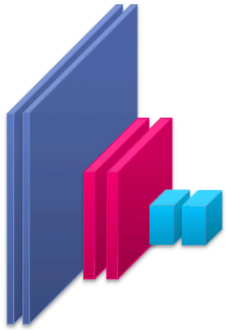
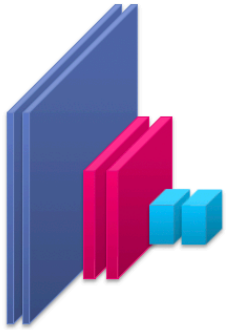
Glance and Focus: a Dynamic Approach to Reducing Spatial Redundancy in Image Classification

NeurIPS 2020

Yulin Wang, Kangchen Lv, Rui Huang, Shiji Song, Le Yang, Gao Huang

Department of Automation, Tsinghua University

Large Inputs for CNNs

Input Images (224^2, 336^2 and 560^2)	**CNN** (*EfficientNet-B0*)	**Accuracy** (*ImageNet*)	**FLOPs** (*Inference*)
		76.3%	0.39B
		78.1%	0.89B
		**79.6%**	**2.45B**

High resolution inputs give high accuracy.

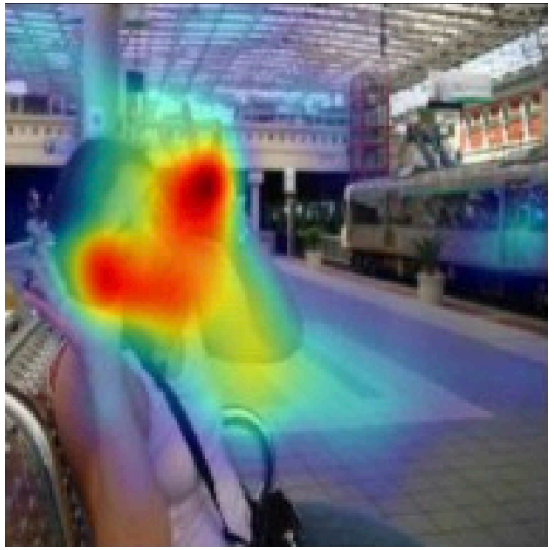
High resolution inputs are **computationally expensive** !

Spatial Redundancy

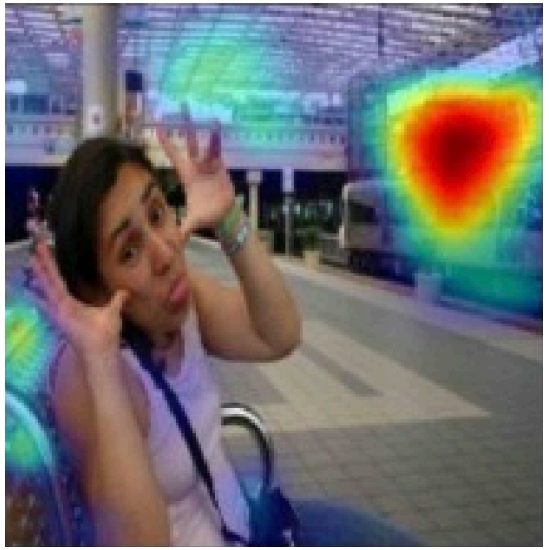
<u>Question:</u> *How do humans recognize objects in large images efficiently?*

- *We (sequentially) attend to some* ***important regions.***

Original Image

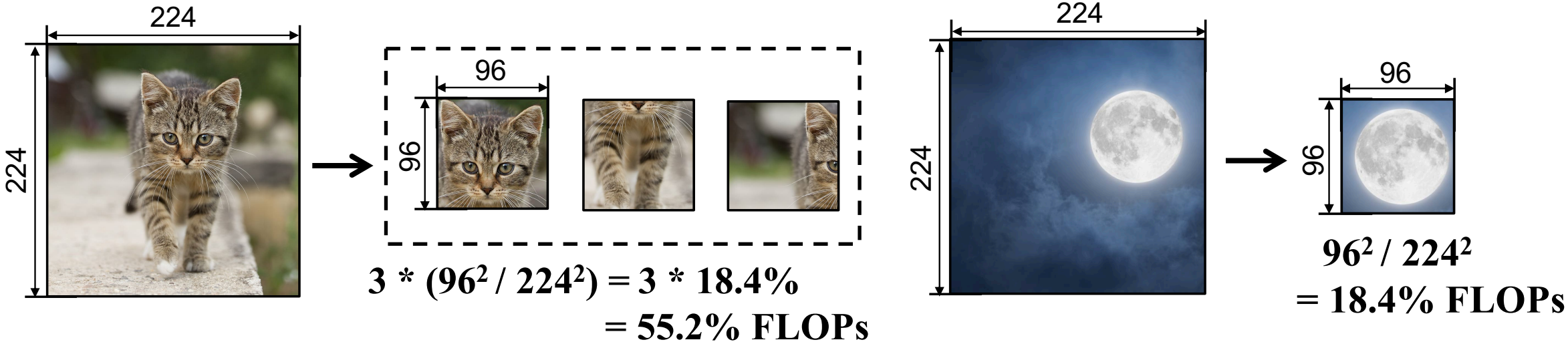


Recognize Person



Recognize Train

Leveraging Discriminative Image Patches



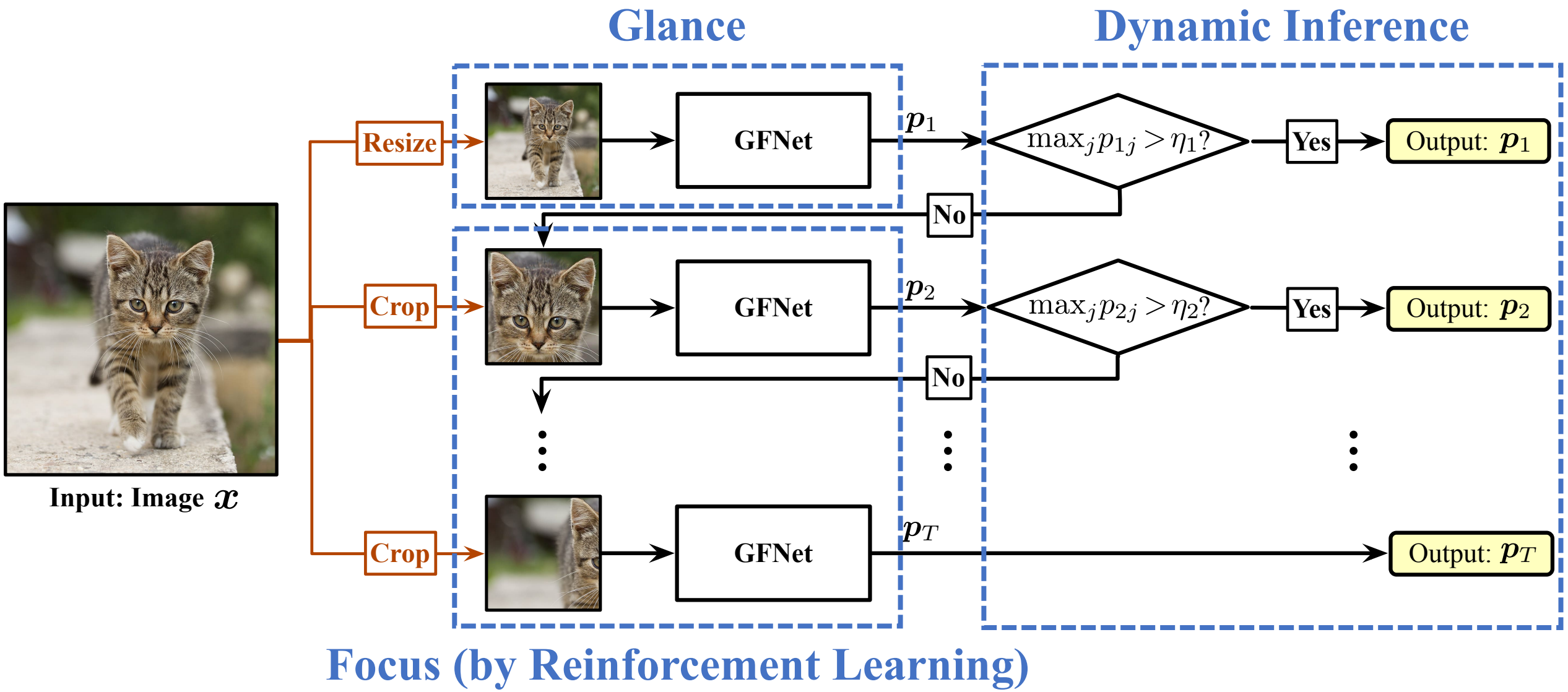
- **Challenge I:** *How to **identify** class-discriminative regions?*

  (***without*** *ground truth bounding boxes*)

- **Challenge II:** *How to determine the **number** of class-discriminative regions?*

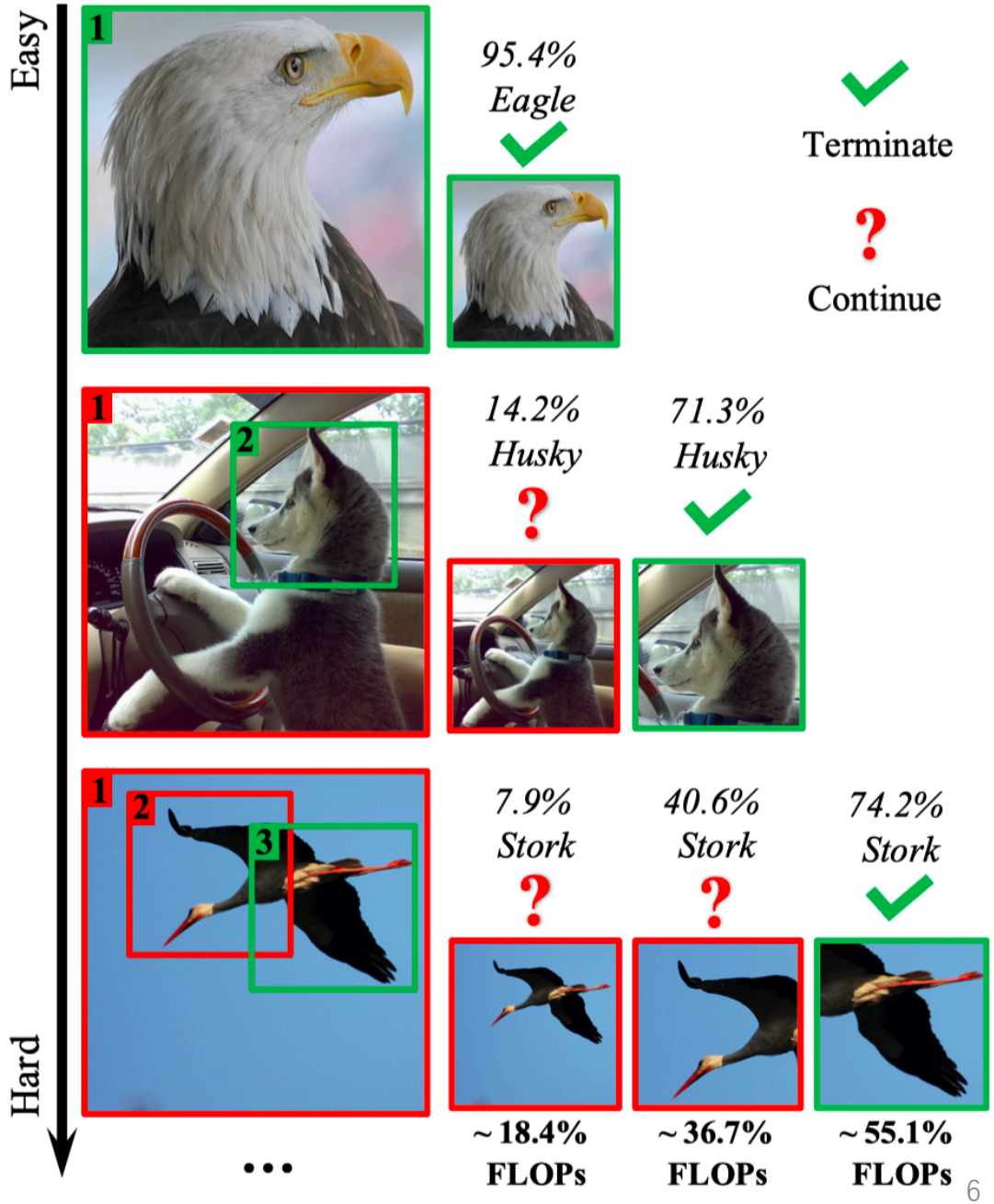
  (*to **adaptively** allocate computation across different inputs*)

A Two-stage Framework

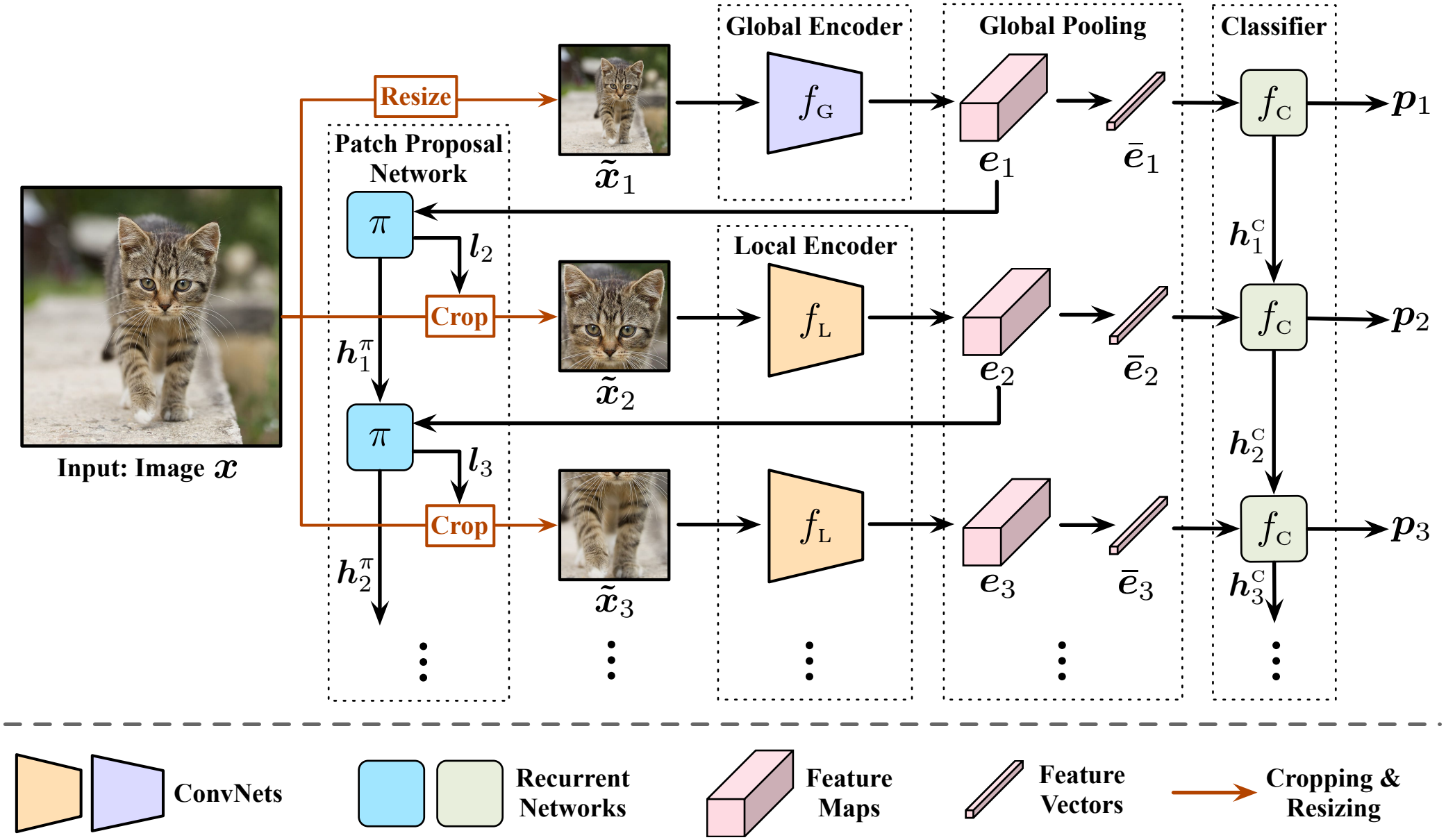


Dynamic Inference

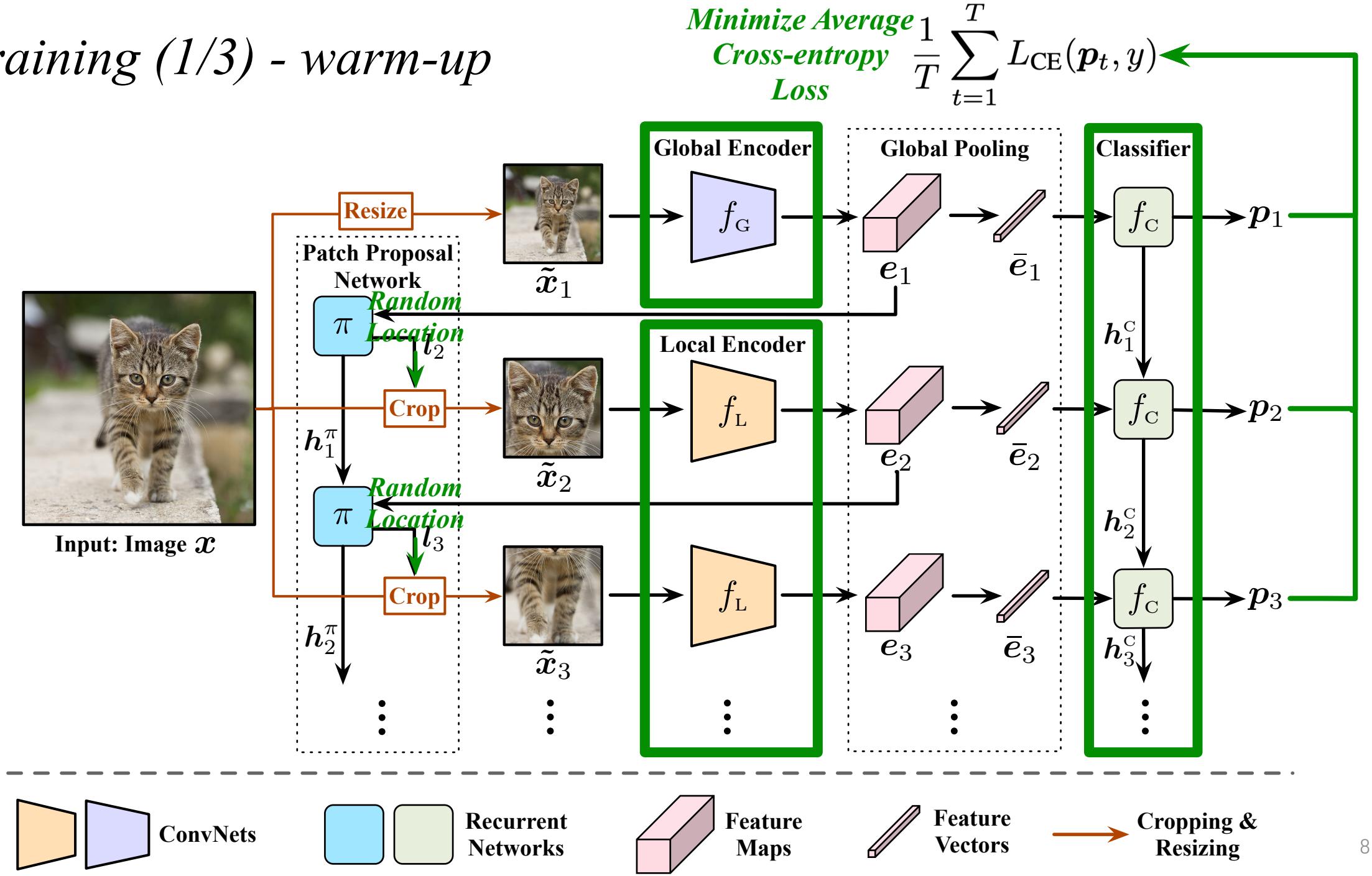
- Allocating computation **unevenly** across different samples



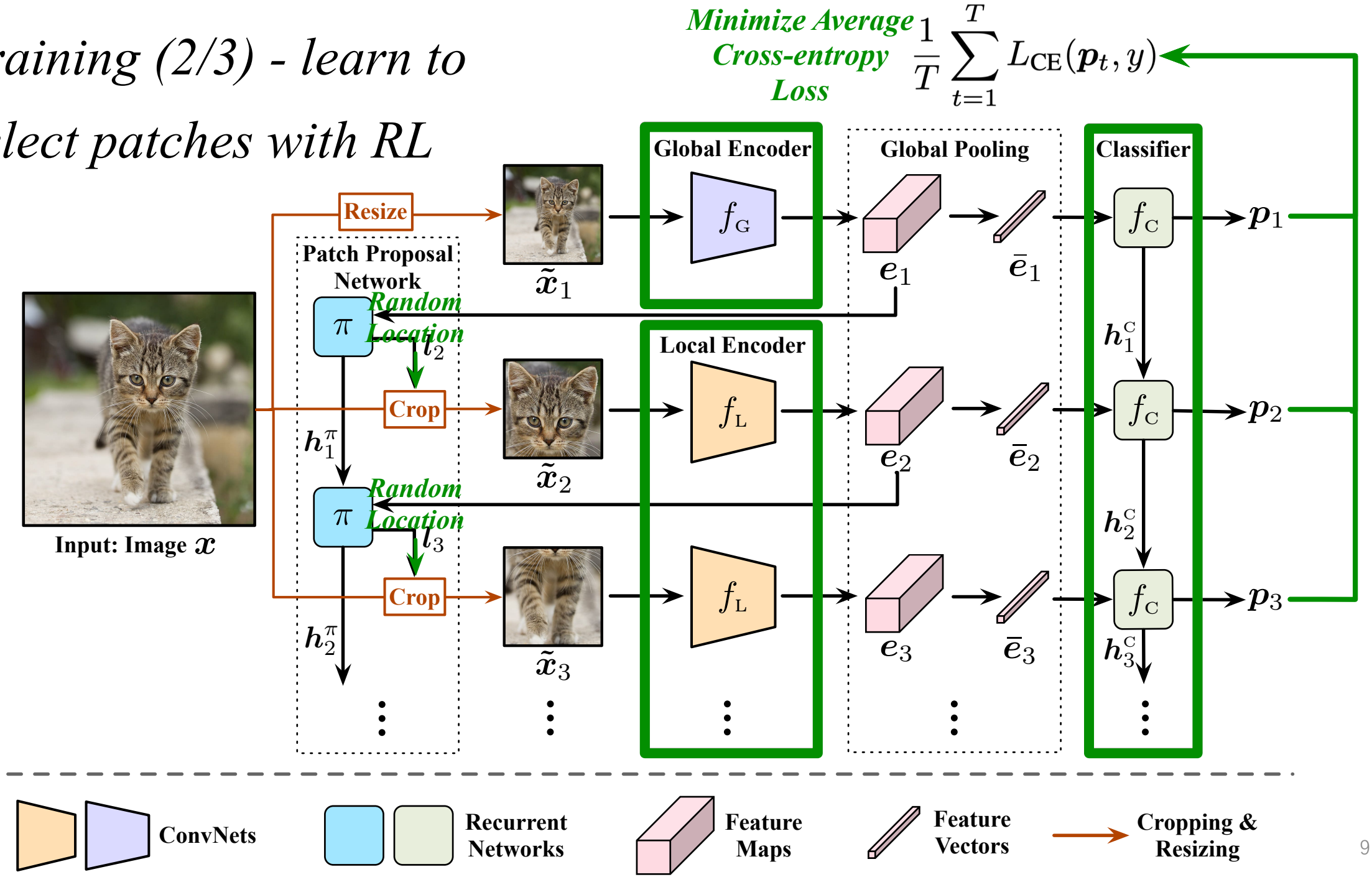
Network Architecture



Training (1/3) - warm-up



Training (2/3) - learn to select patches with RL

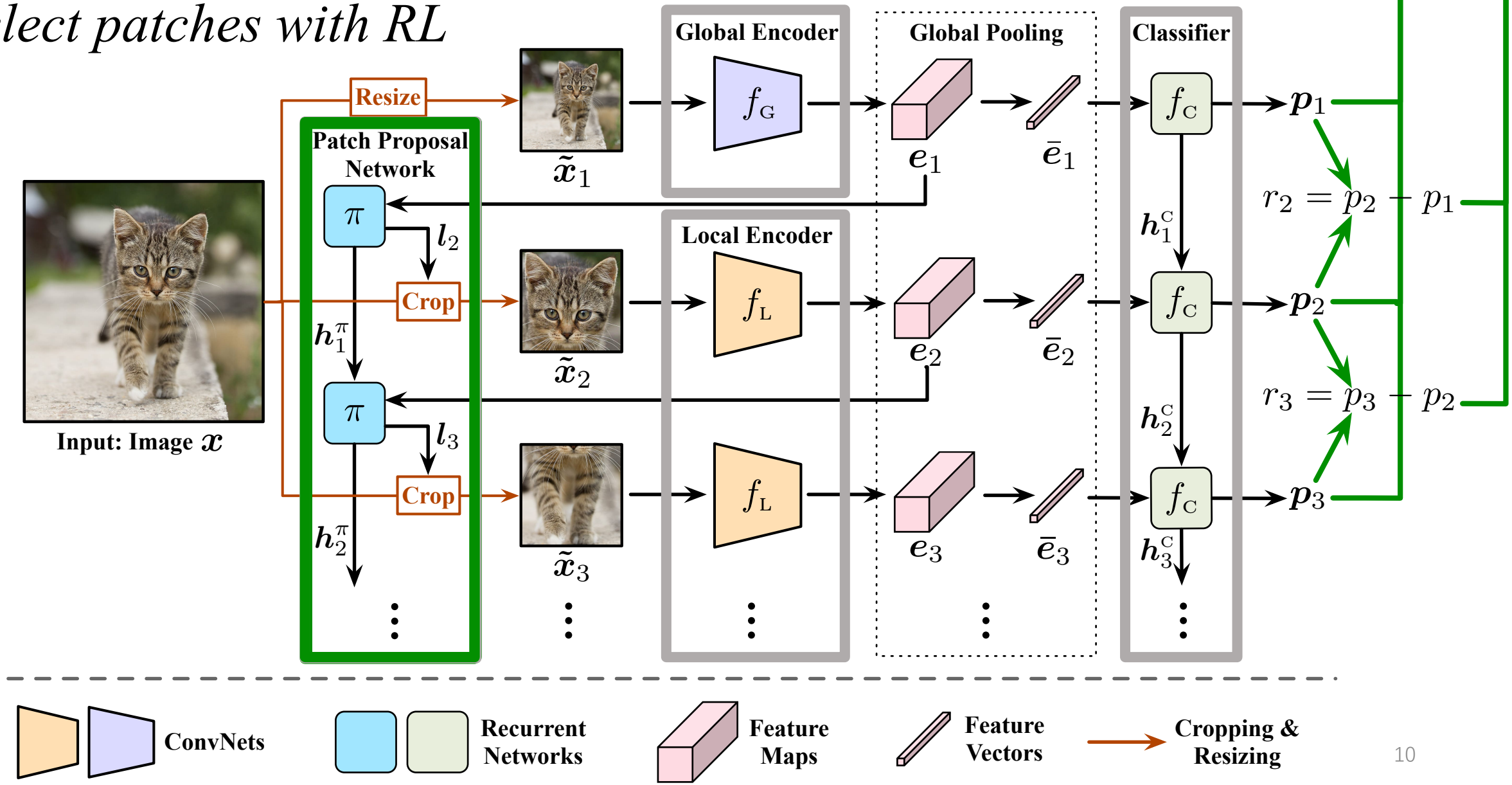


Training (2/3) - learn to select patches with RL

Minimize Average Cross-entropy Loss

Maximize the Discounted Rewards

$$\max_{\pi} \mathbb{E}\left[\sum_{t=2}^{T} \gamma^{t-2} r_t\right]$$

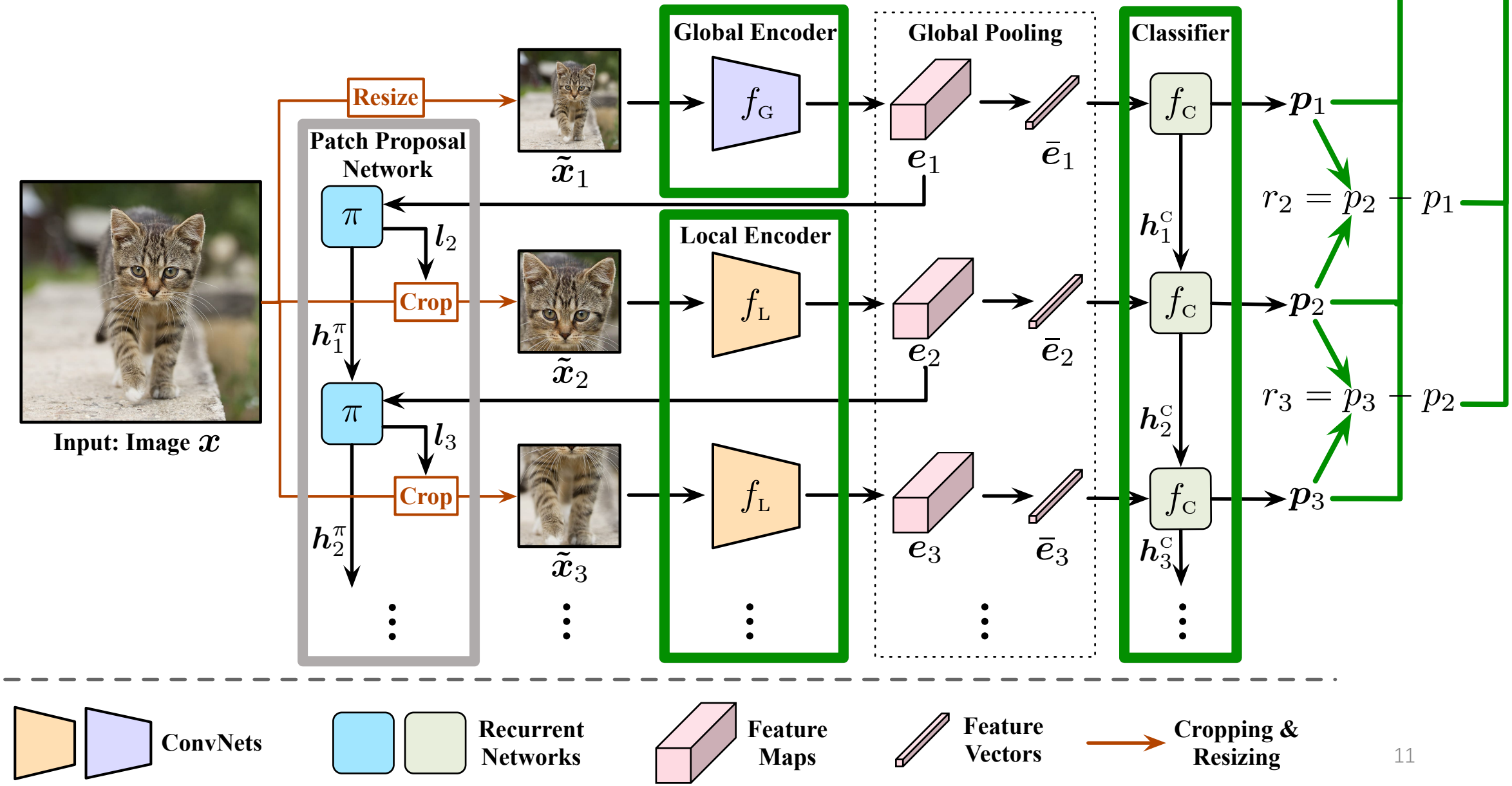


Training (3/3) - finetune CNNs

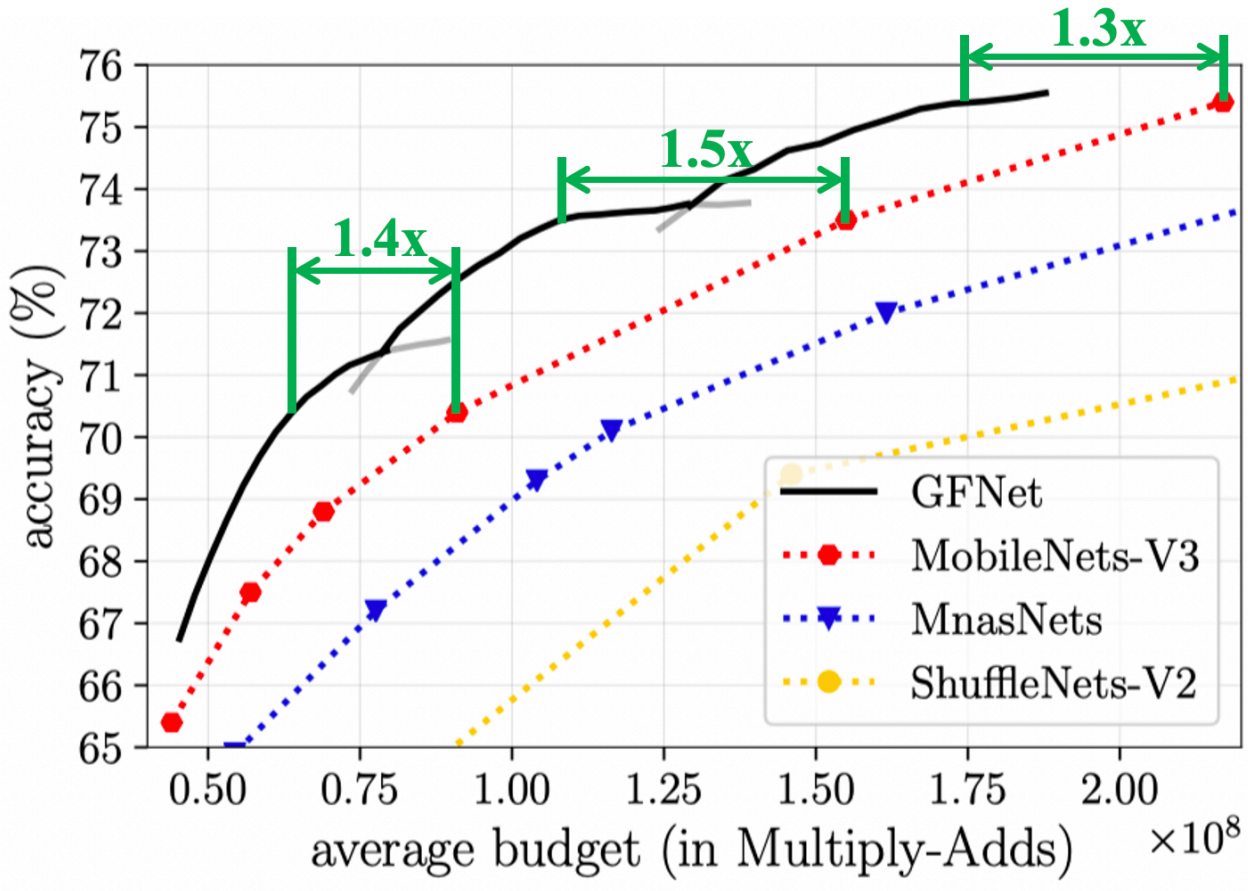
Minimize Average Cross-entropy Loss

Maximize the Discounted Rewards

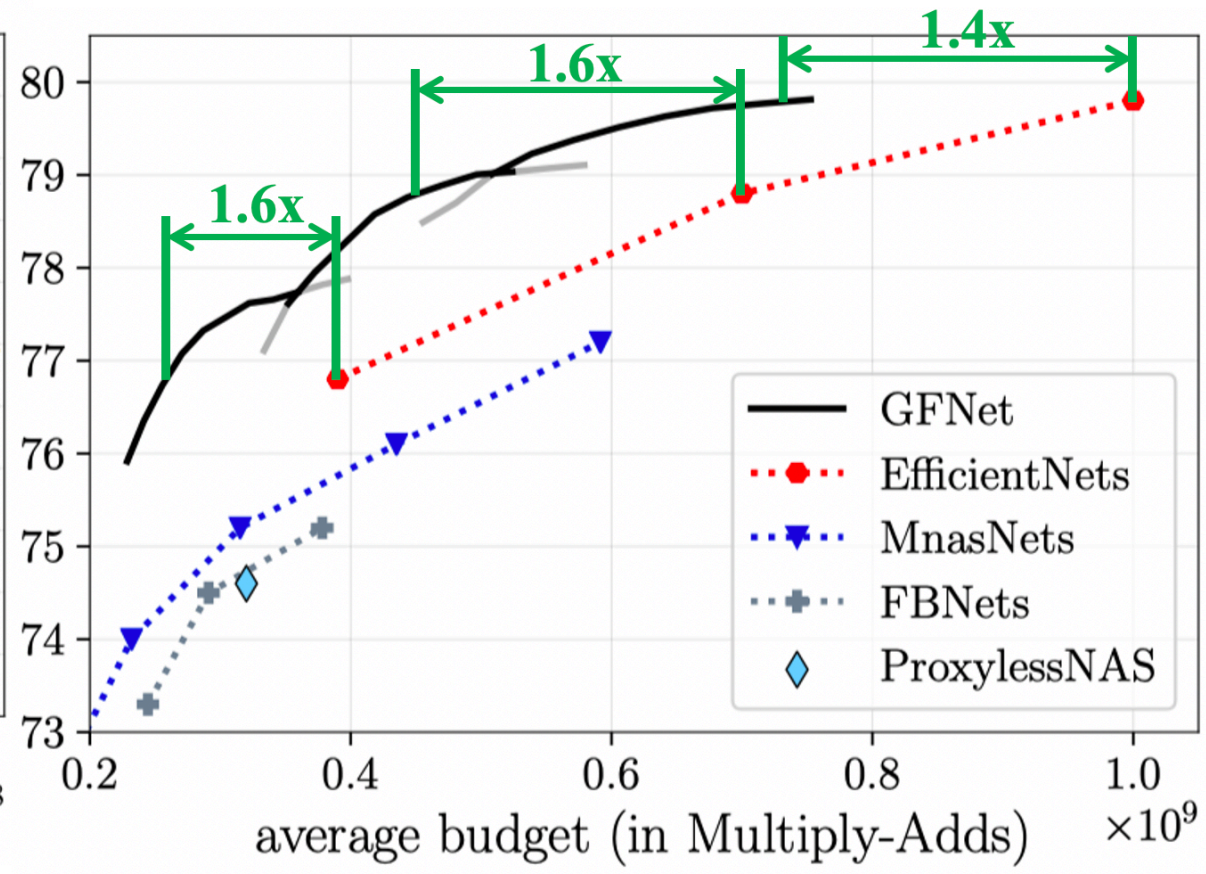
$$\max_{\pi} \mathbb{E}\left[\sum_{t=2}^{T} \gamma^{t-2} r_t\right]$$



Results (FLOPs)

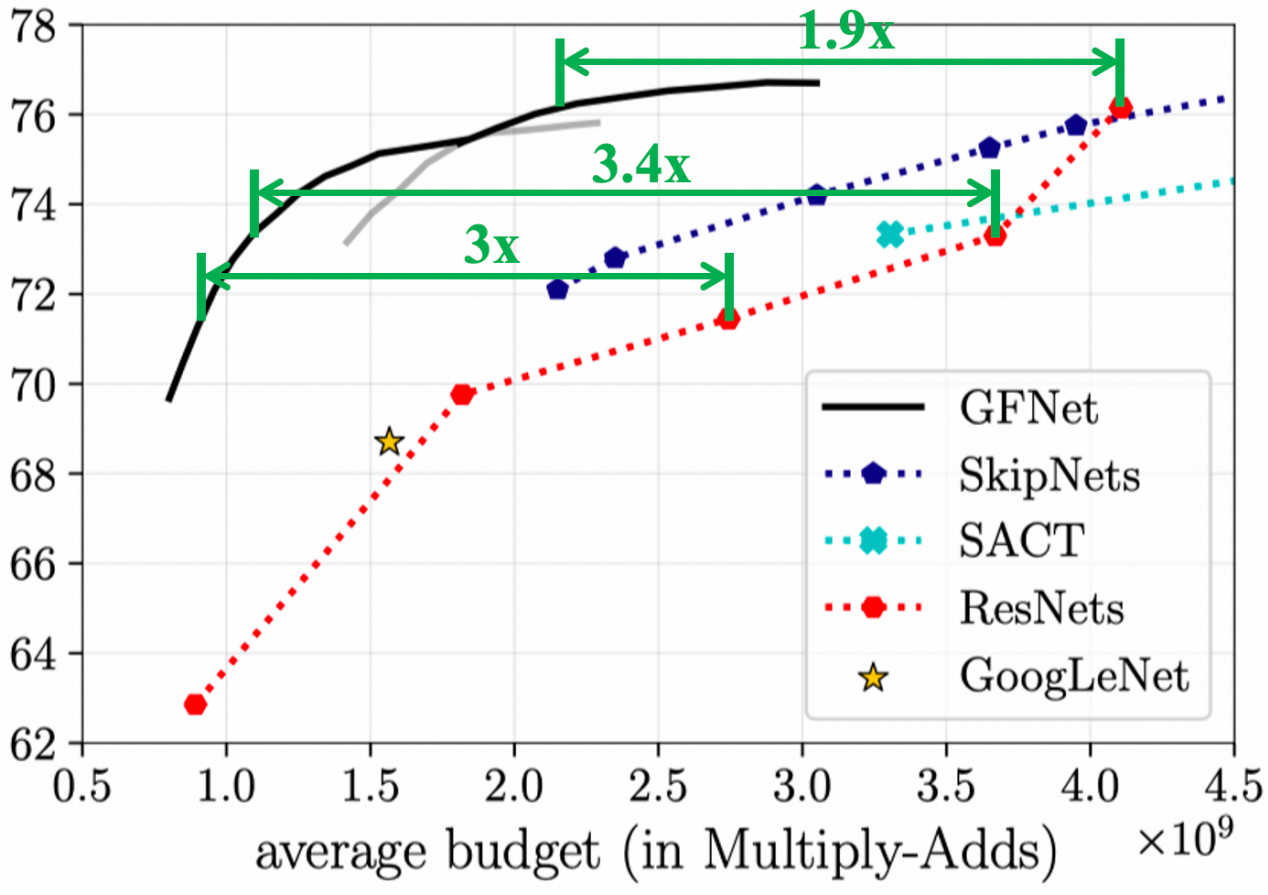


GF-MobileNet-V3



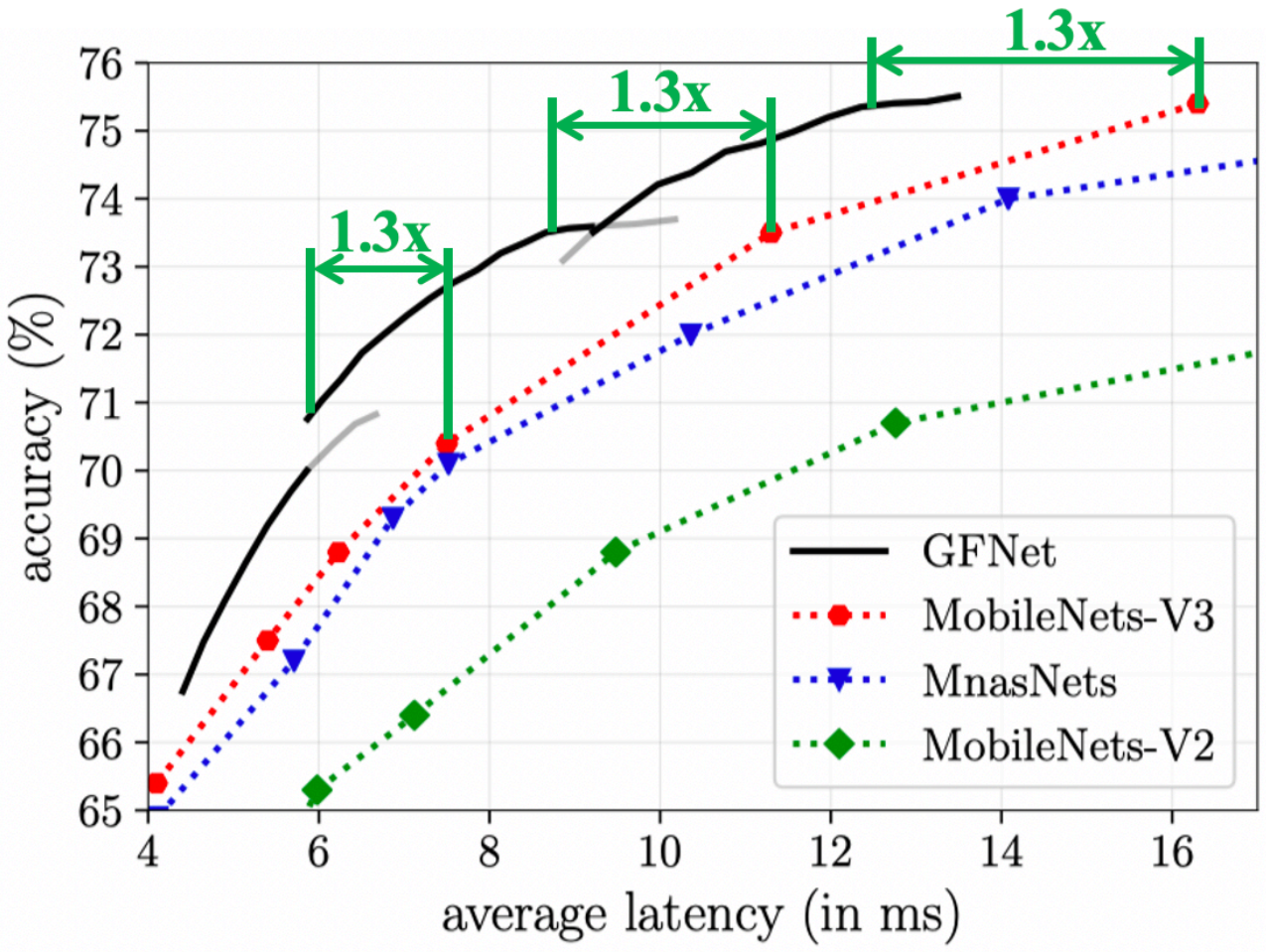
GF-EfficientNet

Results (FLOPs)

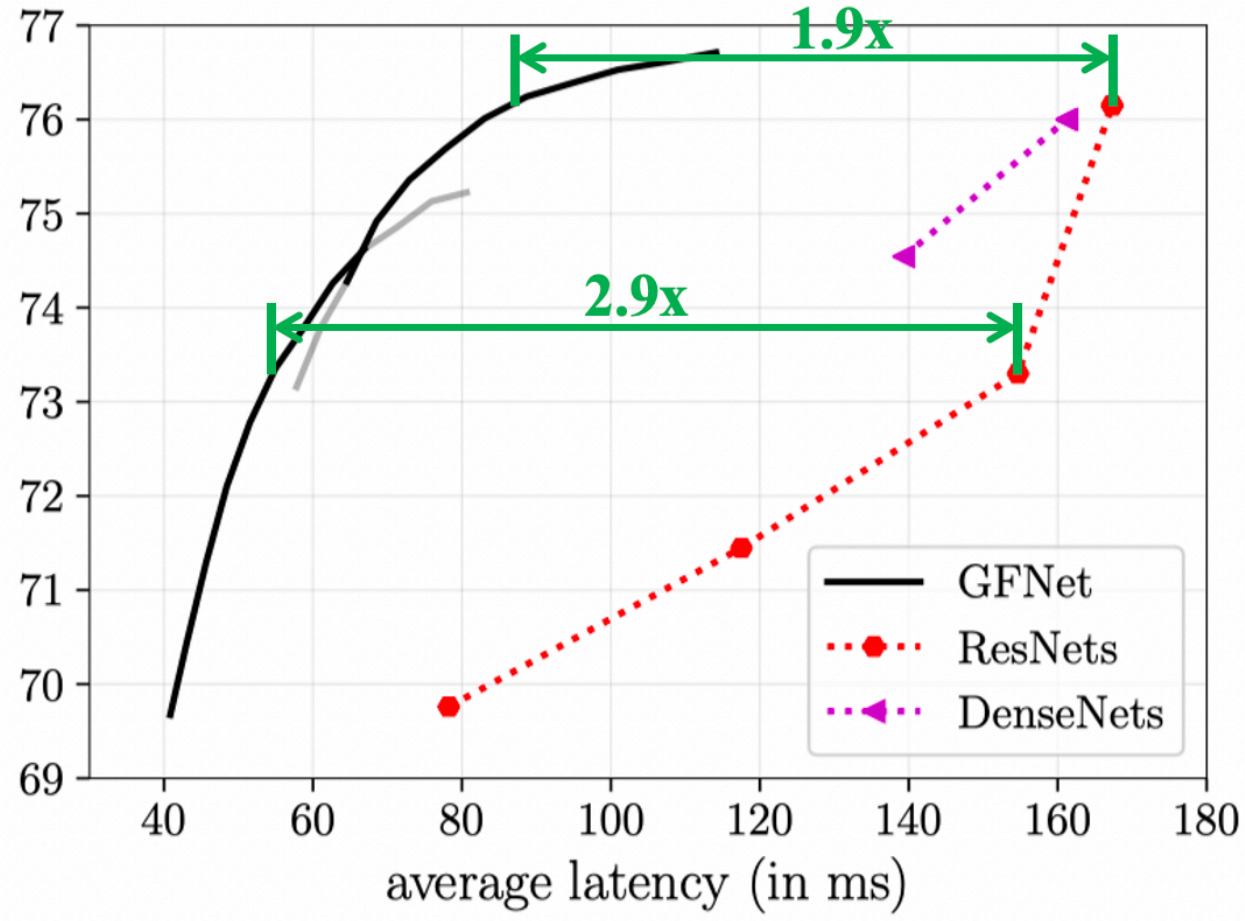


GF-ResNet

Results (iPhone XS Max)

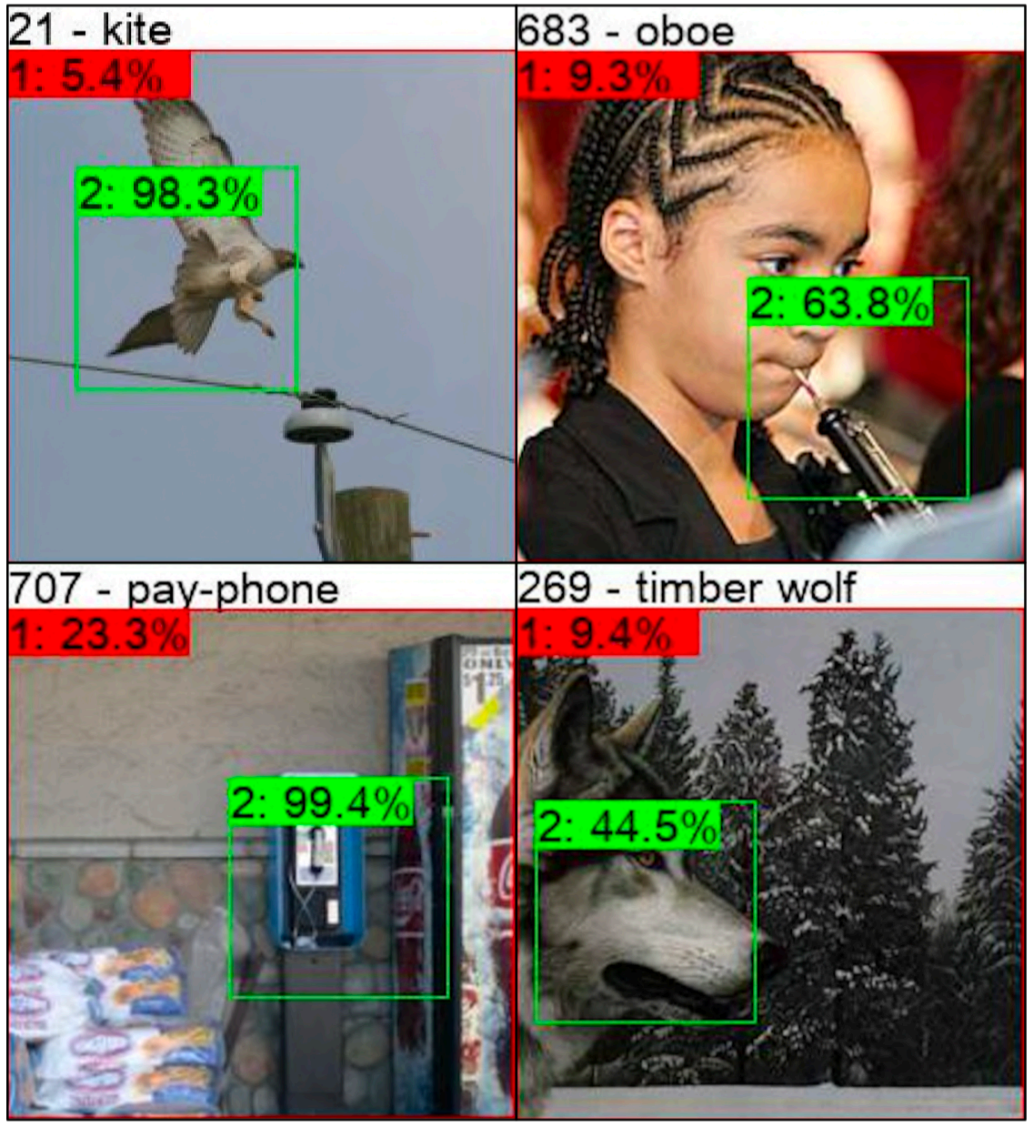


GF-MobileNet-V3

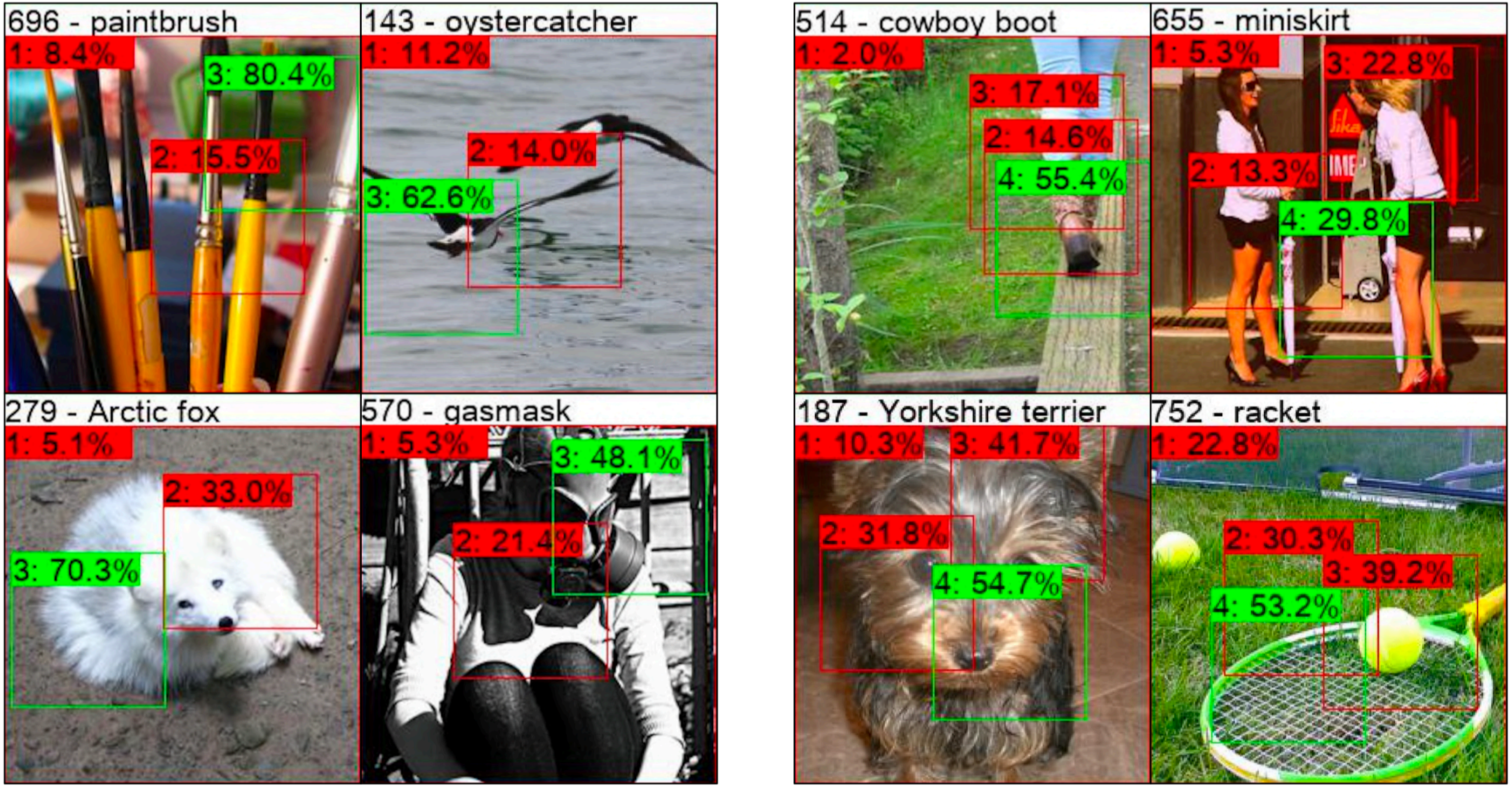


GF-ResNet

Results (visualization)



Results (visualization)



Thank You for the Time!

For more details, please refer to our paper.

Paper

Code

Contact us
(My Homepage)