Department of Computer Science & Engineering, Kyung Hee University, Republic of Korea

Ph.D. Dissertation Presentation

Towards Image Semantic Segmentation and Classification using Bracket-style Convolutional Neural Network and Its Variants

CAM-HAO HUA

Advisors:

Prof. SUNG-HO BAE

Prof. SUNGYOUNG LEE

October 29, 2021

Table of Contents

- Introduction
  - Definitions
  - Motivation
  - Background
  - Problem Statement
  - Taxonomy
- Related Works
  - Highlights & Limitations
  - Unsolved issues & Proposed solution
- Proposed Approach
  - To-be vs. As-is
  - Overview of Themes
  - Theme #1-1 – Architecture Overview
  - Theme #1-1 – Detailed Architecture
  - Theme #1-2 – Architectural Variant
  - Theme #2 – Architectural Variant
  - Theme #2 – Architecture Overview
  - Theme #2 – Detailed Architecture
- Experiments
  - Benchmark Datasets
  - Evaluation Metrics
  - Training Configurations
  - Ablation Study
  - Comparison Results
- Conclusions & Future Work
- Publications
- References

Definitions: Image Classification & Semantic Segmentation

Image Classification is:

- Problem of **understanding image-level content**, i.e., whole image is assigned an object class.
- Applied into images with **single-object contents**

Tiger cat

Pizza

Sports Car

Figure: Top: Input images. Bottom: Desired **classification labels***

Image Semantic Segmentation is:

- Problem of **understanding an image at pixel level**, i.e., each pixel is assigned an object class.
- Applied into images with **more complicated (single- & multi-object) contents**

Figure: Top: Input images. Bottom: Desired **segmentation outputs****

Source: * ImageNet [76]; ** PASCAL VOC 2012 dataset [20]

Motivation (1/2)

- Emergence of Deep Learning in computer vision with breakthrough performance on complex & multi-modal data.
  - Frontiers of Deep Learning are being expanded tremendously.

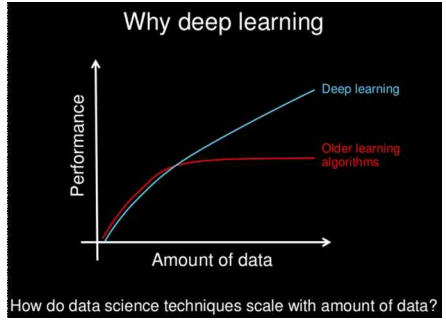


Figure: Power of deep learning along with increasing amount of data. (referred from Andrew Ng, Baidu research)

- Image Semantic Segmentation plays a fundamental role in many modern applications: *autonomous driving, computational photography, …*
  - Breakthrough of image semantic segmentation-related applications using Deep Learning

Figure: Typical applications of semantic segmentation: (left) autonomous driving*, (right) computational photography**.

Source: * CAMVID dataset; ** Copyright © 2016 NVIDIA

Motivation (2/2)

- Applications of Deep Learning into Classification and/or Semantic Segmentation tasks in medical image-based diagnostic systems: *Retinal blood vessel segmentation, Diabetic Retinopathy (DR) severity grading, …*
  - Breakthrough of medical image-based diagnosis systems using Deep Learning for healthcare

Figure: Illustrations of: (left) Retinal blood vessel segmentation*, (right) Diabetic Retinopathy severity grading**.

- Extension of Deep Learning into diverse applications related to human-computer interactions: *Facial Expression Recognition, …*
  - Breakthrough of image classification using Deep Learning for human-computer interactions applications

Figure: Examples of different facial emotions considered for human-computer interactions***

Source: * DRIVE dataset; ** Kaggle DR Detection dataset; *** RAF-DB dataset

Background: Convolutional Neural Network – Abstract Concept

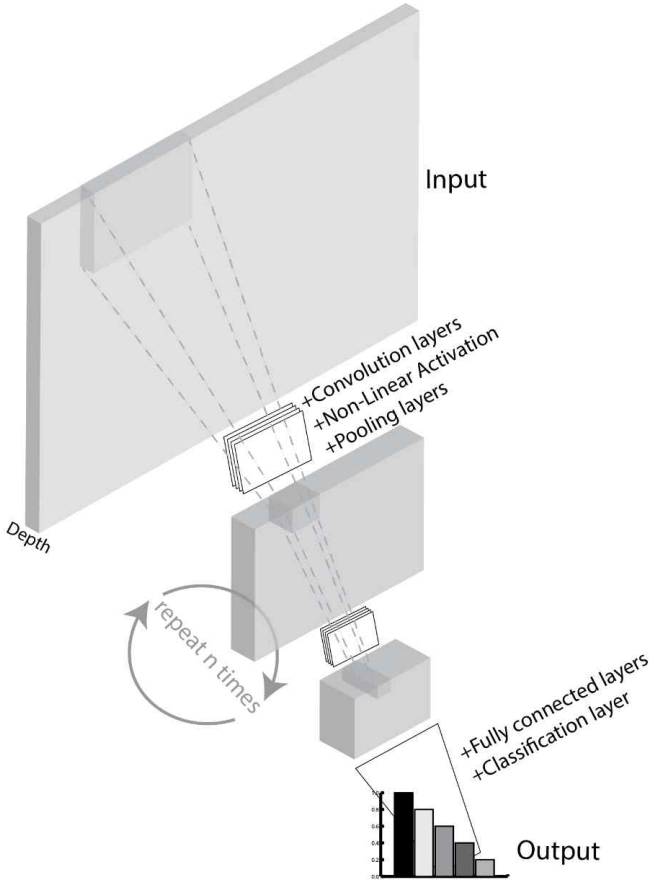


Figure: Compact concept of a CNN for image classification

Convolutional Neural Network (CNN), a typical Deep Learning model widely-used in Computer vision:

- Built by sequential stacks of multiple layers (i.e., convolution, pooling, fully connected, …).
- Every layer extracts representational features from the received input, wherein
  - Shallow (early) layers → Low-level (local details) features
  - Deep (later) layers → High-level (global context) features
- The vanilla version is firstly applied to **Classification** task **(**i.e., **labeling the image's contents)**.

Background: Image Classification vs. Semantic Segmentation

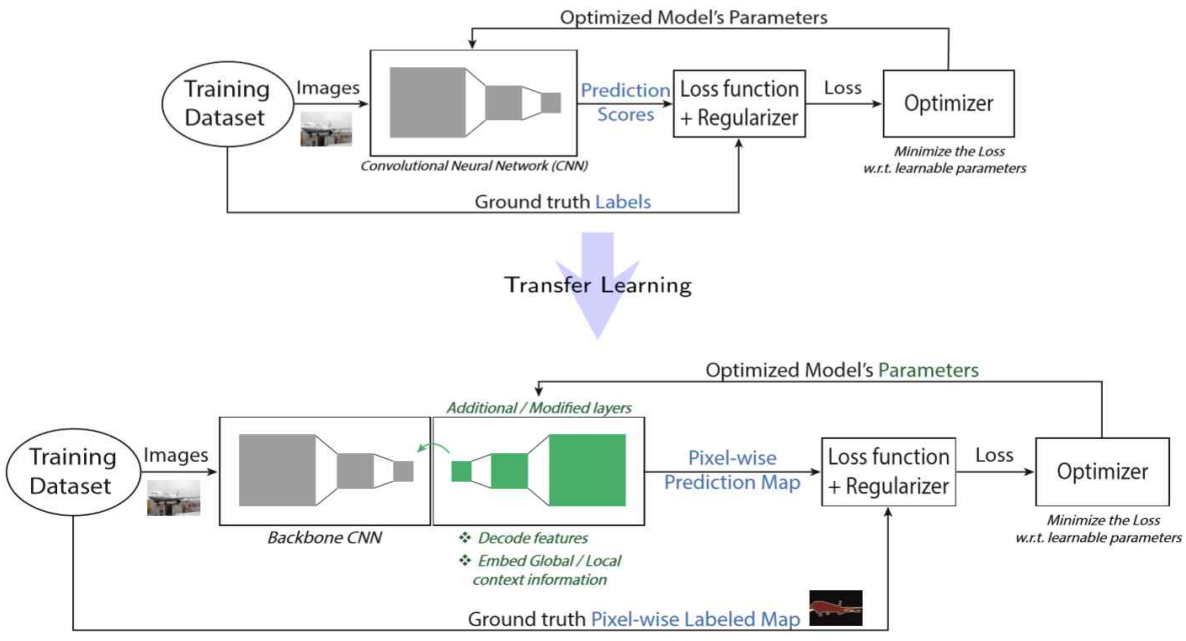


Figure: Abstract workflow of training a CNN-based **classification (top)** and **semantic segmentation (bottom)** model

Major Components for training a deep learning (e.g., CNN) model:

- Training Dataset
- CNN Model
- Loss Function and/or Regularizer
- Optimizer

Transfer Learning from a Classification to Semantic Segmentation model:

- Additional / Modified layers are attached to the Backbone CNN for
  - Decoding extracted features
  - Embedding global into local contextual information

Problem Statement (1/2)

Figure: Demonstrations of how fine-to-coarse features are extracted from a vanilla CNN

Observations from Feature Encoding Phase	Feature Decoding Phase (with additional/modified layers)
Spatial dimension of extracted feature maps **decreases gradually.** (for *abstract feature representation + saving computational cost*) → **Loss of spatial information of objects.**	***Decoding*** scheme is ***needed*** to ***reconstruct spatial representations of the extracted multi-scale features***
Low-level feature maps: → Generated by layers seeing inputs in **narrowed field of view**. → *Capture **locally structural (finely-patterned) details**.*	Effective Decoding
High-level feature maps: → Generated by layers seeing inputs in **large field of view.** → *Capture **global contextual (semantically-rich) information**.*	***Semantically-rich*** **features** and ***Finely-patterned*** **features** are **balancedly combine** for **meaningful fine-grained representations** in Image Semantic Segmentation and/or Classification

Problem Statement (2/2)

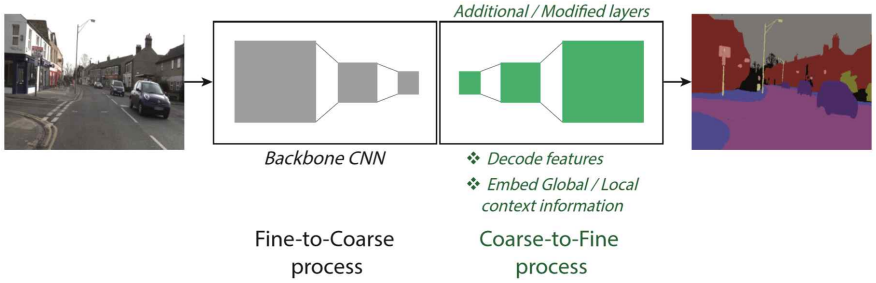


Figure: CNN-based **semantic segmentation architecture**

Architectural Variants

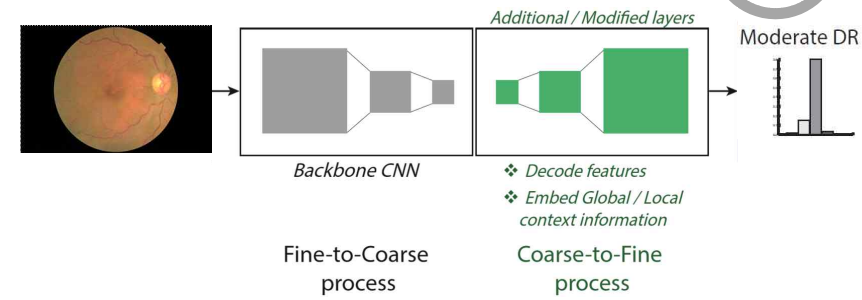


Figure: CNN-based **classification variants** (in specialized domains like Diabetic Retinopathy (DR) severity grading, facial expression recognition)

Problem Statement

How to design an optimal **decoding strategy** being able to balancedly combine local information (finely-patterned features) with global context (semantically-rich features) extracted from shallow-to-deep layers of the backbone CNN?

Goal

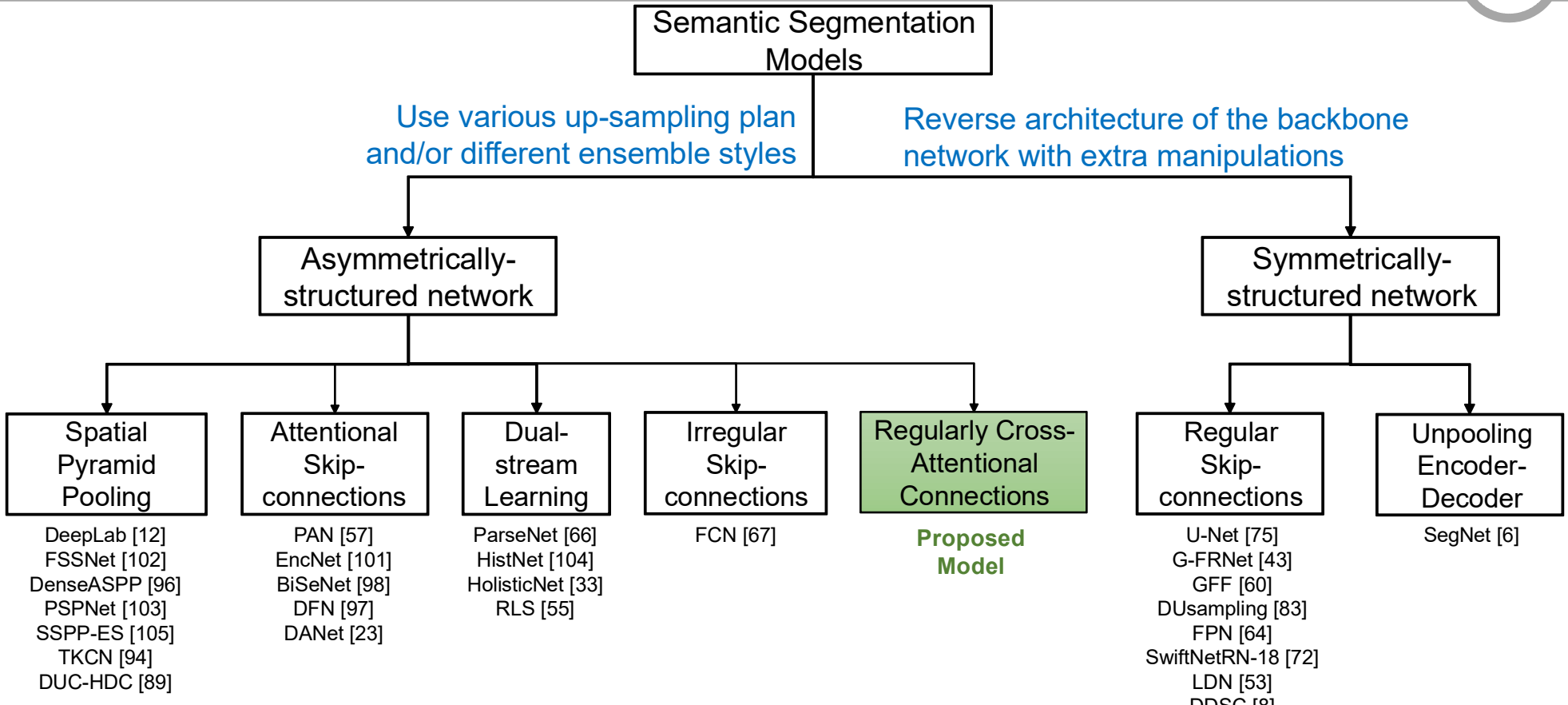
The proposed model can

- effectively generate **pixel-wise labeled map** for predefined **image semantic segmentation** tasks
- be flexibly extended as variants for **image-level recognition** for predefined **image classification** tasks (specialized domains where image labels are heuristically defined by the combination of various structural factors)

Challenges

1. Difficulties in smoothly **coordinating** representational features of **multiple scales** (fine-to-coarse resolutions)
2. Heavy ambiguities of finely-patterned details in **low-level features** for direct utilization in the final classifier

Taxonomy



[Hua, 2020b] Cam-Hao Hua, Thien Huynh-The, Sung-Ho Bae and Sungyoung Lee, "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences, Vol.539, pp.277-294, 2020

Related Work: Highlights & Limitations

Branch	Typical Approach	Highlights	Limitations
Asymmetrically-structured network			
Spatial pyramid pooling	Pyramid Scene Parsing Network (**PSPNet** [103])	• Utilize **dilated convolution** in backbone CNN. • Apply a **pyramid parsing module** to harvest different sub-region representations followed by upsampling and concatenation layers to form the final feature representation.	Local context inconsistency due to dilated convolution.
Attentional Skip-connections	Context Encoding Network (**EncNet** [101])	• Utilize **dilated convolution** in backbone CNN. • Proposed a **Context Encoding module** (i.e., **Channel-wise attention**) at the end for embedding semantic details back into the backbone CNN's features.	Involvement of channel-wise attention mechanism for high-level features only → Loss of attention in low-level counterparts
Dual-stream learning	Recalling Holistic Information for Semantic Segmentation (**HolisticNet** [33])	• Deploy **pixel network stream** for retrieving locally semantic information. • And **patch network stream** for collecting globally knowledge to eliminate noisy predictions.	Heavy cost in terms of learnable parameters and training resources.
Irregular Skip-connections	Fully Convolutional Networks for Semantic Segmentation (**FCN** [67])	• Change fully connected layers into conv. Layers in the backbone CNN. • Up-sample smaller-sized feature maps by implementing **learnable transpose convolutions.** • **Fuse** them for creating pixel-wise prediction map.	Does not take into account low-level feature maps. → Unbalanced aggregation between recovered local and global information.
Symmetrically-structured network			
Regular Skip-connections	U-net: Convolutional networks for biomedical image segmentation (**U-Net** [75])	• Consist of a **contracting path** to capture context and a **symmetric expanding path (+ shortcut with channel-wise concatenation)** that enables precise localization.	Heavy reliance on data preprocessing (augmentation).
Unpooling Encoder-Decoder	Segnet: A deep convolutional encoder-decoder architecture for image segmentation (**SegNet** [6])	• Build a structure with (**encoder** ~ backbone CNN) + (**decoder** ~ layer-wise reversed version of the encoder). • Use **max-pooling indices for up-sampling** to maintain high response features and merits fewer trainable parameters.	High risk of losing neighbor information in the final prediction map.

Related Work: Unsolved issues & proposed solution

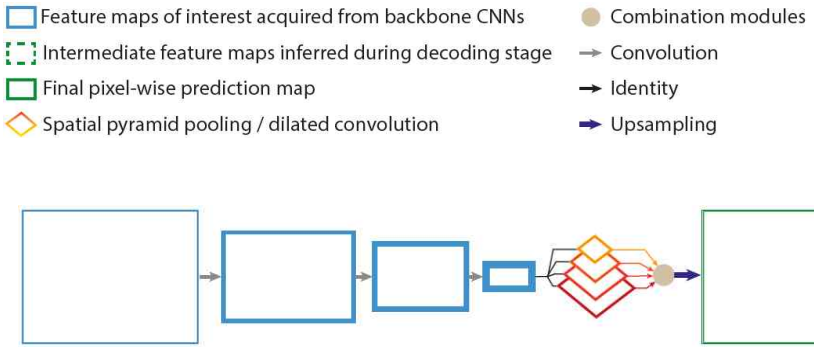


Figure: General **concept** of **asymmetrically-structured** network (spatial pyramid pooling [103])

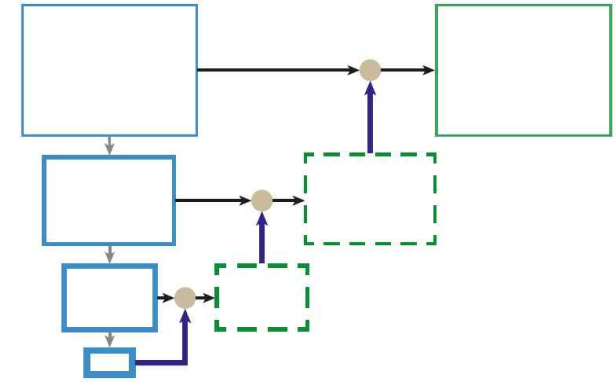


Figure: General **concept** of **symmetrically-structured** network [75, 83, 8, 43, 53, 60, 63, 57]

Only the coarsest feature map is up-sampled **with multi-scale scheme in the pyramid mode**

Observations

Only the coarsest feature map is up-sampled **at each staircase in the ladder manner**

Features extracted from the **middle layers** are not exploited effectively.

Unsolved Issues

Features extracted from the **middle layers** just perform **single role** of linking with up-sampled versions of lower-resolution maps.

Proposed Solution:

Leverage utilization of features learned at middle layers for boosting accuracy.

Proposed Approach: To-be vs. As-is for Semantic Segmentation (1/2)

CNN: Convolutional Neural Network

Feature maps of interest acquired from backbone CNNs
Intermediate feature maps inferred during decoding stage
Final pixel-wise prediction map
Spatial pyramid pooling / dilated convolution
Combination modules
Convolution
Identity
Upsampling

AS-IS

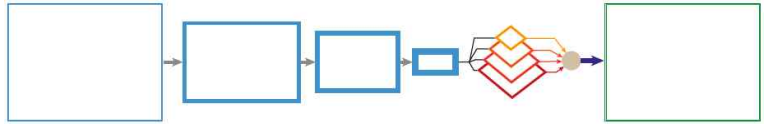


Figure: General **concept** of **asymmetrically-structured** network (spatial pyramid pooling [103])

Features extracted from the **middle layers** are not exploited effectively.

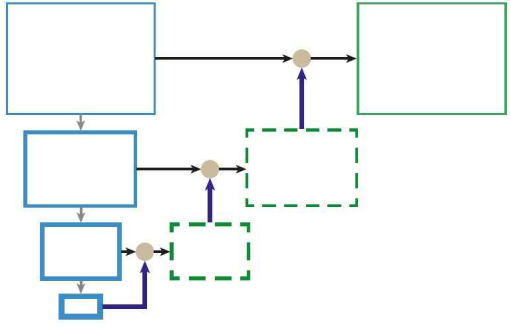
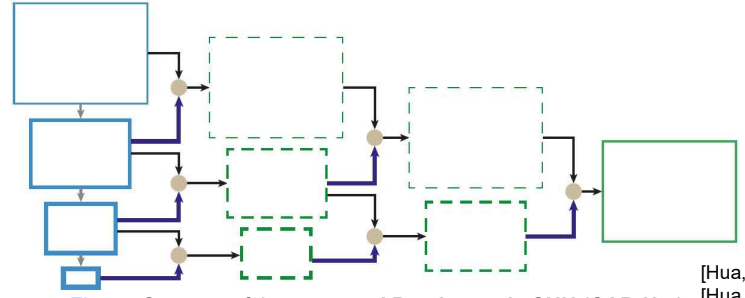


Figure: General **concept** of **symmetrically-structured** network [75, 83, 8, 43, 53, 60, 63, 57, 97, 101]

Features extracted from the **middle layers** just perform **single role** of combining with up-sampled versions of lower-resolution maps.

TO-BE



[Hua, 2018]
[Hua, 2020b]

Figure: **Concept** of the **proposed Bracket-style CNN (CAB-Net)**
(Asymmetrically-structured** network **with regularly cross-attentional connections)

Uniqueness

- **Every feature map combines** with its adjacent higher-resolution version through **cross-attentional fusion** to yield finer ones at each round.
- The newly finer maps **continuously repeat** same manner round-by-round until achieving the final feature map having finest resolution.

Benefits

Contextual information in **middle-scale features** is **exploited exhaustively**.
→**Richer semantic information** in the final segmentation map.
→**Improve** semantic segmentation **accuracy**.

Proposed Approach: To-be vs. As-is for Semantic Segmentation (2/2)

CNN: Convolutional Neural Network

Feature maps of interest acquired from backbone CNNs
Intermediate feature maps inferred during decoding stage
Final pixel-wise prediction map
Combination modules
Convolution
Identity
Upsampling

Concat: Channel-wise Concatenation
Spa. Att.: Spatially Attentional Block
Cha. Att.: Channel-wisely Attentional Block
T. Conv.: Transpose Convolution layer
Sep. Conv.: Separable Convolution layer

x = - {dash}: spatial size equals to 1×1
{x, d}: Feature map having stride of x (i.e., its spatial dimension is 1/x that of the input image) and d channels

AS-IS

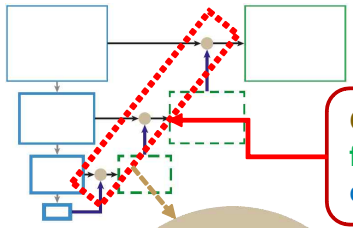
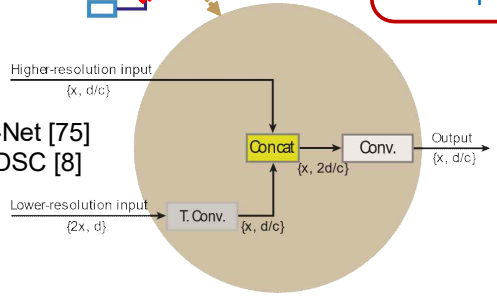


Figure: General **concept** of **symmetrically-structured** network [75, 8, 97]

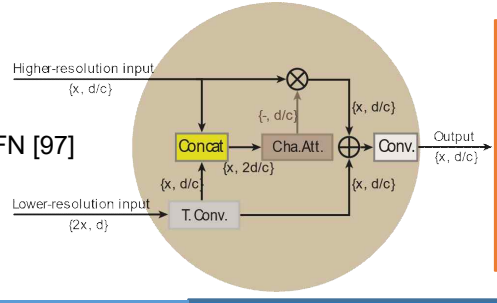
Combination modules: Refine upsampled features using finer-grained knowledge in the corresponding features of same spatial size

U-Net [75]
DDSC [8]



- **Channel-wise Concatenation** → **large tensor's depth size** during upsampling process.
- **Lack of attention** → **Weakly representational** features.

DFN [97]



- **Channel-wise Concatenation** → **large tensor's depth size** during upsampling process.
- Higher-resolution input only manipulated by its **own** and lower-resolution input's **semantic context** → **Suboptimal learning**.

TO-BE

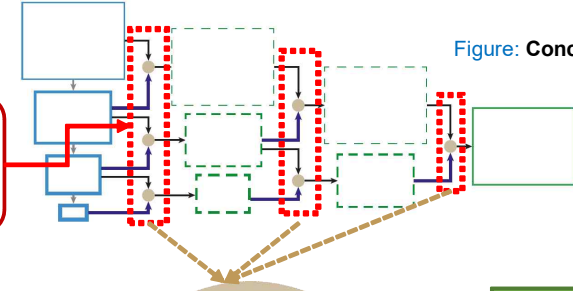
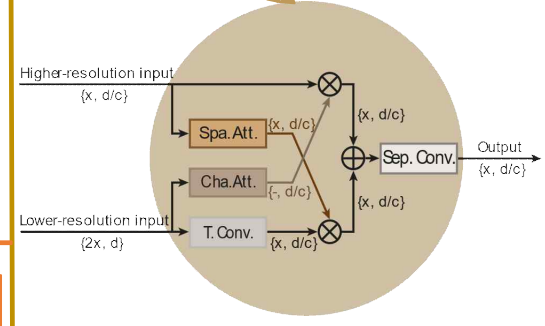


Figure: **Concept** of the **proposed Bracket-style CNN (CAB-Net)** [Hua, 2020b]



Benefits

Cost-effective Inference of **robustly representational features** during the *round-by-round* **tournament of semantic segmentation.**

Uniqueness

Cross-Attentional Fusion modules:

- Higher-resolution input manipulated by **semantically-rich context** of lower-resolution input.
- Upsampled version of Lower-resolution input manipulated by **spatially-rich details** of higher-resolution input.
- Element-wise summation between **cross-attended** features.

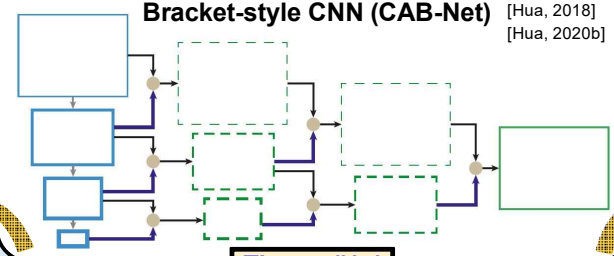
Proposed Approach: Overview of Themes

Challenges

1. Difficulties in smoothly coordinating representational features of multiple scales (fine-to-coarse resolutions)
2. Heavy ambiguities of finely-patterned details in low-level features for direct utilization in the final classifier

Proposed Solution

Bracket-style CNN (CAB-Net) [Hua, 2018] [Hua, 2020b]

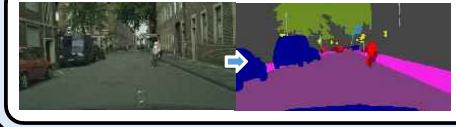


Theme #1-1

Natural Image Semantic Segmentation
for Computational photography, Autonomous driving

PASCAL VOC 2012 Dataset [20]
mIoU = 83.6%

CAMVID Dataset [9]
mIoU = 76.4%

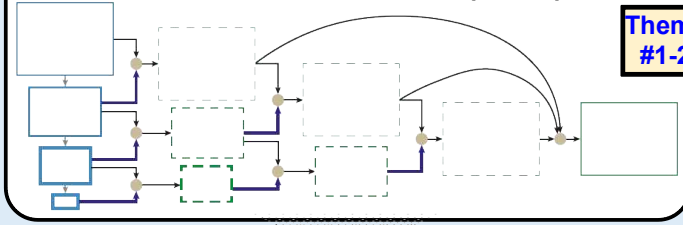


Cityscapes Dataset [18]
mIoU = 78.3%

Objectives

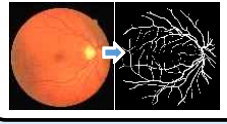
1. More robust representations using Bracket-style combinations
2. Semantically-rich finely-patterned details from cross-attentional refinements

CAB-Net with Round-wise Feature Aggregation
(Architectural Variant) [Hua, 2019]



Theme #1-2

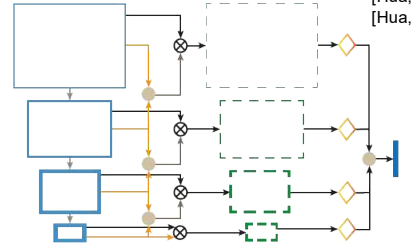
Medical Image Semantic Segmentation
for Retinal blood vessel segmentation



DRIVE Dataset [80]

Sensitivity = 79.32%
Specificity = 97.41%
Accuracy = 95.11%
AUROC = 97.32%

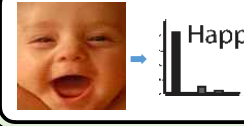
Single-mode Bracket-style CNN (sCAB-Net)
(Architectural Variant) [Hua, 2020a] [Hua, 2020c] [Hua, 2021]



Theme #2

Medical / Natural Image Classification
for Disease diagnosis, Facial expression recognition

DR Kaggle Detection Dataset [48]
QWK = 85.6%



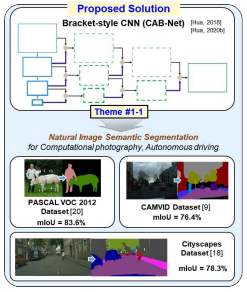
RAF-DB Dataset [59]
Mean Class Accuracy = 79.4%%

Feature maps of interest acquired from backbone CNNs
Intermediate feature maps inferred during decoding stage
Final pixel-wise prediction map
Spatial pooling
Combination modules
Convolution
Identity
Upsampling
Final feature vector
Element-wise multiplication
Attentional features extractor
Fully connected layer

Legends & Abbreviations

CNN: Convolutional Neural Network
DR: Diabetic Retinopathy
mIoU: mean Intersection of Union metric
QWK: Quadratic Weighted Kappa metric
AUROC: Area Under the Receiver Operating Characteristics metric

Proposed Approach: Theme #1-1 – Architecture Overview



[Hua, 2020b] C.-H. Hua et al., "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences, Vol.539, pp.277-294, 2020.
[Hua, 2018] C.-H. Hua et al., "Convolutional Networks with Bracket-style Decoder for Semantic Scene Segmentation", IEEE SMC, Oct 7-10, 2018.

Spa. Att.: Spatially Attentional Block
Cha. Att.: Channel-wisely Attentional Block
T. Conv.: Transpose Convolution layer
Sep. Conv.: Separable Convolution layer
{x, d}: Feature map having stride of x (i.e., its spatial dimension is 1/x that of the input image) and d channels
x = - (dash): spatial size equals to 1×1

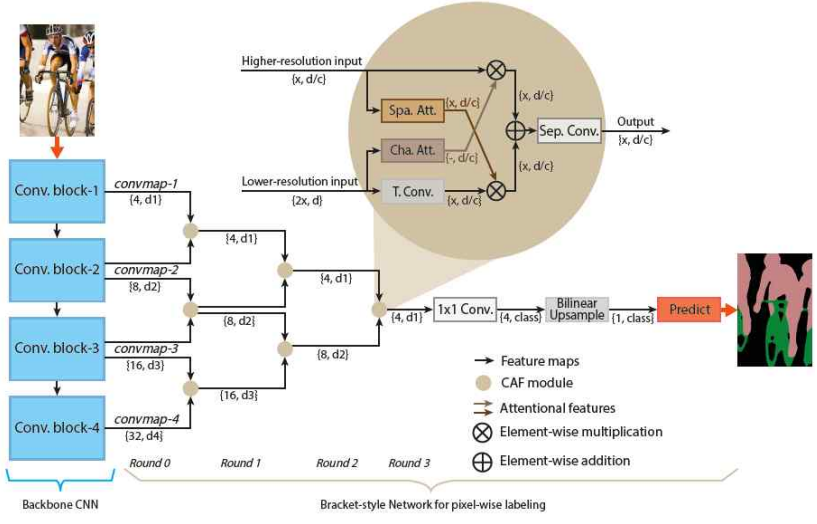


Figure: CAB-Net for **natural** image semantic segmentation [Hua, 2020b]

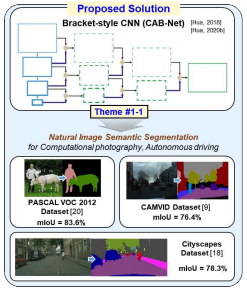
Table: **Brief description** of processes in CAB-Net

Input	Process		Output
Natural Image	**Background CNN (Round 0)**		**Feature maps** of different scales - $F_n^0, where\ n = 1, 2, 3, 4$
4 Feature maps after 0th round	**Bracket-style Network** $F_n^r = C(F_n^{r-1}, F_{n+1}^{r-1})$ *Notes:* $F_n^r = n^{th}$ feature map at r^{th} round $C(.)$: **Cross-Attentional Fusion**	Round 1	**Feature maps** after 1st round - $F_n^1, where\ n = 1, 2, 3$
3 Feature maps after 1st round		Round 2	**Feature maps** after 2nd round - $F_n^2, where\ n = 1, 2$
2 Feature maps after 2nd round		Round 3	**Feature map** after 3rd round - F_1^3
1 Feature map after 3rd round	**Finalization**	Up-sampling	**Final segmentation map** - F_{final}

❖**Backbone Convolutional Neural Network (CNN)** (e.g., **ResNet-101** [29]):
- Extract multi-scale representational feature maps.

❖**Bracket-style Network:**
- **In each round, every** of the **N** input feature maps **combines with the adjacent counterpart** via **Cross-Attentional Fusion modules** to generate **N-1** output feature maps.
  - ➢ This procedure **continuously repeats *round-by-round*** until the **final feature map** (w/ spatial size same as that of the input) is generated.
- Finally, **softmax classifier** assigns the most appropriate label to each pixel of the prediction map.

Proposed Approach: Theme #1-1 – Detailed Architecture (1/9)



[Hua, 2020b] C.-H. Hua et al., "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences, Vol.539, pp.277-294, 2020.
[Hua, 2018] C.-H. Hua et al., "Convolutional Networks with Bracket-style Decoder for Semantic Scene Segmentation", IEEE SMC, Oct 7-10, 2018.

Spa. Att.: Spatially Attentional Block
Cha. Att.: Channel-wisely Attentional Block
T. Conv.: Transpose Convolution layer
Sep. Conv.: Separable Convolution layer
{x, d}: Feature map having stride of x (i.e., its spatial dimension is 1/x that of the input image) and d channels
x = - (dash): spatial size equals to 1×1

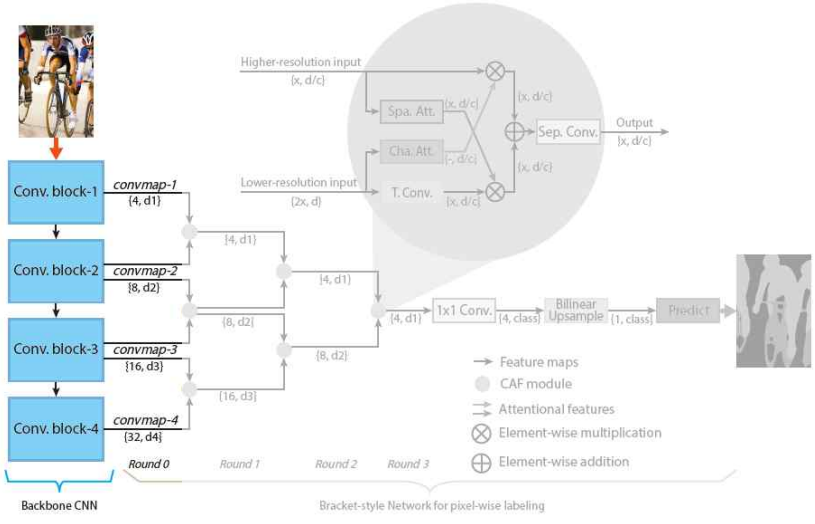
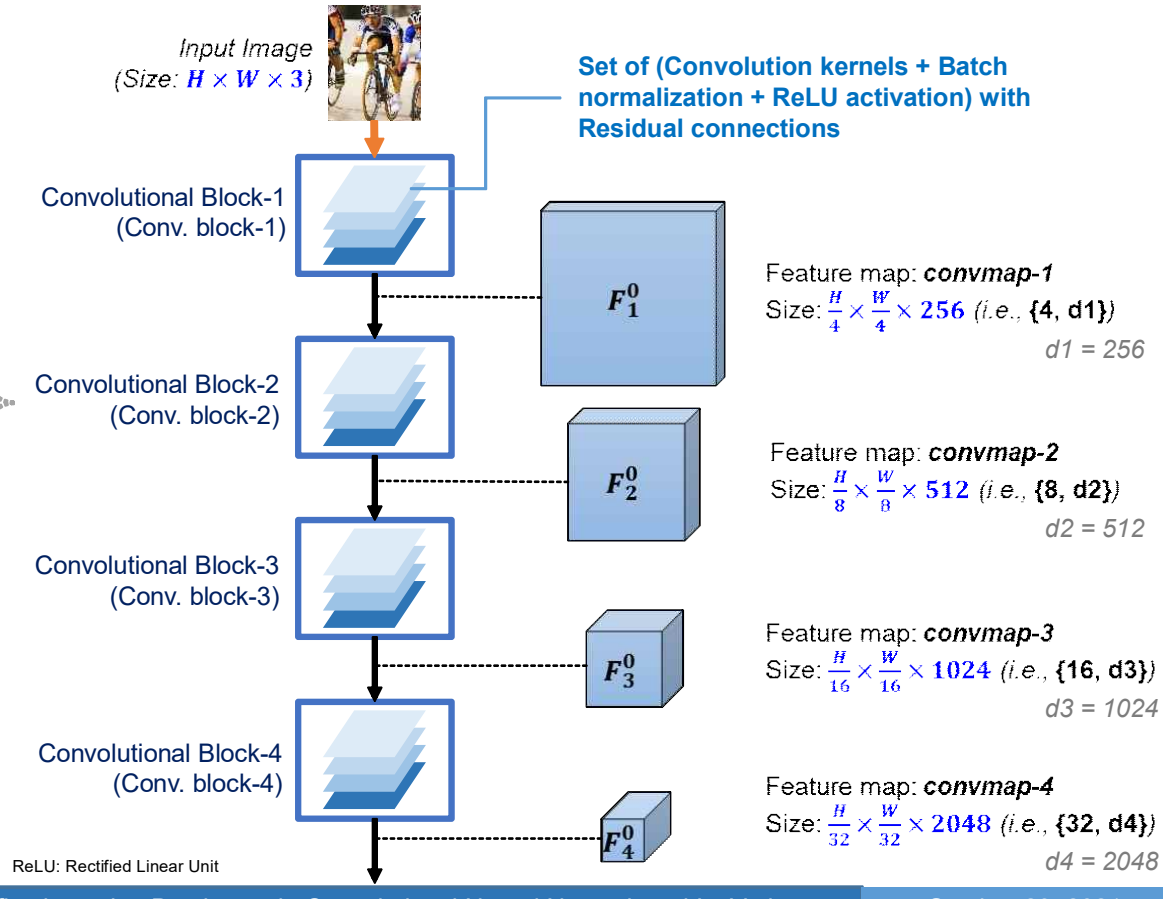


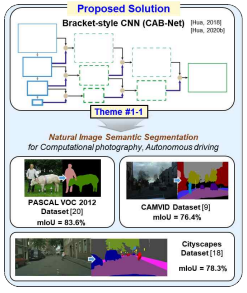
Figure: CAB-Net for **natural** image semantic segmentation [Hua, 2020b]

Backbone CNN (e.g., ResNet-101 [29])
→ **Extract multi-scale representational feature maps**



ReLU: Rectified Linear Unit

Proposed Approach: Theme #1-1 – Detailed Architecture (2/9)



[Hua, 2020b] C.-H. Hua et al., "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences, Vol.539, pp.277-294, 2020.
[Hua, 2018] C.-H. Hua et al., "Convolutional Networks with Bracket-style Decoder for Semantic Scene Segmentation", IEEE SMC, Oct 7-10, 2018.

Spa. Att.: Spatially Attentional Block
Cha. Att.: Channel-wisely Attentional Block
T. Conv.: Transpose Convolution layer
Sep. Conv.: Separable Convolution layer
{x, d}: Feature map having stride of x (i.e., its spatial dimension is 1/x that of the input image) and d channels
x = - (dash): spatial size equals to 1×1

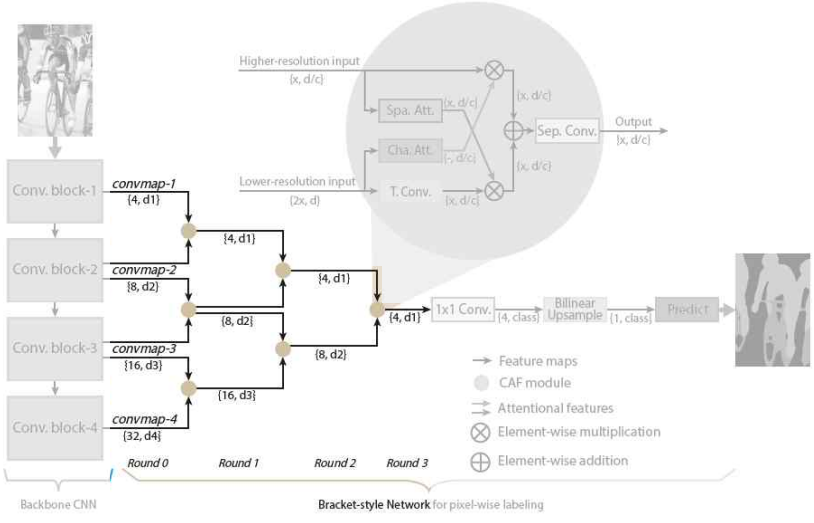


Figure: CAB-Net for **natural** image semantic segmentation [Hua, 2020b]

d1 = 256; d2 = 512; d3 = 1024; d4 = 2048

Bracket-style Network

→ **Integrate Semantically-rich context to finely-patterned details for pixel-wise labeling**

F_n^r

- $F_n^r = n^{th}$ feature map at r
- $r = 1, \ldots, N-1$ (N = No. f
- $C(.)$: Combination module

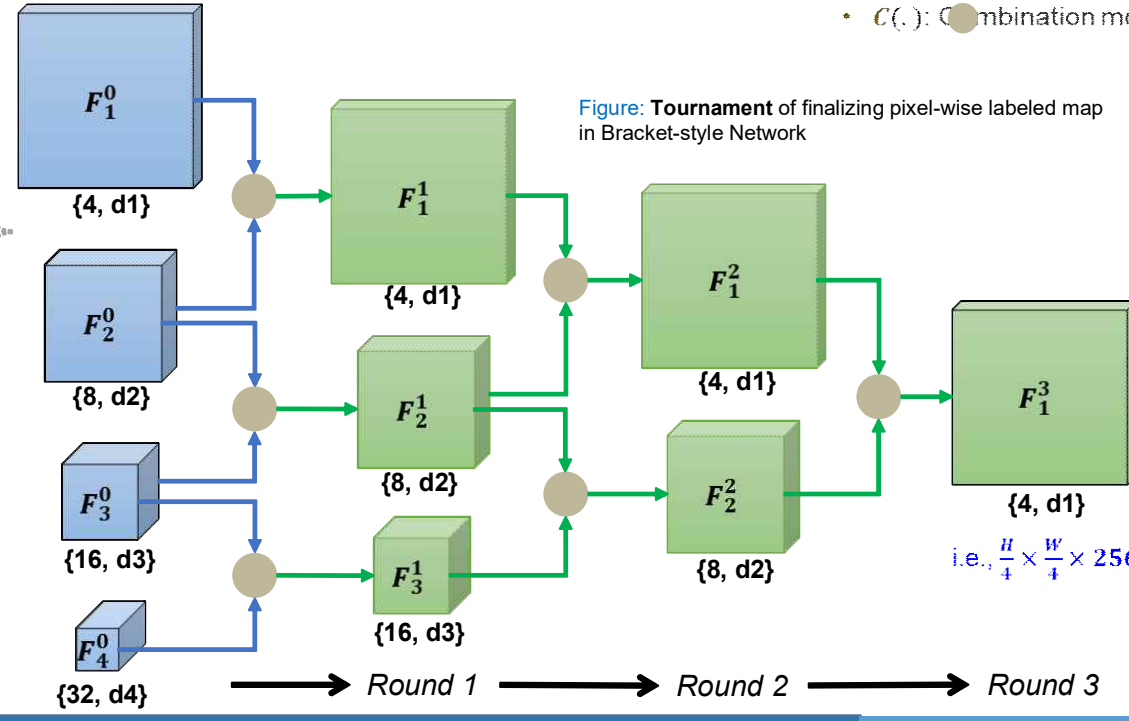
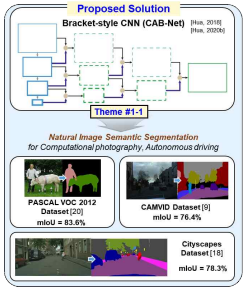


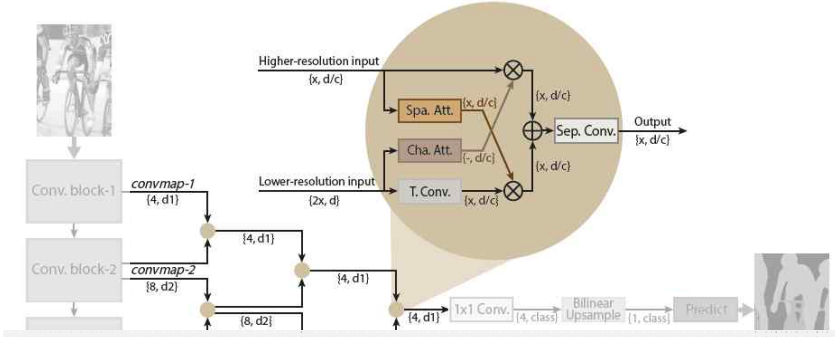
Figure: **Tournament** of finalizing pixel-wise labeled map in Bracket-style Network

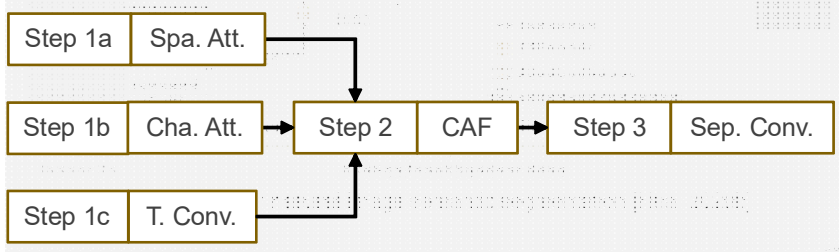
Proposed Approach: Theme #1-1 – Detailed Architecture (3/9)

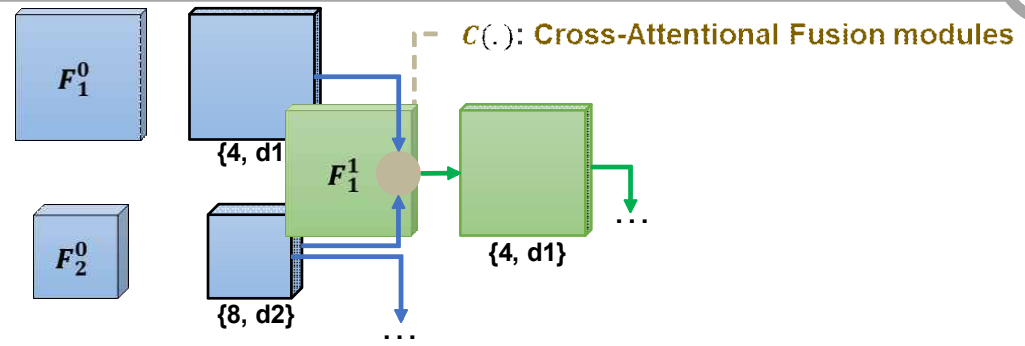


[Hua, 2020b] C.-H. Hua et al., "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences, Vol.539, pp.277-294, 2020.
[Hua, 2018] C.-H. Hua et al., "Convolutional Networks with Bracket-style Decoder for Semantic Scene Segmentation", IEEE SMC, Oct 7-10, 2018.

Spa. Att.: Spatially Attentional Block
Cha. Att.: Channel-wisely Attentional Block
T. Conv.: Transpose Convolution layer
Sep. Conv.: Separable Convolution layer
{x, d}: Feature map having stride of x (i.e., its spatial dimension is 1/x that of the input image) and d channels
x = - (dash): spatial size equals to 1×1







Input:

- ❑ High-resolution (shallower / lower-level) feature map (e.g., F_1^0)
  - ✓ **Finer-patterned** representations
- ❑ Low-resolution (deeper / higher-level) feature map (e.g., F_2^0)
  - ✓ **Semantically-richer** representations

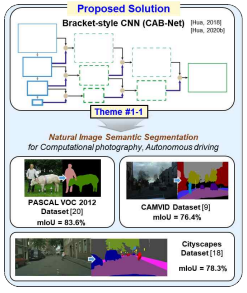
Process: Cross-Attentional Fusion modules $C(.)$, which have following components:

- ❖ Spatially Attentional (Spa. Att.) Block
- ❖ Channel-wisely (Cha. Att.) Block
- ❖ Transpose Convolution
- ❖ Element-wise multiplication followed by Element-wise addition

Output:

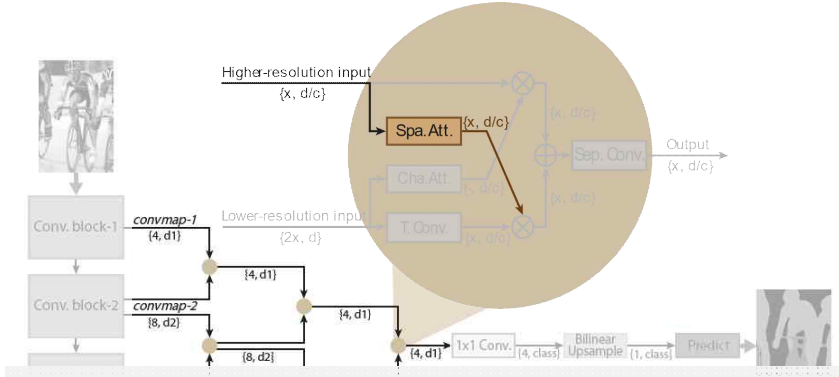
- ❑ Reconstructed feature map (e.g., F_1^1)
  - ✓ **semantically-richer** context **integrated into finer-patterned** representations

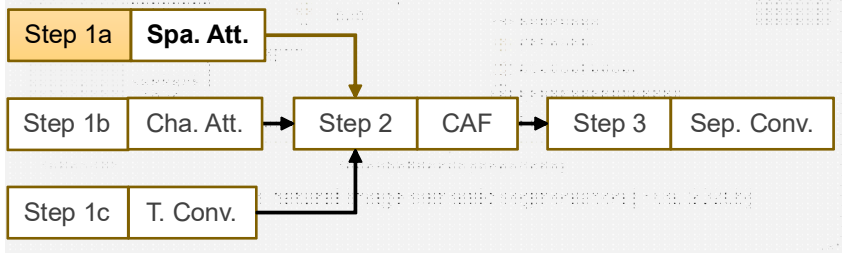
Proposed Approach: Theme #1-1 – Detailed Architecture (4/9)



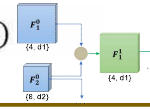
[Hua, 2020b] C.-H. Hua et al., "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences, Vol.539, pp.277-294, 2020.
[Hua, 2018] C.-H. Hua et al., "Convolutional Networks with Bracket-style Decoder for Semantic Scene Segmentation", IEEE SMC, Oct 7-10, 2018.

Spa. Att.: Spatially Attentional Block
Cha. Att.: Channel-wisely Attentional Block
T. Conv.: Transpose Convolution layer
Sep. Conv.: Separable Convolution layer
{x, d}: Feature map having stride of x (i.e., its spatial dimension is 1/x that of the input image) and d channels
x = - (dash): spatial size equals to 1×1





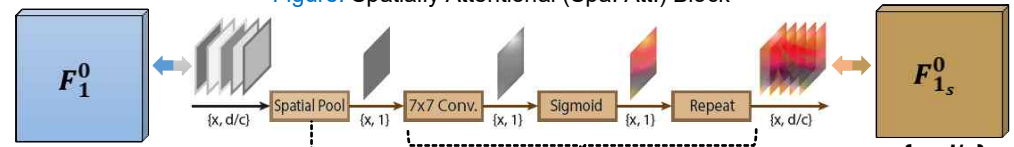
Process: Cross-Attentional Fusion modules $C(.)$



Suppose x = 4; d1 = d/c = 256; d2 = d = 512

Step 1a Extract Spatially Attentional Features from higher-resolution input

Figure: Spatially Attentional (Spa. Att.) Block



Finer-patterned representations

Spatially Attentional Features

Compute average (pooled) value across channels of feature map

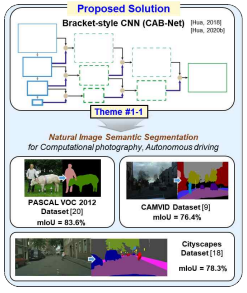
$$F_{n_{sp}}^{r-1}{}_{h,w} = \frac{c}{d}\sum_{z=1}^{\frac{d}{c}} F_n^{r-1}{}_{h,w,z}$$

Learn spatially attentional features (i.e., express spatially discriminative levels of pooled features)

$$F_{n_s}^{r-1} = \text{Repeat}(\sigma(W_{7\times7} * F_{n_{sp}}^{r-1}))$$

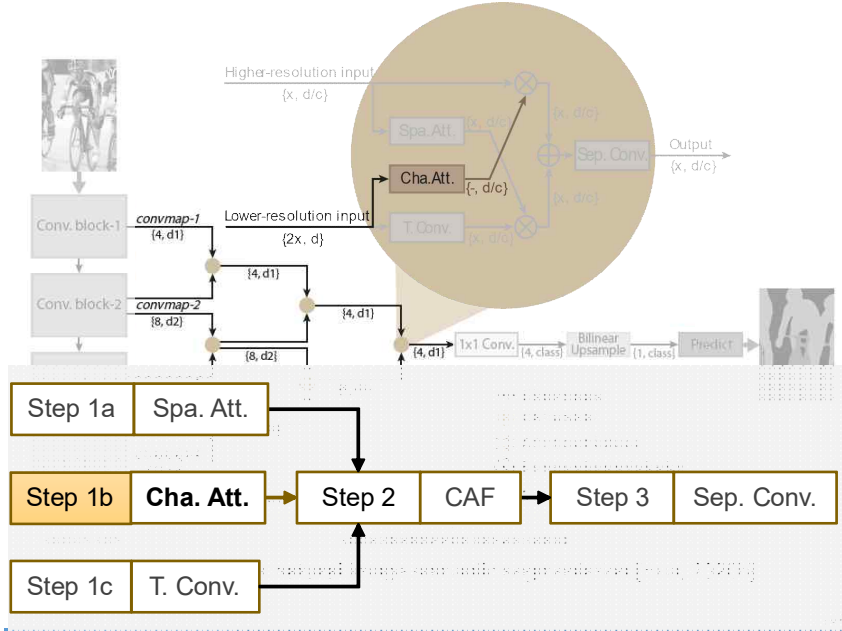
$F_n^r = n^{th}$ feature map at r^{th} round
(h, w, z): pixel coordinates of feature map F
$*$: convolution operator
$W_{7\times7} \in R^{1\times7\times7\times1}$: learnable weights of 7x7 Convolution filters
σ: Sigmoid activation function
Repeat: duplication of feature map along channel dimension

Proposed Approach: Theme #1-1 – Detailed Architecture (5/9)

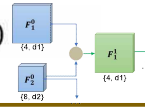


[Hua, 2020b] C.-H. Hua et al., "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences, Vol.539, pp.277-294, 2020.
[Hua, 2018] C.-H. Hua et al., "Convolutional Networks with Bracket-style Decoder for Semantic Scene Segmentation", IEEE SMC, Oct 7-10, 2018.

Spa. Att.: Spatially Attentional Block
Cha. Att.: Channel-wisely Attentional Block
T. Conv.: Transpose Convolution layer
Sep. Conv.: Separable Convolution layer
{x, d}: Feature map having stride of x (i.e., its spatial dimension is 1/x that of the input image) and d channels
x = - (dash): spatial size equals to 1×1



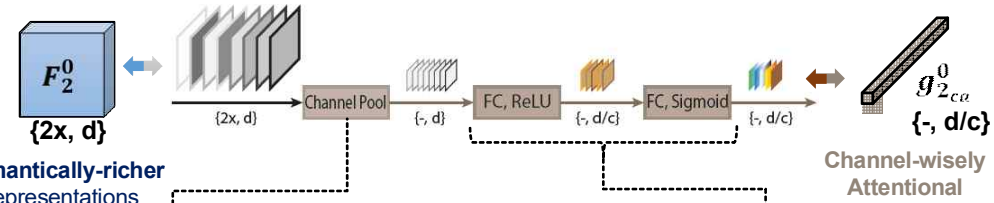
Process: Cross-Attentional Fusion modules $C(.)$



Suppose x = 4; d1 = d/c = 256; d2 = d = 512

Step 1b Extract Channel-wisely Attentional Features from lower-resolution input

Figure: Channel-wisely Attentional (Cha. Att.) Block



Semantically-richer representations

Channel-wisely Attentional Features

Compute spatial average (pooled) values at each channel of feature map

$$G_d(\boldsymbol{F}_{n+1}^{r-1}) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \boldsymbol{F}_{n+1\,h,w,d}^{r-1}$$

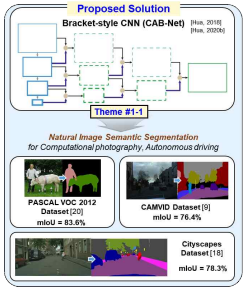
$$\boldsymbol{g} = [G_1(\boldsymbol{F}_{n+1}^{r-1}), \ldots, G_d(\boldsymbol{F}_{n+1}^{r-1})]^T$$

Learn channel-wisely attentional features (i.e., express cross-channel discriminative levels of pooled features)

$$\boldsymbol{g}_{n+1_{ca}}^{r-1} = \sigma\left(\boldsymbol{W}_{fc2}\left(ReLU\left(\boldsymbol{W}_{fc1}\boldsymbol{g} + \boldsymbol{B}_{fc1}\right)\right) + \boldsymbol{B}_{fc2}\right)$$

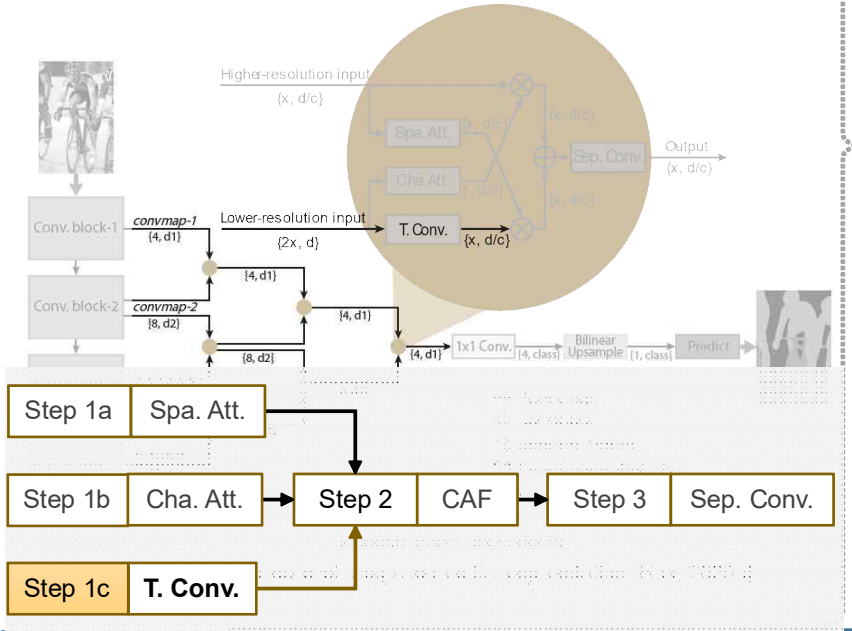
$\boldsymbol{F}_n^r = n^{th}$ feature map at r^{th} round
$G_d(\boldsymbol{F}_n^r)$: Channel Pool operator at d^{th} channel of feature map $\boldsymbol{F}_n^r$
$\boldsymbol{W}_{fc1} \in R^{\frac{d}{c} \times d}, \boldsymbol{B}_{fc1} \in R^{\frac{d}{c}}$: weights and biases of 1st Fully-connected layer
$\boldsymbol{W}_{fc2} \in R^{\frac{d}{c} \times \frac{d}{c}}, \boldsymbol{B}_{fc2} \in R^{\frac{d}{c}}$: weights and biases of 2nd Fully-connected layer
$ReLU$: Rectified Linear Unit activation function
σ: Sigmoid activation function

Proposed Approach: Theme #1-1 – Detailed Architecture (6/9)

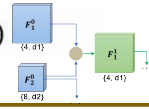


[Hua, 2020b] C.-H. Hua et al., "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences, Vol.539, pp.277-294, 2020.
[Hua, 2018] C.-H. Hua et al., "Convolutional Networks with Bracket-style Decoder for Semantic Scene Segmentation", IEEE SMC, Oct 7-10, 2018.

Spa. Att.: Spatially Attentional Block
Cha. Att.: Channel-wisely Attentional Block
T. Conv.: Transpose Convolution layer
Sep. Conv.: Separable Convolution layer
{x, d}: Feature map having stride of x (i.e., its spatial dimension is 1/x that of the input image) and d channels
x = - (dash): spatial size equals to 1×1



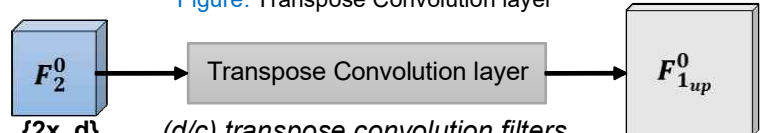
Process: Cross-Attentional Fusion modules $C(.)$



Suppose x = 4; d1 = d/c = 256; d2 = d = 512

Step 1c Upsample lower-resolution input using Transpose Convolution layer

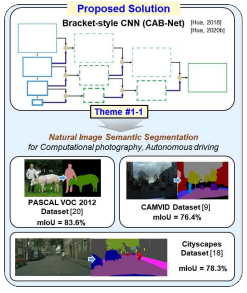
Figure: Transpose Convolution layer



$$W_u *^u F_{n+1}^{r-1} + B_u = F_{n_{up}}^{r-1}$$

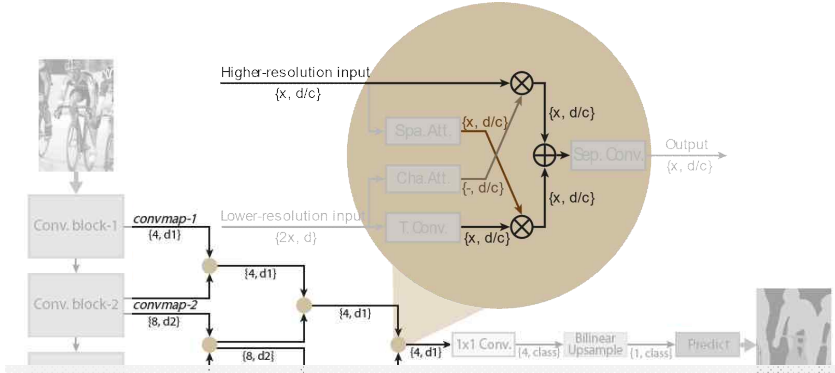
$F_n^r = n^{th}$ feature map at r^{th} round
$*^u$: fractionally-strided convolution operator
$W_u \in R^{\frac{d}{c} \times 3 \times 3 \times d}, B_u \in R^{\frac{d}{c}}$: weights and biases of Transpose Convolution filters

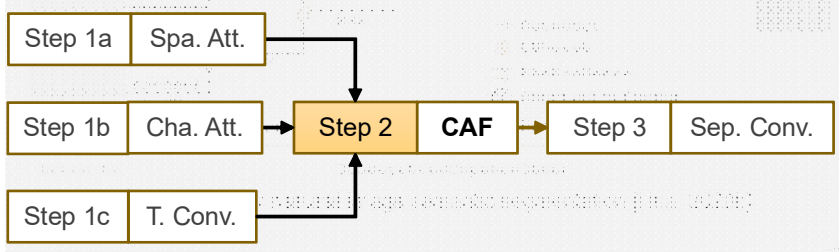
Proposed Approach: Theme #1-1 – Detailed Architecture (7/9)



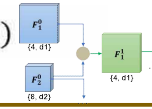
[Hua, 2020b] C.-H. Hua et al., "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences, Vol.539, pp.277-294, 2020.
[Hua, 2018] C.-H. Hua et al., "Convolutional Networks with Bracket-style Decoder for Semantic Scene Segmentation", IEEE SMC, Oct 7-10, 2018.

Spa. Att.: Spatially Attentional Block
Cha. Att.: Channel-wisely Attentional Block
T. Conv.: Transpose Convolution layer
Sep. Conv.: Separable Convolution layer
{x, d}: Feature map having stride of x (i.e., its spatial dimension is 1/x that of the input image) and d channels
x = - (dash): spatial size equals to 1×1



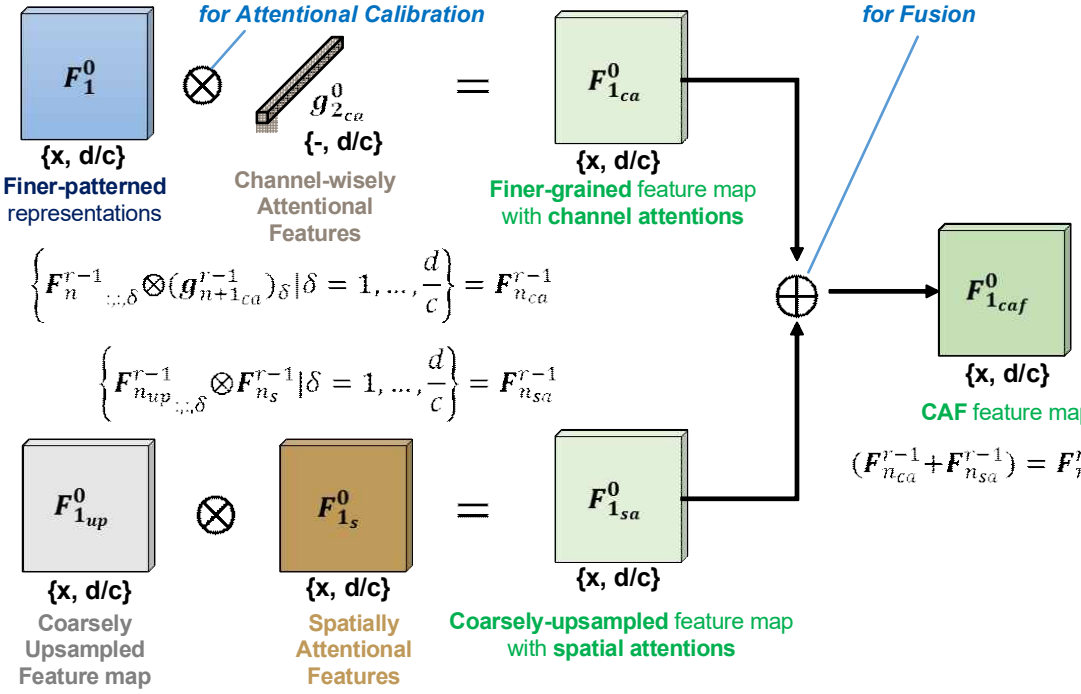


Process: Cross-Attentional Fusion modules $C(.)$



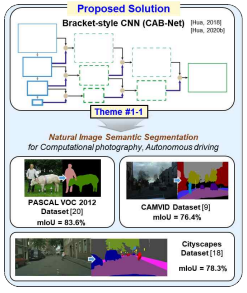
Suppose x = 4; d1 = d/c = 256; d2 = d = 512

Step 2 Perform Cross-Attentional Fusion (CAF)



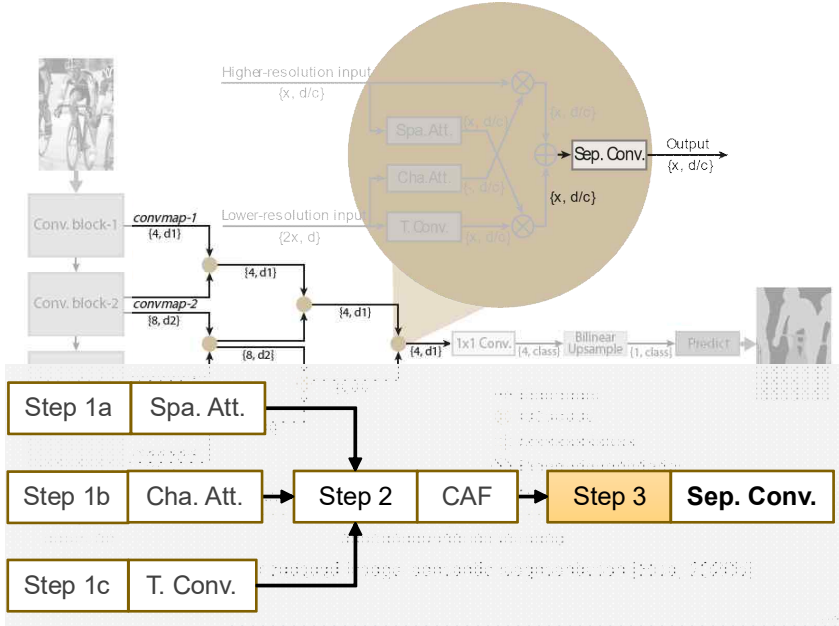
$(g^{r-1}_{n+1_{ca}})_\delta$: δ-th channel of $(n+1)^{th}$ feature vector at $(r-1)^{th}$ round
$F^r_{n\ :,:,\delta}$: δ-th channel of the feature map F^r_n with its full spatial dimension
$\otimes$: Element-wise multiplication

Proposed Approach: Theme #1-1 – Detailed Architecture (8/9)

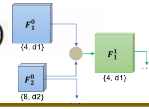


[Hua, 2020b] C.-H. Hua et al., "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences, Vol.539, pp.277-294, 2020.
[Hua, 2018] C.-H. Hua et al., "Convolutional Networks with Bracket-style Decoder for Semantic Scene Segmentation", IEEE SMC, Oct 7-10, 2018.

Spa. Att.: Spatially Attentional Block
Cha. Att.: Channel-wisely Attentional Block
T. Conv.: Transpose Convolution layer
Sep. Conv.: Separable Convolution layer
{x, d}: Feature map having stride of x (i.e., its spatial dimension is 1/x that of the input image) and d channels
x = - (dash): spatial size equals to 1×1



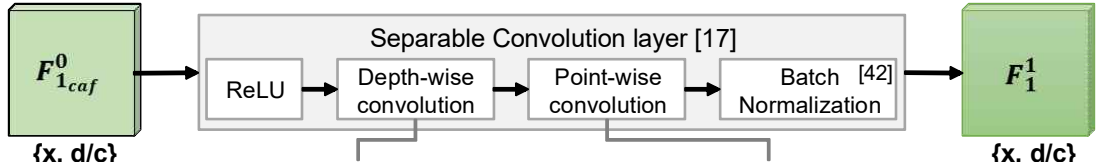
Process: Cross-Attentional Fusion modules $C(.)$



Suppose x = 4; d1 = d/c = 256; d2 = d = 512

Step 3 Learn CAF feature map using Separable Convolution layer

Figure: Separable Convolution layer



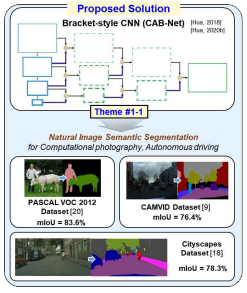
$$BN(\boldsymbol{W}_{sc} * ReLU(\boldsymbol{F}^{r-1}_{n_{caf}})) = \boldsymbol{F}^r_n$$

$\boldsymbol{W}_{sc} = \left\{\boldsymbol{W}_{df} \in R^{\frac{d}{c}\times 3\times 3\times 1}, \boldsymbol{W}_{pf} \in R^{\frac{d}{c}\times 1\times 1\times \frac{d}{c}}\right\}$: weights of depth-wise and point-wise convolution filters, respectively

$ReLU$: Rectified Linear Unit activation function

BN: Batch Normalization layer

Proposed Approach: Theme #1-1 – Detailed Architecture (9/9)



[Hua, 2020b] C.-H. Hua et al., "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences, Vol.539, pp.277-294, 2020.
[Hua, 2018] C.-H. Hua et al., "Convolutional Networks with Bracket-style Decoder for Semantic Scene Segmentation", IEEE SMC, Oct 7-10, 2018.

Spa. Att.: Spatially Attentional Block
Cha. Att.: Channel-wisely Attentional Block
T. Conv.: Transpose Convolution layer
Sep. Conv.: Separable Convolution layer
{x, d}: Feature map having stride of x (i.e., its spatial dimension is 1/x that of the input image) and d channels
x = - (dash): spatial size equals to 1×1

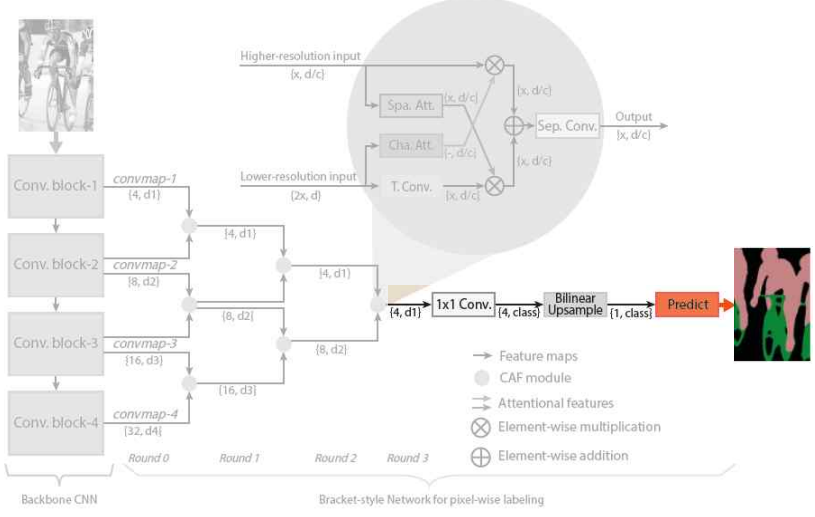


Figure: CAB-Net for **natural** image semantic segmentation [Hua, 2020b]

Finalization

F_1^3 {4, 256} **Feature map after 3rd round**
d1 = 256

1x1 Convolution layer $F_{class} = W_{11} * ReLU(F_1^3)$

$W_{11} \in R^{class \times 1 \times 1 \times \frac{d}{c}}$: weights of 1x1 Convolution filters
$ReLU$: Rectified Linear Unit activation function

F_{class} {4, class} **Compressed-channel Feature map**
*(which has **channel dimension** equal to number of **supervised classes**)*

Bilinear upsample

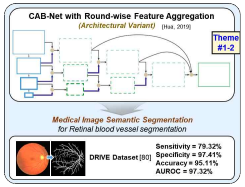
F_{final} {1, class} $H \times W \times class$
Final Segmentation Score map
*(which has **same spatial dimension** as that of the input image)*

Predict

$$\hat{F}_{final} = Softmax(F_{final}) = \left\{ \hat{F}_{final_i} = \frac{e^{F_{final_i}}}{\sum_{j=1}^{Class} e^{F_{final_j}}} | i = 1, \ldots, Class \right\}$$

Final pixel-wise segmentation map
$H \times W \times 1$
*(Each pixel is assigned the **label (class)** having **highest score** along the **channel dimension** of $\hat{F}_{final}$)*

Proposed Approach: Theme #1-2 – Architectural Variant

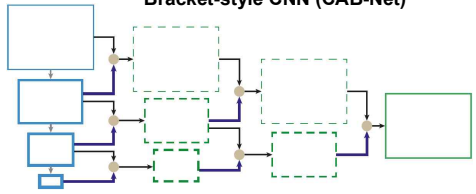


[Hua, 2019] C.-H. Hua et al., "Retinal Vessel Segmentation using Round-wise Features Aggregation on Bracket-shaped Convolutional Neural Networks", IEEE EMBC, July 23-27, 2019.

- Feature maps of interest acquired from backbone CNNs
- Intermediate feature maps inferred during decoding stage
- Final pixel-wise prediction map
- Spatial pyramid pooling / dilated convolution
- Combination modules
- Convolution
- Identity
- Upsampling

Theme #1-1

Figure: **Concept** of the **proposed Bracket-style CNN (CAB-Net)**



Observations:

Ratio between the number of **blood vessel** and **background pixels** is **massively imbalanced** in fundus images.

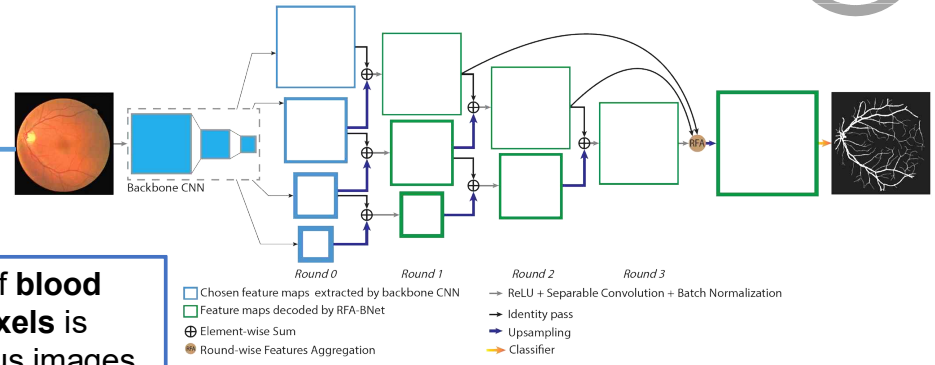


Figure: **CAB-Net** with **Round-wise Feature Aggregation (RFA-BNet)** for **medical (fundus)** image semantic segmentation [Hua, 2019]

Hypothesis

Leveraging the **utilization** of **round-wise finely-patterned features** is capable of effectively labeling small objects in case of heavily class-imbalancing issue.

Architectural Variants

Theme #1-2

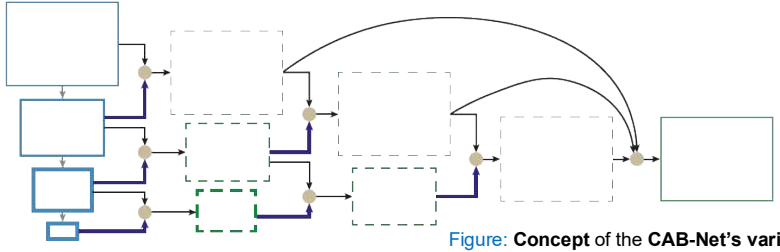


Figure: **Concept** of the **CAB-Net's variant**, namely **RFA-BNet**

Table: **Differences** between **baseline** CAB-Net for **natural** image semantic segmentation and its **variant** for **medical** image

Variants	Baseline CAB-Net for natural image	Variant for medical (fundus) image
Combination module $\mathcal{C}(.)$	Cross-Attentional Fusion (CAF) (*including Upsampling, Spatial & Channel-wise attention*)	Upsampling + Element-wise Sum (***simplicity*** *for binary labeling*)
Finalization $\rightarrow F_{final}$	$Upsampling(F_1^3)$	$Upsampling(Concat(F_1^3, F_1^2, F_1^1))$

Round-wise Feature Aggregation using **depth-wise concatenation** ***(Concat)*** for aggregating multi-round finely patterned features in cases of labeling diversely and irregularly present objects (e.g., blood vessel) under heavily class-imbalancing issue **(without necessity of costly processing image patches like [22, 31, 45, 65])**.

Proposed Approach Variant: To-be vs. As-is for Semantic Segmentation

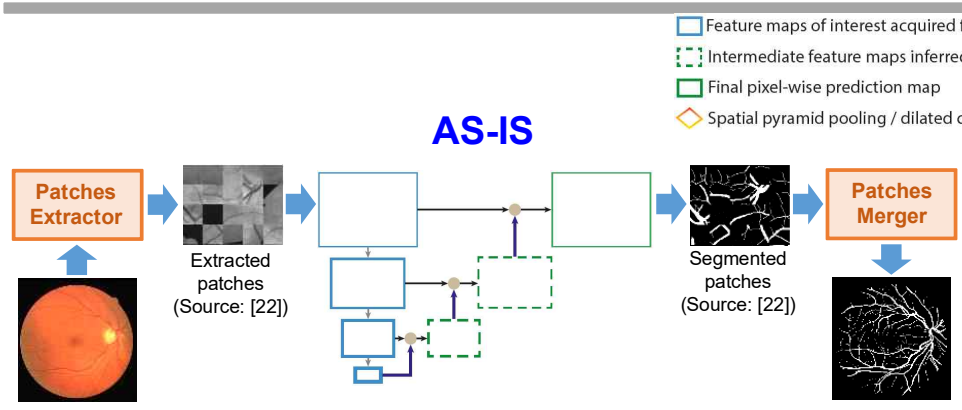


Figure: General **concept** of **symmetrically-structured** network [75] with **costly pre/post-processing of image patching** [22, 31, 45, 65] for Retinal Blood Vessel Segmentation

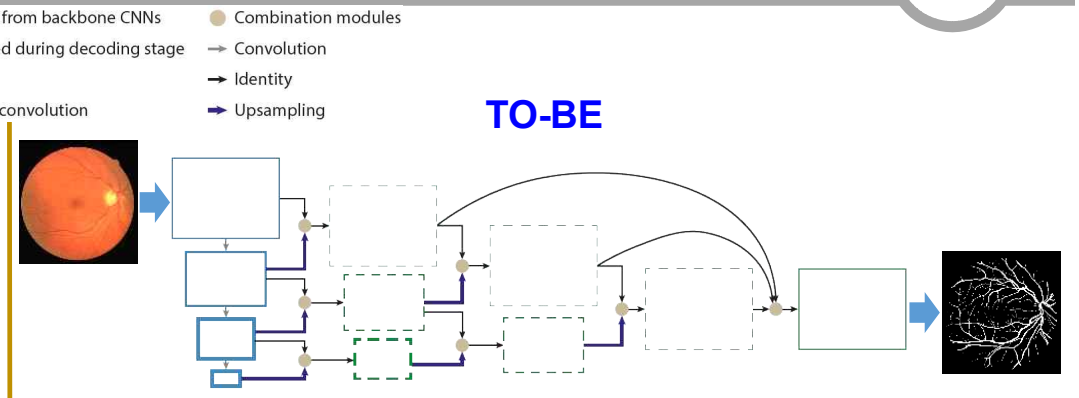


Figure: **Concept** of the **proposed CAB-Net with Round-wise Feature Aggregation (RFA-BNet)** for Retinal Blood Vessel Segmentation [Hua, 2019]

Ratio between the number of **blood vessel** and **background pixels** is **massively imbalanced** in fundus images.

Additional pre/post-processing steps related to the generation and integration of small patches from the input and into the segmentation output, respectively.

Pros: Patch-based approaches facilitate the segmentation model to learn features more effectively.
Cons: **Expensive computation** and **complications** in data pre/post-processing, training, and execution stages for trading-off better performance.

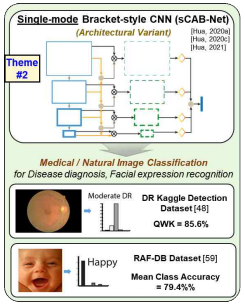
Uniqueness

- **Round-wise Feature Aggregation** attached to the baseline **CAB-Net** for **leveraging** the **utilization of finely-patterned features** across different decoding rounds.

Benefits

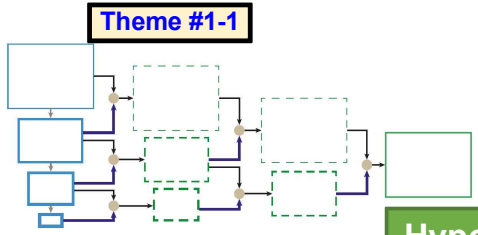
Contextual information in **low-level features of earlier rounds** is **additionally leveraged** besides the middle-level counterparts..
→**Attain comparable performance** with existing patch-based approaches **without** considering costly **image patching**.

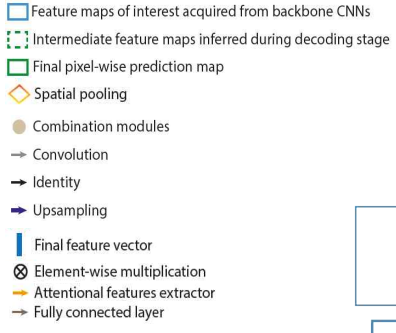
Proposed Approach: Theme #2 – Architectural Variant



[Hua, 2020a] C.-H. Hua et al., "Convolutional Network with Densely Backward Attention for Facial Expression Recognition", IMCOM, Jan 3-5, 2020.
[Hua, 2020c] C.-H. Hua et al., "DRAN: Densely Reversed Attention based Convolutional Network for Diabetic Retinopathy Detection", IEEE EMBC, July 20-24, 2020.
[Hua, 2021] C.-H. Hua et al., "Convolutional Network with Twofold Feature Augmentation for Diabetic Retinopathy Recognition from Multi-modal Images", IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2686-2697, July 2021.

Figure: **Concept** of the **proposed Bracket-style CNN (CAB-Net)**





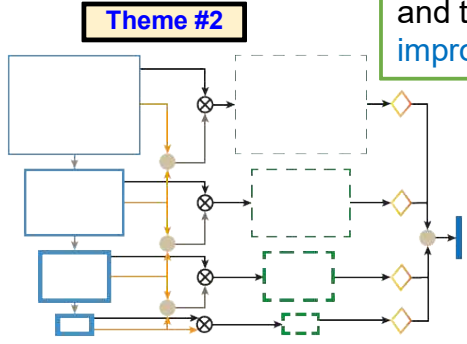


Figure: **Concept** of the **proposed Single-mode Bracket-style CNN (sCAB-Net)**

Hypothesis

Pair-wisely refining fine-grained details of **low-level** features using semantic context of **high-level** ones and then **amalgamating** them can improve recognition performance

Observations:

Diabetic Retinopathy (DR): Severity scales are determined by various structural factors such as existence of microaneurysms , exudates , hemorrhages, neovascularization, etc.

Facial Expression Recognition: Emotions via a person's face are represented by the combination of various muscular modalities (e.g., deformation of eyes, eyebrow, nose wrinkle, mouth, …).

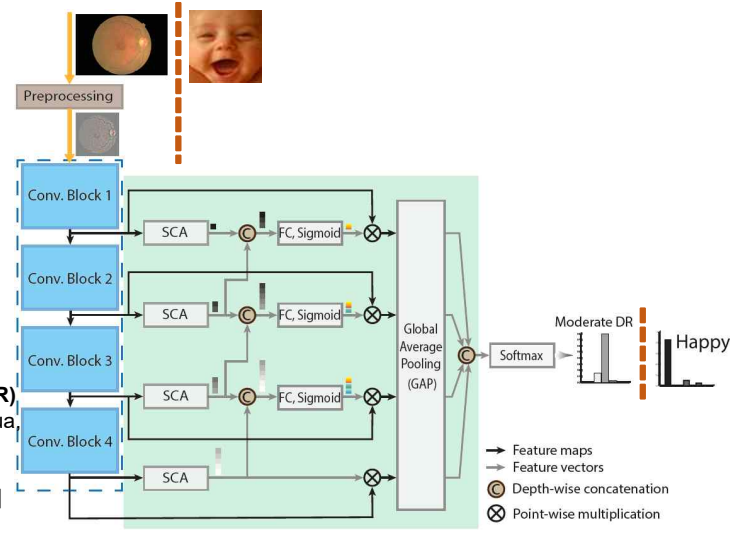


Figure: sCAB-Net for **Diabetic Retinopathy (DR) Severity Recognition** [Hua, 2020a], [Hua, 2021] or **Facial Expression Recognition** [Hua, 2020c]

Proposed Approach Variant: To-be vs. As-is for Image Classification

CNN: Convolutional Neural Network
NP: Nonproliferative
DR: Diabetic Retinopathy
FER: Facial Expression Recognition

Source: * Kaggle DR Detection dataset [48];
** RAF-DB dataset [59]

Feature maps of interest acquired from backbone CNNs
Intermediate feature maps inferred during decoding stage
Final feature vector
Spatial pooling

Combination modules
Element-wise multiplication
Convolution
Identity
Attentional features extractor
Fully connected layer

AS-IS

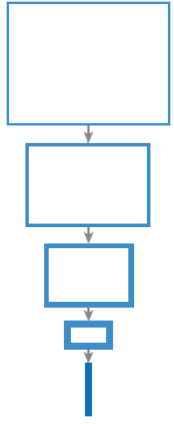


Figure: General **concept** of **conventional CNN** [44, 85]

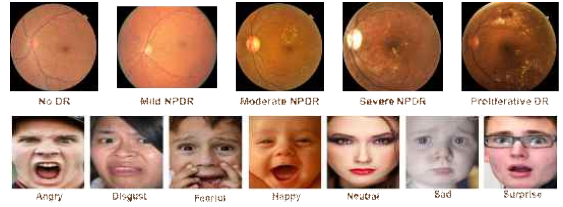
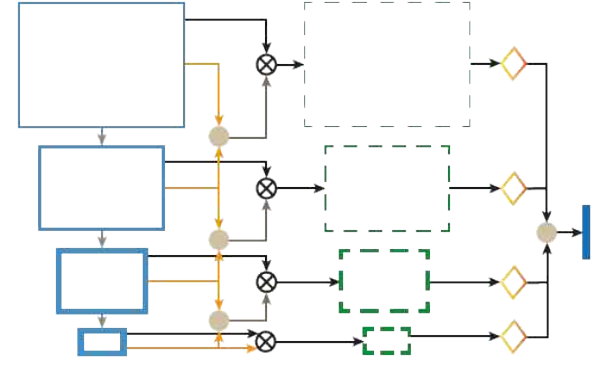


Figure: **Specialized classification domains**:
(top) DR severity levels*, (bottom) Facial expressions**

TO-BE



[Hua, 2020a]
[Hua, 2020c]
[Hua, 2021]

Figure: **Concept** of the **proposed Variant of Bracket-style CNN (i.e., Single-mode CAB-Net or sCAB-Net)**

Target labels are decided by the **amalgamation of various spatial factors** inside the image.

Multiple downsampling stages during feedforward process of the vanilla CNN

Loss of certain spatial correlations between **structural details**, which are hardly encoded along channel dimension, in the input image

Only relying on the highest-level feature for the final classifier is **not optimal**

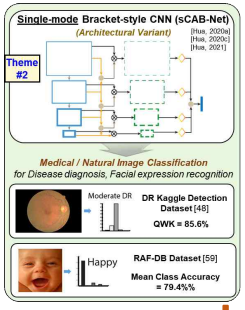
Uniqueness

- Channel-wise attentional information of semantically-rich features are **integrated** into the finely-patterned counterparts **reversely**.
- Feature maps of different scales can then be amalgamated for **extensively involving spatially-rich representations** to the final predictions.

Benefits

Contextual information in **middle- and low-level features** is **exploited exhaustively.**
✓ **Spatially-rich information** can be extensively involved **without ambiguities.**
✓ **Improve** classification **accuracy** in **specialized domains** (e.g., DR, FER).

Proposed Approach: Theme #2 – Architecture Overview



[Hua, 2020a] C.-H. Hua et al., "Convolutional Network with Densely Backward Attention for Facial Expression Recognition", IMCOM, Jan 3-5, 2020.
[Hua, 2020c] C.-H. Hua et al., "DRAN: Densely Reversed Attention based Convolutional Network for Diabetic Retinopathy Detection", IEEE EMBC, July 20-24, 2020.
[Hua, 2021] C.-H. Hua et al., "Convolutional Network with Twofold Feature Augmentation for Diabetic Retinopathy Recognition from Multi-modal Images", IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2686-2697, July 2021.

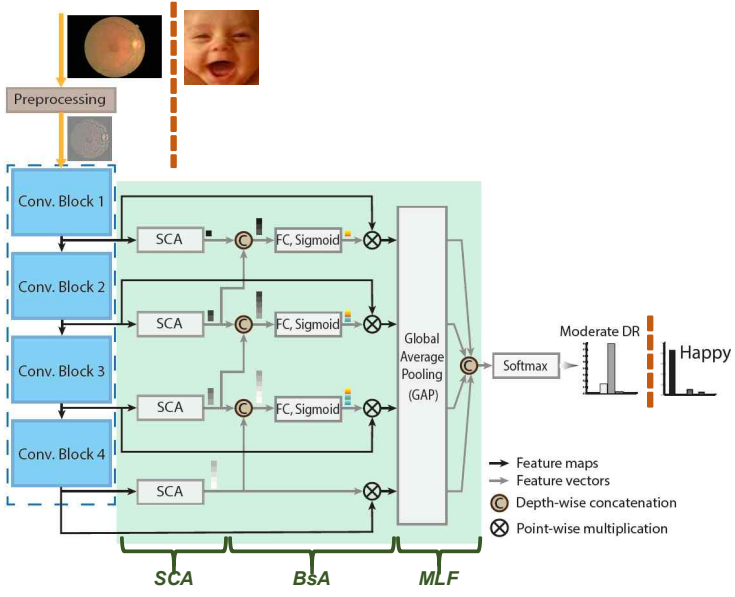


Figure: sCAB-Net for **Diabetic Retinopathy (DR) Severity Recognition** [Hua, 2020a], [Hua, 2021] or **Facial Expression Recognition** [Hua, 2020c]

Table: **Brief description** of processes in sCAB-Net

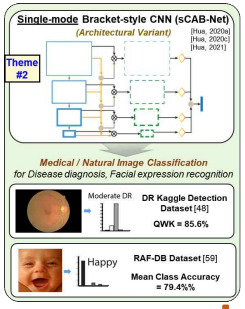
Input	Process		Output
DR-related image / Face image	**Backbone CNN**		**Feature maps** of different scales - $F_n, where\ n = 1, \dots, 4$
Feature maps of different scales - F_n	**SCA**	GAP + (FC,ReLU) + (FC,Sigmoid)	**Feature vectors** of **self-**attentional context - s_n
Feature vectors of self-attentional context - s_n	**BsA**	Depth-wise Concatenation	Cross-level **concatenated feature vectors** – $C[s_n, s_{n+1}]$, $where\ n = 1,2,3$
Cross-level concatenated feature vectors – $C[s_n, s_{n+1}]$		(FC,Sigmoid)	**Feature vectors** of **cross-attentional** context – s'_n, $where\ n = 1,2,3;\ s'_4 = s_4$
Feature vectors of **cross-**attentional context – s'_n		Point-wise multiplication	**Refined** multi-level **feature maps** - $F_{bsa_n} = F_n \otimes s'_n$
Refined multi-level feature maps - F_{bsa_n}	**MLF**	GAP + Depth-wise Concatenation	**Mixture** of multi-level **feature maps** - $F_{final} = C[F_{bsa_1}, F_{bsa_2}, F_{bsa_3}, F_{bsa_4}]$

sCAB-Net Constituents:

- Backbone CNN (dashed-line region) (e.g., ResNet-101 [29])
- Channel-wisely Cross-Attentional Scheme (green region)
  1. Self-Context Aggregation (SCA)
  2. Bracket-style Attention (BsA)
  3. Multi-level Fusion (MLF)

CNN: Convolutional Neural Network; Conv. Block: Convolutional Block; GAP: Global Average Pooilng; FC: Fully Connected layer; ReLU: Rectified Linear Unit

Proposed Approach: Theme #2 – Detailed Architecture (1/4)



[Hua, 2020a] C.-H. Hua et al., "Convolutional Network with Densely Backward Attention for Facial Expression Recognition", IMCOM, Jan 3-5, 2020.
[Hua, 2020c] C.-H. Hua et al., "DRAN: Densely Reversed Attention based Convolutional Network for Diabetic Retinopathy Detection", IEEE EMBC, July 20-24, 2020.
[Hua, 2021] C.-H. Hua et al., "Convolutional Network with Twofold Feature Augmentation for Diabetic Retinopathy Recognition from Multi-modal Images", IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2686-2697, July 2021.

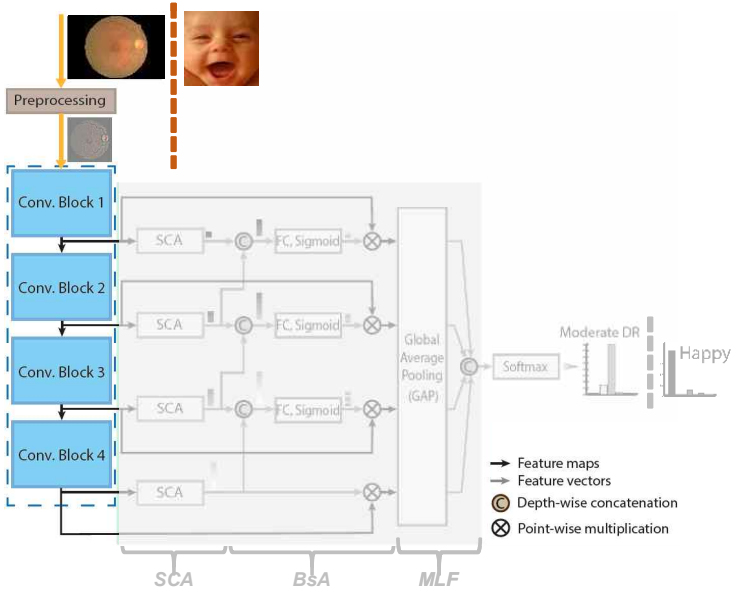
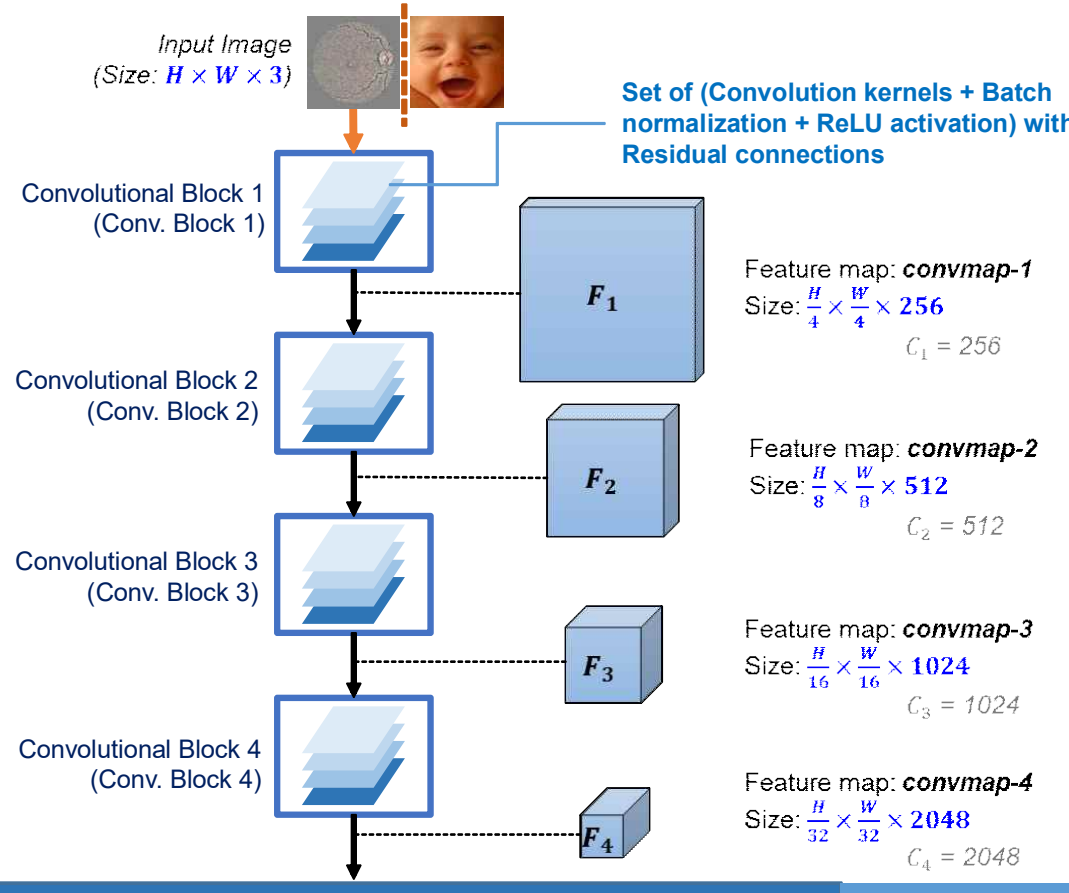
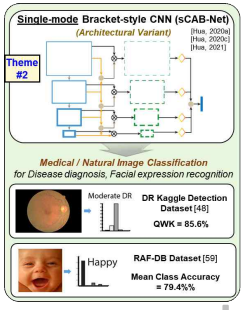


Figure: sCAB-Net for **Diabetic Retinopathy (DR) Severity Recognition** [Hua, 2020a], [Hua, 2021] or **Facial Expression Recognition** [Hua, 2020c]

Backbone CNN (e.g., ResNet-101 [29])
→ **Extract multi-scale representational feature maps**



Proposed Approach: Theme #2 – Detailed Architecture (2/4)



[Hua, 2020a] C.-H. Hua et al., "Convolutional Network with Densely Backward Attention for Facial Expression Recognition", IMCOM, Jan 3-5, 2020.
[Hua, 2020c] C.-H. Hua et al., "DRAN: Densely Reversed Attention based Convolutional Network for Diabetic Retinopathy Detection", IEEE EMBC, July 20-24, 2020.
[Hua, 2021] C.-H. Hua et al., "Convolutional Network with Twofold Feature Augmentation for Diabetic Retinopathy Recognition from Multi-modal Images", IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2686-2697, July 2021.

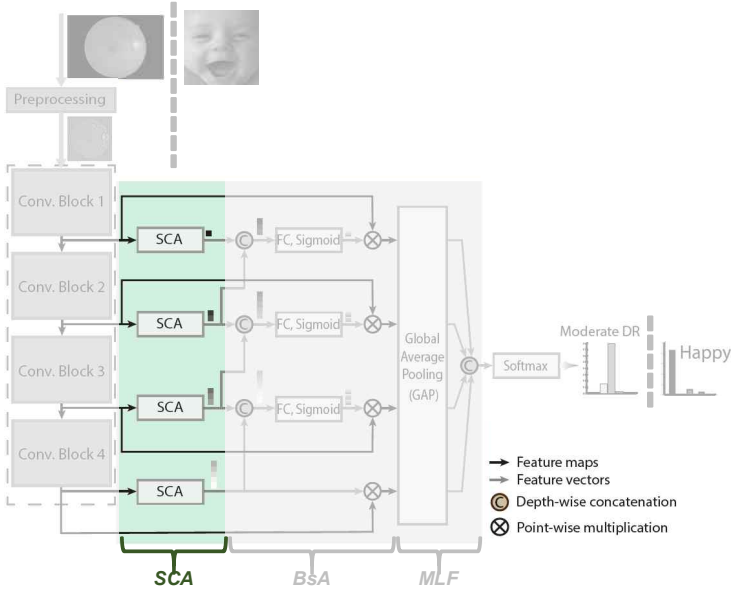


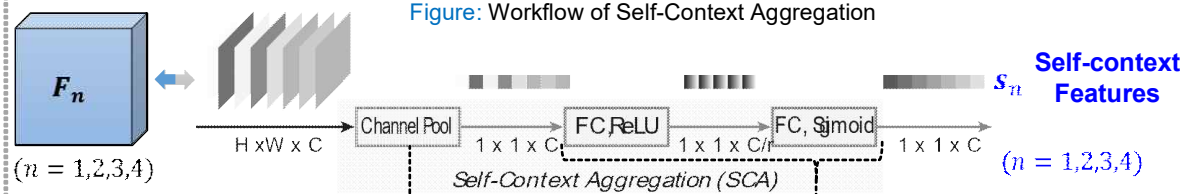
Figure: sCAB-Net for **Diabetic Retinopathy (DR) Severity Recognition** [Hua, 2020a], [Hua, 2021] or **Facial Expression Recognition** [Hua, 2020c]

Channel-wisely Cross-Attentional Scheme

1. Self-Context Aggregation (SCA)

→ **Extract Channel-wisely Attentional features at multiple levels**

Figure: Workflow of Self-Context Aggregation



Compute spatial average (pooled) values at each channel of feature map

Learn channel-wisely attentional features (i.e., express cross-channel discriminative levels of pooled features)

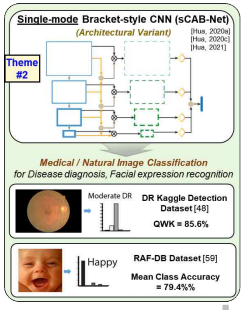
$$G_d(\boldsymbol{F}_n) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \boldsymbol{F}_{n\,h,w,d}$$

$$\boldsymbol{g}_n = \left[G_1(\boldsymbol{F}_n), \ldots, G_d(\boldsymbol{F}_n), \ldots, G_{C_n}(\boldsymbol{F}_n) \right]^T$$

$$\boldsymbol{s}_n = \sigma\left(\boldsymbol{W}_{fc2} \left(ReLU\left(\boldsymbol{W}_{fc1} \boldsymbol{g}_n + \boldsymbol{B}_{fc1} \right) \right) + \boldsymbol{B}_{fc2} \right)$$

$\boldsymbol{F}_n = n^{th}$ feature map having channel dimension of C_n where $n = 1,2,3,4$
(h, w, d): pixel coordinates of feature map $\boldsymbol{F}_n$
$G_d(\boldsymbol{F}_n)$: Channel Pool operator at d^{th} channel of feature map $\boldsymbol{F}_n$
$\boldsymbol{g}_n$: Channel pooled feature of $\boldsymbol{F}_n$
$r = 16$: compression rate (=16) for saving computation cost
$\boldsymbol{W}_{fc1} \in R^{\frac{C}{r} \times C}, \boldsymbol{B}_{fc1} \in R^{\frac{C}{r}}$: weights and biases of 1st Fully-connected (FC) layer
$\boldsymbol{W}_{fc2} \in R^{C \times \frac{C}{r}}, \boldsymbol{B}_{fc2} \in R^{C}$: weights and biases of 2nd Fully-connected (FC) layer
ReLU: Rectified Linear Unit activation function
σ: Sigmoid activation function

Proposed Approach: Theme #2 – Detailed Architecture (3/4)



[Hua, 2020a] C.-H. Hua et al., "Convolutional Network with Densely Backward Attention for Facial Expression Recognition", IMCOM, Jan 3-5, 2020.
[Hua, 2020c] C.-H. Hua et al., "DRAN: Densely Reversed Attention based Convolutional Network for Diabetic Retinopathy Detection", IEEE EMBC, July 20-24, 2020.
[Hua, 2021] C.-H. Hua et al., "Convolutional Network with Twofold Feature Augmentation for Diabetic Retinopathy Recognition from Multi-modal Images", IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2686-2697, July 2021.

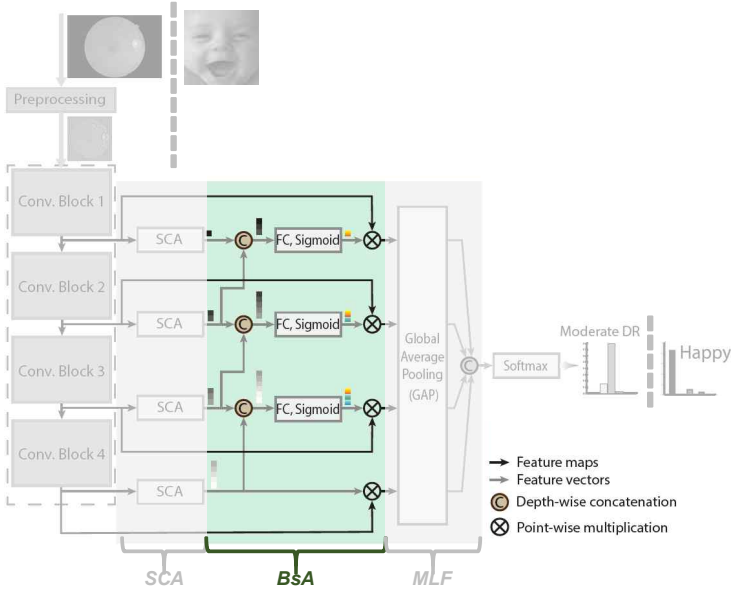


Figure: sCAB-Net for **Diabetic Retinopathy (DR) Severity Recognition** [Hua, 2020a], [Hua, 2021] or **Facial Expression Recognition** [Hua, 2020c]

Channel-wisely Cross-Attentional Scheme

2. Bracket-style Attention (BsA)

→ **Integrate semantically-richer representations of deeper features into the shallower counterparts in Bracket-style manner**

Figure: Workflow of Bracket-style Attention

Combine the adjacent self-context features in **pairwise** manner → Cross-level Concatenation

$$s_{1\sqcup 2} = Concat(s_1, s_2)$$
$$s_{2\sqcup 3} = Concat(s_2, s_3)$$
$$s_{3\sqcup 4} = Concat(s_3, s_4)$$

Manage the integration of semantically richer self-context into adjacently lower-level version → FC, Sigmoid

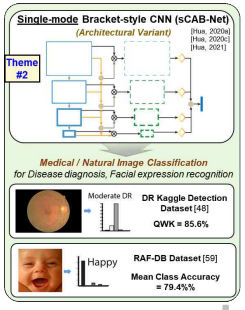
$$s_{12} = \sigma(W_{fc_1} s_{1\sqcup 2} + B_{fc_1})$$
$$s_{23} = \sigma(W_{fc_2} s_{2\sqcup 3} + B_{fc_2})$$
$$s_{34} = \sigma(W_{fc_3} s_{3\sqcup 4} + B_{fc_3})$$

Re-calibrate the multi-scale feature maps from backbone network via point-wise multiplication with cross-context features → Attentional Re-calibration

$$F_{bsa_1} = \left\{ F_{1_{:,:,\delta}} \otimes (s_{12})_\delta \middle| \delta = 1, \ldots, C_1 \right\}$$
$$F_{bsa_2} = \left\{ F_{2_{:,:,\delta}} \otimes (s_{23})_\delta \middle| \delta = 1, \ldots, C_2 \right\}$$
$$F_{bsa_3} = \left\{ F_{3_{:,:,\delta}} \otimes (s_{34})_\delta \middle| \delta = 1, \ldots, C_3 \right\}$$
$$F_{bsa_4} = \left\{ F_{4_{:,:,\delta}} \otimes (s_4)_\delta \middle| \delta = 1, \ldots, C_4 \right\}$$

$s_n = n^{th}$ self-context feature where $n = 1,2,3,4$
$Concat$: Depth-wise concatenation
$W_{fc_n} \in R^{C_n \times (C_n + C_{n+1})}, B_{fc_n} \in R^{C_n}$: weights and biases of Fully-connected (FC) layer that learns $s_{n\sqcup(n+1)}$
σ: Sigmoid activation function
$F_{n_{:,:,\delta}}$: δ-th channel of the feature map F_n with its full spatial dimension
$(s_n)_\delta$: δ-th entry (length) of the context-based feature vector s_n
$\otimes$: Point-wise multiplication

Proposed Approach: Theme #2 – Detailed Architecture (4/4)



[Hua, 2020a] C.-H. Hua et al., "Convolutional Network with Densely Backward Attention for Facial Expression Recognition", IMCOM, Jan 3-5, 2020.
[Hua, 2020c] C.-H. Hua et al., "DRAN: Densely Reversed Attention based Convolutional Network for Diabetic Retinopathy Detection", IEEE EMBC, July 20-24, 2020.
[Hua, 2021] C.-H. Hua et al., "Convolutional Network with Twofold Feature Augmentation for Diabetic Retinopathy Recognition from Multi-modal Images", IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2686-2697, July 2021.

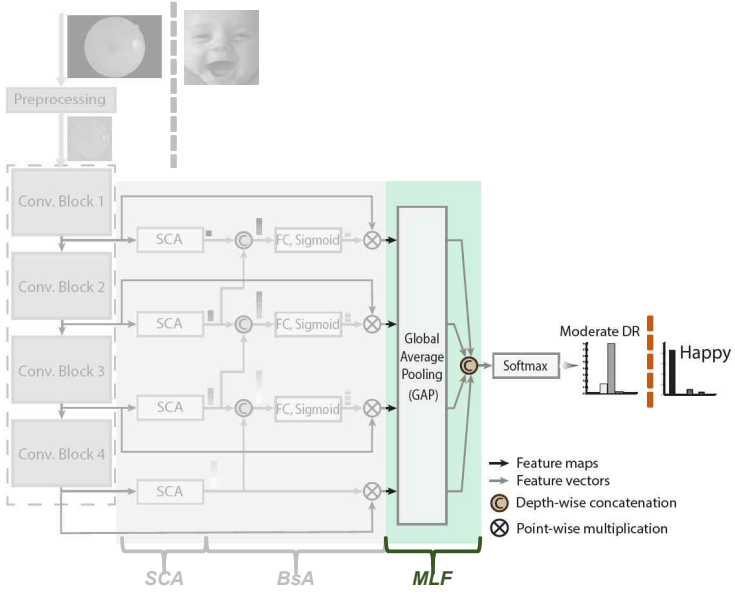


Figure: sCAB-Net for **Diabetic Retinopathy (DR) Severity Recognition** [Hua, 2020a], [Hua, 2021] or **Facial Expression Recognition** [Hua, 2020c]

Channel-wisely Cross-Attentional Scheme

3. Multi-level Fusion (MLF)

→ **Gain the mixture of multi-level context**

Figure: Workflow of Multi-level Fusion

Compute spatial average (pooled) values at each channel of feature map

Global Average (Channel) Pooling

$$G_d(\boldsymbol{F}_{bsa_n}) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \boldsymbol{F}_{bsa_n}{}_{h,w,d}$$

$$\boldsymbol{g_n} = \left[G_1(\boldsymbol{F}_{bsa_n}), \ldots, G_{C_n}(\boldsymbol{F}_{bsa_n})\right]^T$$

$$(n = 1,2,3,4)$$

Aggregate Bracket-style Attentional feature maps of different semantic levels

All-level Concatenation

$$\boldsymbol{F_{final}} = Concat(\boldsymbol{g_1}, \boldsymbol{g_2}, \boldsymbol{g_3}, \boldsymbol{g_4})$$

$$(\boldsymbol{F_{final}} \in R^{class})$$

Assign the label (class) having highest score along the channel dimension of $\hat{\boldsymbol{F}}_{\boldsymbol{final}}$ as final prediction

Prediction (using Softmax function)

$$\hat{\boldsymbol{F}}_{\boldsymbol{final}} = Softmax(\boldsymbol{F_{final}})$$

$$= \left\{\hat{\boldsymbol{F}}_{\boldsymbol{final}_i} = \frac{e^{F_{final_i}}}{\sum_{j=1}^{Class} e^{F_{final_j}}} \middle| i = 1, \ldots, Class\right\}$$

$F_{bsa_n} = n^{th}$ Bracket-style Attentional feature map having channel dimension of C_n where $n = 1,2,3,4$
(h, w, d): pixel coordinates of feature map F_{bsa_n}
$G_d(F_{bsa_n})$: Global Average (Channel) Pool operator at d^{th} channel of feature F_{bsa_n}
g_n: Global Average (Channel) pooled feature of Bracket-style Attentional feature F_{bsa_n}
$Concat$: Depth-wise concatenation

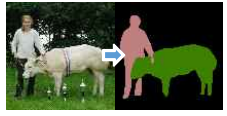
Experiments: Benchmark Datasets

Theme #1-1

Natural Image Segmentation

Common object

PASCAL VOC 2012 Dataset [20]



No. training images: 10,582
No. validation images: 1,449
No. testing images: 1,456
Input size: 513 x 513
No. semantic classes: 20

(i.e., groups of person, animal, vehicle, and indoor context.)

Street scenes

Cityscapes Dataset [18]

No. training images: 2,975
No. validation images: 500
No. testing images: 1,525
Input size: 768 x 768
No. semantic classes: 19

(i.e., road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, bicycle.)

CAMVID Dataset [9]

No. training images: 367
No. validation images: 101
No. testing images: 233
Input size: 360 x 480
No. semantic classes: 12

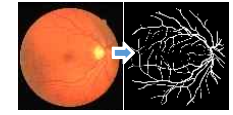
(i.e., building, tree, sky, car, sign-symbol, road, pedestrian, fence, pole, sidewalk, bicyclist, background.)

Theme #1-2

Medical (Fundus) Image Segmentation

Retinal blood vessel

DRIVE Dataset [80]



No. training images: 20
No. testing images: 20
Input size: 584 x 565
No. semantic classes: 2

(i.e., retinal blood vessel and background.)

Theme #2

Natural Image Classification

Facial Expression Recognition

RAF-DB Dataset [59]

No. training images: 12,271
No. testing images: 3,068
Input size: 100 x 100
No. output classes: 7 (i.e., angry, disgust, fearful, happy, neutral, sad, surprise emotions.)

Medical (Fundus) Image Classification

Diabetic Retinopathy (DR) Severity Recognition

Kaggle DR Detection Dataset [48]

No. training images: 35,000
No. validation images: 11,000
No. testing images: 43,000
Input size: 448 x 448
No. output classes: 5 (i.e, No DR, mild Nonproliferative DR (NPDR), moderate NPDR, severe NPDR, proliferative DR.)

Experiments: Evaluation Metrics

Theme #1-1 — **Mean Intersection of Union**

$$mIoU = \frac{1}{L}\sum_{x=1}^{L}\frac{p_{xx}}{\sum_{y=1}^{L} p_{xy} + \sum_{y=1}^{L} p_{yx} - p_{xx}}$$

(Ratio between the intersection and union of 2 sets, i.e., ground truth and the prediction)

PASCAL VOC 2012 Dataset [20]
Cityscapes Dataset [18]
CAMVID Dataset [9]

p_{xy}: pixels having ground-truth label x is predicted as label y
L: total number of labels

Theme #1-2

DRIVE Dataset [80]

Sensitivity $Sen = TP/(TP + FN)$

Specificity $Spe = TN/(TN + FP)$

Accuracy $Acc = (TP + TN)/(TP + TN + FP + FN)$

AUROC (Area Under the Receiver Operating Characteristic curve)

Where ROC curve shows the trade-off between TP rate and FP rate across different decision thresholds.

Ground-truth \ Prediction	0	1
0	True Negative (TN)	False Positive (FP)
1	False Negative (FN)	True Positive (TP)

Table: Confusion matrix of binary classification

Theme #2

RAF-DB Dataset [59] — **Mean Class Accuracy**

$$mCA = \frac{1}{L}\sum_{x=1}^{L}\frac{I_{xx}}{\sum_{y=1}^{L} I_{xy}}$$

(Ratio of correct predictions in per-label basis and then averaged over all labels)

I_{xy}: Images having ground-truth label x is predicted as label y
L: total number of labels

Kaggle DR Detection Dataset [48] — **Quadratic Weighted Kappa**

$$QWK = 1 - \frac{\sum_{x=1}^{L}\sum_{y=1}^{L} W_{x,y} O_{x,y}}{\sum_{x=1}^{L}\sum_{y=1}^{L} W_{x,y} E_{x,y}}$$

(Agreement degree of classification results between two raters, i.e., the group of grading experts and the prediction of learning model)

$W_{x,y}$: weighting matrix showing penalty of difference between the predictions (in x indices) and corresponding ground-truth labels (in y indices)
$O_{x,y}$: observed confusion matrix computed from the classifier's results
$E_{x,y}$: expected matrix inferred by the outer product between the L-length ground-truth and prediction vectors, which carry occurrences of counted predicted and actual labels
L: total number of labels

Experiments: Training Configurations

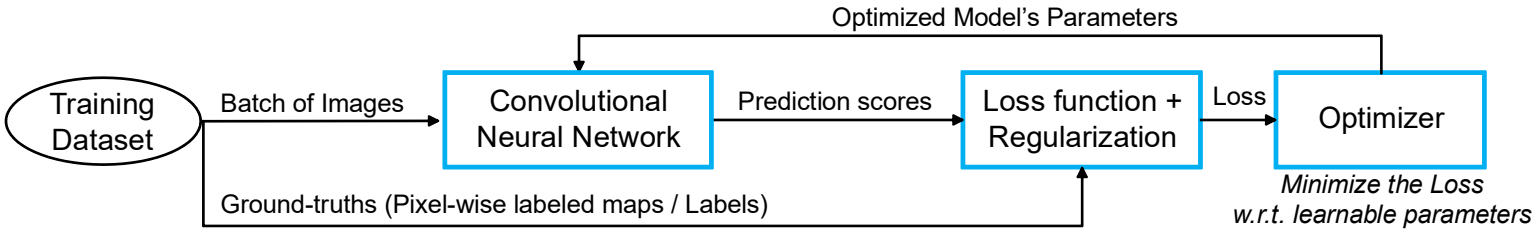
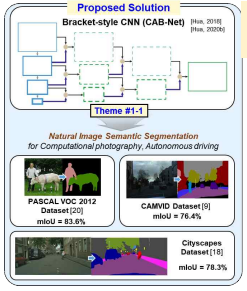


Figure: Abstract workflow of training a CNN for image semantic segmentation or classification

Training Configurations

- **Initialization of parameters in Convolutional Neural Network (CNN):**
  - **Backbone CNN:** Those pretrained with ImageNet [76]
  - **Bracket-style Network:** *He* Initialization [30]
- **Loss function:** Softmax (cross-entropy)
- **Regularization:** Weight decay with coefficient of $1e-5$ (for Themes #1-1 & #1-2) or $5e-4$ (for Theme #2)
- **Optimizer:**
  - Gradient Descent algorithm with momentum of 0.9
  - Initial learning rate $\alpha_0 = 0.01$ (for Themes #1-1 & #1-2) or 0.005 (for Theme #2)
  - Learning rate decay schedule $\alpha_i = \alpha_0 \left(1 - \frac{i}{I}\right)^{0.9}$ (where α_i is learning rate at i^{th} training iteration given $0 \leq i \leq I$)

Experiments: Ablation Study (1/4)



Theme #1-1:AS1 Contribution of backbone CNN to final segmentation performance

Table: **mIoU (%)** on **Pascal VOC 2012 [20]** validation set and number of parameters with **various strategies of attentional mechanism**

Backbone CNN	Depth sizes {d1, d2, d3, d4}	mIoU (%)	No. parameters		
			Backbone	Bracket	Total
VGG-16 [79]	{128, 256, 512, 512}	75.24	14.72M	7.13M	21.85M
Xception-65 [17]	{128, 256, 728, 2048}	77.96	20.81M	21.06M	41.87M
ResNet-50 [29]	{256, 512, 1024, 2048}	78.27	23.51M	38.97M	62.48M
ResNet-101 [29]	{256, 512, 1024, 2048}	80.37	42.50M	38.97M	81.47M

❖ **Backbone Convolutional Neural Network (CNN)**

- Extract multi-scale representational feature maps.

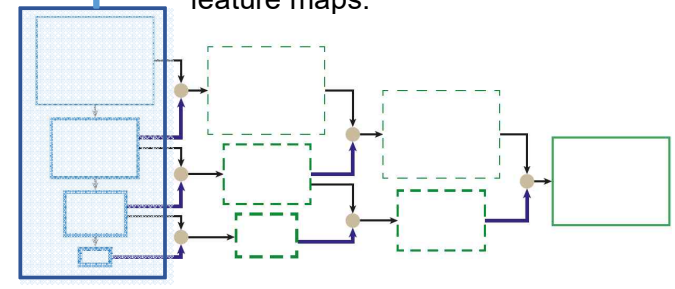
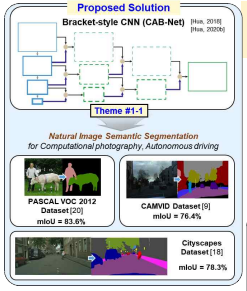


Figure: Concept of the proposed Bracket-style CNN (CAB-Net)

- Deeper architectures attain better mIoU performance (up to ~5.3% ResNet-101 vs. VGG-16).
- Model complexity is enlarged (>1.5x) due to increment of backbone CNN's capacity and depth sizes of feature maps involved in the Bracket-style decoding stage.
- Compared to Xception-65, ResNet-50 is slightly better while ResNet-101 outperforms by 2.41%.

→ Depth-wisely representational abilities of involved features (via depth sizes d1, …, d4) strongly impact on the final segmentation performance.

Experiments: Ablation Study (2/4)



Theme #1-1:AS2 Effectiveness of Bracket-style over Ladder/U-shaped Decoding Network

Table: **mIoU (%)** on **Pascal VOC 2012 [20]** validation set and number of parameters with **Bracket-style vs. Ladder/U-shaped** decoding network

Feature Combination Strategy	mIoU (%)	No. parameters
Ladder/U-shaped	77.30	72.81M
Bracket-style	80.37	81.47M

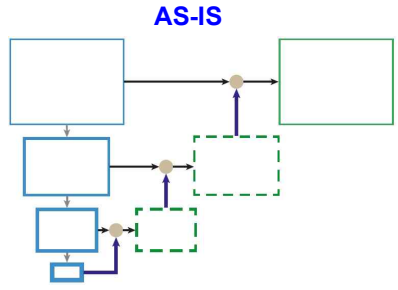


Figure: Ladder/U-shaped Feature Combination (Symmetrically-structured network)

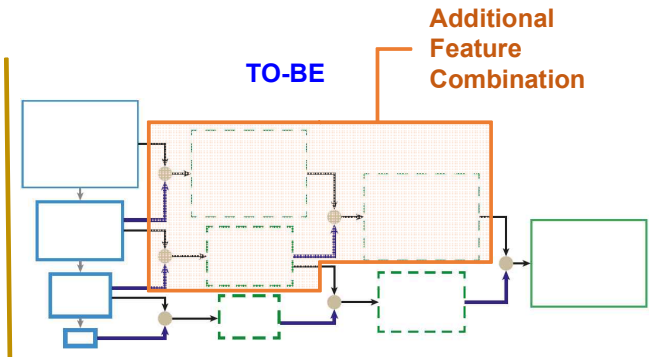


Figure: **Bracket-style** Feature Combination (Asymmetrically-structured network)

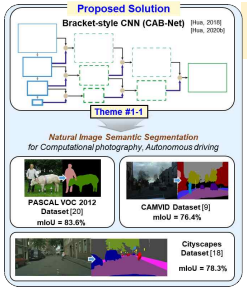
- Performance improved by the utilization of Bracket-style feature combination is considerable: ↑3.07%.

→ Continual and extensive utilization of middle-level features in Bracket-structured manner along the decoding process brings in better segmentation performance

- Additional feature combination modules involved in the proposed Bracket-style network results in an increment of No. parameters by ~11.89%.

mIoU: mean Intersection of Union

Experiments: Ablation Study (3/4)



Theme #1-1:AS3 Coordination between Bracket-shaped Network and CAF-based Connections

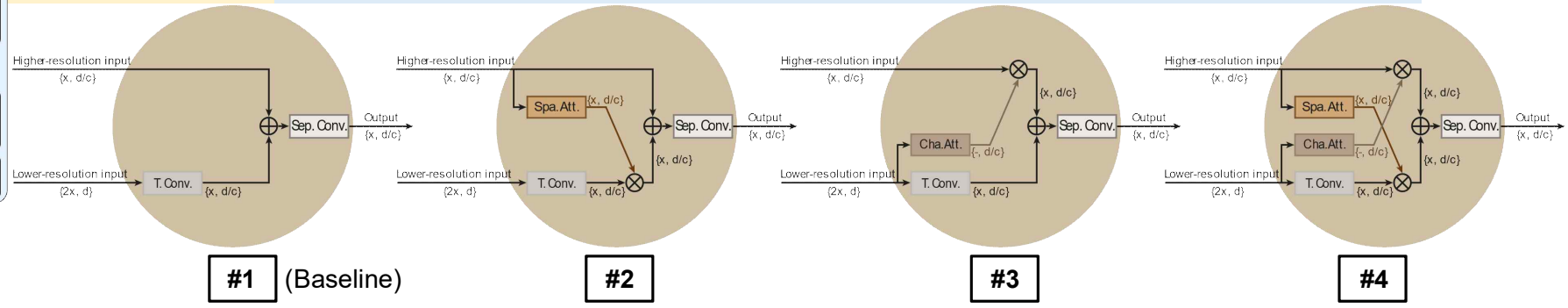


Figure: Various strategies of attentional mechanism

Table: **mIoU (%)** on **Pascal VOC 2012 [20]** validation set and number of parameters with **various strategies of attentional mechanism**

#	Strategy		mIoU (%)	No. parameters
	Cha. Att.	Spa. Att.		
1			76.73	33.66M
2		√	77.86	33.66M
3	√		79.45	38.97M
4	√	√	80.37	38.97M

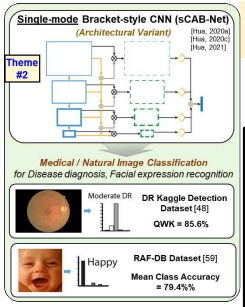
CAF: Cross-Attentional Fusion
Cha. Att.: Channel-wise Attentional Block
Spa. Att.: Spatially Attentional Block
mIoU: mean Intersection of Union

- Performance improved by the involvement of attentional mechanism is considerable: 1.13% (spatial-based vs. baseline) & 2.72% (channel-based vs. baseline).
- With cross-attentional strategy, mIoU is further elevated by ~1.0%.

→ Powerful coordination between Bracket-structured network and CAF-based connections.

- Additional operations in Spa. Att. have nearly no impact on the model complexity.
- Those in Cha. Att. increases No. parameters by ~15.8% due to the dependence of hidden nodes' amount in Fully Connected layer on high-level feature's channel size.

Experiments: Ablation Study (4/4)[1/3]



Theme #2:AS1 Effectiveness of Channel-wisely Cross-Attentional (CCA) Stream for Classification-based CNN

Table: **Mean Class Accuracy (%)** on **RAF-DB [59]** test set; **Quadratic Weighted Kappa (%)** on **Kaggle DR Detection [48]** test set; and number of parameters with **various strategies of backbone CNN and attentional mechanism**

Backbone CNN	Strategy: Baseline	Strategy: SCA+MLF	Strategy: SCA+BsA+MLF (i.e., CCA)	Mean Class Accuracy (%)	Quadratic Weighted Kappa (QWK) (%)	No. Parameters
VGG-16 [79]	√			74.96	84.9	134.30M
		√		77.35	85.4	14.81M
			√	78.81	86.3	15.59M
ResNet-101 [29]	√			77.10	85.4	42.51M
		√		77.48	86.1	43.23M
			√	79.33	86.7	47.36M
DenseNet-161 [38]	√			77.21	85.5	26.49M
		√		77.75	86.5	27.78M
			√	79.37	86.9	39.56M

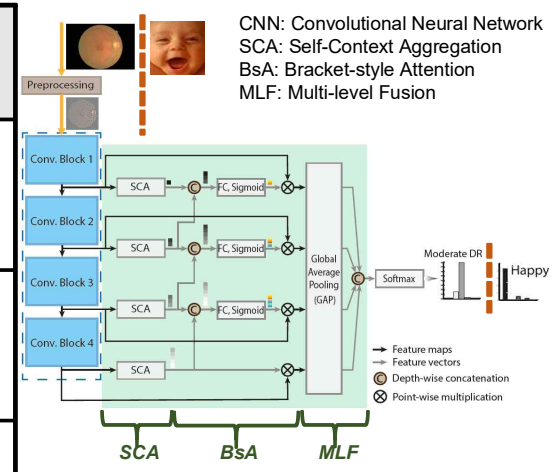
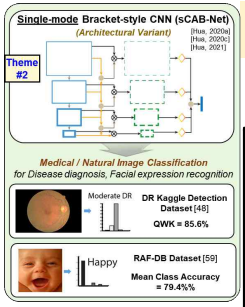


Figure: sCAB-Net Architecture

- **Generally:** Deeper architectures attain better recognition performance
- **Model Complexity** in terms of **No. parameters**: → Increment is majorly caused by BsA
  - SCA+MLF vs. Baseline: ↑1.69% for ResNet-101; ↑4.87% for DenseNet-161.
  - SCA+BsA+MLF vs. SCA+MLF: ↑9.55% for ResNet-101; ↑42.40% for DenseNet-161.

Note: The proposed CCA strategy with backbone VGG-16 reduces the number of parameters by approximately 88.4% because of **not** involving expensive Fully Connected layers at the end of the baseline network.

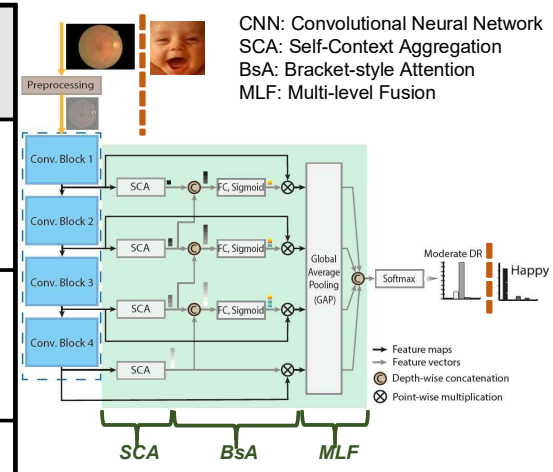
Experiments: Ablation Study (4/4)[2/3]



Theme #2:AS1 Effectiveness of Channel-wisely Cross-Attentional (CCA) Stream for Classification-based CNN

Table: **Mean Class Accuracy (%)** on **RAF-DB [59]** test set; **Quadratic Weighted Kappa (%)** on **Kaggle DR Detection [48]** test set; and number of parameters with **various strategies of backbone CNN and attentional mechanism**

Backbone CNN	Strategy			Mean Class Accuracy (%)	Quadratic Weighted Kappa (QWK) (%)	No. Parameters
	Baseline	SCA+MLF	SCA+BsA+MLF (i.e., CCA)			
VGG-16 [79]	√			74.96	84.9	134.30M
		√		77.35	85.4	14.81M
			√	78.81	86.3	15.59M
ResNet-101 [29]	√			77.10	85.4	42.51M
		√		77.48	86.1	43.23M
			√	79.33	86.7	47.36M
DenseNet-161 [38]	√			77.21	85.5	26.49M
		√		77.75	86.5	27.78M
			√	79.37	86.9	39.56M



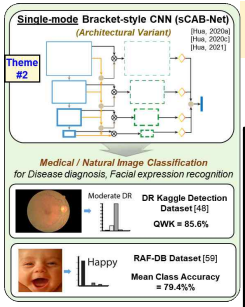
CNN: Convolutional Neural Network
SCA: Self-Context Aggregation
BsA: Bracket-style Attention
MLF: Multi-level Fusion

Figure: sCAB-Net Architecture

- **Mean Class Accuracy** on **RAF-DB** dataset:
  - 1a. SCA+MLF and CCA (i.e., SCA+BsA+MLF) outperform Baseline by 0.38-2.39% and 2.16-3.85% for all backbone networks.
- **QWK** on **Kaggle DR Detection** dataset:
  - 1b. SCA+MLF and CCA (i.e., SCA+BsA+MLF) outperform Baseline by 0.5-1.0% and 1.3-1.4% for all backbone networks.

→ Engagement of attention scheme at multi-scale features and subsequent depth-wise aggregation of corresponding outcomes are plausible for image classification in these specialized domains.

Experiments: Ablation Study (4/4)[3/3]



Theme #2:AS1 Effectiveness of Channel-wisely Cross-Attentional (CCA) Stream for Classification-based CNN

Table: **Mean Class Accuracy (%)** on **RAF-DB [59]** test set; **Quadratic Weighted Kappa (%)** on **Kaggle DR Detection [48]** test set; and number of parameters with **various strategies of backbone CNN and attentional mechanism**

Backbone CNN	Strategy: Baseline	Strategy: SCA+MLF	Strategy: SCA+BsA+MLF (i.e., CCA)	Mean Class Accuracy (%)	Quadratic Weighted Kappa (QWK) (%)	No. Parameters
VGG-16 [79]	√			74.96	84.9	134.30M
		√		77.35	85.4	14.81M
			√	78.81	86.3	15.59M
ResNet-101 [29]	√			77.10	85.4	42.51M
		√		77.48	86.1	43.23M
			√	79.33	86.7	47.36M
DenseNet-161 [38]	√			77.21	85.5	26.49M
		√		77.75	86.5	27.78M
			√	79.37	86.9	39.56M

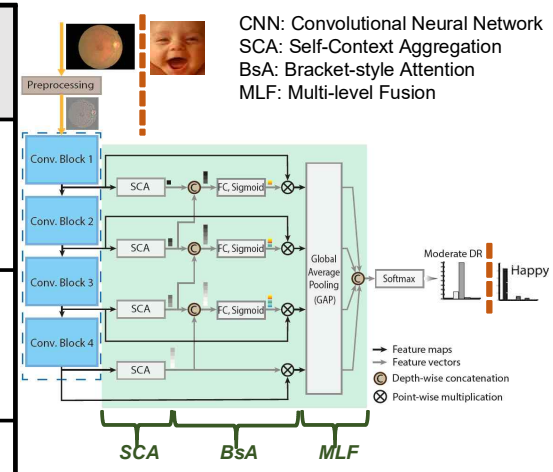
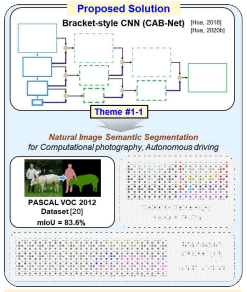


Figure: sCAB-Net Architecture

- ~~**Mean Class Accuracy** on **RAF-DB** dataset:~~
  - 2a. BsA embedded between SCA and MLF further improves 1.46-1.62%.
- **QWK** on **Kaggle DR Detection** dataset:
  - 2b. BsA embedded between SCA and MLF further improves 0.4-0.9%.

→ Advantage of integrating higher-level attentional context to recalibrate lower-level features for leveraging their contributions of structural representations to the softmax classifier.

Experiments: Comparison Results – Theme #1-1 (1/3)



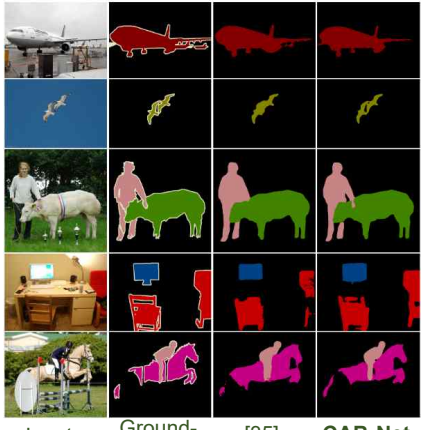
Theme #1-1

Approach (Structure type)	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU (%)
FCN [67] (A)	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
BNetVGG-LCM [35] (A)	92.0	42.9	92.3	73.3	77.5	91.4	86.4	91.5	42.7	81.9	61.6	84.4	85.8	88.4	90.1	65.5	86.4	60.0	86.1	72.5	78.5
G-FRNet [43] (S)	91.4	44.6	91.4	69.2	78.2	95.4	88.9	93.3	37.0	89.7	61.4	90.0	91.4	87.9	87.2	63.8	89.4	59.9	87.0	74.1	79.3
DDSC [8] (S)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	81.2
WideResNet [95] (S)	94.4	72.9	94.9	68.8	78.4	90.6	90.0	92.1	40.1	90.4	71.7	89.9	93.7	**91.0**	89.1	71.3	90.7	61.3	87.7	**78.1**	82.5
PSPNet [103] (A)	91.8	71.9	94.7	71.2	75.8	95.2	89.9	**95.9**	39.3	90.7	71.7	90.5	94.5	88.8	89.6	**72.8**	89.6	64.0	85.1	76.3	82.6
DANet [23] (A)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	82.6
DFN [97] (A)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	82.7
EncNet [101] (A)	94.1	69.2	**96.3**	**76.7**	**86.2**	96.3	**90.7**	94.2	38.8	90.7	73.3	90.0	92.5	88.8	87.9	68.7	**92.6**	59.0	86.4	73.4	82.9
TKCN [94] (A)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	83.2
CAB-Net (A)	**96.0**	**75.6**	94.3	69.1	79.9	**97.1**	89.8	94.8	**40.4**	**91.2**	**74.6**	89.4	**94.7**	87.2	**91.7**	69.6	92.1	**65.5**	**88.8**	76.9	**83.6**

Table: **mIoU** on test set of **PASCAL VOC 2012 Dataset** [20]

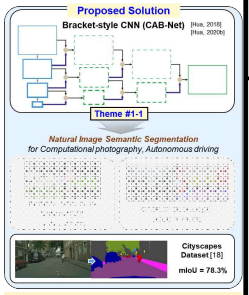
Boldface values indicate the best performance at each criterion
(A): Asymmetrically-structured Network
(S): Symmetrically-structured Network

Figure:
Qualitative results



- The proposed approach achieves competitive mIoU of 83.6% compared with that of the state-of-the-art methods.
- Regarding the class-wise results, the CAB-Net attains the top performance with significant margin (up to 3.7%) for 10/20 semantic objects (*aeroplane, bike, bus, chair, cow, table, horse, person, sofa, train*) ranging from small to large scale.
- Besides that, several qualitative results of the proposed method are demonstrated in the Figure to show its effectiveness in semantic segmentation.

Experiments: Comparison Results – Theme #1-1 (2/3)



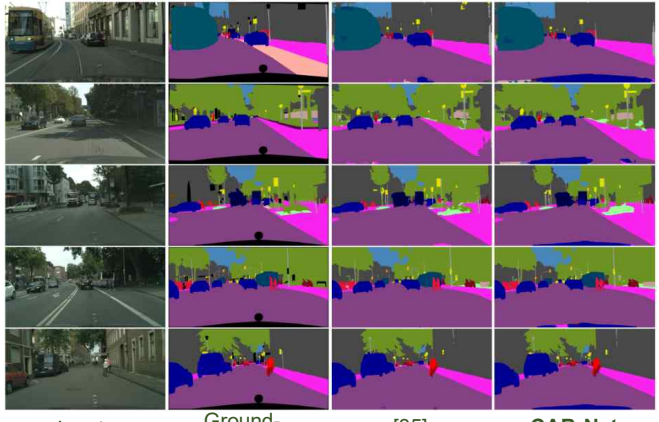
Theme #1-1

Approach (Structure type)	road	swalk	build.	wall	fence	pole	tlight	sign	veg.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU (%)
SegNet [6] (S)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	56.1
FSSNet [102] (A)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	58.8
FCN [67] (A)	97.4	78.4	89.2	34.9	44.2	47.4	60.1	65.0	91.4	69.3	93.9	77.1	51.4	92.6	35.3	48.6	46.5	51.6	66.8	65.3
DeepLab-CRF [12] (A)	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4	94.2	79.8	59.8	93.7	56.5	67.5	57.5	57.7	68.8	70.4
RefineNet [63] (S)	98.2	83.3	91.3	47.8	50.4	56.1	66.9	71.3	92.3	70.3	94.8	80.9	63.3	94.5	64.6	76.1	64.3	62.2	70.0	73.6
BiSeNet [98] (A)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	74.7
SwiftNetRN-18 [72] (S)	98.3	83.9	92.2	46.3	52.8	63.2	70.6	75.8	93.1	70.3	95.4	84.0	64.5	95.3	63.9	78.0	71.9	61.6	73.6	75.5
BNetVGG-LCM [35] (A)	98.4	84.8	92.4	55.1	55.5	62.1	71.7	76.3	93.3	71.4	95.0	85.1	67.9	95.6	60.5	72.0	62.4	67.3	74.9	75.9
DUC-HDC [89] (A)	98.5	85.5	92.8	**58.6**	55.5	**65.0**	73.5	77.9	93.3	72.0	95.2	84.8	68.5	95.4	70.9	78.8	68.7	65.9	73.8	77.6
PSPNet [103] (A)	**98.6**	**86.2**	**92.9**	50.8	58.8	64.0	**75.6**	**79.0**	**93.4**	**72.3**	**95.4**	**86.5**	**71.3**	**95.9**	68.2	79.5	**73.8**	69.5	**77.2**	**78.4**
CAB-Net (A)	98.5	85.4	92.8	55.6	**59.1**	63.3	70.9	75.6	**93.4**	71.1	95.2	86.4	**71.3**	**95.9**	**72.3**	**82.2**	72.3	**70.4**	76.5	78.3

Table: **mIoU** on test set of **Cityscapes Dataset** [18]

Boldface values indicate the best performance at each criterion
(A): Asymmetrically-structured Network
(S): Symmetrically-structured Network

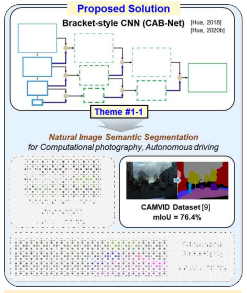
Figure:
Qualitative results



Input | Ground-truth | [35] | **CAB-Net**

- The proposed approach achieves competitive mIoU of 78.3% compared with that of the state-of-the-art methods.
- Regarding the class-wise results, the CAB-Net attains the superior performance (up to 2.7%) in segmenting 7/19 semantic objects (*fence, vegetation, rider, car, truck, bus,* and *motorbike*) over the compared methods.
  - The performance of remaining categories, except for small-scale *traffic light* and *sign symbol,* has average lower IoU of ~0.6% compared with those of the state-of-the-art PSPNet [103].
- Besides that, several qualitative results of the proposed method are demonstrated in the Figure to show its effectiveness in semantic segmentation.

Experiments: Comparison Results – Theme #1-1 (3/3)



Theme #1-1

Approach (Structure type)	building	tree	sky	car	sign	road	pedestrian	fence	pole	sidewalk	bicyclist	mIoU (%)
SegNet [6] (S)	-	-	-	-	-	-	-	-	-	-	-	60.1
DeepLab-LFOV [14] (A)	81.5	74.6	89.0	82.2	42.3	92.2	48.4	27.2	14.3	75.4	50.1	61.6
Dilation8 [99] (A)	82.6	76.2	89.9	84.0	46.9	92.2	56.3	35.8	23.4	75.3	55.5	65.3
Dilation+FSO-DF [54] (A)	84.0	77.2	91.3	85.6	49.9	92.5	59.1	37.6	16.9	76.0	57.2	66.1
BNetVGG-LCM [35] (A)	81.4	75.3	92.8	82.5	42.8	89.2	60.8	47.8	**36.3**	66.4	54.8	66.4
G-FRNet [43] (S)	82.5	76.8	92.1	81.8	43.0	94.5	54.6	47.1	33.4	82.3	59.4	68.0
BiSeNet [98] (A)	83.0	75.8	92.0	83.7	46.5	94.6	58.8	53.6	31.9	81.4	54.0	68.7
DDSC [8] (S)	-	-	-	-	-	-	-	-	-	-	-	70.9
LDN121 16→2 [53] (S)	-	-	-	-	-	-	-	-	-	-	-	75.8
CAB-Net	**91.1**	**88.9**	**95.7**	**93.0**	**64.8**	**94.7**	**66.5**	**70.5**	29.8	**85.3**	**60.3**	**76.4**

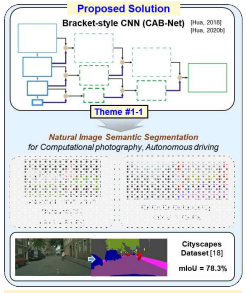
Table: **mIoU** on test set of **CAMVID Dataset** [9]

Boldface values indicate the best performance at each criterion
(A): Asymmetrically-structured Network
(S): Symmetrically-structured Network

Figure: Qualitative results

- The proposed approach achieves state-of-the-art mIoU of 76.4% compared with that of the state-of-the-art methods.
- Regarding the class-wise results, the CAB-Net attains the top performance in 10/11 semantic objects (only except for *pole* category) with significant margins to the second places (0.1-16.9%).
- Besides that, several qualitative results of the proposed method are demonstrated in the Figure to show its effectiveness in semantic segmentation.

Experiments: Computational Complexity Comparisons



Theme #1-1

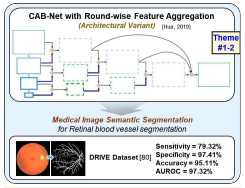
Network Structure	Typical Approach	GPU	mIoU (%)	Inference speed (fps)	No. parameters
Symmetric	SegNet [6]	Titan X	56.1	24	29.46M
	SwiftnetRN-18 [72]	GTX 1080Ti	75.5	39	11.80M
Asymmetric	PSPNet [103]	GTX 1080Ti	**78.4**	11	65.60M
	BiSeNet [98]	Titan Xp	74.7	**65.5**	49.00M
Bracket	B-Net-VGG-LCM [35]	GTX 1080Ti	75.9	27	25.92M
	CAB-Net	GTX 1080Ti	78.3	20	81.47M

Boldface values indicate the best performance at each criterion

Table: Comparison of **mIoU (%)**, **inference speed (frames per second – fps)**, and **number of model parameters** for an input image with resolution of 1024×2048 in **Cityscapes Dataset** [18]

- Comparison with typical *symmetrically-structured* networks:
  - Both SegNet [6] and SwiftnetRN-18 [72] have faster segmentation speeds of 4 and 19 fps than CAB-Net due to the employment of much lower-capacity CNNs, i.e., VGG-16 [79] and ResNet-18 [29], respectively.
  - Meanwhile, CAB-Net greatly outperforms SegNet [6] by mIoU of 22.2% and SwiftNetRN-18 [72] by mIoU of 2.8%.
  - B-Net-VGG-LCM [35], another representative of Bracket-style structure, attains higher mIoU (↑19.8%) and processing rate (↑3 fps) while having same backbone CNN but fewer parameters (↓12%) in comparison with SegNet [6].
- Comparison with typical *asymmetrically-structured* networks:
  - CAB-Net reaches comparable mIoU (with trivially ↓0.1%) and nearly double speed (20 vs. 11 fps) in comparison with PSPNet [103].
    - ❖ Manifold concatenation of the deepest feature maps in ResNet-101 [29] for various pooling rates followed by conventional convolutional layers in PSPNet [103] → heavily elaborates the volume of operations (comprising multiply, add, max-value calculations), which subsequently reduces inference speed.
  - CAB-Net reaches higher mIoU (↑3.6%) but much lower inference speed (20 vs. 65.5 fps) in comparison with BiSeNet [98].
    - ❖ BiSeNet [98] targets at processing rapidity more favorably, it is built upon the lightweight backbone ResNet-18 [72] with an attached dual network stream for amalgamating global context and local details in a cost-efficient way.

Experiments: Comparison Results – Theme #1-2



Theme #1-2

RFA: Round-wise Feature Aggregation
(A): Asymmetrically-structured Network
(S) Symmetrically-structured Network

Approach (Structure type)	Sensitivity	Specificity	Accuracy	AUROC
Liskowski et al. [65] (A)	0.7763	0.9768	0.9495	0.9720
Jiang et al. [45] (A)	0.7540	0.9825	**0.9624**	**0.9810**
Feng et al. [22] (S)	0.7811	**0.9839**	0.9560	0.9792
He et al. [31] (S)	0.7761	0.9792	0.9519	N/a
Baseline (w/o RFA) (A)	0.7807	0.9667	0.9484	0.9659
RFA-Bnet (A)	**0.7932**	0.9741	0.9511	0.9732

Boldface values indicate the best performance at each criterion

Table: **Sensitivity, Specificity, Accuracy, AUROC** on test set of **DRIVE Dataset** [80]

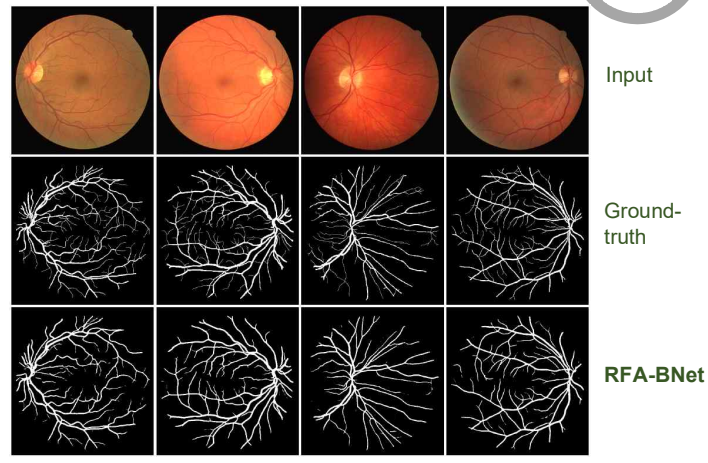
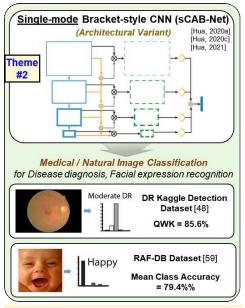


Figure: Qualitative results

- **Theme#1-2:AS1** → Compared to the baseline concept, the involvement of RFA scheme improves 0.0027-0.0125 for all the evaluation metrics.
- The proposed Bracket-style network variant is tuned to achieve state-of-the-art Sensitivity (0.7932) compared to the existing patch-based approaches (i.e., aiming at accurately recognizing pixels of real retinal vessels).
- The performance in terms of Specificity, Accuracy, and AUROC is still comparable to that of the patch-based methods as follows:
  - Specificity: 0.9741 (↓0.0098 compared to the best performance reported in [22])
  - Accuracy: 0.9511 (↓0.0113 compared to [45])
  - AUROC: 0.9732 (↓0.0078 compared to [45])
- Besides that, several qualitative results of the proposed method are demonstrated in the Figure to show its effectiveness in segmenting retinal blood vessels, which appear diversely and irregularly under various illumination conditions of input images.

Experiments: Comparison Results – Theme #2



Theme #2

Boldface values indicate the best performance at each criterion

Table: **Quadratic Weighted Kappa (QWK)** on test set of **Kaggle DR Detection Dataset** [48]

Approach	No. params	QWK (%)
11-layer CNN [85]	10.93M	76.7
SI2DRNet-v1 [15]	10.6M	80.4
18-layer CNN [44]	18.9M	85.1
Zoom-in-Net [92]	55.8M	**85.7**
sCAB-Net (VGG-16 [79]**)**	15.59M	84.9
sCAB-Net (ResNet-101 [29]**)**	47.36M	85.4
sCAB-Net (DenseNet-161 [38]**)**	39.56M	85.6

- sCAB-Net with different backbone network settings reaches competitive QWK rates with the state-of-the-art methods for DR severity recognition.
- Regarding model complexity (No. parameters) trading-off QWK rate:
  - Compared to Zoom-in-Net [92]: sCAB-Net (ResNet-101 & DenseNet-161) has No. parameters less than ~15-29% while achieving comparable QWK (↓0.1-0.3%) .
  - Compared to 18-layer CNN [44]: sCAB-Net (VGG-16) has No. parameters less than ~17.5% while achieving comparable QWK (↓0.2%) .

Approach	Mean Class Accuracy (%)
DLP-CNN [59]	74.20
3DMFA [62]	75.73
ResiDen [47]	76.54
MRE-CNN [21]	76.73
Capsule-based Net [24]	77.48
Double C*d*-LBP [78]	78.60
sCAB-Net (VGG-16 [79]**)**	78.81
sCAB-Net (ResNet-101 [29]**)**	79.33
sCAB-Net (DenseNet-161 [38]**)**	**79.37**

Table: **Mean Class Accuracy** on test set of **RAF-DB Dataset** [59]

Remark: The compared works do not report number of parameters in their models

- sCAB-Net with backbone VGG-16 achieves higher rates of 0.21-4.61% than those of the existing methods.
- Mean Class Accuracies are further improved for sCAB-Net with deeper backbone networks like ResNet-101 (↑0.52%) and DenseNet-161 (↑0.56%) to gain state-of-the-art performance of facial expression recognition.

Conclusions

Thesis Contributions

- An end-to-end trainable deep learning model: **Bracket-style Convolutional Neural Network**, which
  - **Round-by-round combine** semantically-rich information (of the lower-resolution inputs) with finely-patterned features (of the higher-resolution counterparts) through **cross-attentional fusion** mechanism.
  - **Exhaustively exploit** contextual information in **middle- and low-level features** along the tournament of generating final features.
  - **Flexible to coordinate** with different backbone Convolutional Neural Networks for multi-scale feature representations.
  - **Extensible to variants** for different tasks of image semantic segmentation and classification in computer vision.
- Achievements of **impressive results** in comparison with state-of-the-art methods on various benchmark datasets:
  - **Image Semantic Segmentation:** PASCAL VOC 2012 (mIoU = 83.6%), Cityscapes (mIoU = 78.4%), CAMVID (mIoU = 76.4%), DRIVE (Sensitivity = 79.32%).
  - **Image Classification:** RAF-DB (Mean class accuracy = 79.37%), Kaggle DR Detection (QWK = 85.6%).

Uniqueness

- **Bracket-shaped Convolutional Neural Network** and **variants (Round-wise Feature Aggregation; Single-mode structure)** to manipulate feature maps on the tournament of image semantic segmentation or classification.
- **Cross-Attentional Fusion** mechanism to efficiently amalgamate semantically-rich context with finely-patterned representations.

Limitations & Future Work

- More operational computations leading to difficult in meeting the requirements of inference with very high frame rate or usage on mobile platforms.
  - **Future work:** constructing *fast* and *compact* versions of the proposed deep learning architecture using *Knowledge Distillation* to adapt trade-off prerequisites of accuracy, latency, and resource capacity.

- The final performance proportionally relies on the completeness and size of training dataset for any deep learning models in common.
  - **Future work:** applying the strategy of *Unsupervised Domain Adaptation* with the proposed Bracket-structured network to address the lack of well-labeled and big visual data.
    - Utilizing large-scale labeled data available from computer games or computer graphics programs to train models while overcoming domain-shift issue for pixel- and/or image-level classification of real-world images with same contents.

- Besides image semantic segmentation and classification, the Bracket-style network concept can be manipulated to manage more complicated tasks.
  - **Future work:** covering other perception-related problems such as *object detection*, *panoptic segmentation* (which performs instance and semantic segmentation simultaneously), *image super-resolution*, *etc.*

Publications

SCI(E) Journal: 6

- **First author: 3**
  - International Journal of Medical Informatics (Elsevier publisher) (IF: 4.046, 2019)
  - Information Sciences (Elsevier publisher) (IF: 6.795, 2020)
  - IEEE Journal of Biomedical and Health Informatics (IF: 5.772, 2021)
- Co-author: 3

Non-SCI(E) Journal: 1

- Co-author: 1

International Conference: 7

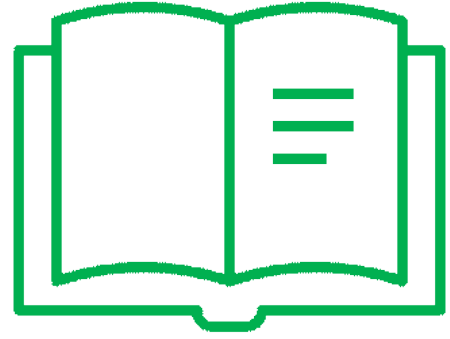
- **First author: 4** (IEEE SMC 2018, IEEE EMBC 2019, IMCOM 2020, IEEE EMBC 2020)
- Co-author: 3

Domestic Conference: 3

- First author: 3

Patent: 3

- **Domestic (Registered): 1** (No. 10-2215757, Feb 8, 2021)
- **International (Registered): 2**
  - JP - No. (JP)6890345, May 27, 2021
  - US - No. 11,145,061, Nov 12, 2021



Total publications = 18

References

[Hua, 2018], [35] Cam-Hao Hua, Thien Huynh-The and Sungyoung Le, "Convolutional Networks with Bracket-style Decoder for Semantic Scene Segmentation", 2018 IEEE Conference on System, Man and Cybernetics (SMC), Oct 7-10, 2018.
[Hua, 2019] Cam-Hao Hua, Thien Huynh-The and Sungyoung Lee, "Retinal Vessel Segmentation using Round-wise Features Aggregation on Bracket-shaped Convolutional Neural Networks", 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, July 23-27, 2019.
[Hua, 2020a] Cam-Hao Hua, Thien Huynh-The, Hyunseok Seo, and Sungyoung Lee, "Convolutional Network with Densely Backward Attention for Facial Expression Recognition", The 14th International Conference on Ubiquitous Information Management and Communication (IMCOM 2020), Taichung, Taiwan, Jan 3-5, 2020.
[Hua, 2020b] Cam-Hao Hua, Thien Huynh-The, Sung-Ho Bae and Sungyoung Lee, "Cross-Attentional Bracket-shaped Convolutional Network for Semantic Image Segmentation", Information Sciences (SCI, IF:5.524), Vol.539, pp.277-294, 2020.
[Hua, 2020c] Cam-Hao Hua, Thien Huynh-The and Sungyoung Lee, "DRAN: Densely Reversed Attention based Convolutional Network for Diabetic Retinopathy Detection", 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Montreal, QC, Canada, July 20-24, 2020.
[Hua, 2021] Cam-Hao Hua, Kiyoung Kim, Thien Huynh-The, Jong In You, Seung-Young Yu, Thuong Le-Tien, Sung-Ho Bae and Sungyoung Lee, "Convolutional Network with Twofold Feature Augmentation for Diabetic Retinopathy Recognition from Multi-modal Images", IEEE Journal of Biomedical and Health Informatics (SCI, IF:5.772), vol. 25, no. 7, pp. 2686-2697, July 2021.
[9] Gabriel J. Brostow et al. "Segmentation and Recognition Using Structure from Motion Point Clouds". In: Computer Vision – ECCV 2008. Springer Berlin Heidelberg, 2008, pp. 44–57.
[18] M. Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016, pp. 3213–3223.
[20] M. Everingham et al. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. 2012.
[48] "Kaggle: Diabetic Retinopathy Detection". In: (https://www. kaggle.com/c/diabetic-retinopathydetection).
[59] Shan Li andWeihong Deng. "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition". In: IEEE Transactions on Image Processing 28.1 (2019), pp. 356–370.
[76] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: International Journal of Computer Vision (IJCV) 115.3 (2015), pp. 211–252.
[80] J. Staal et al. "Ridge-based vessel segmentation in color images of the retina". In: IEEE Transactions on Medical Imaging 23.4 (2004), pp. 501–509. ISSN: 0278-0062.
[12] L. C. Chen et al. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". In: IEEE Transactions on Pattern Analysis and Machine Intelligence 40.4 (2018), pp. 834–848.
[96] M. Yang et al. "DenseASPP for Semantic Segmentation in Street Scenes". In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, pp. 3684–3692.
[102] X. Zhang et al. "Fast Semantic Segmentation for Scene Perception". In: IEEE Transactions on Industrial Informatics 15.2 (2019), pp. 1183–1192.
[103] H. Zhao et al. "Pyramid Scene Parsing Network". In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, pp. 6230–6239.
[105] Feng Zhou, Yong Hu, and Xukun Shen. "Scale-aware spatial pyramid pooling with both encoder-mask and scale-attention for semantic segmentation". In: Neurocomputing 383 (2020), pp. 174 –182. ISSN: 0925-2312.
[23] Jun Fu et al. "Dual Attention Network for Scene Segmentation". In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. 2019, pp. 3146–3154.
[57] Hanchao Li et al. "Pyramid Attention Network for Semantic Segmentation". In: British Machine Vision Conference 2018, BMVC. 2018, p. 285.
[97] C. Yu et al. "Learning a Discriminative Feature Network for Semantic Segmentation". In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, pp. 1857–1866.
[98] Changqian Yu et al. "BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation". In: Computer Vision – ECCV 2018. 2018, pp. 334–349. ISBN: 978-3-030-01261-8.
[101] H. Zhang et al. "Context Encoding for Semantic Segmentation". In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, pp. 7151–7160.
[33] Hexiang Hu et al. "Recalling Holistic Information for Semantic Segmentation". In: CoRR abs/1611.08061 (2016). arXiv: 1611.08061.
[55] T. H. N. Le et al. "Reformulating Level Sets as Deep Recurrent Neural Network Approach to Semantic Segmentation". In: IEEE Transactions on Image Processing 27.5 (2018), pp. 2393–2407.
[66] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. "ParseNet: Looking Wider to See Better". In: CoRR abs/1506.04579 (2015). arXiv: 1506.04579.
[104] Wang Zhe et al. "Learnable Histogram: Statistical Context Features for Deep Neural Networks". In: Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016, pp. 246–262.
[67] J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation". In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015, pp. 3431–3440.
[17] F. Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions". In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, pp. 1800–1807.
[42] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: Proceedings of the 32nd International Conference on Machine Learning. Vol. 37. 2015, pp. 448–456.
[34] J. Hu, L. Shen, and G. Sun. "Squeeze-and-Excitation Networks". In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, pp. 7132–7141.
[11] L. Chen et al. "SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning". In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, pp. 6298–6306.

References (cont.)

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla. “SegNet: A Deep Convolutional Encoder- Decoder Architecture for Image Segmentation”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 39.12 (2017), pp. 2481–2495.
[8] Piotr Bilinski and Victor Prisacariu. “Dense Decoder Shortcut Connections for Single-Pass Semantic Segmentation”. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR. 2018, pp. 6596–6605.
[43] M. A. Islam et al. “Gated Feedback Refinement Network for Dense Image Labeling”. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, pp. 4877–4885.
[53] Ivan Kreso, Josip Krapac, and Sinisa Segvic. “Efficient Ladder-style DenseNets for Semantic Segmentation of Large Images”. In: CoRR abs/1905.05661 (2019). arXiv: 1905.05661.
[60] Xiangtai Li et al. “GFF: Gated Fully Fusion for Semantic Segmentation”. In: CoRR abs/1904.01803 (2019). arXiv: 1904.01803.
[63] G. Lin et al. “RefineNet: Multi-Path Refinement Networks for Dense Prediction”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2019), pp. 1–1.
[64] T. Y. Lin et al. “Feature Pyramid Networks for Object Detection”. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, pp. 936–944.
[72] Marin Orsic et al. “In Defense of Pre-Trained ImageNet Architectures for Real-Time Semantic Segmentation of Road-Driving Images”. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. 2019, pp. 12607–12616.
[75] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, 2015, pp. 234–241.
[83] Zhi Tian et al. “Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation”. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. 2019, pp. 3126–3135.
[95] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. “Wider or Deeper: Revisiting the ResNet Model for Visual Recognition”. In: Pattern Recognition 90 (2019), pp. 119 –133. ISSN: 0031-3203.
[94] T. Wu et al. “Tree-Structured Kronecker Convolutional Network for Semantic Segmentation”. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). 2019, pp. 940–945.
[89] P. Wang et al. “Understanding Convolution for Semantic Segmentation”. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). 2018, pp. 1451–1460.
[14] Liang-Chieh Chen et al. “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs”. In: 3rd International Conference on Learning Representations, ICLR. 2015.
[99] Fisher Yu and Vladlen Koltun. “Multi-Scale Context Aggregation by Dilated Convolutions”. In: 4th International Conference on Learning Representations, ICLR. 2016.
[54] A. Kundu, V. Vineet, and V. Koltun. “Feature Space Optimization for Semantic Video Segmentation”. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016, pp. 3168–3175.
[22] Z. Feng, J. Yang, and L. Yao. “Patch-based fully convolutional neural network with skip connections for retinal blood vessel segmentation”. In: 2017 IEEE International Conference on Image Processing (ICIP). 2017, pp. 1742–1746.
[31] Q. He et al. “Multi-Label Classification Scheme Based on Local Regression for Retinal Vessel Segmentation”. In: 2018 25th IEEE International Conference on Image Processing (ICIP). 2018, pp. 2765–2769.
[45] Zhexin Jiang et al. “Retinal blood vessel segmentation using fully convolutional network with transfer learning”. In: Computerized Medical Imaging and Graphics 68 (2018), pp. 1 –15. ISSN: 0895-6111.
[65] P. Liskowski and K. Krawiec. “Segmenting Retinal Blood Vessels With Deep Neural Networks”. In: IEEE Transactions on Medical Imaging 35.11 (2016), pp. 2369–2380. ISSN: 0278-0062.
[1] D. Acharya et al. “Covariance Pooling for Facial Expression Recognition”. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018, pp. 480–4807.
[21] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. “Multi-region ensemble convolutional neural network for facial expression recognition”. In: International Conference on Artificial Neural Networks. Springer. 2018, pp. 84–94.
[24] S. Ghosh, A. Dhall, and N. Sebe. “Automatic Group Affect Analysis in Images via Visual Attribute and Feature Networks”. In: 2018 25th IEEE International Conference on Image Processing (ICIP). 2018, pp. 1967–1971.
[47] S. Jyoti, G. Sharma, and A. Dhall. “Expression Empowered ResiDen Network for Facial Action Unit Detection”. In: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019). 2019, pp. 1–8.
[59] Shan Li andWeihong Deng. “Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition”. In: IEEE Transactions on Image Processing 28.1 (2019), pp. 356–370.
[62] F. Lin et al. “Facial Expression Recognition with Data Augmentation and Compact Feature Learning”. In: 2018 25th IEEE International Conference on Image Processing (ICIP). 2018, pp. 1957–1961.
[78] F. Shen, J. Liu, and P.Wu. “Double Complete D-LBP with Extreme Learning Machine Auto-Encoder and Cascade Forest for Facial Expression Analysis”. In: 2018 25th IEEE International Conference on Image Processing (ICIP). 2018, pp. 1947–1951.
[29] K. He et al. “Deep Residual Learning for Image Recognition”. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016, pp. 770–778.
[38] G. Huang et al. “Densely Connected Convolutional Networks”. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, pp. 2261–2269.
[79] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: 3rd International Conference on Learning Representations, ICLR. 2015.
[15] Y. Chen et al. “Diabetic Retinopathy Detection Based on Deep Convolutional Neural Networks”. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018, pp. 1030–1034.
[44] S. M. S. Islam, Md M. Hasan, and S. Abdullah. “Deep Learning based Early Detection and Grading of Diabetic Retinopathy Using Retinal Fundus Images”. In: CoRR abs/1812.10595 (2018).
[85] M. C. A. Trivino et al. “Deep Learning on Retina Images as Screening Tool for Diagnostic Decision Support”. In: CoRR abs/1807.09232 (2018). arXiv: 1807.09232.
[92] Z. Wang et al. “Zoom-in-Net: Deep Mining Lesions for Diabetic Retinopathy Detection”. In: Medical Image Computing and Computer Assisted Intervention, MICCAI 2017. 2017, pp. 267–275.
[30] Kaiming He et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). 2015, pp. 1026–1034.

THANK YOU

Comments & Questions