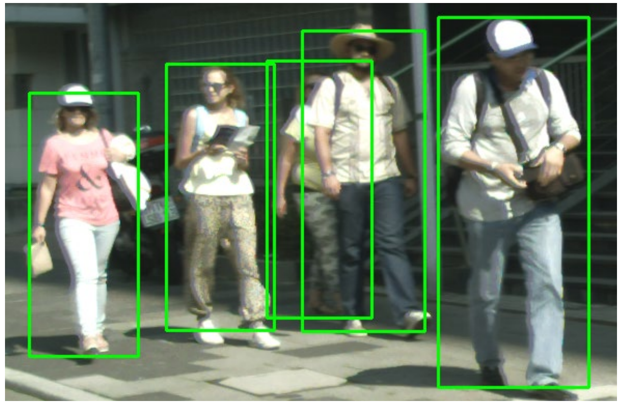
The First Step in Digital Identity in the Wild: Human Detection

Ioannis A. Kakadiaris

University of Houston

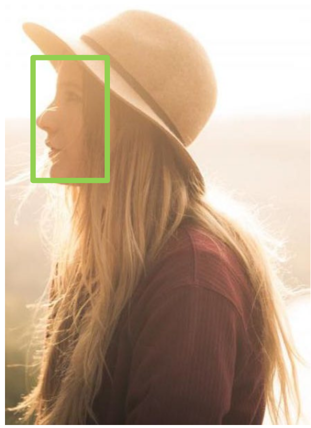
Problem Statement: Human Body Detection

Classification	Classification + Localization	Human Body Detection
		

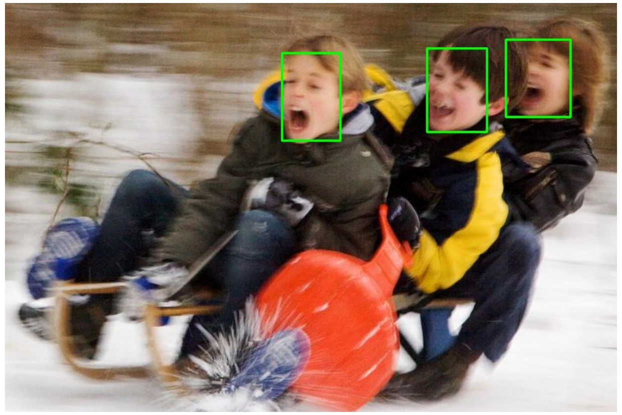
Problem Statement: Face Detection

Classification

Classification + Localization



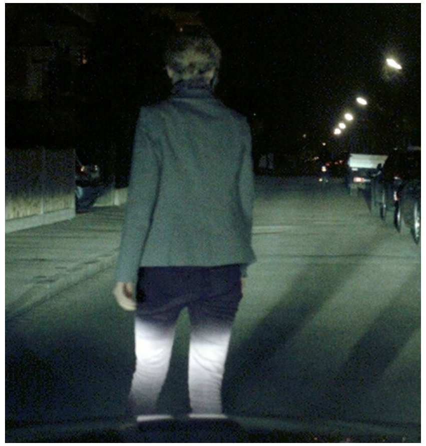
Face Detection



Challenges: Human Body Detection

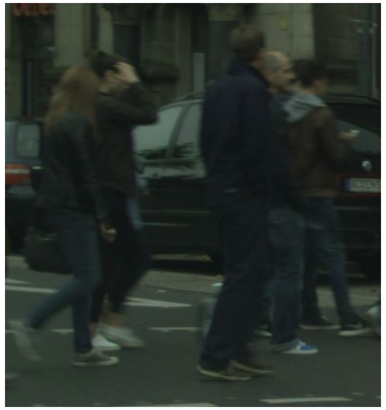
Pose

Illumination



Challenges: Human Body Detection

Occlusion



Multi-scale

Challenges: Face Detection

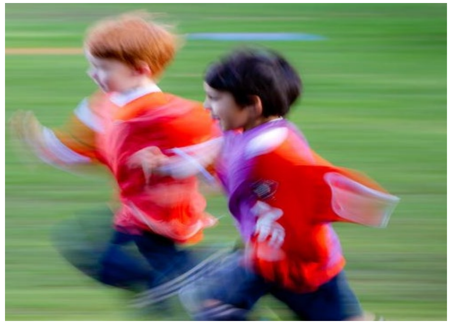
Pose

Expression

Illumination

Challenges: Face Detection

Blur



Occlusion

Multi-scale

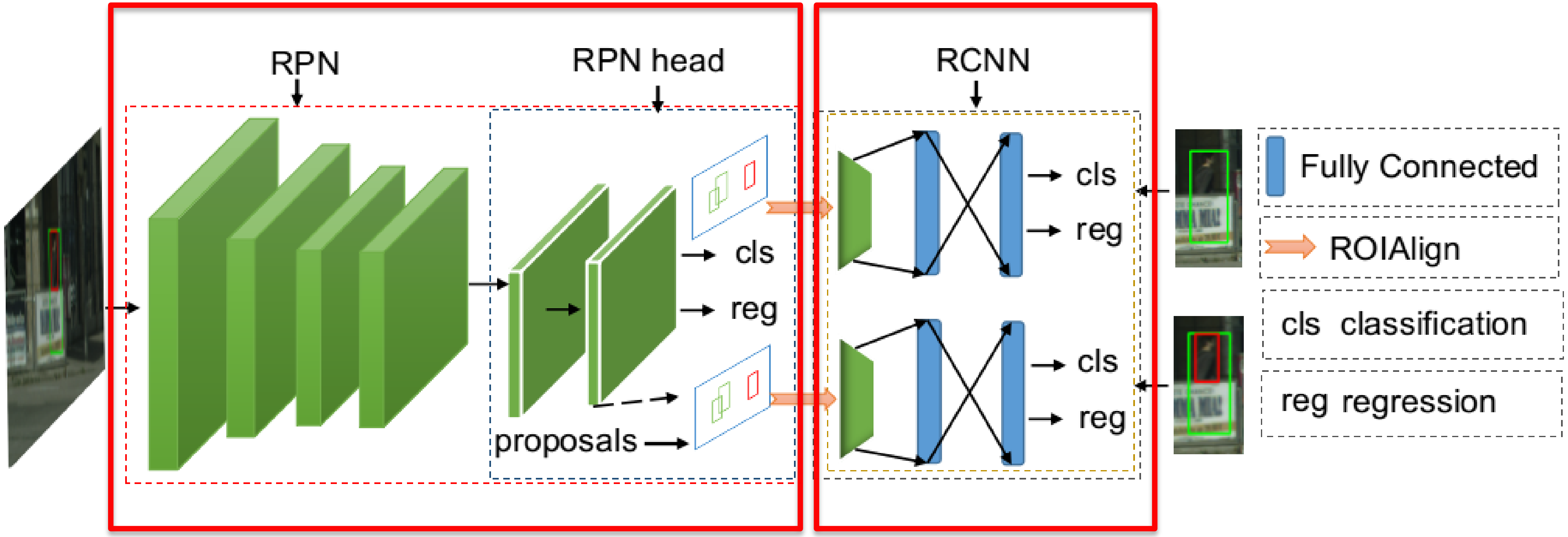
Goal

To design, develop, and evaluate human detection algorithms in th
wild.

Objective 1

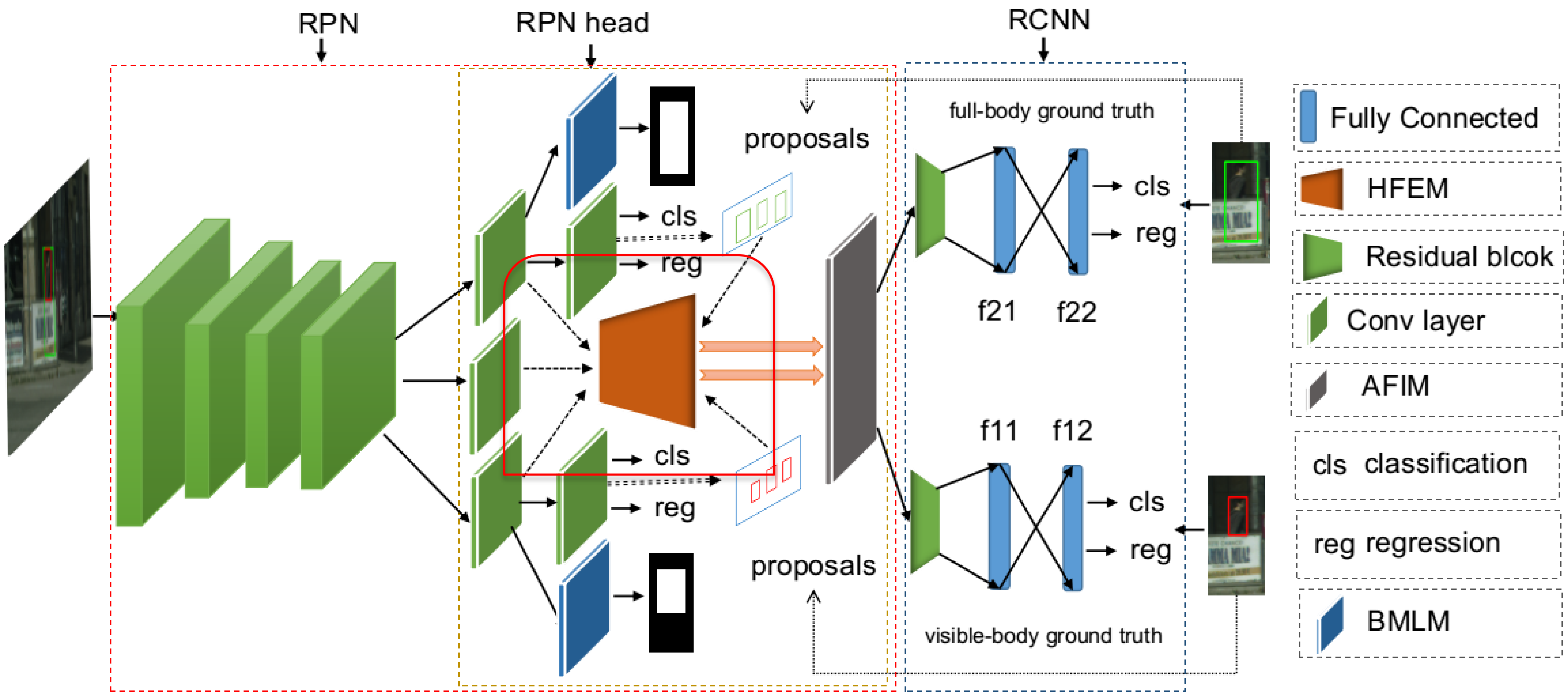
Design, develop, and evaluate a **human detector for** 2D images to overcome occlusion challenge in the wild.

Obj. 1: Related Work



[1] C. Zhu, J. Yuan. Bi-box Regression for Pedestrian Detection and Occlusion Estimation. ECCV, 2018.

DVRNet+


RPN
RPN head
RCNN
proposals
cls
reg
cls
reg
proposals
full-body ground truth
f21
f22
cls
reg
f11
f12
cls
reg
visible-body ground truth
Fully Connected
HFEM
Residual blcok
Conv layer
AFIM
cls classification
reg regression
BMLM

Head-aware Feature Enhancement Module

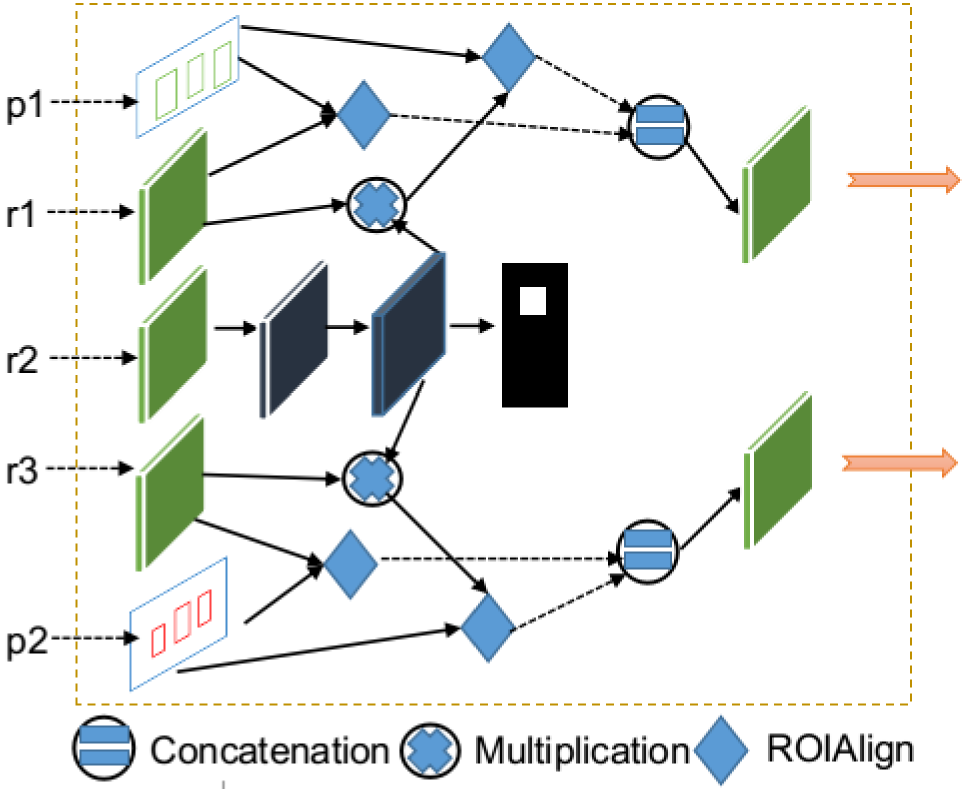
- Use a head supervision signal and a supervised attention mechanism jointly in the RPN stage (HFEM) to provide stable and discriminative information for the network to learn human features

Rationale

- The head could provide more stable (than the visible-body) information to the network because it is rarely occluded
- The head appearance is more discriminative than the visible body

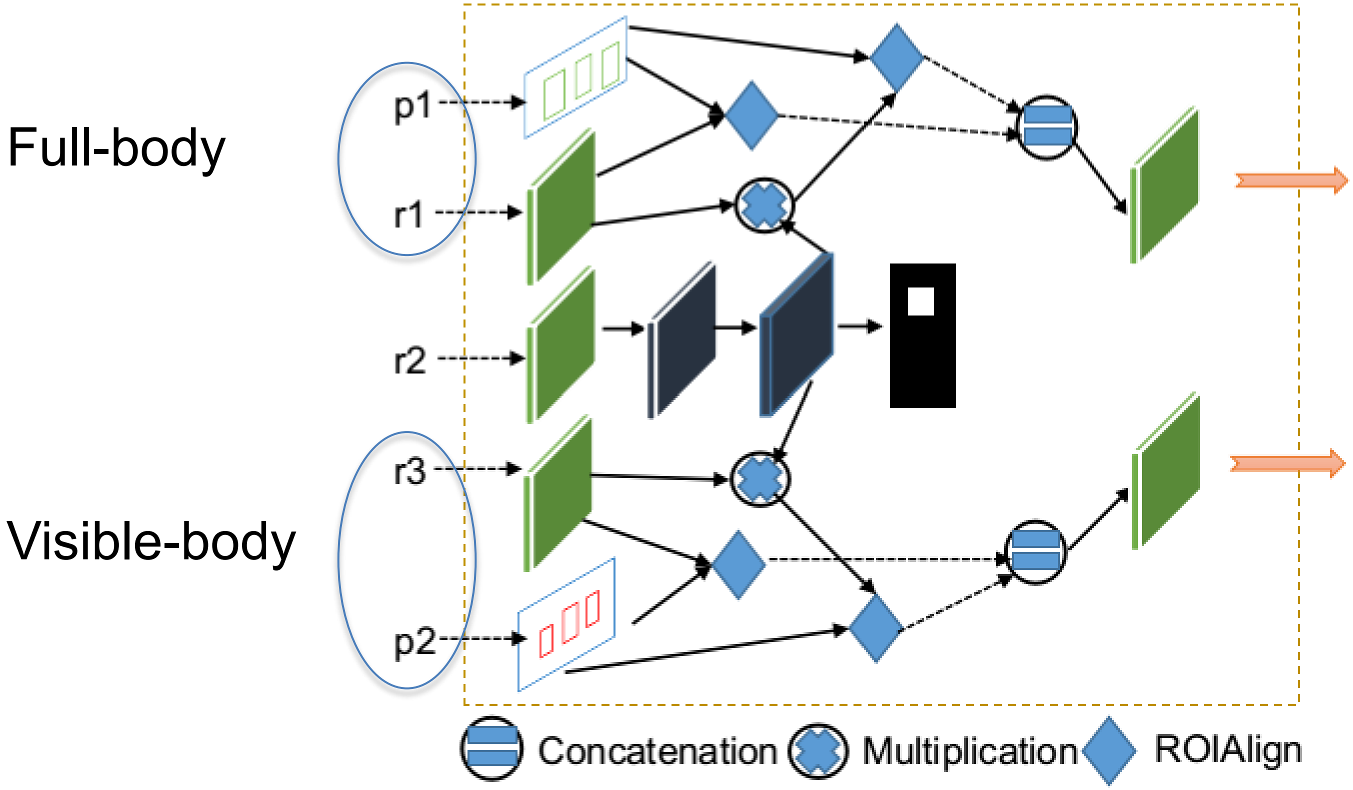
HFEM

- Depiction of the architecture of the Head-aware Feature Enhancement Module. All convolutional layers have the same kernel size of 3x3, padding of 1, and stride of 1.



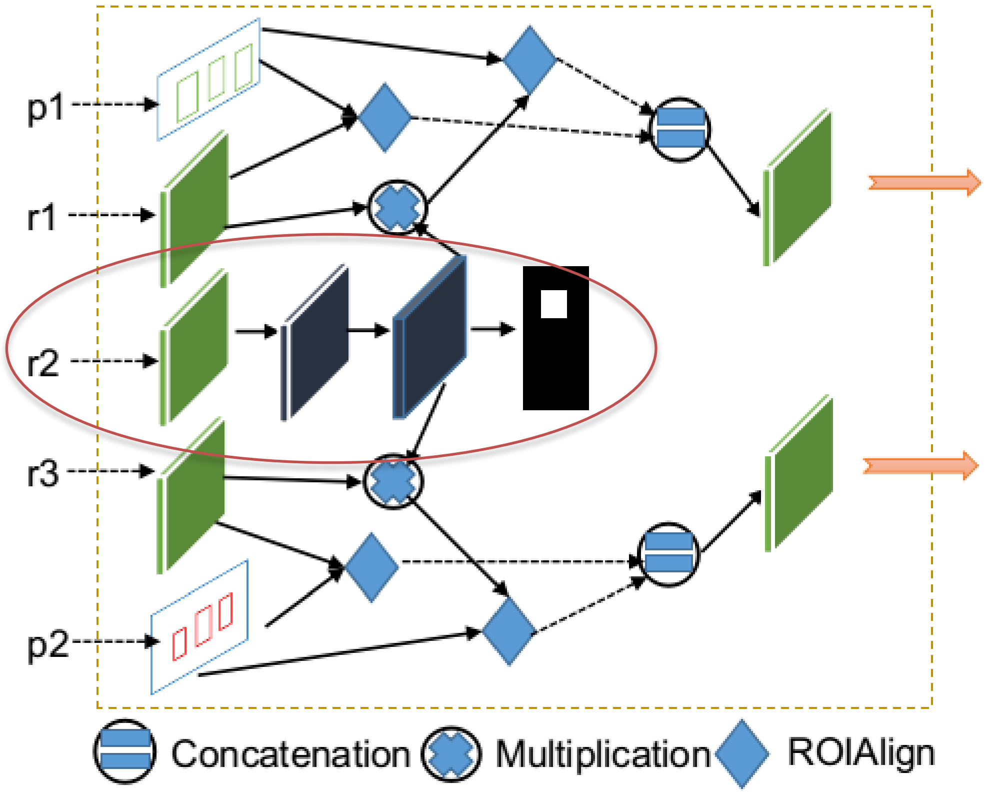
HFEM (2)

- Depiction of the architecture of the Head-aware Feature Enhancement Module. All convolutional layers have the same kernel size of 3x3, padding of 1, and stride of 1.



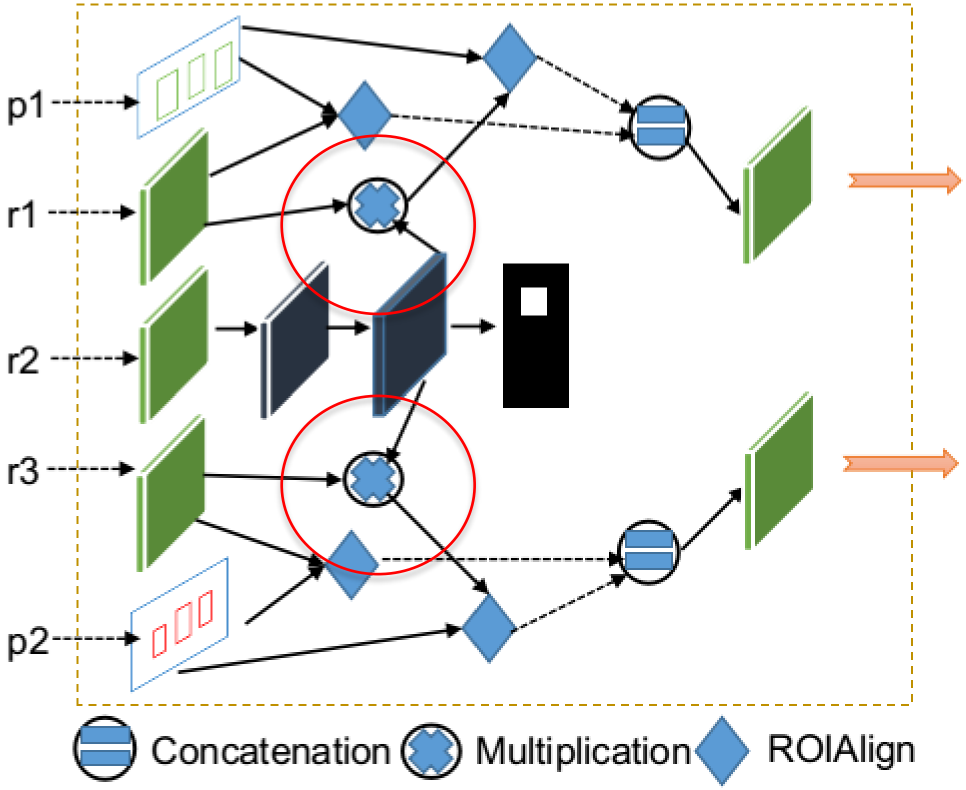
HFEM (3)

- Depiction of the architecture of the Head-aware Feature Enhancement Module. All convolutional layers have the same kernel size of 3x3, padding of 1, and stride of 1.



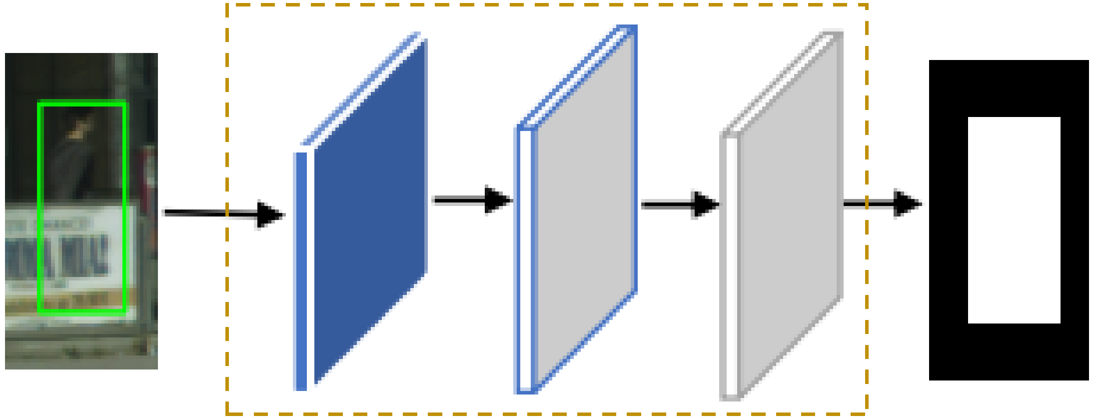
HFEM (4)

- Depiction of the architecture of the Head-aware Feature Enhancement Module. All convolutional layers have the same kernel size of 3x3, padding of 1, and stride of 1.



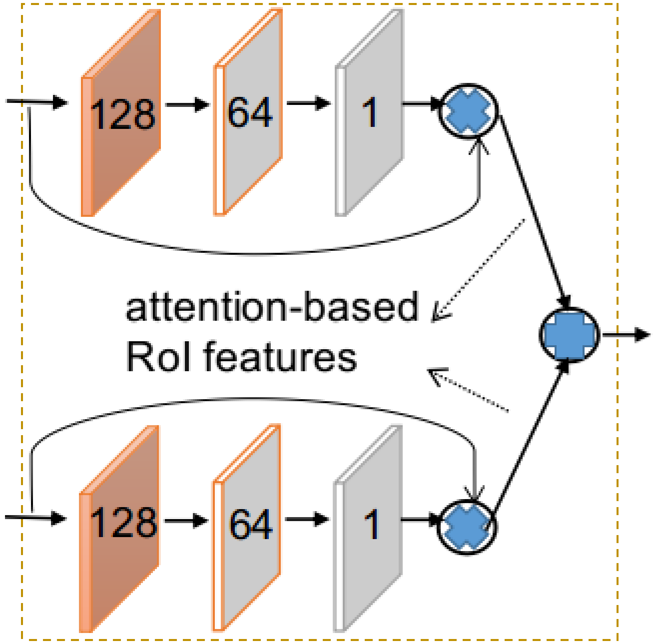
Binary Mask Learning Module

- Depiction of the architecture of the Binary Mask Learning Module. All convolutional layers have the same kernel size of 3x3, padding of 1, and stride of 1.



Attention-based Feature Interleaver Module

- Depiction of the architecture of Attention-based Feature Interleaver Module. All convolutional layers have the same kernel size of 3x3, padding of 1, and stride of 1.



Dataset

- Most widely used dataset for human detection

Dataset	Set	Images
CityPersons	Training	2,975
CityPersons	Validation	500
CityPersons	Testing	1,525

Dataset

- Multi-modality (VIS-NIR) dataset for which the performance is not saturated

Dataset	Set	Images
EDGE20	Day(VIS)	2,694
EDGE20	Night(NIR)	797

Evaluation Metric

Log average miss-rate (MR^{-2}): Log average miss-rate is calculated by averaging miss rate (MR) at ten FPPI rates evenly spaced in log-space in the range 10^{-2} to 10^{0}.

FPPI = False Positive / number of tested images
MR = False Negative / number of ground truth boxes

Lower is better

Evaluation Metric (2)

Input:
1. detected bounding boxes
2. ground truth bouding boxes

Output: MR^{-2}

1. Match detection bounding boxes with ground truth bounding boxes in terms of IoU value threshold (0.5). The matched detection bounding box is true positive, the dismatched detection bounding box is false positive.

2. Compute FPPI and MR
   FPPI = False Positive / number of tested images
   MR = False Negative / number of ground truth boxes

3. Compute MR^{-2}
   Averaging miss rate at ten FPPI rates evenly spaced in log-space in the range 10^{-2} to 10^{0} ([0.0100, 0.0178, 0.03160, 0.0562, 0.1000, 0.1778, 0.3162, 0.5623, 1.000]).

Baselines

Paper Source	Abbreviation
Zhou et al. ECCV 2018	Bi-Box
Wang *et al*. CVPR 2018	Repulsion Loss
Zhang *et al*. ECCV 2018	OR-CNN
Liu *et al*. CVPR 2019	Adaptive-NMS
Pang *et al*. ICCV 2019	MGAN

Quantitative Results

MR^{-2} on subsets of validation set of CityPersons

Methods	Backbone	MR^{-2}
Bi-Box	VGG-16	11.24
OR-CNN	VGG-16	11.0
Repulsion Loss	ResNet-50	10.9
Adaptive-NMS	ResNet-50	10.8
MGAN	VGG16	10.5
DVRNet+	ResNet-50	10.5

Comparison of Model Complexity

Methods	Params(M)	GFLOPs
MGAN	133	15.5
DVRNet+	26	3.80

DVRNet+ has lower model complexity than MGAN.

Qualitative Results

MR^{-2} on the EDGE20

Methods	Day	Night
Mod-Bi-box	23.2	100
DVRNet+	18.7	85.8

Statistical Results

P-value	Day	Night
Mod-Bi-box DVRNet$^+$	$2.287e^{-32}$	$7.419e^{-110}$

F test

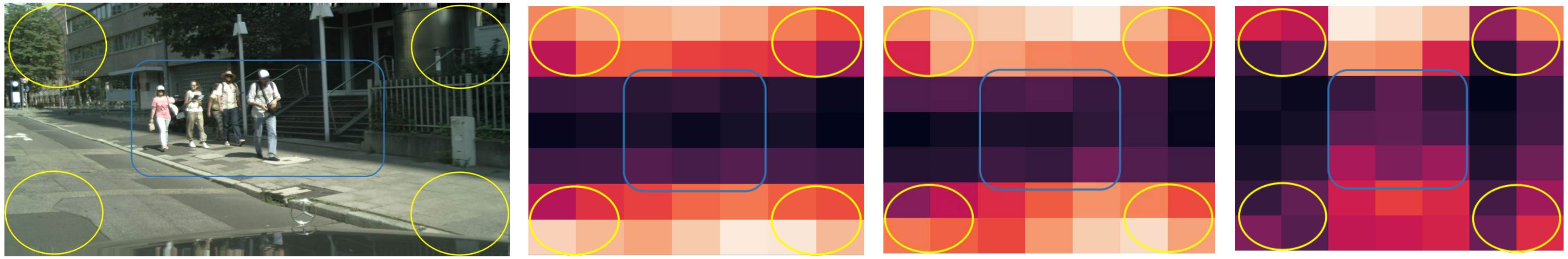
Statistical Results

P-value	Day	Night
Mod-Bi-box DVRNet^{+}	$2.287e^{-32}$	$7.419e^{-110}$

F test

DVRNet+ improves the baseline statistically significantly.

Heatmaps



(a) (b) (c) (d)

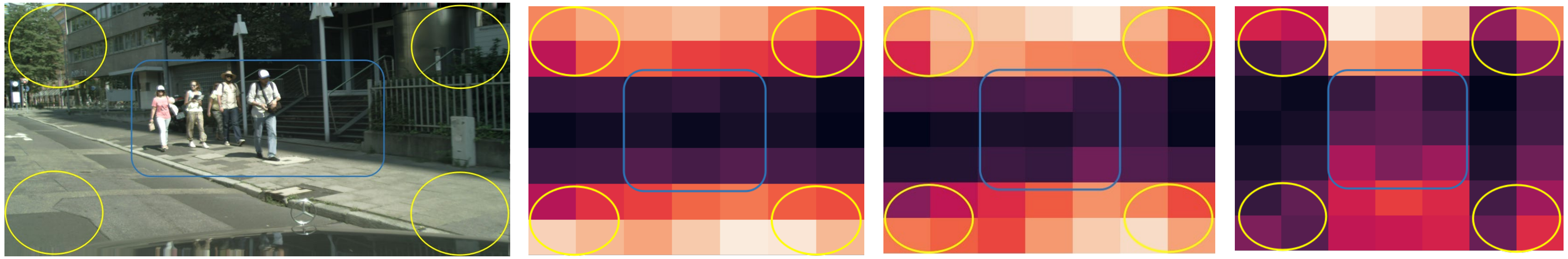
Depiction of an input image and the heatmaps of input RoI features and fused attention-based RoI features in the AFIM.

Human features

Background features

(a) The input image

Heatmaps

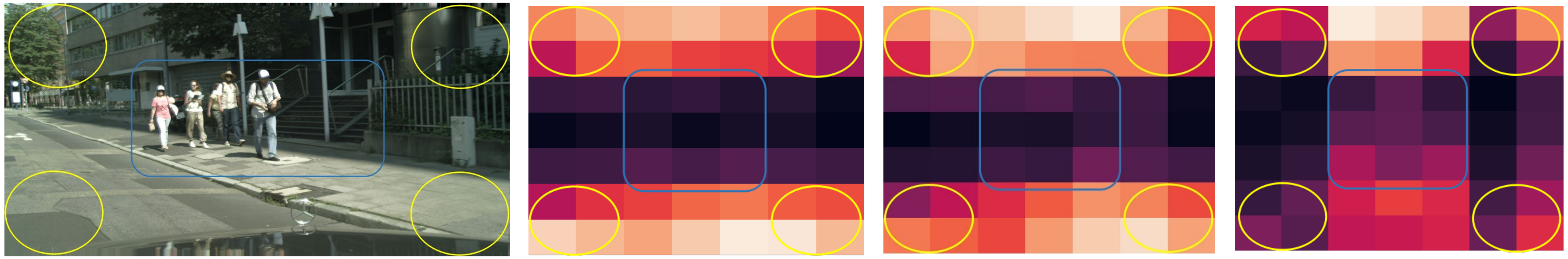


(a) (b) (c) (d)

Depiction of an input image and the heatmaps of input features and fused attention-based RoI features in the AFIM.

(a) The input image
(b) The heatmap of the input features used for predicting the full-body

Heatmaps

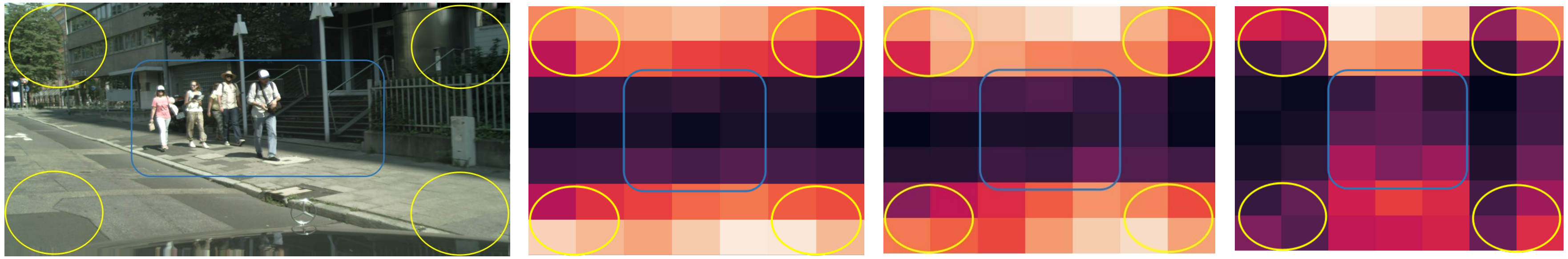


(a) (b) (c) (d)

Depiction of an input image and the heatmaps of input RoI features and fused attention-based RoI features in the AFIM.

(a) The input image
(b) The heatmap of the input features used for predicting the full-body
(c) The heatmap of the input features used for predicting the visible-body

Heatmaps

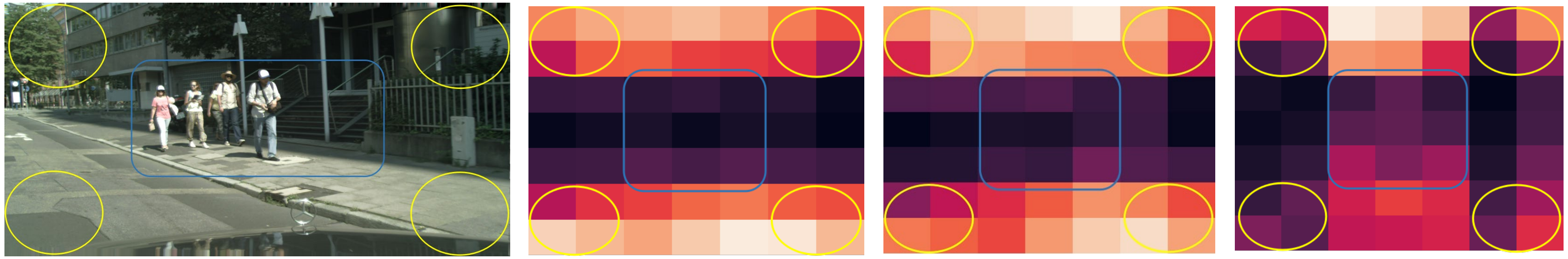


(a) (b) (c) (d)

Depiction of an input image and the heatmaps of input RoI features and fused attention-based RoI features in the AFIM.

(a) The input image
(b) The heatmap of the input features used for predicting the full-body
(c) The heatmap of the input features used for predicting the visible-body
(d) The heatmap of the fused attention-based features obtained by AFIM

Heatmaps

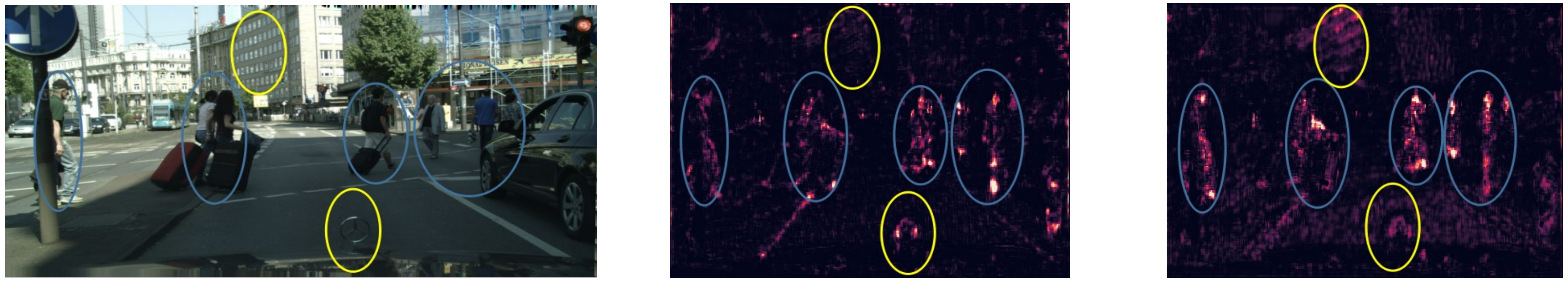


(a) (b) (c) (d)

Depiction of an input image and the heatmaps of input RoI features and fused attention-based RoI features in the AFIM.
(a) The input image
(b) The heatmap of the input RoI features used for predicting the full-body
(c) The heatmap of the input RoI features used for predicting the visible-body
(d) The heatmap of the fused attention-based RoI features obtained by AFIM

The AFIM increases the contrast of the human features and background features.

Heatmaps



(a) (b) (c)

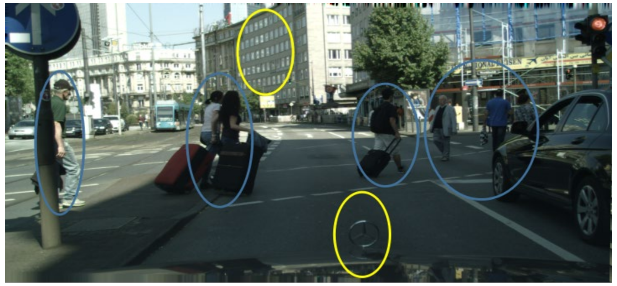
Depiction of an input image and the corresponding heatmaps of RPN feature maps, which are learned by RPN with and without BMLM.

Human features

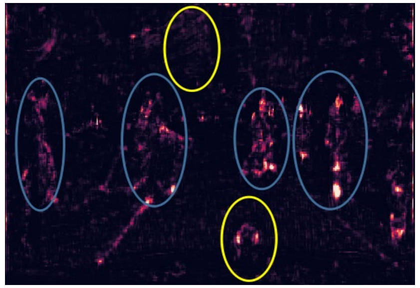
Background features

(a) The input image

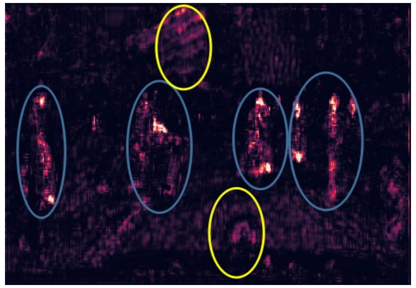
Heatmaps



(a)



(b)

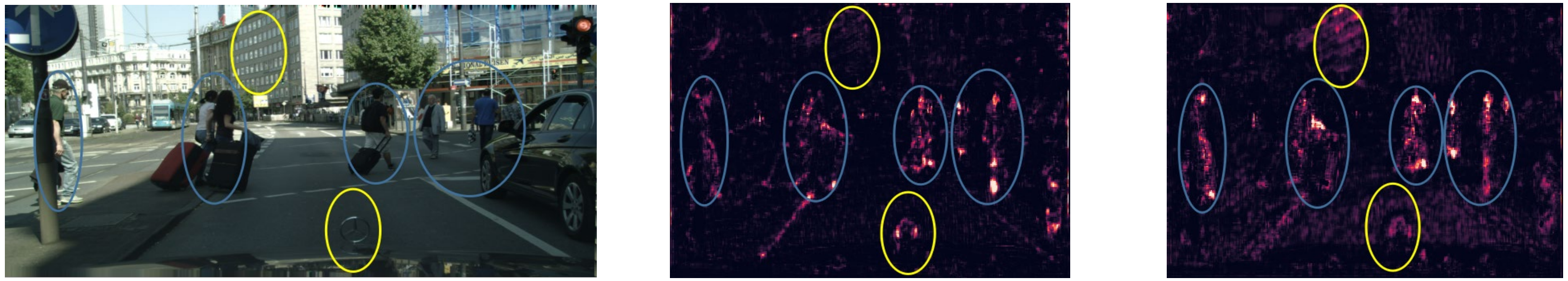


(c)

Depiction of an input image and the corresponding heatmaps of RPN feature maps, which are learned by RPN with and without BMLM.

(a) The input image
(b) The heatmap of the RPN features without BMLM

Heatmaps

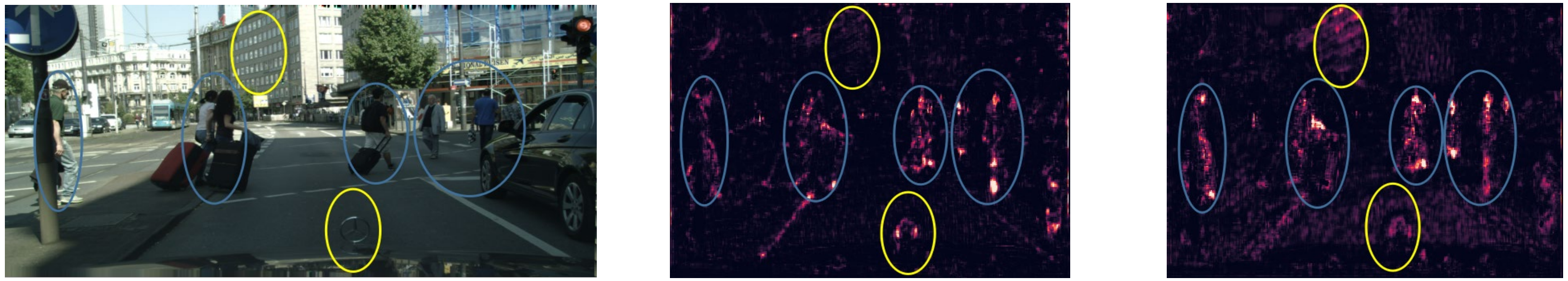


(a) (b) (c)

Depiction of an input image and the corresponding heatmaps of RPN feature maps, which are learned by RPN with and without BMLM

(a) The input image
(b) The heatmap of the RPN features without BMLM
(c) The heatmap of the RPN features with BMLM

Heatmaps



(a) (b) (c)

Depiction of an input image and the corresponding heatmaps of RPN feature maps, which are learned by RPN with and without BMLM.

(a) The input image
(b) The heatmap of the RPN features without BMLM
(c) The heatmap of the RPN features with BMLM

The BMLM increases the contrast of the human features and background features.

Heatmaps

(a) (b) (c)

Depiction of an input image and the heatmaps of RPN feature learned with and without head supervision signal.

Human features

Background features

(a) The input image.

Heatmaps

(a) (b) (c)

Depiction of an input image and the heatmaps of RPN feature learned with and without head supervision signal.

(a) The input image.
(b) The heatmap of the RPN features without head supervision signal.

Heatmaps

(a) (b) (c)

Depiction of an input image and the heatmaps of RPN feature learned with and without head supervision signal.

(a) The input image.
(b) The heatmap of the RPN features without head supervision signal.
(c) The heatmap of the RPN features with the head supervision signal.

Heatmaps

(a) (b) (c)

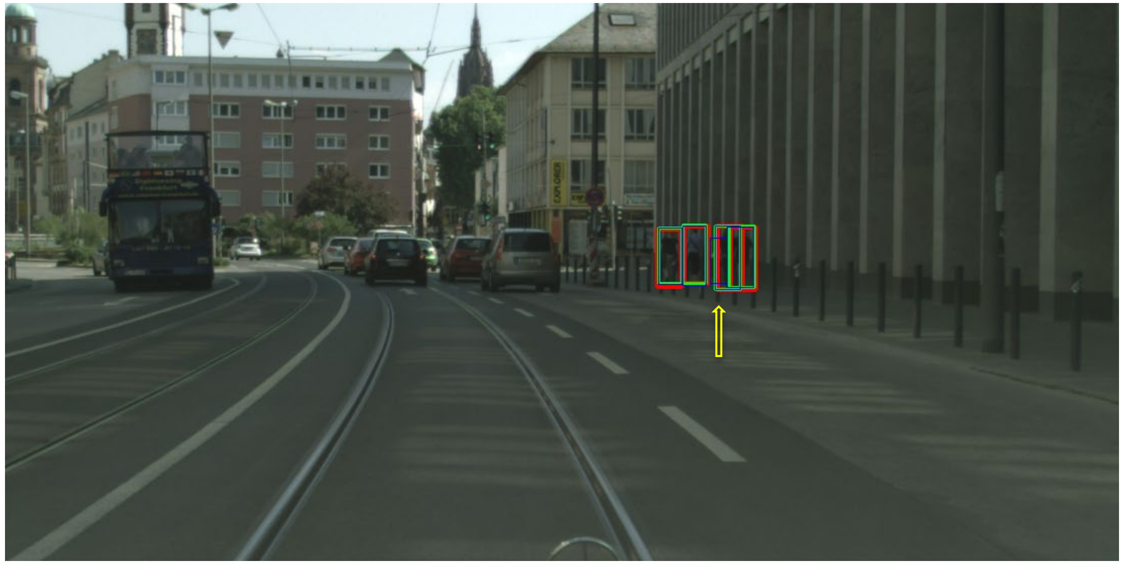
Depiction of an input image and the heatmaps of RPN feature learned with and without head supervision signal.

(a) The input image.
(b) The heatmap of the RPN features without head supervision signal.
(c) The heatmap of the RPN features with the head supervision signal.

Head supervision signal is more powerful than visible-body and full-body supervision signals.

✓ Head could provide more stable and discriminative information than visible-body.

Qualitative Results (Good)

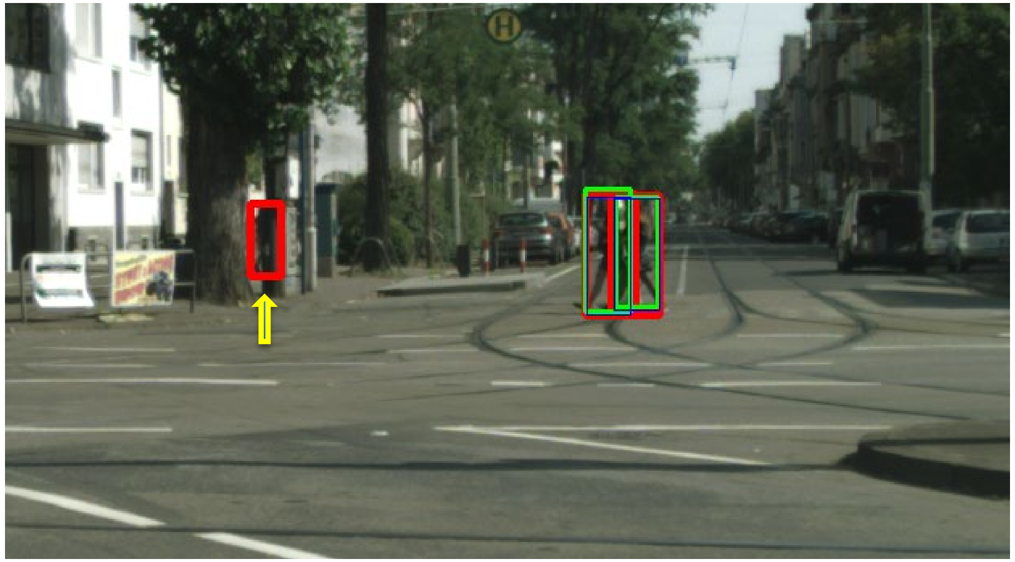


Ground truth

DVRNet

DVRNet+

Qualitative Results (Bad)

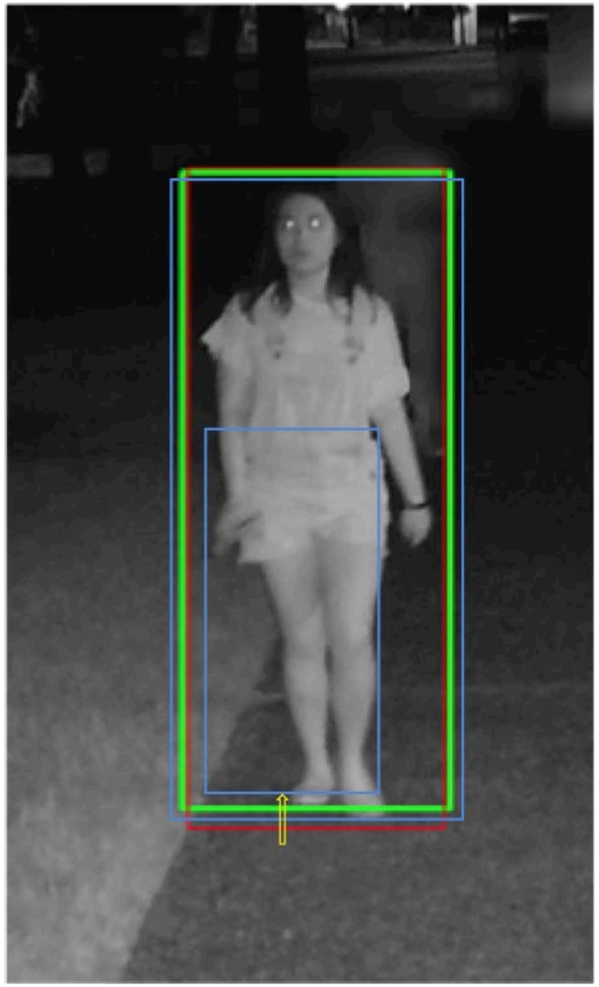


Ground truth

DVRNet

DVRNet+

Qualitative Results (Good)

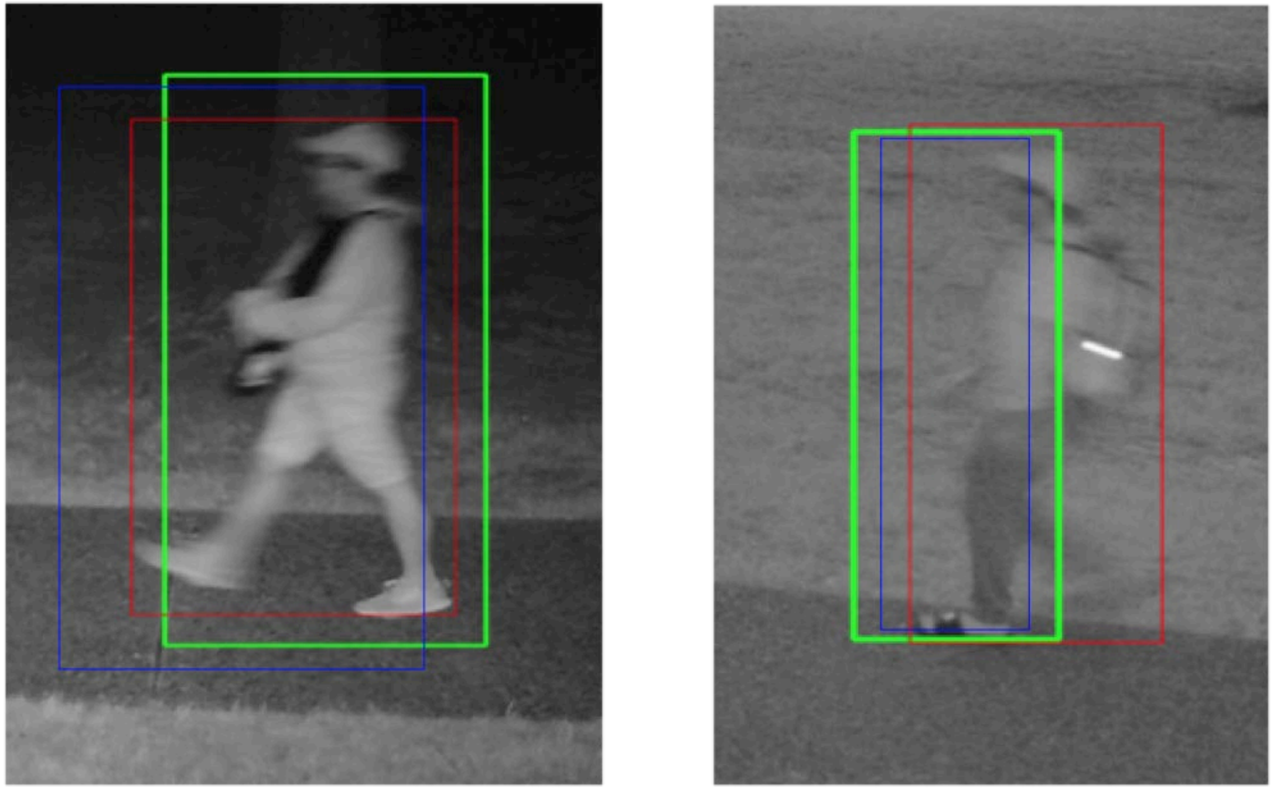


Ground truth

DVRNet

DVRNet+

Qualitative Results (Bad)



Ground truth

DVRNet

DVRNet+

Key Things To Remember

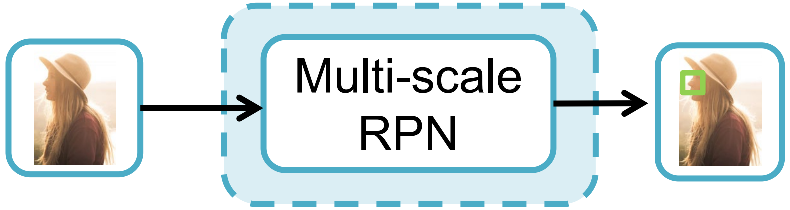
Human-body detection

- Hierarchical relationship inference: head-> visible-body-> full-body
- Discriminative human features
- Understand how the network learns features from visible and near-infrared images.

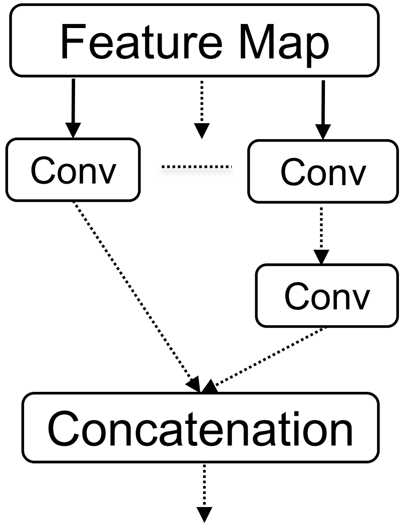
Objective 2

Design, develop, and evaluate a single stage face detector for 2D images to overcome scale challenge in the wild.

Obj. 2: Related Work



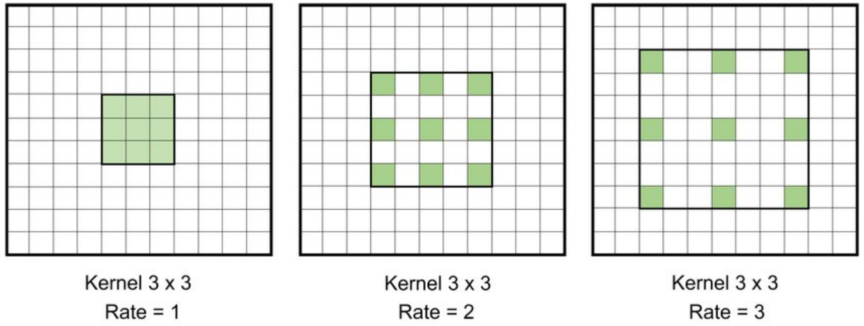
single-stage Detector

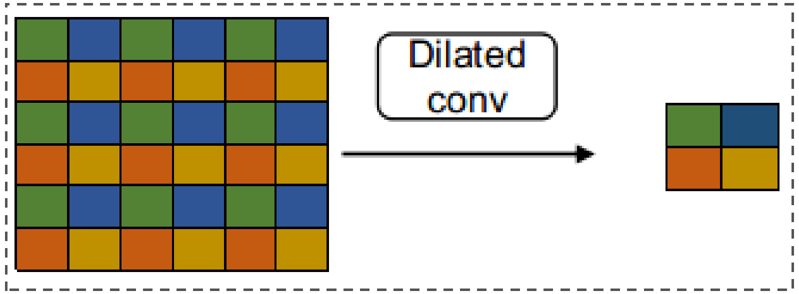


Context Aggregation

Obj. 2: Related Work Limitations

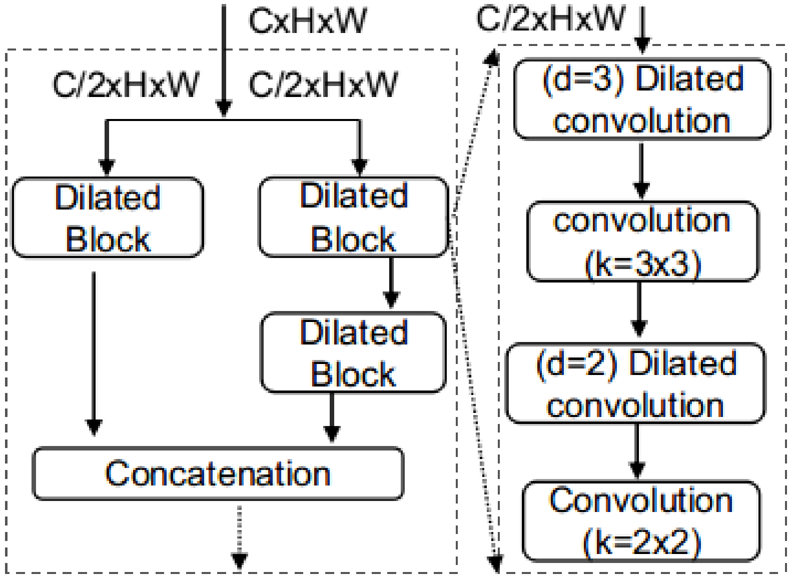
- Gridding artifacts problem





The DConv has kernel size of 3x3, strides of 1, and dilation rate of 2. The green pixels in the right feature map are obtained by nine green pixels in the left feature map. The pixels with other different colors share the same idea. Therefore, neighboring four pixels in the right feature map are obtained by completely separate four sets of units in the left feature map.

Smoothed Context Enhancement Module



Baselines

Paper Source	Abbreviation
Ren et al. NeurIPS 2015	Faster R-CNN
Zhang *et al*. ICCV 2017	S3FD
Peng *et al*. CVPR 2017	HR
Najibi *et al*. CVPR 2017	SSH

Qualitative Results

mAP (%) on subsets of UFDD on each condition

Methods	Rain	Snow	Haze	Blur	Illumination	Lens impediments
Faster R-CNN	54.8	54.9	46.4	68.0	57.9	52.6
SSH	73.5	71.3	65.4	80.6	72.0	59.4
S3FD	75.9	72.3	71.9	83.8	78.0	60.7
HR-ER	75.9	74.3	72.5	84.4	77.2	68.5
SANet	78.7	77.2	75.3	87.8	82.7	69.4

Dataset

- Multi-modality (VIS-NIR) dataset for which the performance is not saturated

Dataset	Set	Images
EDGE20	Day	2,694
EDGE20	Night	797

Qualitative Results

mAP (%) on the EDGE20

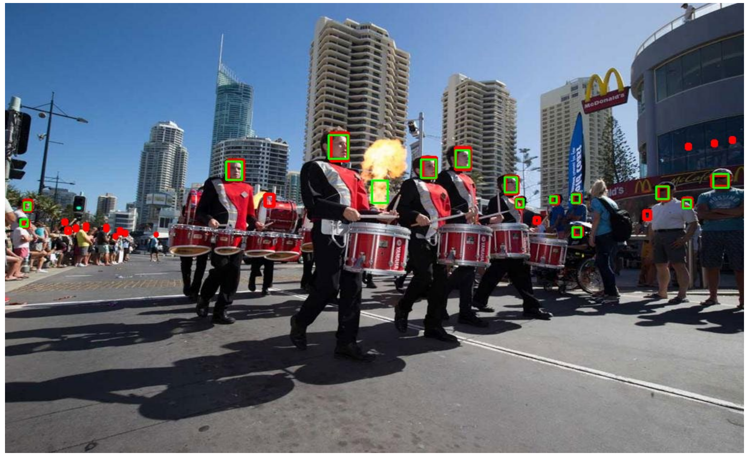
Method	Day	Night
SANet	85.5	22.0
S3FD	84.0	17.5

Qualitative Results (Good)

Ground truth

Our prediction

Qualitative Results (Bad)

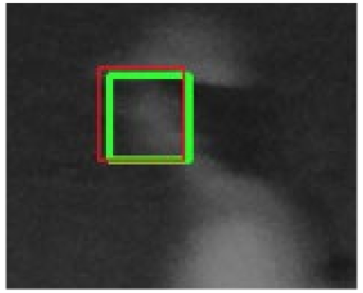


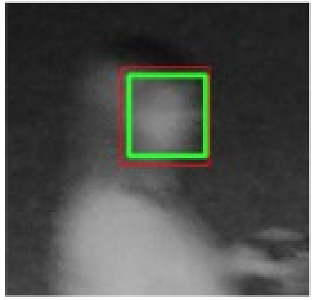
Ground truth

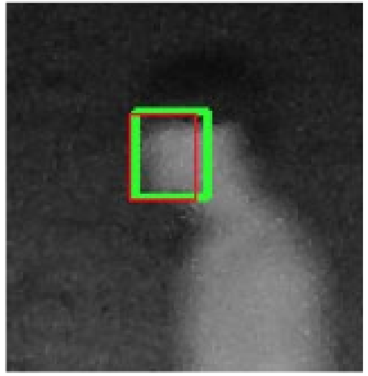
Our prediction

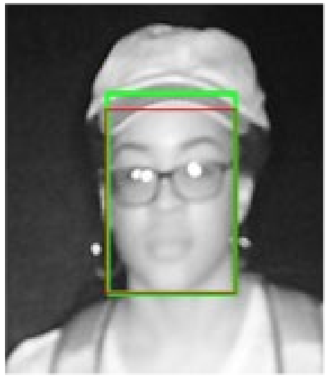
Qualitative Results (Good)

Qualitative Results (Good)







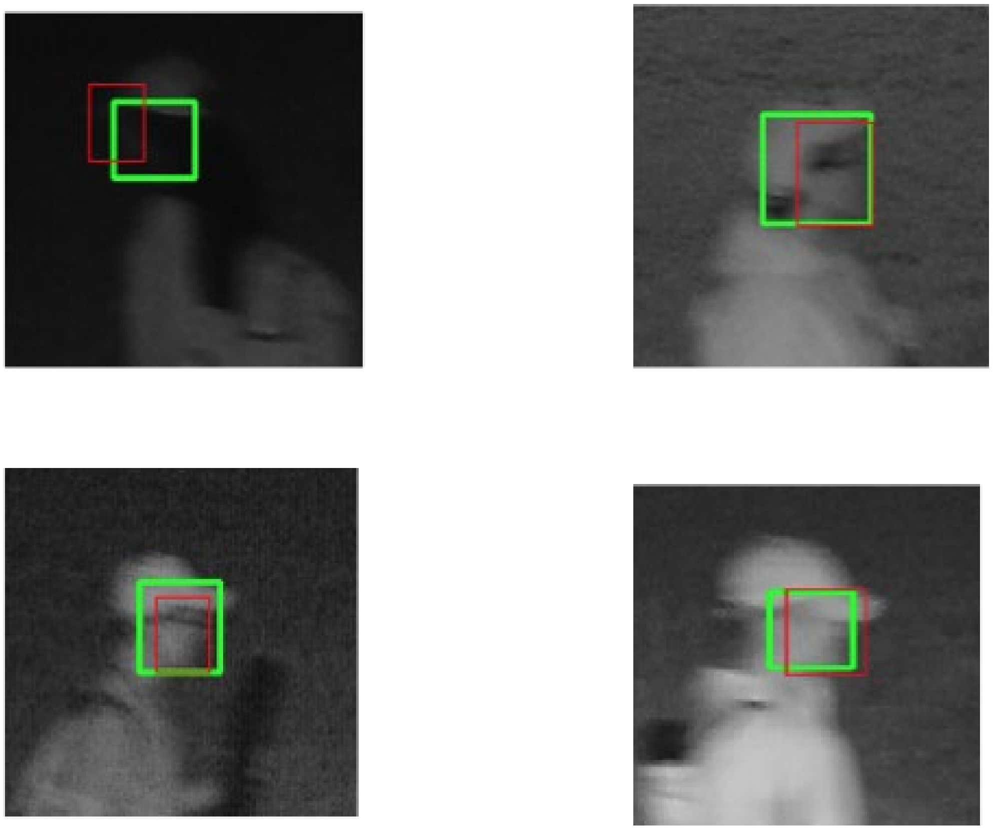


Ground truth

Our prediction

Qualitative Results (Bad)

Qualitative Results (Bad)



Ground truth

Our prediction

Conclusion

For human-body detection:

1. Compared to visible-body and full-body, the head provides more discriminative information to the network.
2. Feature interaction is an effective way of improving performance during training. For each iteration, the network could employ additional contextua information to learn discriminative features.
3. Pixel-wise classification task is a good complement of the region-wise classification task.

For face detection:

1. Larger receptive field size is more important than consistent local information for detecting multi-scale faces.
2. Contextual information is always an effective way of solving the scale problem.

Thank you!