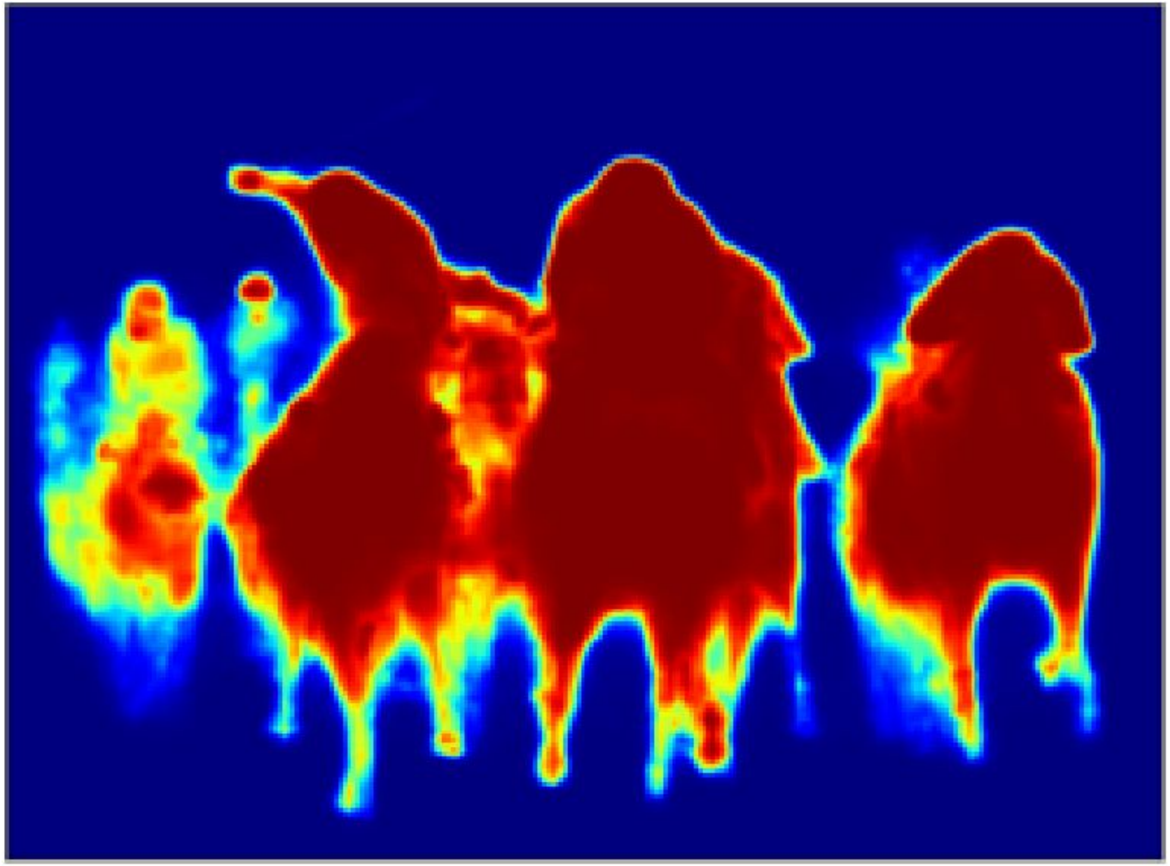
Source: *Pascal in Detail Dataset* - https://sites.google.com/view/pasd

Normal Estimation



Source: http://www.cs.cmu.edu/~aayushb/marrRevisited/Friends/Random06/

Saliency Estimation



Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

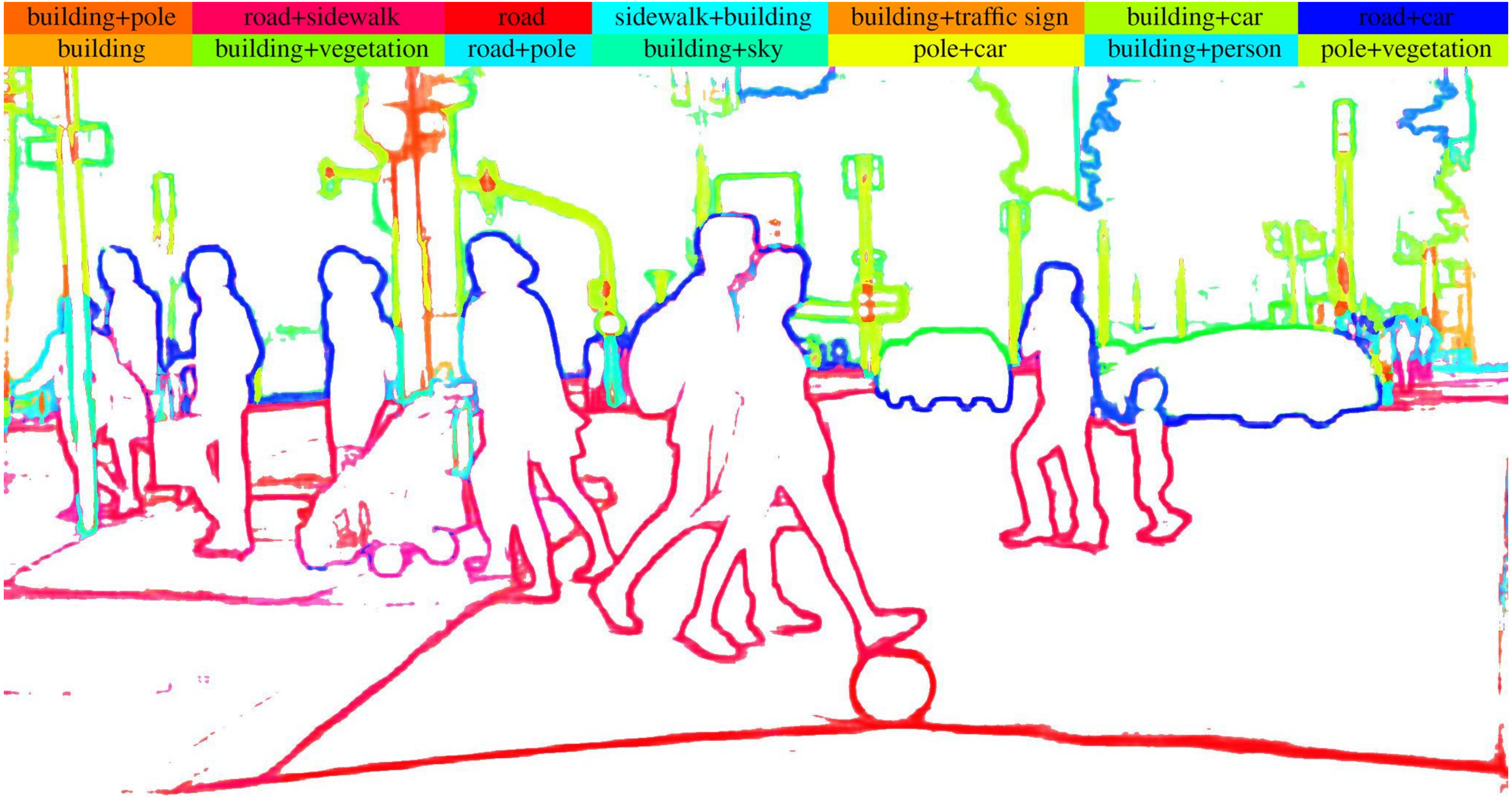
Semantic Segmentation

Source: *Pascal in Detail Dataset* - https://sites.google.com/view/pasd

Human Part Segmentation

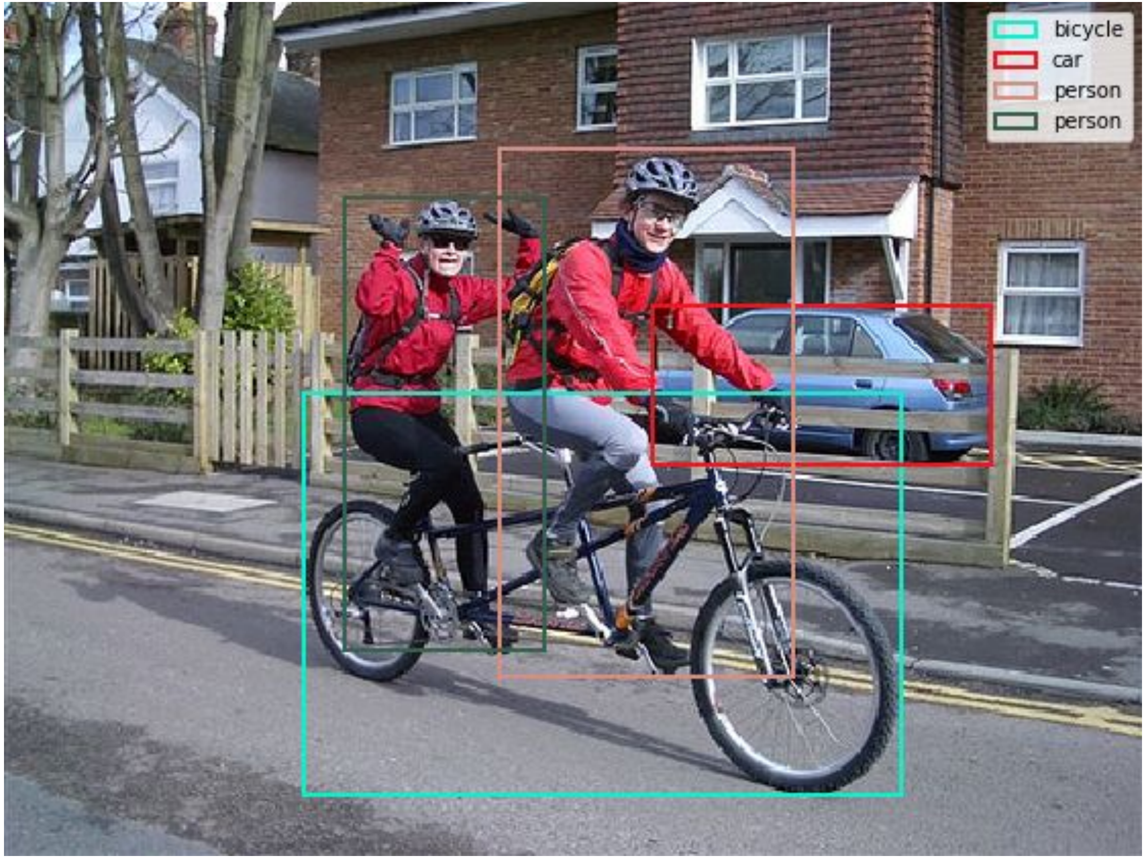
Source: *Pascal in Detail Dataset* - https://sites.google.com/view/pasd

Semantic Boundary Detection



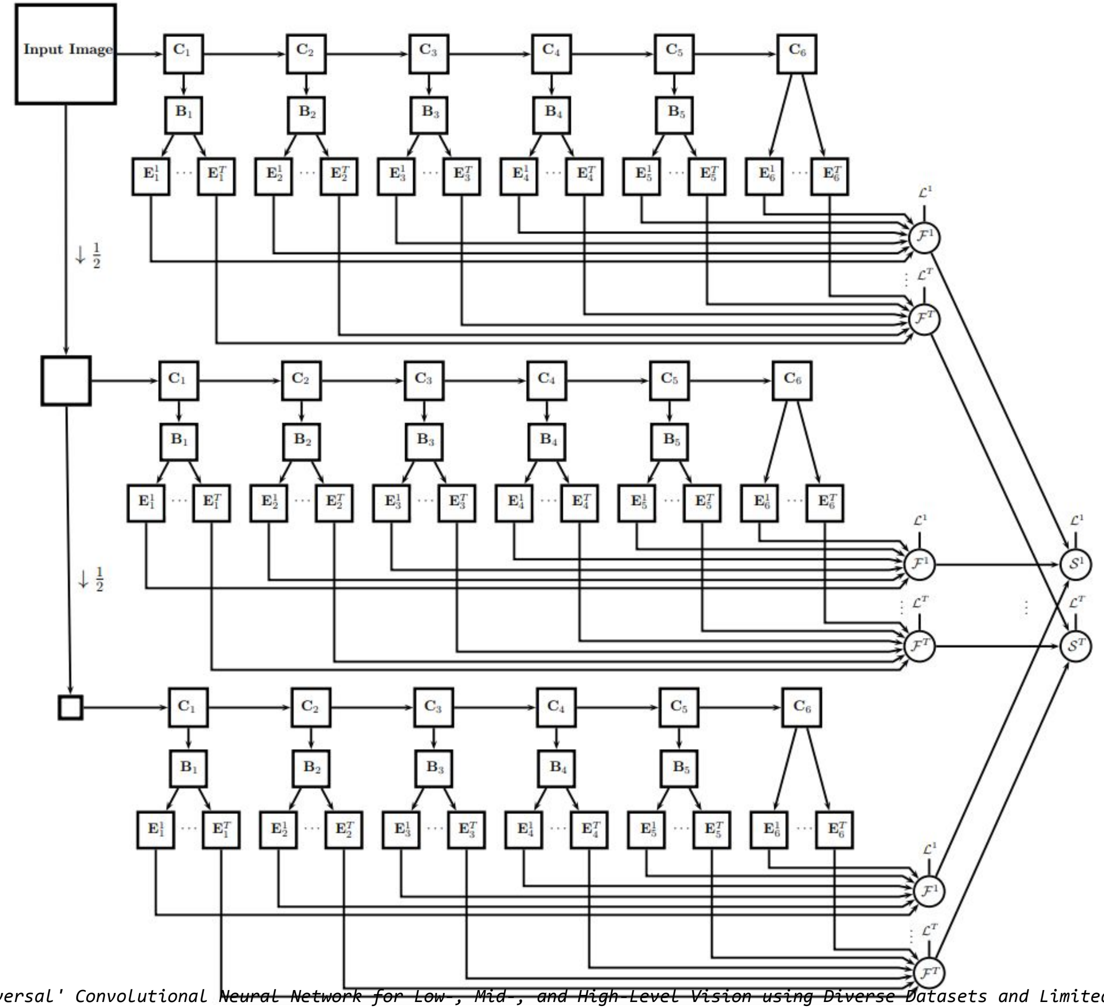
Source: https://chrisding.github.io/projects.htm *CASENet: Deep Category-Aware Semantic Edge Detection (MERL)*

Region Proposal Generation and Object Detection

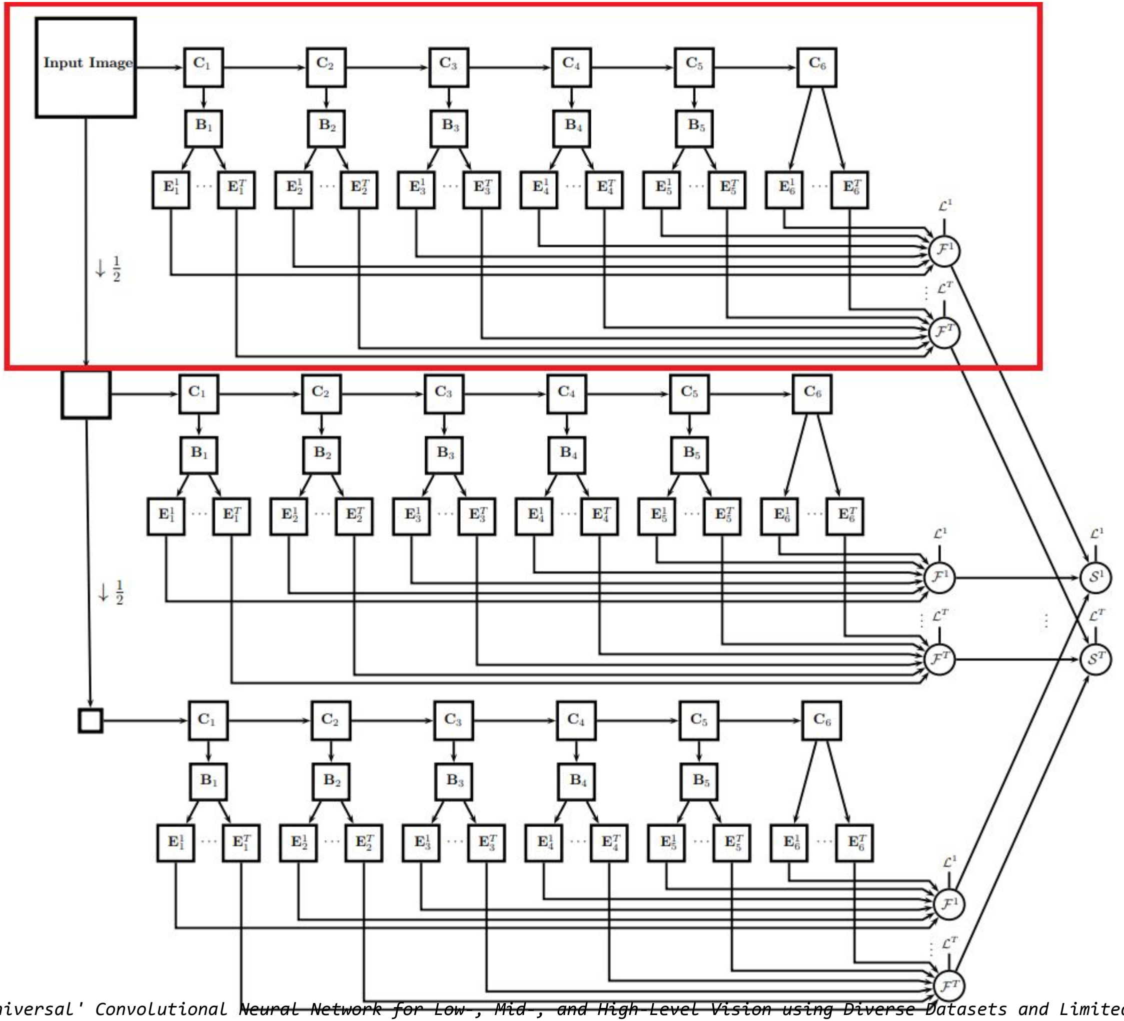


Source: *Pascal in Detail Dataset* - https://sites.google.com/view/pasd

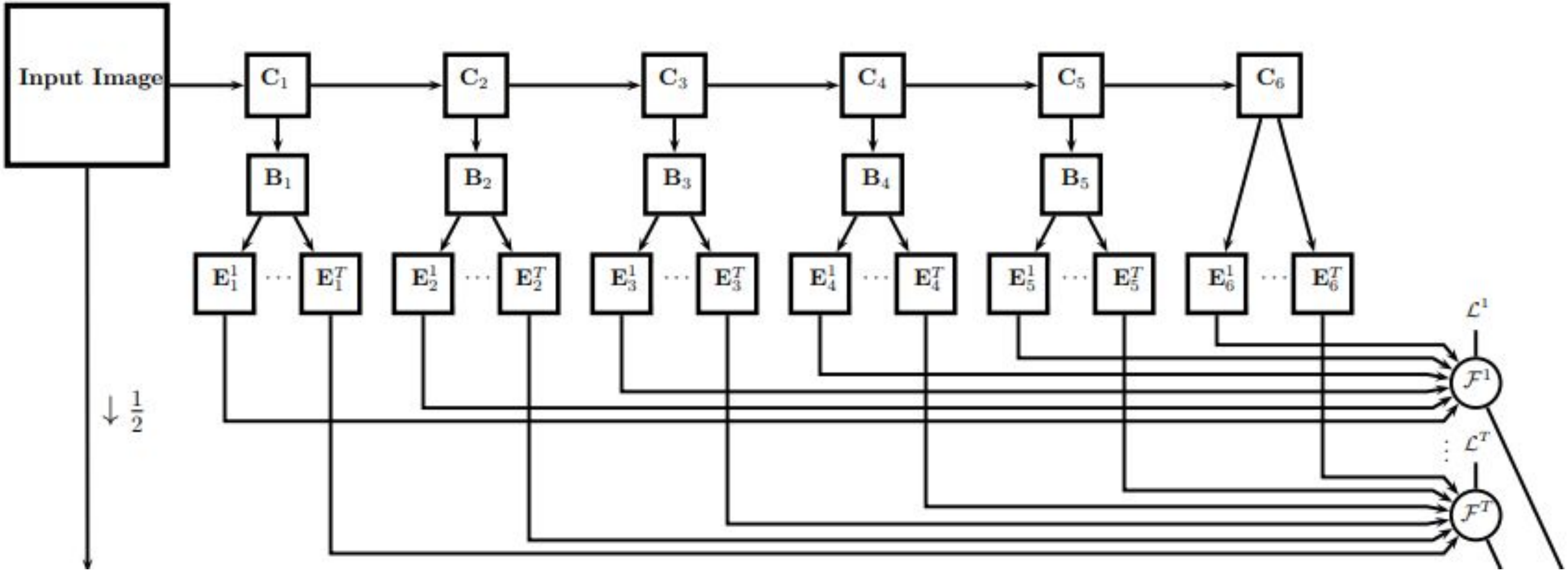
Architecture



Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

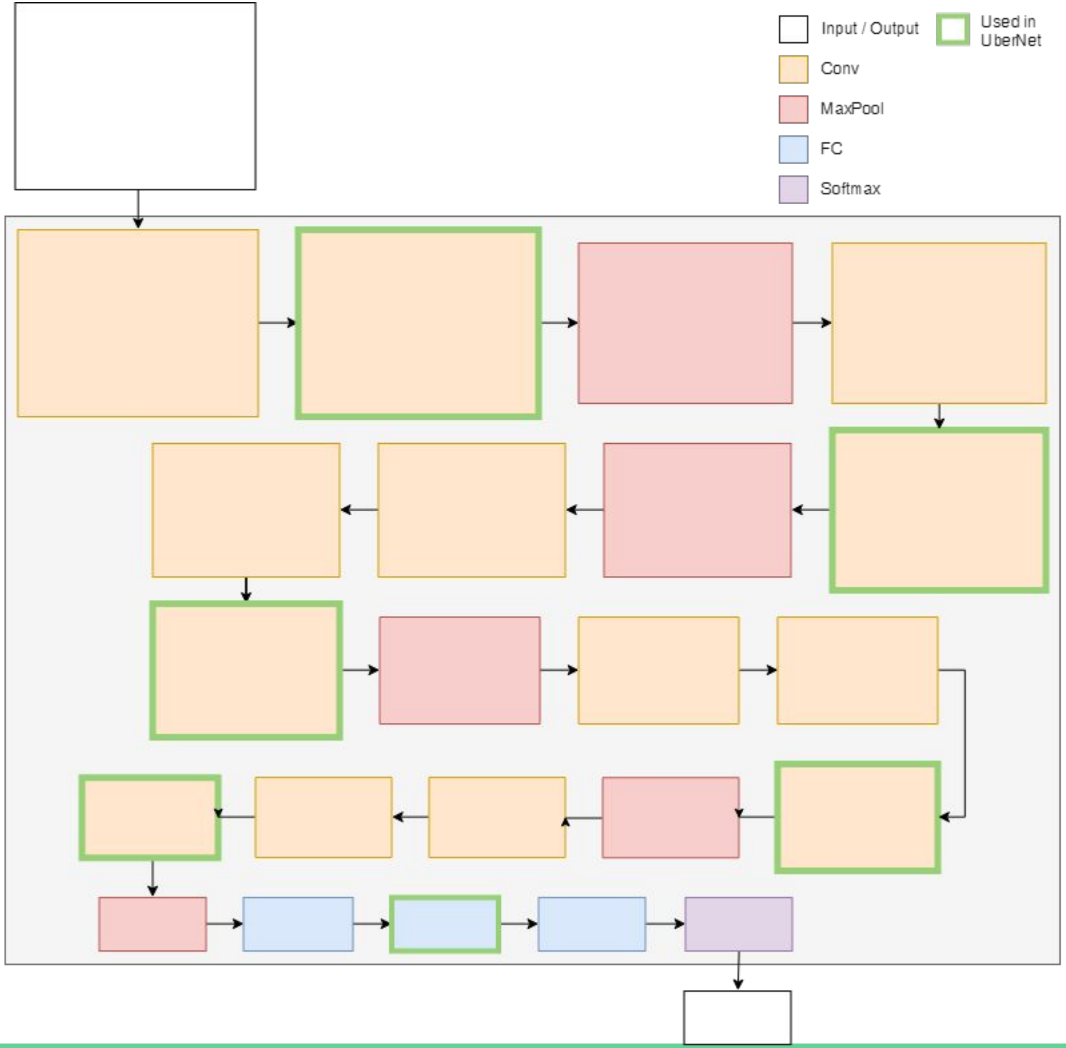


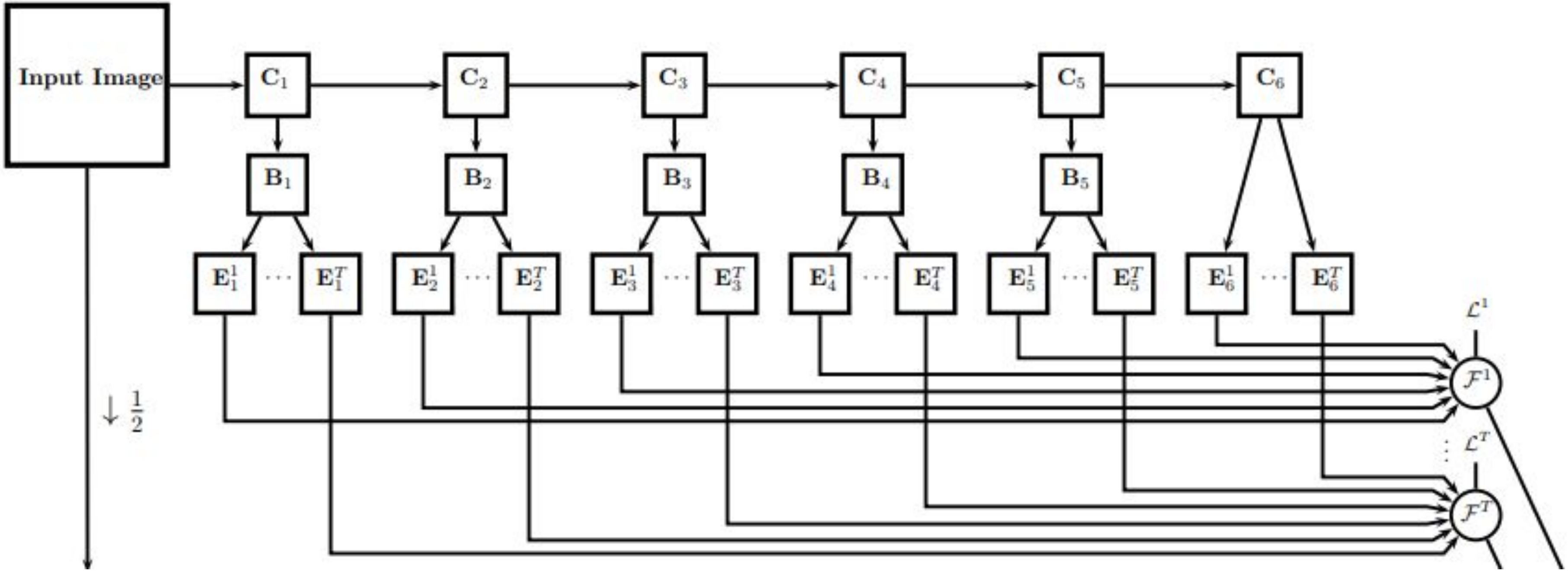
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132



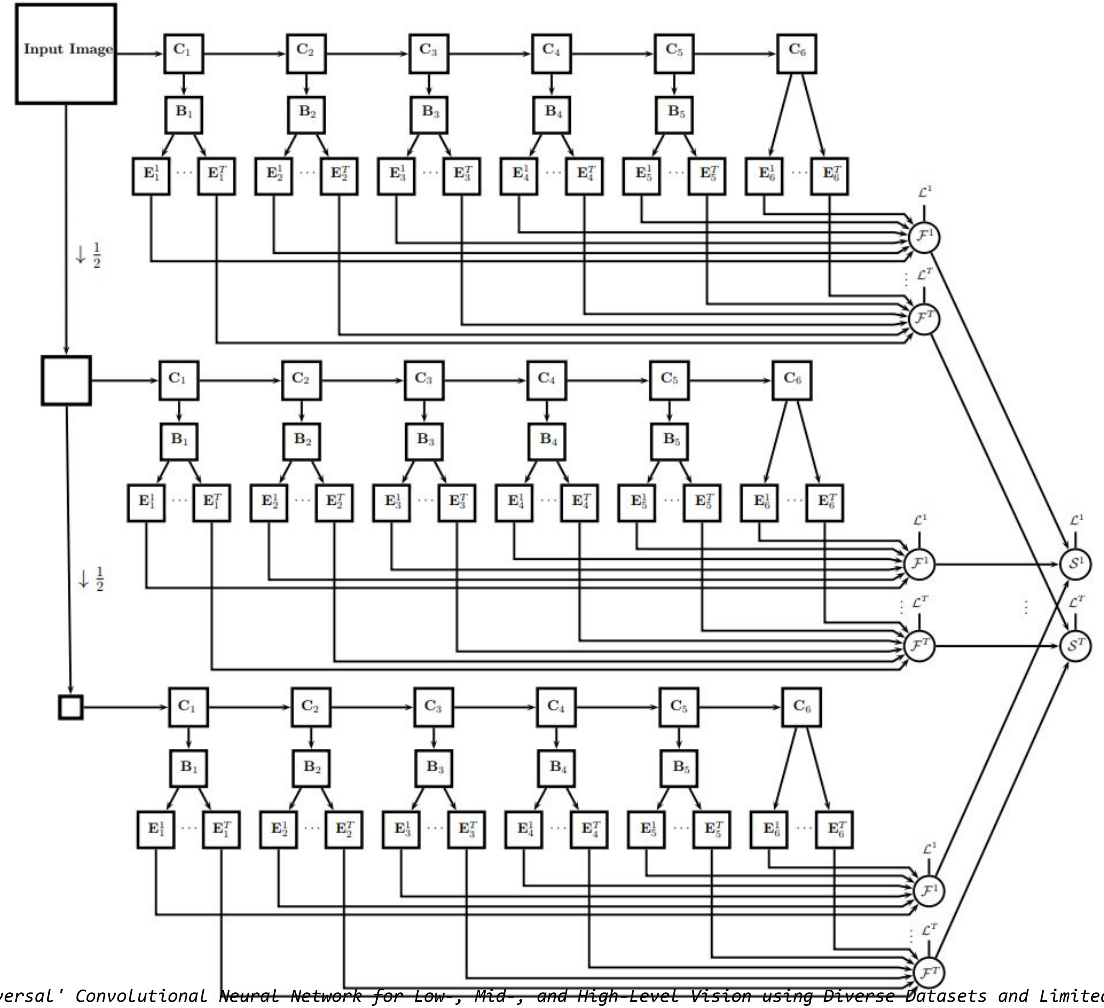
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

VGG-16

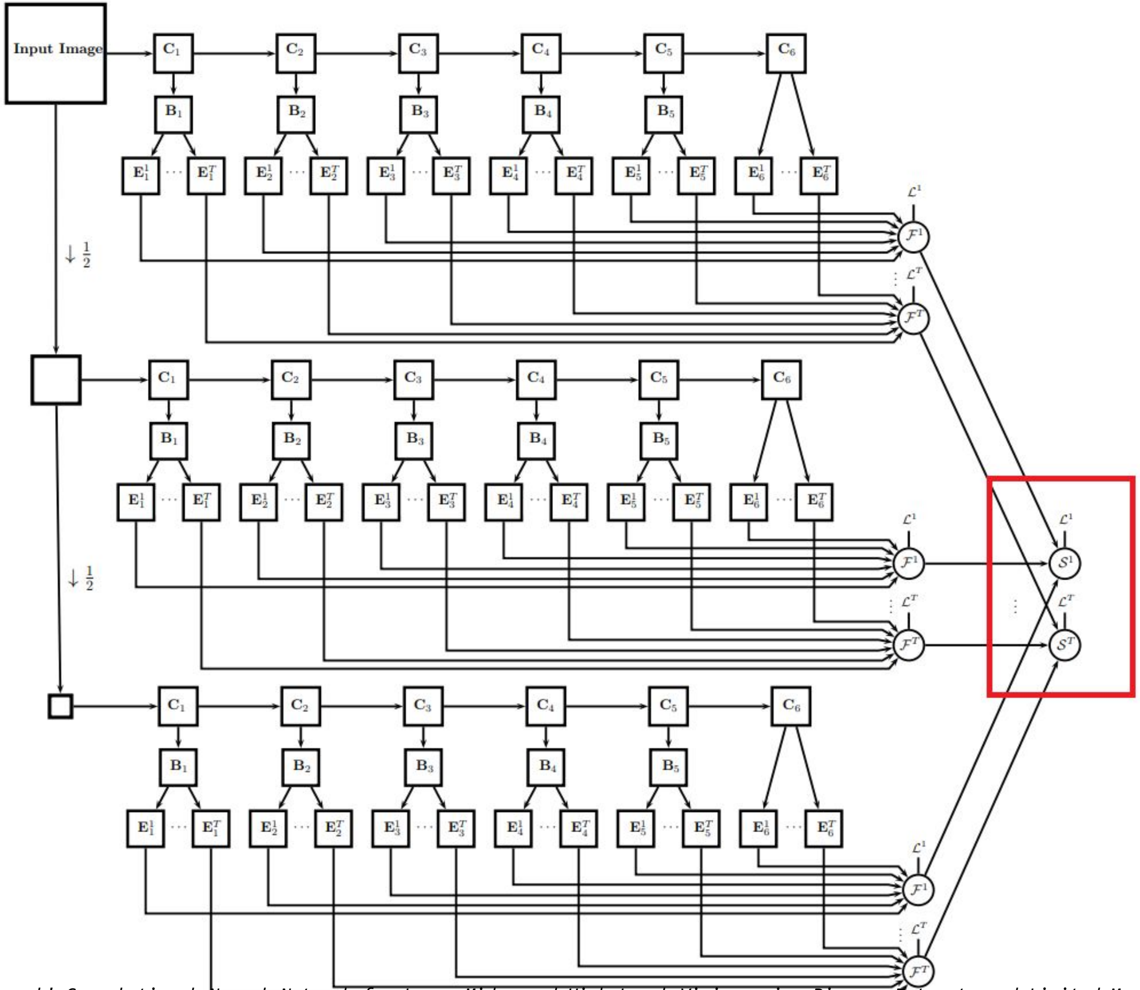




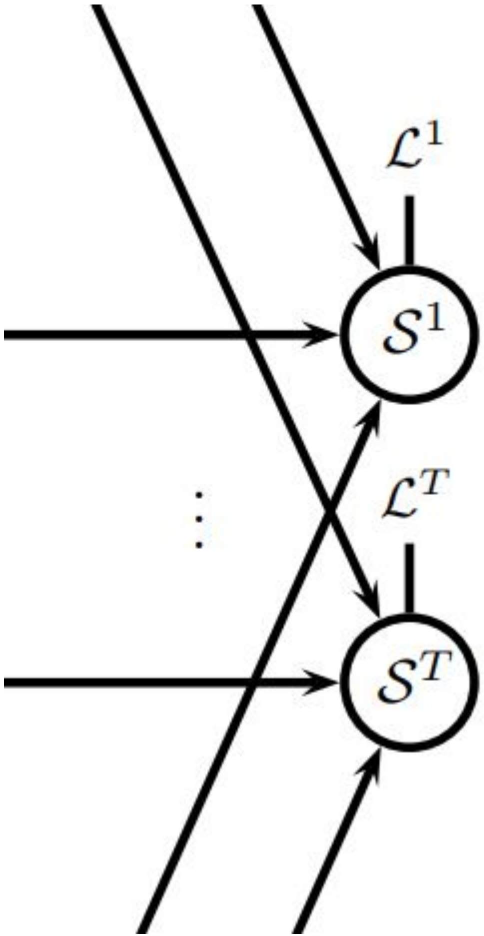
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132



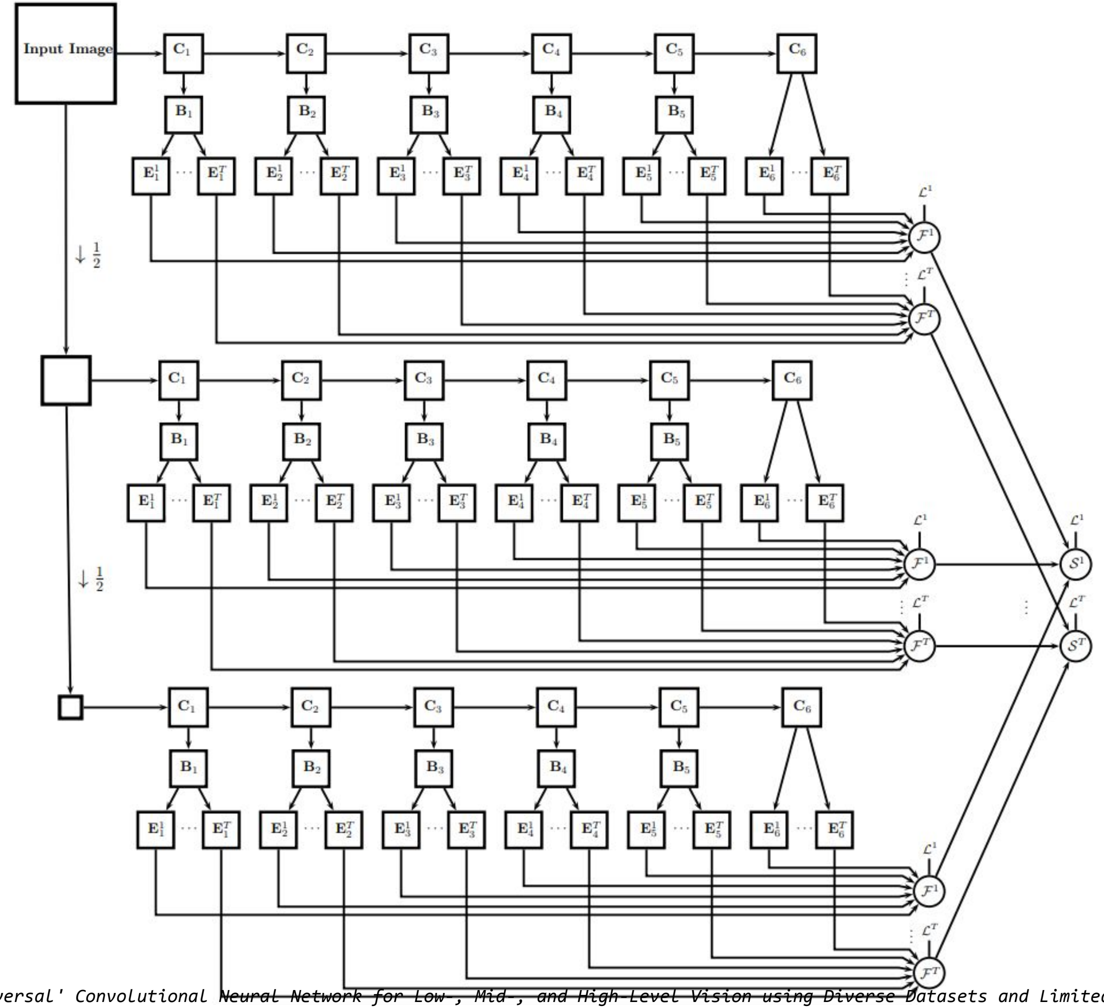
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132



Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

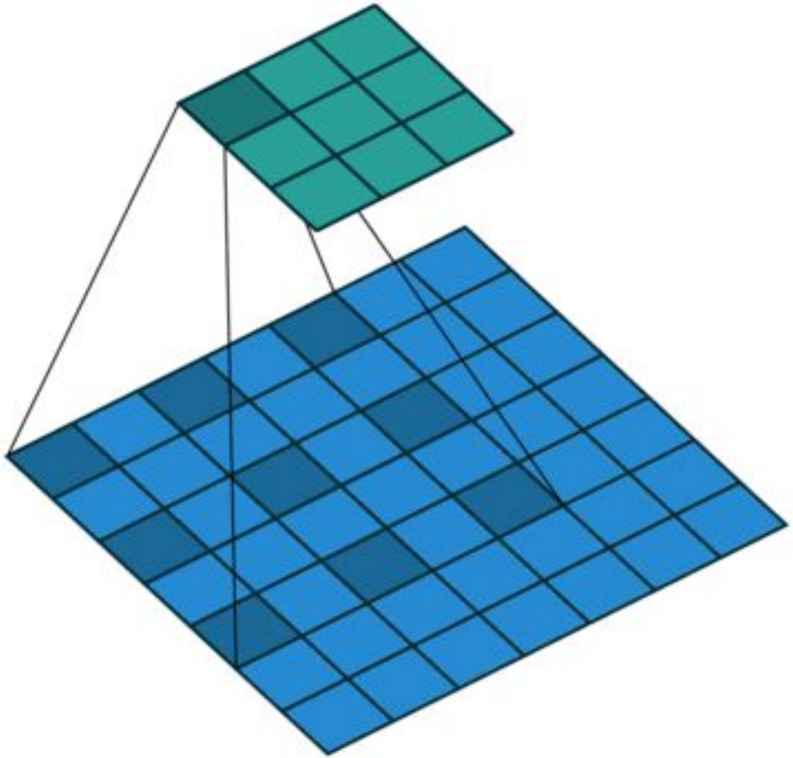


Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132



Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Atrous/Dilated Convolutions



Source: https://towardsdatascience.com/types-of-convolutions-in-deep-learning-717013397f4d

Architecture

- Skip layers
- Skip-layer normalization
- Cumulative task-specific operations
- Fusion layers
- Atrous convolution
- Multi-resolution CNN
- Task-specific deviations

Multi-Task Training

Objective: minimize the loss

$$\mathcal{L}(\mathbf{w}_{0,t_1,\ldots,t_T}) = \mathcal{R}(\mathbf{w}_0) + \sum_{t=1}^{T} \gamma_{t_k} \left(\mathcal{R}(\mathbf{w}_t) + L_t\left(\mathbf{w}_0, \mathbf{w}_t\right)\right)$$

$$L_t\left(\mathbf{w}_0, \mathbf{w}_t\right) = \frac{1}{N} \sum_{i=1}^{N} \delta_{t,i} L_t\left(\mathbf{f}_t^i(\mathbf{w}_0, \mathbf{w}_t), \mathbf{y}_t^i\right)$$

Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Many task = many different type annotations

Problem: Diverse datasets

There isn’t a dataset,

which contains annotations for all the problems

Divers datasets

	Imagenet [88]	VOC'07 [26]	VOC'10 [26]	VOC'12 [26]	MS-COCO [62]	NYU [74]	MSRA10K [18]	BSD [69]
Detection	Partial	Yes	Yes	Yes	Yes	No	No	No
Semantic Segmentation	No	Partial	[36, 71]	Partial	Yes	Yes*	No	No
Instance Segmentation	No	Partial	[36, 71]	Partial	Yes	No	No	No
Human parts	No	No	[17]	No	No	No	No	No
Human landmarks	No	No	[10]	No	Yes	No	No	No
Surface Normals	No	No	No	No	No	Yes	No	No
Saliency	No	No	No	No	No	No	Yes	No
Boundaries	No	No	[71]	No	No	No	No	Yes
Symmetry	No	No	No	No	Partial, [91]	No	No	[97]

Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Divers datasets

	VOC'07 trainval	VOC'12 train	VOC'12 val	NYU	MSRA10K	BSD
	5011	5717	5823	23024	10000	5100
Detection	5011	5717	5823	0	0	0
S. Segmentation	422	4998	5105	0	0	0
S. Boundaries	0	4998	5105	0	0	0
Human Parts	0	1716	1817	0	0	0
Normals	0	0	0	23024	0	0
Saliency	0	0	0	0	10000	0
Boundaries	0	4998	5105	0	0	5100

Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

PASCAL in Detail

	10100
Semantic Segmentation	10100
Human Parts	3548
Occlusion	10094
Boundaries	10100

Many task = many different annotation types

Problem: Diverse datasets

There isn't a dataset,

which contains annotations for all the problems

Synchronous SGD

for $m = 1$ to M **do**

$\{$construct minibatch$\}$

$\mathcal{B} \leftarrow \{i_1, \ldots, i_B\}$ with $i_i \sim U[1, N]$

$\{$initialize gradient accumulators$\}$

$\mathbf{dw}_0 \leftarrow 0, \mathbf{dw}_1 \leftarrow 0, \ldots, \mathbf{dw}_T \leftarrow 0$

for $i \in \mathcal{B}$ **do**

$\{$cnn gradients$\}$

$\mathbf{dw}_0 \leftarrow \mathbf{dw}_0 + \sum_t \delta_{t,i} \gamma_t \nabla_{\mathbf{w}_0} L_t \left(\mathbf{f}_t^i(\mathbf{w}_0, \mathbf{w}_t), \mathbf{y}_t^i \right)$

$\{$Task gradients, $t = 1, \ldots, T\}$

$\mathbf{dw}_t \leftarrow \mathbf{dw}_t + \delta_{t,i} \gamma_t \nabla_{\mathbf{w}_0} L_t \left(\mathbf{f}_t^i(\mathbf{w}_0, \mathbf{w}_t), \mathbf{y}_t^i \right)$

end for

for $p \in \{0, 1, \ldots, T\}$ **do**

$\mathbf{w}_p \leftarrow \mathbf{w}_p - \epsilon \left(\lambda \mathbf{w}_p + \frac{1}{B} \mathbf{dw}_p \right)$

end for

end for

Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Asynchronous SGD

{initialize gradient accumulators}

$\mathbf{dw}_0 \leftarrow 0, \mathbf{dw}_1 \leftarrow 0, \ldots, \mathbf{dw}_T \leftarrow 0$

{initialize counters}

$\mathbf{c}_0 \leftarrow 0, \mathbf{c}_1 \leftarrow 0, \ldots, \mathbf{c}_T \leftarrow 0$

for $m = 1$ to $M \cdot B$ **do**

  Sample $i \sim U[1, N]$

  {cnn gradients & counter: always updated}

  $\mathbf{c}_0 \leftarrow \mathbf{c}_0 + 1$

  $\mathbf{dw}_0 \leftarrow \mathbf{dw}_0 + \sum_t \delta_{t,i}\gamma_t \nabla_{\mathbf{w}_0} L_t\left(\mathbf{f}_t^i(\mathbf{w}_0, \mathbf{w}_t), \mathbf{y}_t^i\right)$

  for $t \in \{1, \ldots, T\}$ **do**

    if $\delta_{t,i} = 1$ **then**

      {update accumulator and counter for task t if the current sample is relevant}

      $\mathbf{c}_t \leftarrow \mathbf{c}_t + 1$

      $\mathbf{dw}_t \leftarrow \mathbf{dw}_t + \gamma_t \nabla_{\mathbf{w}_t} L_t\left(\mathbf{f}_t^i(\mathbf{w}_0, \mathbf{w}_t), \mathbf{y}_t^i\right)$

    end if

  end for

  for $p \in \{0, 1, \ldots, T\}$ **do**

    if $\mathbf{c}_p = B_p$ **then**

      {update parameters if we have seen enough}

      $\mathbf{w}_p \leftarrow \mathbf{w}_p - \epsilon\left(\lambda \mathbf{w}_p + \frac{1}{B_p}\mathbf{dw}_p\right)$

      $\mathbf{c}_p \leftarrow 0, \mathbf{dw}_p \leftarrow 0$

    end if

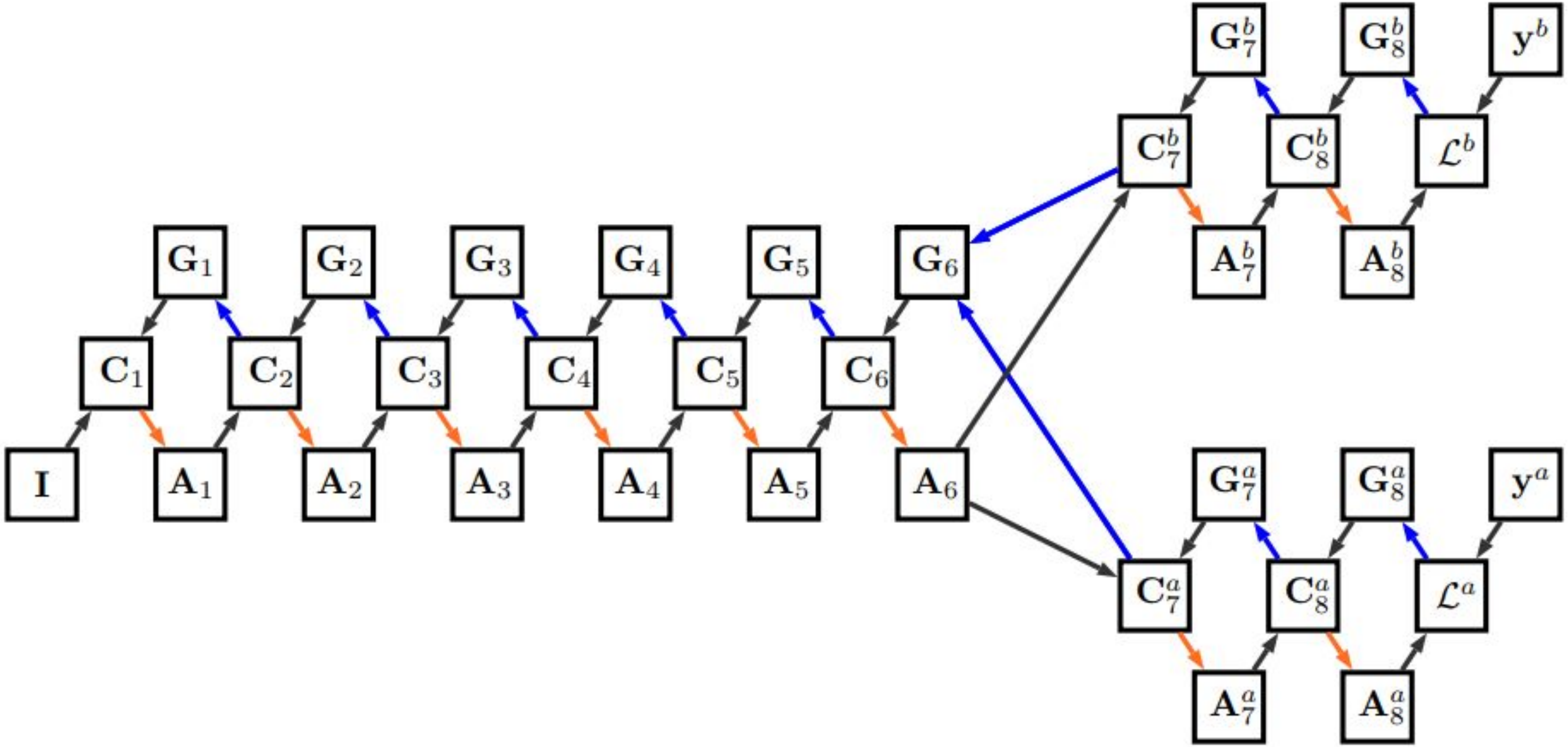
  end for

end for

Source: UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory https://arxiv.org/abs/1609.02132

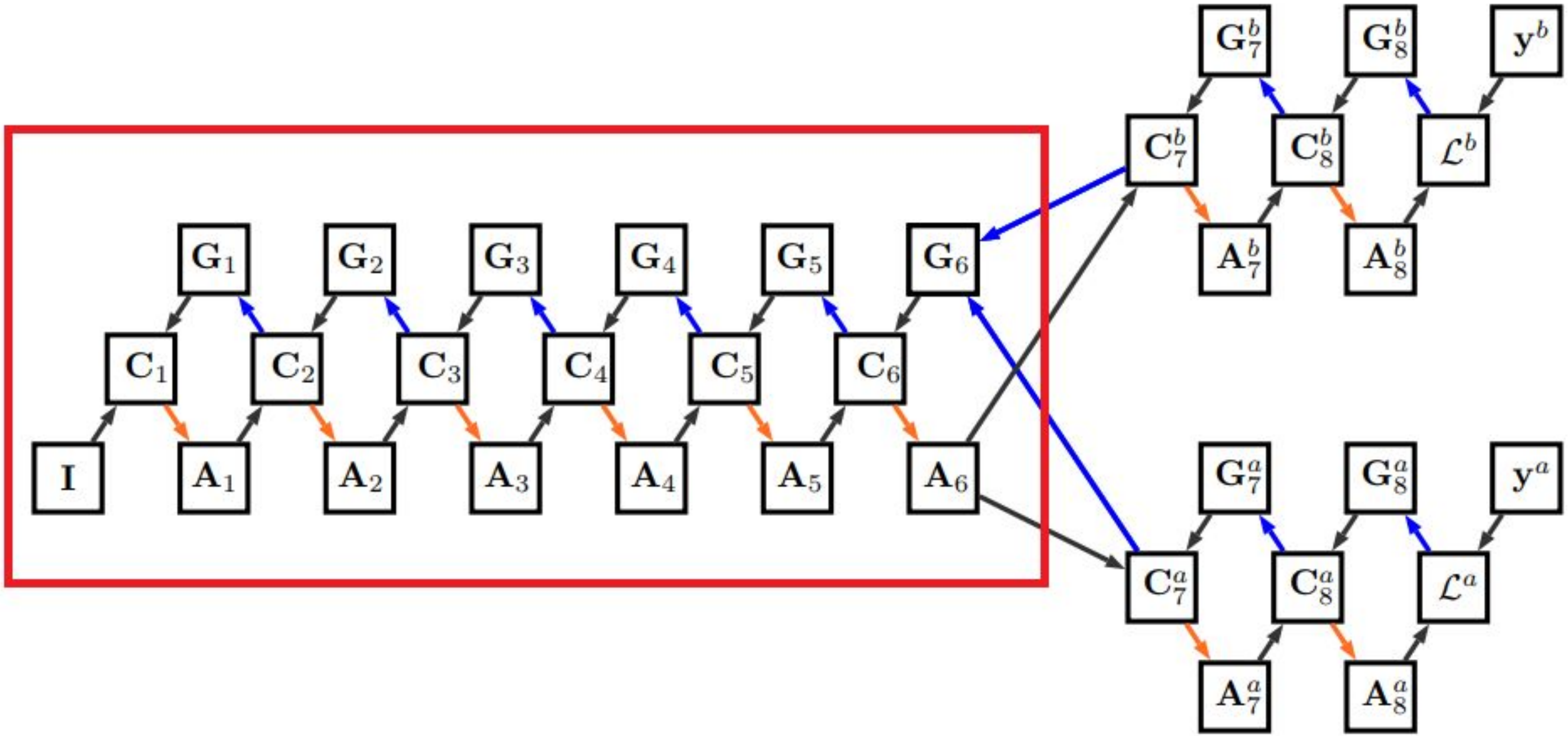
Low memory usage

Normal memory forward and backward pass



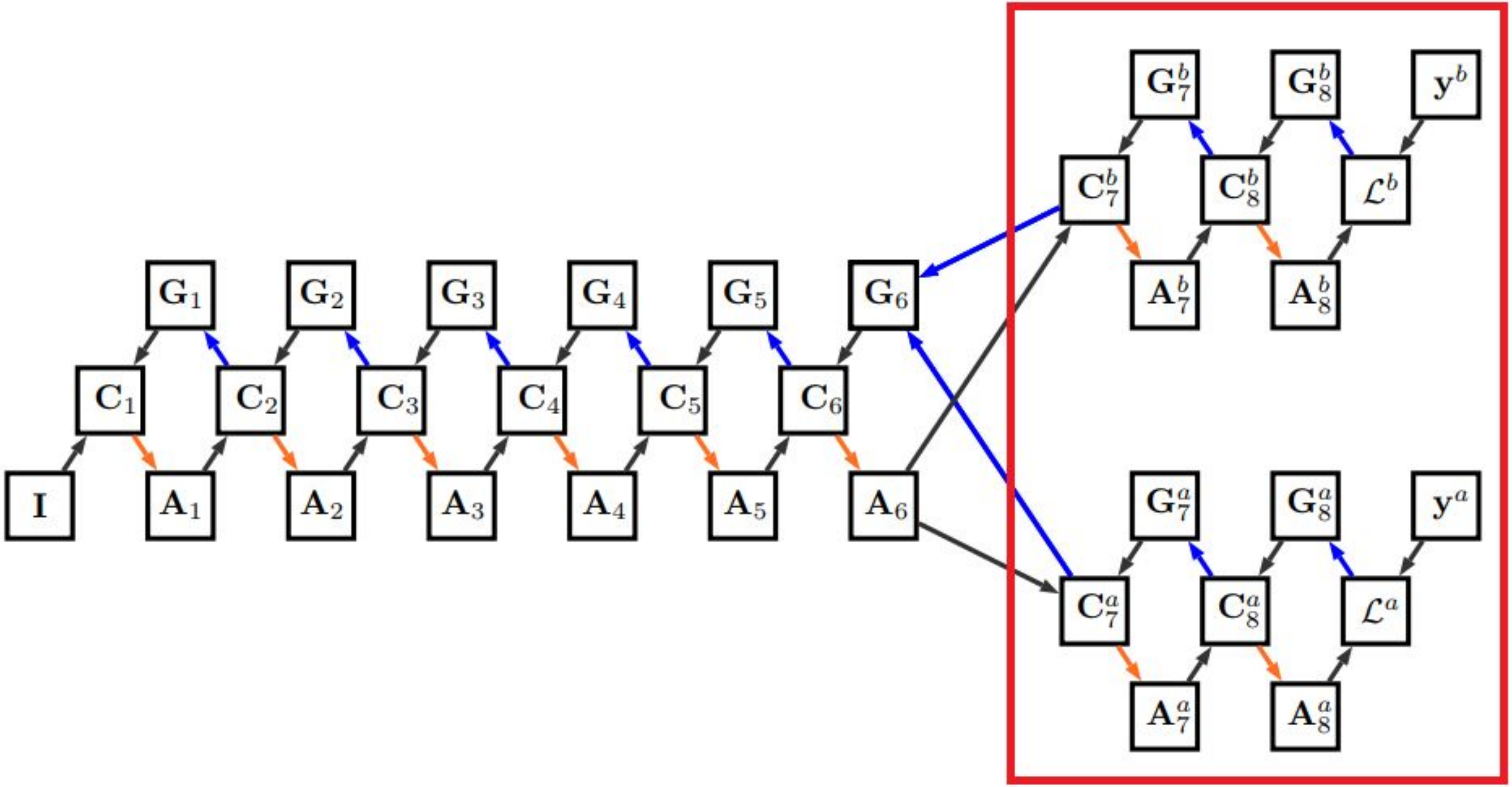
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Normal memory forward and backward pass



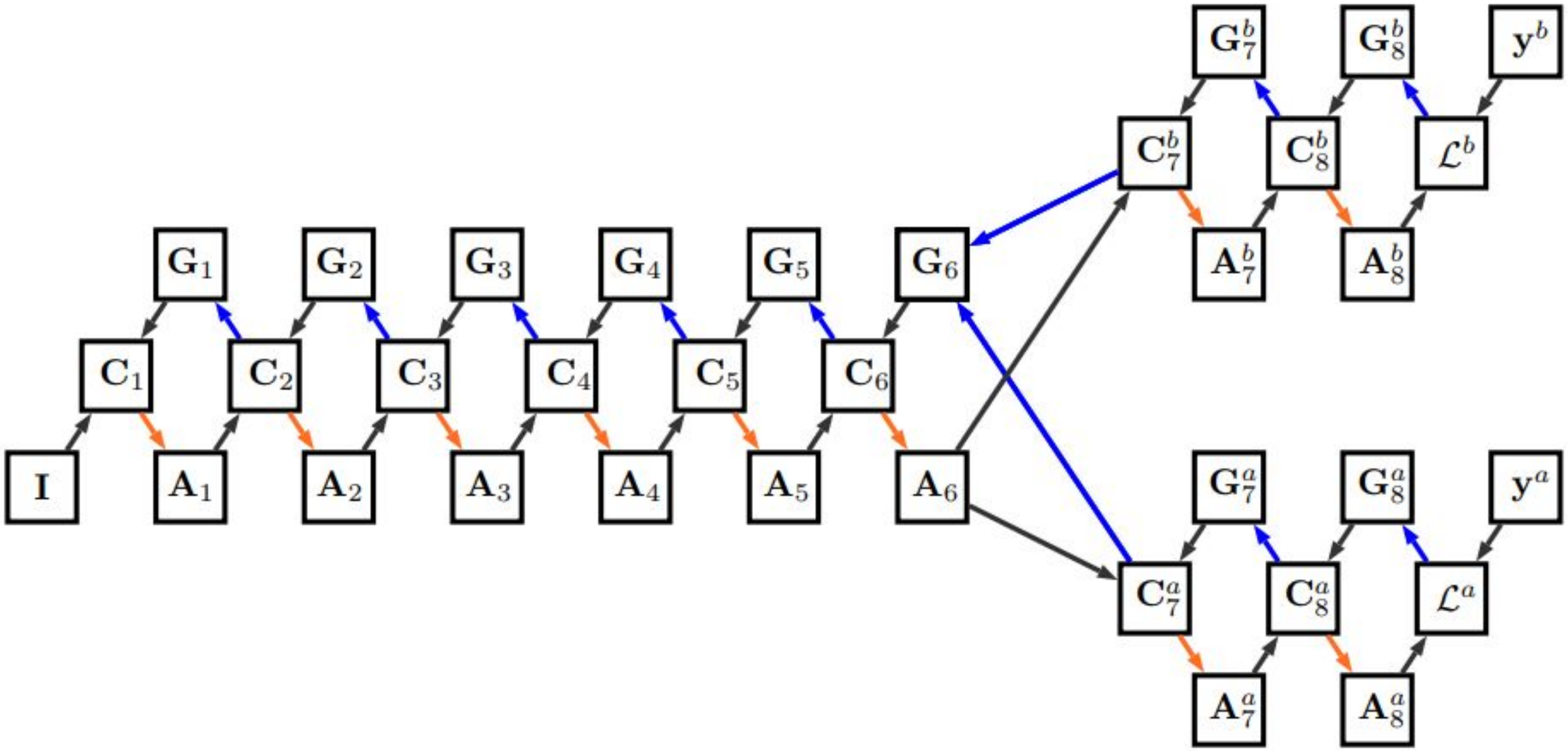
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Normal memory forward and backward pass



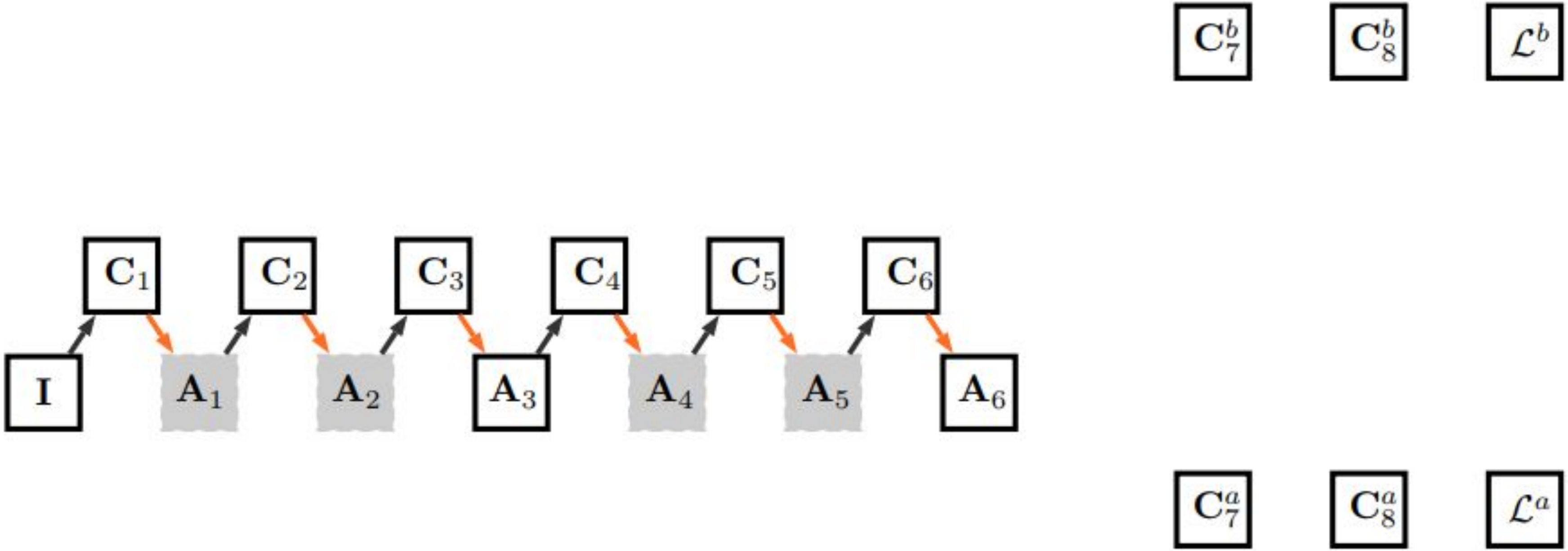
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Normal memory forward and backward pass



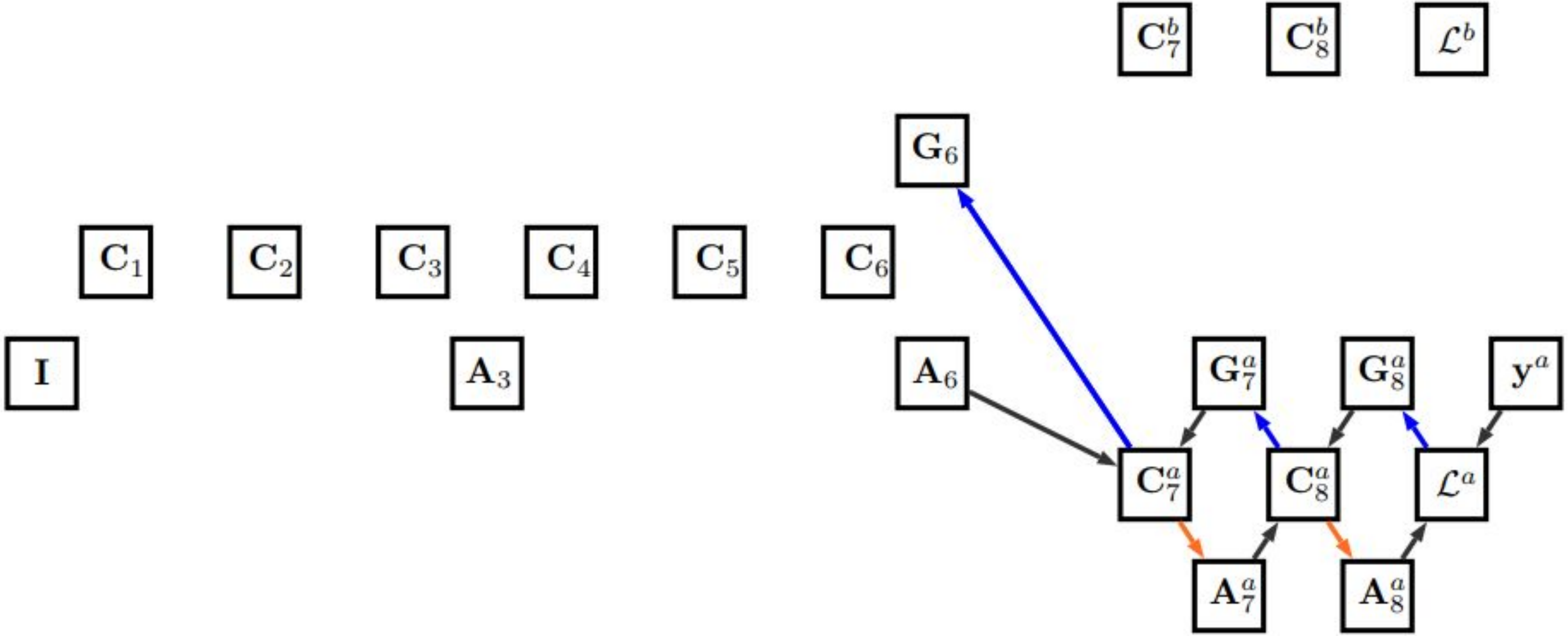
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Low memory forward and backward pass



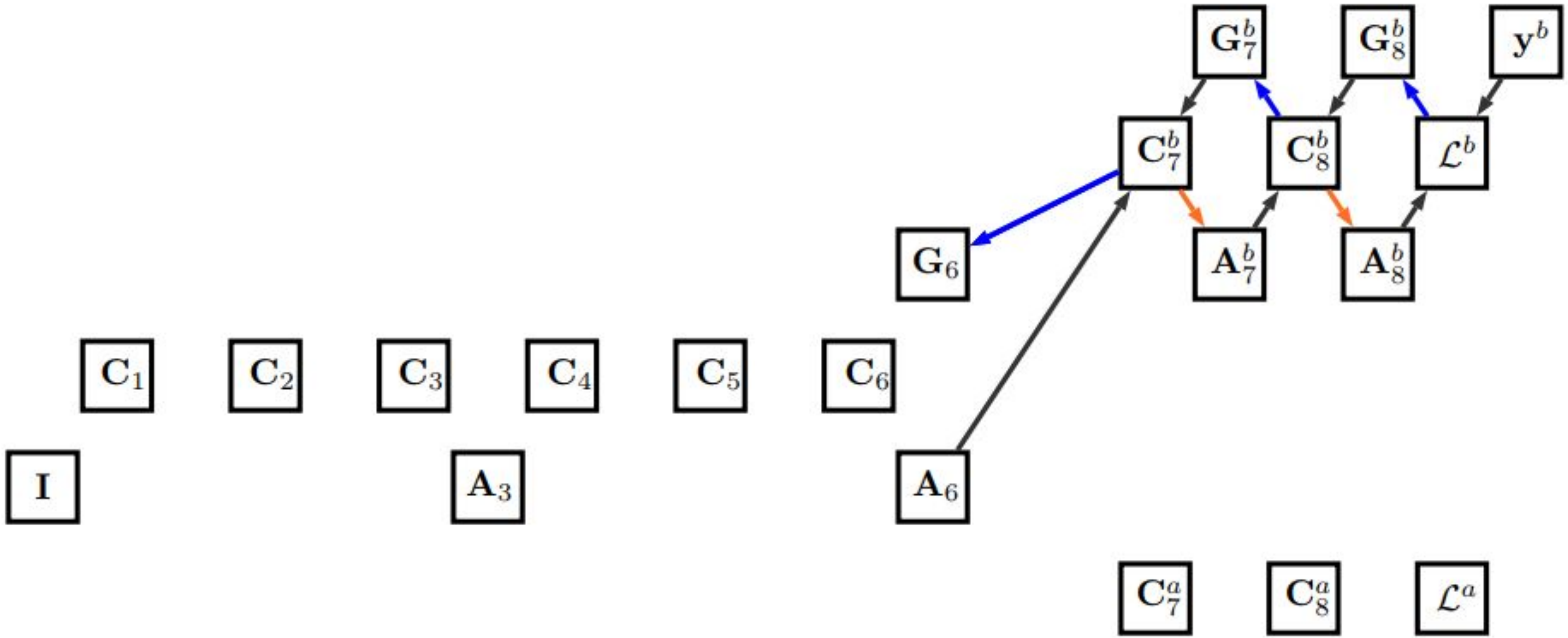
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Low memory forward and backward pass



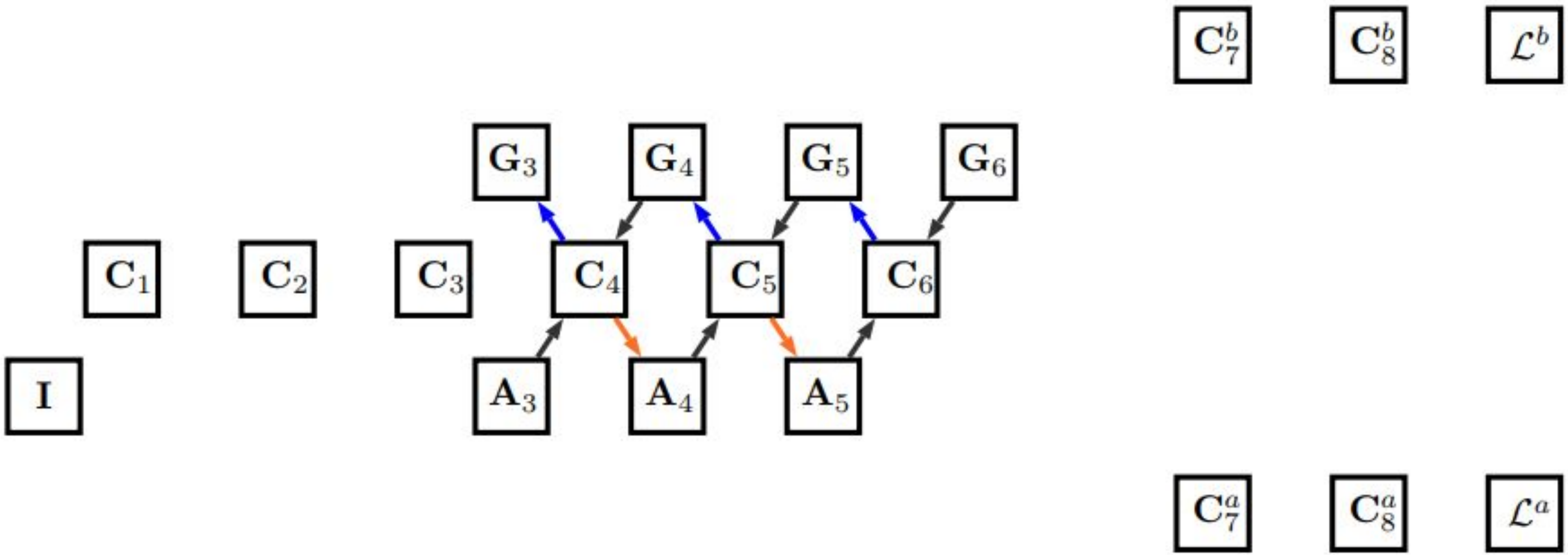
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Low memory forward and backward pass



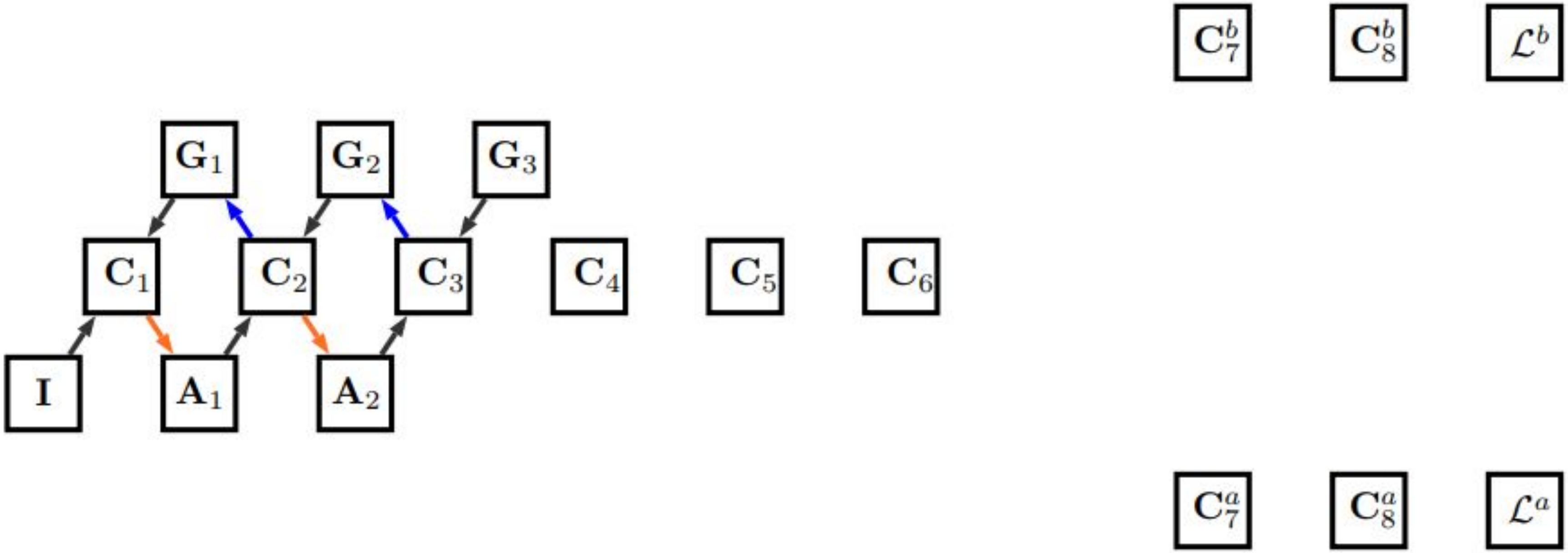
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Low memory forward and backward pass



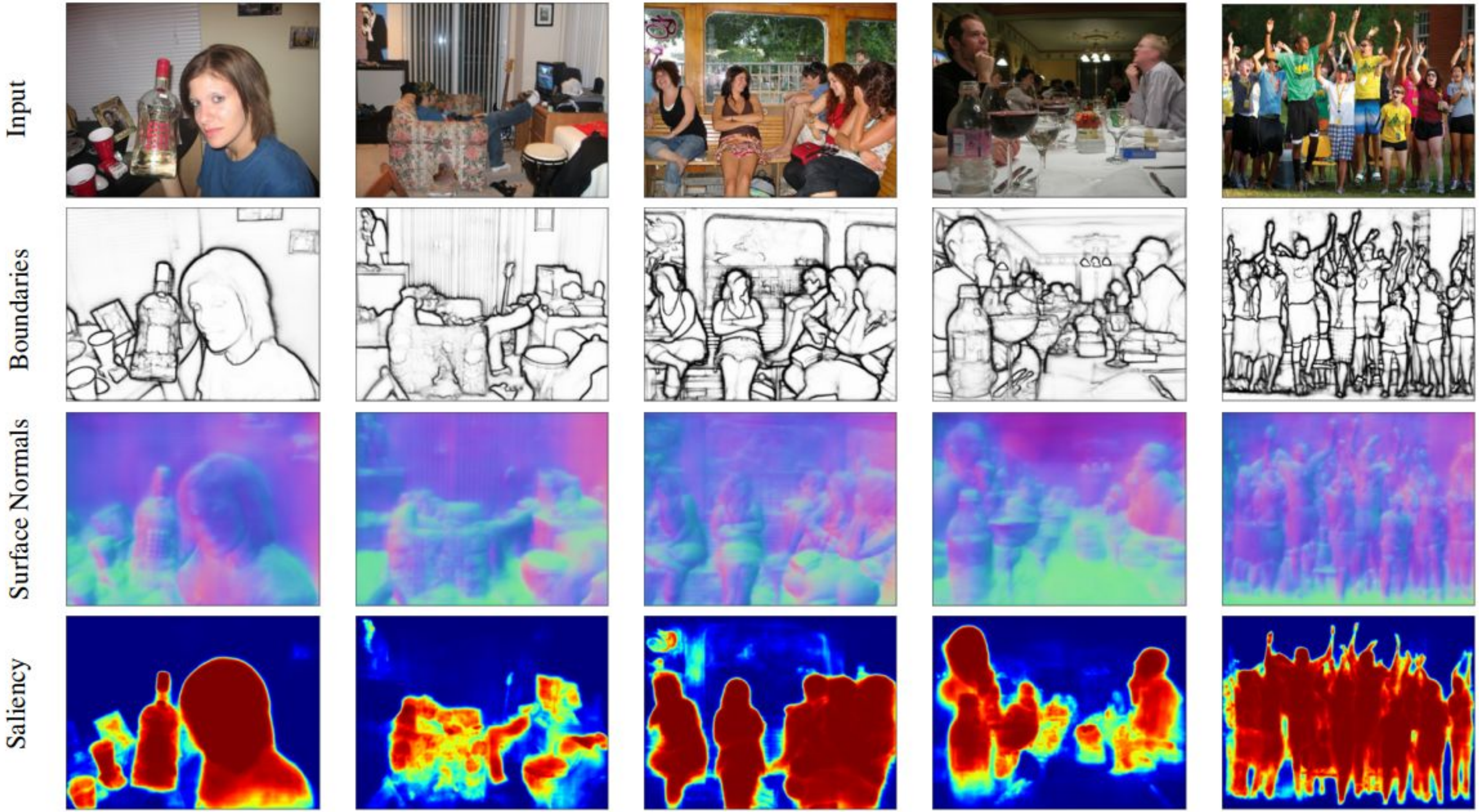
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Low memory forward and backward pass



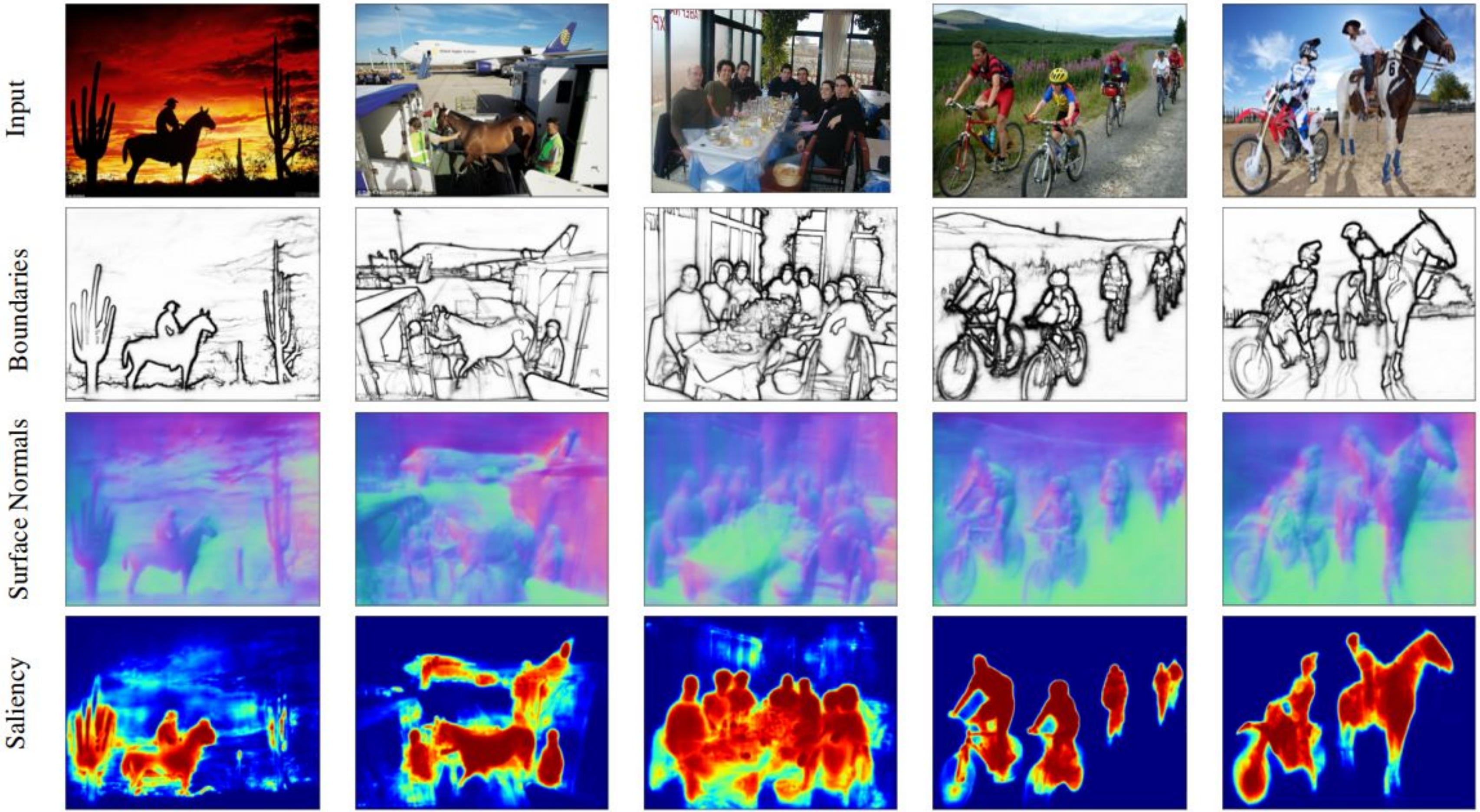
Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Results

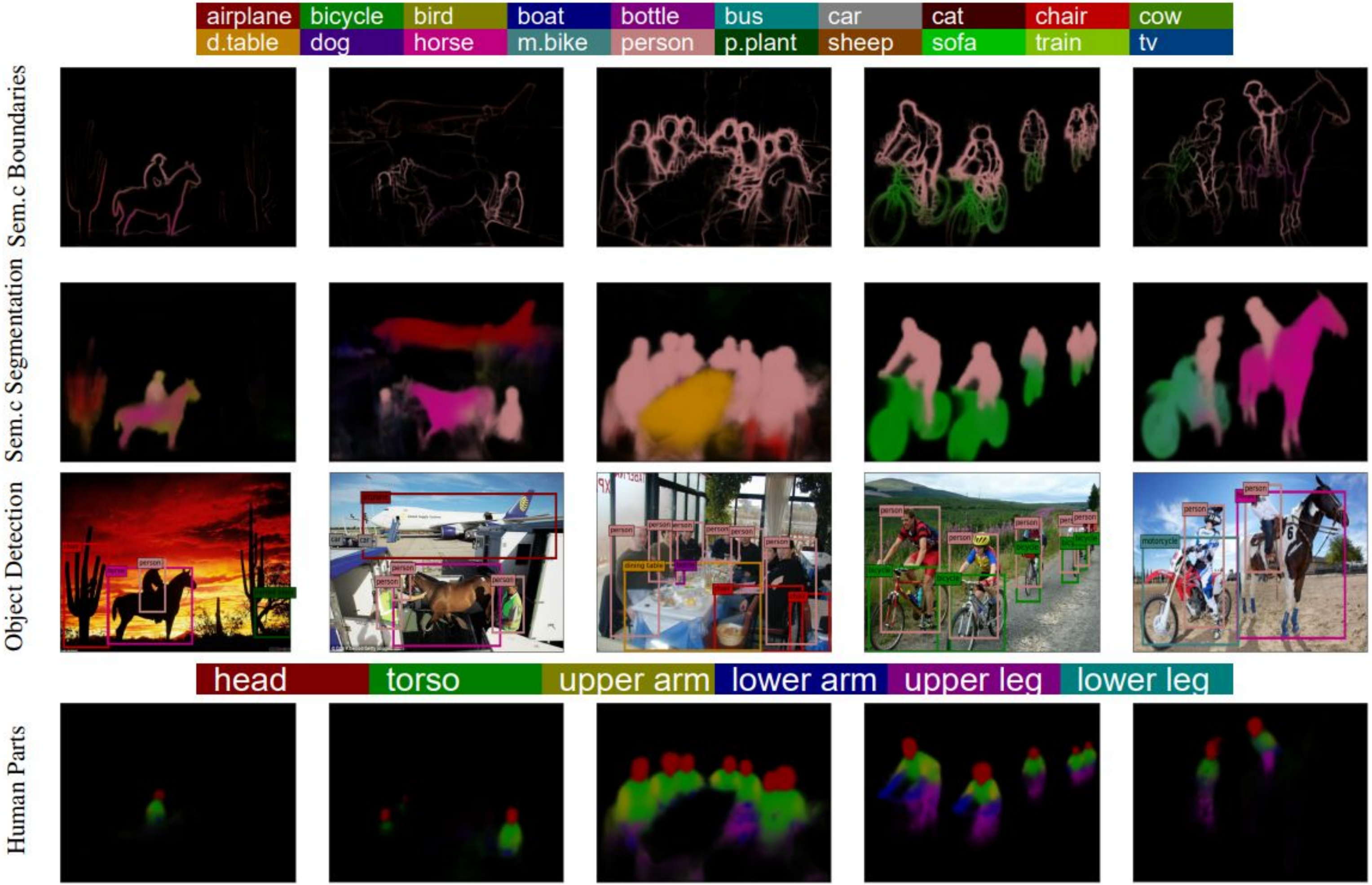


Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132



Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132



Source: *UberNet: Training a `Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory* https://arxiv.org/abs/1609.02132

Bibliography

- UberNet: Training a ‘Universal’ Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory https://arxiv.org/abs/1609.02132
- Very Deep Convolutional Networks for Large-Scale Image Recognition https://arxiv.org/abs/1409.1556
- CVPR’17 PASCAL in Detail Challenge https://sites.google.com/view/pasd
- Wikipedia:
  - Saliency map https://en.wikipedia.org/wiki/Normal_(geometry)
  - Normal (geometry) https://en.wikipedia.org/wiki/Saliency_map
- An Introduction to different Types of Convolutions in Deep Learning https://towardsdatascience.com/types-of-convolutions-in-deep-learning-717013397f4d

Thank you!