IMAGENET
ILSVRC 2015 CLS-LOC.

Multi-Class AttentionNet.

D. Yoo[1], K. Paeng[1], S. Park[1], S. Hwang[2], H. E. Kim[2], J. Lee[2],
M. Jang[2], A. S. Paek[2], K. K. Kim[1], S. D. Kim[1], I. S. Kweon[1].
[1]KAIST, [2]Lunit Inc.

Lunit KAIST

State-of-the-art methods for object localization.

State-of-the-art methods for object localization.

1) Box-regression with a CNN.

[Szegedy et al., NIPS'13],

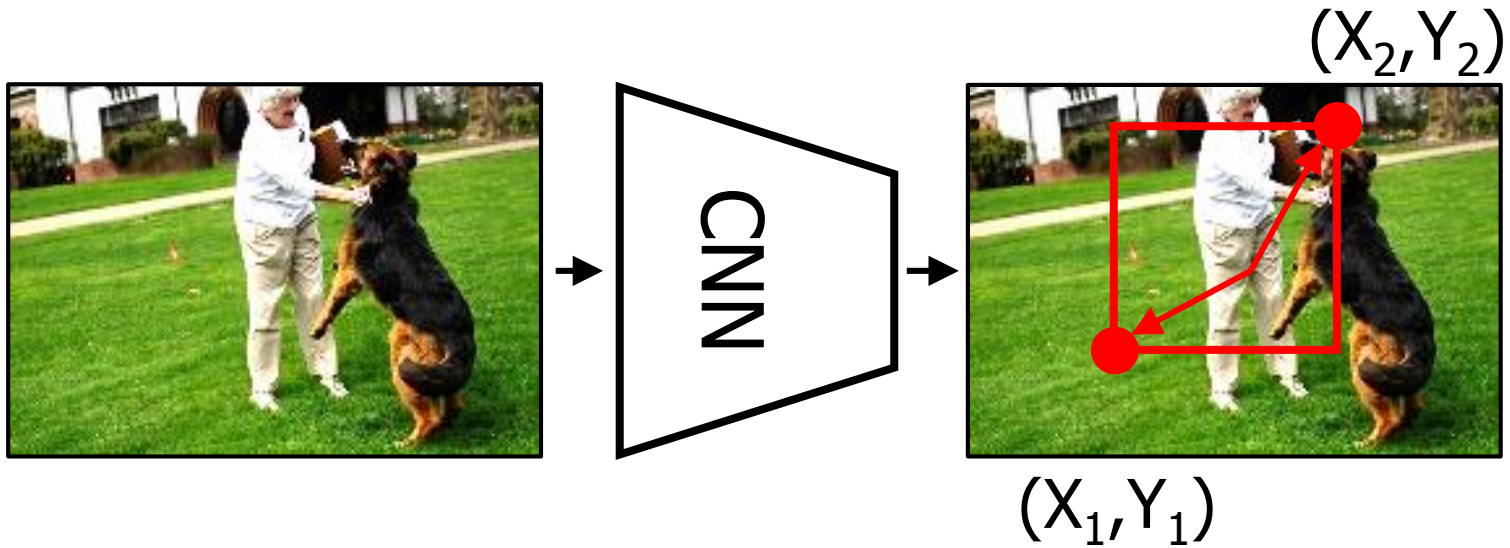
DeepMultiBox [Erhan et al., CVPR'14],
OverFeat [Sermanet et al., ICLR'14],

…

State-of-the-art methods for object localization.

1) Box-regression with a CNN.

(–) Direct mapping from an image to an exact bounding box is relatively difficult for a CNN.



State-of-the-art methods for object localization.

2) Region proposal + classifier.

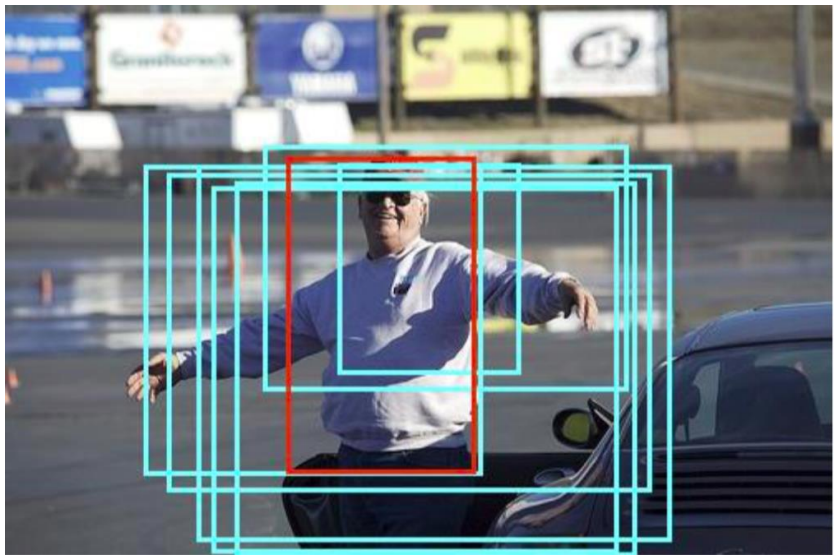
R-CNN [Gkioxari et al., CVPR'14],
Fast R-CNN [Gkioxari, ICCV'15],
Faster R-CNN [Ren et al., NIPS'15],
DeepMultiBox [Erhan et al., CVPR'14],

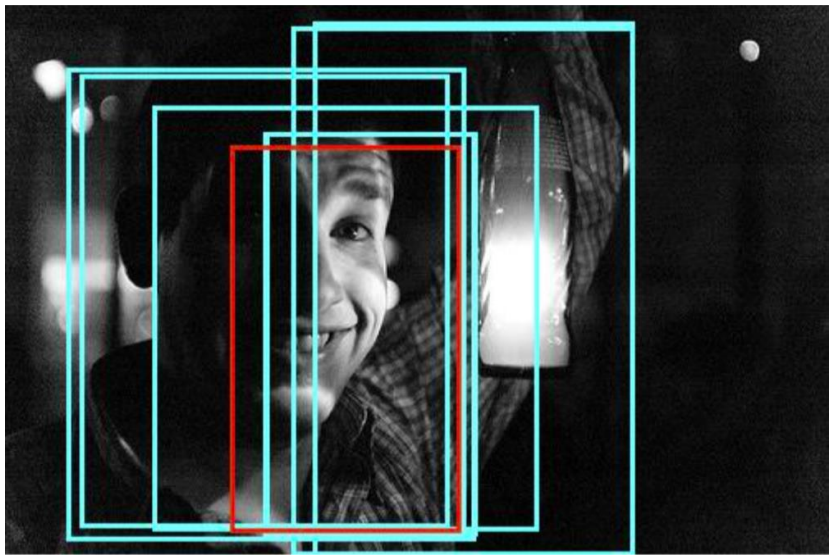
...

State-of-the-art methods for object localization.

2) Region proposal + classifier.

(–) Prone to focus on discriminative part (e.g. face) rather than entire object (e.g. human body).





Idea:
Ensemble of weak directions.

Idea:
Ensemble of weak directions.

Idea:
Ensemble of weak directions.

Idea:
Ensemble of weak directions.

Idea:
Ensemble of weak directions.

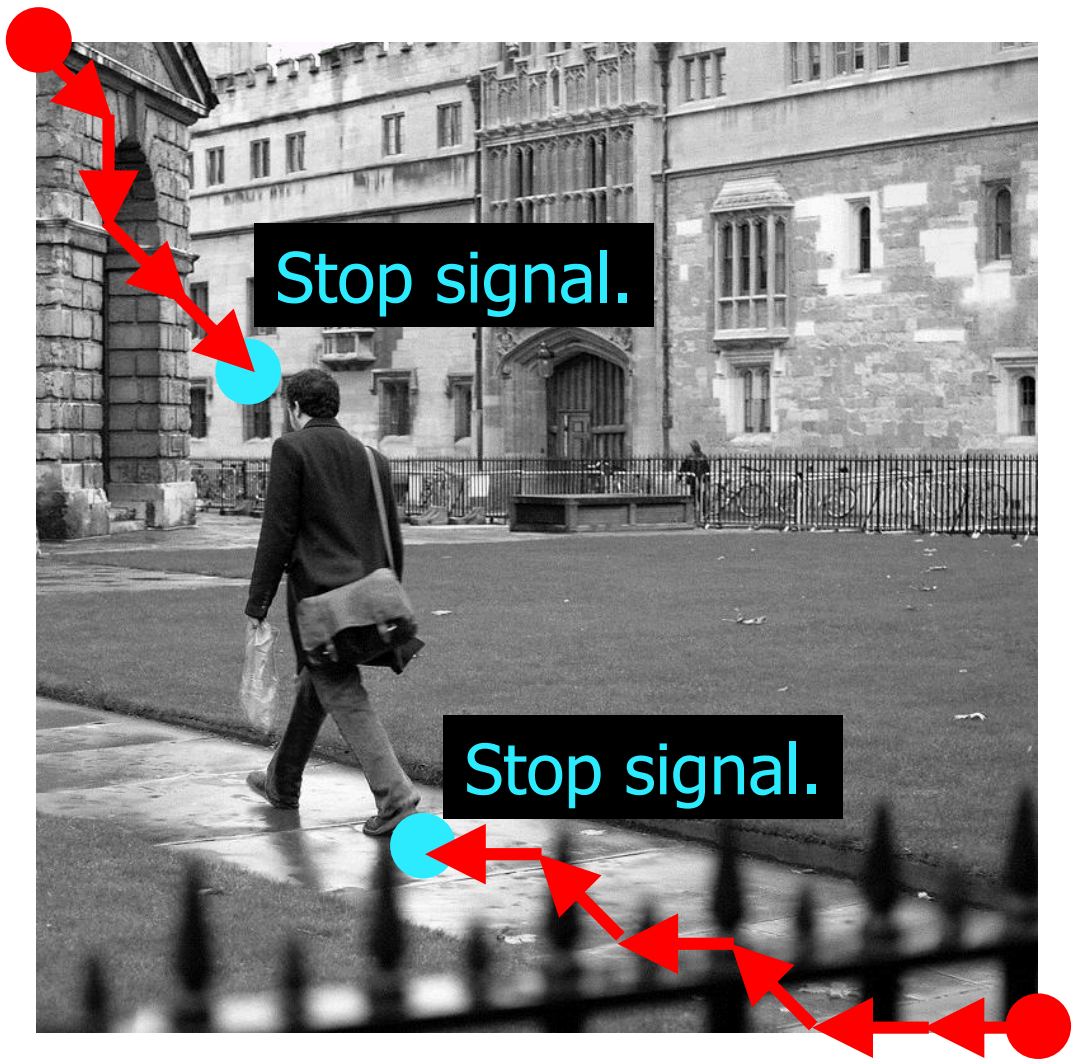
Idea:
Ensemble of weak directions.

Idea:
Ensemble of weak directions.

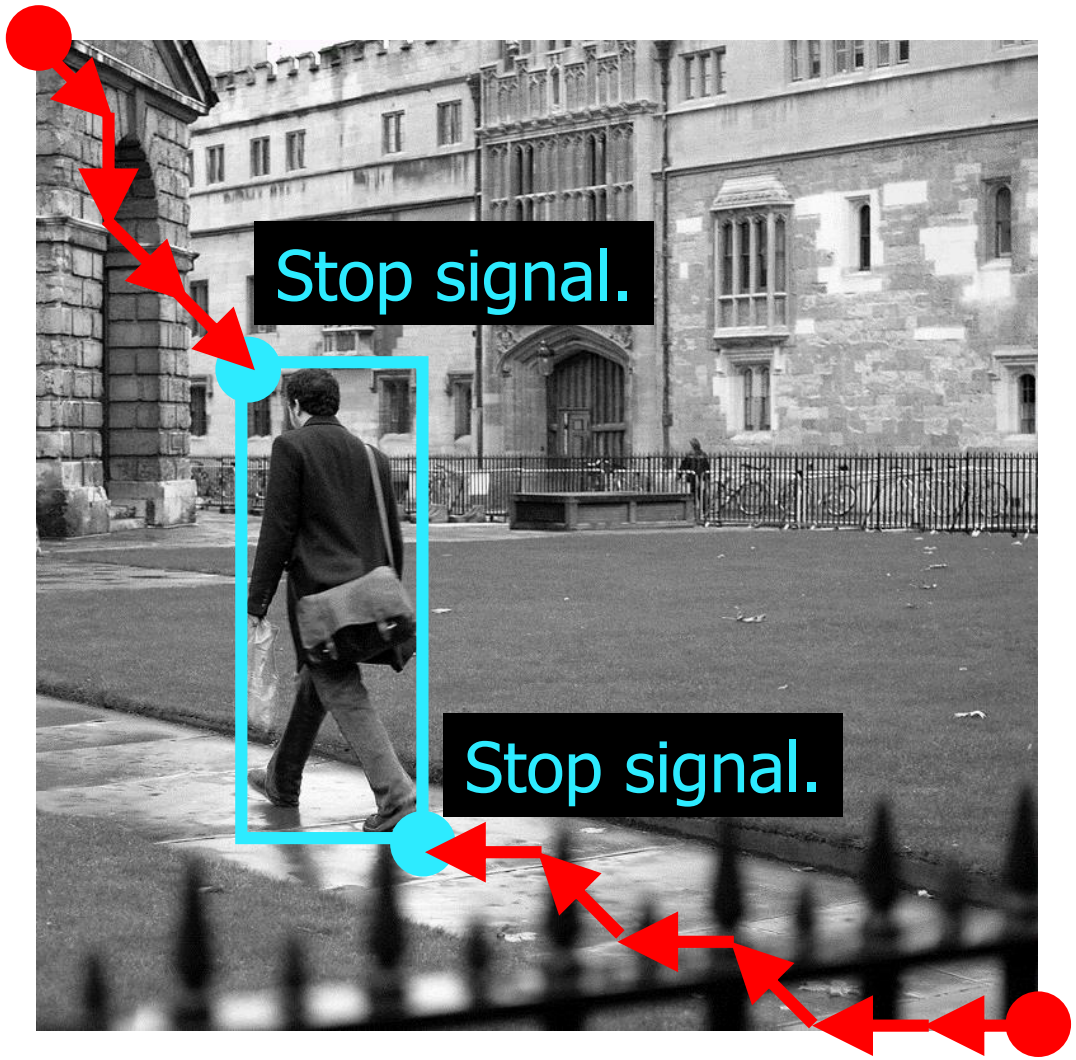
Idea:
Ensemble of weak directions.

Idea:
Ensemble of weak directions.

Idea:
Ensemble of weak directions.



Idea:
Ensemble of weak directions.



Model:

Model:

(CNN regression model)

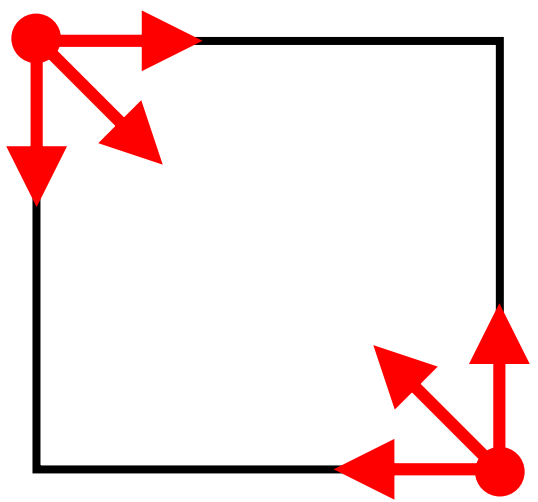
Model:

Rather than CNN regression model,

we use *CNN classification* model.

Model:
Rather than CNN regression model,
we use *CNN classification* model.

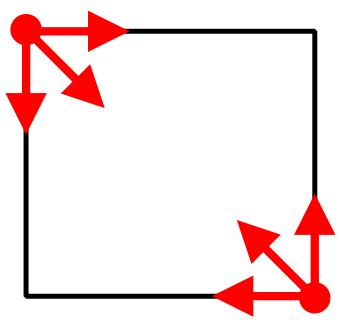
Define weak directions:
fixed length, and quantized.



Strength to the previous methods.

Box-regression:
(−) Relatively
difficult for a CNN.

Weak direction:
(+) Relatively
easy for a CNN.

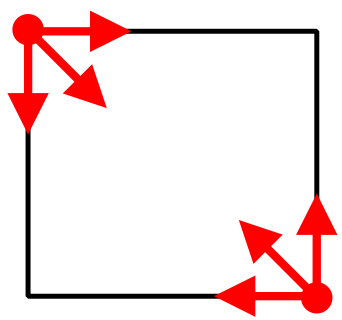


Strength to the previous methods.

Box-regression:
(−) Relatively difficult for a CNN.

↓

Weak direction:
(+) Relatively easy for a CNN.

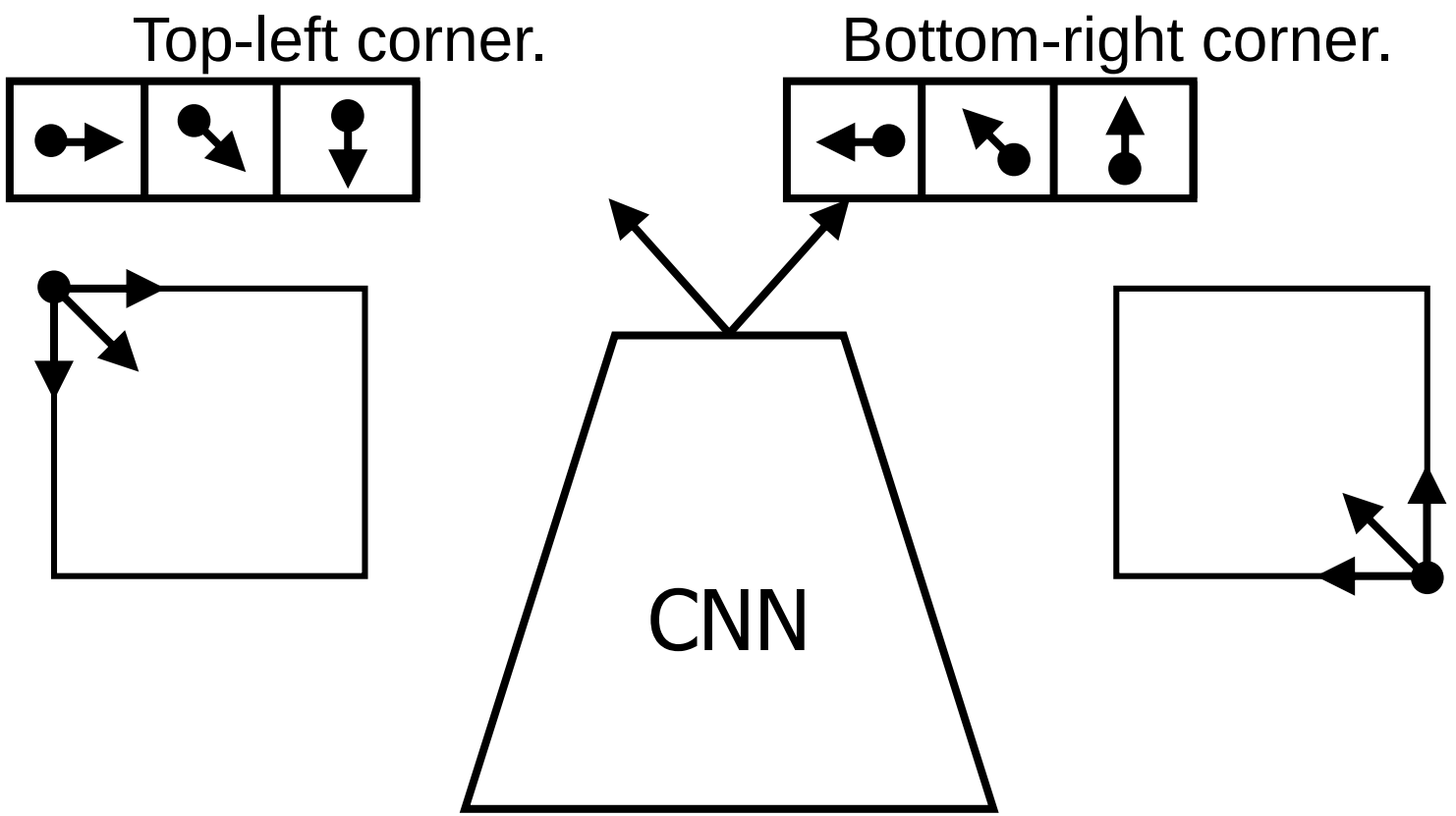


R-CNN:
(−) Focuses on distinctive parts.

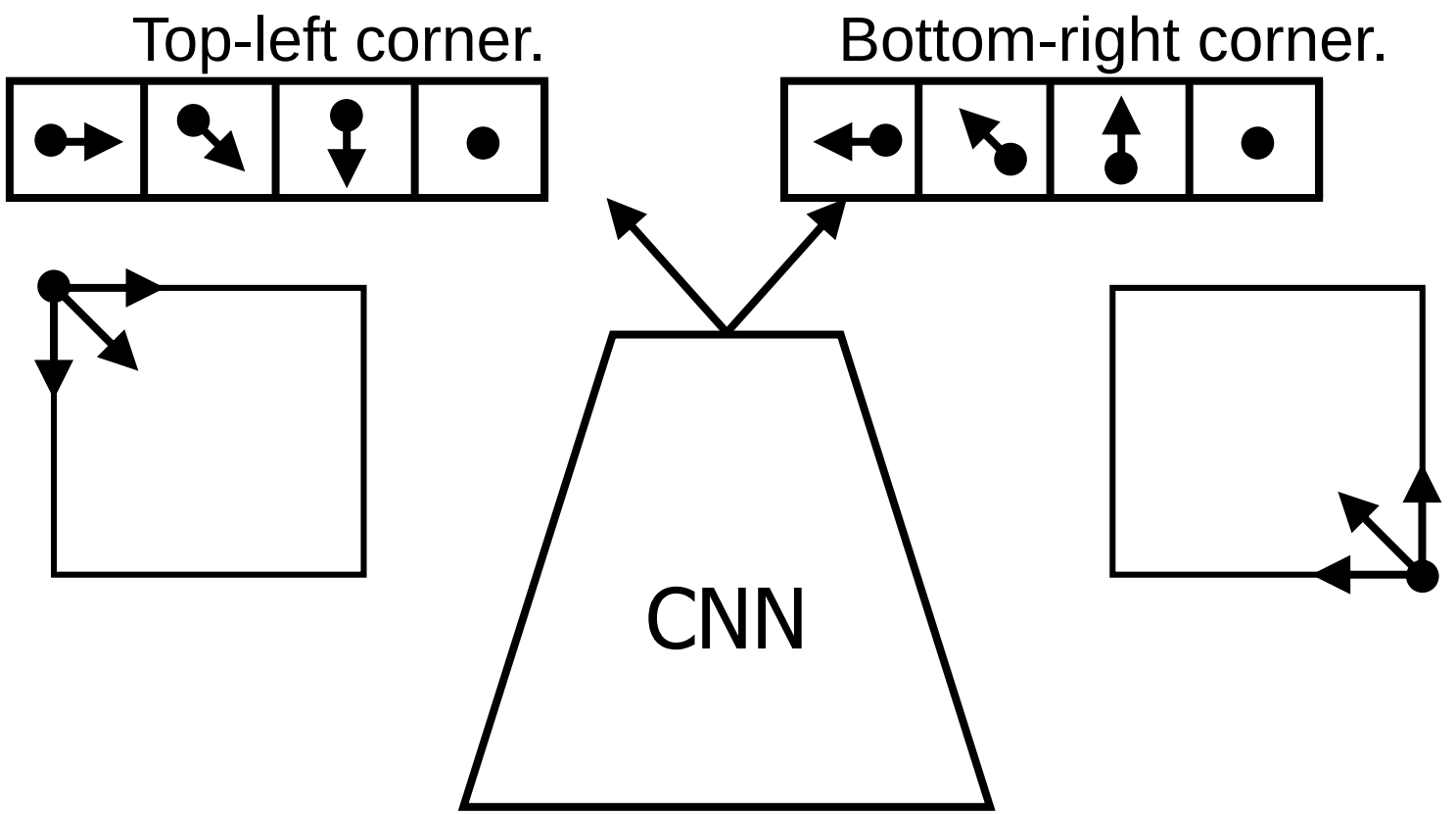
↓

Stop signal:
(+) Supervision of clear terminal point.

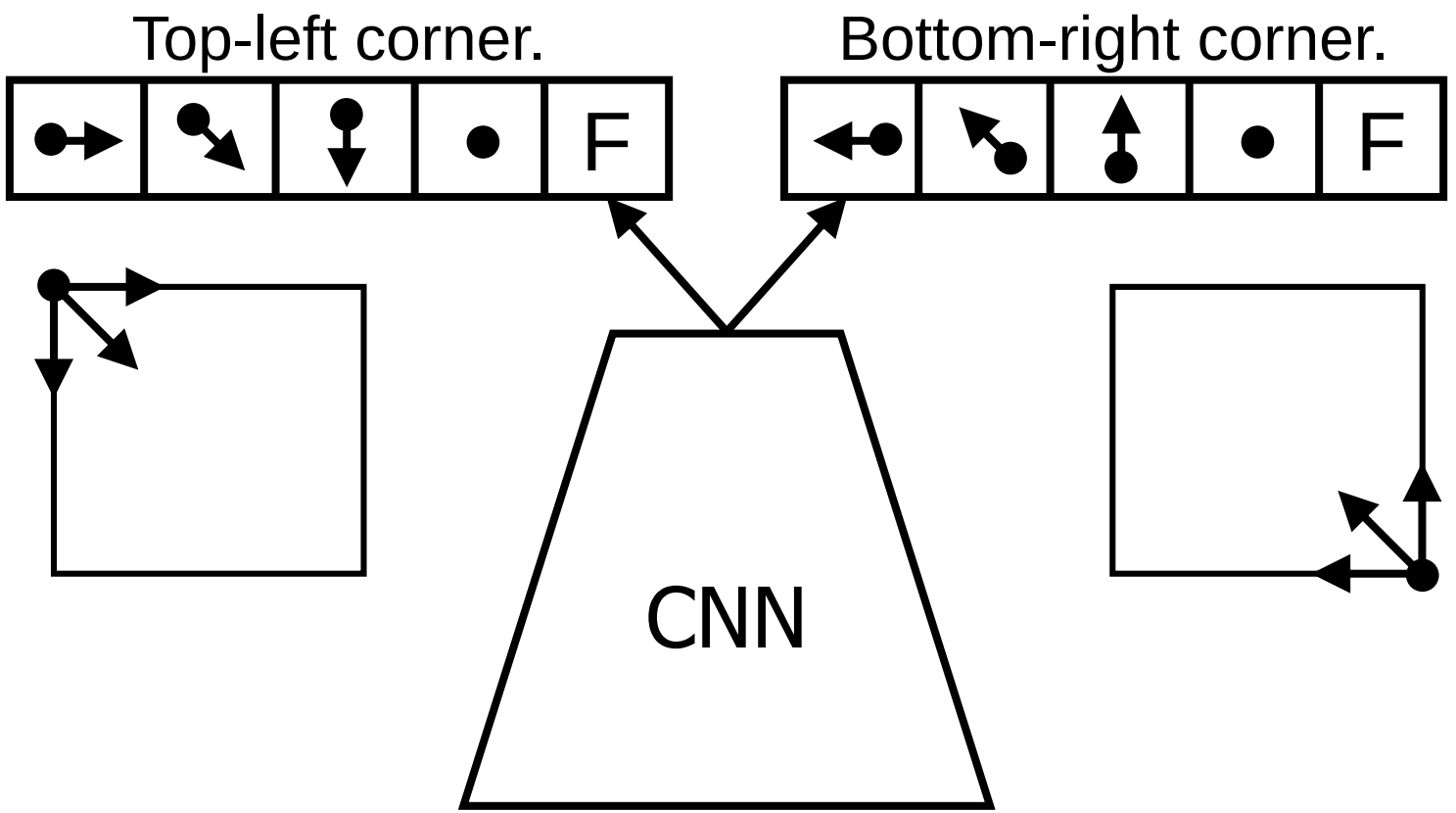
AttentionNet:
Two layers for each corner.



AttentionNet:
Two layers for each corner.

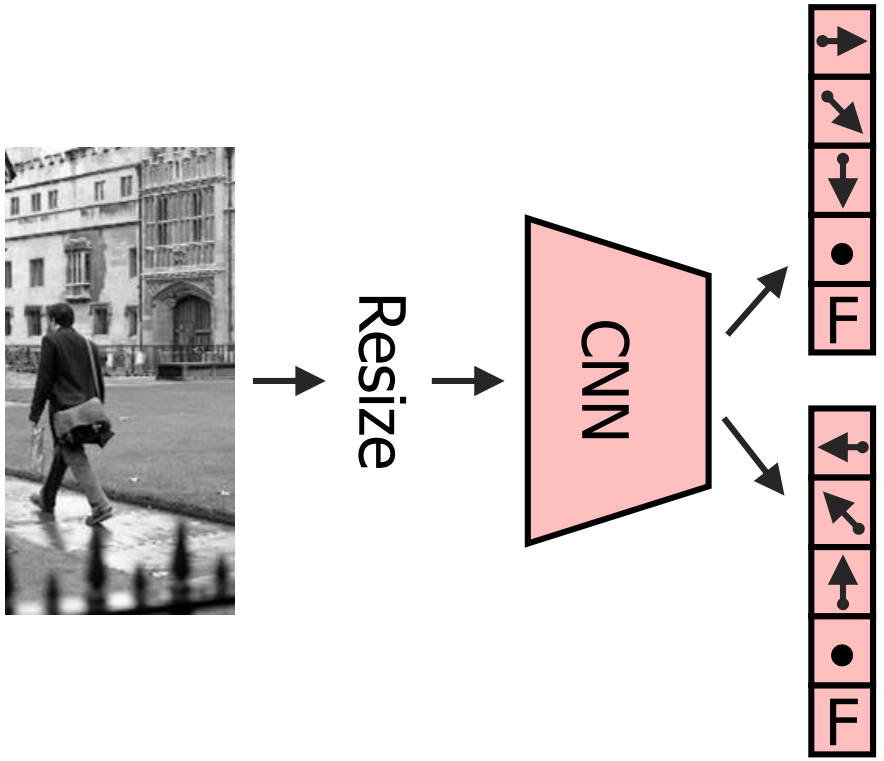


AttentionNet:
Two layers for each corner.



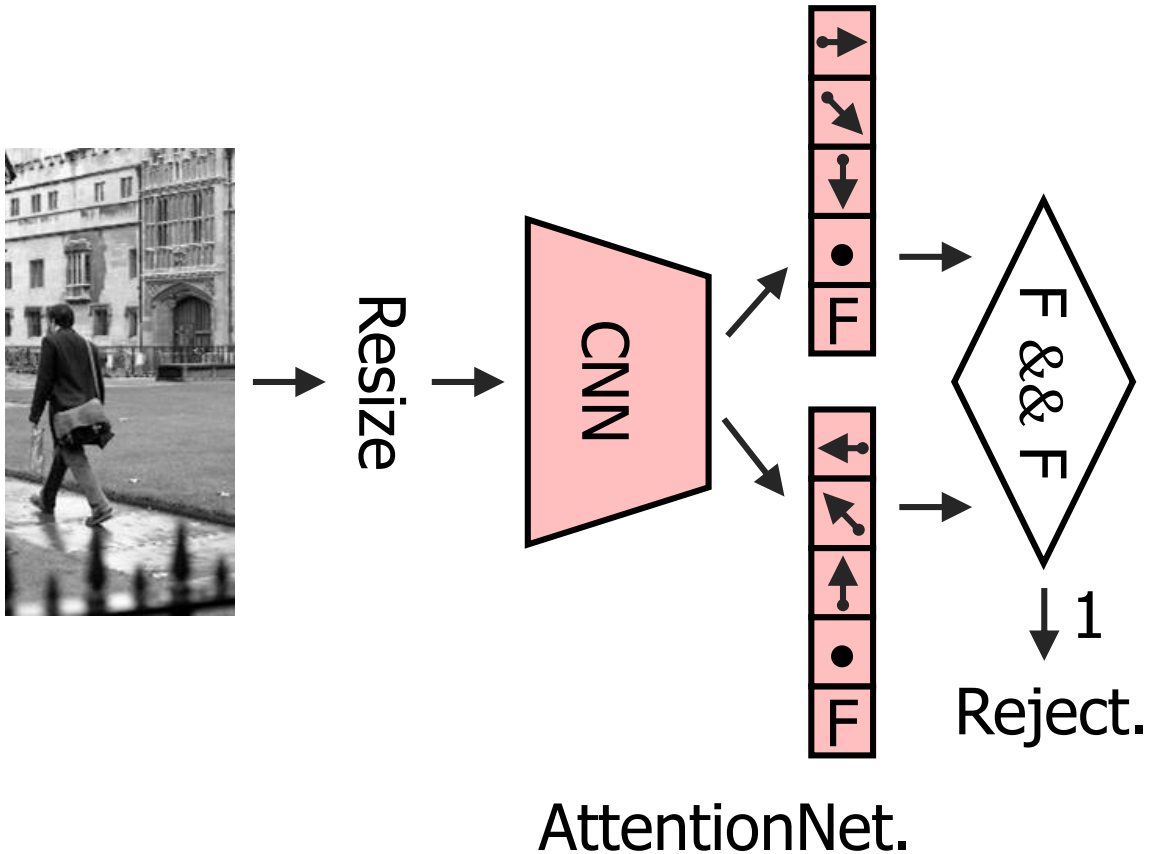
AttentionNet: iterative classification.

AttentionNet: iterative classification.



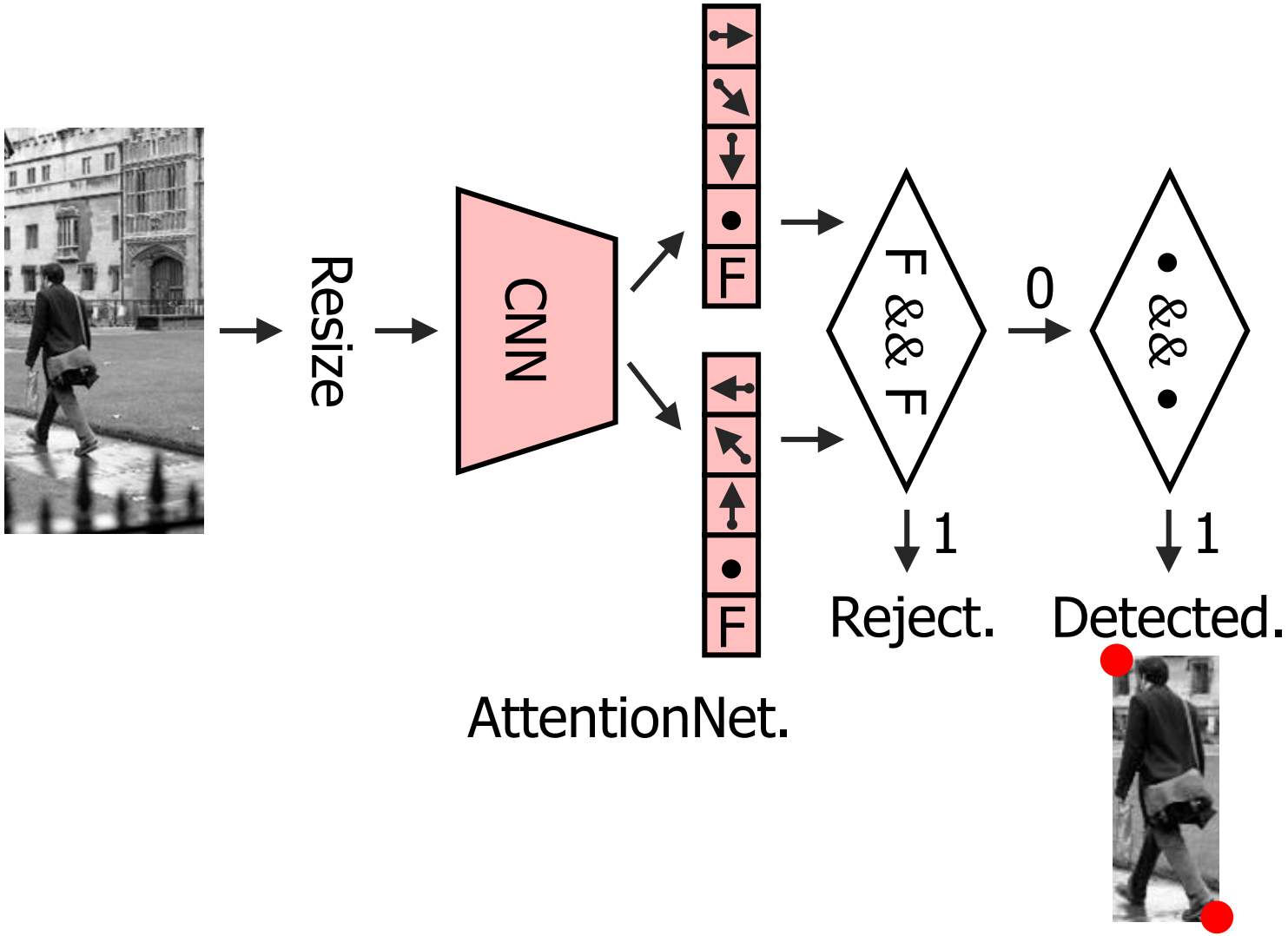
AttentionNet.

AttentionNet: iterative classification.

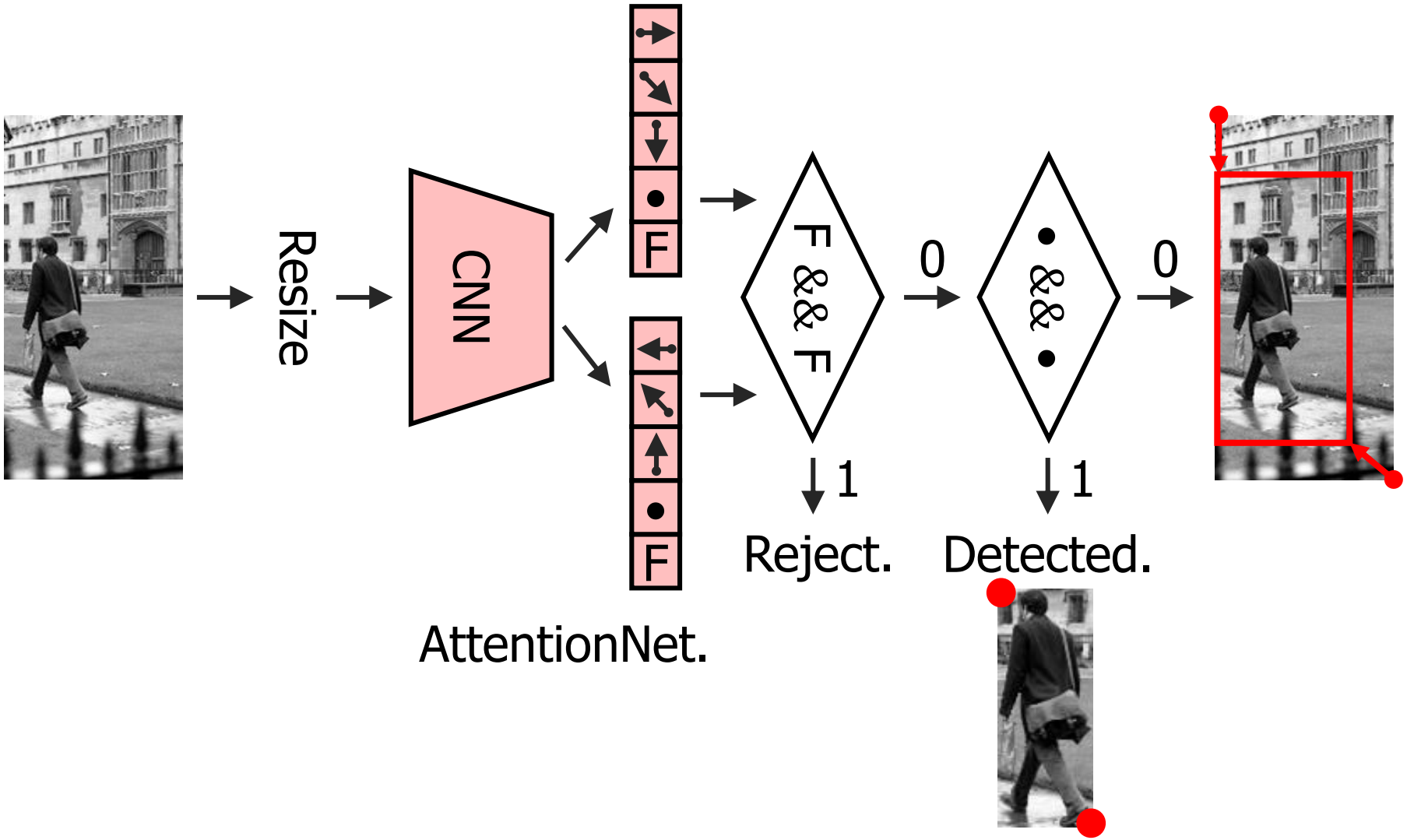


AttentionNet.

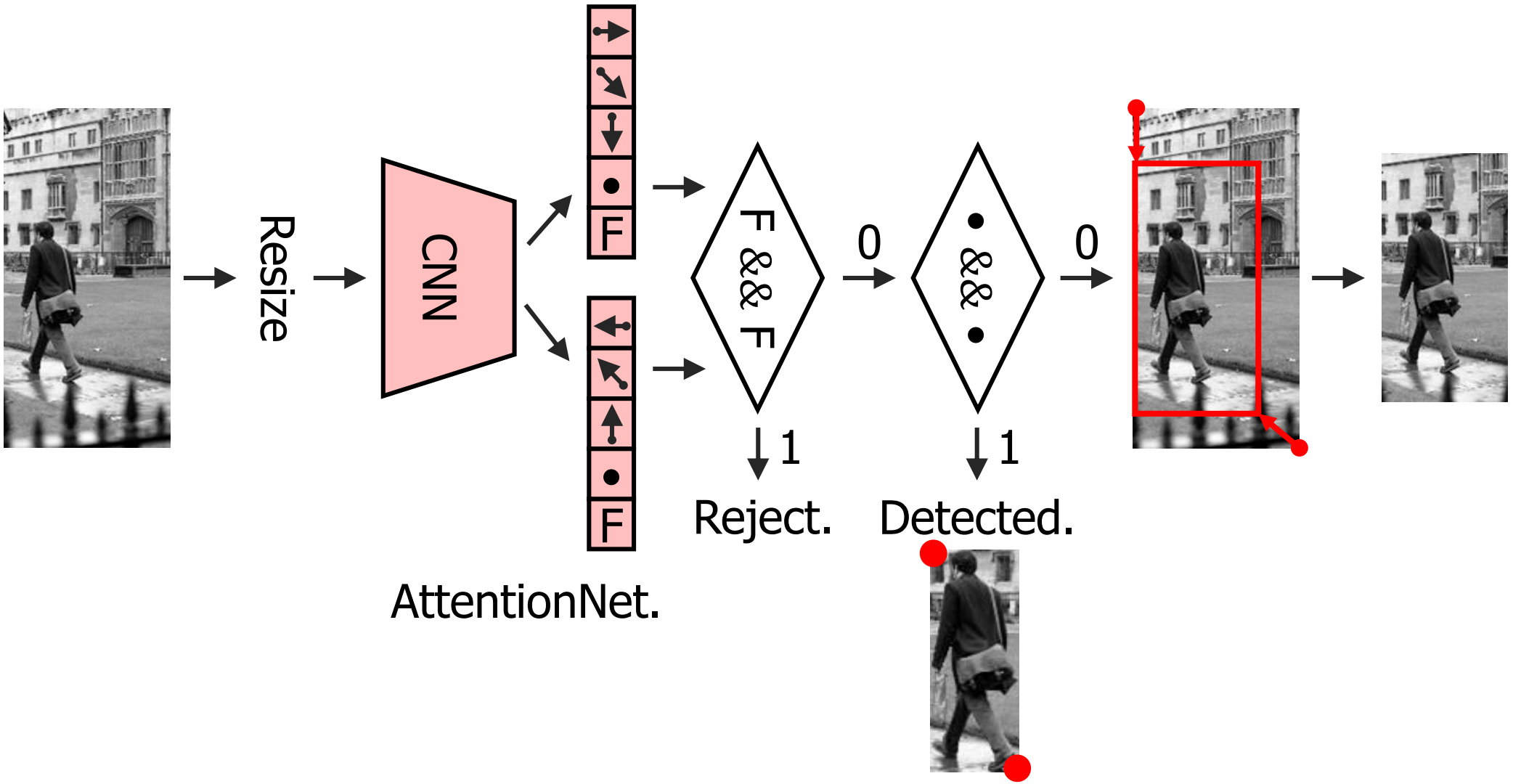
AttentionNet: iterative classification.



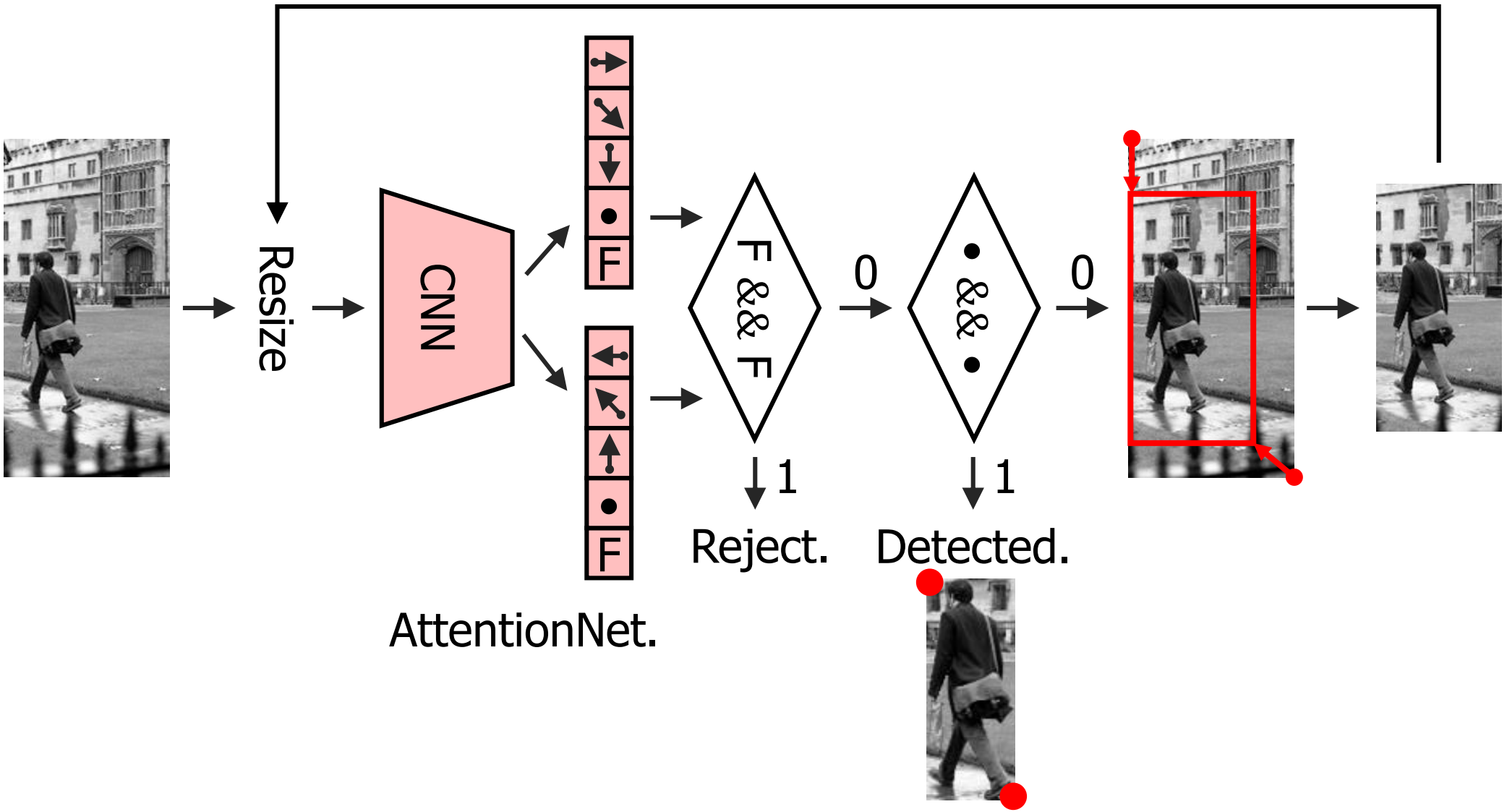
AttentionNet: iterative classification.



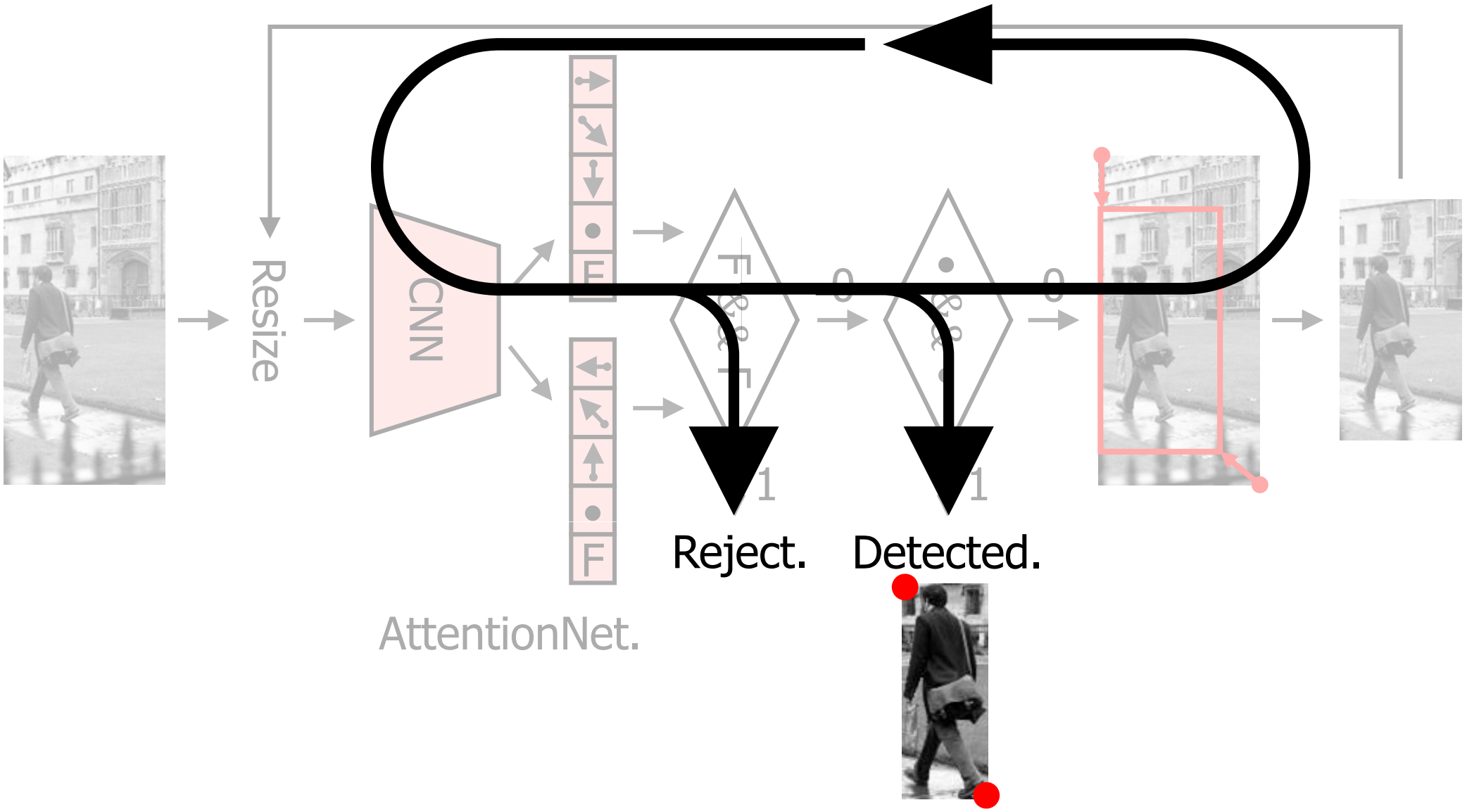
AttentionNet: iterative classification.

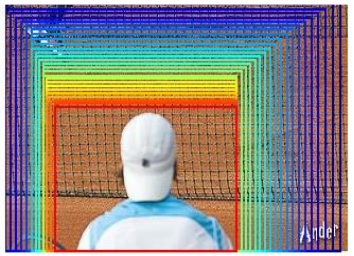
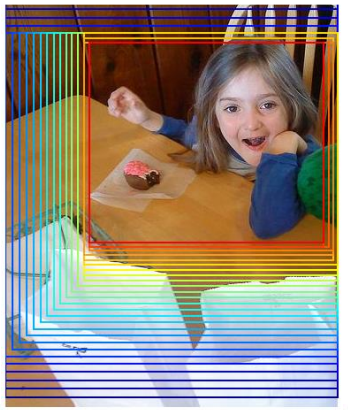


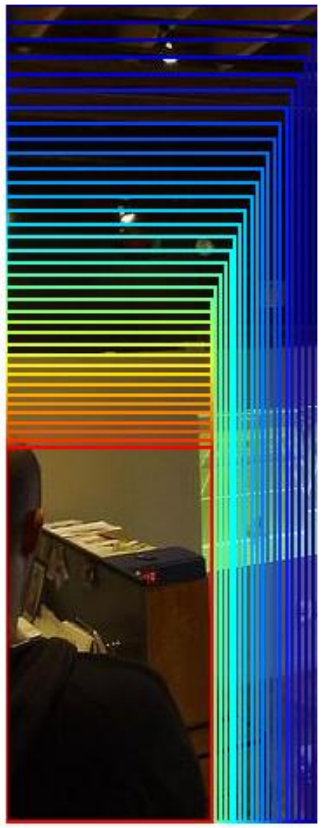
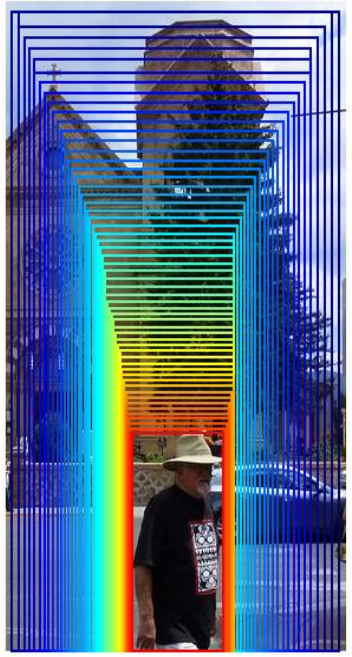
AttentionNet: iterative classification.

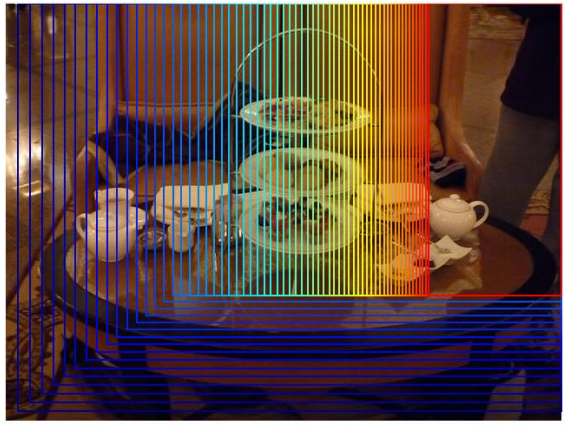
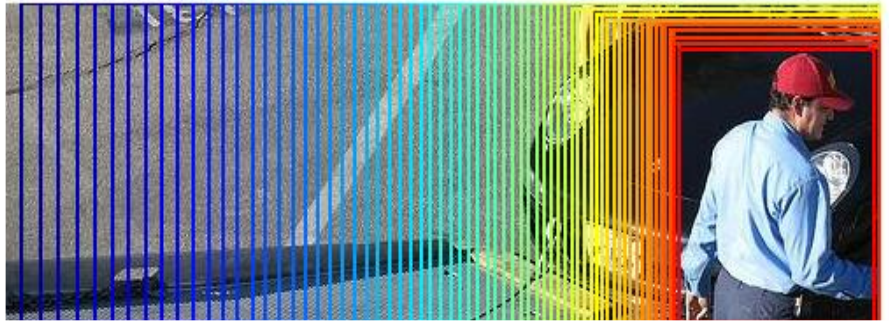
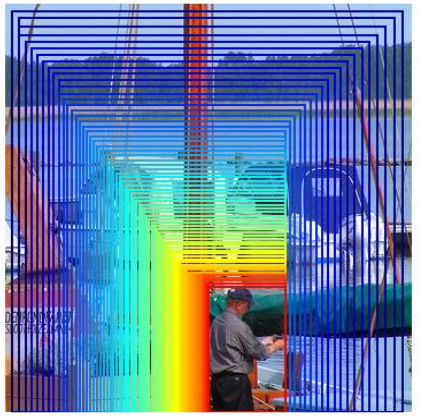


AttentionNet: iterative classification.



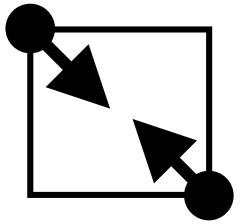





AttentionNet: Aggregating Weak Directions for Accurate Object Detection

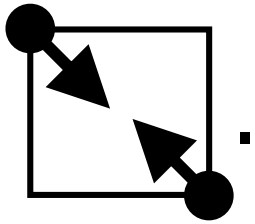
Human detection examples on PASCAL VOC 2007

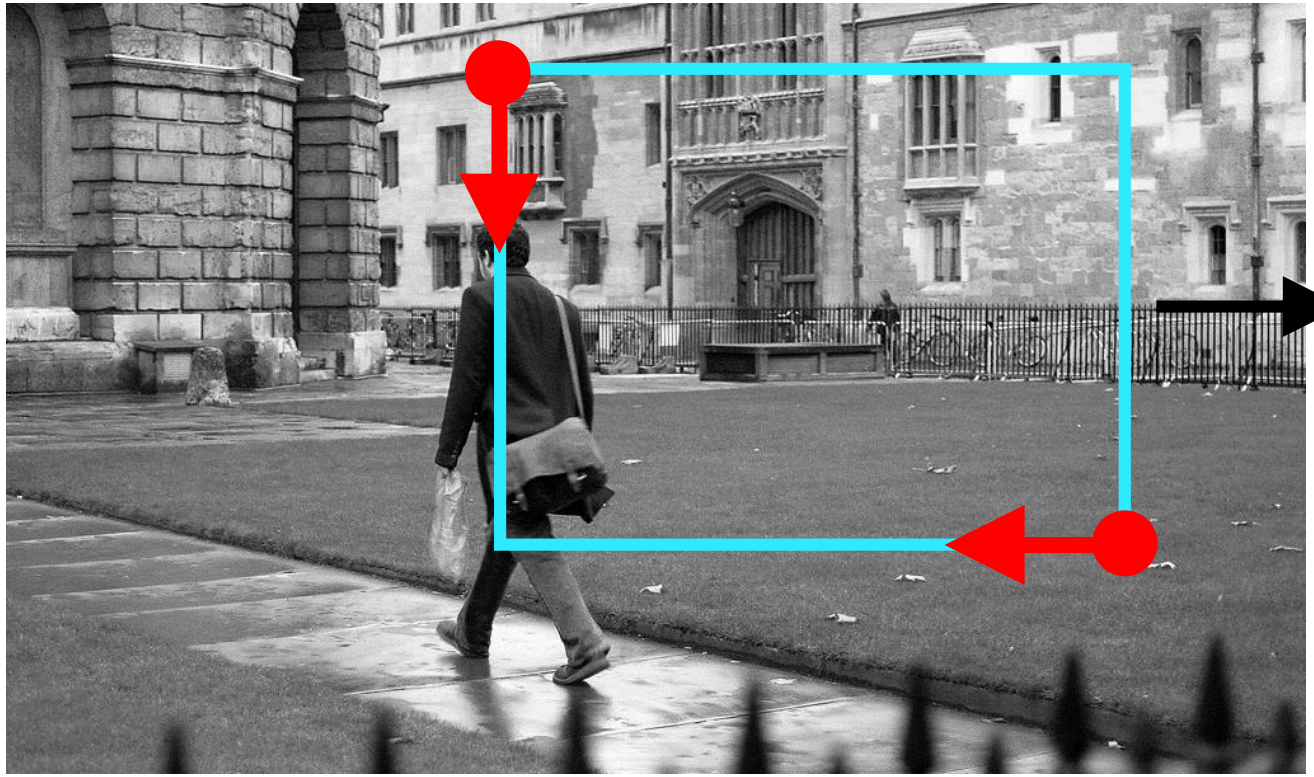
Initial box proposal:

Initial box proposal:

Boxes satisfying  .

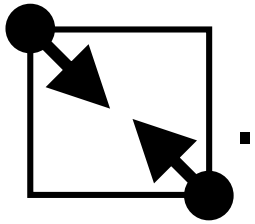
Initial box proposal:

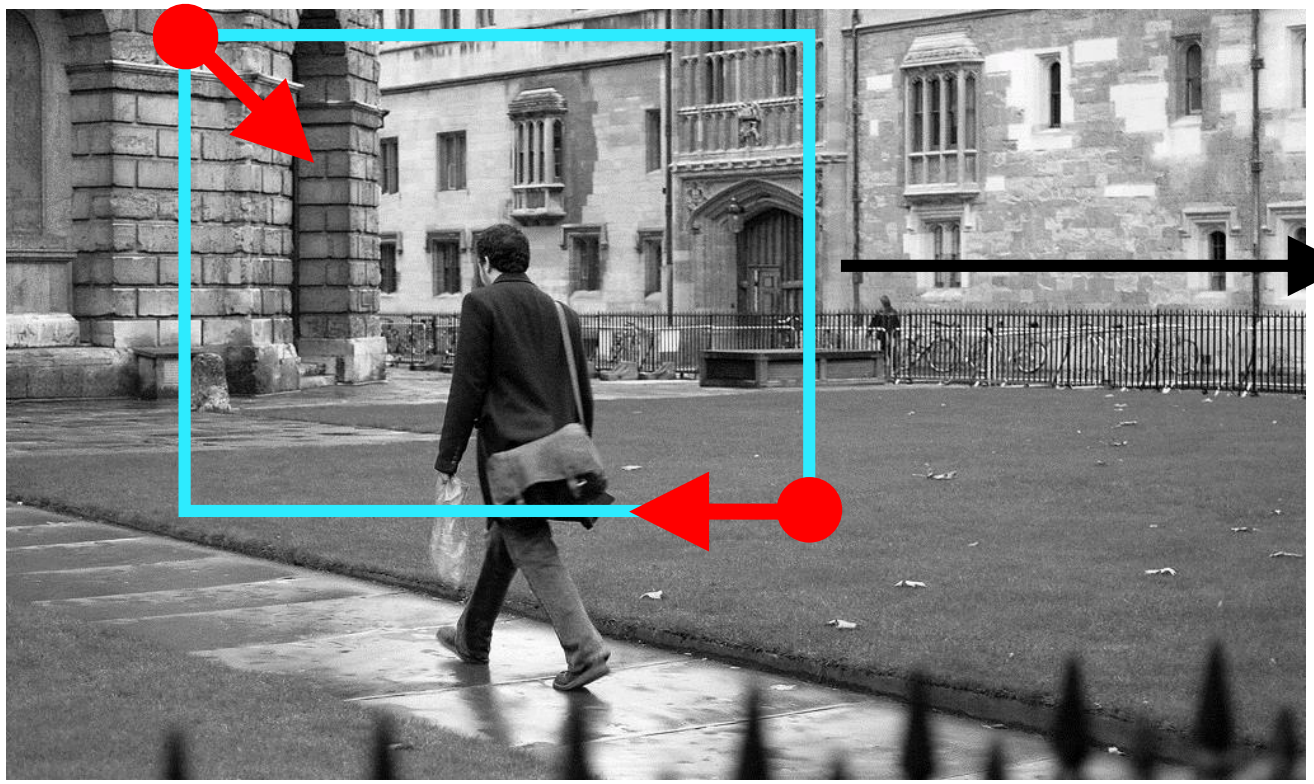
Boxes satisfying  .



Rejected.

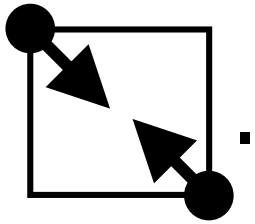
Initial box proposal:

Boxes satisfying  .



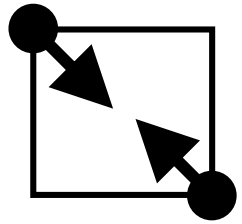
Rejected.

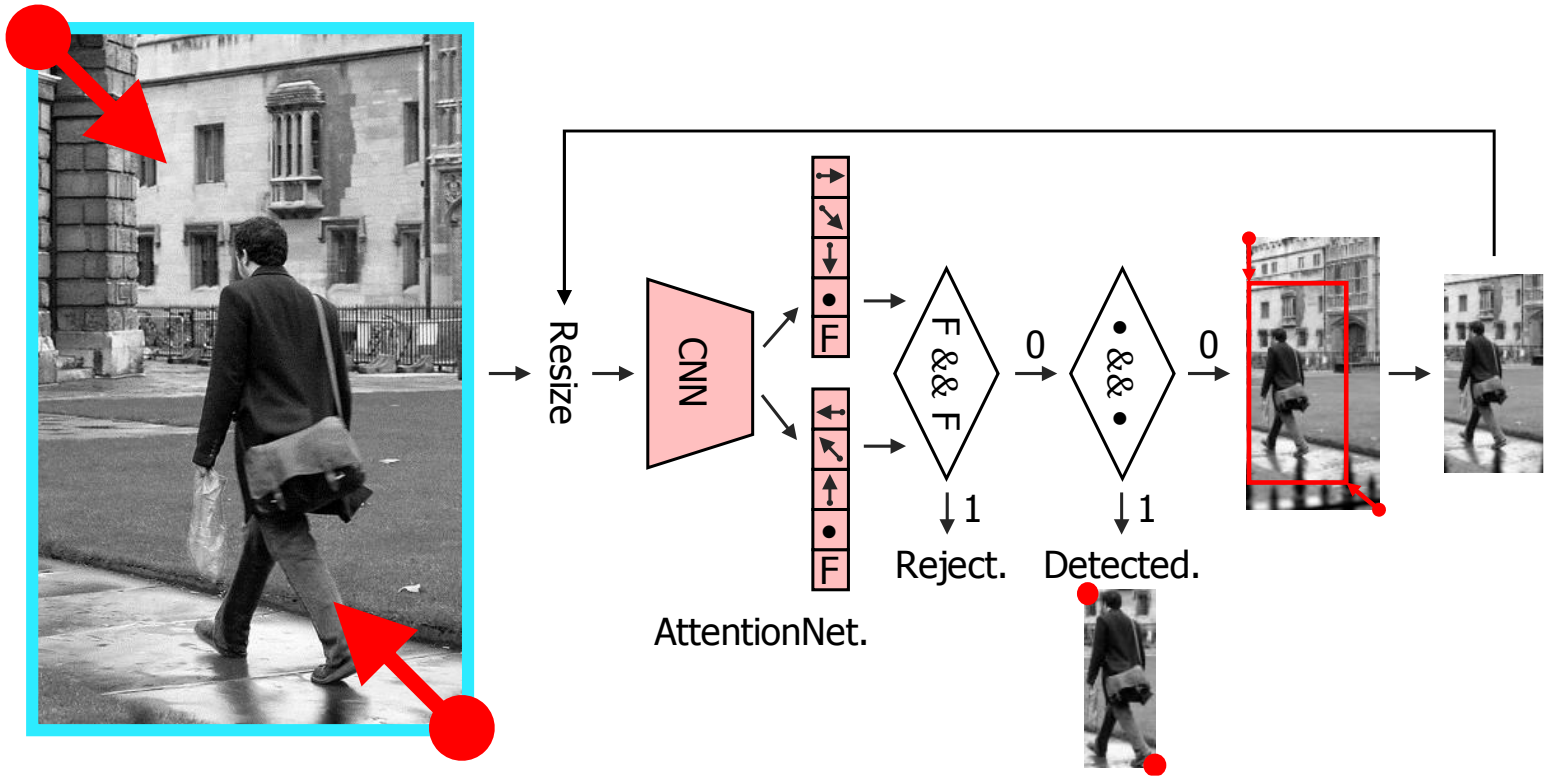
Initial box proposal:

Boxes satisfying  .

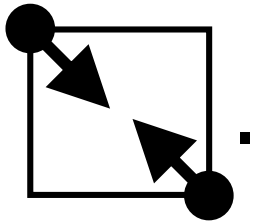
Continue.

Initial box proposal:

Boxes satisfying  .

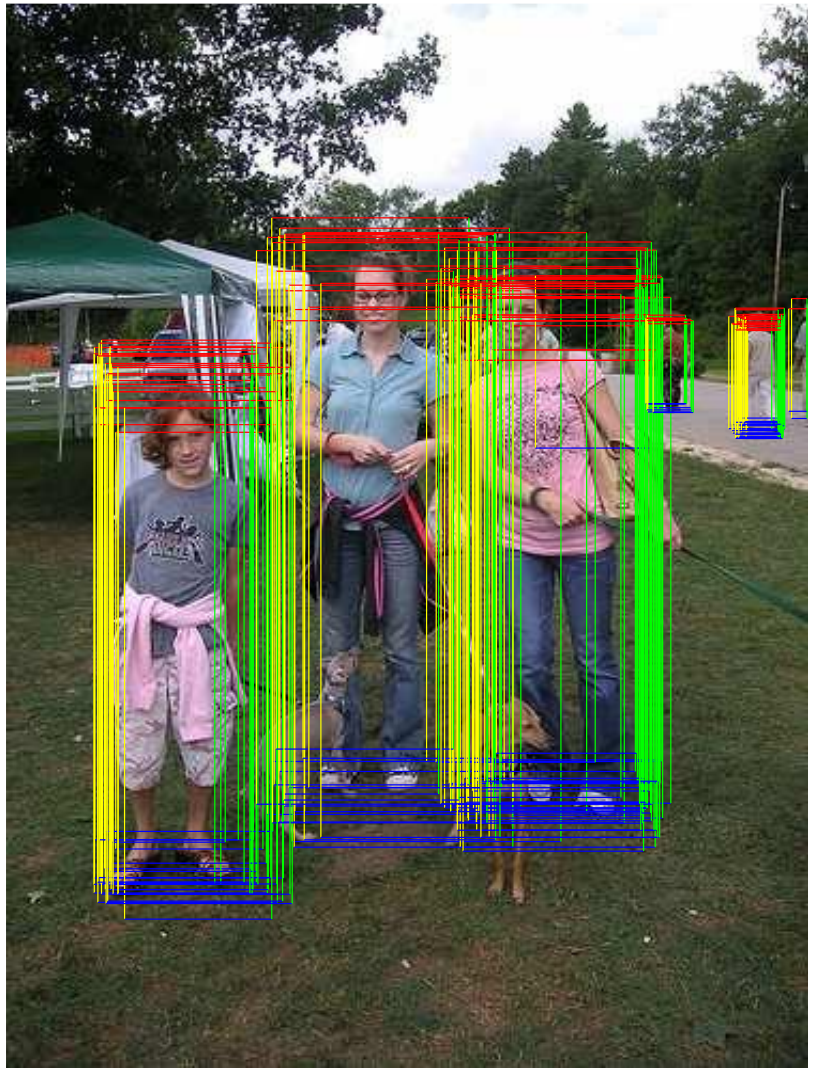


Initial box proposal:

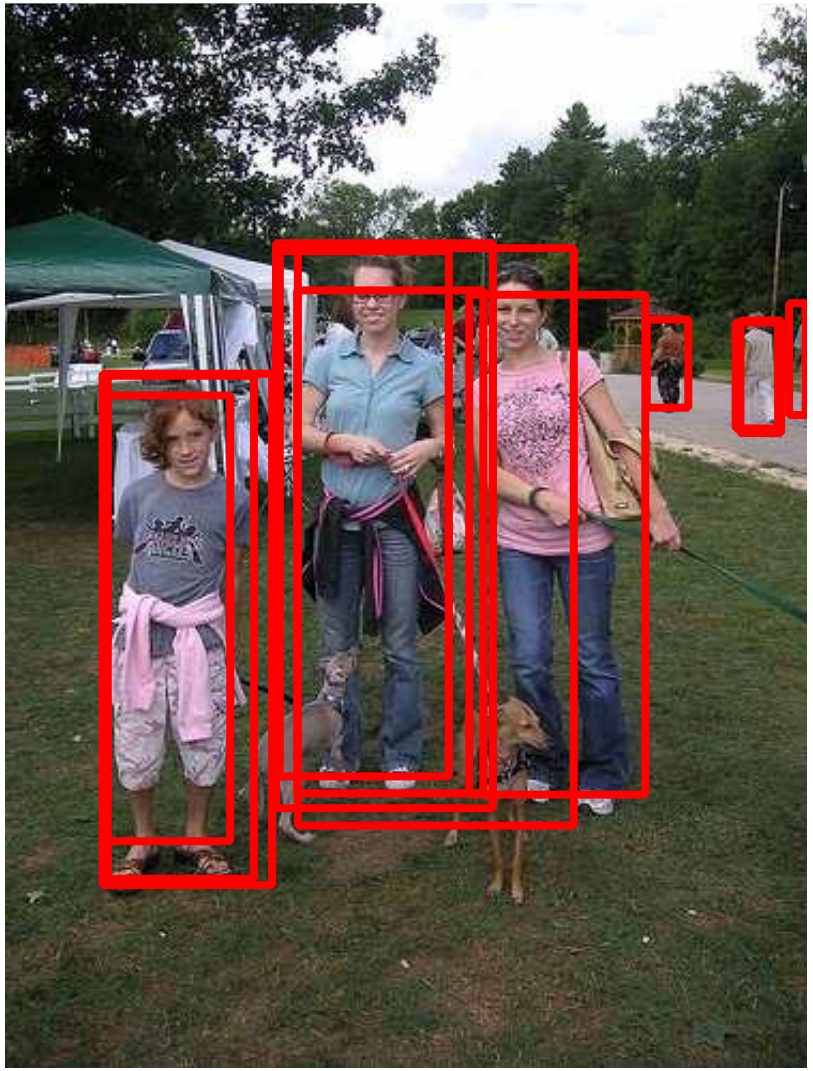
Boxes satisfying  .

Multi-{scale, aspect ratio} sliding window search using ***fully-convolutional network***.

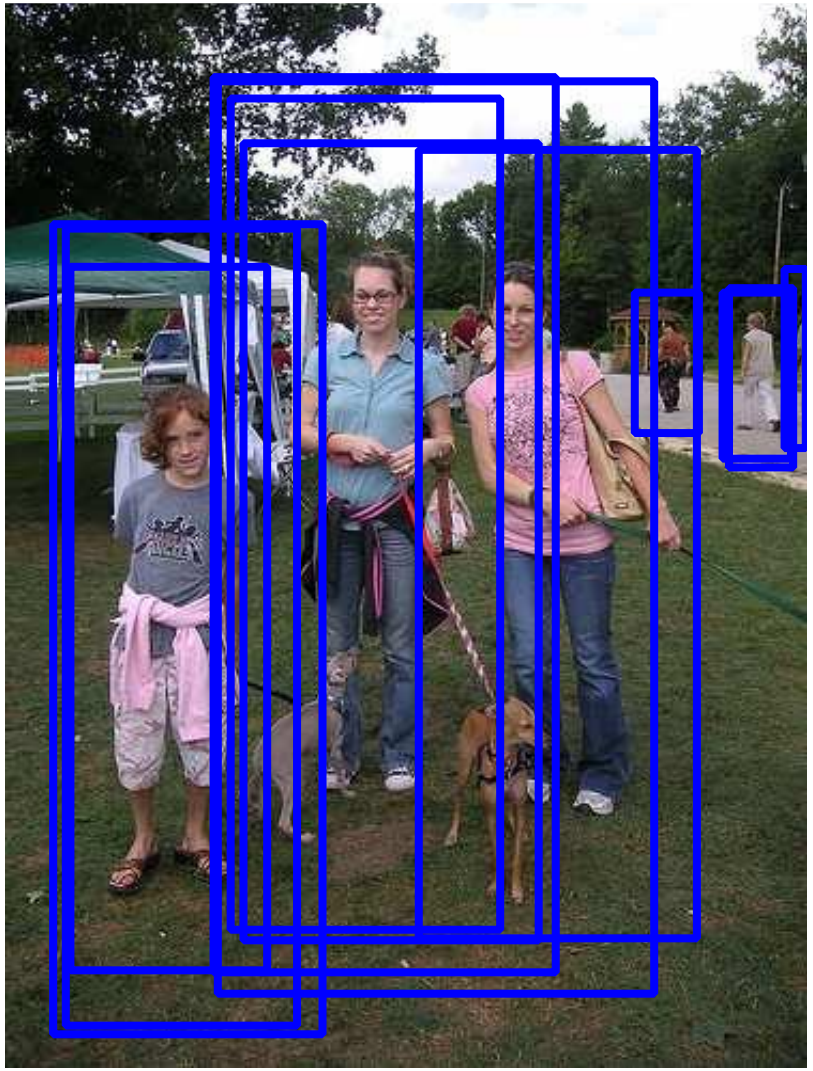
Initial detection and refinement.



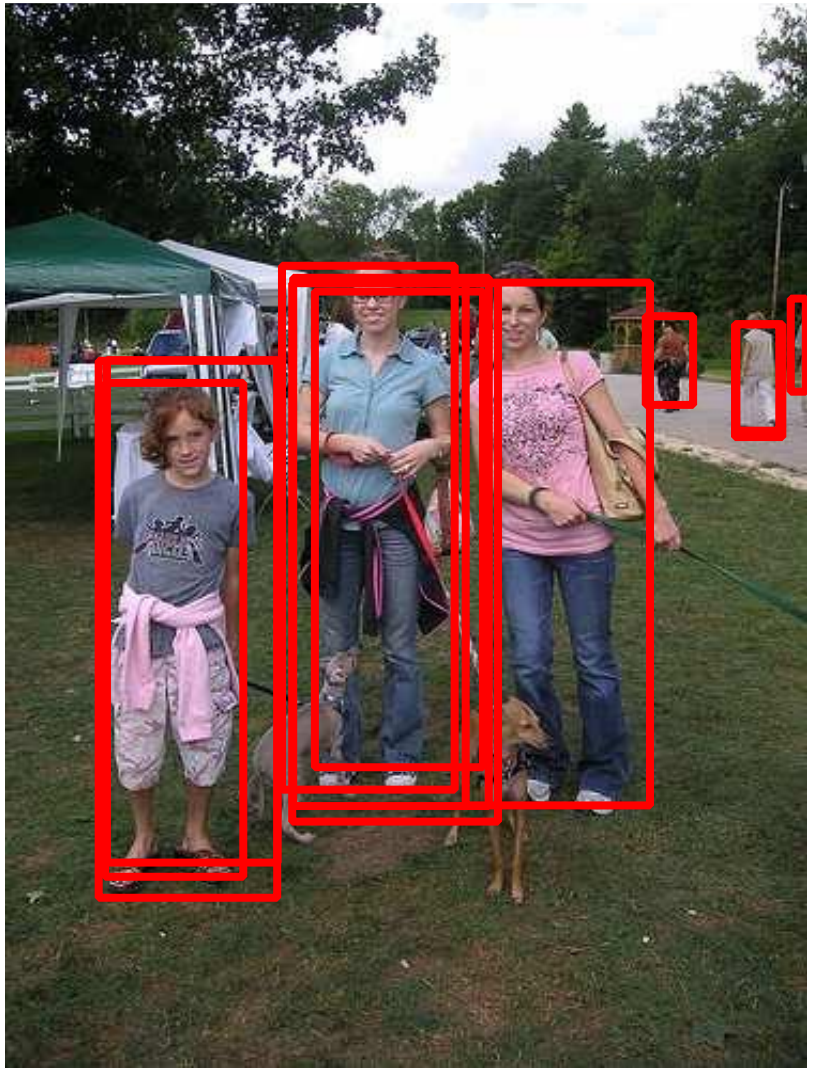
Initial detection and refinement.



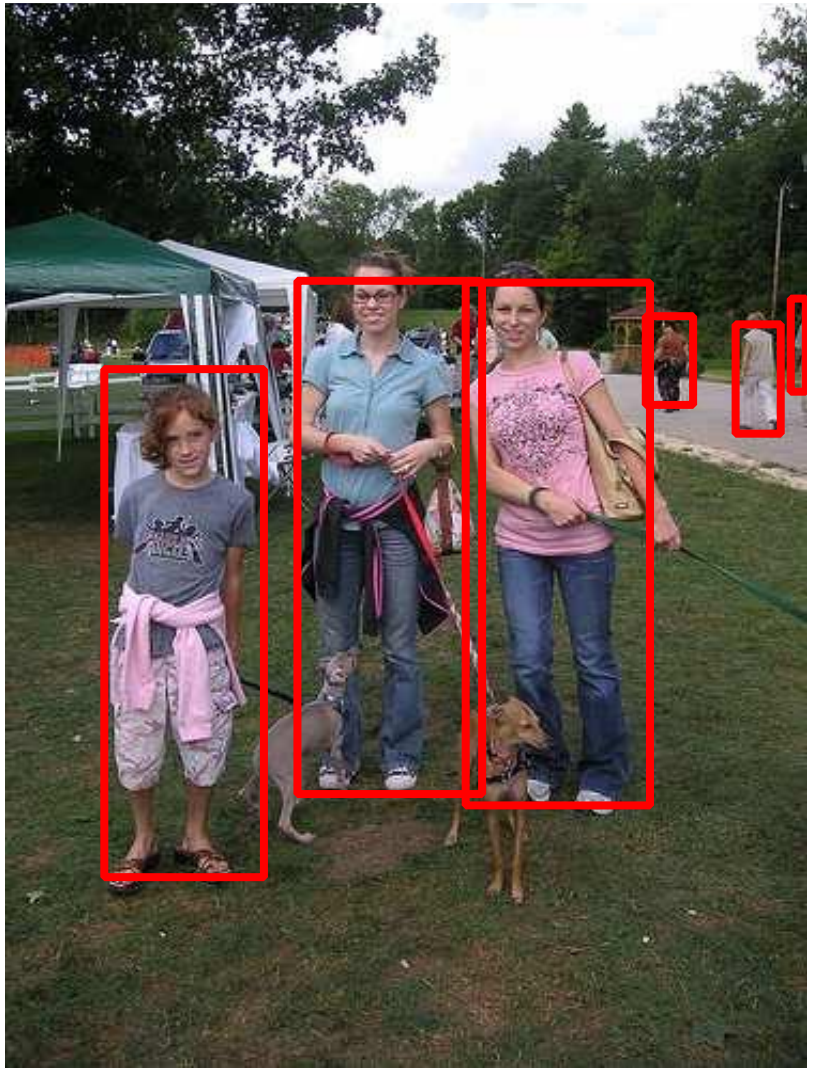
Initial detection and refinement.



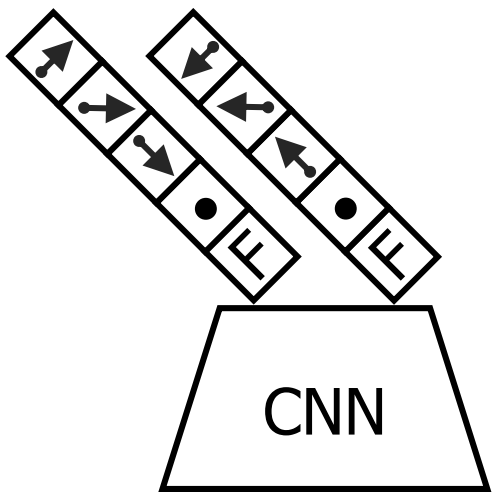
Initial detection and refinement.



Initial detection and refinement.



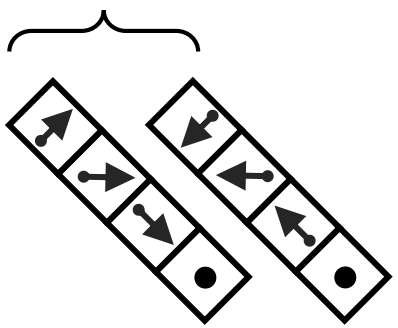
Extension to multiple classes.



AttentionNet.

Extension to multiple classes.

Class 1.



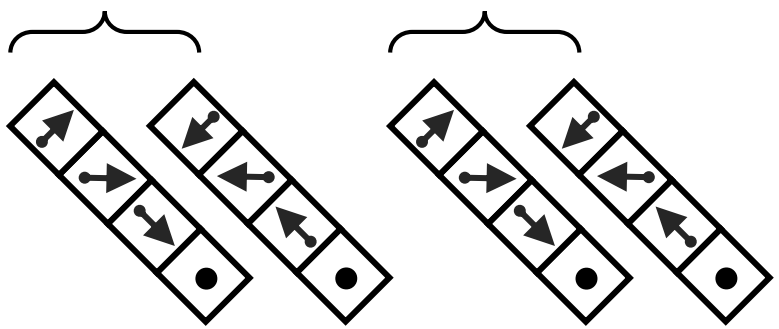
CNN

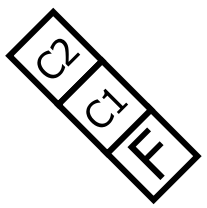
AttentionNet.

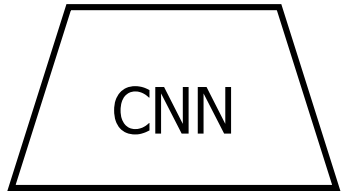
Extension to multiple classes.

Class 1.

Class 2.

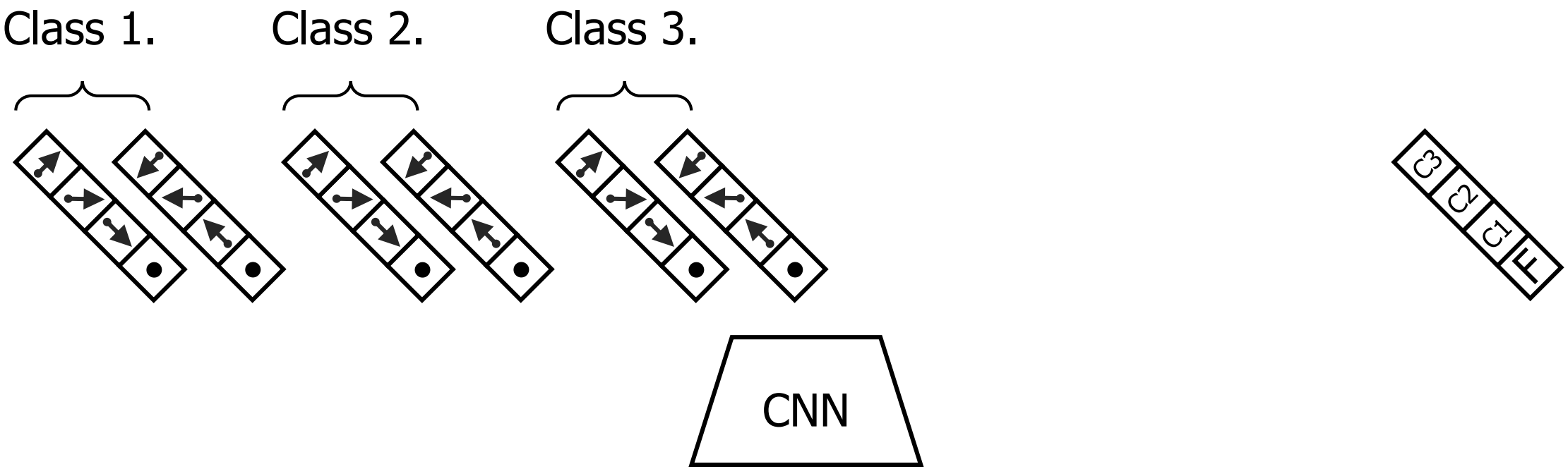






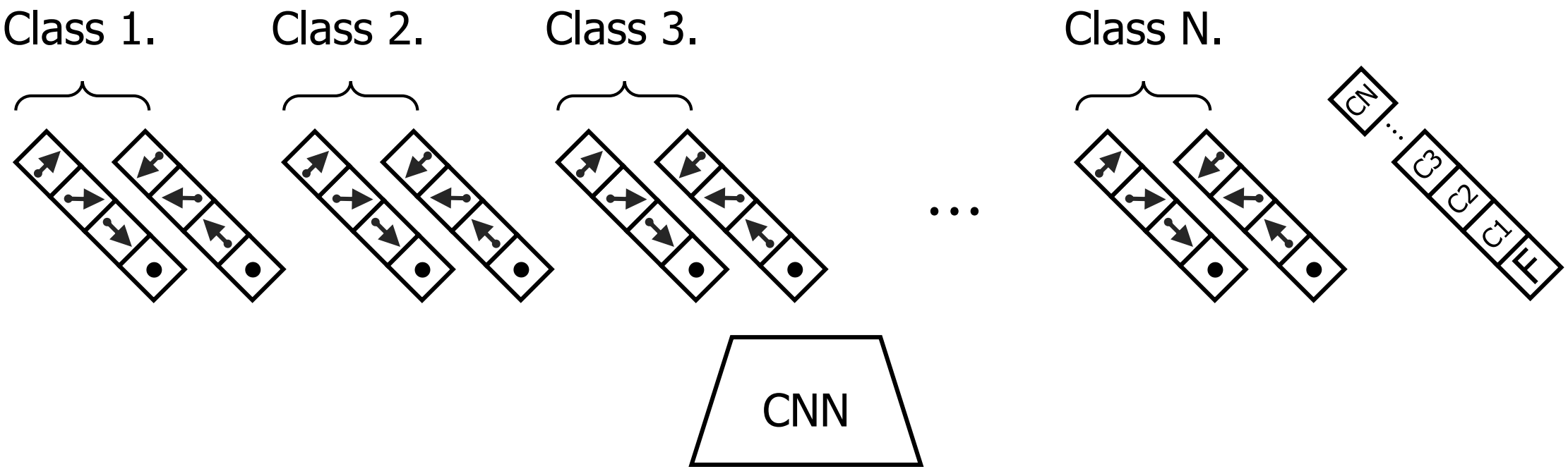
Multi-class AttentionNet.

Extension to multiple classes.



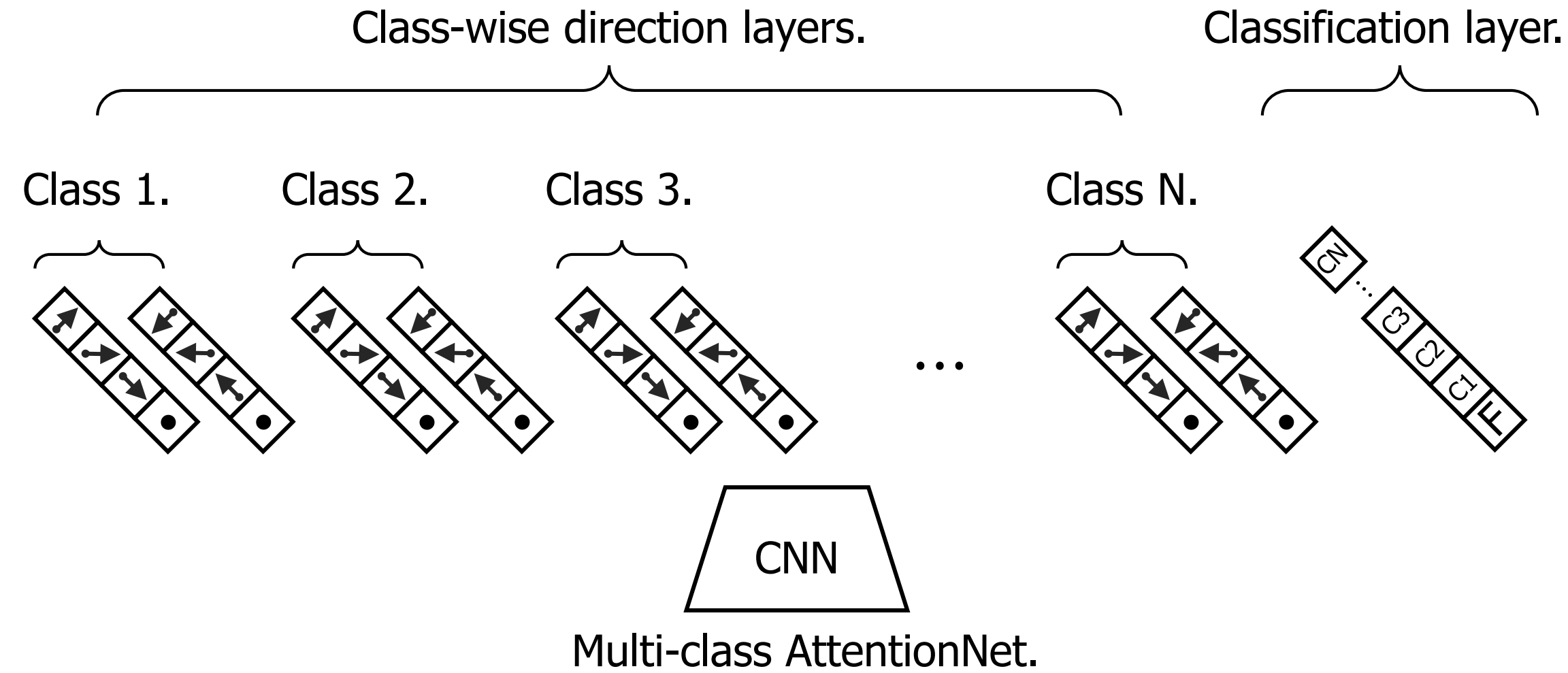
Multi-class AttentionNet.

Extension to multiple classes.



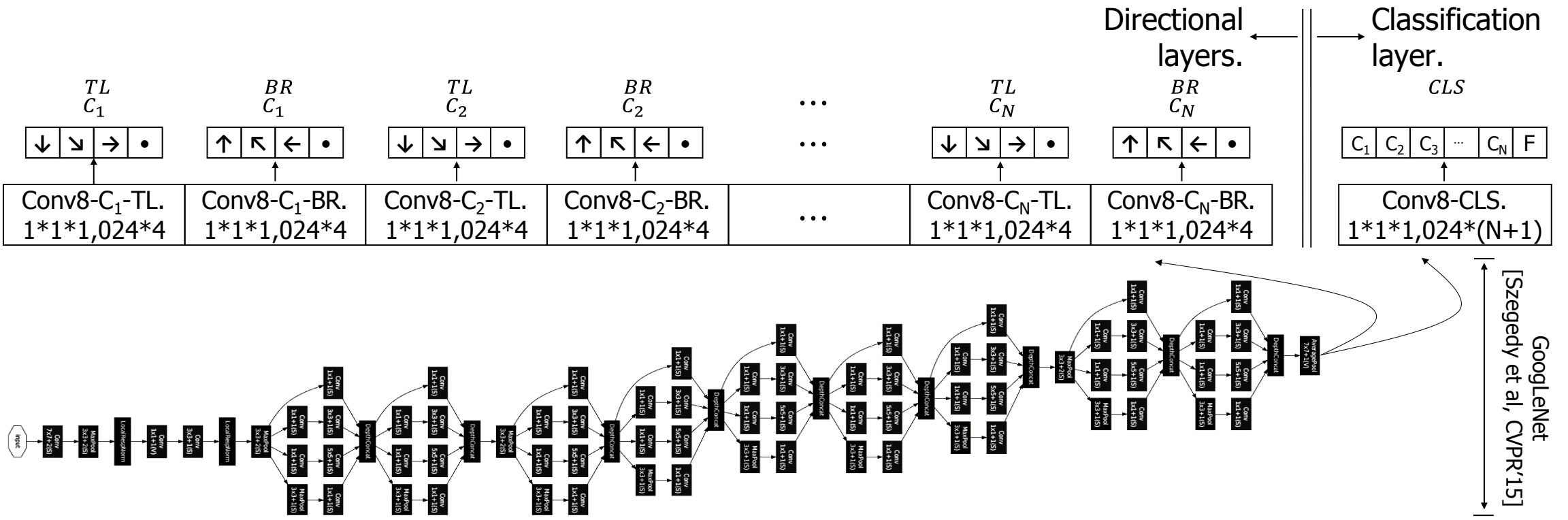
Multi-class AttentionNet.

Extension to multiple classes.



Multi-class AttentionNet.

Final architecture.



Training multi-class AttentionNet.

Training multi-class AttentionNet.

- Pre-training.
  - GoogLeNet [Szegedy et al, CVPR'15].
  - ILSVRC-CLS dataset.

Training multi-class AttentionNet.

- Pre-training.
  - GoogLeNet [Szegedy et al, CVPR'15].
  - ILSVRC-CLS dataset.
- Fine-tuning.
  - # epochs: 5.
  - # training region: 22M. (randomly generated.)
  - Learning rate of the classification layer: 0.01.
  - Learning rate of the 2K(=1K+1K) directional layers: 0.01.
  - Learning rate of the layers from conv1 to conv21: 0.001.

Training multi-class AttentionNet.

$$Loss = \underbrace{\frac{1}{3}Loss^{TL} + \frac{1}{3}Loss^{BR}}_{\text{Directional terms.}} + \underbrace{\frac{1}{3}Loss^{CLS}}_{\text{Classification term.}},$$

Training multi-class AttentionNet.

$$Loss = \frac{1}{3} Loss^{TL} + \frac{1}{3} Loss^{BR} + \frac{1}{3} Loss^{CLS},$$

$$Loss^{TL} = \frac{1}{N} \sum_{i=1}^{N} \left(t_{c_i}^{TL} \neq 0 \right) \cdot SoftMaxLoss\left(y_{c_i}^{TL}, t_{c_i}^{TL} \right),$$

$$Loss^{BR} = \frac{1}{N} \sum_{i=1}^{N} \left(t_{c_i}^{BR} \neq 0 \right) \cdot SoftMaxLoss\left(y_{c_i}^{BR}, t_{c_i}^{BR} \right),$$

$$Loss^{CLS} = SoftMaxLoss(y^{CLS}, t^{CLS}).$$

Test:
Given top-5 class predictions,
we detect the classes by AttentionNet.

Test:
Given top-5 class predictions, we detect the classes by AttentionNet.

- Top-5 class prediction (7% Err):
  Ensemble of GoogLeNet, GoogLeNet-BN, VGG-16.

Test:
Given top-5 class predictions, we detect the classes by AttentionNet.

- Top-5 class prediction (7% Err):
  Ensemble of GoogLeNet, GoogLeNet-BN, VGG-16.

- Number of multi-{scale, aspect ratio} inputs: 6.

Results on validation set.

Results on validation set.

Method.	Top-5 CLS-LOC Error.
OverFeat [Sermanet et al., ICLR'14]	30.00%
VGG [Simonyan and Zisserman, ICLR'15]	26.90%
GoogLeNet [Szegedy et al, CVPR'15]	26.70% (test set)

Results on validation set.

Method.	Top-5 CLS-LOC Error.
OverFeat [Sermanet et al., ICLR'14]	30.00%
VGG [Simonyan and Zisserman, ICLR'15]	26.90%
GoogLeNet [Szegedy et al, CVPR'15]	26.70% (test set)
A single "Multi-class AttentionNet", without test augmentation.	16.11%

Results on validation set.

Method.	Top-5 CLS-LOC Error.
OverFeat [Sermanet et al., ICLR'14]	30.00%
VGG [Simonyan and Zisserman, ICLR'15]	26.90%
GoogLeNet [Szegedy et al, CVPR'15]	26.70% (test set)
A single "Multi-class AttentionNet", without test augmentation.	16.11%
A single "Multi-class AttentionNet", with test augmentation (original and flip).	**14.96%**

Results on validation set.

Method.	Top-5 CLS-LOC Error.
OverFeat [Sermanet et al., ICLR'14]	30.00%
VGG [Simonyan and Zisserman, ICLR'15]	26.90%
GoogLeNet [Szegedy et al, CVPR'15]	26.70% (test set)
A single "Multi-class AttentionNet", without test augmentation.	16.11%
A single "Multi-class AttentionNet", with test augmentation (original and flip).	**14.96%**

Note that we use a SINGLE "Multi-class AttiontionNet".

Related publication:
Donggeun Yoo, Sunggyun Park, Joon-Young Lee, Anthony S. Paek, In So Kweon,
AttentionNet: Aggregating Weak Directions for Accurate Object Detection,
In ICCV, 2015.