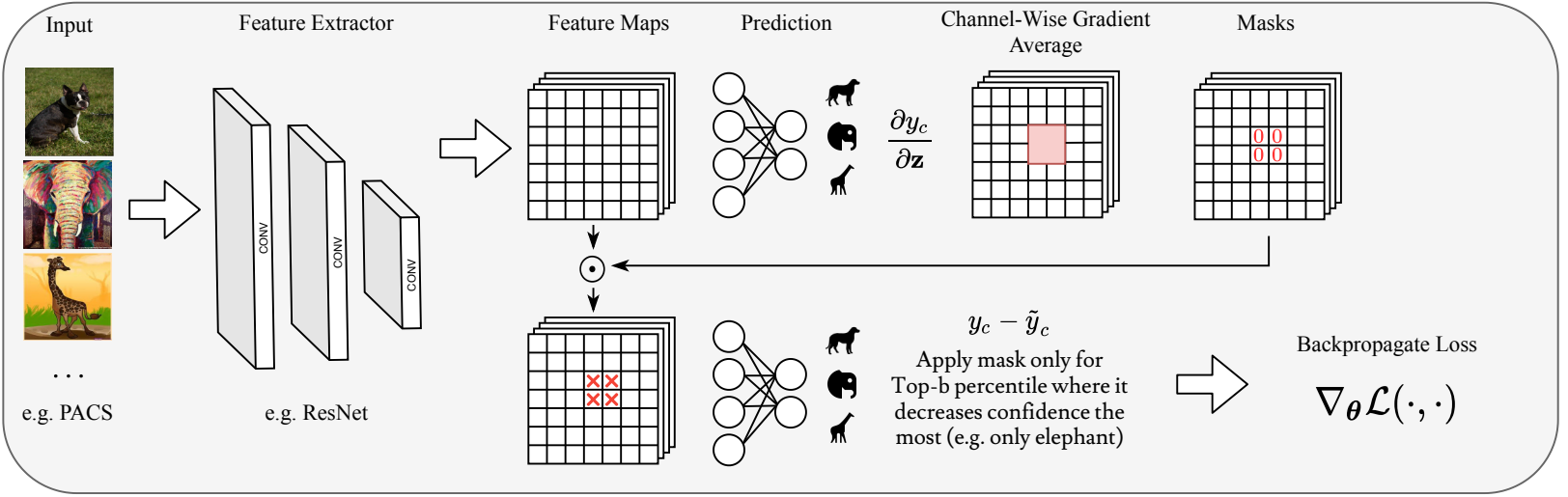
Input
Feature Extractor
Feature Maps
Prediction
Channel-Wise Gradient Average
Masks
CONV
CONV
CONV
$\frac{\partial y_c}{\partial \mathbf{z}}$
0 0
0 0
$y_c - \tilde{y}_c$
Apply mask only for Top-b percentile where it decreases confidence the most (e.g. only elephant)
Backpropagate Loss
$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\cdot, \cdot)$
. . .
e.g. PACS
e.g. ResNet