Generalization in Machine Learning for Counting Problems

Dr. Minglun Gong

School of Computer Science, University of Guelph

(Based on the PhD. dissertation of Dr. Mingjie Wang)

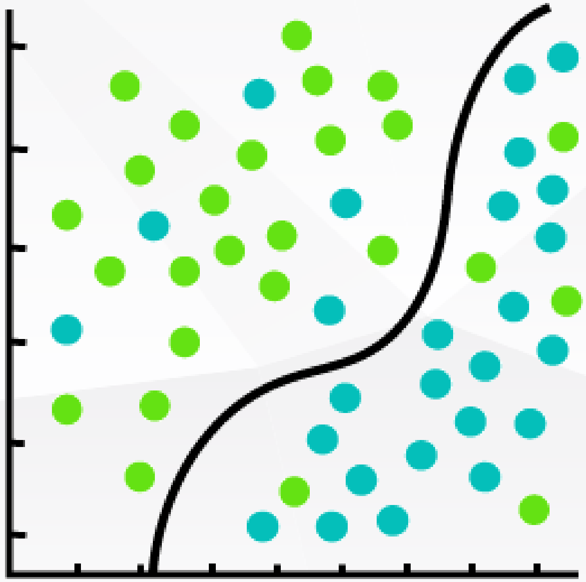
UNIVERSITY of GUELPH

UNIVERSITY of GUELPH

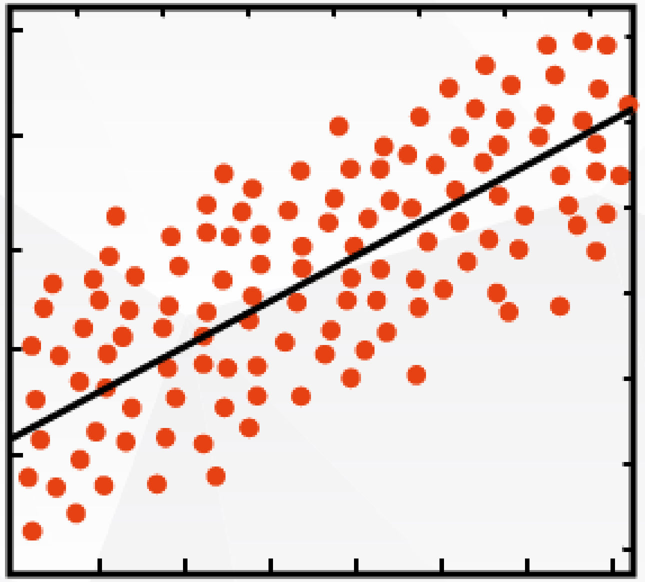
- Introduction
- Scale-Invariant Counter
- Background-Aware Counter
- Weakly-Supervised Counter
- General Object Counter
- Conclusion

Two Types of Machine Learning Problems

- Classification predicts discrete labels:
  - Find accurate decision boundaries;
  - Evaluate using accuracy.
- Computer vision problems:
  - Object/identify recognition
  - Semantic/instance segmentation

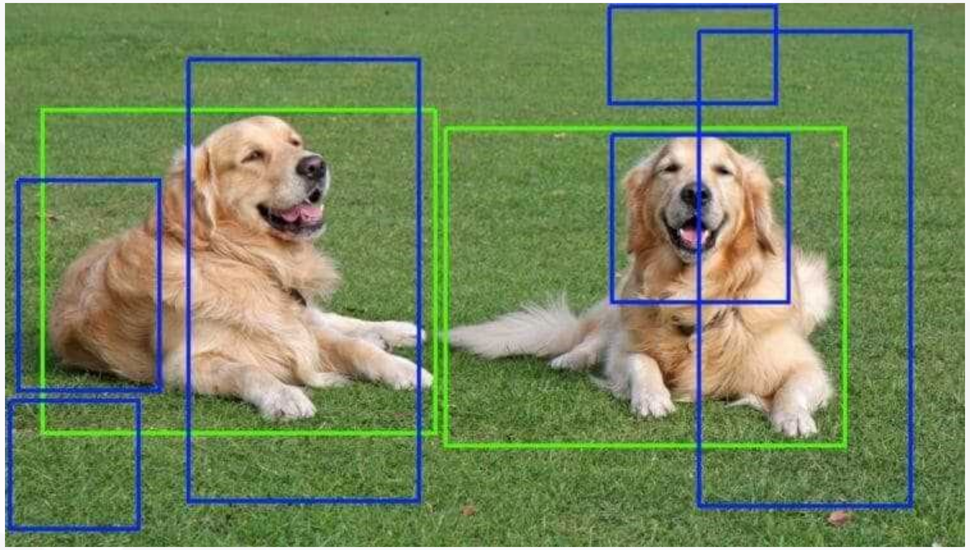


- Regression predicts continuous quantities;
  - Find the best fitting lines/curves;
  - Evaluate using root mean squared error.
- Computer vision problems:
  - Age estimation
  - Object localization



Regression Problems

- Many regression problems have constrained output:
  - Age estimation:
    - *A single age number*
  - Object localization:
    - *Coordinates of 1 or more objects*
  - Depth estimation:
    - *A depth value at each pixel location*
  - Image synthesis:
    - *A color at each pixel location*
- Counting problem is unconstrained:
  - Output a single value in [0,+infinity)
  - Referred as open set problem



UNIVERSITY of GUELPH

Crowd Counting Problem

Introduction

Scale-Invariant Counter

Background-Aware Counter

Weakly-Supervised Counter

General Object Counter

Conclusion

- Estimate the number of people in an image or a video stream
  - Detection-based approach:
    - *Perform poorly on congested scenes.*
  - Regression-based approach:
    - *Map input images to density maps then integrate.*
    - *Map to direct count (new trend).*
- Challenges:
  - Severe occlusions
  - Scale variation & density shift
  - Noisy background
  - Overfitting

Evaluation Datasets for Crowd Counting

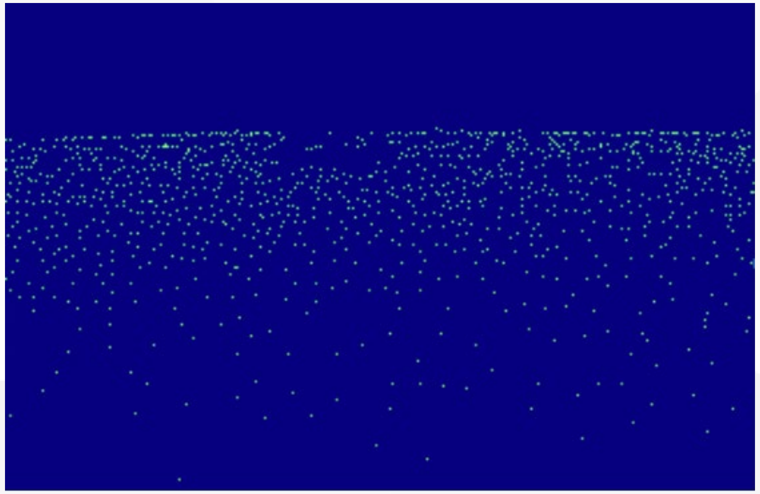
- ShanghaiTech:
  - 1198 images & 330,165 annotated people
  - Classical Part A contains 482 congested internet images.
  - Part B consists of 716 sparse images with a fixed size of 768×1024 captured from outside streets.
- UCF_QNRF:
  - 1,553 images collected from the Web, with a total of 1,252,642 people
  - 1201 images are for training & 334 for testing.
  - Has a wider range of diverse densities, backgrounds, & perspectives.
- UCF_CC_50:
  - Includes 50 images taken from the Internet with an average of 1280 individuals per image.
  - Challenging due to the severely small set of samples & huge density shifts.
- JHU-CROWD++:
  - A large-scale crowd dataset including 4,372 images
  - 2,722 images are for training, 1,600 scenes for testing, & 500 samples are used for validation.
  - Has a total of 1.51 million annotations on centroids of people & the number of crowd ranges from 0 to 25,791.

JHU-CROWD++: A Large-scale Unconstrained Dataset

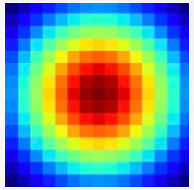
Ground Truth Annotation

Input Crowd Image

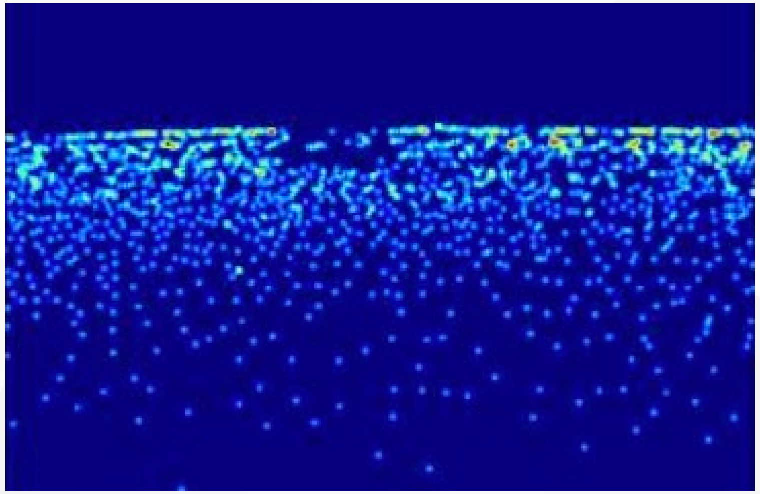
$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x)$$



Dot Map



Gaussian Kernel, G



Density Map

Evaluation Metrics

- Mean Absolute Error (MAE):
  - $MAE = \frac{1}{N}\sum_{i=1}^{N}|C_i - C_i^{GT}|$
- Mean Square Error (MSE):
  - $MSE = \frac{1}{N}\sum_{i=1}^{N}\left(C_i - C_i^{GT}\right)^2$
- Root Mean Square Error (RMSE):
  - $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(C_i - C_i^{GT}\right)^2}$
- Notations:
  - N is the number of test images;
  - C_i denotes the predicted count;
  - C_i^{GT} is the ground truth of count number annotated for the i_{th} input image.

Leaderboard (ShanghaiTech Part B)

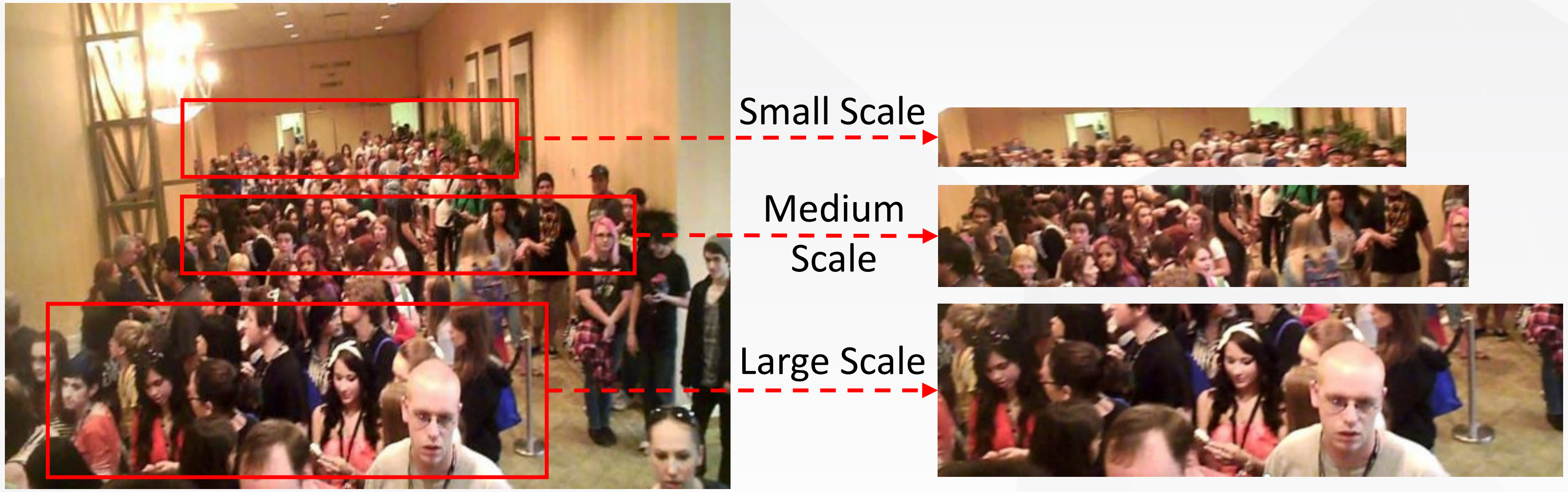
Conference/Journal	Methods	MAE	MSE
2019-ICCV	DSSINet	60.63	96.04
2019-ICCV	MBTTBF-SCFB	60.2	94.1
2019-ICCV	RANet	59.4	102.0
2019-ICCV	SPANet+SANet	59.4	92.5
2019-TIP	PaDNet	59.2	98.1
2019-ICCV	S-DCNet	58.3	95.0
2020-ICPR	M-SFANet	57.55	94.48
2019-ICCV	PGCNet	57.0	86.0
2020-ECCV	AMSNet	56.7	93.4
2020-CVPR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

https://github.com/gjy3035/Awesome-Crowd-Counting

Scale-invariant Transformation & Random Data Augmentation

Neurocomputing, 2021

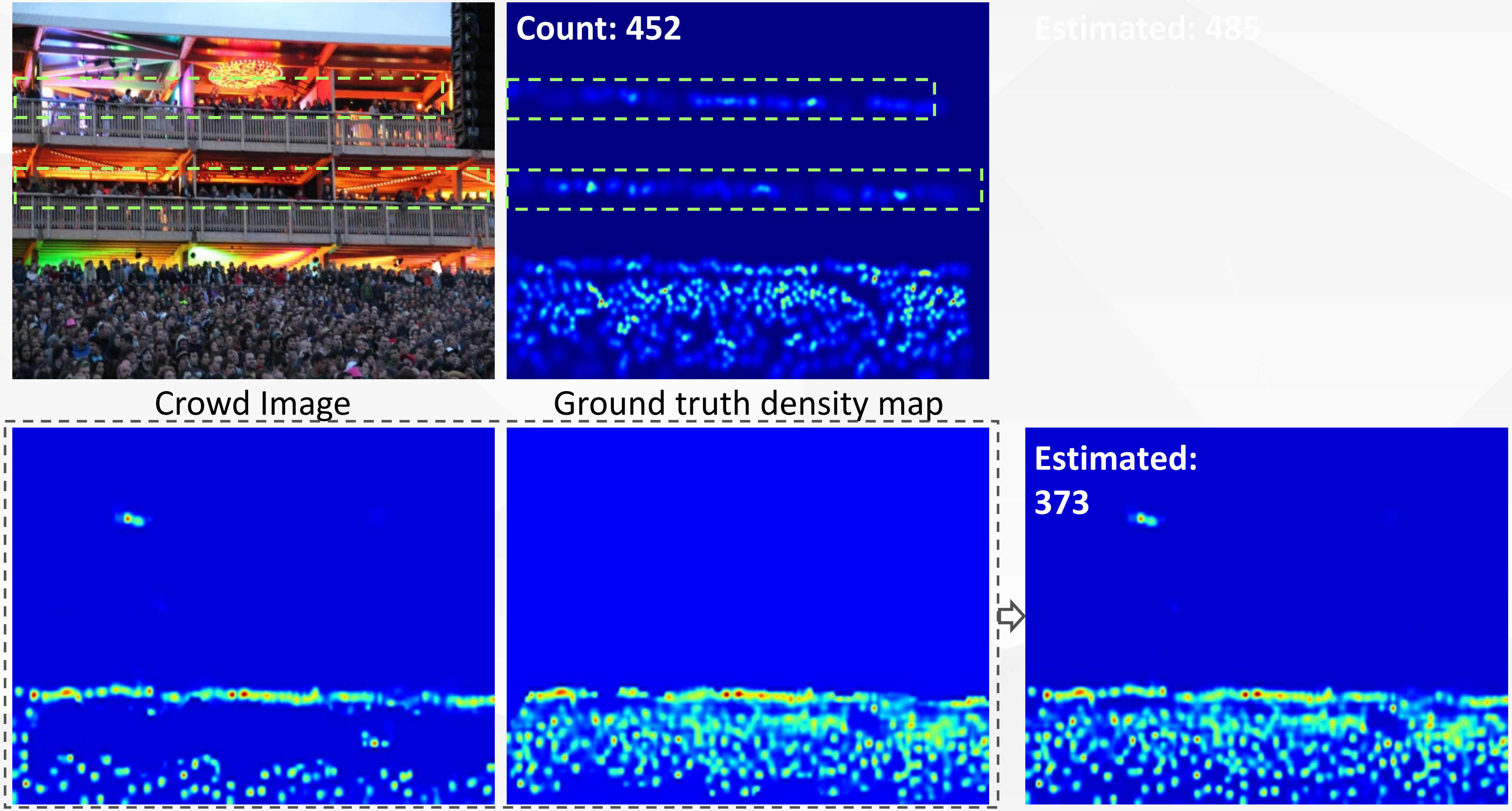
Challenges: Scale Variation & Density Shift



(a) Scale Variations

(b) Density Shifts

Previous Work: ASNet [CVPR 2020]

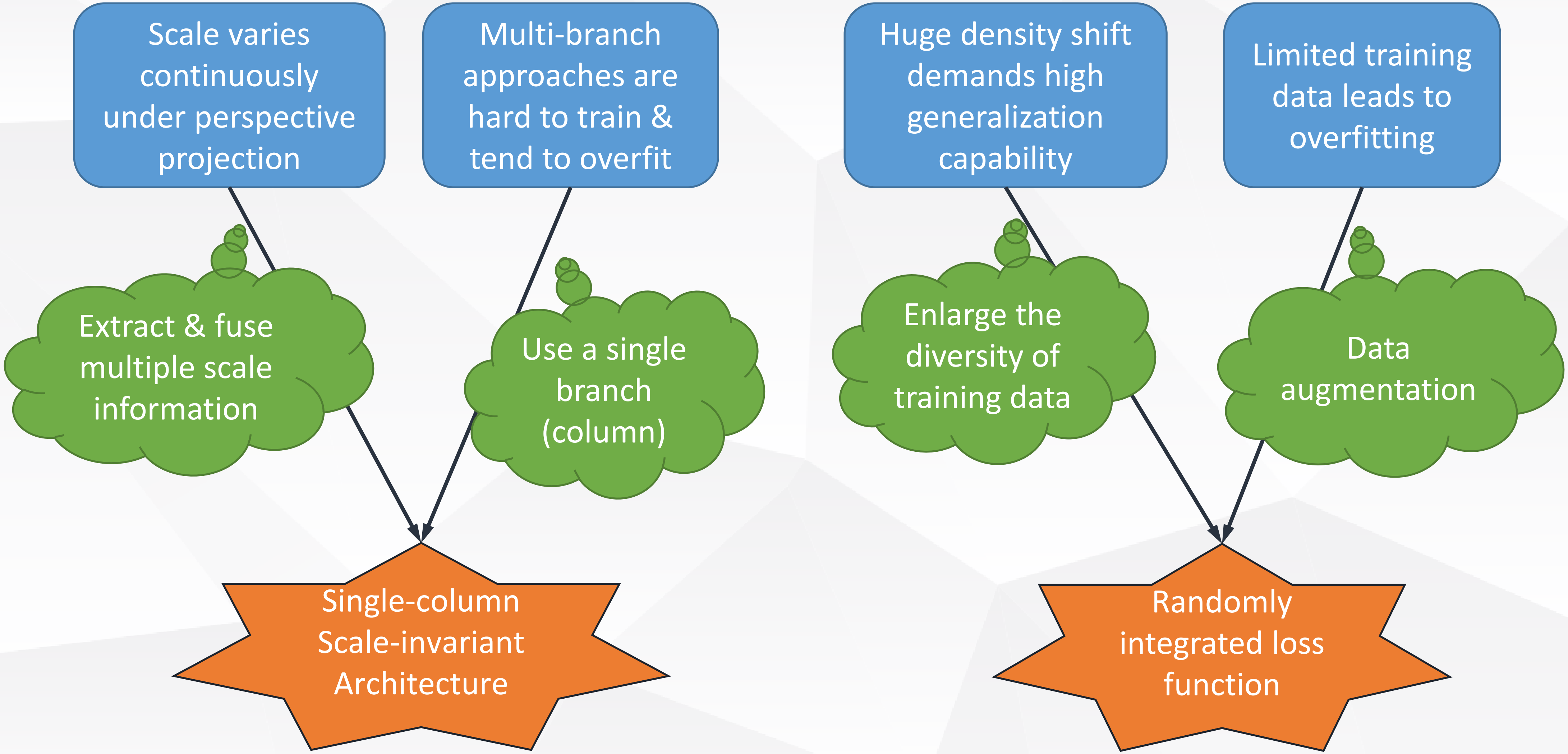


Crowd Image | Ground truth density map

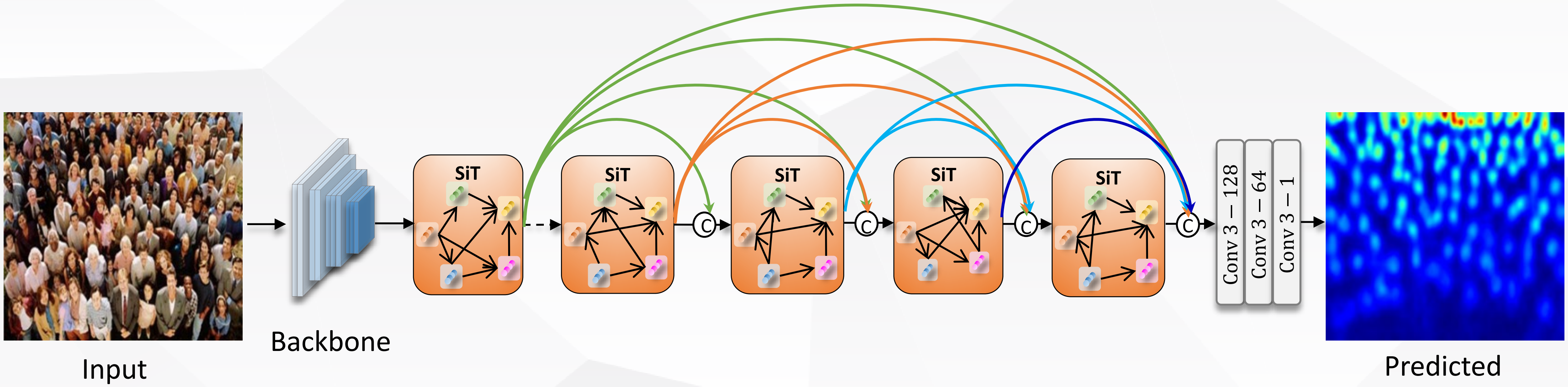
ASNet Branch1 | ASNet Branch2 | ASNet

Motivations

Introduction
Scale-Invariant Counter
Background-Aware Counter
Weakly-Supervised Counter
General Object Counter
Conclusion



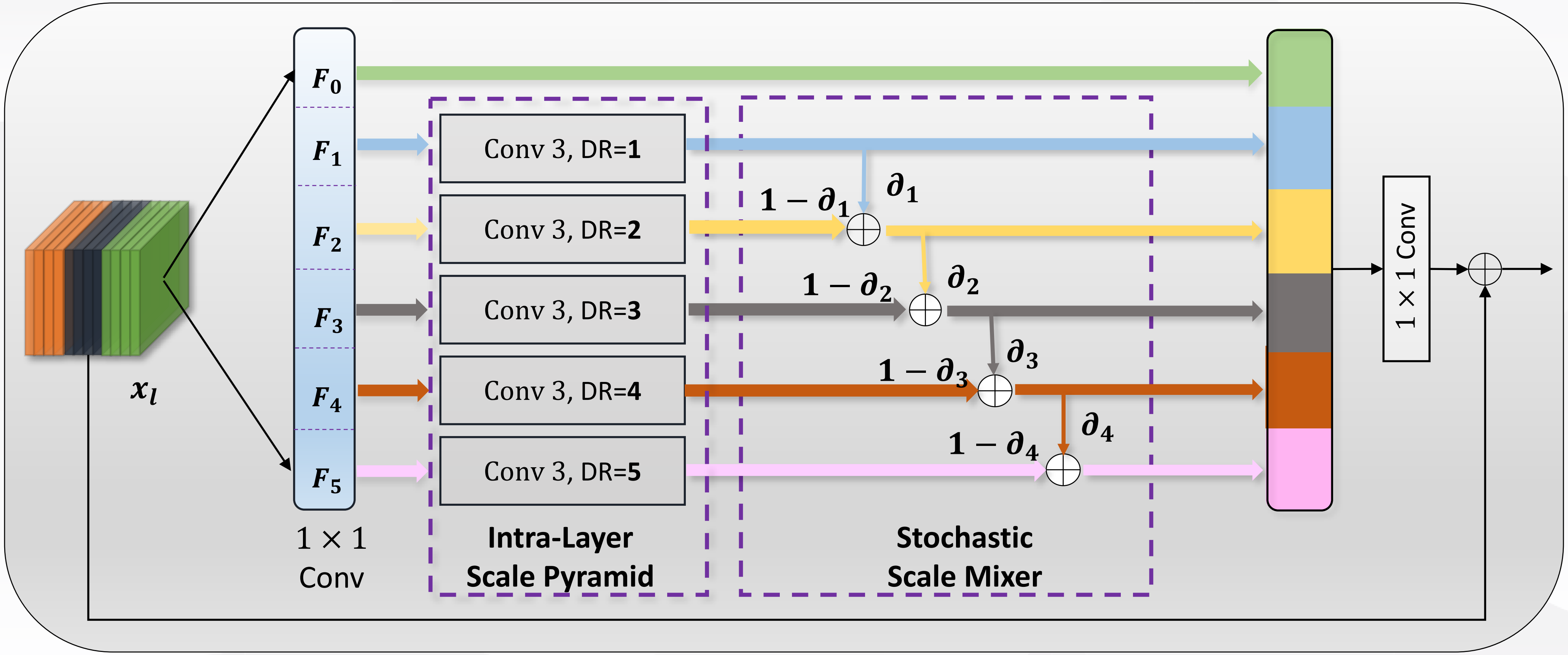
Overall Architecture w/ Scale-invariant Modules



SiT: Scale-invariant Transformation

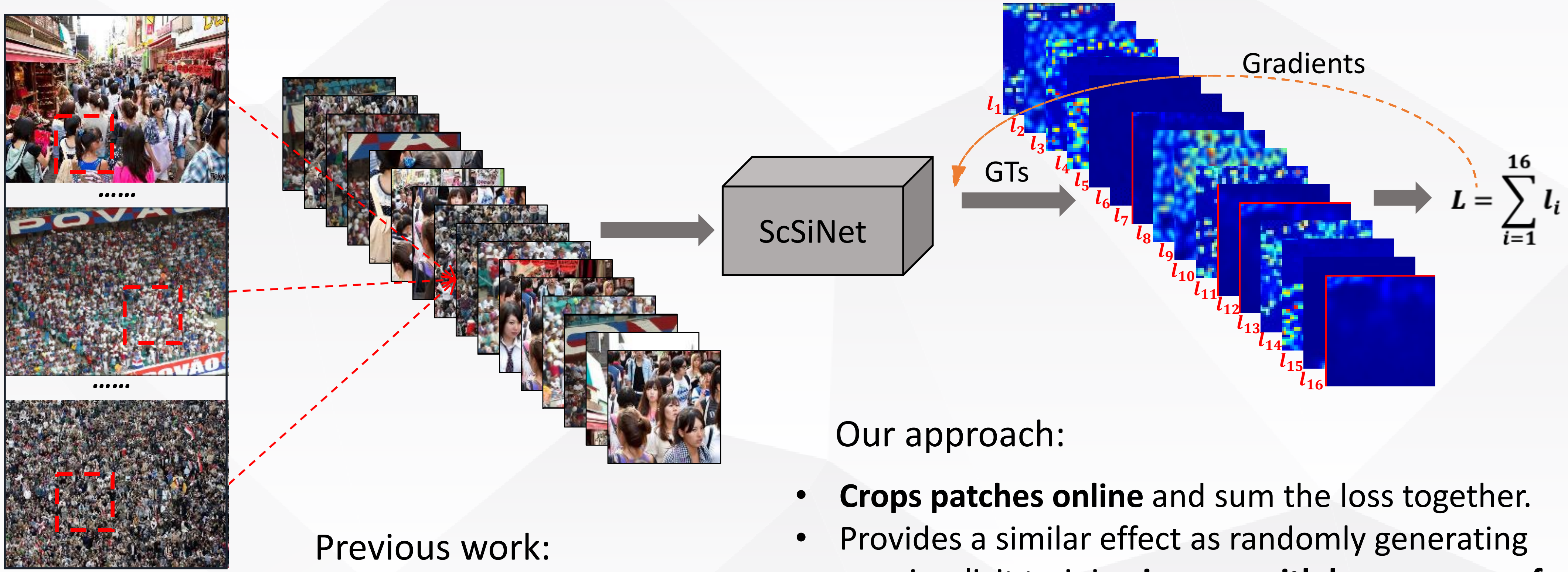
- The dense connections between different SiTs provide inter-layer (coarse grain) scale aggregation.

Scale-invariant Transformation (SiT)



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

Randomly Integrated Loss



Previous work:

- Crop patches off-line and use them repetitively.
- Use average loss for training.

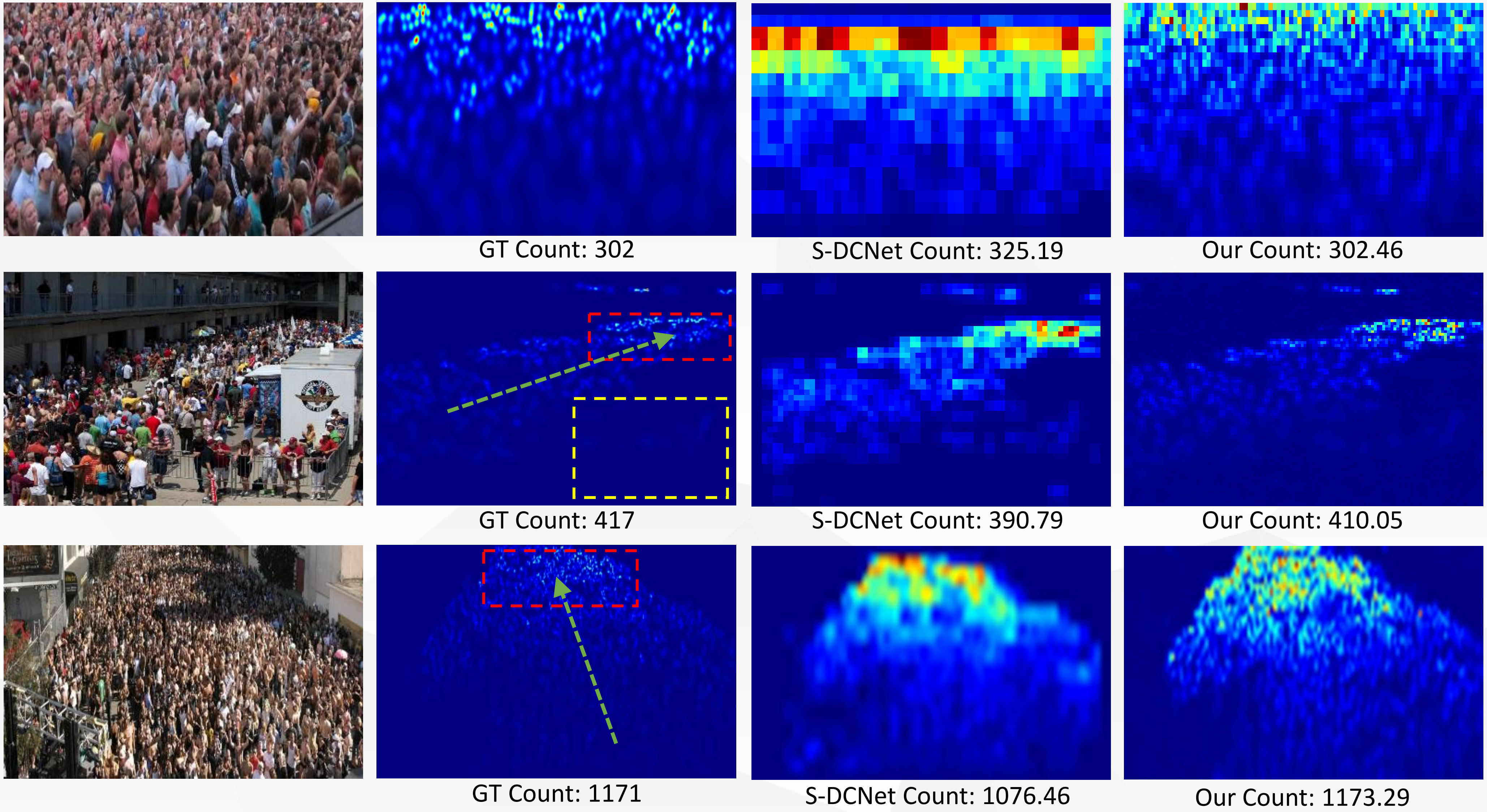
Our approach:

- **Crops patches online** and sum the loss together.
- Provides a similar effect as randomly generating new implicit training **images with larger range of density levels**.
- Takes datasets with different image resolutions **without the needs for resizing**.

Quantitative Evaluation

Methods	Part A		Part B		UCF-QNRF		UCF_CC_50		AR
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	
TEDNet [53]	64.2	109.1	8.2	12.8	113	188	249.4	354.5	11.5
ADCrowdNet [77]	63.2	98.9	7.6	13.9	-	-	257.1	363.5	10.7
PACNN + CSRNet [114]	62.4	102.0	7.6	11.8	-	-	241.7	320.7	9.3
CAN [79]	62.3	100.0	7.8	12.2	107	183	212.2	243.7	8.0
CFF [115]	65.2	109.4	7.2	12.2	-	-	-	-	11.0
SPN+L2SM [156]	64.2	98.4	7.2	11.1	104.7	173.6	188.4	315.3	7.3
MBTTBF-SCFB [123]	60.2	94.1	8.0	15.5	97.5	165.2	233.1	300.9	7.3
PGCNet [159]	57.0	**86.0**	8.8	13.7	-	-	244.6	361.2	9.0
HyGNN [87]	60.2	94.5	7.5	12.7	100.8	185.3	184.4	270.1	5.0
BL [91]	62.8	101.8	7.7	12.7	**88.7**	**154.8**	229.3	308.2	7.0
DSSINet [76]	60.63	96.04	6.85	10.34	99.1	159.2	216.9	302.4	5.3
SPANet+SANet [15]	59.4	92.5	**6.5**	**9.9**	-	-	232.6	311.7	4.3
S-DCNet [155]	58.3	95.0	6.7	10.	104.4	176.1	204.2	301.3	3.8
ScSiNet (proposed)	**55.77**	90.23	6.79	10.95	89.69	178.46	**154.87**	**199.42**	**1.8**

Comparison on Density Maps



Introduction

Scale-Invariant Counter

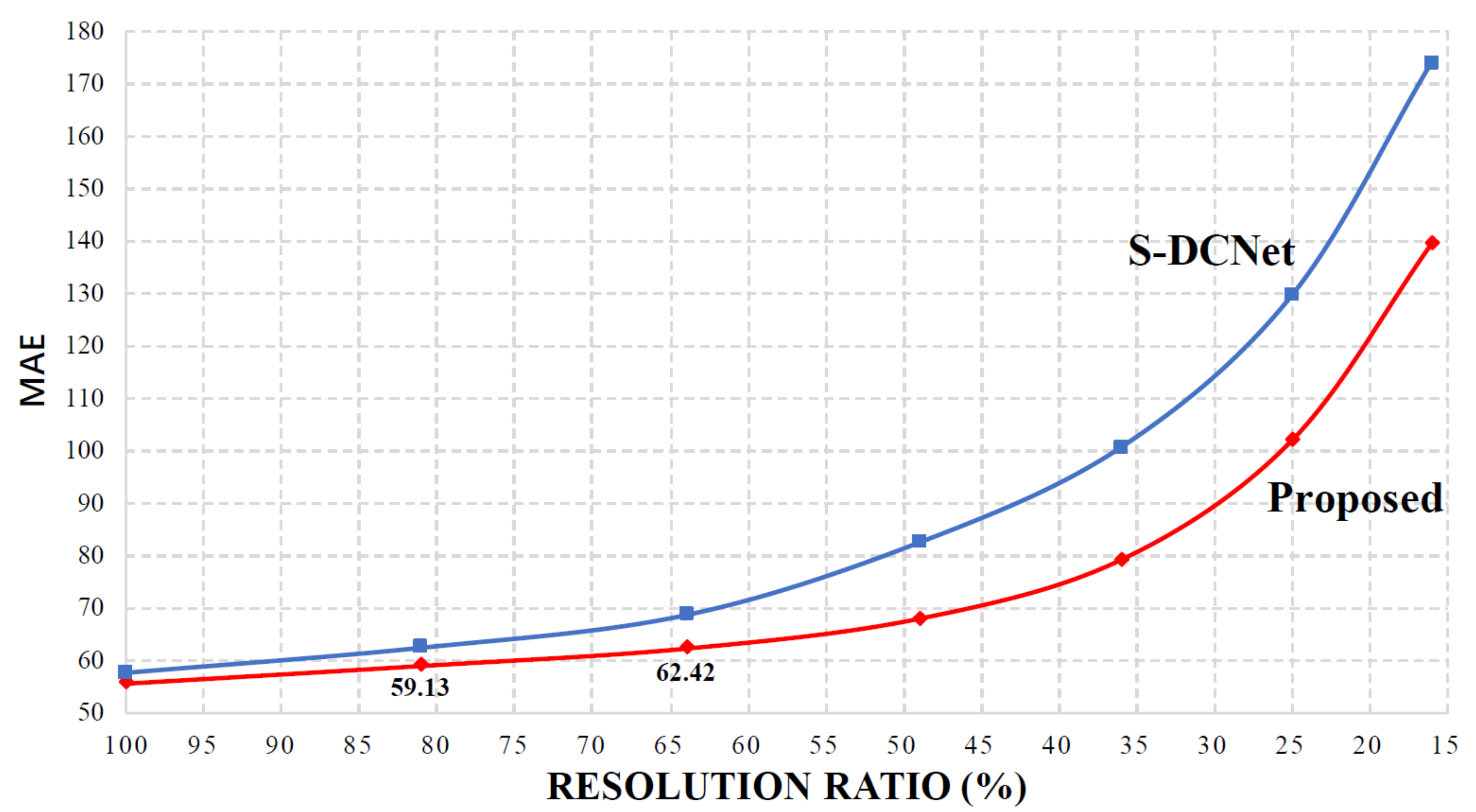
Background-Aware Counter

Weakly-Supervised Counter

General Object Counter

Conclusion

Scale-Invariance Test



UNIVERSITY of GUELPH

Ablation Studies

Introduction

Scale-Invariant Counter

Background-Aware Counter

Weakly-Supervised Counter

General Object Counter

Conclusion

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

The effectiveness of stochastic scale mixer

The effectiveness of stochastic mixer strategy

	Concat		Attention-based		Scale Mixer	
	MAE	MSE	MAE	MSE	MAE	MSE
Part A	61.05	100.27	58.56	94.35	55.77	90.23
Part B	7.01	11.53	6.82	11.03	6.79	10.95
UCF-QNRF	91.12	170.80	91.01	169.43	89.69	178.46

	w.o. Mini-Batch		w. Mini-Batch	
	MAE	MSE	MAE	MSE
Part A	60.33	107.24	55.77	90.23
UCF_CC_50	195.34	248.62	154.87	199.42

The effect of randomly integrated loss

UNIVERSITY of GUELPH

Tree-based Scale Enhancer & Auxiliary Supervision on Background Awareness

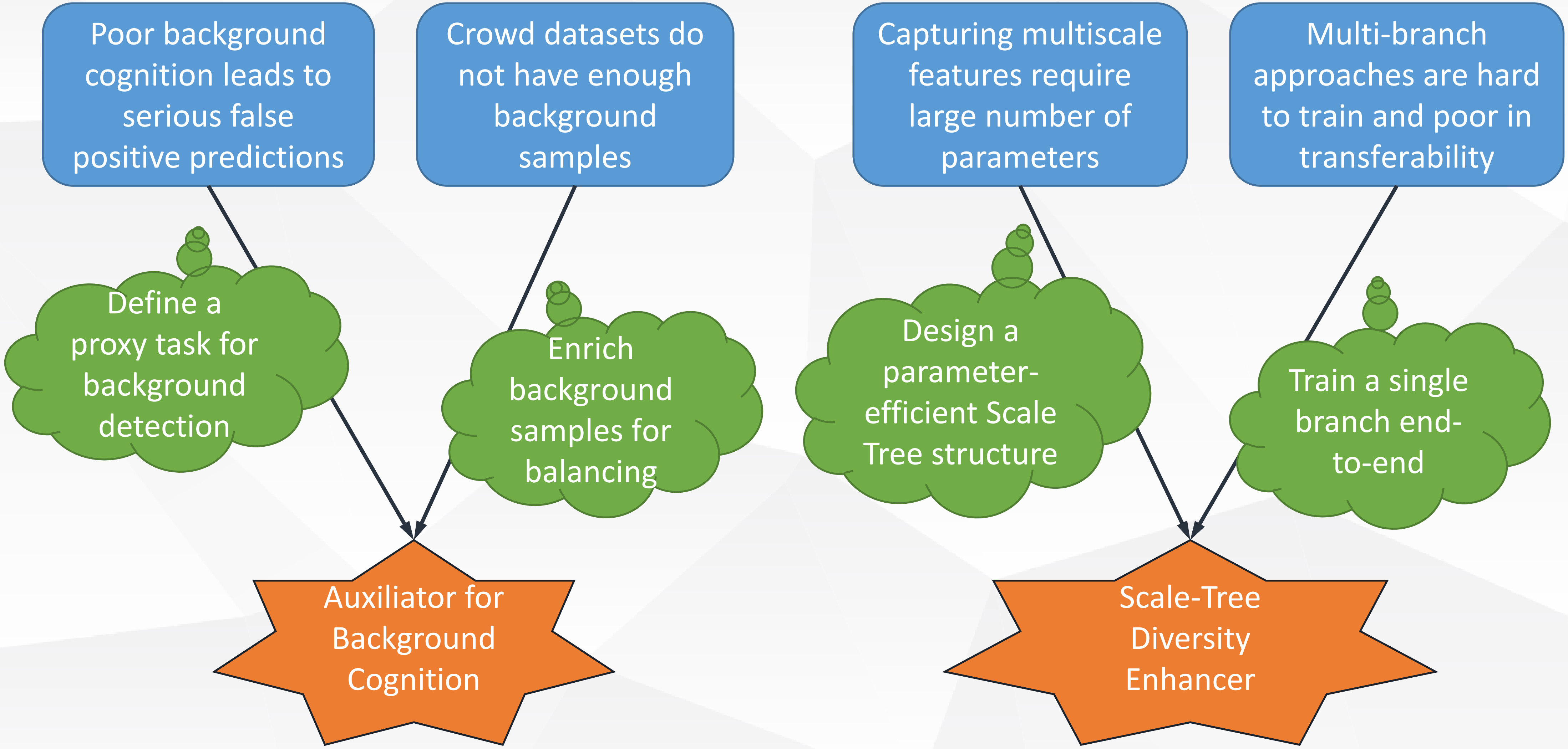
IEEE Transactions on Multimedia, 2021

Challenge: Counting under Different Scene Contexts

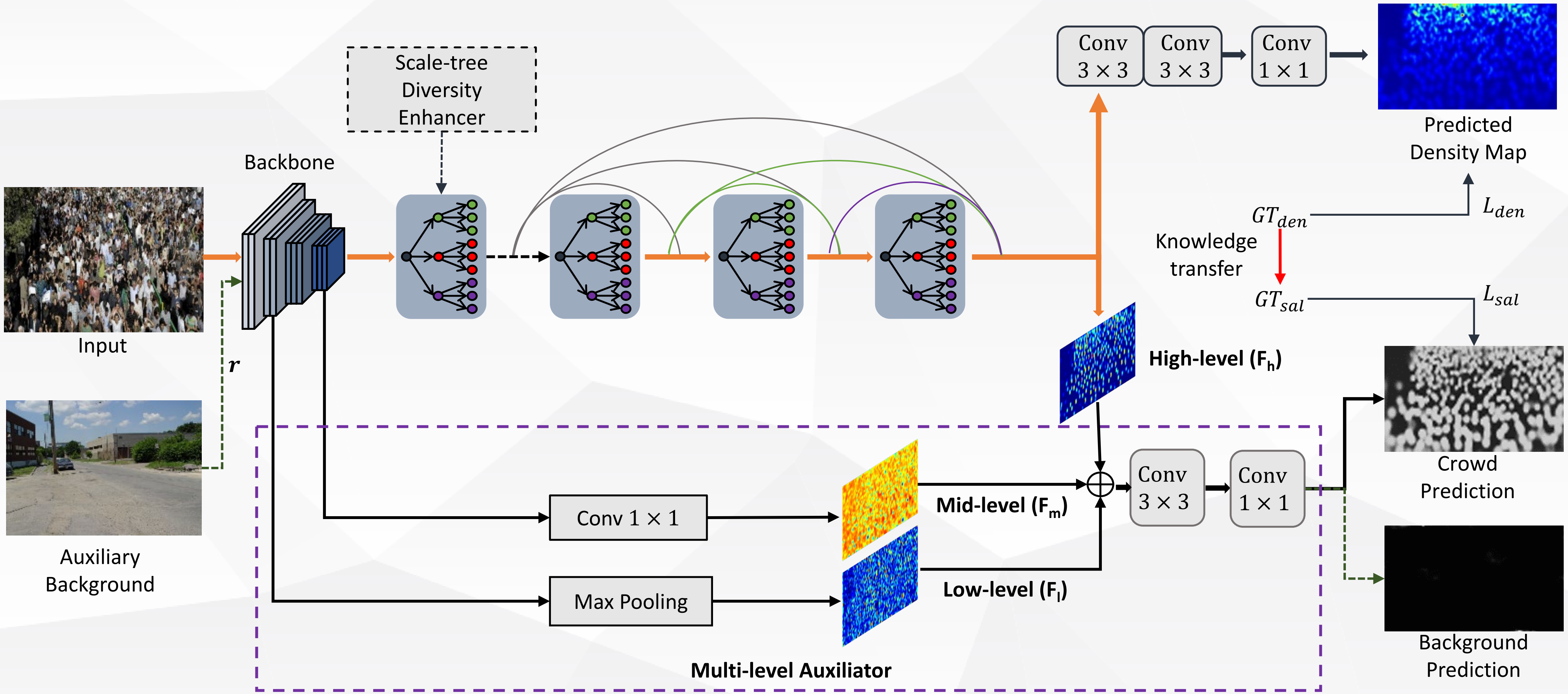
- To handle real world counting problem, the network needs to be robust under various scenes.
  - Needs to focus on crowd (foreground) and ignore all other objects (background).
- Most training images in crowd counting dataset are scenes with lots of people.
  - Lack of training samples for background.

UNIVERSITY of GUELPH

Motivations



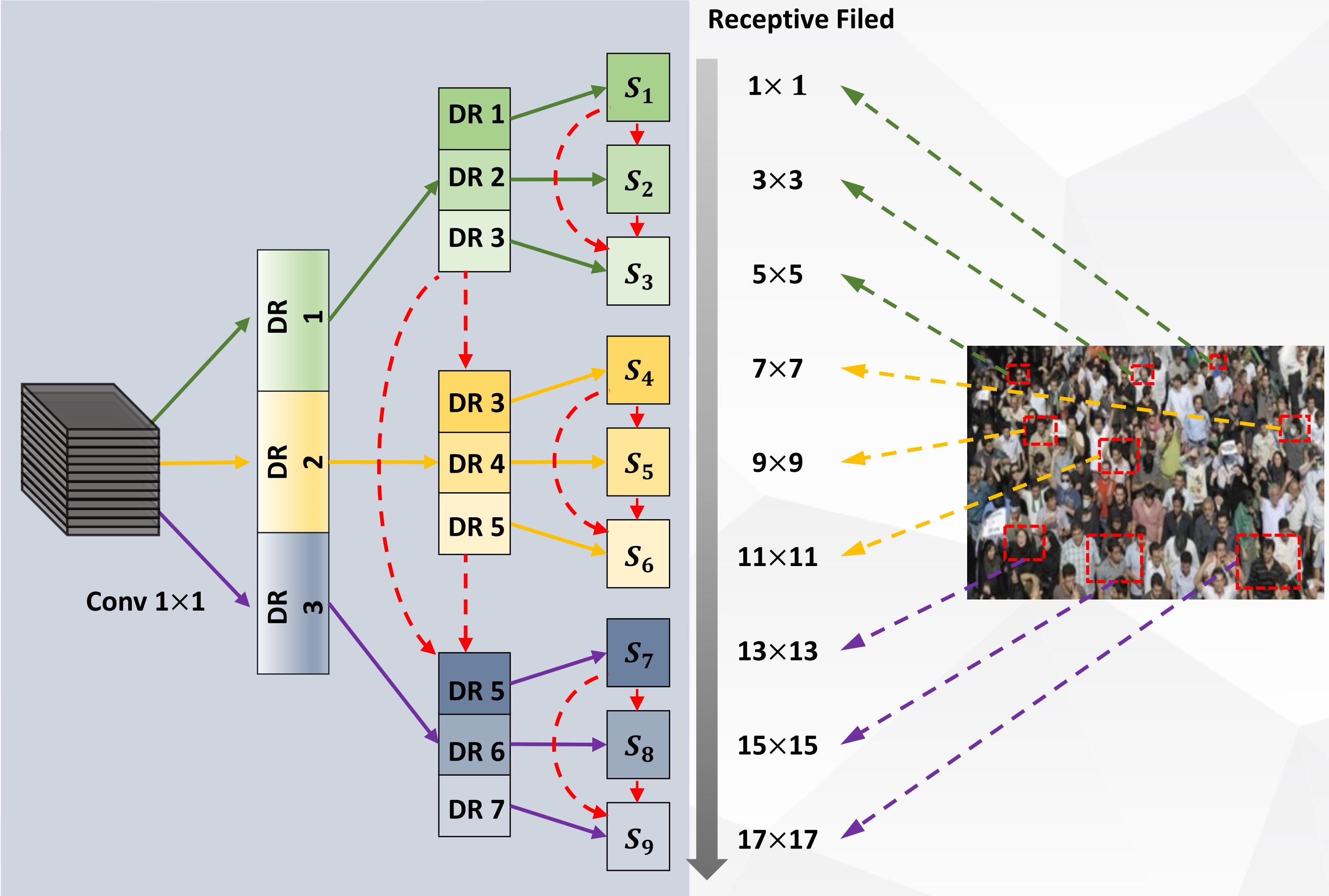
Overall Architecture w/ Background Cognition

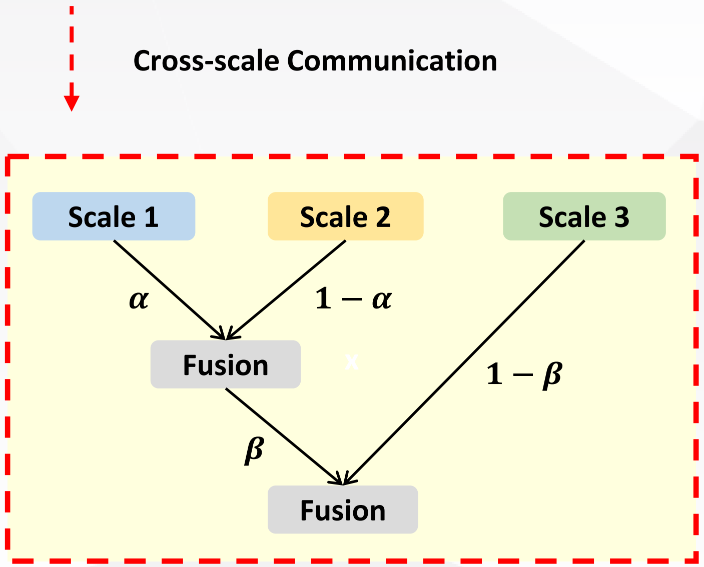


UNIVERSITY of GUELPH

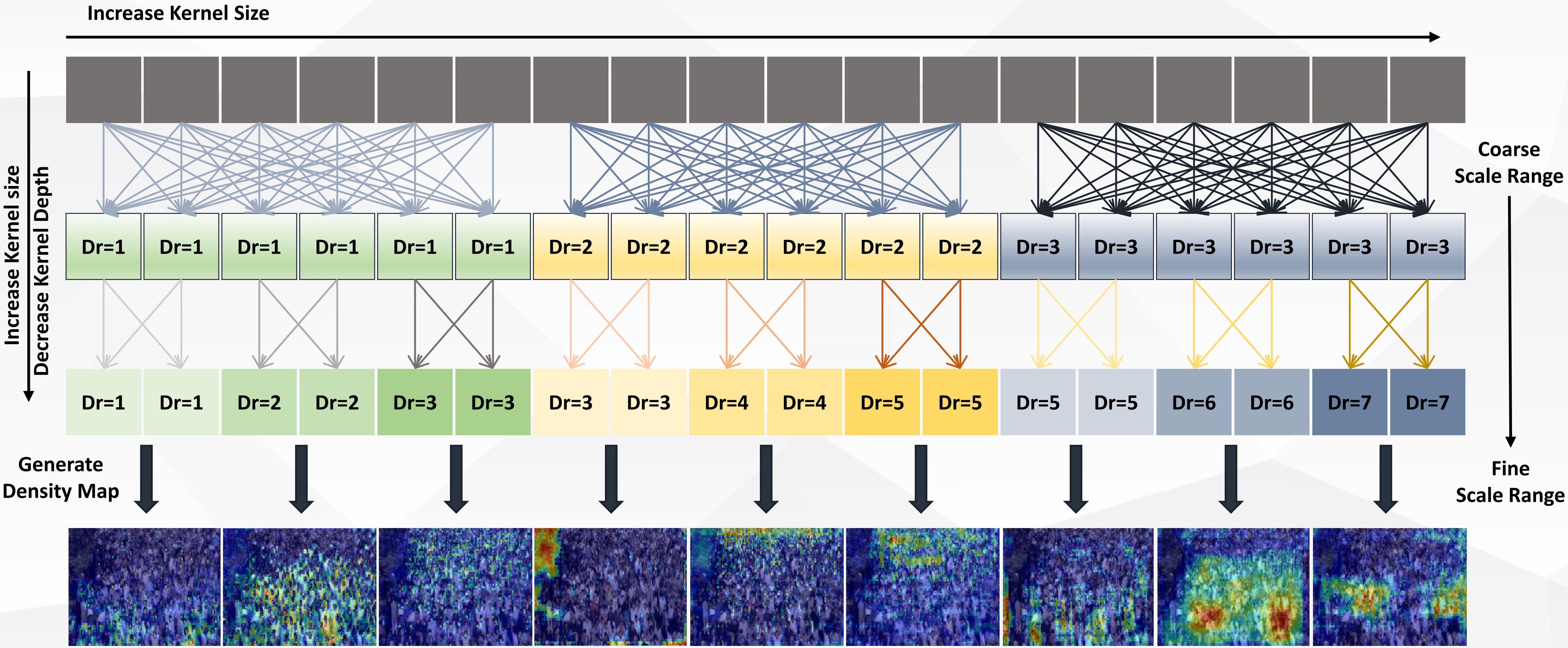
Scale Tree Diversity Enhancer

- Introduction
- Scale-Invariant Counter
- Background-Aware Counter
- Weakly-Supervised Counter
- General Object Counter
- Conclusion





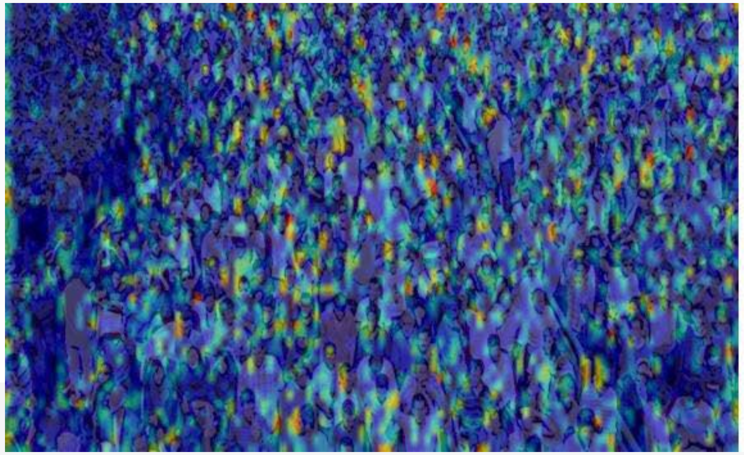
Visualize Scale Tree Diversity Enhancer



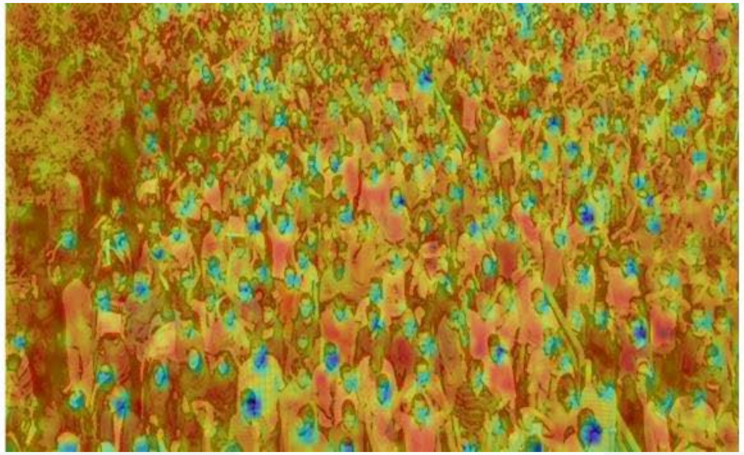
Background Cognition using Multi-level Input

Raw image

- Information from all 3 levels are useful for the auxiliary (background detection) task.
- Feeding loss to low-middle-levels also helps to train useful features.



Low-level



Middle-level

High-level

UNIVERSITY of GUELPH

Quantitative Evaluation

Introduction

Scale-Invariant Counter

Background-Aware Counter

Weakly-Supervised Counter

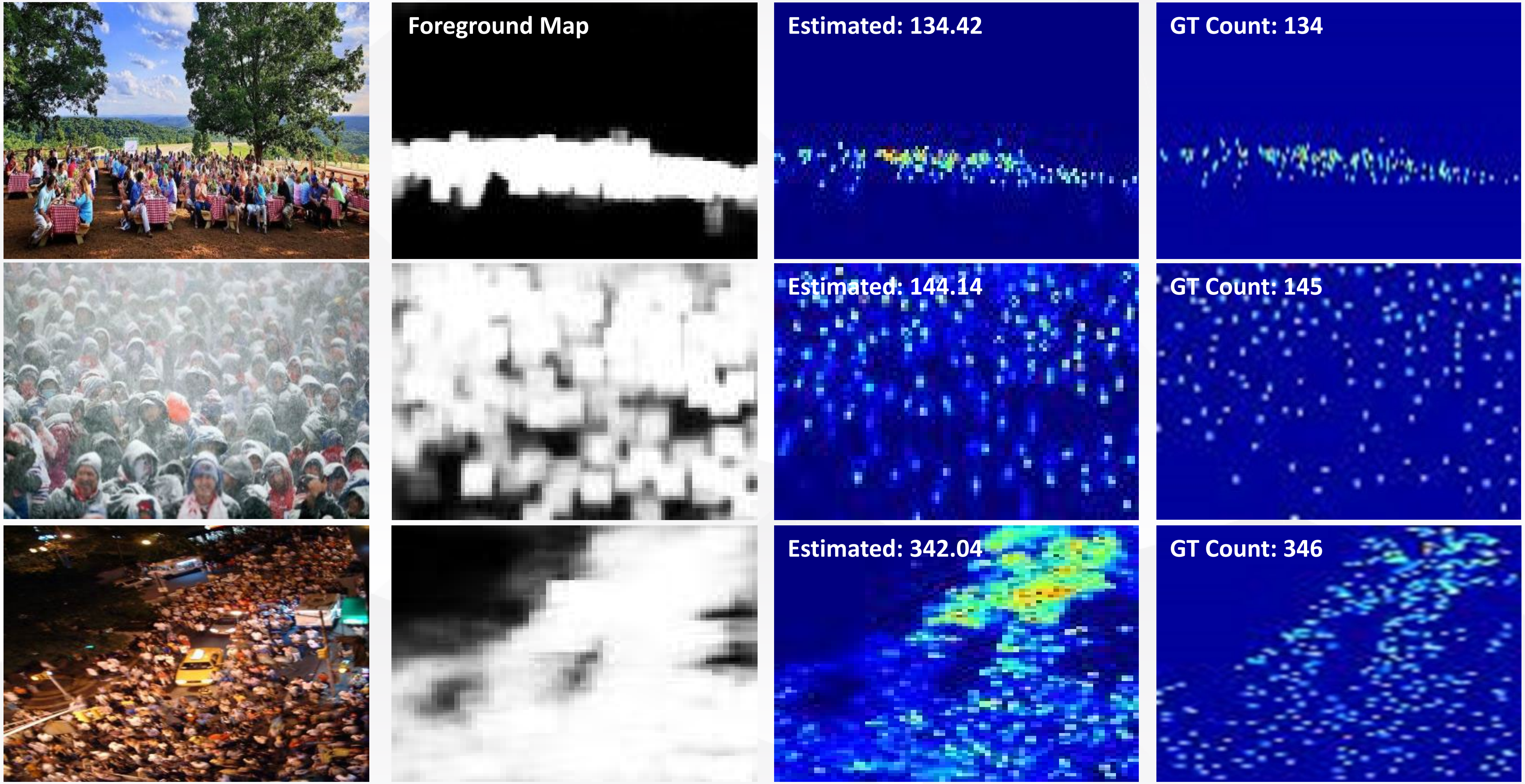
General Object Counter

Conclusion

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**

UNIVERSITY of GUELPH

Predicted Foreground & Density Maps



Ablation Studies

	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

	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

The Effect of Multi-level Auxiliator

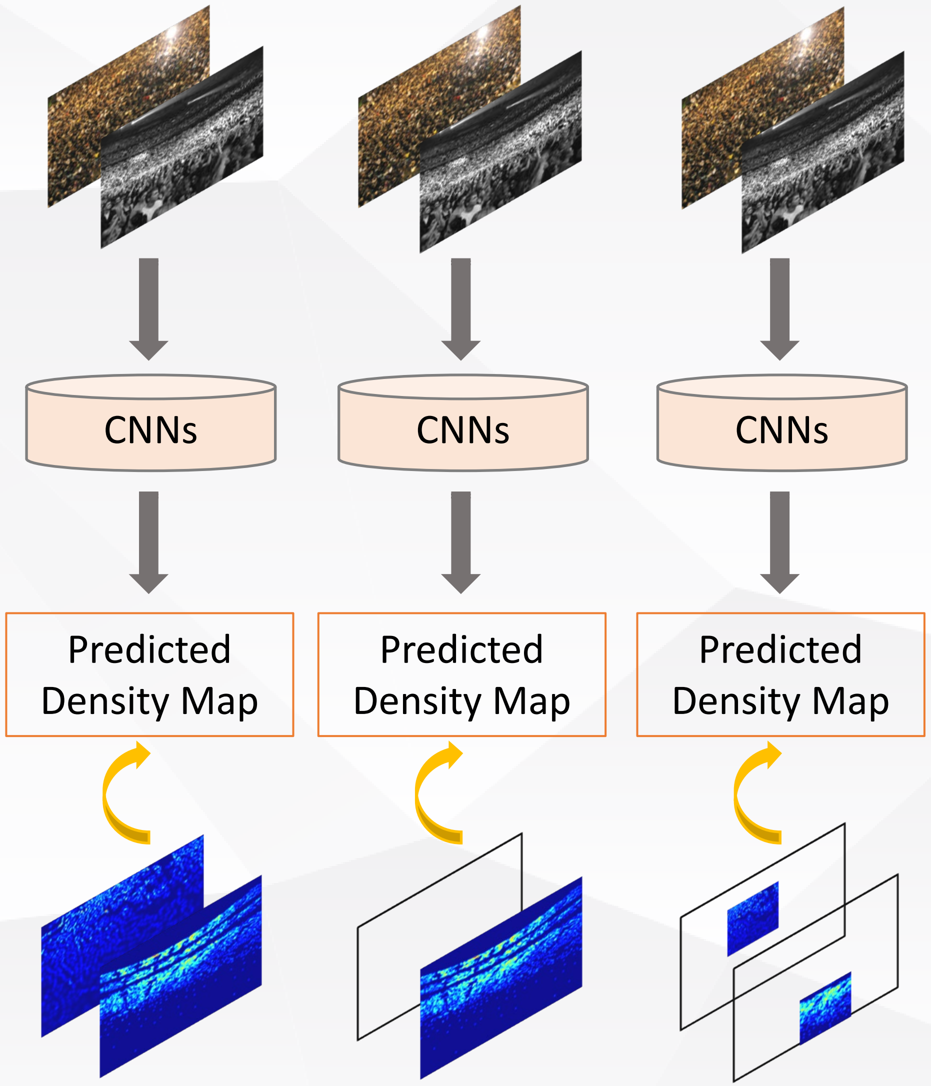
	w/o Balancing		w/ Balancing	
	MAE	MSE	MAE	MSE
Part A	56.47	97.19	52.85	83.64
Part B	6.71	11.36	6.25	10.30
UCF-QNRF	89.81	168.29	87.88	166.44

Semi Supervision ablation of the STNet

MLP-based Weakly-supervised Counter & Self-supervised Proxy Task

Pattern Recognition, 2023

Existing Location-level Supervision Methods



- Collecting location-level annotation is time-consuming & labour-intensive.
  - Efforts have been made to reduce annotation needs.
- Strongly supervised:
  - Learn from an entire set of location-level density maps.
- Weakly supervised:
  - Learn from density maps of partially selected images.
  - Learn from density maps of partial regions in crowd scenes.

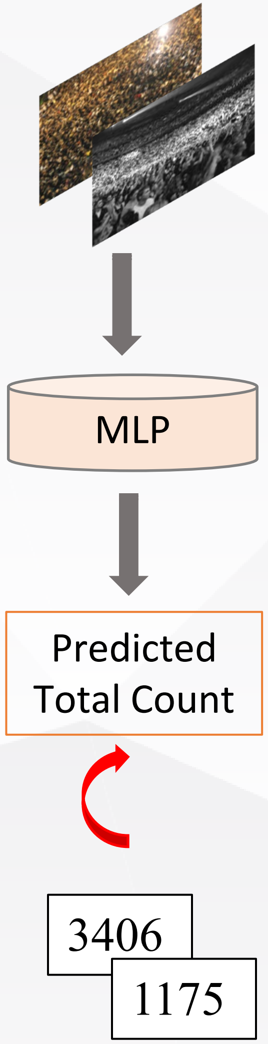
UNIVERSITY of GUELPH

- Introduction
- Scale-Invariant Counter
- Background-Aware Counter
- Weakly Supervised Counter
- General Object Counter
- Conclusion

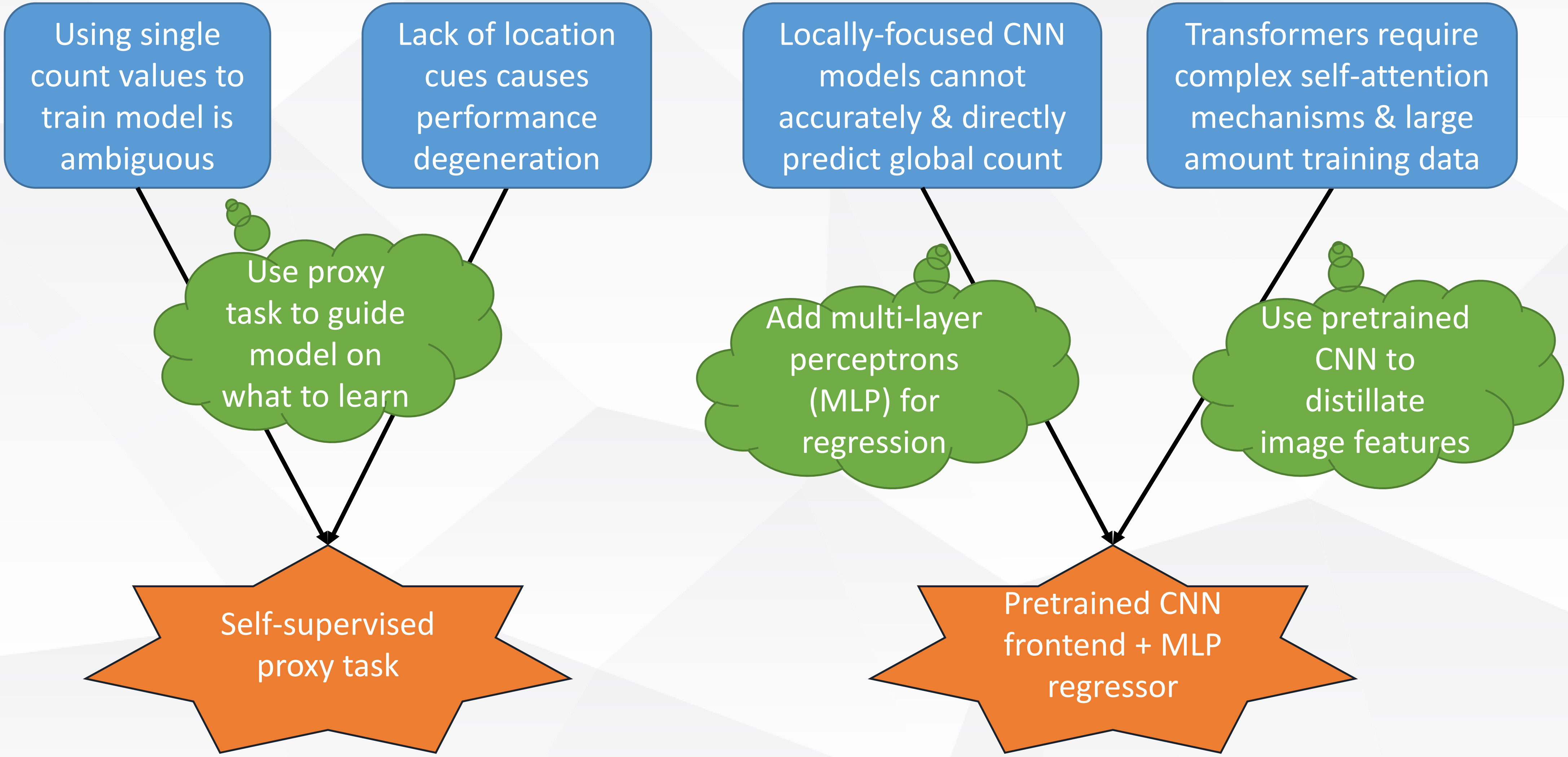
Limitations of Location-level Supervision

- Introduces a domain gap between training & inference phases
  - Models are trained to predict accurate2D density maps but evaluated solely on total counts.
- 2D density maps are derived from ground-truth dot maps through a heuristic procedure
  - May inductive bias that confuses the models on what to learn
- Collecting single counts can be much easier in many real-world scenarios.
  - E.g., scenes with controlled access

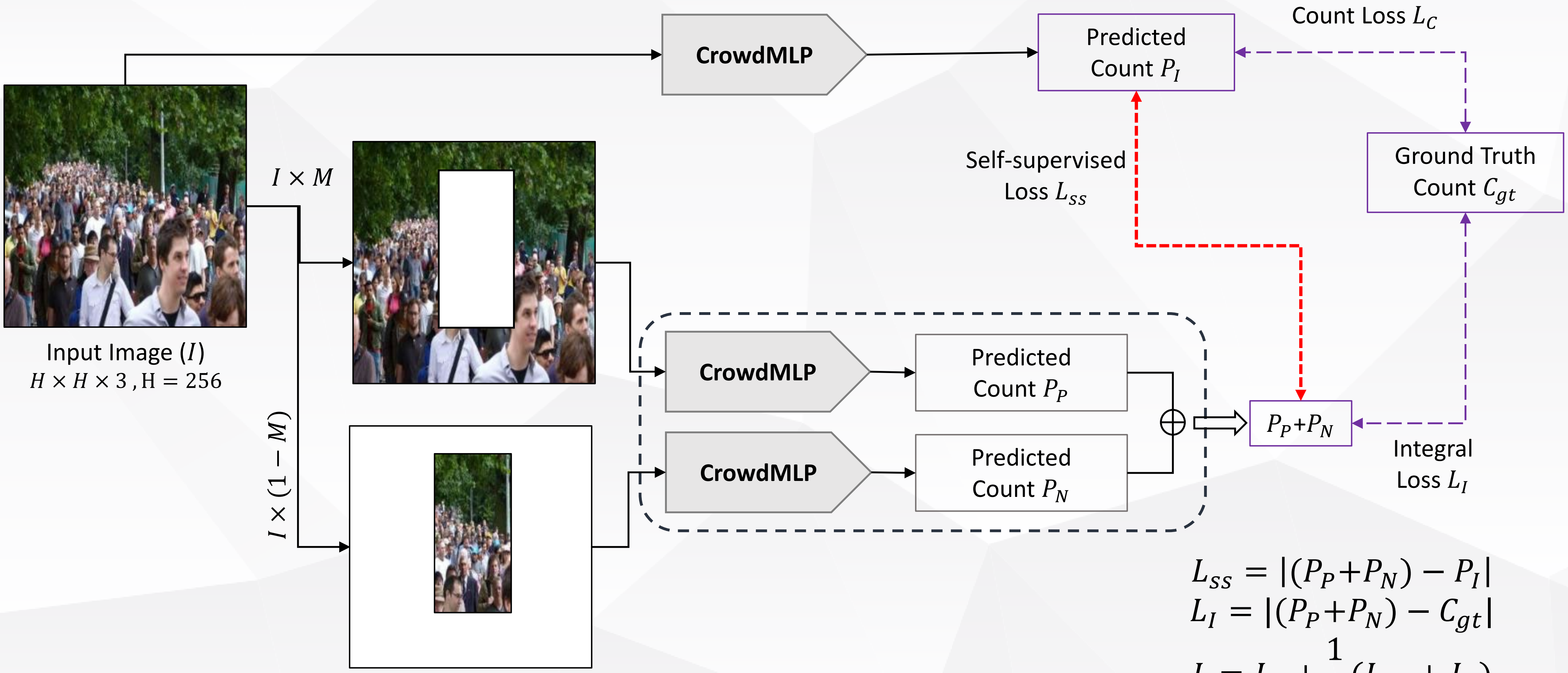
Can we train using total counts only?



Motivations

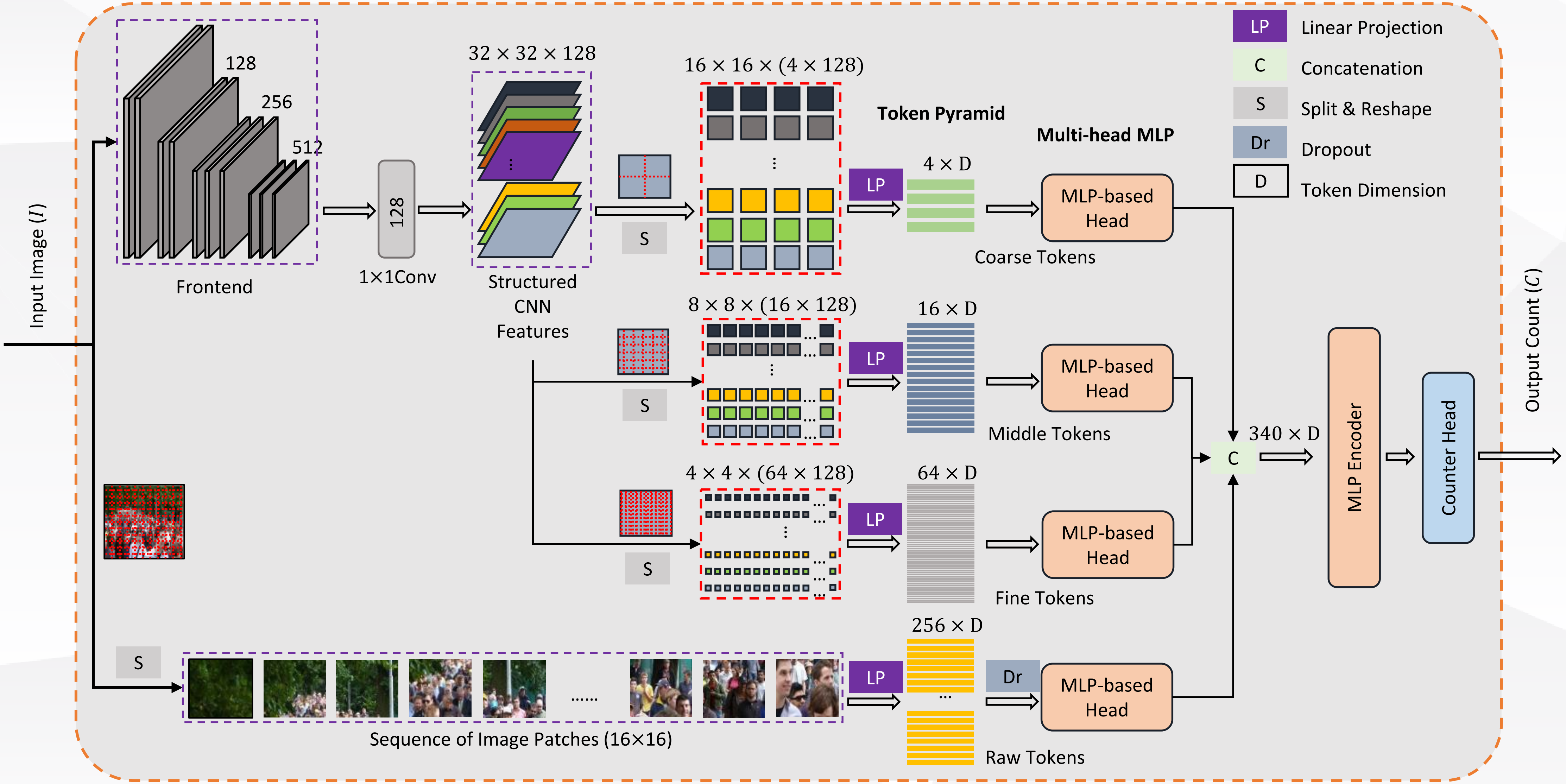


Overall Architecture w/ Split-Counting



$$L_{ss} = |(P_P+P_N) - P_I|$$
$$L_I = |(P_P+P_N) - C_{gt}|$$
$$L = L_c + \frac{1}{2}(L_{SS} + L_I)$$

Details of the CrowdMLP Counter

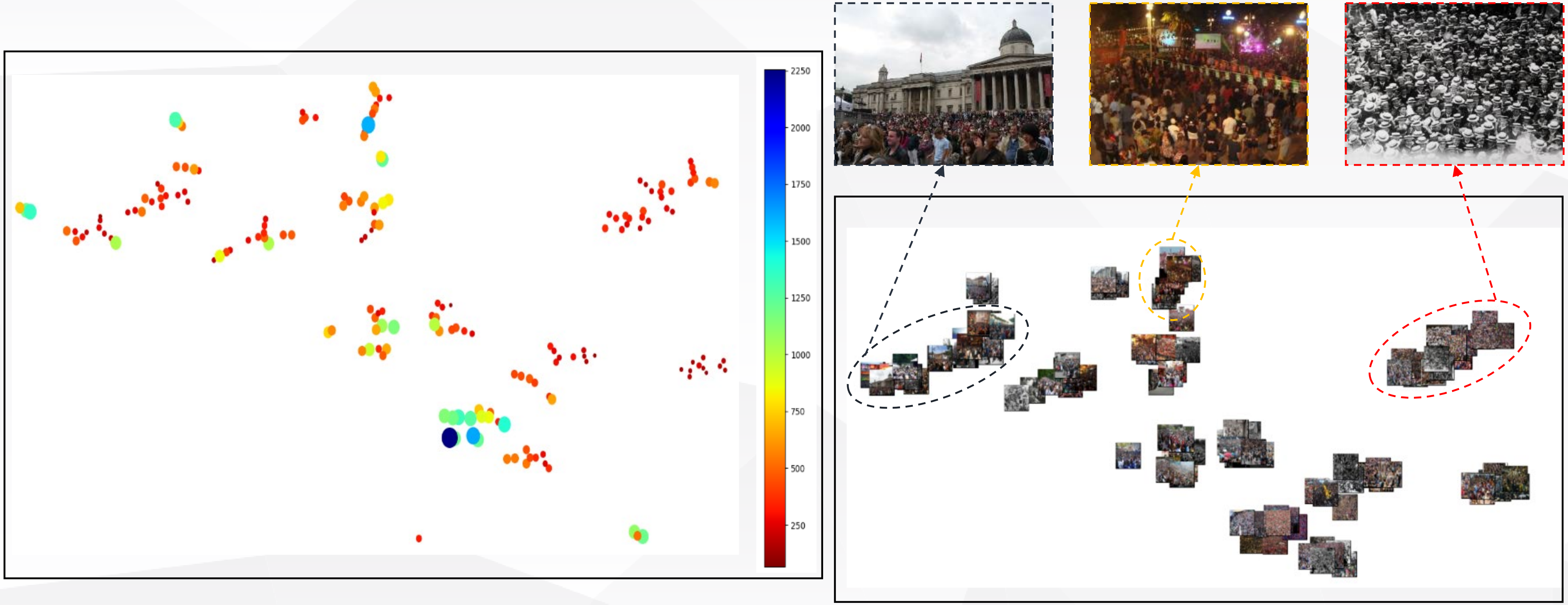


UNIVERSITY of GUELPH

Quantitative Evaluation

Methods	Location Label	Part A		Part B		UCF-QNRF		JHU++	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
ADCrowdNet [77]	✓	63.2	98.9	7.6	13.9	-	-	-	-
PACNN [114]	✓	62.4	102.0	7.6	11.8	-	-	-	-
CAN [79]	✓	62.3	100.0	7.8	12.2	107	183	100.1	314.0
MBTTBF [123]	✓	60.2	94.1	8.0	15.5	97.5	165.2	81.8	**299.1**
DSSINet [76]	✓	60.63	96.04	6.85	10.34	99.1	159.2	133.5	416.5
S-DCNet [155]	✓	58.3	95.0	6.7	10.7	104.4	176.1	-	-
ASNet [54]	✓	57.78	90.13	-	-	91.59	159.71	-	-
AMRNet [81]	✓	61.59	98.36	7.02	11.00	86.6	152.2	-	-
DM-Count [146]	✓	59.7	95.7	7.4	11.8	85.6	**148.3**	-	-
BL [91]	✓	62.8	101.8	7.7	12.7	88.7	154.8	**75.0**	299.9
P2PNet [126]	✓	**52.7**	**85.0**	**6.2**	**9.9**	**85.3**	154.5	-	-
MATT [64]	×	80.1	129.4	11.7	17.5	-	-	-	-
Sorting [161]	×	104.6	145.2	12.3	21.2	-	-	-	-
TransCrowd-T [73]	×	69.0	116.5	10.6	19.7	98.9	176.1	76.4	319.8
TransCrowd-G [73]	×	66.1	105.1	9.3	16.1	97.2	**168.5**	74.9	295.6
CrowdMLP	×	**57.8**	**84.4**	**7.6**	**12.1**	**94.1**	170.3	**67.5**	**256.1**

T-SNE Visualization on Feature Clusters



Ablation Studies

Methods	Parameters	Part A	
		MAE	MSE
w.o. Raw Token	23.70M	60.411	88.613
w.o. 16×16 Token	17.65M	61.896	93.894
w.o. 8×8 Token	23.74M	61.697	87.717
w.o. 4×4 Token	24.56M	59.294	87.483
Baseline	26.63M	57.828	84.412

The Effects of Coarse-to-fine Token Streams

Using Different Learning Paradigms

Methods	Parameters	Part A	
		MAE	MSE
Crowd-CNN	26.63M	159.925	251.649
Crowd-Transformer	36.43M	61.121	92.922
Baseline	26.63M	57.828	84.412

The effect of Proxy Split-Counting: 0.55% & 0.32% improvements on MAE & MSE

General Counting through Zero-shot Self-similarity Learning

Submitted to IEEE Transactions on Image Processing

UNIVERSITY of GUELPH

Existing Class-specific & Class-agnostic Counting

Introduction

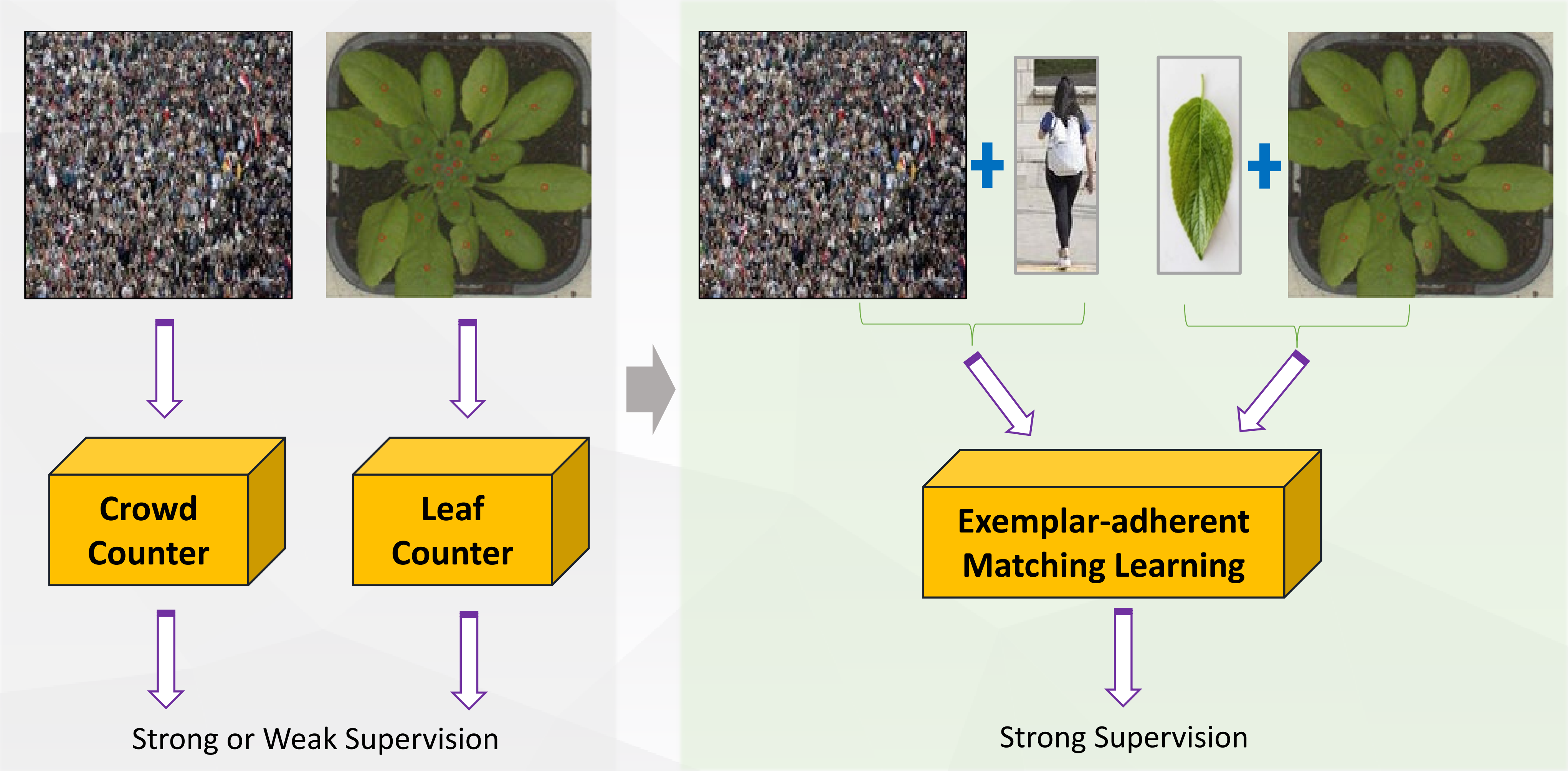
Scale-Invariant Counter

Background-Aware Counter

Weakly-Supervised Counter

General Object Counter

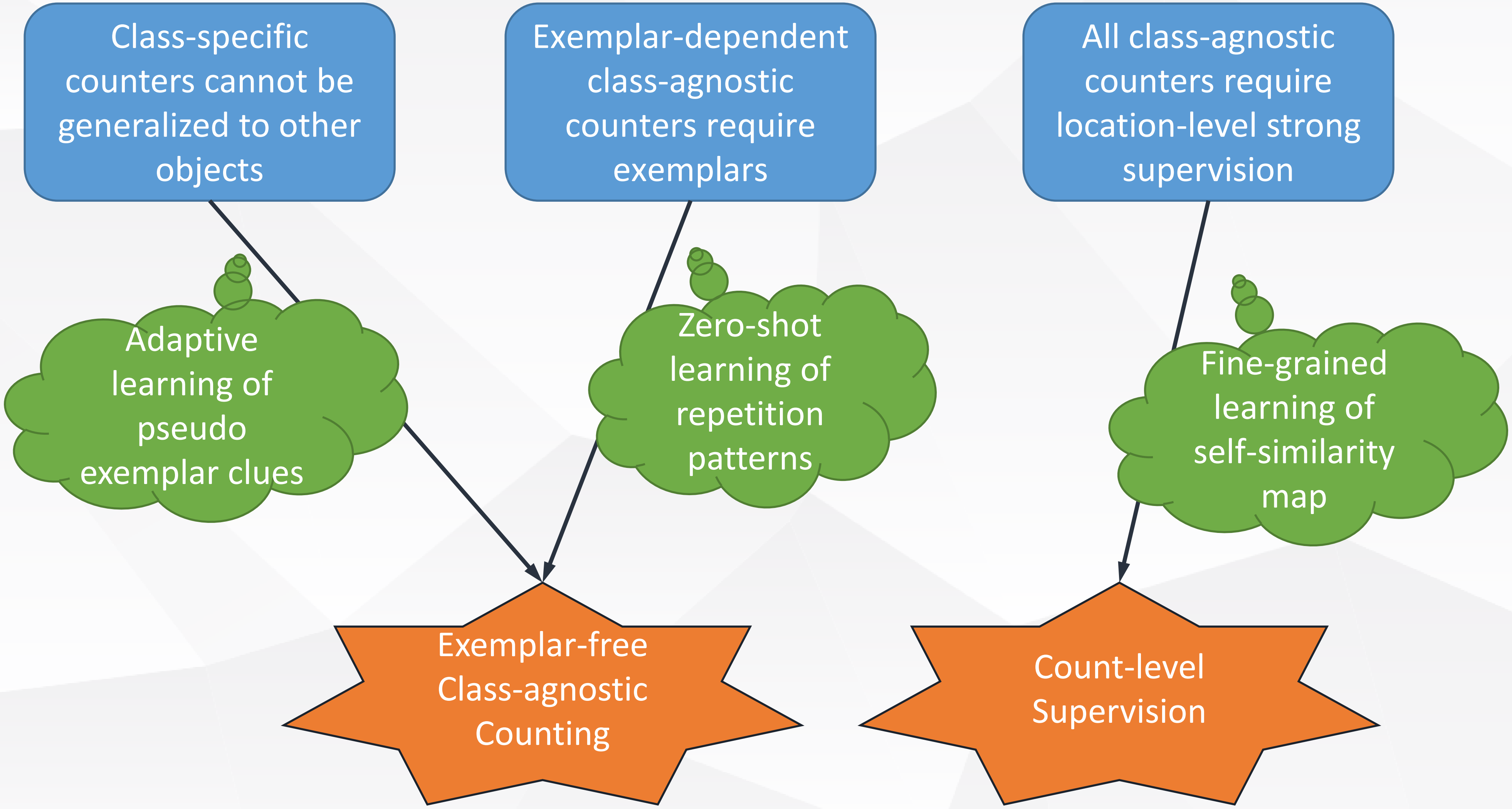
Conclusion



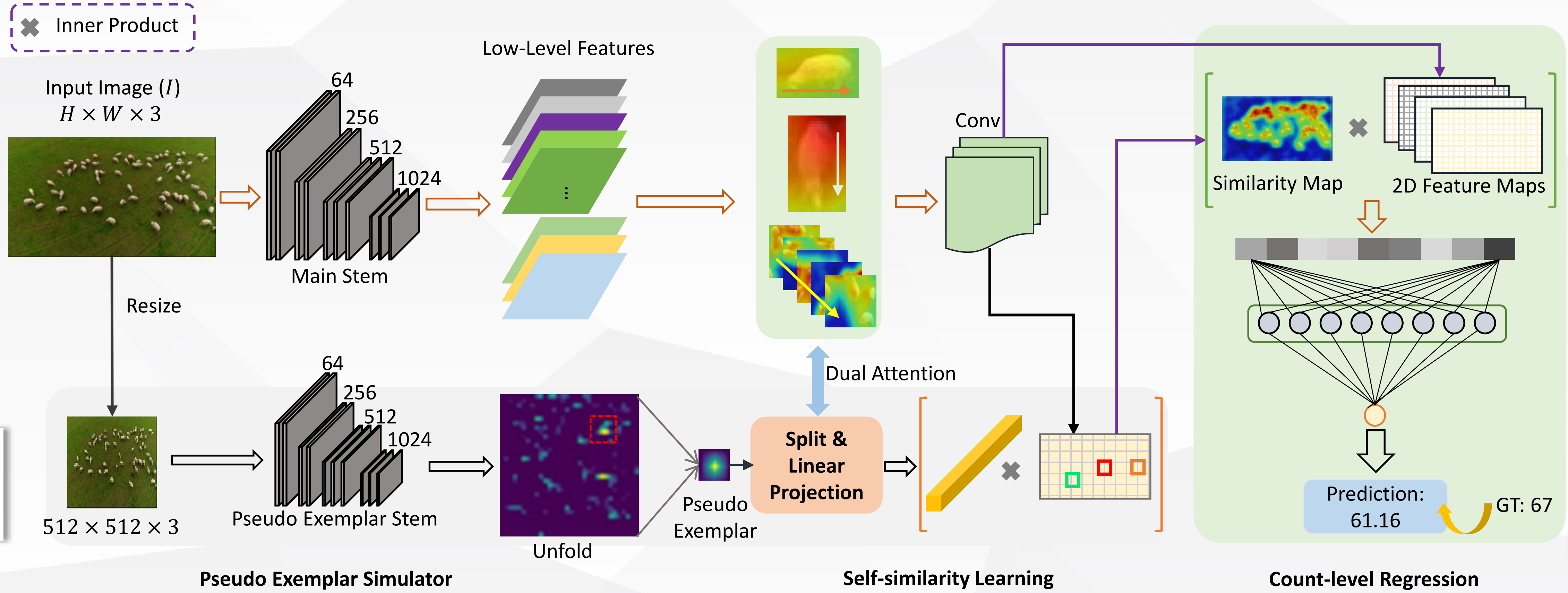
(a) Class-specific Counting

(b) Exemplar-dependent Class-agnostic Counting

Motivations

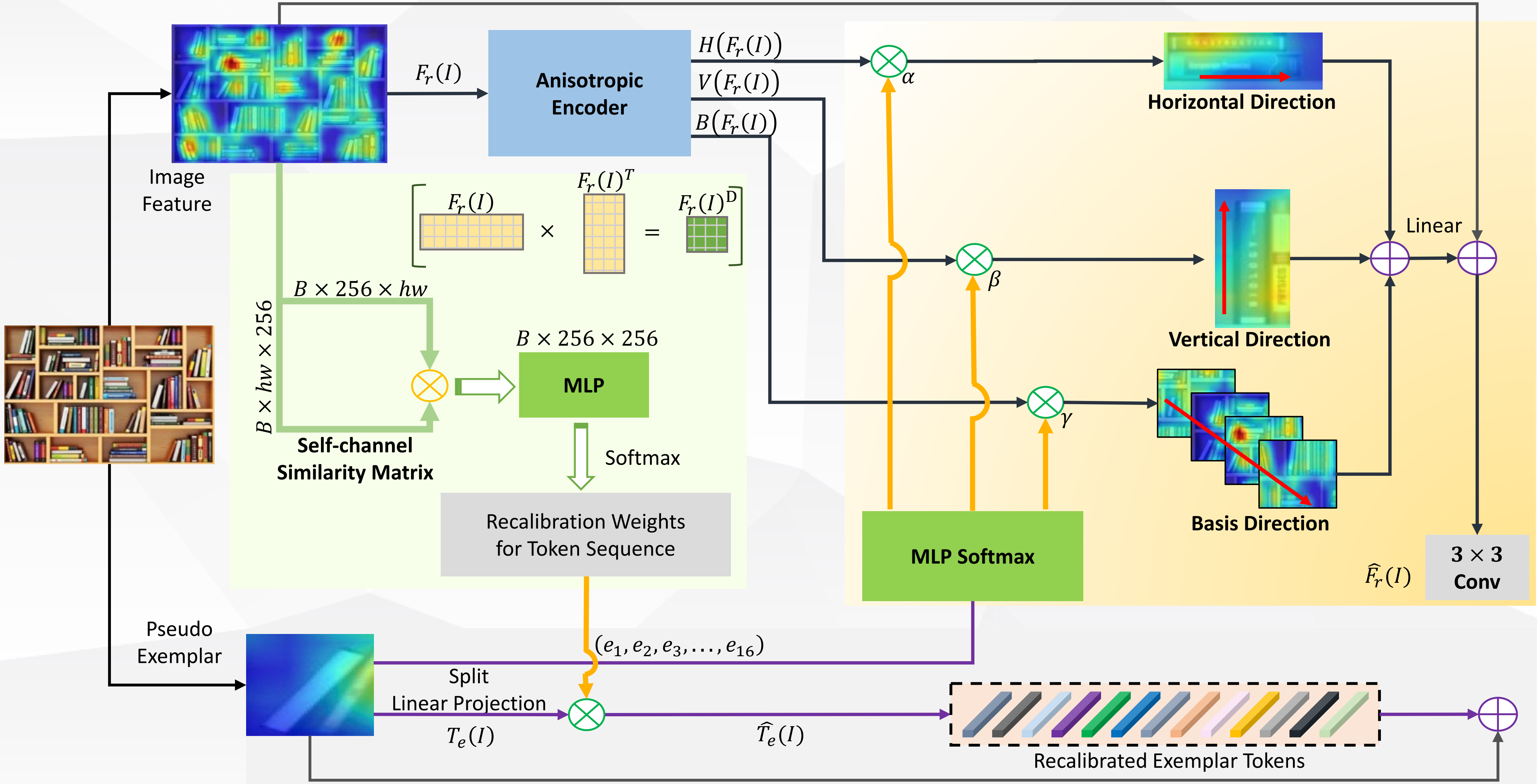


Overall Architecture for General Counter

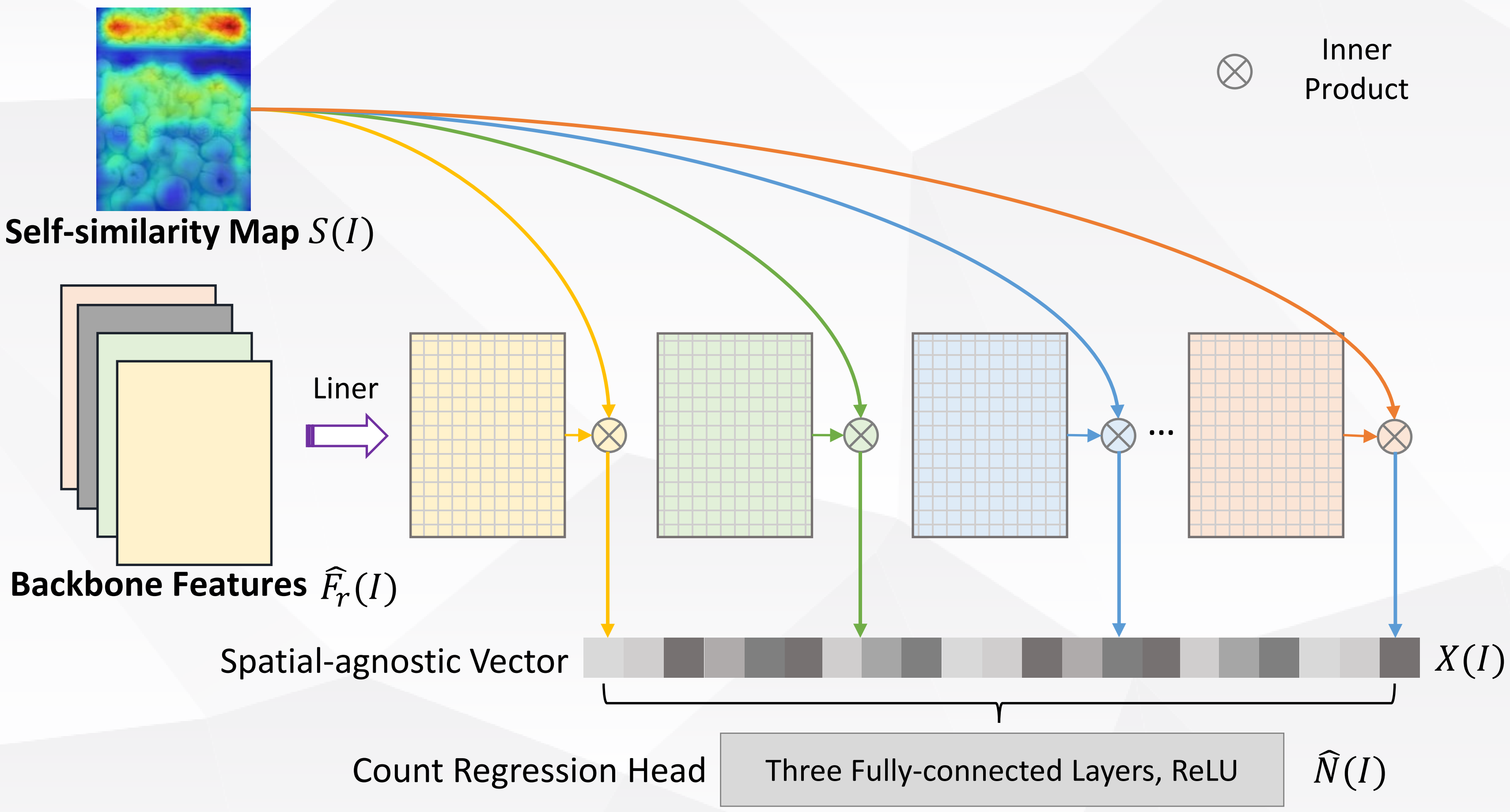


Details of Self-similarity Learning

- Introduction
- Scale-Invariant Counter
- Background-Aware Counter
- Weakly-Supervised Counter
- **General Object Counter**
- Conclusion



Weakly-supervised Location-aware Counter

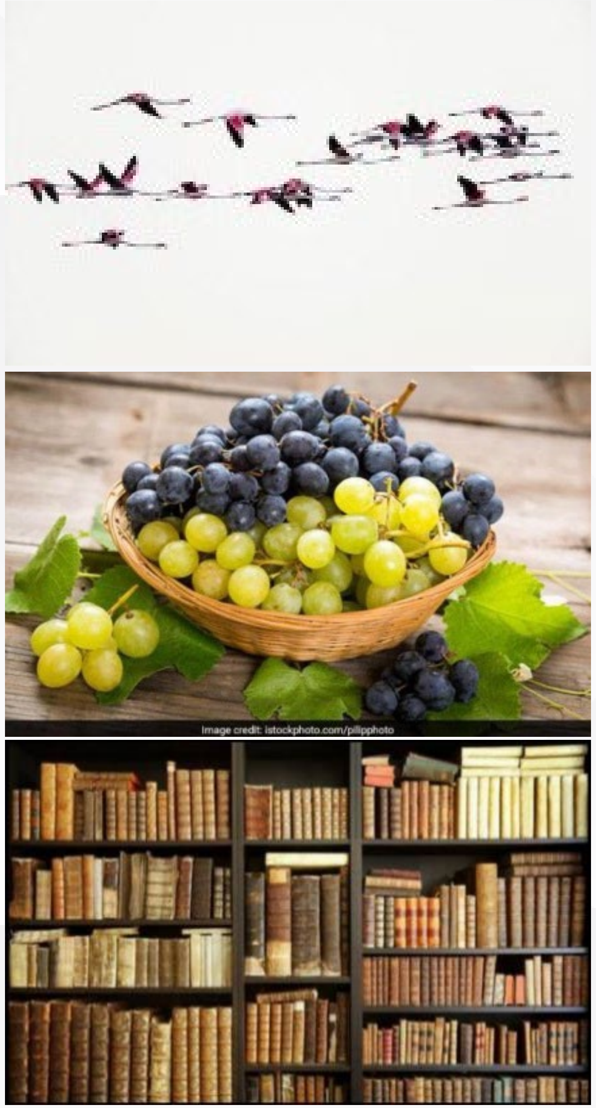


Quantitative Evaluation

Frameworks	Exemplar	Location	**MAE** for *Val*	**MSE** for *Val*	**MAE** for *Test*	**MSE** for *Test*
GMN [4]	✓	✓	29.66	89.81	26.52	124.57
FamNet [4]	✓	✓	24.32	70.94	22.56	101.54
FamNet+ [4]	✓	✓	23.75	69.07	22.08	**99.54**
CFOCNet [5]	✓	✓	21.19	**61.41**	22.10	112.71
BMNet [7]	✓	✓	**19.06**	67.95	**16.71**	103.31
RepRPN-Counter [54]	×	×	29.24	98.11	26.66	129.11
Baseline	×	×	23.14	77.30	21.88	112.37
GCNet+Exemplar (ours)	✓	×	19.61	66.22	17.86	106.98
GCNet (ours)	×	×	**19.50**	**63.13**	**17.83**	**102.89**

UNIVERSITY of GUELPH

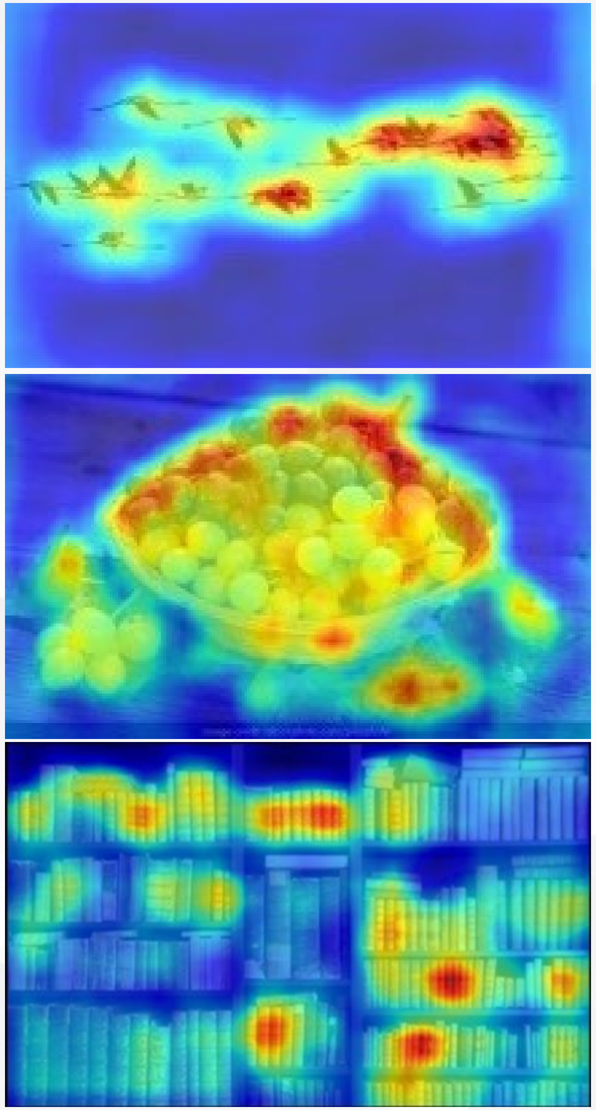
Results on General Objects



GT: 22

GT: 80

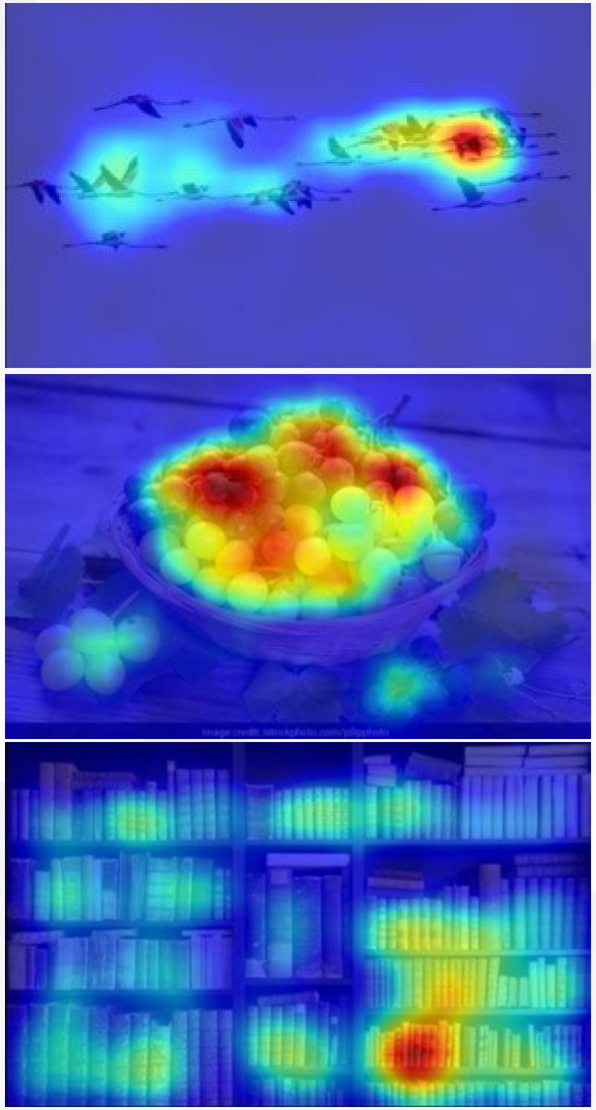
GT: 230



Ours: 21.45

Ours: 79.49

Ours: 232.35



Baseline: 28.76

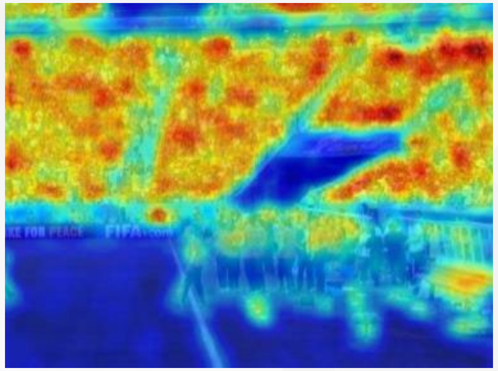
Baseline: 58.37

Baseline: 170.32

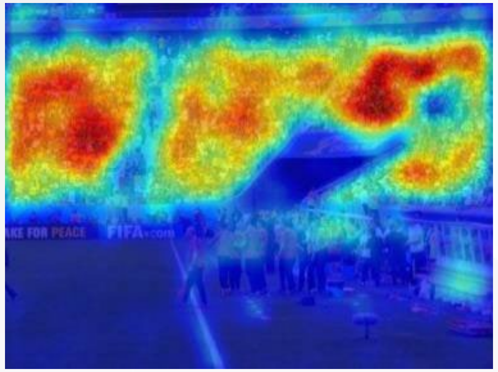
Results on Crowd Counting

GT: 1357

Ours: 1233.55

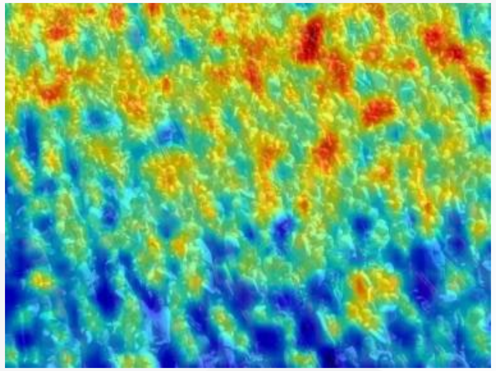


Baseline: 357.00

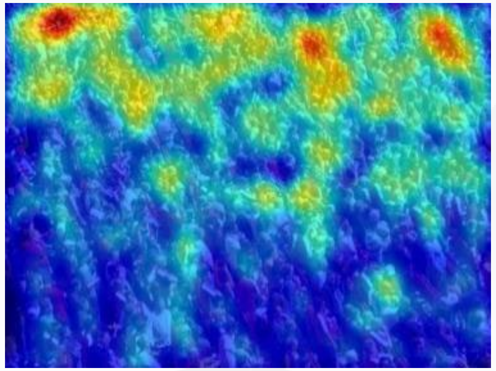


GT: 817

Ours: 818.97

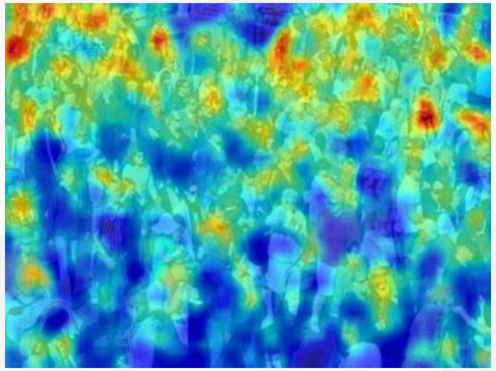


Baseline: 412.92

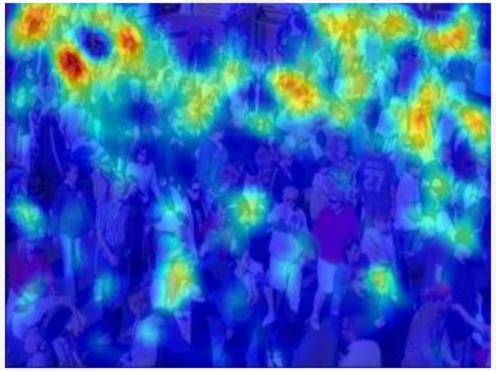


GT: 170

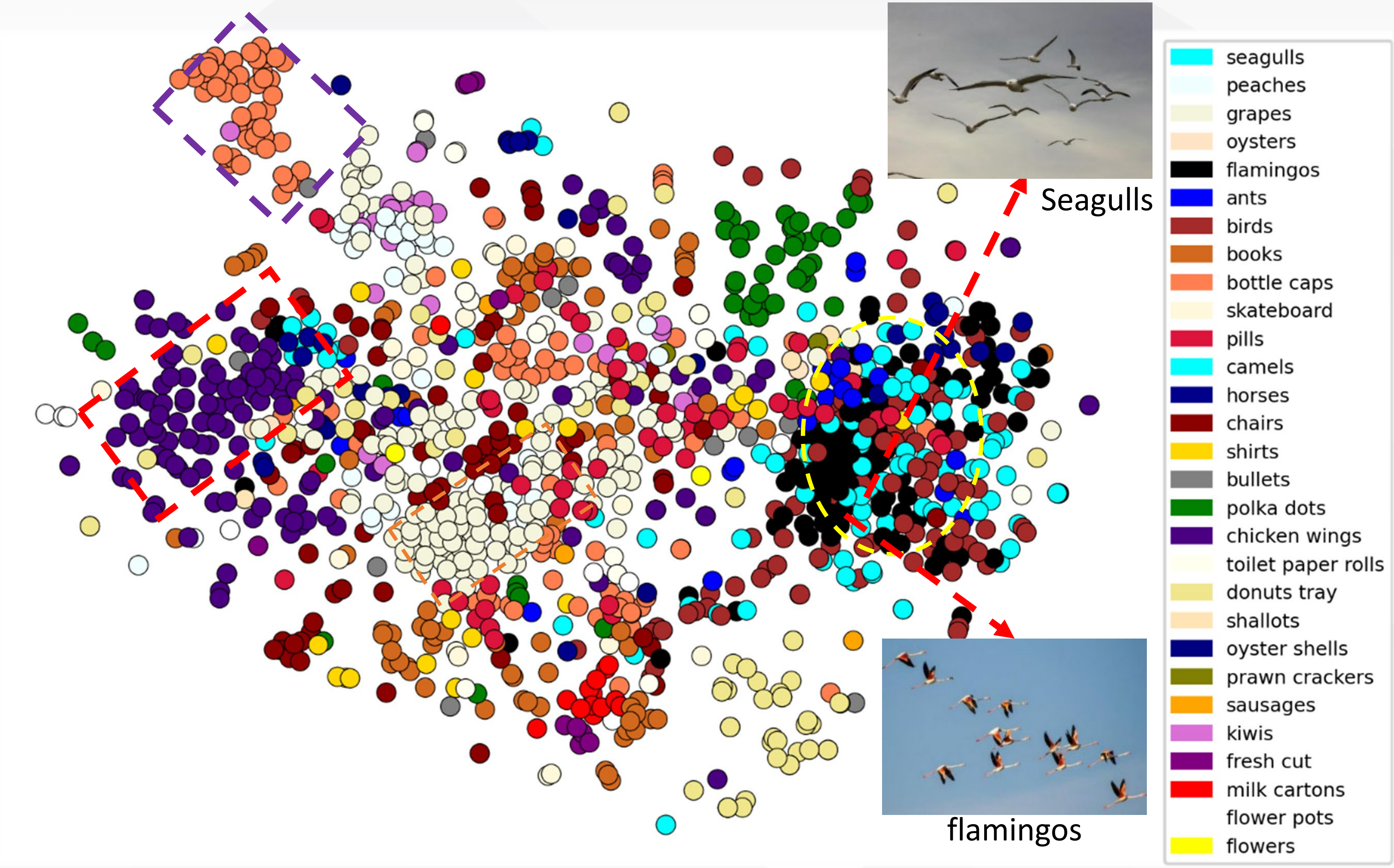
Ours: 168.65



Baseline: 116.10



Visualization of Intermediate Features

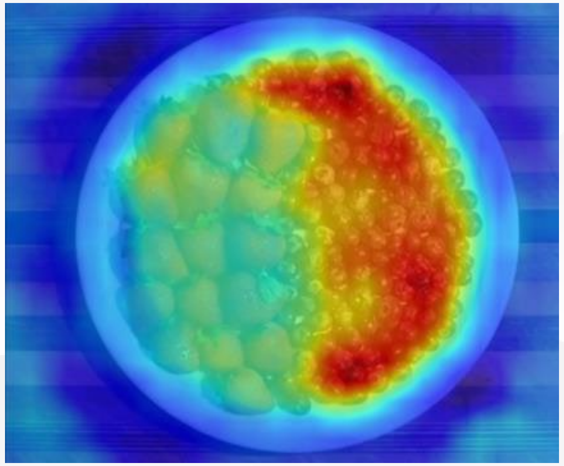


Failure Cases

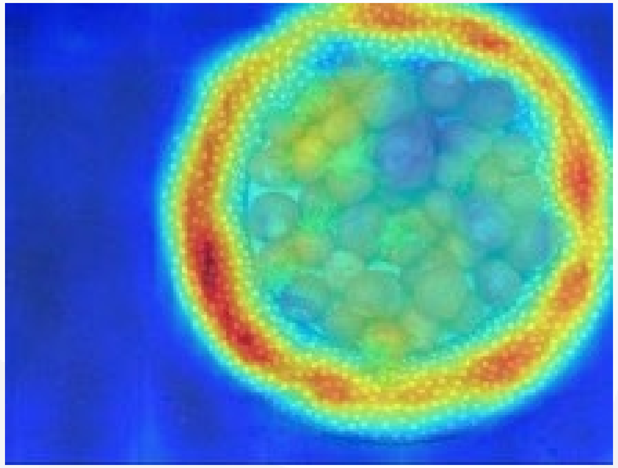
GT: 14

GT: 16

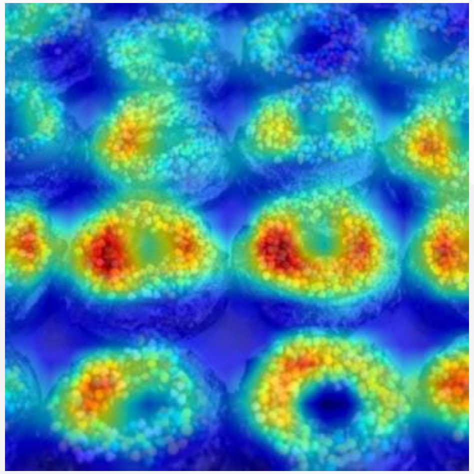
GT: 43



Ours: 63.88



Ours: 85.56



Ours: 137.43

Conclusions

- Several strategies are explored to generalize counting networks:
- Generalization on object scales:
  - Scale-Invariant modules and Tree Diversity Enhancer can effectively improve the models' generalization on counting objects with different scales/densities.
- Generalization on scene contexts (background)
  - Properly designed proxy tasks can guide the networks to recognize various background and focus on counting correct objects.
- Generalization on location-agnostic counting
  - With global MLP-based regressor, it is possible to train counters using only global count information.
- Generalization on class-agnostic counting
  - Exploring self-similarity learning enables the models to adaptively learn pseudo exemplar clues from inherent repetition patterns.

UNIVERSITY of GUELPH

Introduction

Scale-Invariant Counter

Background-Aware Counter

Weakly-Supervised Counter

General Object Counter

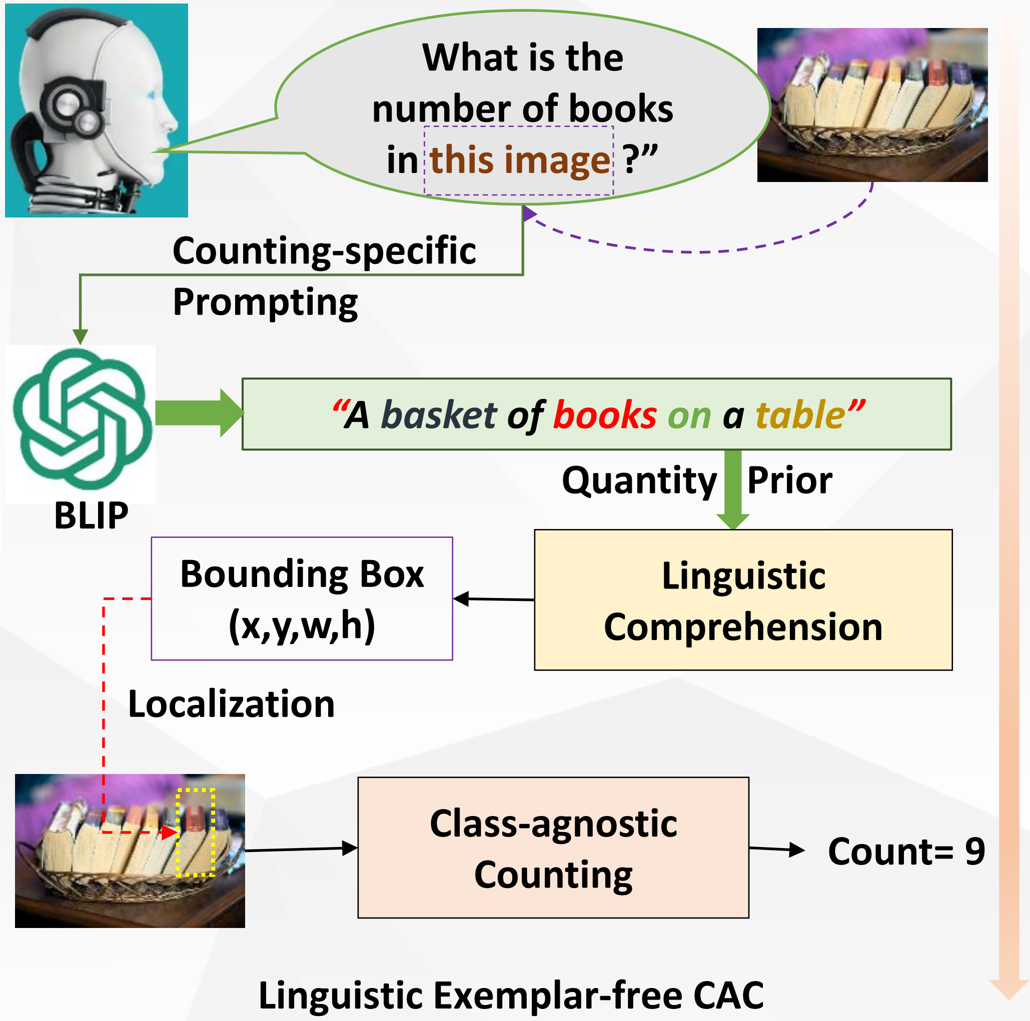
Conclusion

Future Work

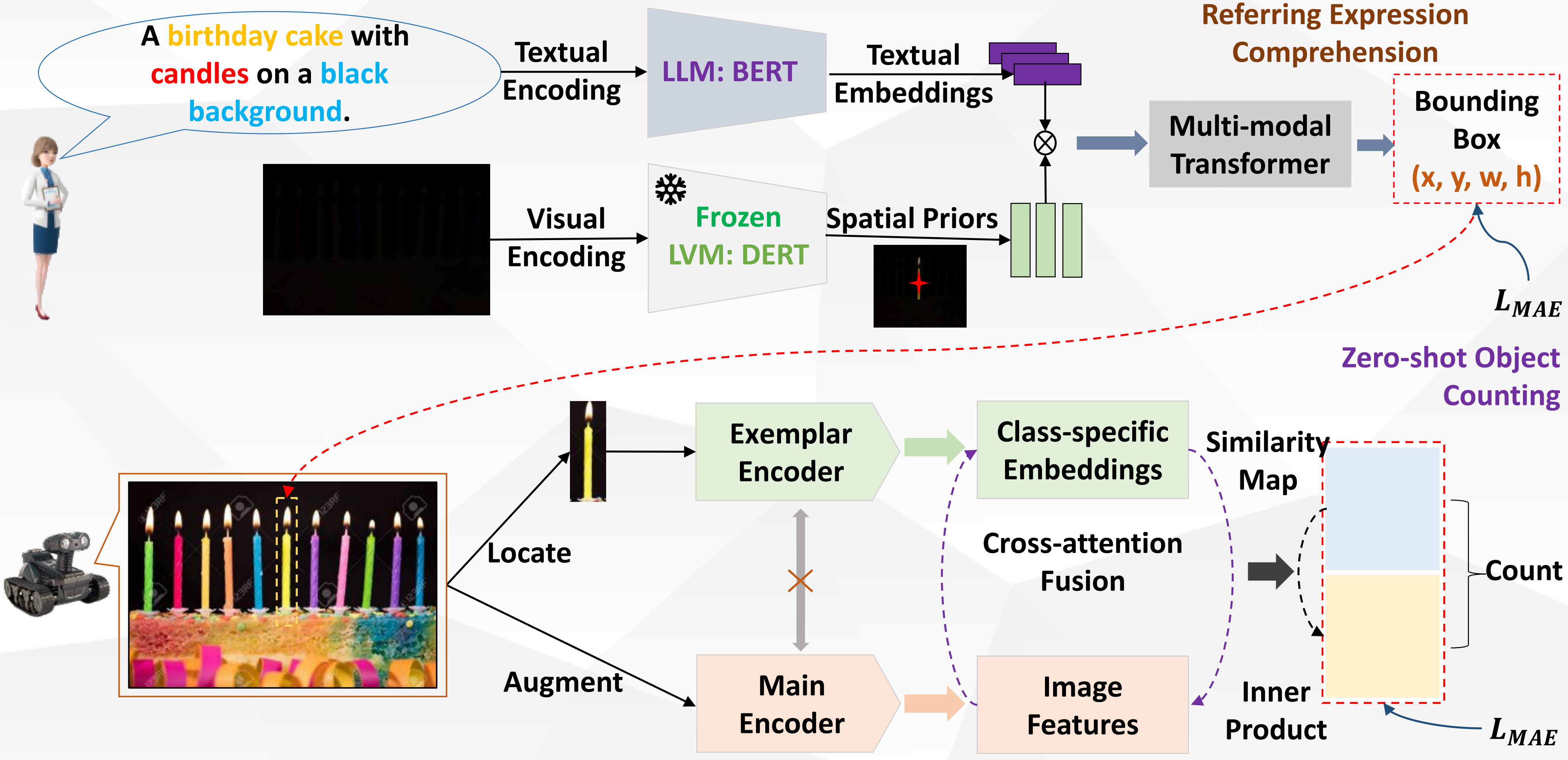
- The imbalance in regression problem:
  - Crowd scene counts exhibit a long-tailed distribution, so relative count error may be more meaningful.
- Ambiguities on counting objects:
  - Use text to provide additional cues on what to count.
- Extend counting to multi-camera:
  - Handle occlusions & redundancies.
- Extend counting to videos:
  - Handle dynamic scenes and enforce temporal coherence in video frames.
  - Capture how objects (e.g., crowds) flow through the scene.
- Real-world applications:
  - Agriculture:
    - *Digital agriculture, crop yield estimation, pest monitoring, plant disease detection, livestock management, weed control*
  - Manufacturing:
    - *Components on a circuit board, defects in a product*
  - Medical Imaging:
    - *Cancer cells in a biopsy sample, white blood cells in a blood smear, or brain lesions in a MRI scan*
  - Traffic Analysis:
    - *Vehicles, pedestrians, or bicycles*
  - Wildlife Monitoring

Multi-Modal Learning for Zero-Shot Counting

- Existing Exemplar-Free Counting (EFC) methods are commonly trained using rudimentary image/class name signals.
  - Large-scale Language Models provide much detailed descriptions for objects in scenes than image/class names.
  - Steer zero-shot counting via referring expression comprehension.
- EFC approaches often extract exemplars by scanning the entire input image.
  - Use language cue (e.g., "a basket of books on a table") to find exemplar location quickly.



Overall Architecture for ExpressCount



- This talk introduces strategies for achieving generalization in machine learning counting problems, a challenging area requiring output of a single, unrestricted value. Traditional methods often overfit similar-density inputs, need location-level annotations, and depend on explicit exemplars for diverse object counts. Our approaches mitigate these issues by managing density variations, eliminating the need for location annotation, and identifying adaptive exemplars in input images.
- We address large scale variation and density shift in test data via scale-invariant features and hierarchical scale information parsing. Our model eschews location annotation reliance by training with total counts only, using a novel counter and a multigranularity MLP regressor. To overcome sample limitation and spatial hint shortages, we propose a self-supervised task, Split-Counting.
- For counting different object categories without specific exemplars, a zero-shot generalized counter is introduced, which learns pseudo exemplar clues from repetition patterns and captures spatial location hints via a self-similarity learning strategy.
- Our extensive experiments validate these strategies' efficacy in enhancing generalization for counting problems, substantiated by ablation studies and visualization analysis.