HUAWEI

# MixACM: Mixup-Based Robustness Transfer via Distillation of Activated Channel Maps

Awais Muhammad∗, Fengwei Zhou∗, Chuanlong Xie∗, Jiawei Li, Sung-Ho Bae, Zhenguo Li

Kyung-Hee University

**tl;dr:** a new method to transfer robustness from robust pre-trained models, without generating adversarial examples.

## Why Adversarial Robustness

Adversarial robustness is important for many reasons such as security, improvement in robustness to noise, learning better, more interpretable features, etc.

## Two Types of Prior Works

Adversarial Training Train model on adversarially augmented examples instead of clean ones.
**Pros:** more robust.
**Cons:** requires 1) more data, 2) more compute power, 3) larger models and 4) decreases clean accuracy.
Distillation-based Use labels or input gradients of a pre-trained teacher for robustness distillation.
**Pros**: Less compute
**Cons:** less robust, still requires extra gradient

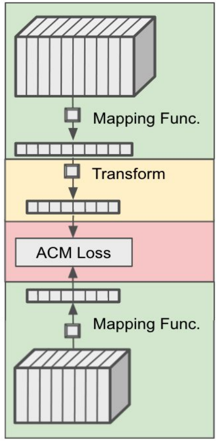
## Robustness Transfer via MixACM

We propose to transfer robustness via activated channel maps of a teacher, based on the hypothesis that activated channels are crucial for robustness.

Get activated channel maps of teacher and student.

Transform to match size

Minimize Loss between teacher and student activated channel maps along with other losses



## Theory: How to transfer robustness

**Result 1:** Students' adv. error can be bounded by it's adv. loss, adv. distillation loss between teacher and student, and distillation complexity.

Adversarial Error ≤ Adversarial Loss + Adv. Distillation Loss + Adv. Distillation Complexity

**Result 2:** adv. loss and adv. distillation is bounded by their normal counterparts on mixup examples.

Adversarial Loss + Adv. Distillation Loss ≤ Normal Loss on Mixup Examples + Distillation Loss on Mixup Examples

## Comparison with SOTA

WRN34-10 for CIFAR10

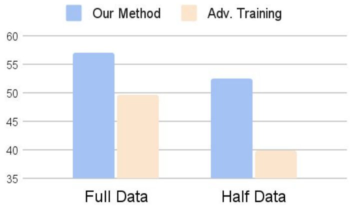
| Method | Acc. | Rob. |
|---|---|---|
| AT | 87.14 | 47.04 |
| AWP | 85.36 | 59.12 |
| FreeAT | 86.11 | 46.19 |
| TRADES | 84.92 | 56.43 |
| FAT | 84.52 | 54.32 |
| IGAM | 88.70 | 43.00 |
| Ours | **90.76** | **56.65** |

## ImageNet Results

ResNet50

| Method | Acc. | Rob. |
|---|---|---|
| Free-AT | 60.21 | 31.82 |
| Fast-AT | 55.45 | 31.39 |
| Ours | 62.05 | 33.63 |

## Less Data Results



* equal contribution