Enhancing Learning Capability of Convolutional Neural Networks for Fundamental Vision Problems

Minglun Gong

School of Computer Science, University of Guelph

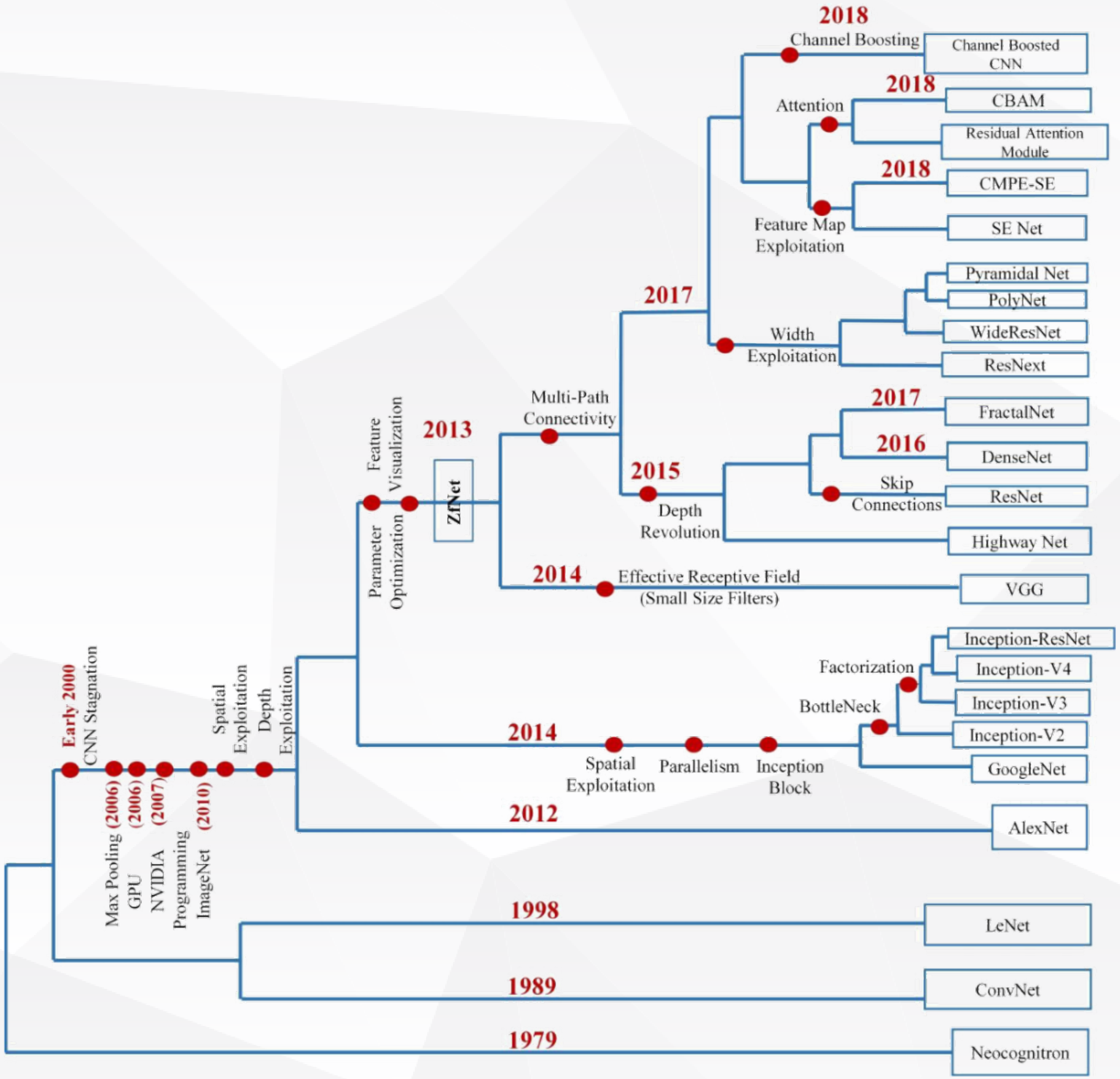
Based on the PhD thesis work of Mingjie Wang

UNIVERSITY of GUELPH

Most Influential CNN Architectures

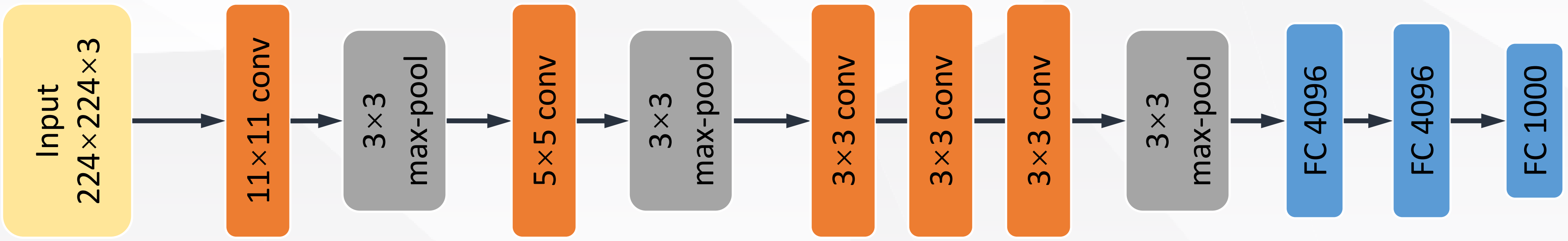
- Introduction
- Classification #1
- Classification #2
- Regression #1
- Regression #2
- Regression #3
- Summary

- LeNet: 1998
- AlexNet: 2012
- VGG-16: 2014
- GoogLeNet: 2014
- ResNet: 2015
- SqueezeNet: 2016
- DenseNet: 2017
- ShuffleNet: 2018
- SENet: 2018
- EfficientNet: 2019



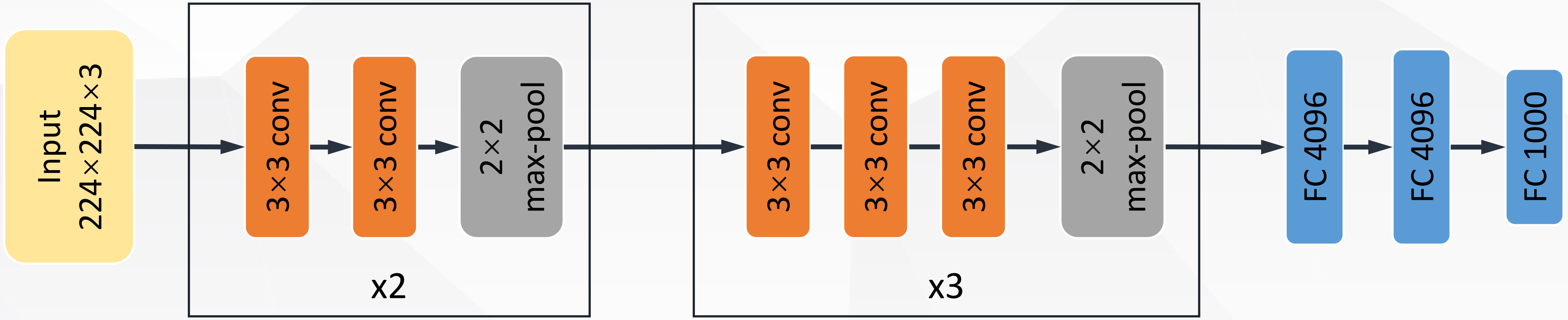
UNIVERSITY of GUELPH

AlexNet (2012): 60M Parameters



- Popularize Rectified Linear Unit (ReLU) to replace sigmoid or tanh as activation function.
  - Overcome the gradient saturation problem.
  - Allow models to learn faster and perform better.
- Apply dropout, which randomly removes neurons from propagation.
  - Breaks co-adaptation among neural units.

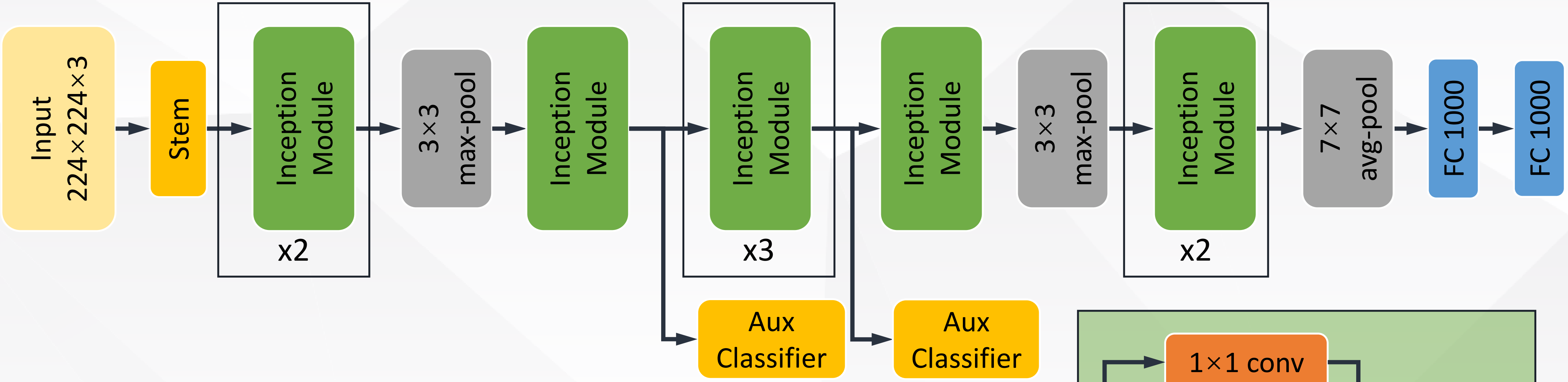
VGG-16 (2014): 138M Parameters

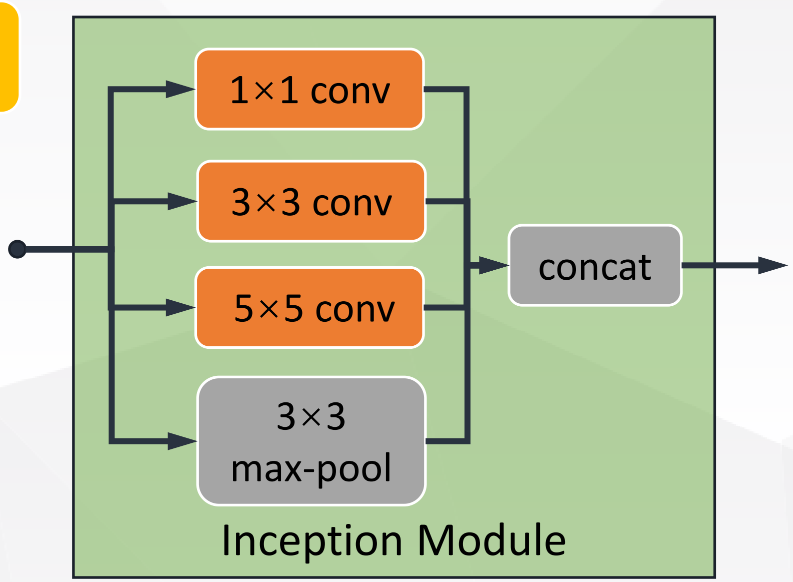


- Train much deeper network.
    - Roughly twice as deep as AlexNet.
- Replace convolution kernels of different sizes by stacking uniform (3×3) convolutions.
    - Reduce hyperparameters.
    - Save computational cost by using small kernel sizes.
    - Has been adopted as the pretrained model for most of downstream tasks.

UNIVERSITY of GUELPH

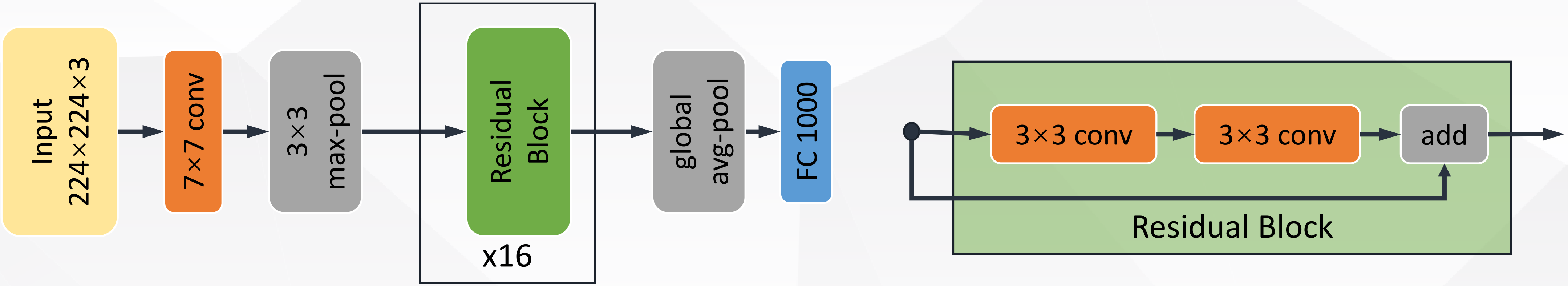
GoogLeNet/Inception-v1 (2014): 7M Parameters





- Introduce inception module
  - Stack together outputs of pooling layer and convolution layers with different filter sizes.
  - Strengthen the regularization and scale invariance of extracted features.
- Use 1×1 convolutions for dimensionality reduction.
- Introduce two auxiliary classifiers.

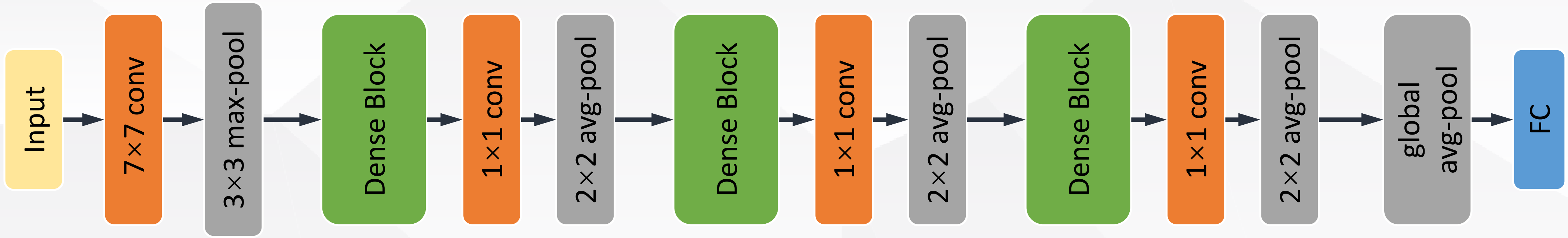
ResNet (2015): 26M Parameters



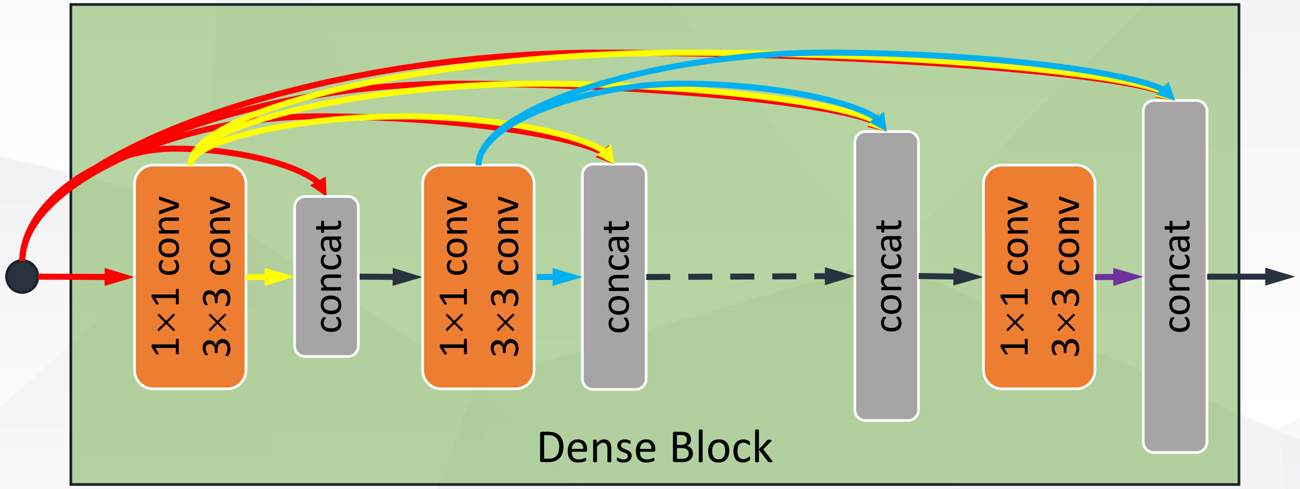
- Popularise skip connections to form residual blocks.
- Allow training very deep networks (1,000+ layers).
    - Additional layers won’t hurt the performance due to skip connections.
- Among the first to use batch normalisation to handle internal covariate shift.

UNIVERSITY of GUELPH

DenseNet (2017):

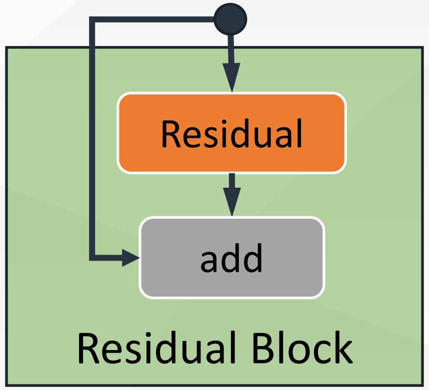


- Provide a new paradigm of refining features in a densely-connected manner.
- Within each dense block, features from all preceding-layer are reused.
  - Concatenated together instead of added.

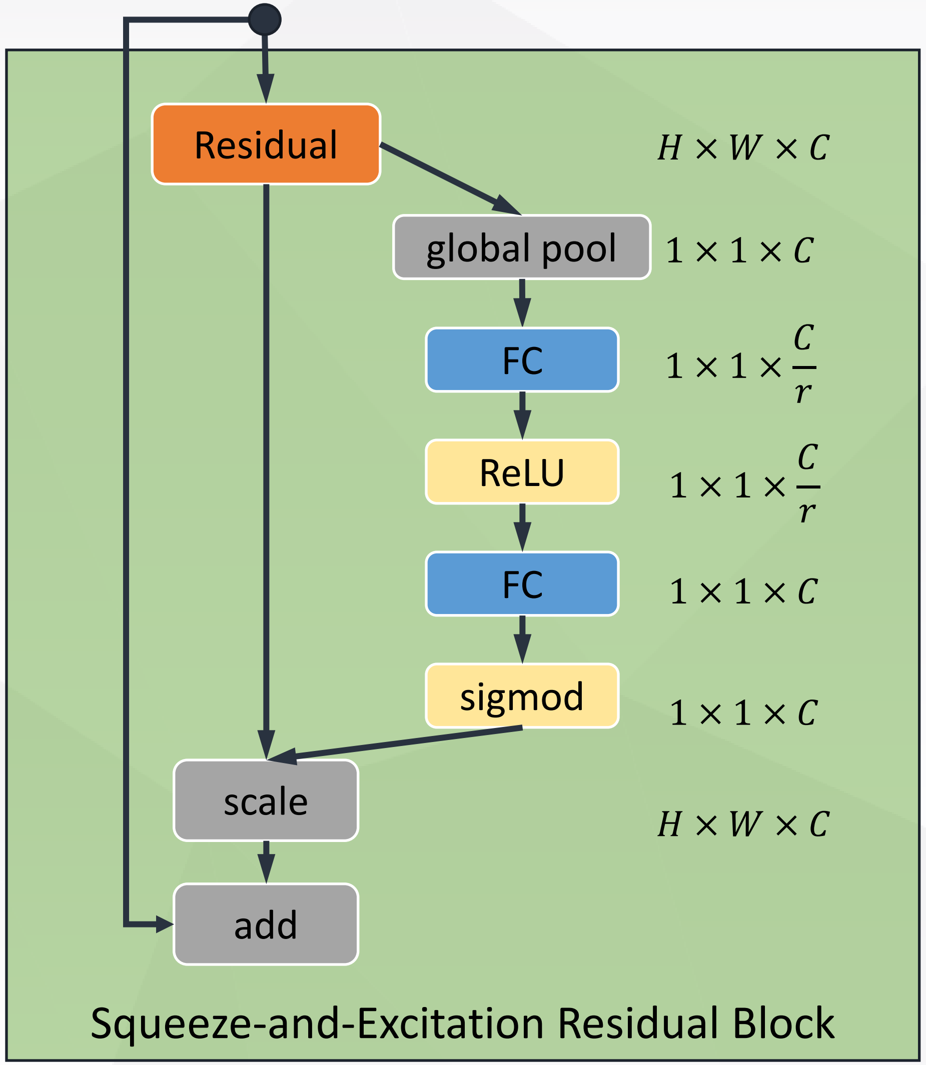


SENet (2018):

- Explicitly model the inter-dependencies between the channels of convolutional features.
- Perform dynamic channel-wise feature recalibration.
  - Use global information to emphasize informative features and suppress less useful ones



Residual Block



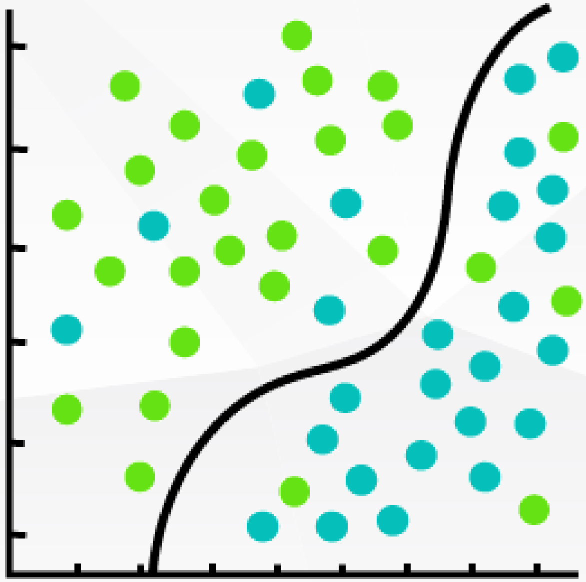
Squeeze-and-Excitation Residual Block

UNIVERSITY of GUELPH

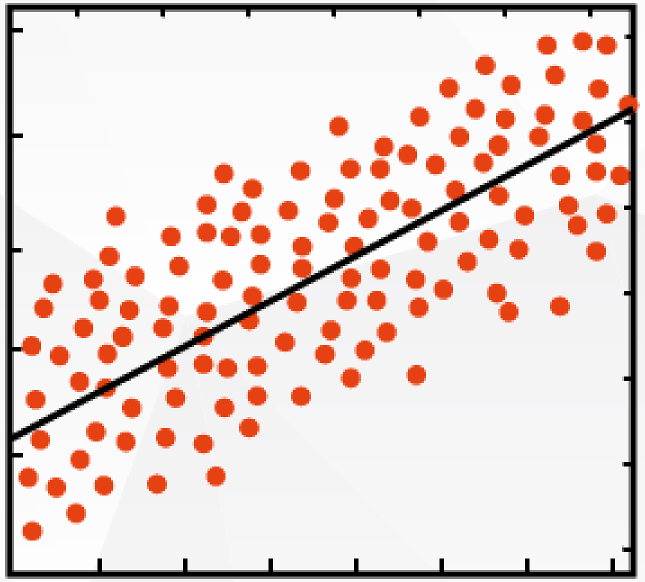
Two Types of Learning Problems

- Introduction
- Classification #1
- Classification #2
- Regression #1
- Regression #2
- Regression #3
- Summary

- Classification:
  - Predict discrete labels.
  - Find accuracy decision boundaries.
- Computer vision problems:
  - Image Classification
  - Gender Detection
  - Semantic/instance Segmentation…



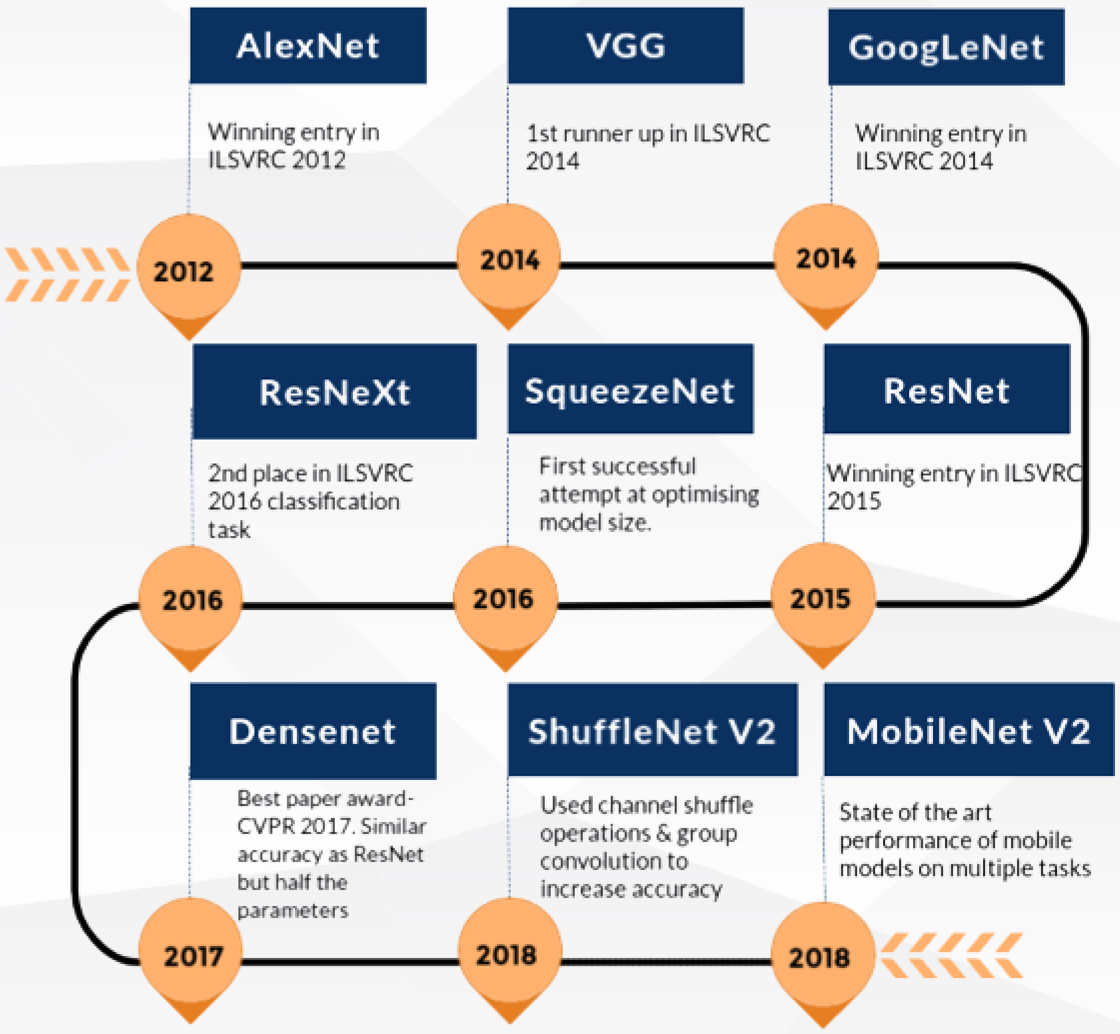
- Regression:
  - Predict continuous quantities.
  - Find the best fitting line.
- Computer vision problems :
  - Crowd Counting
  - Age Estimation
  - Object Localization…

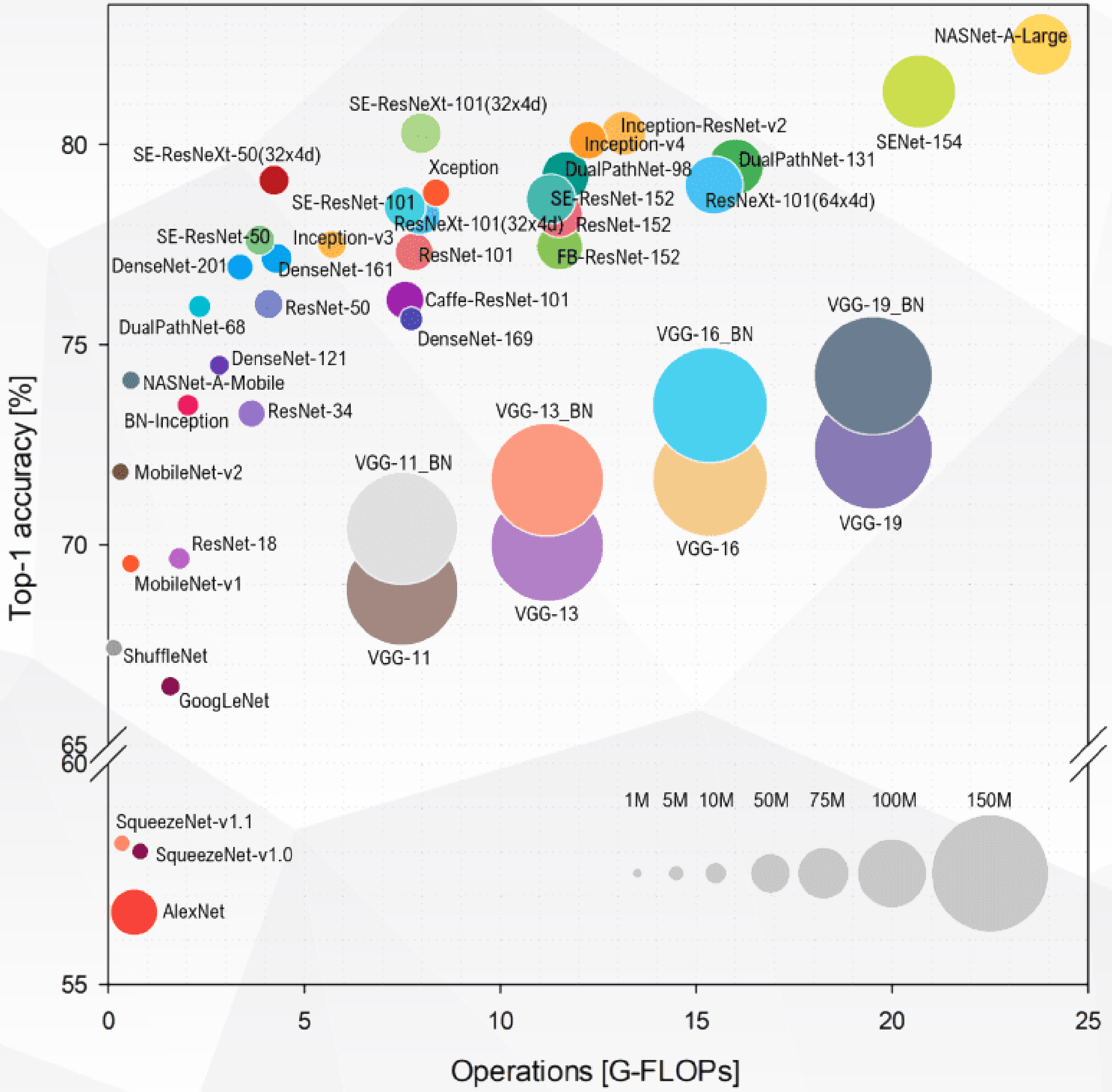


UNIVERSITY of GUELPH

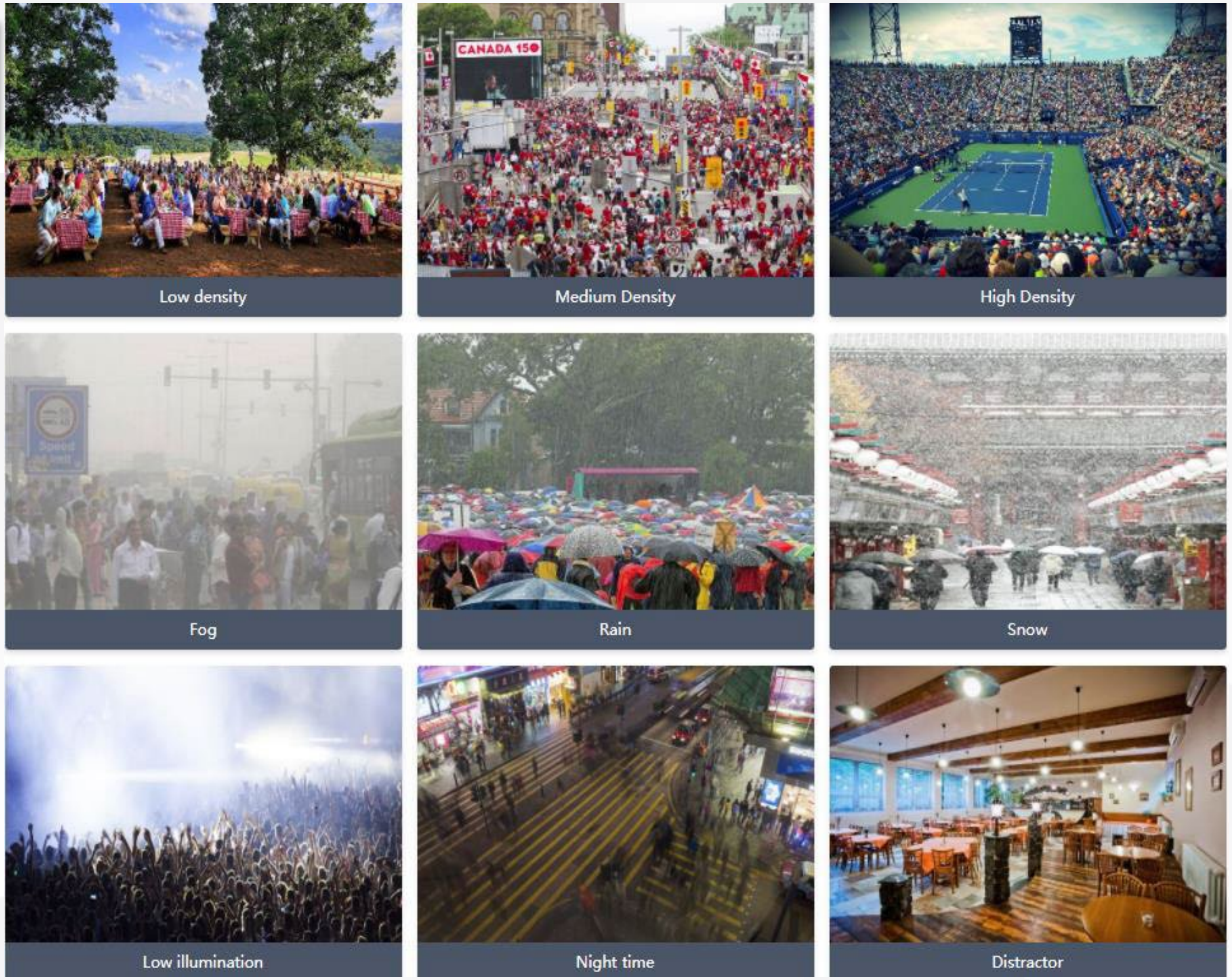
Image Classification

- **Introduction**
- Classification #1
- Classification #2
- Regression #1
- Regression #2
- Regression #3
- Summary





Crowd Counting



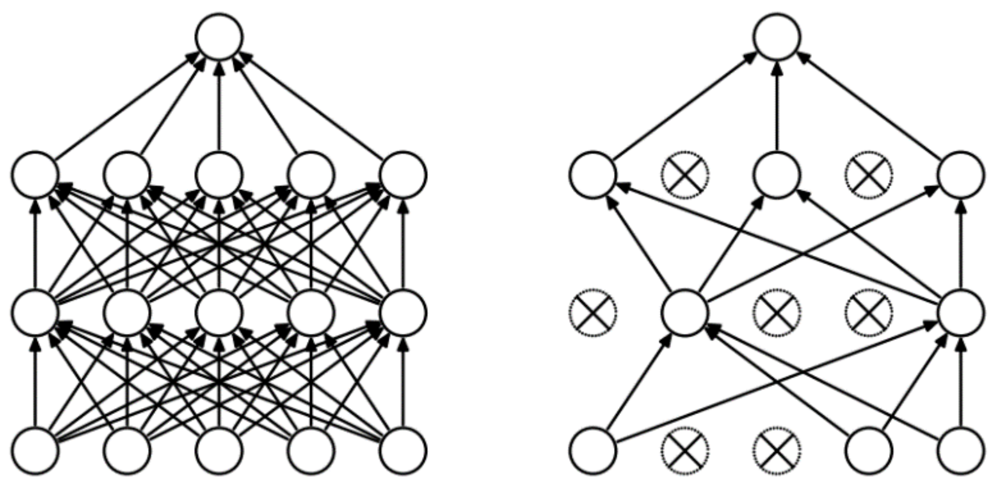
JHU-CROWD++: A large-scale unconstrained crowd counting dataset

Year-Conference/Journal	Methods	MAE	MSE
2019--ICCV	DSSINet	60.63	96.04
2019--ICCV	MBTTBF-SCFB	60.2	94.1
2019--ICCV	RANet	59.4	102.0
2019--ICCV	SPANet+SANet	59.4	92.5
2019--TIP	PaDNet	59.2	98.1
2019--ICCV	S-DCNet	58.3	95.0
2020--ICPR	M-SFANet+M-SegNet	57.55	94.48
2019--ICCV	**PGCNet**	57.0	**86.0**
2020--ECCV	AMSNet	56.7	93.4
2020--CVPR	ADSCNet	55.4	97.7
2021--AAAI	**SASNet**	**53.59**	88.38

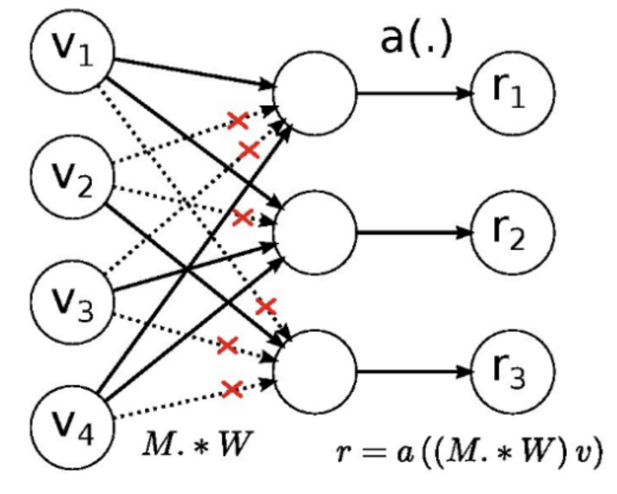
Multi-scale Convolution Aggregation and Stochastic Feature Reuse for DenseNets

WACV 2019

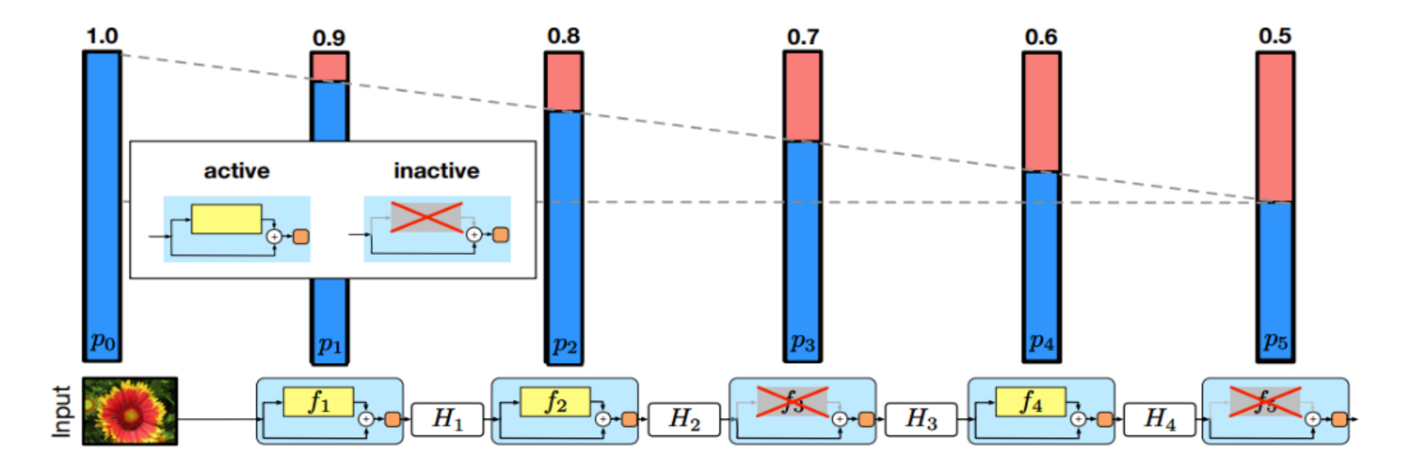
Background on Stochastic Regularization



(a) Dropout (neurons)



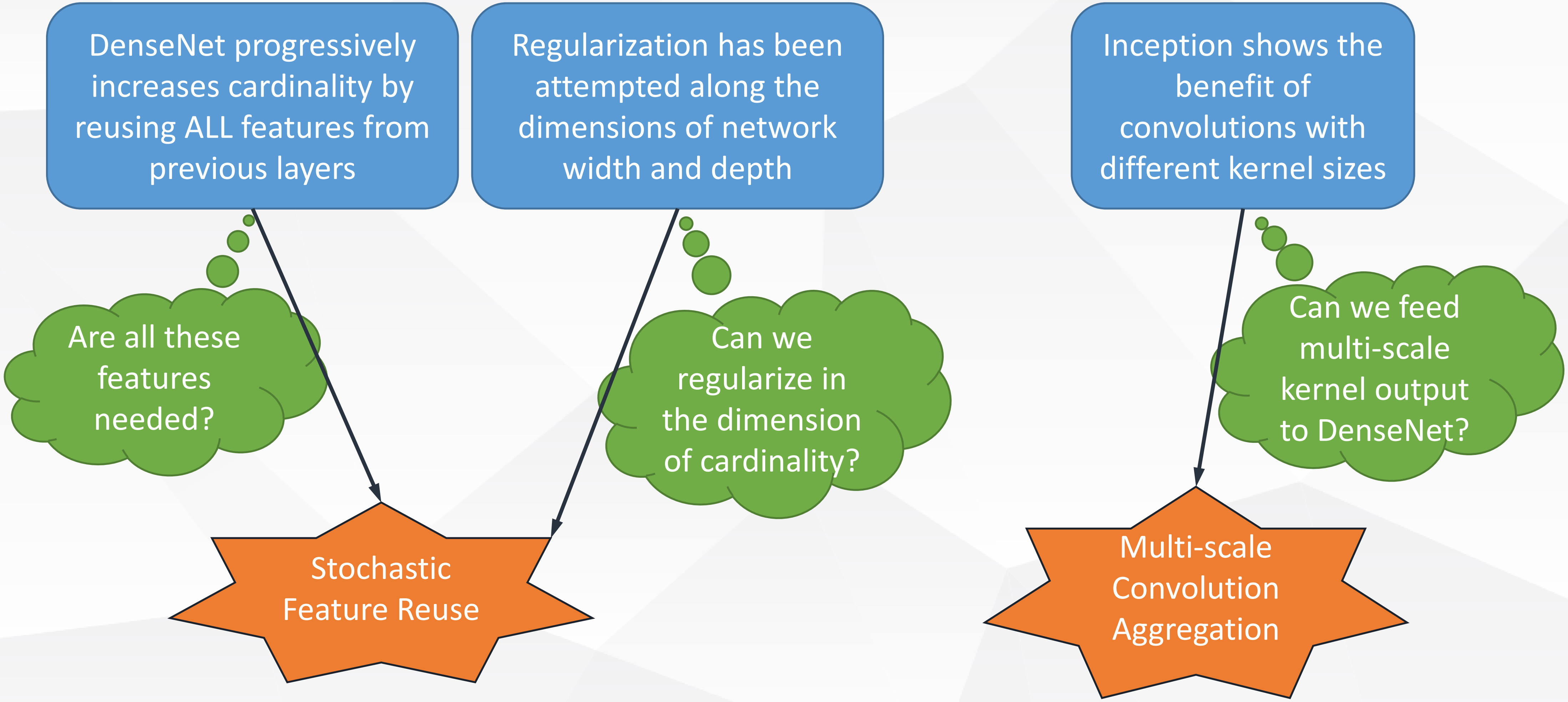
(b) Drop Connect (connections)



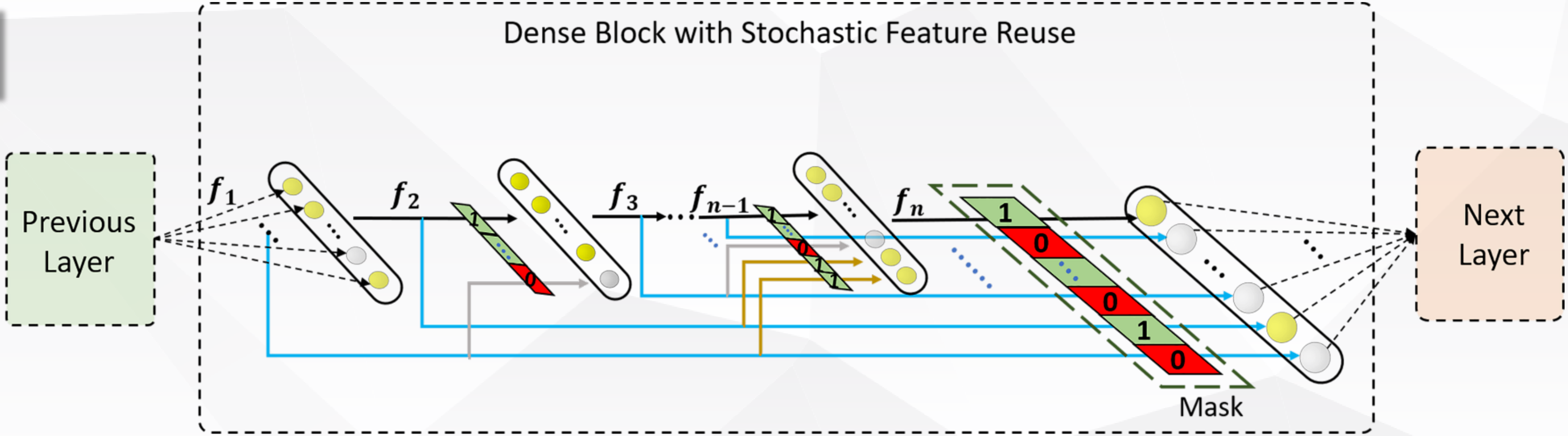
(c) Stochastic Depth (depth)

UNIVERSITY of GUELPH

Motivations



Stochastic Feature Reuse (SFR)



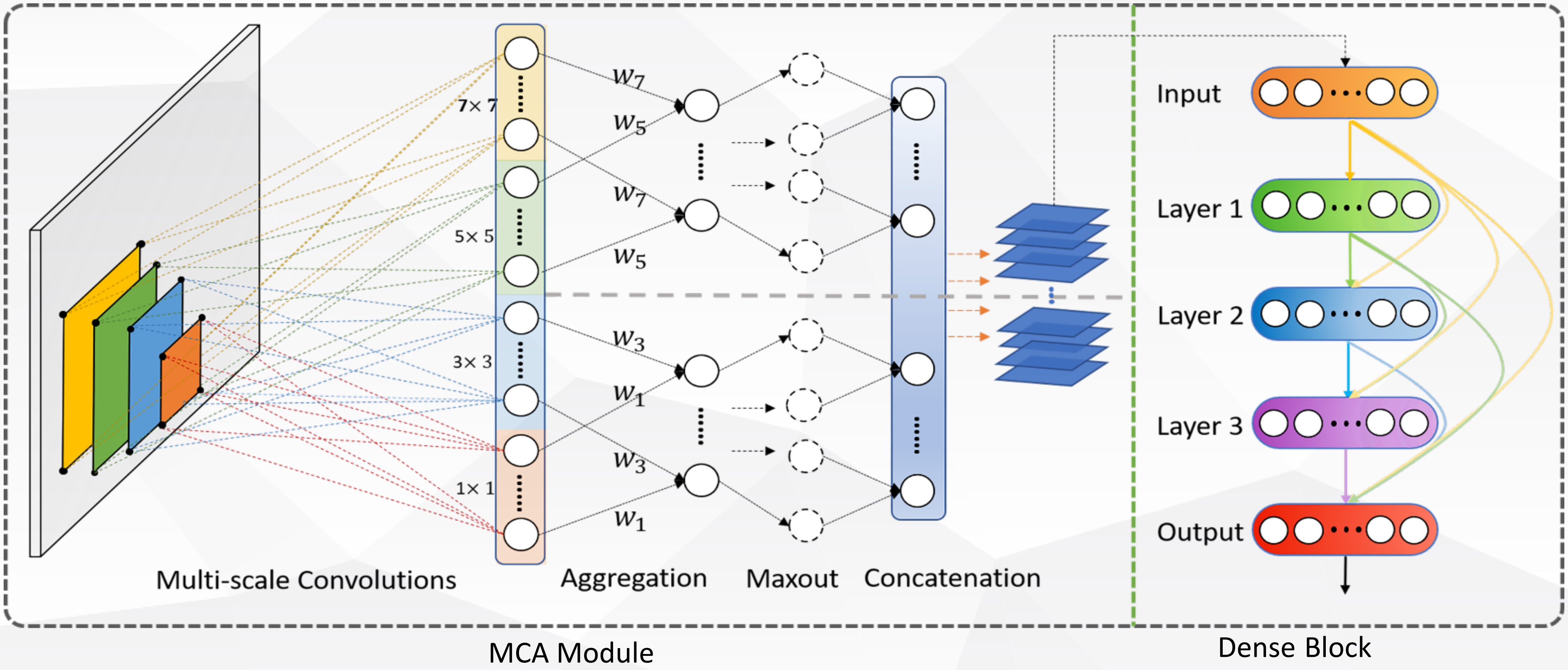
$$f_{l+1} = M_l x_l = M_l \cdot H_l([f_0, f_1, \dots, f_{l-1}])$$

- Mask tensor M_l obeying Bernoulli distribution is randomly generated for each layer during each mini batch.

UNIVERSITY of GUELPH

Multi-scale Convolution Aggregation (MCA) Module

- Introduction
- Classification #1
- Classification #2
- Regression #1
- Regression #2
- Regression #3
- Summary



Formulation of MCA Module

- Involve 3 steps:
  - Multi-scale Convolutions
    - $\mathrm{M}(\mathrm{x}, \mathrm{W}) = \mathrm{Concat}(\mathrm{G}_{1\times1}(\mathrm{x}, \mathrm{W}_1), \mathrm{G}_{3\times3}(\mathrm{x}, \mathrm{W}_3), \mathrm{G}_{5\times5}(\mathrm{x}, \mathrm{W}_5), \mathrm{G}_{7\times7}(\mathrm{x}, \mathrm{W}_7))$
  - Cross-scale Aggregation
    - $\mathrm{M}(\mathrm{x}, \mathrm{W}) = \mathrm{Concat}(\mathrm{w}_1\mathrm{G}_{1\times1}(\mathrm{x}, \mathrm{W}_1) + \mathrm{w}_3\mathrm{G}_{3\times3}(\mathrm{x}, \mathrm{W}_3), \mathrm{w}_5\mathrm{G}_{5\times5}(\mathrm{x}, \mathrm{W}_5) + \mathrm{w}_7\mathrm{G}_{7\times7}(\mathrm{x}, \mathrm{W}_7))$
  - Maxout Activation
    - $\mathrm{M}(\mathrm{x}, \mathrm{W}) = \mathrm{Concat}(\mathrm{Maxout}(\mathrm{w}_1\mathrm{G}_{1\times1}(\mathrm{x}, \mathrm{W}_1) + \mathrm{w}_3\mathrm{G}_{3\times3}(\mathrm{x}, \mathrm{W}_3)),$
    - $\mathrm{Maxout}(\mathrm{w}_5\mathrm{G}_{5\times5}(\mathrm{x}, \mathrm{W}_5) + \mathrm{w}_7\mathrm{G}_{7\times7}(\mathrm{x}, \mathrm{W}_7)))$
- The final output of MCA module is fed it into the first Dense Block.
  - Replace the original Initial Layer with a highly non-linear transformation between input image and the dense block.

UNIVERSITY of GUELPH

Datasets and Training Details

- CIFAR-10, CIFAR-100 and SVHN
  - CIFAR-10: training (50,000), testing (10,000), 32×32 resolution, 10 classes.
  - CIFAR-100: training (50,000), testing (10,000), 32×32 resolution, 100 classes.
  - Street View House Number: training (73,257), testing (26,032), additional training (53,1131), 32×32 resolution, 10 classes.
- Normalization Methods
  - CIFAR datasets: Subtract mean values and divide standard deviations.
  - SVHN: Divided by 255.
- Details of Training Configurations
  - 350 epochs for CIFAR and 40 epochs for SVHN.
  - Initial learning rate is 0.1 and divided by 10 at epochs 150, 225 and 300 for CIFAR and epochs 20 and 30 for SVHN.
  - Weight decay 0.0001 and Nesterov momentum 0.9.
  - Dropout probability 0.8 and He Initialization of weights.

UNIVERSITY of GUELPH

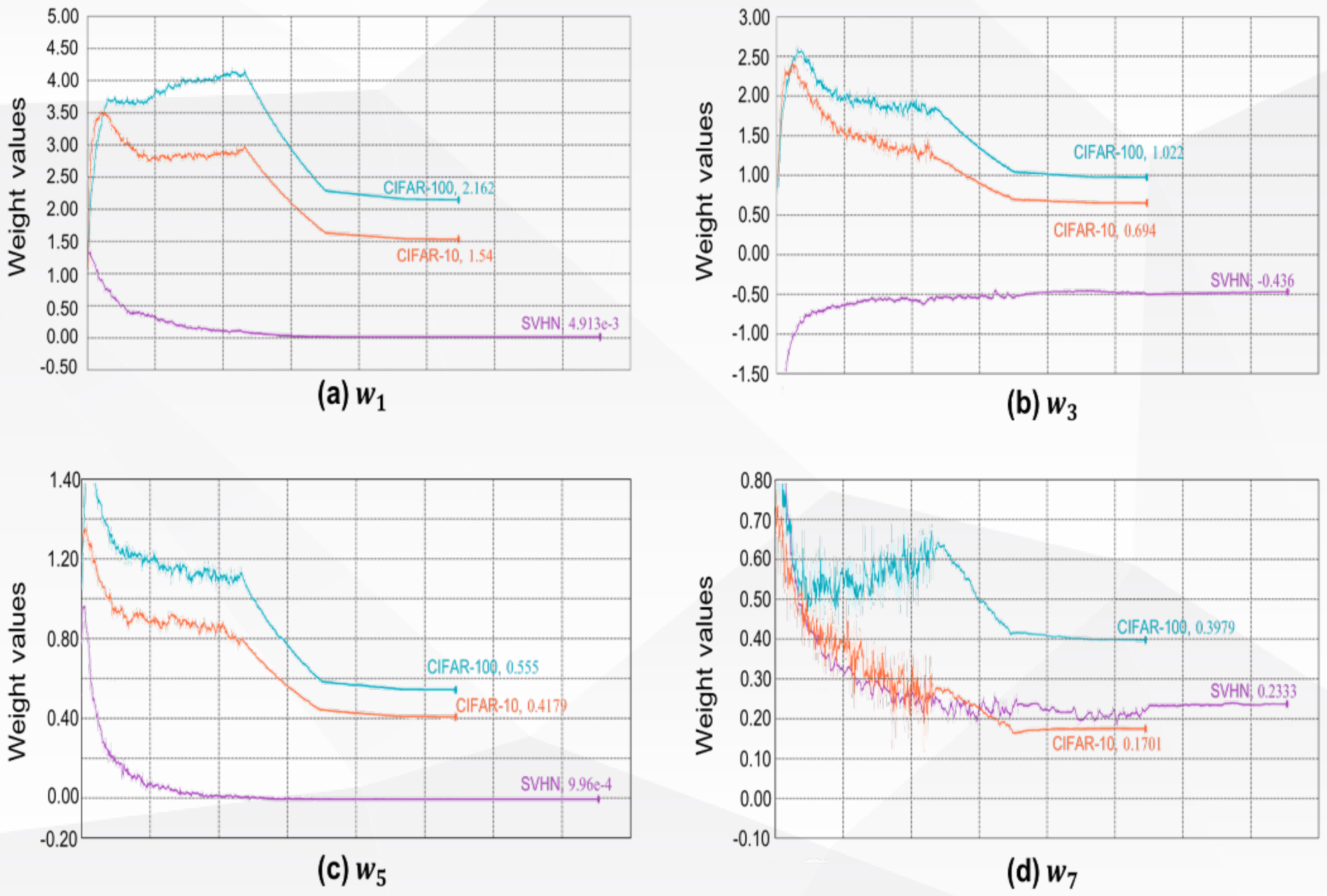
Quantitative Evaluation on Classification Accuracy

Model	Depth	Params.	C10(%)	C100(%)	SVHN(%)
Stochastic Pooling [36]	-	-	15.13	42.51	2.80
Maxout Networks [4]	-	-	11.68	38.57	2.47
Network in Network [18]	-	-	10.41	35.68	2.35
Deeply Supervised Net [16]	-	-	9.69	34.57^{+}	1.92
Competitive Multi-scale [17]	-	4.48M	6.87	27.56	1.76
Highway Network [24]	-	-	7.72^{+}	32.39^{+}	-
Fractal Network [15]	21	38.6M	10.18	35.34	2.01
FractalNet with Drop-path [15]	21	38.6M	7.33	28.20	1.87
ResNet [6]	110	1.7M	6.61^{+}	-	-
Stochastic Depth [10]	110	1.7M	11.66	37.80	1.75
ResNet(pre-activation) [7]	164 1001	1.7M 10.2M	11.26 10.56	35.58 33.47	- -
DenseNet($k = 12$) [9]	40	1.0M	7.00	27.55	1.79
DenseNet($k = 24$) [9]	100	27.2M	5.83	23.42	1.59
DenseNet($k = 24$)[9]	53	7.8M	6.45	24.32	1.78
DenseNet with SFR($k = 24$)	53	7.8M	**6.08**	**23.82**	**1.66**
DenseNet-BC($k = 12$)[9]	100	0.8M	5.92	24.15	1.76
DenseNet-BC with MCA($k = 12$)	100	0.8M	**5.41**	**24.07**	-
DenseNet with MCA($k = 12$)	40	1.0M	6.44	27.44	1.77
DenseNet with MCA($k = 24$)	40	4.2M	**5.38***	**23.78**	**1.66**
DenseNet with MCA($k = 40$)	40	11.6M	5.76	**22.65***	**1.61***

- Under growth rate k=24 and depth=40:
  - Obtain the lowest classification errors on CIFAR-10 (5.38%) and CIFAR-100 (23.78%)
  - Use fewer parameters.
- Under k=40 and depth=40:
  - Get impressive results on CIFAR-100 (22.65%) and 1.61% on SVHN.

UNIVERSITY of GUELPH

Adaptive Weights in MCA



Weights	CIFAR10	CIFAR100	SVHN
w_1	1.54	2.162	4.913e-3
w_3	0.694	1.022	-0.436
w_5	0.4179	0.555	9.96e-4
w_7	0.1701	0.3979	0.2333

- The converged weights for different datasets are drastically different.
- The scales with high discrimination power are preserved, whereas the redundant ones are suppressed.

Ablation Studies

	Block index	Error(%)	Dataset
DenseNet [9]	*None*	6.45	CIFAR-10
SFR	4^{th}	**6.08**	CIFAR-10
SFR	1^{st}	8.99	CIFAR-10
SFR(No Dropout)	4^{st}	10.00	CIFAR-10
DenseNet [9]	None	24.32	CIFAR-100
SFR	4^{st}	**23.82**	CIFAR-100
SFR	1^{st}	26.54	CIFAR-100
SFR(No Dropout)	4^{st}	27.15	CIFAR-100
DenseNet [9]	None	1.78	SVHN
SFR	4^{st}	**1.66**	SVHN
SFR	1^{st}	2.12	SVHN
SFR(No Dropout)	4^{st}	3.02	SVHN

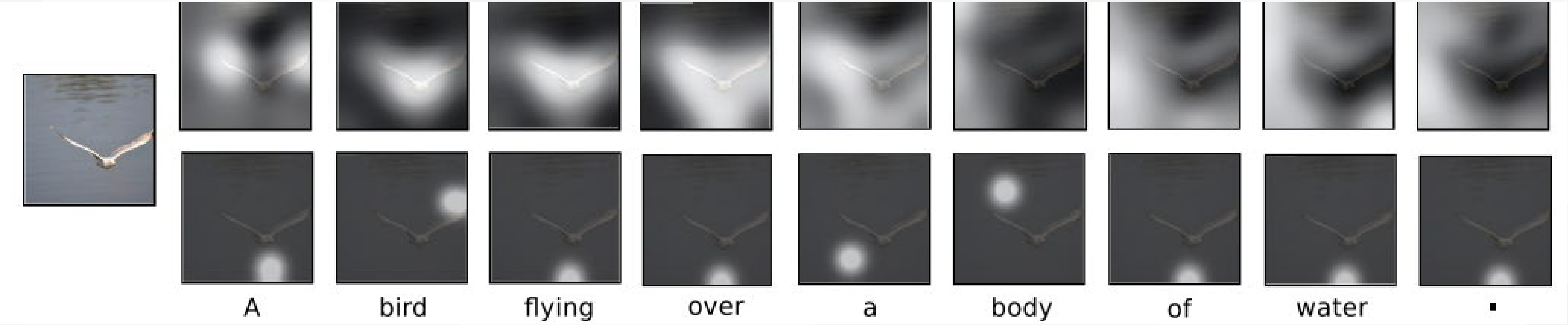
	Width	w.o. SFR	w. SFR	Improve
SFR($k = 12$)	17196	6.93	6.80	0.13
SFR($k = 24$)	34392	6.45	6.08	**0.37**
SFR(WIL)	34536	6.09	5.76	0.33
SFR($k = 40$)	57320	6.53	6.32	0.21

- Place SFR operator at different locations of DenseNets.
  - Adding dense block with SFR on the top of the DenseNet leads to best results.
- Ablation studies on the diverse growth rates of SFR, k=12, 24 and 40.
  - More effective on relatively wider DenseNets.

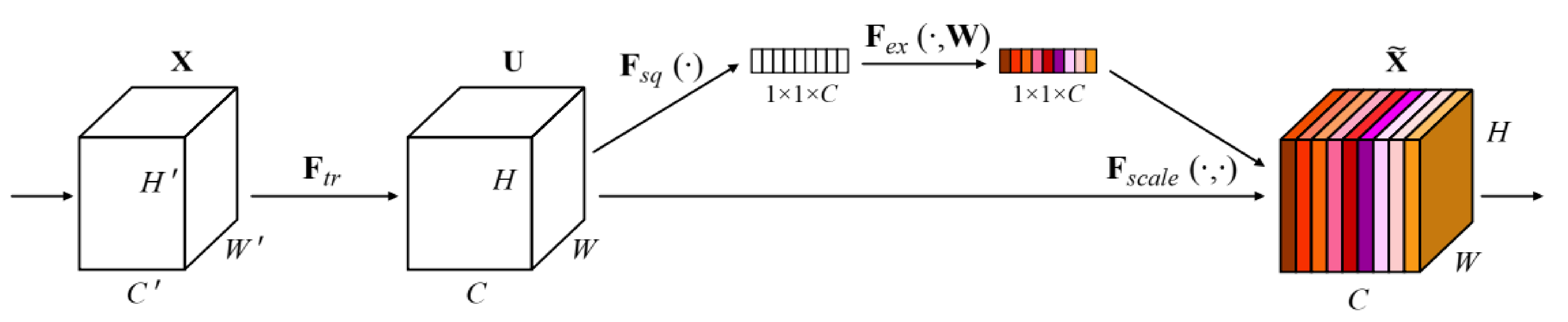
ADNet: Adaptively Dense Convolutional Neural Networks

WACV 2020

Background on Attentions



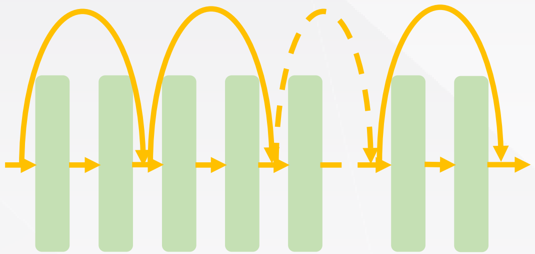
Spatial-wise attention in “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”



Channel-wise attention in “Squeeze-and-Excitation Networks”

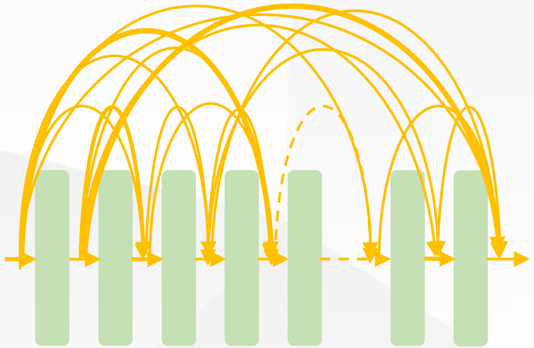
Motivations

ResNet



- Sparse connections
- Add kernel outputs together before activation

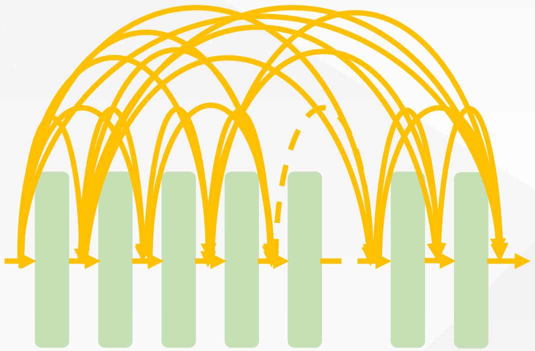
Can a network adaptively determine its connection density?



Can a compact model works equally well?

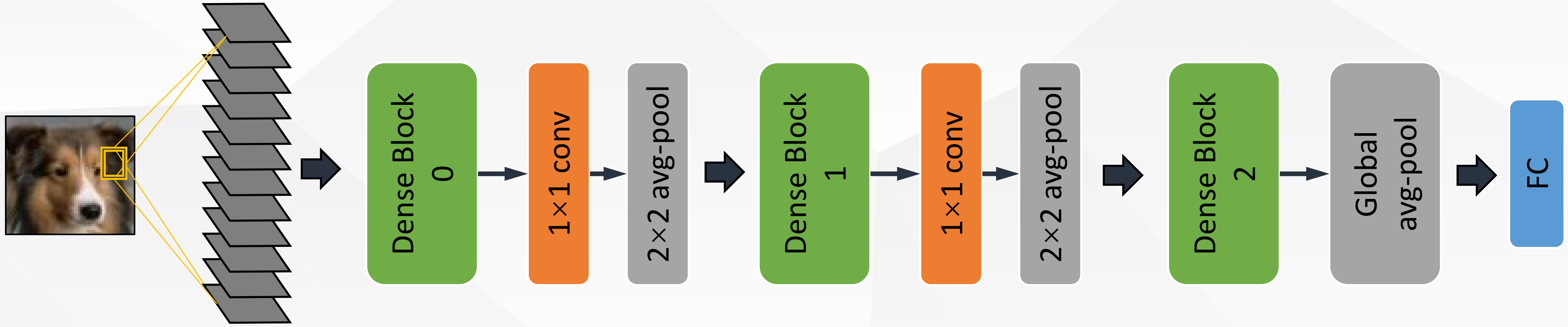
Adaptively Dense Net

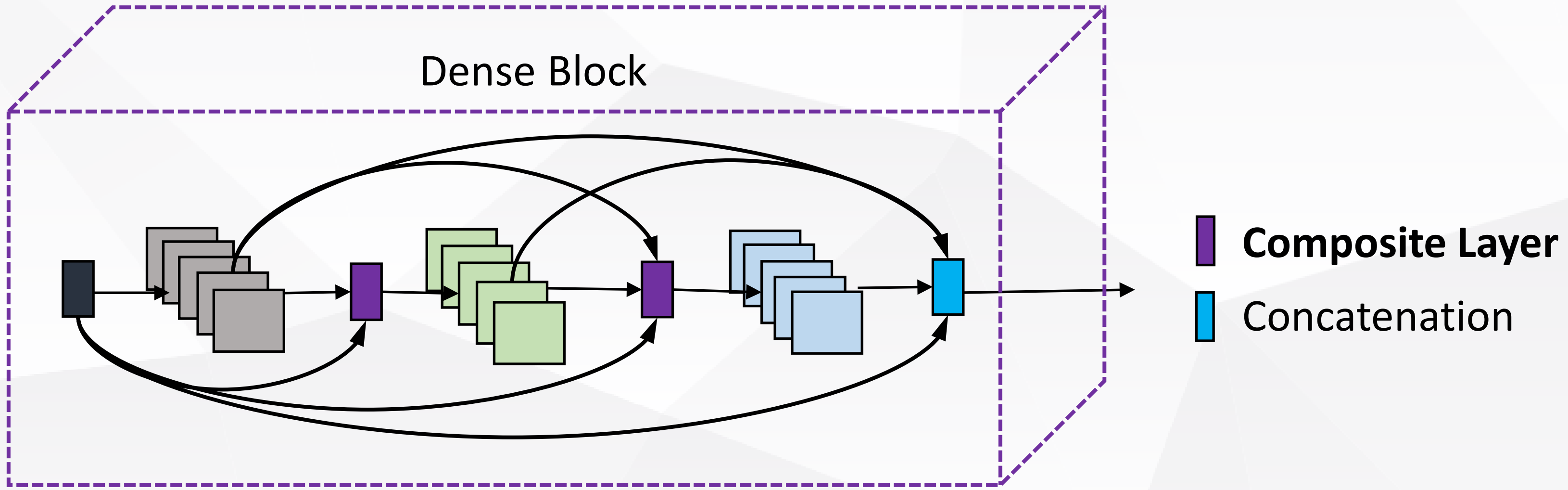
DenseNet



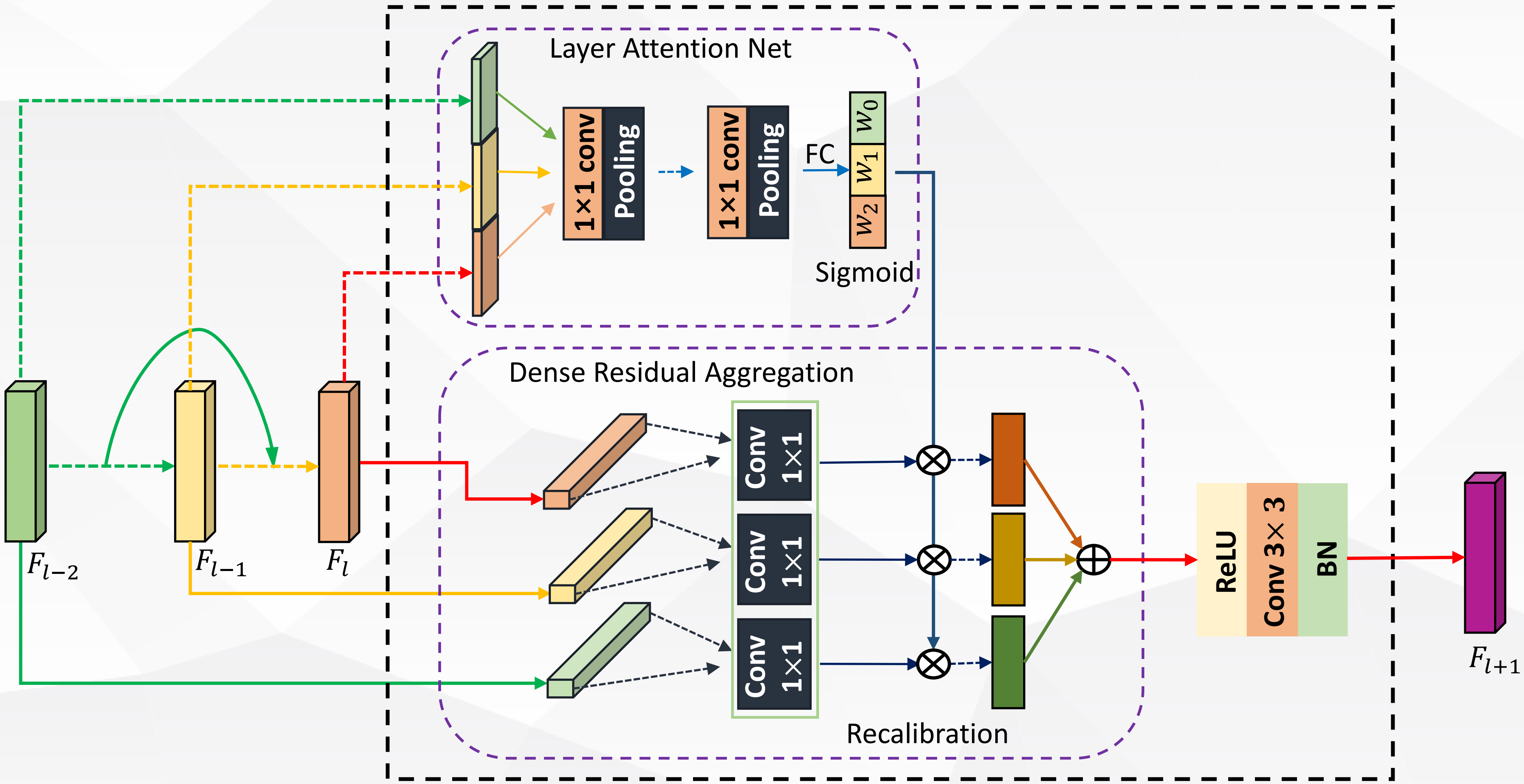
- Dense connections
- Concatenate features together.

Overall Architecture of ADNet





Composite Layer

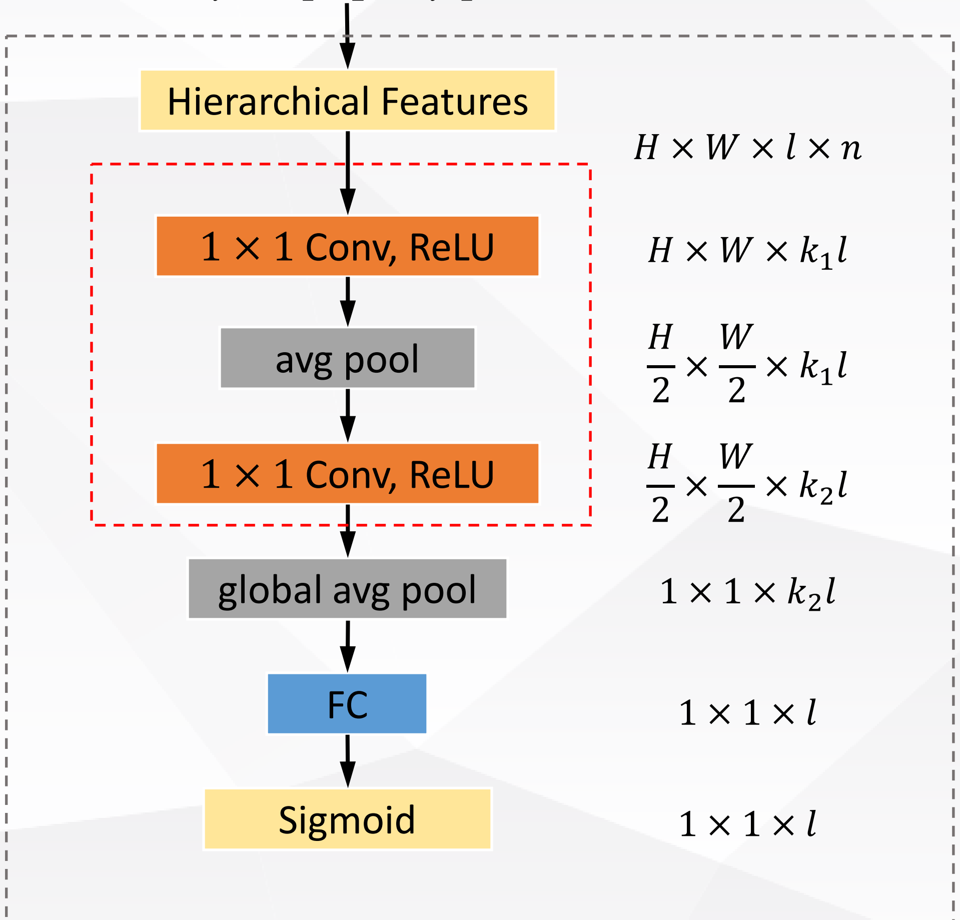


Composite Layer

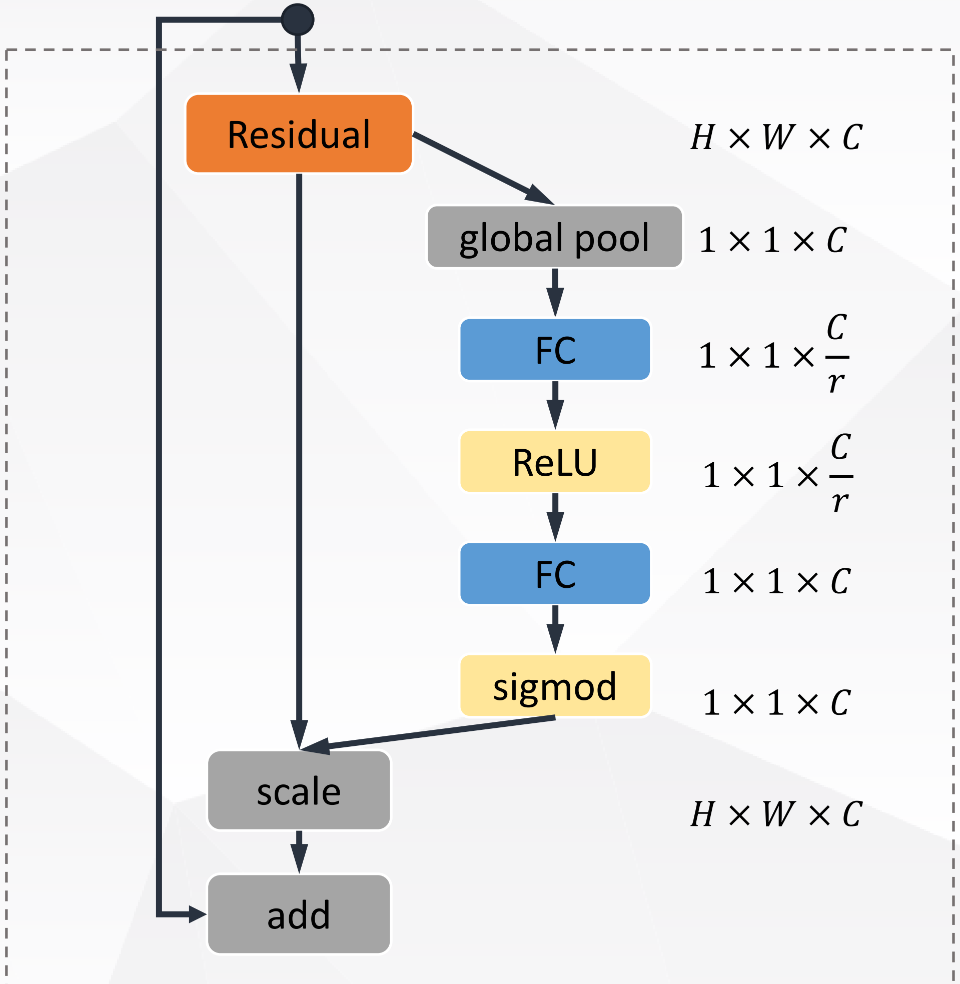
UNIVERSITY of GUELPH

Layer Attention Net

$I_l = [x_1, x_2 \ldots x_{l-1}]$



Layer Attention in our ADNet

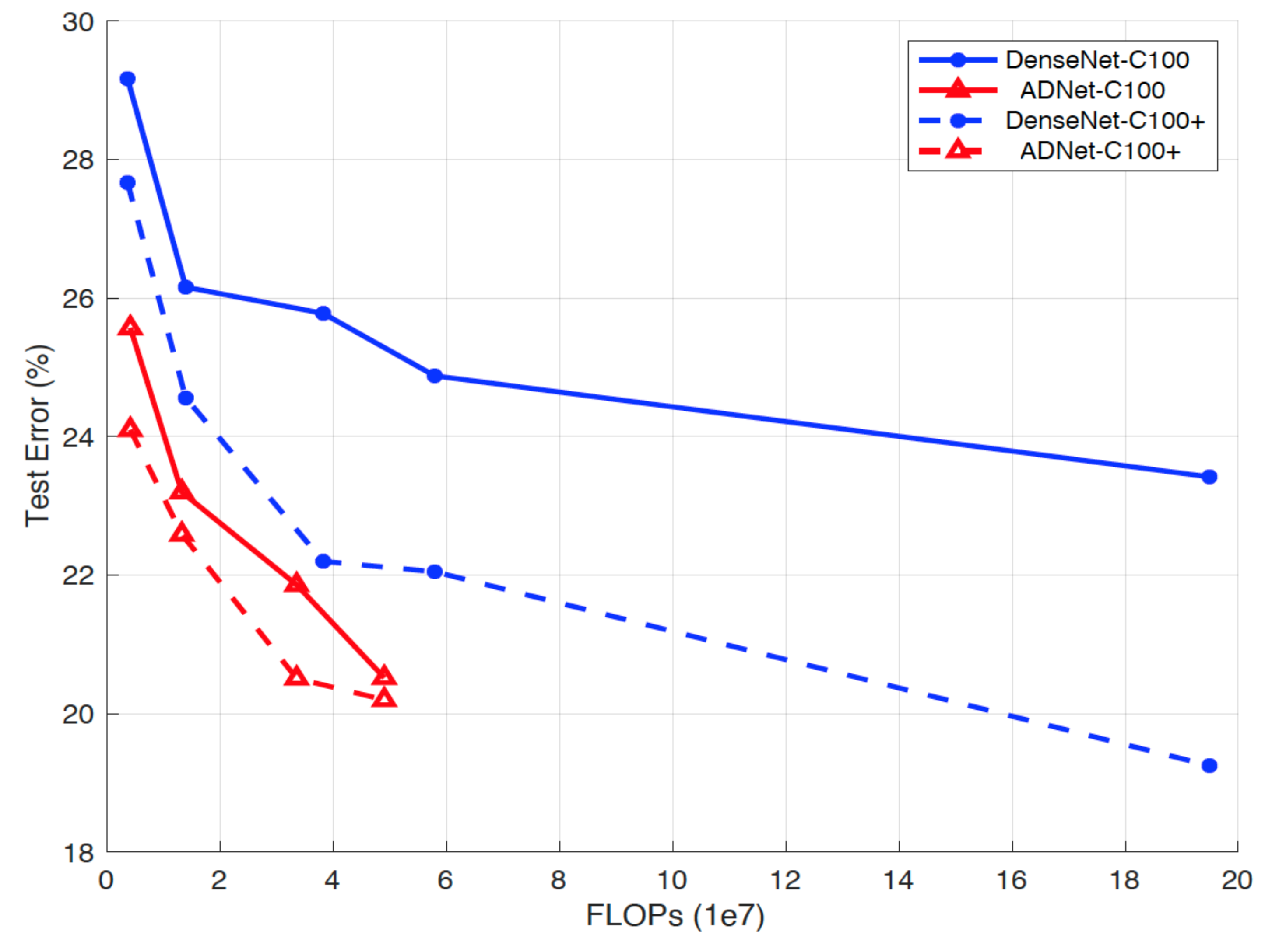


Channel Attention in SENet

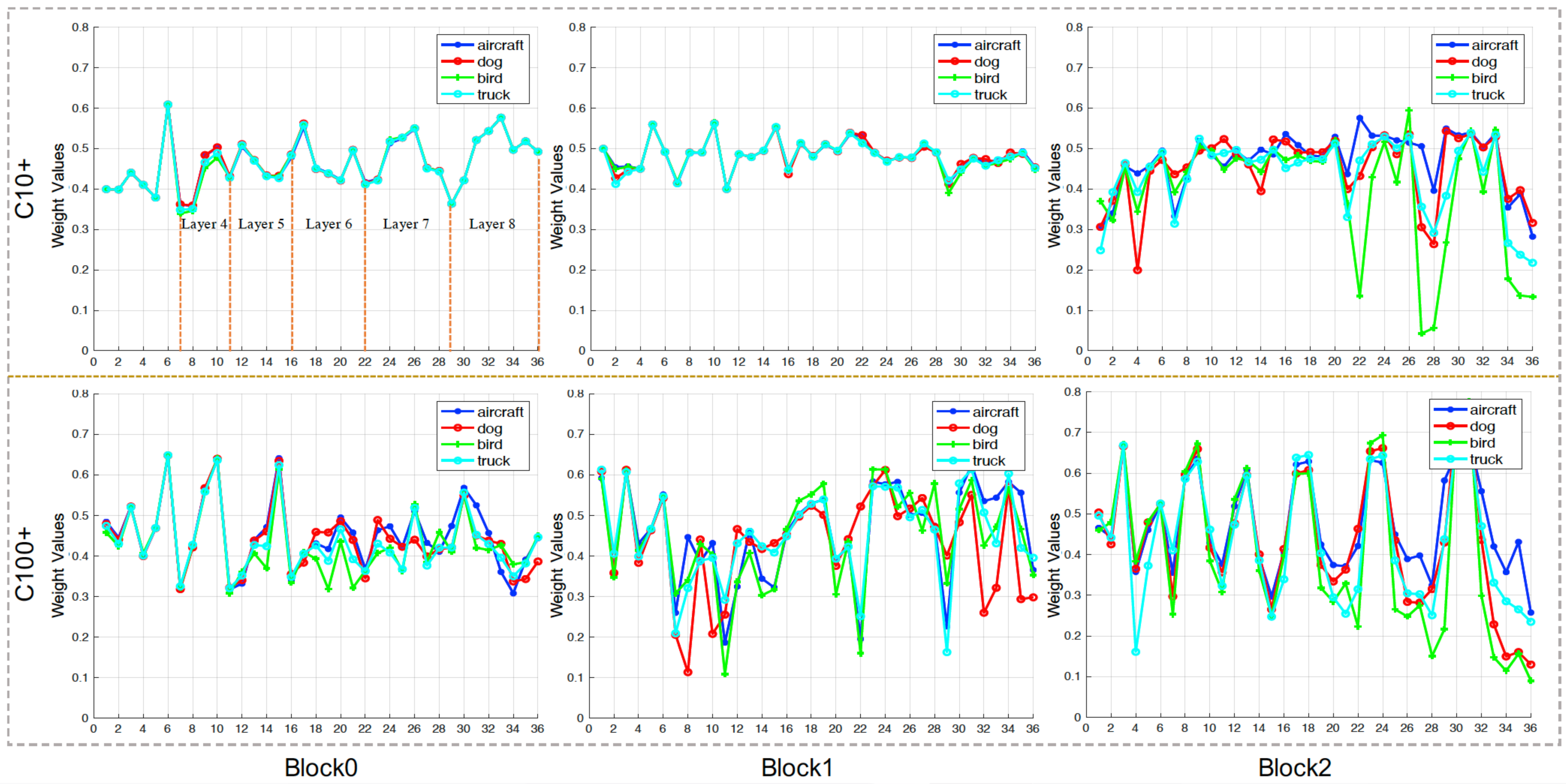
Quantitative Evaluation on Classification Accuracy

Model	Depth	Params	C10	C10+	C100	C100+	SVHN
Fractal Network [22]	21	38.6M	7.33	4.60	28.20	23.73	1.87
ResNet [8]	110	1.7M	13.63	6.41	44.74	27.22	2.01
Stochastic Depth [15]	1002	10.2M	-	4.91	-	-	-
ResNet(pre-act.) [9]	1001	10.2M	10.56	4.62	33.47	22.71	-
WRN-16 [44]	16	11.0M	-	4.81	-	22.07	1.54
WRN-28 [44]	28	36.5M	-	4.17	-	20.50	-
ResNeXt [39]	29	0.8M	-	6.74	-	26.48	-
SparseNet($n = 12$) [47]	40	0.8M	-	5.13	-	24.65	-
SparseNet($n = 24$) [47]	100	2.5M	-	4.64	-	22.41	-
SparseNet($n = 36$) [47]	100	5.7M	-	4.34	-	20.50	-
DenseNet($n = 12$) [14]	40	1.0M	7.00	5.24	27.55	24.42	1.79
DenseNet($n = 12$) [14]	100	7.0M	5.77	4.10	23.79	**20.20**	1.67
DenseNet($n = 12$) [14]	28	0.5M	7.36*	6.09*	29.17*	27.67*	1.82*
Our ADNet($n = 12$)	28	**0.6M**	**5.99**	**5.34**	**25.57**	**24.10**	**1.68**
DenseNet($n = 24$) [14]	28	2.0M	6.58*	4.83*	26.16*	24.56*	1.79*
Our ADNet($n = 24$)	28	**1.9M**	**5.23**	**4.40**	**23.20**	**22.59**	**1.59**
DenseNet($n = 40$) [14]	28	5.4M	5.99*	4.50*	25.78*	22.20*	1.71*
Our ADNet($n = 40$)	28	**4.7M**	**5.20**	**3.84**	**21.86**	**20.51**	**1.59**
DenseNet($n = 40$) [14]	36	8.2M	6.15*	4.30*	24.88*	22.05*	1.63*
Our ADNet($n = 40$)	36	**6.9M**	**5.13**	**3.96**	**20.52**	**20.20**	**1.54**

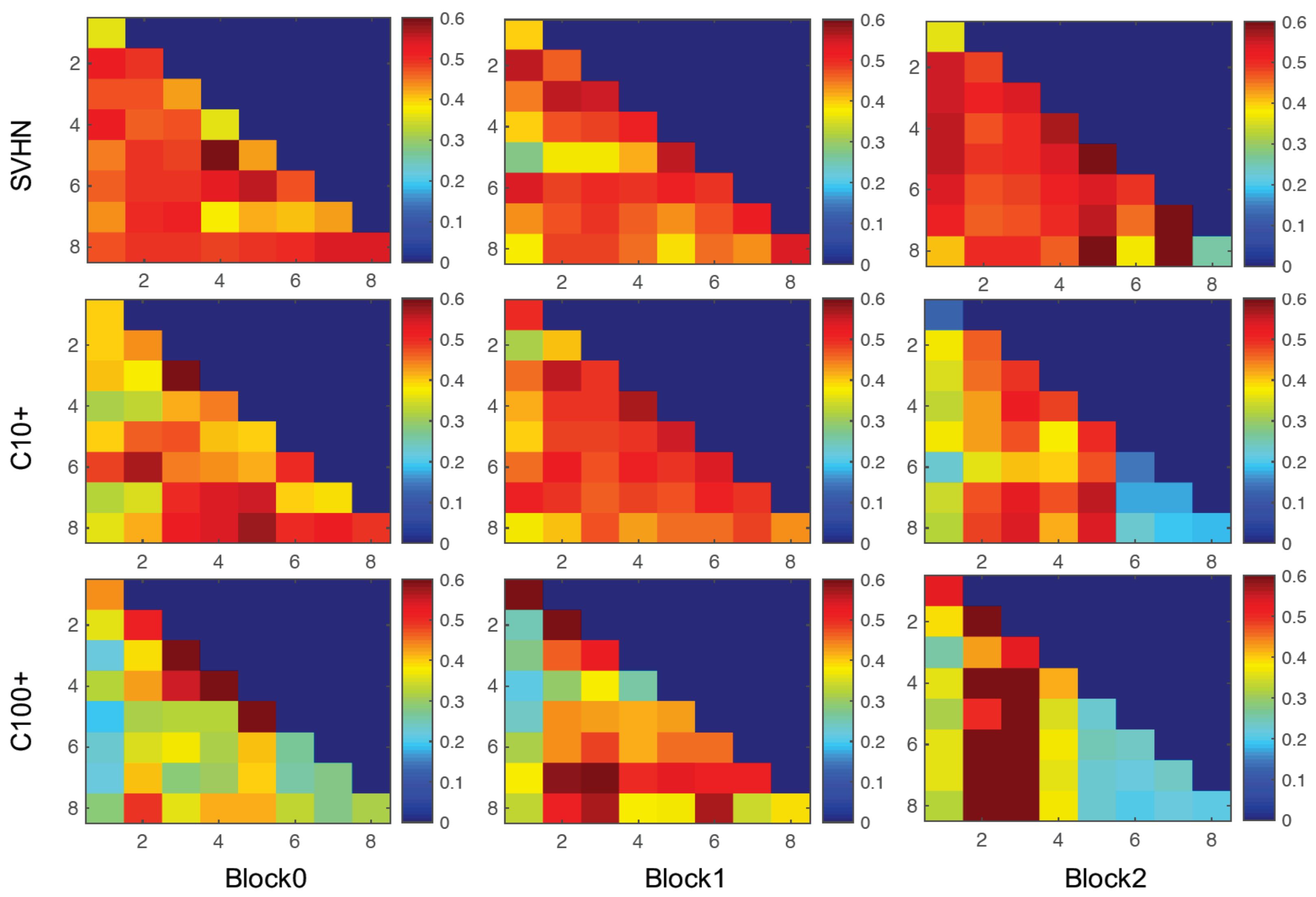
Accuracy vs. Computational Costs



Distributions of Learned Layer Attention Weights



Average Layer Attention Weights



- On simpler dataset (SVHN), the average weights have relatively high values (warmer colours).
- On more complex C100+, the weights show strong variations.
- Suggests that ADNet is capable of automatically determining the status of feature reuse.

Ablation Study on Layer Attention

Test Error (%) of 28-layer ADNet

Dataset	Growth rate	DenseNet	w.o. LA	w. LA
C10	n=12	7.36	6.30	5.99
	n=24	6.58	5.52	5.23
	n=40	5.99	5.42	5.20
C100	n=12	29.17	26.12	25.57
	n=24	26.16	23.89	23.20
	n=40	25.78	22.01	21.86
SVHN	n=12	1.82	1.75	1.68
	n=24	1.79	1.65	1.59
	n=40	1.71	1.64	1.59

Stochastic Multi-Scale Aggregation Network for Crowd Counting

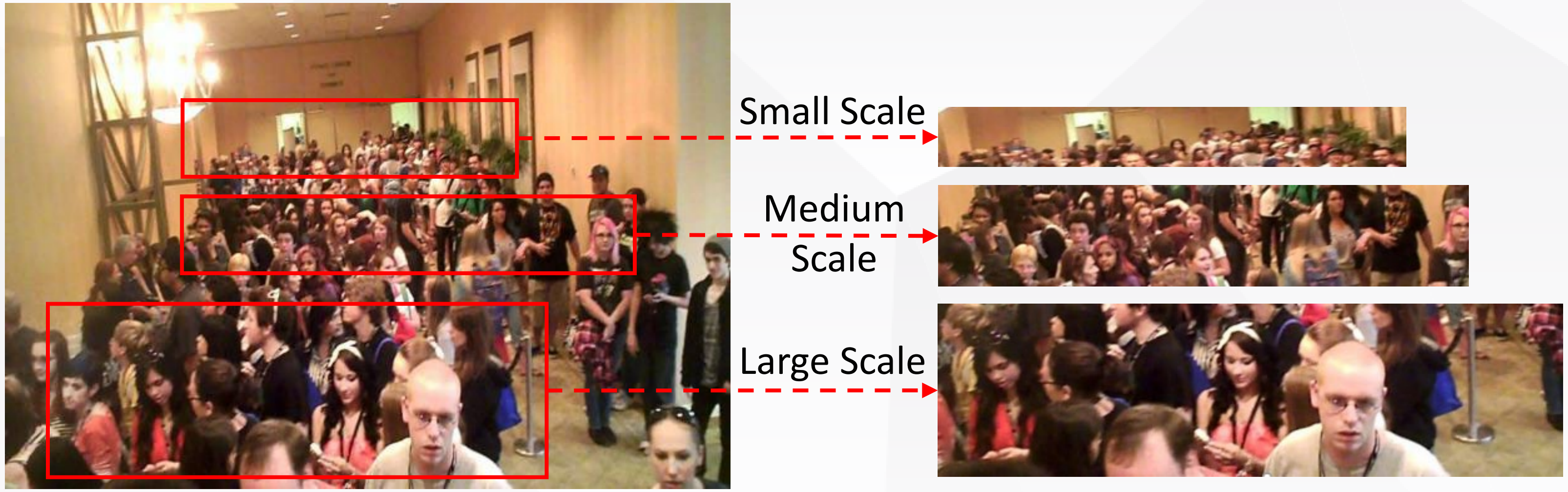
ICASSP 2020

Background on Crowd Counting

- Wide applications:
    - Security
    - Traffic control
    - Social distance monitoring
- Challenges:
    - Severe occlusions
    - Scale variation & density shift
    - Noisy background
    - Overfitting
- Approaches:
    - Detection-based
        - *Perform poorly on congested scenes.*
    - Regression-based
        - *Map to density maps then integrate.*

UNIVERSITY of GUELPH

Scale Variation & Density Shift

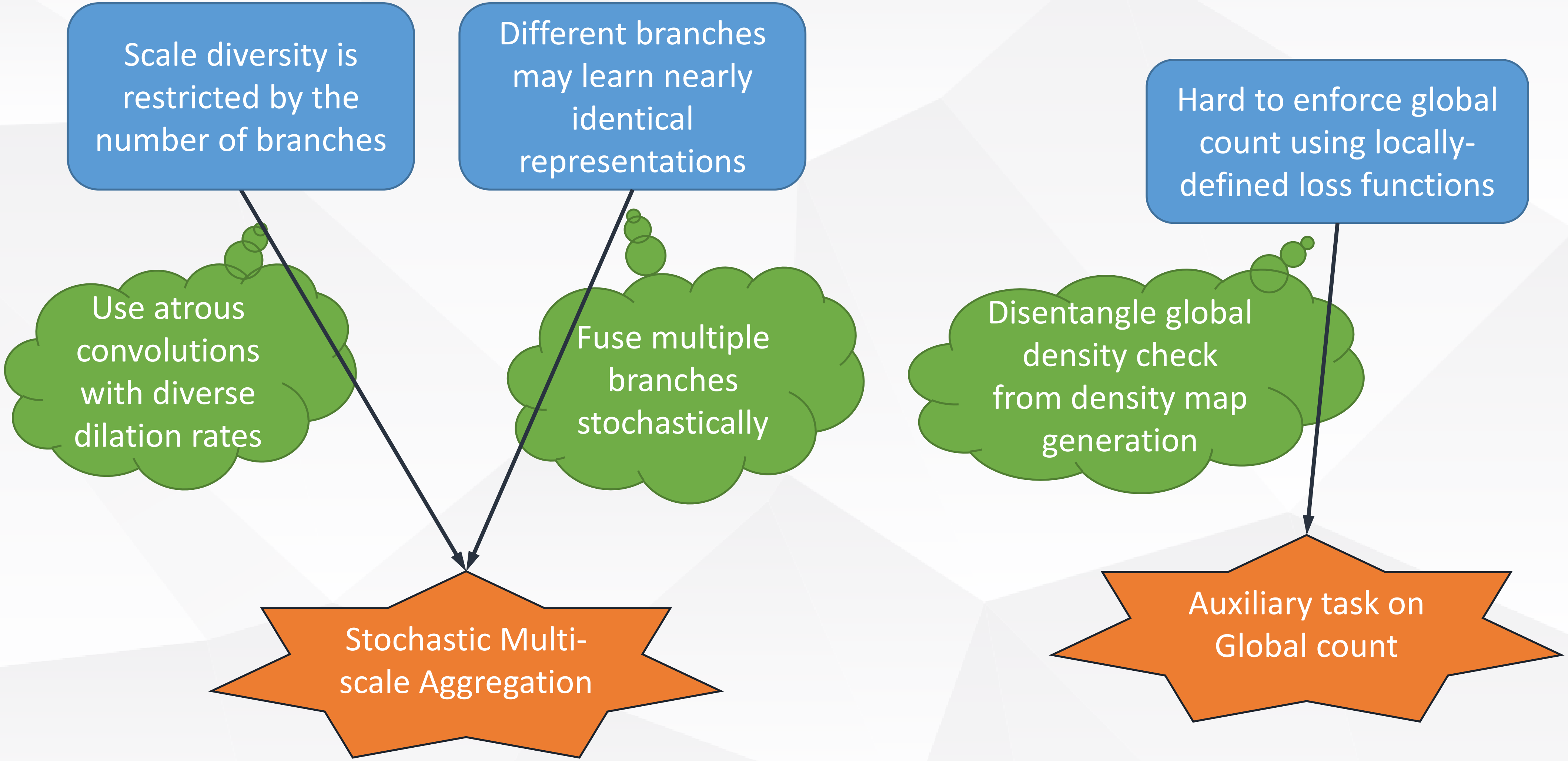


(a) Scale Variation

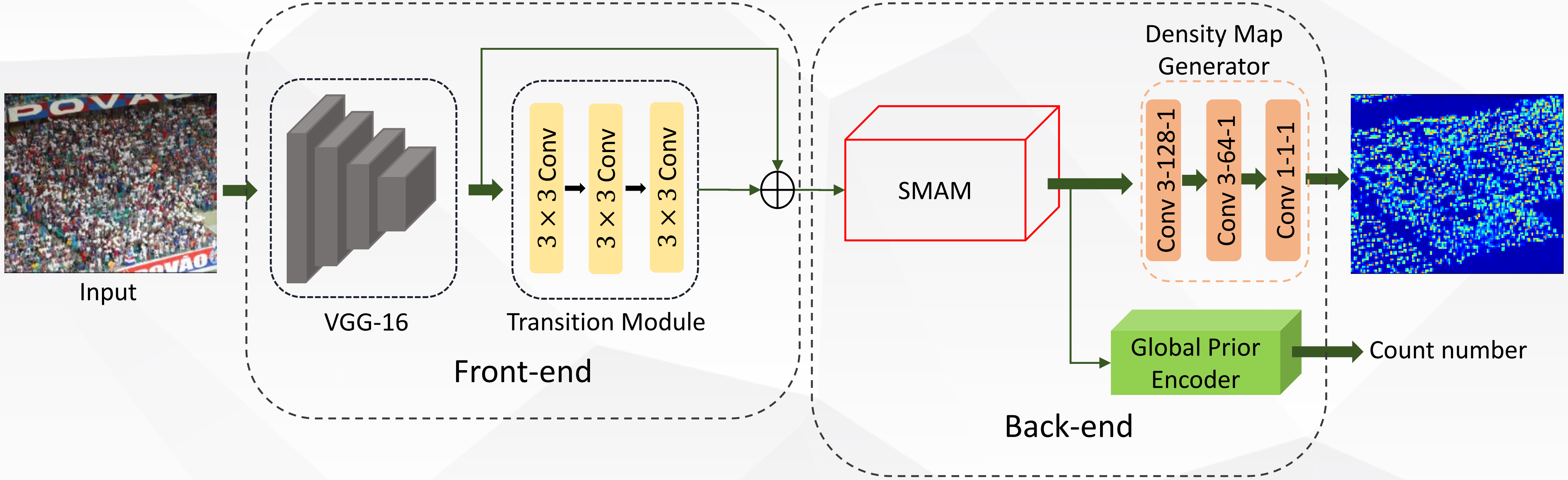
(b) Density Shift

UNIVERSITY of GUELPH

Motivations



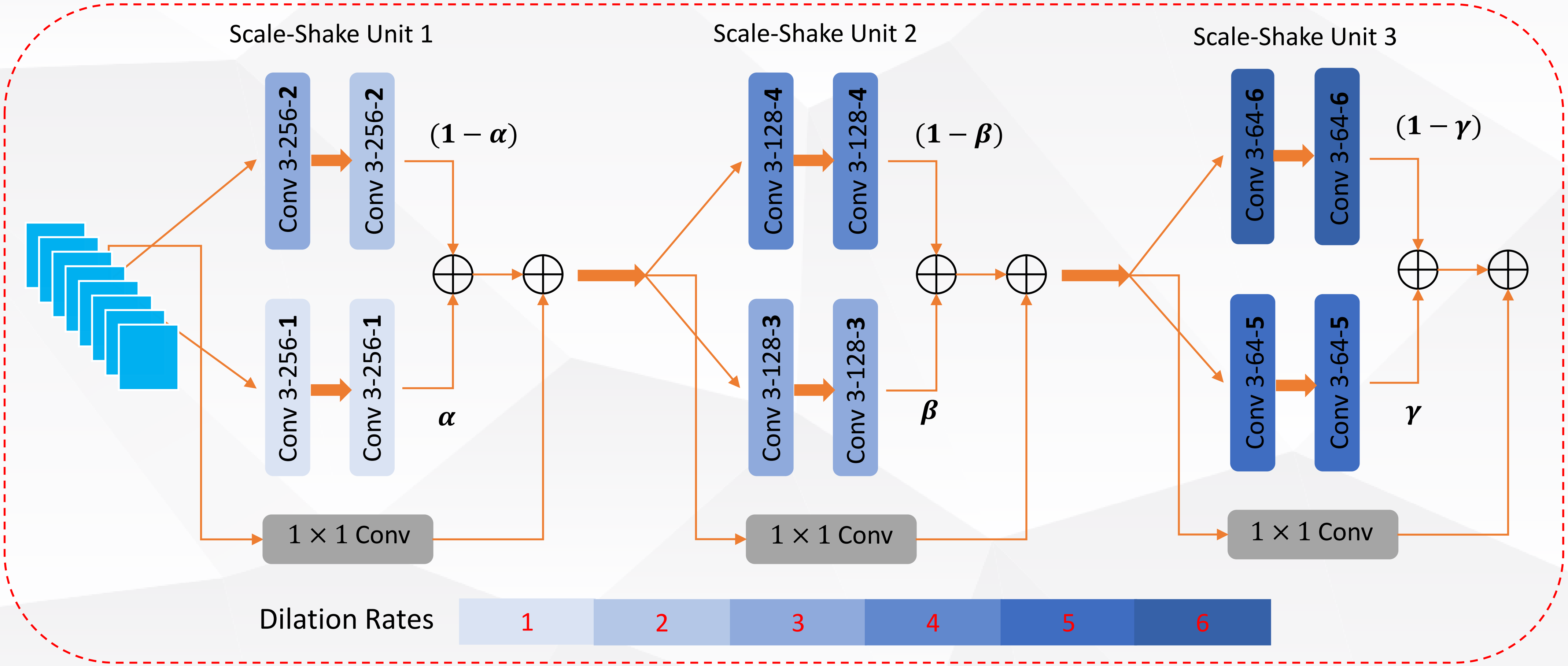
Overall Architecture of SMANet



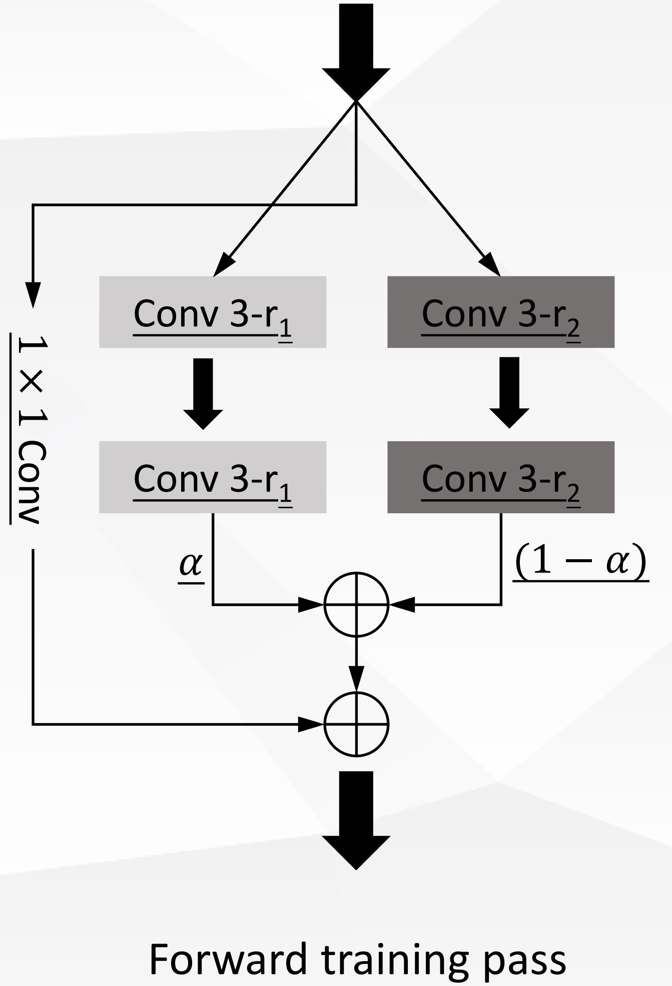
SMAM: Stochastic Multi-scale Aggregation Module
GPE: Global Prior Encoder

Stochastic Multi-scale Aggregation Module (SMAM)

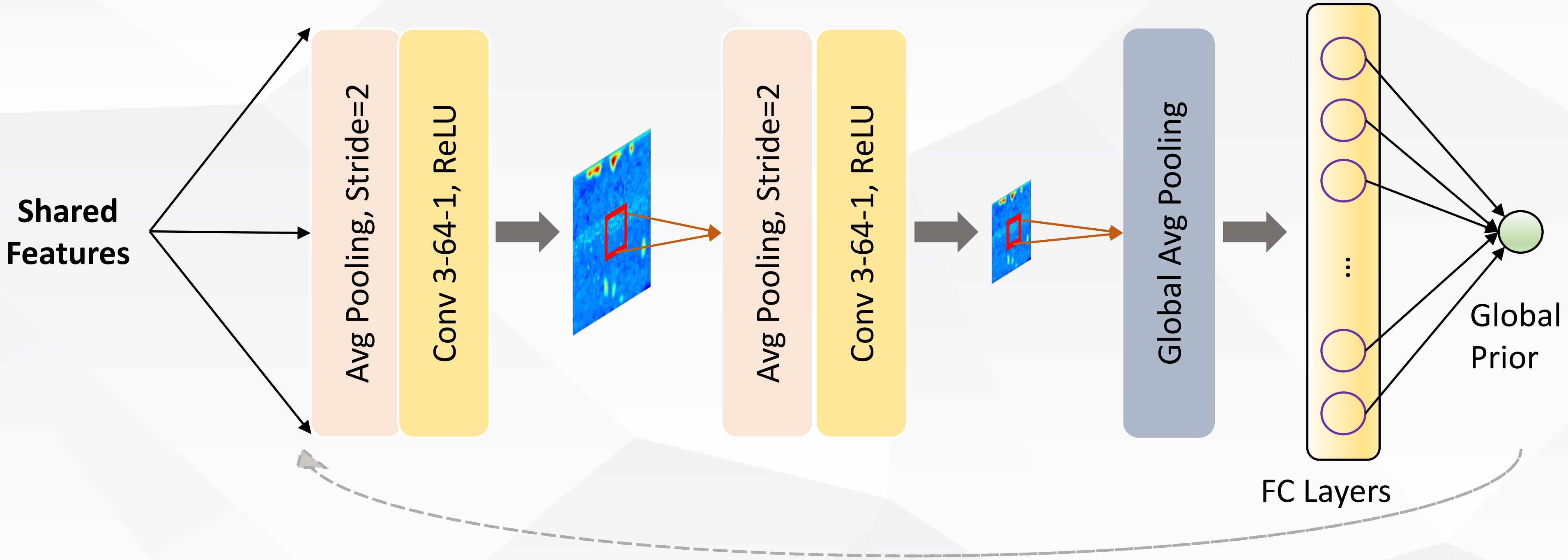
- Introduction
- Classification #1
- Classification #2
- Regression #1
- Regression #2
- Regression #3
- Summary



Training of Scale-Shake Unit



Global Prior Encoder (GPE)



- Directly generate total count from shared features.
    - Provide contraint on a global prior
    - Disentangle overall density level estimation from local density map generation.

Objective Function

$$L = (1-t)L_l + tL_g$$

$$L_l(W^l) = \frac{1}{N}\sum_{i=1}^{N}\left\|M(X_s^i; W^l) - M_i^{GT}\right\|_2^2$$

Local Density Estimation
M^{GT}: Ground truth density map

$$L_g(W^g) = \frac{1}{N}\sum_{i=1}^{N}\left\|P(X_s^i; W^g) - S_i^{GT}\right\|_2^2$$

Global Prior Consistency
S^{GT}: Ground truth count

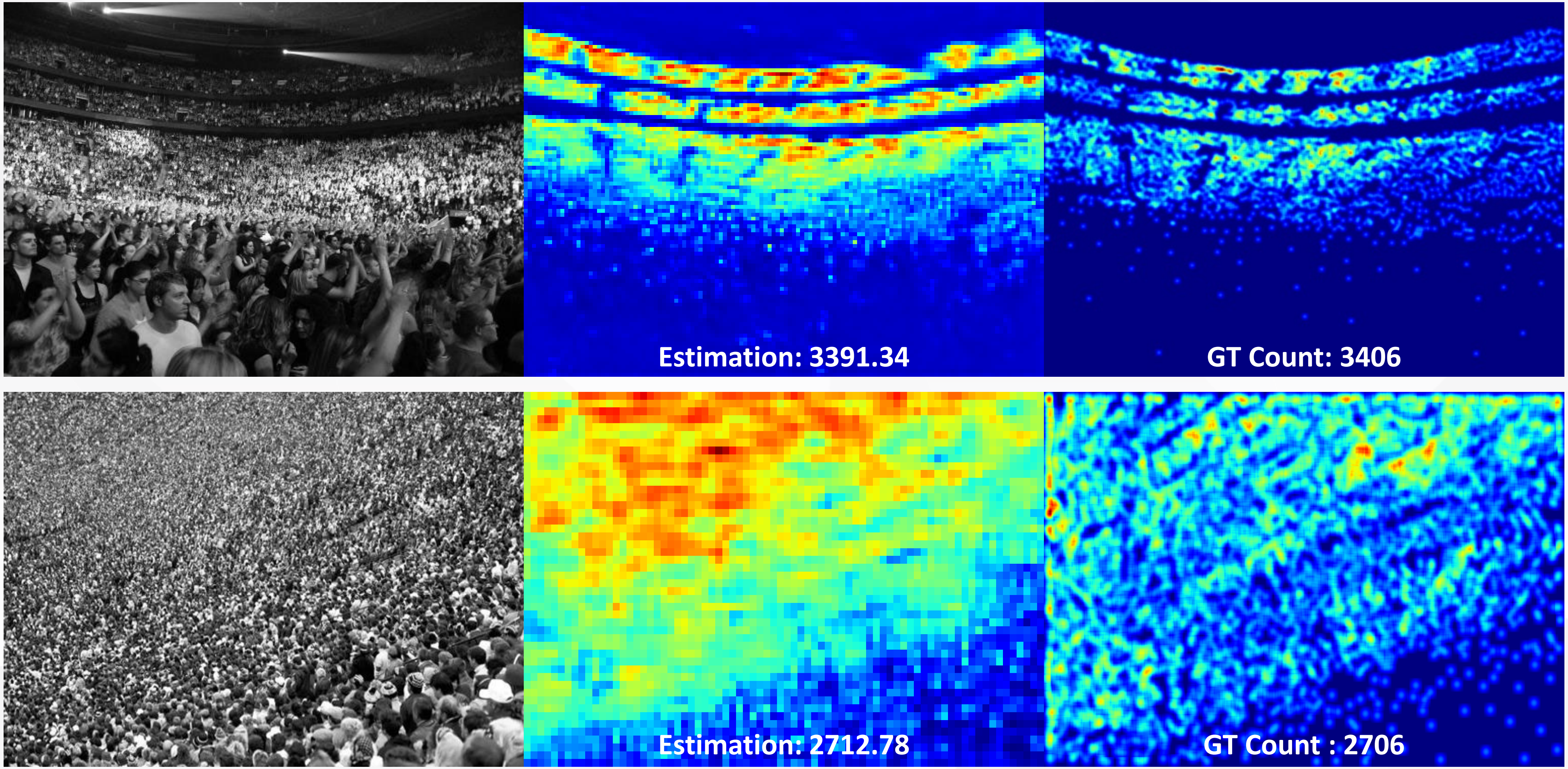
N: Batch size
t: hyper-parameter for balancing between local and global supervisions

Quantitative Comparison on Counting Accuracy

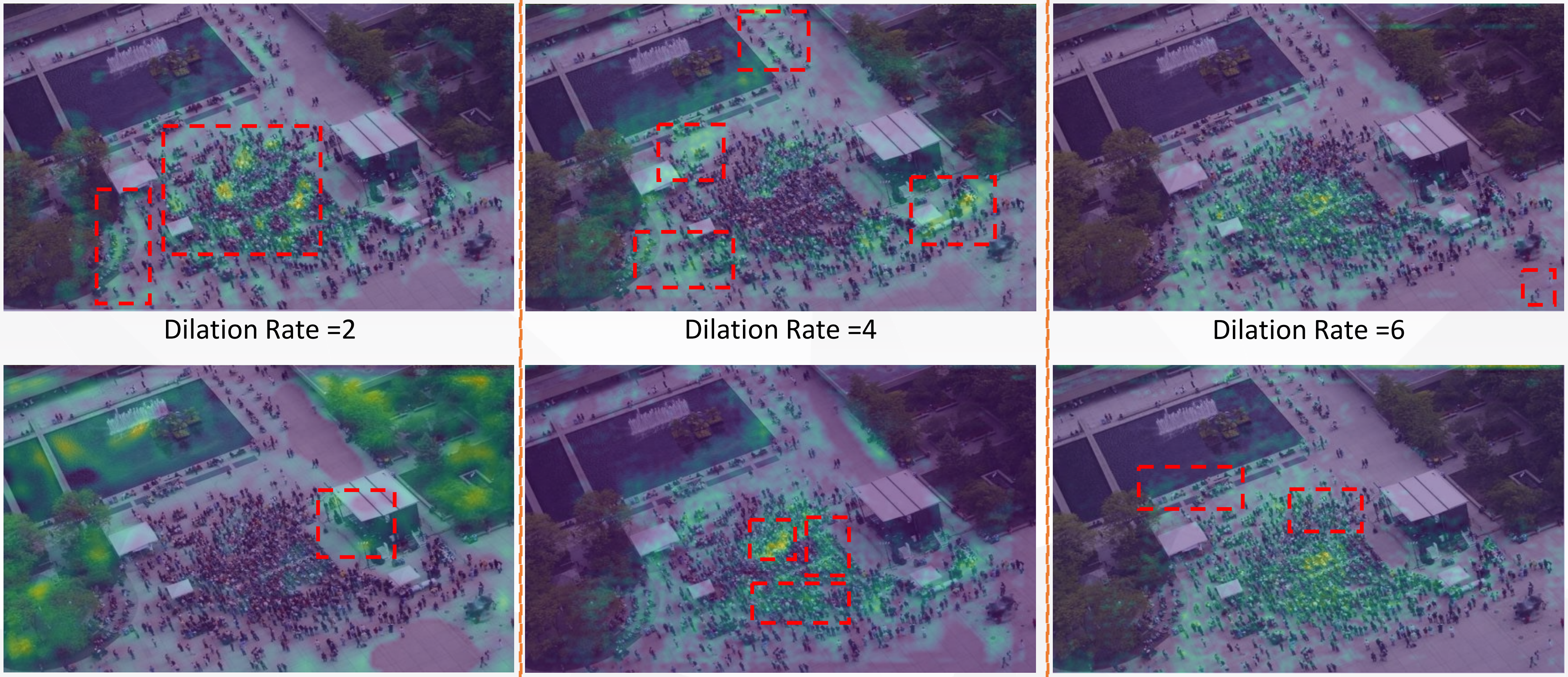
Introduction
Classification #1
Classification #2
Regression #1
Regression #2
Regression #3
Summary

Methods	Part_A MAE	Part_A MSE	Part_B MAE	Part_B MSE	UCF-QNRF MAE	UCF-QNRF MSE	UCF_CC_50 MAE	UCF_CC_50 MSE
D-ConvNet [20] *(CVPR18)*	73.5	112.3	18.7	26.0	-	-	288.4	404.7
CSRNet [17] *(CVPR18)*	68.2	115.0	10.6	16.0	-	-	266.1	397.5
SANet [1] *(ECCV18)*	67.0	104.5	8.4	13.6	-	-	258.4	334.9
TEDNet [26] *(CVPR19)*	64.2	109.1	8.2	12.8	113	188	249.4	354.5
ADCrowdNet [18] *(CVPR19)*	63.2	98.9	7.6	13.9	-	-	257.1	363.5
PACNN + CSRNet [27] *(CVPR19)*	62.4	102.0	7.6	11.8	-	-	241.7	320.7
CAN [14] *(CVPR19)*	62.3	100.0	7.8	12.2	107	183	212.2	243.7
SPN+L2SM [15] *(ICCV19)*	64.2	98.4	7.2	11.1	104.7	173.6	188.4	315.3
MBTTBF-SCFB [28] *(ICCV19)*	60.2	94.1	8.0	15.5	97.5	165.2	233.1	300.9
SMANet (ours)	**59.7**	102.1	7.3	12.9	**92.5**	176.7	**178.4**	256.3

Density Maps for Congested Scenes

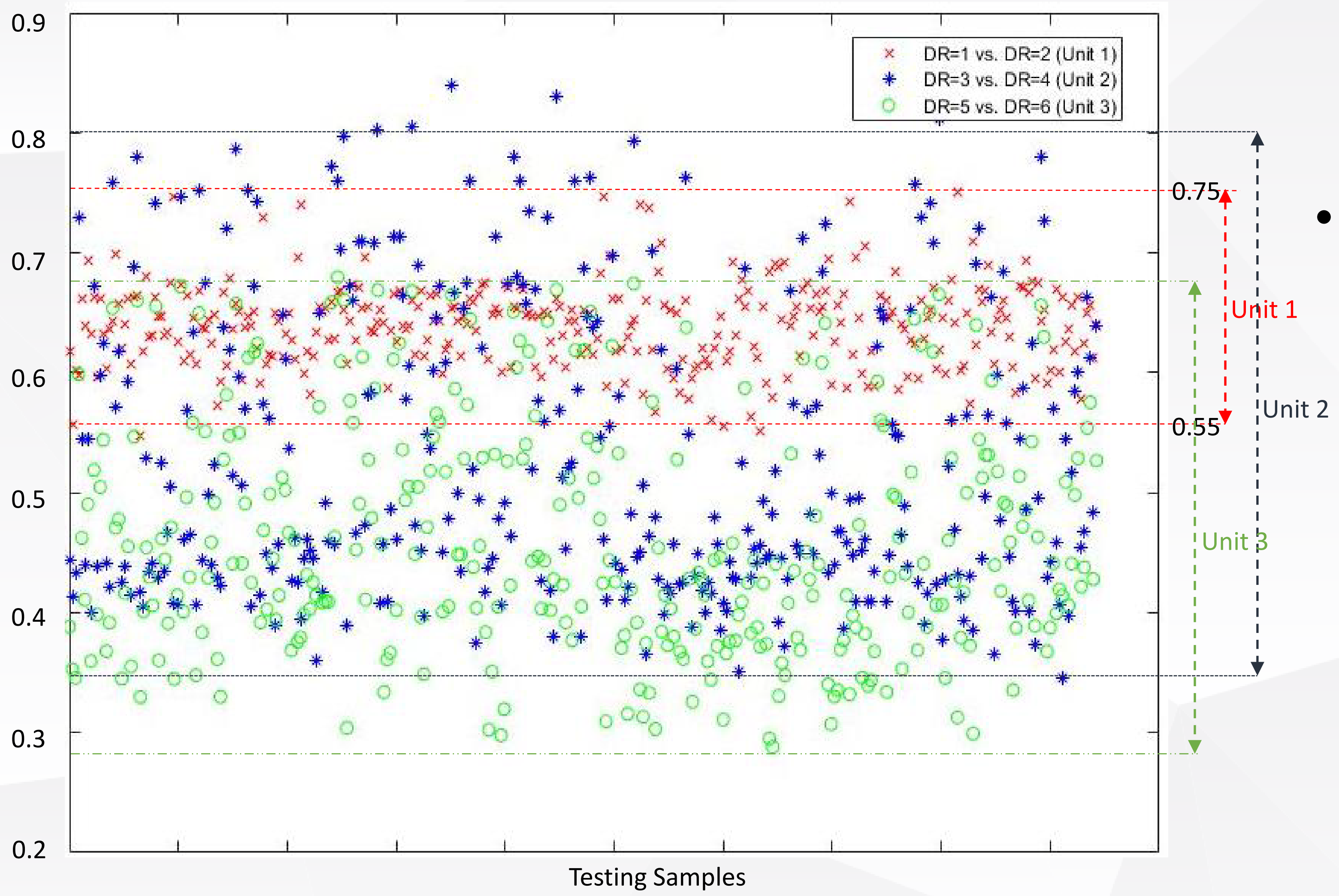


Visualization of Multi-scale Feature Maps



Dilation Rate =2
Dilation Rate =4
Dilation Rate =6
Dilation Rate =1
Dilation Rate =3
Dilation Rate =5

Feature Cosine Similarity between Branches



- Feature maps generated by two branches with different dilation rates are not similar.

Ablation Study on GPE

Datasets	w.o. GPE	w. GPE
Part_A	62.4	59.1 (↓ **5.2%**)
Part_B	7.5	7.2 (↓ **4.5%**)
UCF-QNRF	97.2	92.2 (↓ **5.2%**)

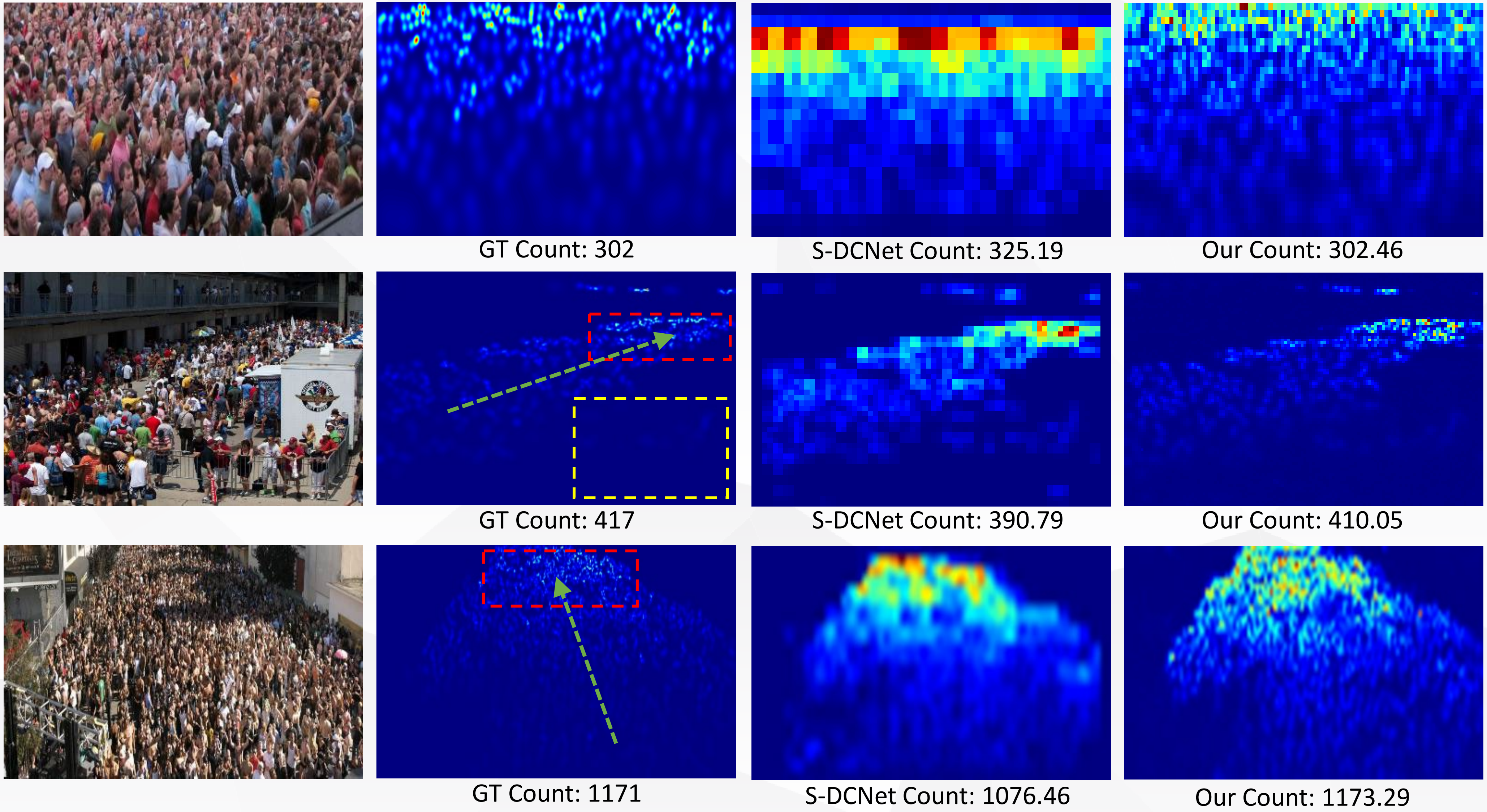
- Using the the global consistency (GPE) constantly performs better than withouth using it.

UNIVERSITY of GUELPH

Interlayer and Intralayer Scale Aggregation for Scale-Invariant Crowd Counting

Neurocomputing 2021

Previous Work: S-DCNet



Motivations

Scale varies continuously in congested scenes

Extract and fuse fine grain scale information

Multi-branch approaches are hard to train and tend to overfit

Use a single branch (column)

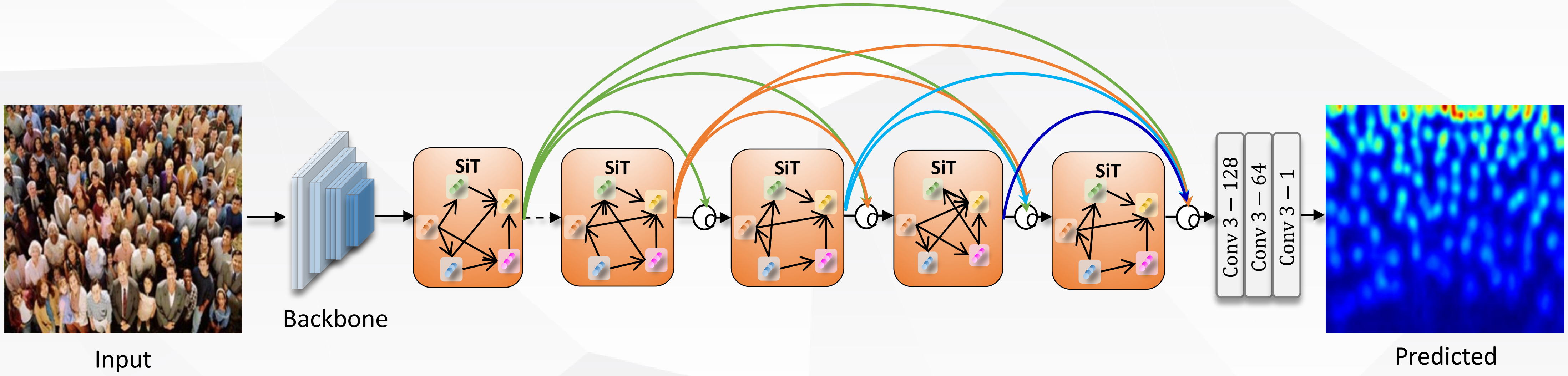
Single-column Scale-invariant Architecture

Huge variation in total count (density shift) demands high generalization capability

Enlarge the diversity of training data

Randomly integrated loss function

Overall Architecture of ScSiNet

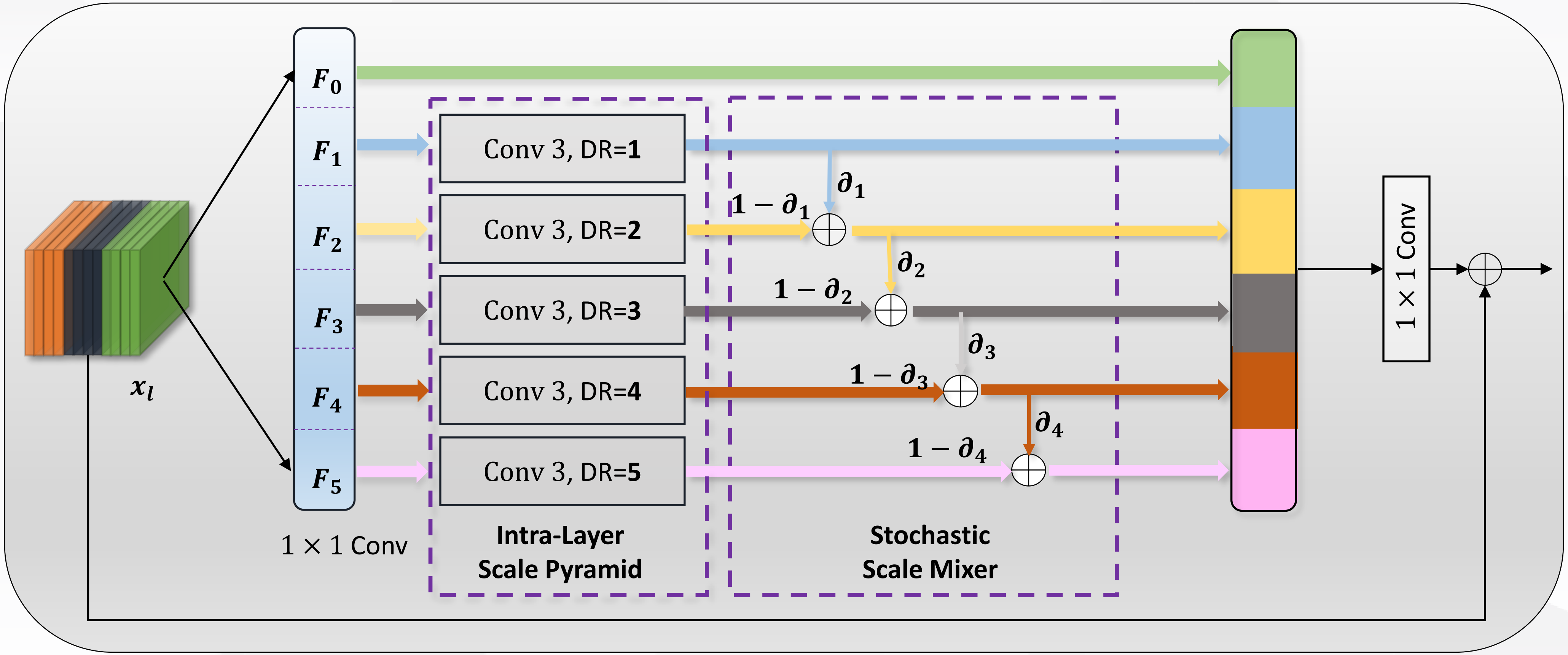


SiT: Scale-invariant Transformation

- The dense connections between different SiTs provide inter-layer (coarse grain) scale aggregation.

UNIVERSITY of GUELPH

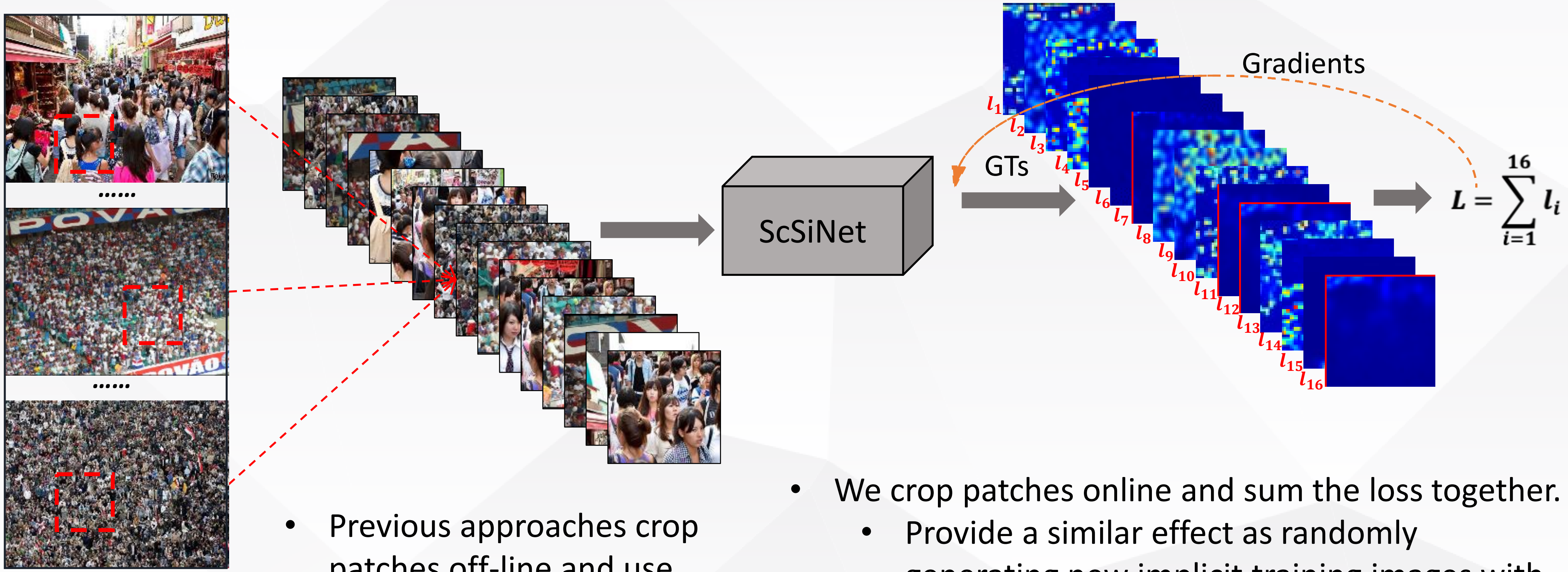
Scale-invariant Transformation



- Each SiT contains a pyramid of dilated convolution filters, which provides intra-layer (fine grain) scale fusion.

UNIVERSITY of GUELPH

Randomly Integrated Loss



- Previous approaches crop patches off-line and use them repetitively.
- Average loss is used for training.

- We crop patches online and sum the loss together.
  - Provide a similar effect as randomly generating new implicit training images with larger range of density levels.
- Use datasets with different image resolutions directly without the needs for resizing.

Quantitative Comparison on Counting Accuracy

Introduction

Classification #1

Classification #2

Regression #1

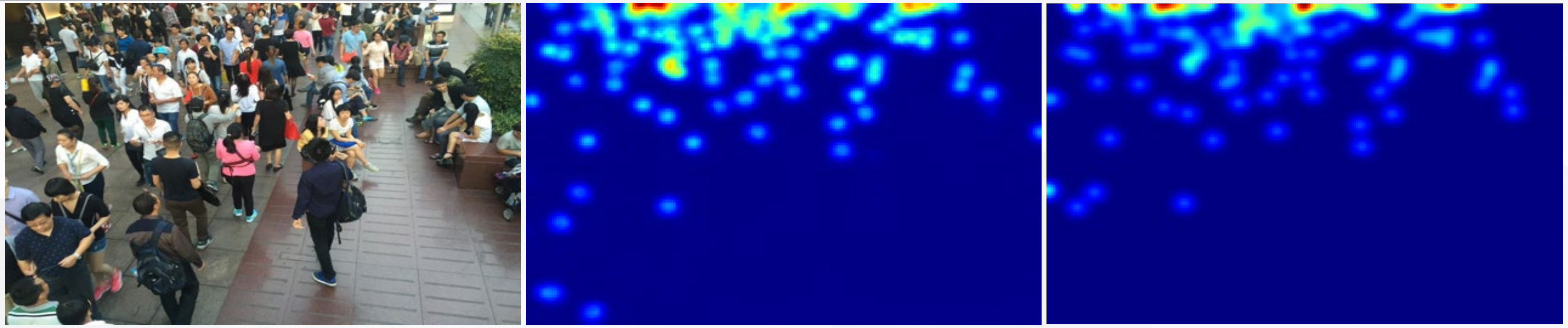
Regression #2

Regression #3

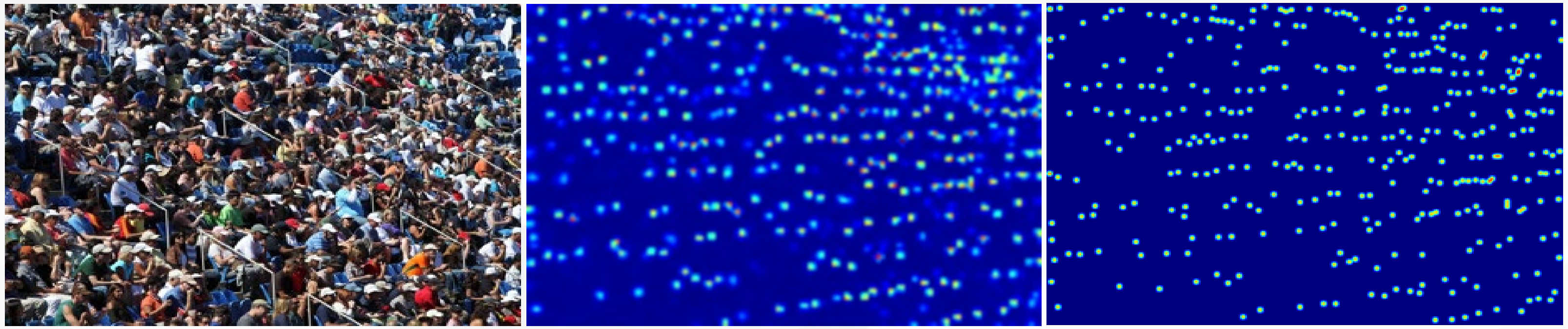
Summary

Methods	Part_A		Part_B		UCF-QNRF		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
TEDNet [13]	64.2	109.1	8.2	12.8	113	188	249.4	354.5
ADCrowdNet [44]	63.2	98.9	7.6	13.9	-	-	257.1	363.5
PACNN + CSRNet [27]	62.4	102.0	7.6	11.8	-	-	241.7	320.7
CAN [15]	62.3	100.0	7.8	12.2	107	183	212.2	243.7
CFF [55]	65.2	109.4	7.2	12.2	-	-	-	-
SPN+L2SM [18]	64.2	98.4	7.2	11.1	104.7	173.6	188.4	315.3
MBTTBF-SCFB [17]	60.2	94.1	8.0	15.5	97.5	165.2	233.1	300.9
PGCNet [56]	57.0	**86.0**	8.8	13.7	-	-	244.6	361.2
BL [42]	62.8	101.8	7.7	12.7	**88.7**	**154.8**	229.3	308.2
DSSINet [16]	60.63	96.04	6.85	10.34	99.1	159.2	216.9	302.4
SPANet+SANet [47]	59.4	92.5	**6.5**	**9.9**	-	-	232.6	311.7
S-DCNet [19]	58.3	95.0	6.7	10.	104.4	176.1	204.2	301.3
ScSiNet (proposed)	**55.77**	90.23	6.79	10.95	89.69	178.46	**154.87**	**199.42**

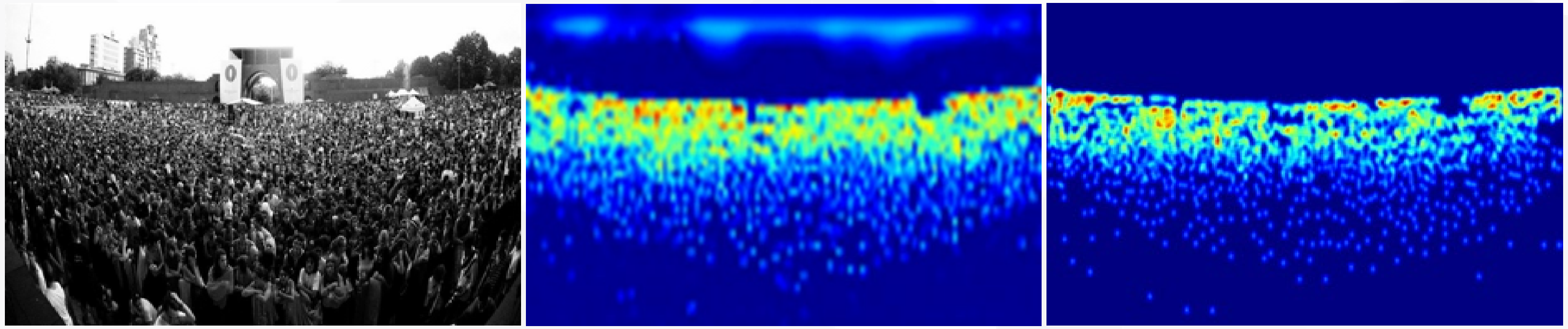
Qualitative Evaluation on Density Maps



Our Count: 109.99 GT Count: 110



Our Count: 353.22 GT Count: 353



Our Count: 1859.39 GT Count: 1858

UNIVERSITY of GUELPH

Ablation Studies

- Introduction
- Classification #1
- Classification #2
- Regression #1
- Regression #2
- Regression #3
- Summary

	No Inter (G=6)	No Intra (G=1)	G=2	G=4	G=6	G=8
Parameters	11.0M	20.9M	11.7M	12.9M	14.1M	15.3 M
MAE	57.00	60.11	60.04	58.85	**55.77**	57.50
MSE	96.77	104.67	104.50	95.83	**90.23**	98.60

- The impacts of inter/intra-layer scale fusion and the number of groups (G) in SiT on ShanghaiTech Part A dataset.

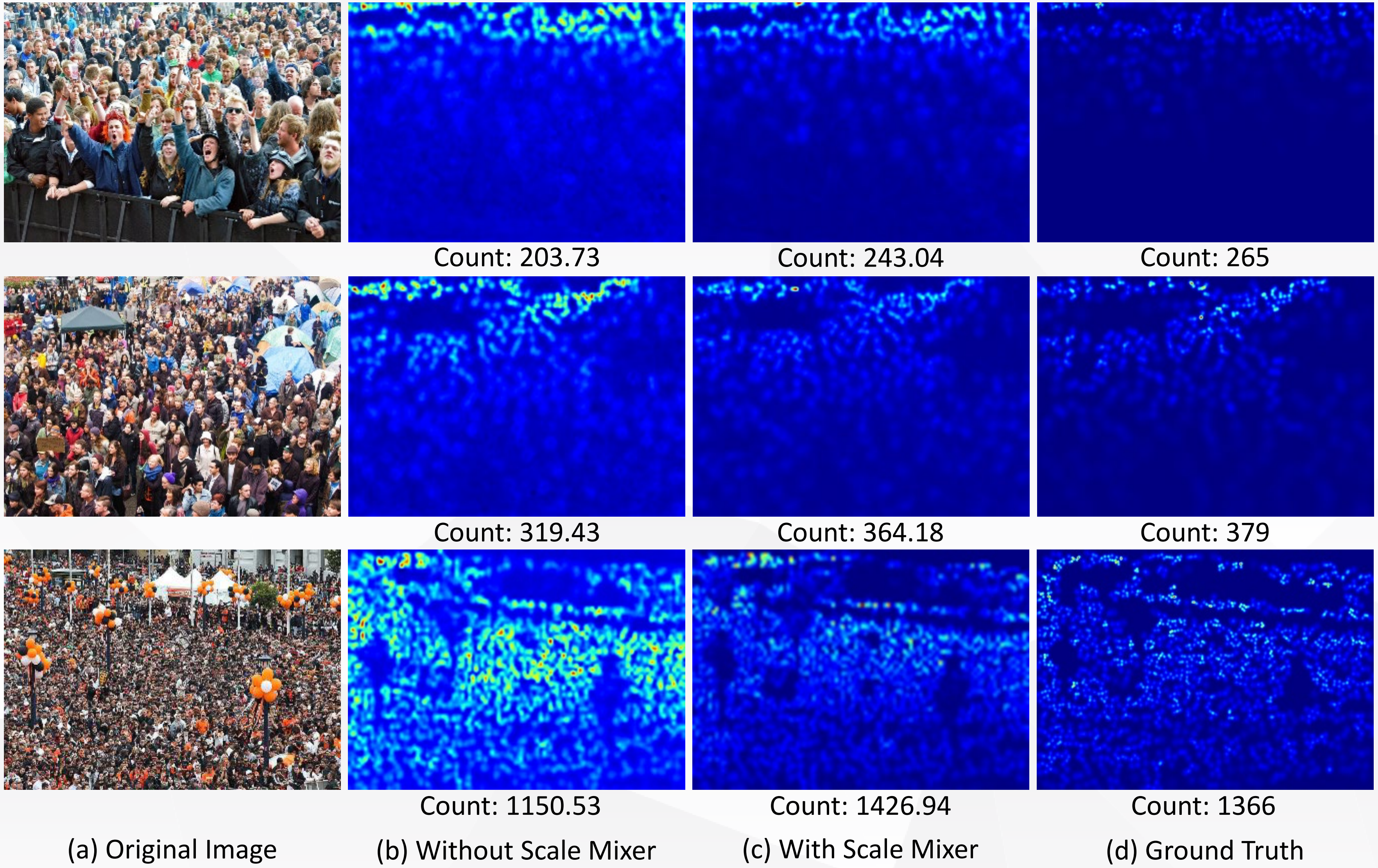
- The effectiveness of stochastic scale mixer on ShanghaiTech Part A (A), Part B (B), and UCF-QNRF (Q) datasets.

	w.o Mixer		w. Mixer (α=1)		w. Mixer (α)	
	MAE	MSE	MAE	MSE	MAE	MSE
A	61.05	100.27	57.60	93.12	55.77	90.23
B	7.01	11.53	6.96	11.21	6.79	10.95
Q	91.12	170.80	-	-	89.69	178.46

	w.o. Mini-Batch		w. Mini-Batch	
	MAE	MSE	MAE	MSE
Part_A	69.2	116.1	55.77	90.23
UCF_CC_50	229.53	343.27	154.87	199.42

- The effectiveness of the proposed uniform mini-batch training on ShanghaiTech Part A and UCF_CC_50.

Effectiveness of Stochastic Scale Mixer



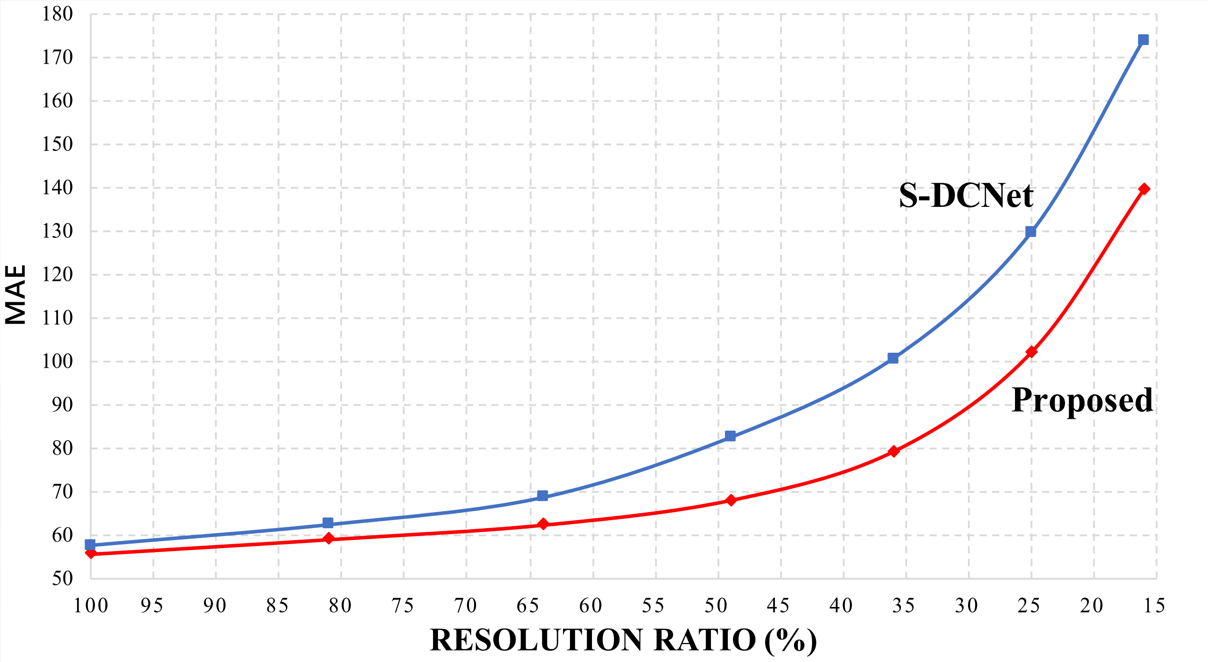
Cross-dataset Transferability and Scale-invariant Tests

Introduction
Classification #1
Classification #2
Regression #1
Regression #2
Regression #3
Summary

Methods	Part_A→Part_B		Part_B→Part_A		Part_A→UCF-QNRF		UCF-QNRF→Part_A	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
D-ConvNet [45]	49.1	99.2	140.4	226.1	-	-	-	-
SPN [18]	23.8	44.2	131.2	219.3	236.3	428.4	87.9	126.3
SPN+L2SM [18]	21.2	38.7	126.8	**203.9**	227.2	405.2	73.4	119.4
ScSiNet	**20.96**	**36.38**	**118.19**	214.13	**194.63**	**370.85**	**69.23**	**107.44**

- Cross-dataset evaluation for transferability comparison.

- Scale-invariant Tests using down-sampled images

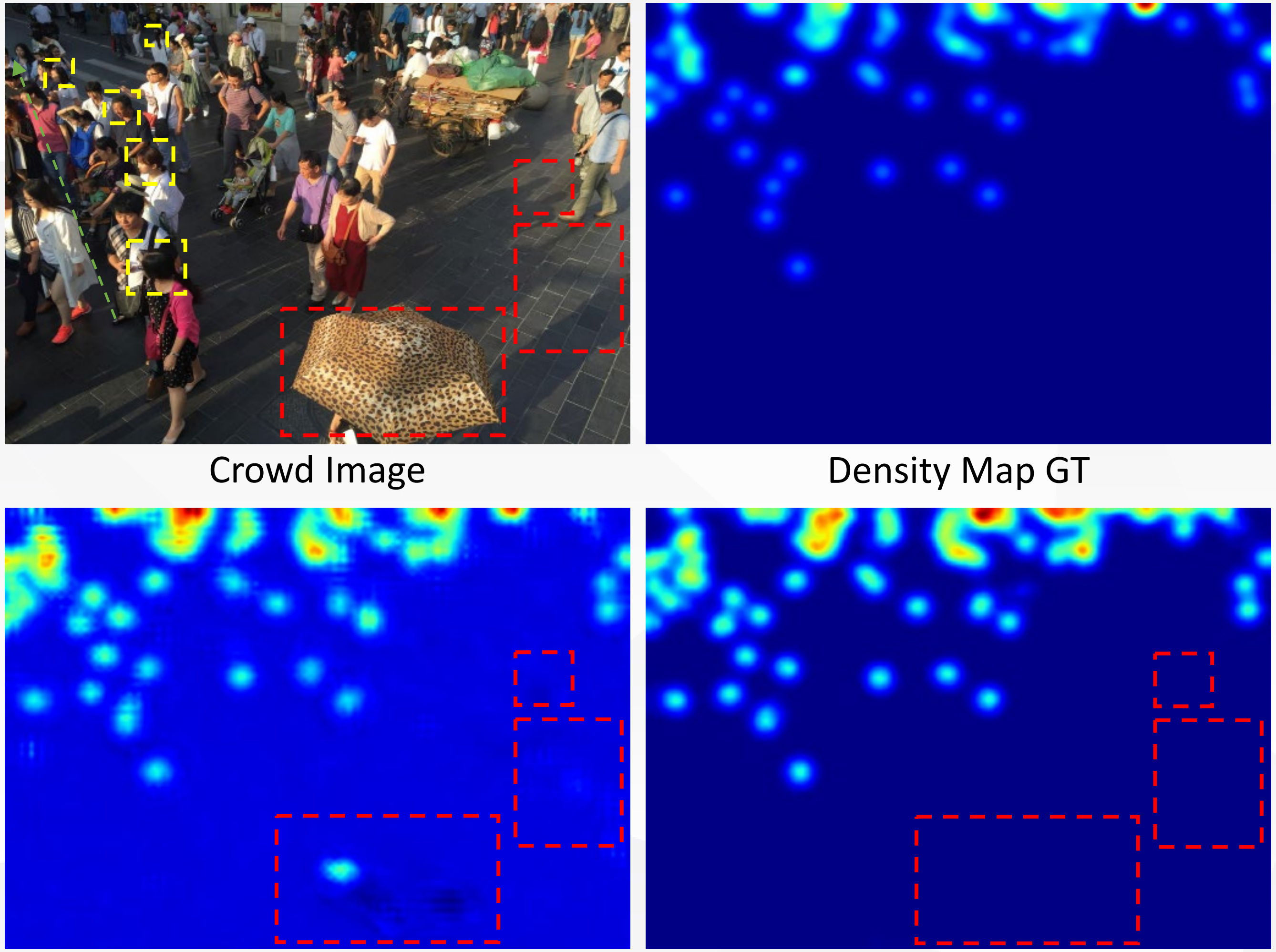


STNet: Scale Tree Network with Multi-level Auxiliator for Crowd Counting

TMM, accepted with minor revision

UNIVERSITY of GUELPH

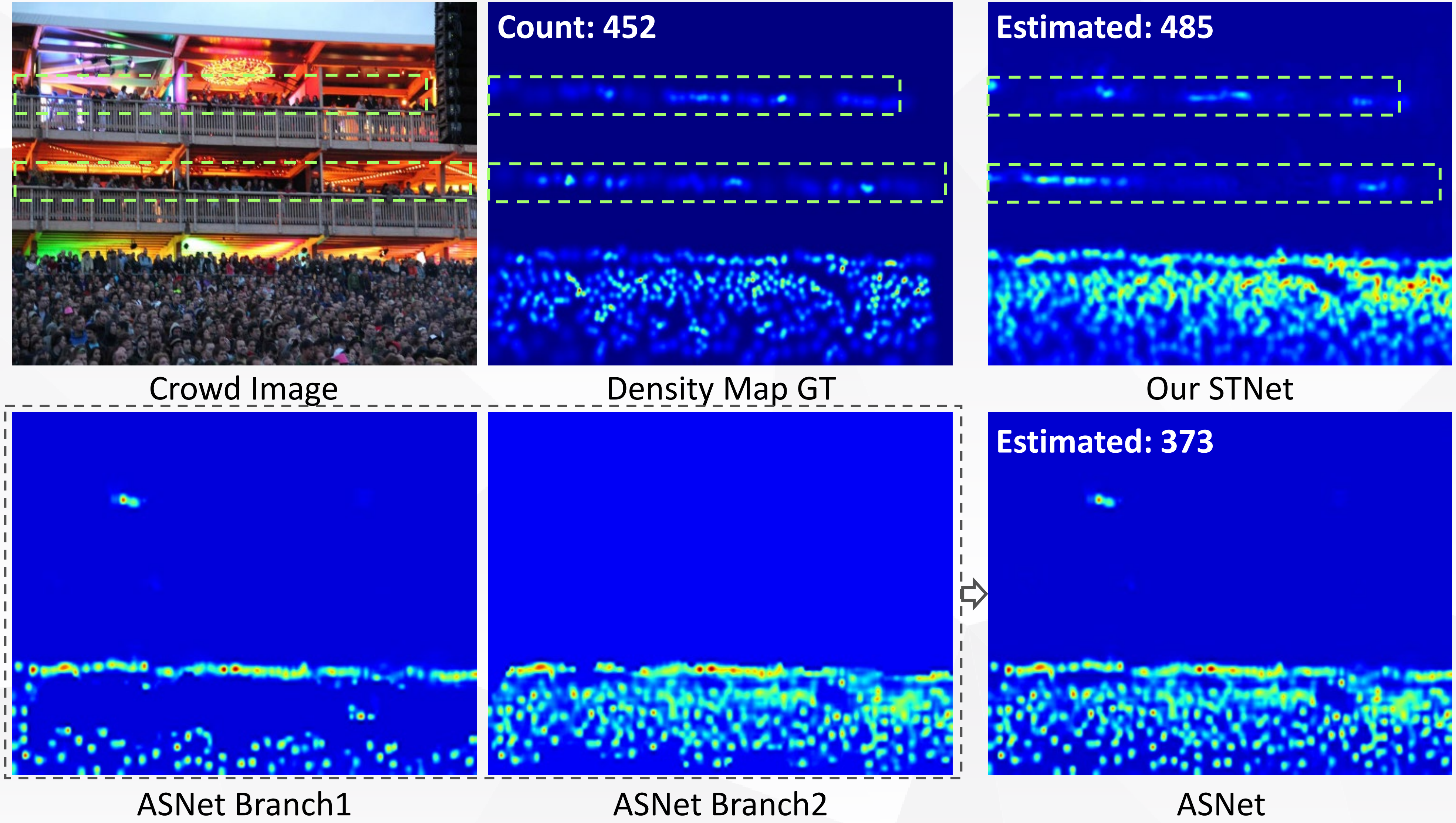
Previous Work: CANet


Crowd Image
Density Map GT
CANet
Our STNet

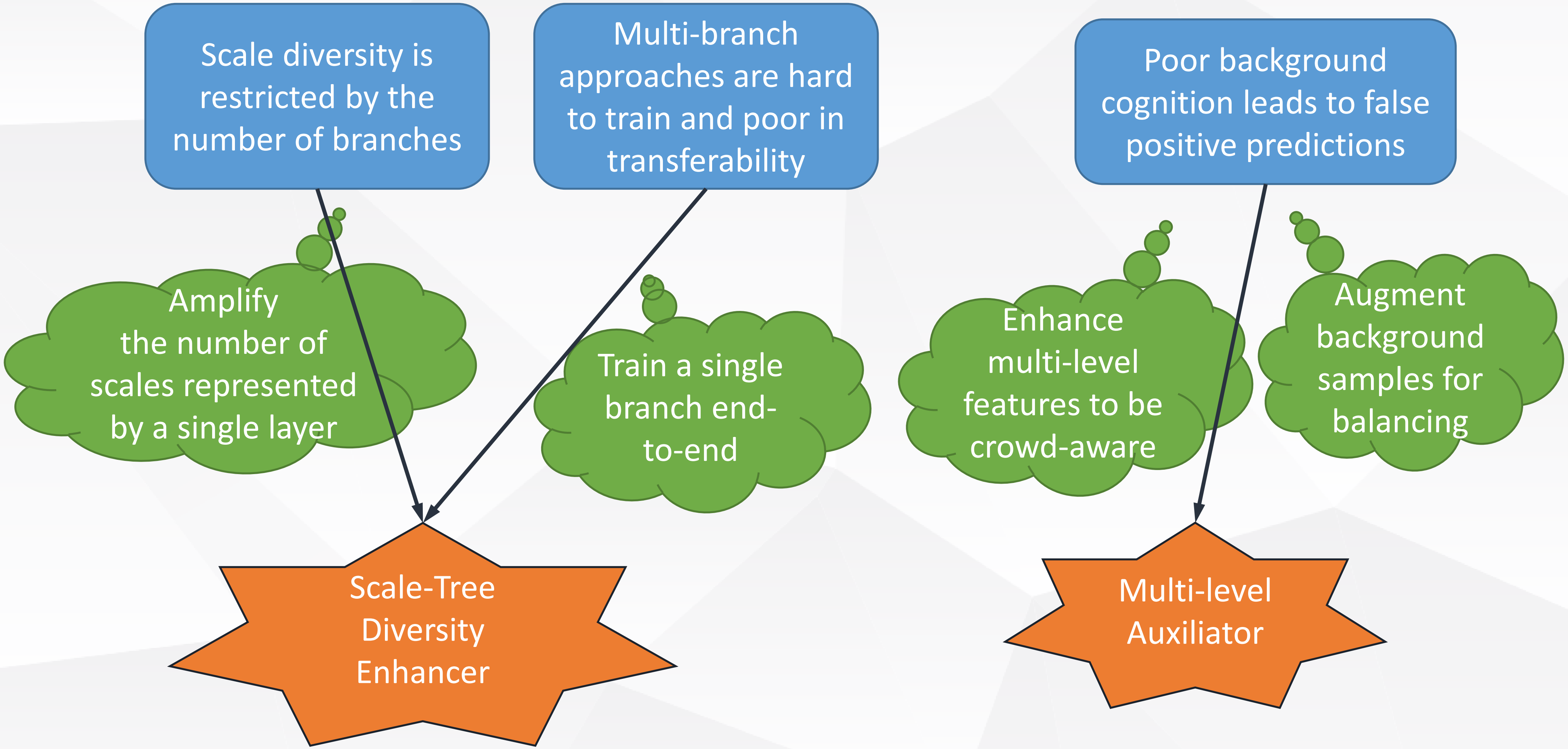
- Confused by complex background noises (red boxes).
- Has difficulty handling scale variations (yellow boxes).

UNIVERSITY of GUELPH

Previous Work: ASNet



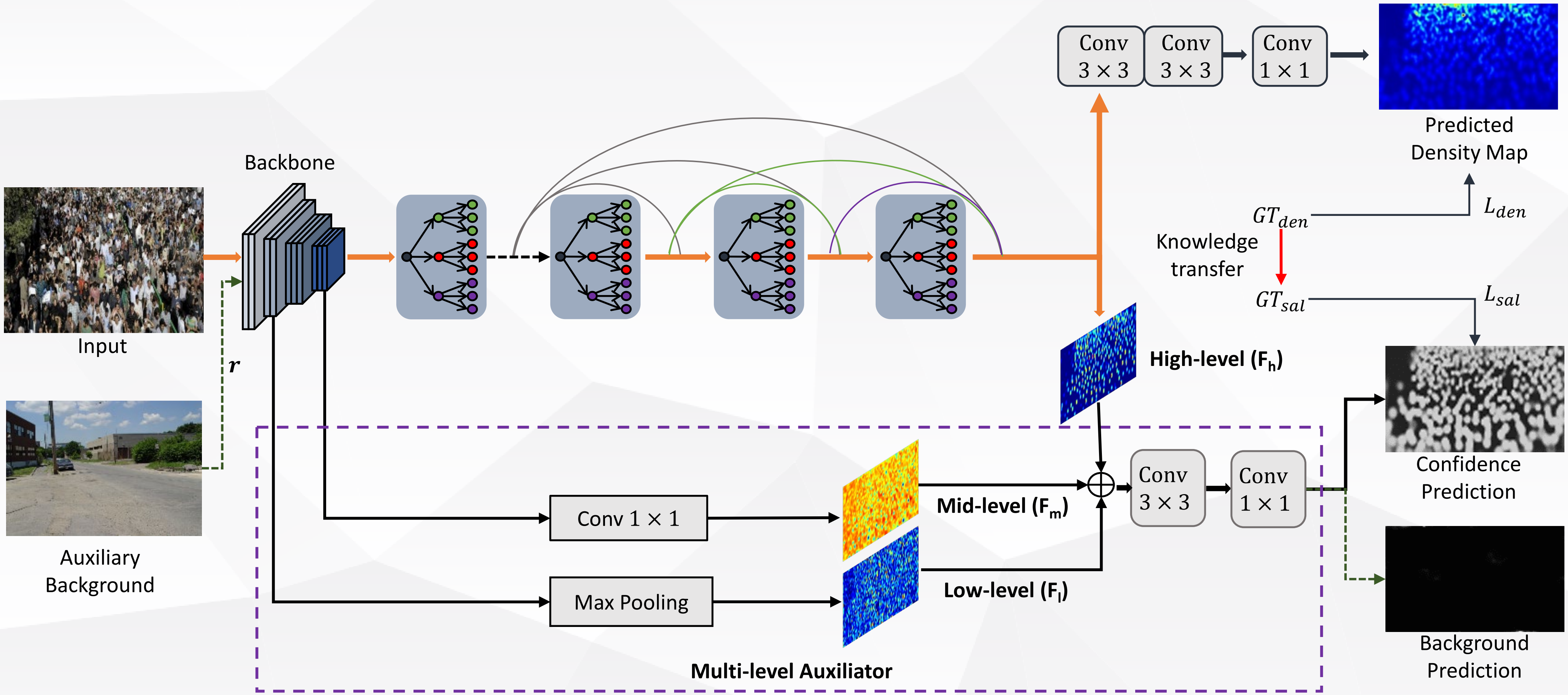
Motivations



UNIVERSITY of GUELPH

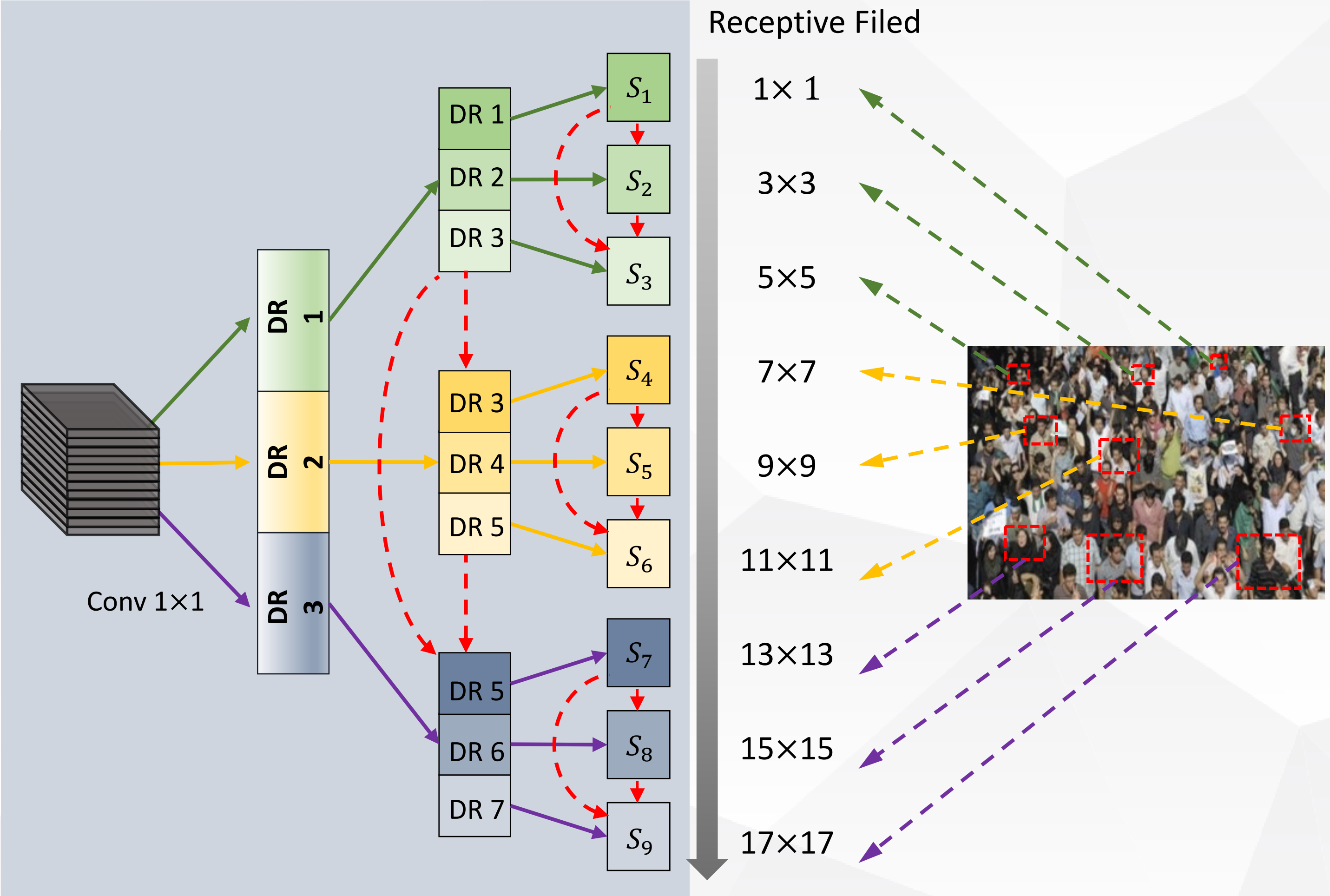
Overall Architecture of STNet

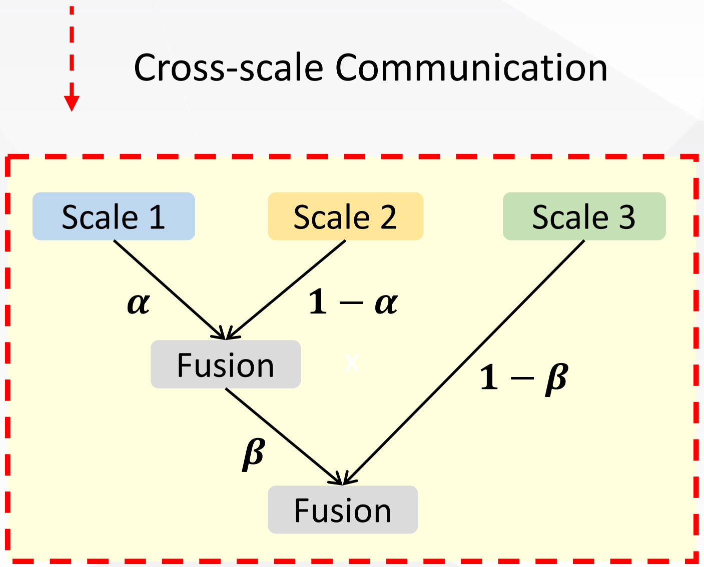
- Introduction
- Classification #1
- Classification #2
- Regression #1
- Regression #2
- Regression #3
- Summary



UNIVERSITY of GUELPH

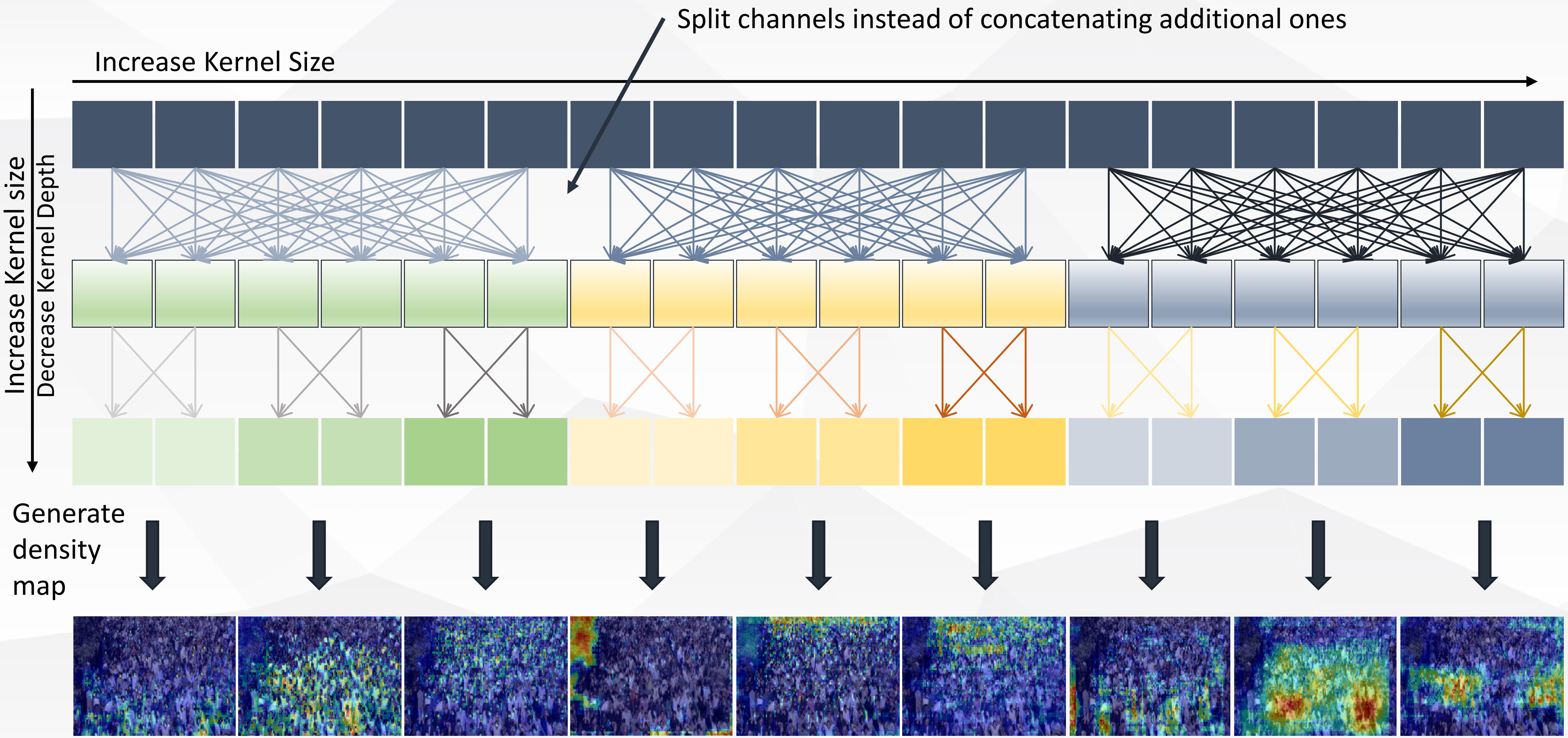
Scale Tree Diversity Enhancer





UNIVERSITY of GUELPH

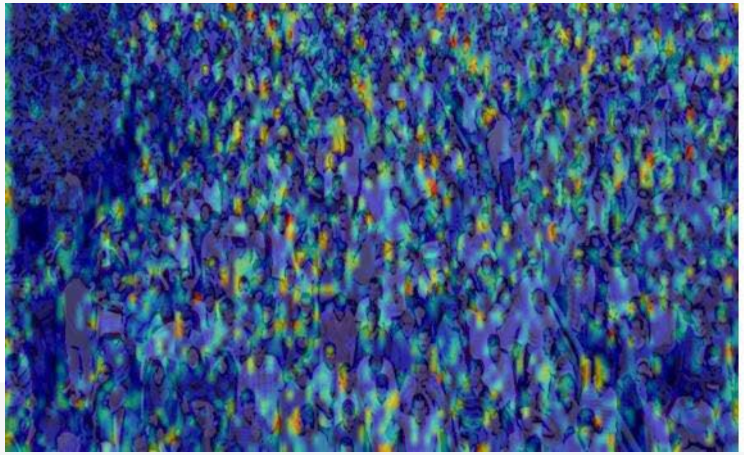
Visualize Scale Tree Diversity Enhancer



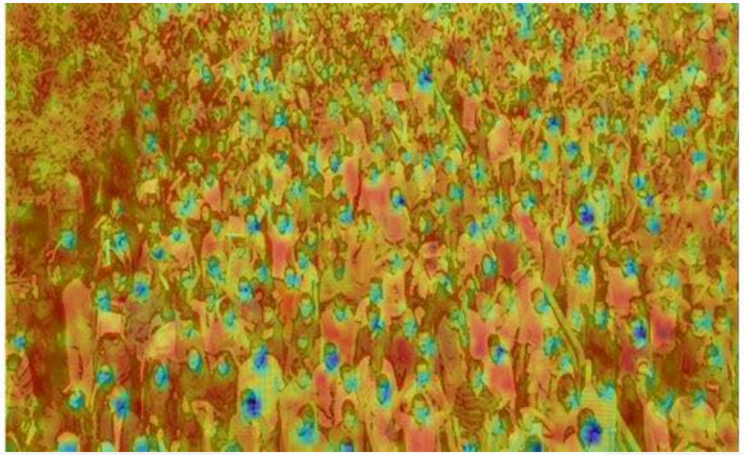
Multi-level Auxiliator

Raw image

- Information from all 3 scales are useful for the auxiliary (background detection) task.
- Back propagating loss to low- and middle-scales helps to train useful features.



Low-level



Middle-level

High-level

UNIVERSITY of GUELPH

Quantitative Comparison on Counting Accuracy

Methods	Part_A		Part_B		UCF-QNRF		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
TEDNet [35]	64.2	109.1	8.2	12.8	113	188	249.4	354.5
ADCrowdNet [18]	63.2	98.9	7.6	13.9	-	-	257.1	363.5
PACNN + CSRNet [36]	62.4	102.0	7.6	11.8	-	-	241.7	320.7
CANet [6]	62.3	100.0	7.8	12.2	107	183	212.2	243.7
SPN+L2SM [37]	64.2	98.4	7.2	11.1	104.7	173.6	188.4	315.3
MBTTBF-SCFB [10]	60.2	94.1	8.0	15.5	97.5	165.2	233.1	300.9
DSSINet [9]	60.63	96.04	6.85	10.34	99.1	159.2	216.9	302.4
S-DCNet [14]	58.3	95.0	6.7	10.7	104.4	176.1	204.2	301.3
ASNet [7]	57.78	90.13	-	-	91.59	159.71	174.84	251.63
ADNet [12]	61.3	103.9	7.6	12.1	90.1	147.1	245.4	327.3
AMRNet [11]	61.59	98.36	7.02	11.00	86.6	152.2	184.0	265.8
AMSNet [38]	58.0	96.2	7.1	10.4	103	165	208.6	296.3
BL [39]	62.8	101.8	7.7	12.7	88.7	154.8	229.3	308.2
ADSCNet [12]	55.4	97.7	6.4	11.3	**71.3**	**132.5**	198.4	267.3
STNet (proposed)	**52.85**	**83.64**	**6.25**	**10.30**	87.88	166.44	**161.96**	**230.39**

Ablation Studies

Introduction
Classification #1
Classification #2
Regression #1
Regression #2
Regression #3
Summary

	CSRNet	STNet-Enh	STNet
Params.	16.26M	25.10M	15.56M
MAE	68.2	58.7	52.8
MSE	115.0	97.2	83.6

- The impacts of Scale Tree Diversity Enhancer on ShanghaiTech Part A dataset.

- Impacts of training with pure background images on ShanghaiTech Part A, Part B and UCF-QNRF datasets.

	w/o Balancing		w/ Balancing	
	MAE	MSE	MAE	MSE
Part_A	56.47	97.19	52.85	83.64
Part_B	6.71	11.36	6.25	10.30
QNRF	89.81	168.29	87.88	166.44

	MAE	MSE
STNet-Enh+ w/o Auxiliator	62.378	100.951
STNet-Enh+ w/ Auxiliator	58.711	97.197
STNet w/ BT=0	61.591	100.313
STNet w/o Auxiliator	59.296	98.699
STNet w/ H.A	57.66	98.93
STNet w/ H.M.A	57.519	98.86
STNet w/ H.M.L.A	52.85	83.64

- Multi-level Auxiliator ablation study on ShanghaiTech Part A.
  - H.A, H.M.A and H.M.L.A indicates the auxiliator uses high-level only, high and middle-level only, and all three levels, respectively.
  - BT stands for binarization threshold.

Summary

- Many common issues limit the performance of CNNs:
  - Vanishing gradient problem:
    - *Hard to train networks that are too deep.*
  - Superfluous feature reuse:
    - *Makes the network too wide without much benefit.*
  - Overfitting:
    - *Require stochastic regularizations.*
  - Scale variation within or between images:
    - *Need to capture features at different scales.*
  - Background noise:
    - *Be aware of foreground/background is important for scene understanding.*

General Principles

- Perform stochastic regularizations along different dimensions:
  - SFR randomly selects layers to feed into dense connections.
  - Scale-Shake Unit randomly blends two branches each time.
  - SiT randomly fuses different convolution groups for different scales.
- Reduce model size:
  - ADNet aggregates different layers together (instead of concatenation) through layer-attention.
  - SiT fuses features from different scales together through a stochastic mixer.
  - STNet splits channels instead of adding additional channels.
- Model fine-grain scale variations:
  - ScSiNet uses SiT to capture intra-layer scale variation.
  - STNet designs scale-tree to encode different scales in different channels.
- Carefully design auxiliary tasks:
  - Auxiliary task in SMANet directly predicts total crowd count.
  - Auxiliary task in STNet uses intermediate features to predict background confidence scores.
- Augment training data:
  - ScSiNet randomly crops patches online.
  - STNet adds pure background to balance foreground/background distribution.

- Convolutional Neural Networks (CNNs) have demonstrated superior performances in many Computer Vision tasks, thanks to their strong learning capabilities. Hence, further enhancing CNNs' learning capability has very broad impacts. This talk introduces several enhancement attempts through mechanisms such as stochastic regularizations, layer-wise attention, multi-scale aggregation, and auxiliary tasks. Since there are essentially two types of learning problems, classification and regression, two fundamental vision tasks are chosen accordingly as targe applications.
- We start with the image classification problem since it has been the gold standard for evaluating different CNN architectures. Inspired by the success of feature reuse in DenseNet and a series of drop-based stochastic regularizations, Stochastic Features Reuse is presented to strengthen capacity and generalization of DenseNet through randomly dropping reused features. Simultaneously, a Multi-scale Convolution Aggregation module is also explored to facilitate learning scale-invariant representations. Albeit promising, the above approach inherits DenseNet's limitations on large model size and superfluous feature reuse. To achieve high discriminative power with compact models, layer-wise attention is designed to form a powerful variant, named Adaptively Dense CNN.
- We then turn to the crowd counting problem since it expects a single, non-constrained value as output, making it more arduous and representative than other regression tasks. Appling the ideas of stochastic regularizations and multi-scale aggregation again, a Stochastic Multi-Scale Aggregation Network is designed to enlarge the scale diversity of feature maps and to combat overfitting. To further boost the capacity of handling large scale variation, a Single-column Scale-invariant Network is presented, which extracts scale-invariant features though both interlayer multi-scale integration and a novel intralayer Scale-invariant Transformation. Finally, an innovative Scale Tree Network is presented to parse scale information hierarchically and efficiently using a tree structure. It also employs a Multi-level Auxiliator to facilitate the recognition of cluttered backgrounds.
- Extensive experiments on widely used benchmarks demonstrate the effectiveness of the proposed strategies in enhancing learning capabilities, thereby achieving superior performances in classification and counting accuracy. Ablation studies and visualization analysis are also performed to better understand the impacts and behaviors of individual components.