Stanford University

# C-Mixup: Improving Generalization in Regression

Huaxiu Yao[1*], Yiping Wang[2*], Linjun Zhang[3]
James Zou[1], Chelsea Finn[1]

[1]Stanford University, [2]Zhejiang University
[3]Rutgers University

# Mixup in Deep Learning

A learning model

$$\mathcal{D}_{tr} = \{x_i, y_i\}_{i=1}^{N} \rightarrow \text{Classifier},$$

Mixup (Zhang et al. 2018):

$$\widetilde{\mathcal{D}}_{tr} = \{\tilde{x}_i, \tilde{y}_i\}_{i=1}^{N} \rightarrow \text{Classifier},$$

where

$$\tilde{x}_i = \lambda x_i + (1-\lambda) x_j, \tilde{y}_i = \lambda y_i + (1-\lambda) y_j$$

$$\lambda \sim \text{Beta}(\alpha, \beta)$$

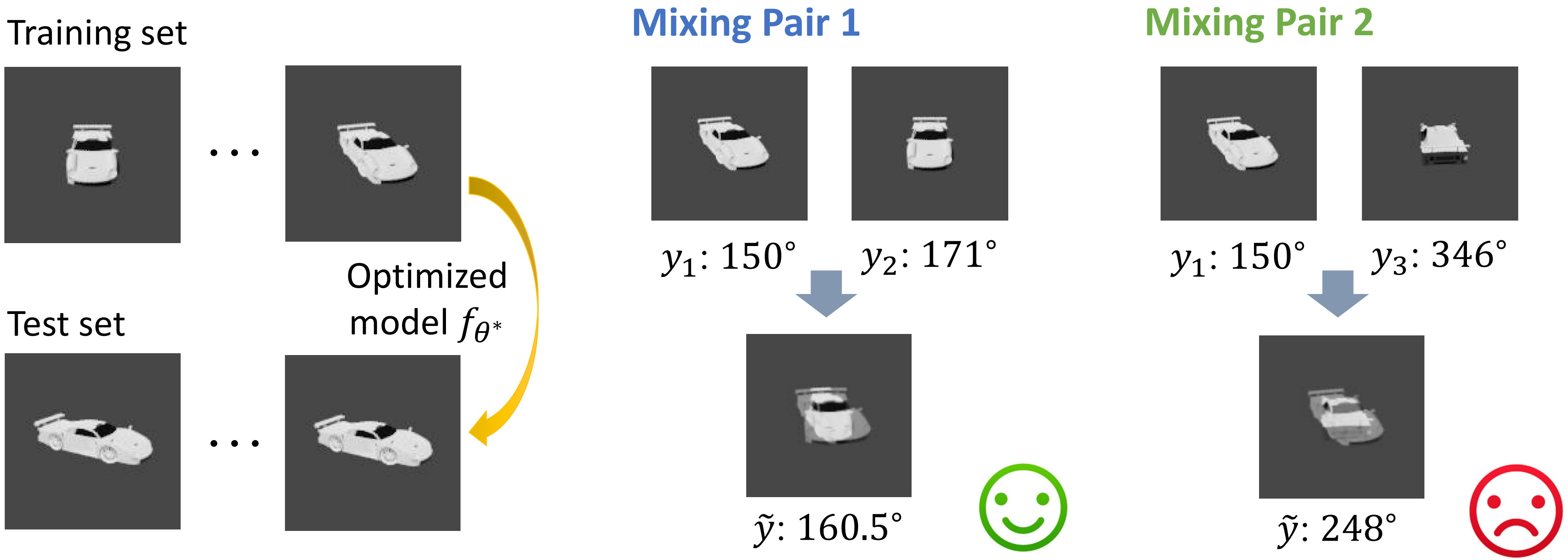
[1.0, 0.0] cat dog    [0.0, 1.0] cat dog    [0.7, 0.3] cat dog

Image credit: https://blog.airlab.re.kr/2019/11/mixup

# Why Mixup May Fail in Regression?

Directly applying mixup in Regression may produce arbitrary labels

Example: Pose Prediction



# How to Apply Mixup to Regression?

Solution: mixing examples with similar labels

Specifically, we change the sampling probability of mixing pairs

$$P((x_j, y_j)|(x_i, y_i)) \propto \exp\left(-\frac{d(i,j)}{2\sigma^2}\right)$$

d: distance between examples $i$ and $j$

Natural way: compute the distance using the input feature $x$

$$d(i,j) = d(x_i, x_j)$$

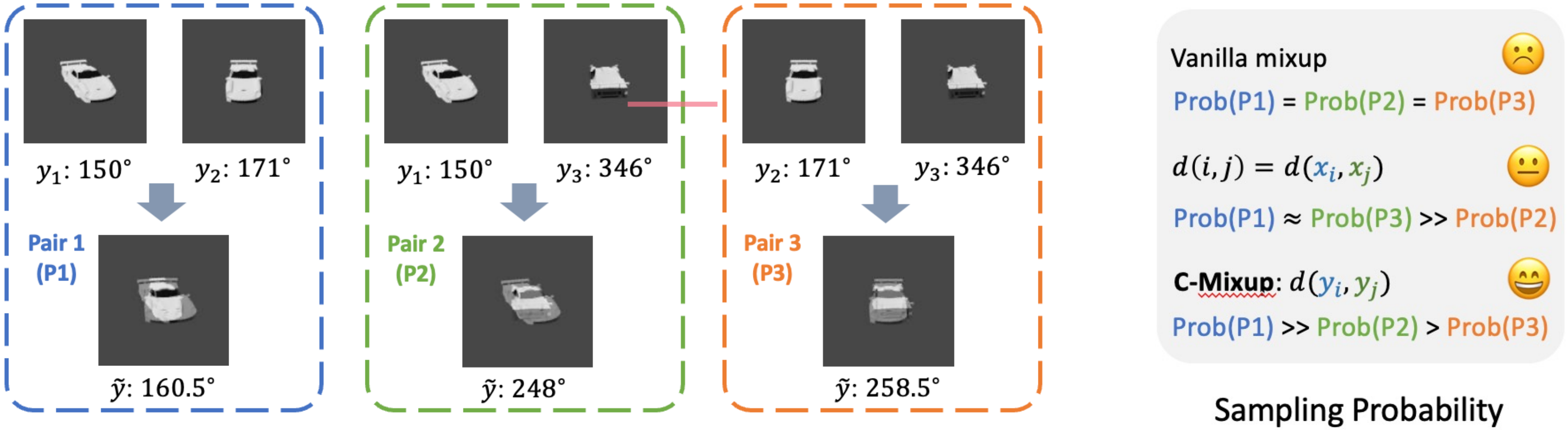
Drawbacks:

- Lacking good distance metrics to capture structured feature information for high dim data
- Distance between features can be easily influenced by feature noise

# C-Mixup

Examples with closer labels ➡ Higher probability to be mixed

$$d(i,j) = d(y_i, y_j)$$




Mixing Pairs ($\lambda = 0.5$)

Sampling Probability Comparison

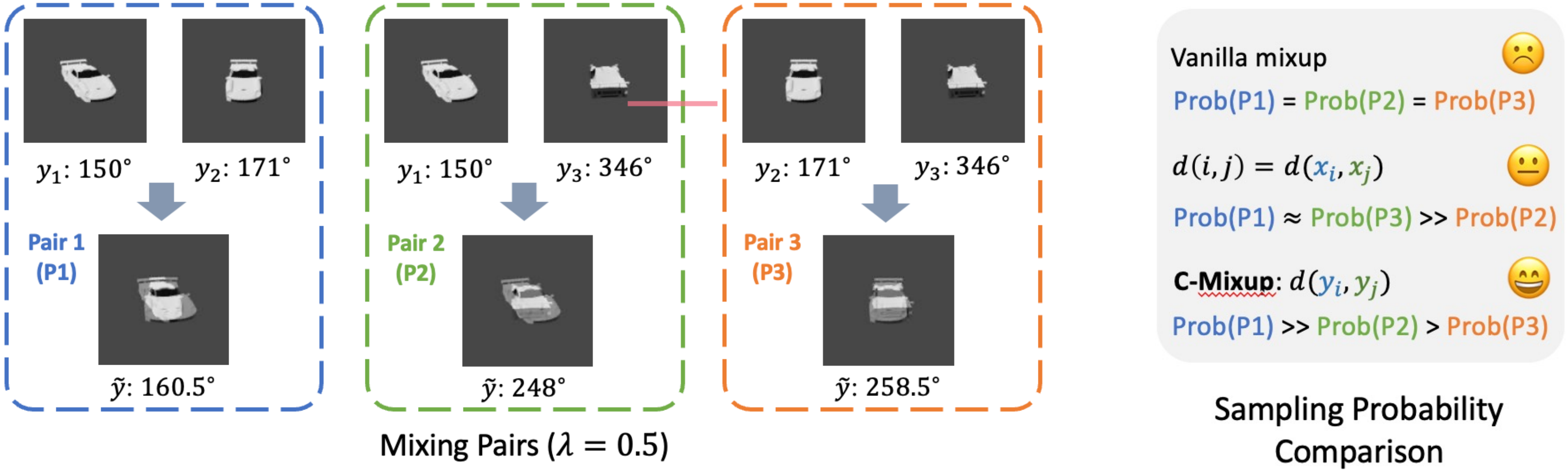
+ benefit both in-distribution and out-of-distribution generalization

+ calculating label distance is computationally efficient

# C-Mixup

Example pairs with closer labels ➡ Higher probability to be mixed

$$d(i,j) = d(y_i, y_j)$$




+ benefit both in-distribution and out-of-distribution generalization

Theory (linear and monotonic non-linear model)

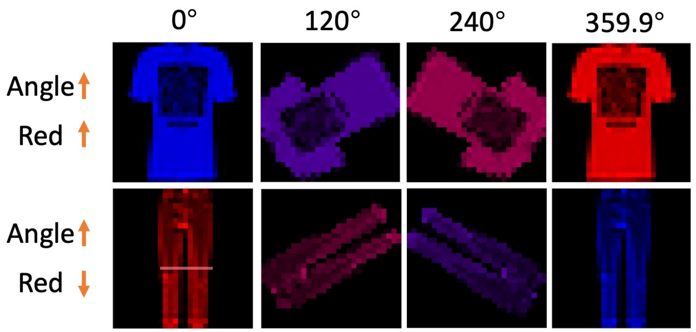
mean square error: C-Mixup < min(ERM, $d(x_i, x_j)$)

# How C-Mixup Performs? In-Distribution Generalization

| | | RMSE ↓ | | |
|---|---|---|---|---|
| | | ERM | Best mixup variant | **C-Mixup** |
| Tabular | Airfoil | 2.901 | 2.938 | **2.717** |
| | NO2 | 0.537 | 0.517 | **0.509** |
| Time-series | Exchange-Rate | 0.0236 | 0.0236 | **0.0203** |
| | Eletricity | 0.0581 | 0.0575 | **0.0570** |
| Video | Echocardiogram | 5.402 | 5.393 | **5.177** |

# How C-Mixup Performs? Out-of-Distribution Generalization

## Subpopulation shift – mitigate spurious correlation




| | Training | Test |
|---|---|---|
| Angle ↑ Red ↑ | 80% | 20% |
| Angle ↑ Red ↓ | 20% | 80% |

RCF-MNIST: angle spuriously correlates with labels

## Domain shift – generalize the model to new domains

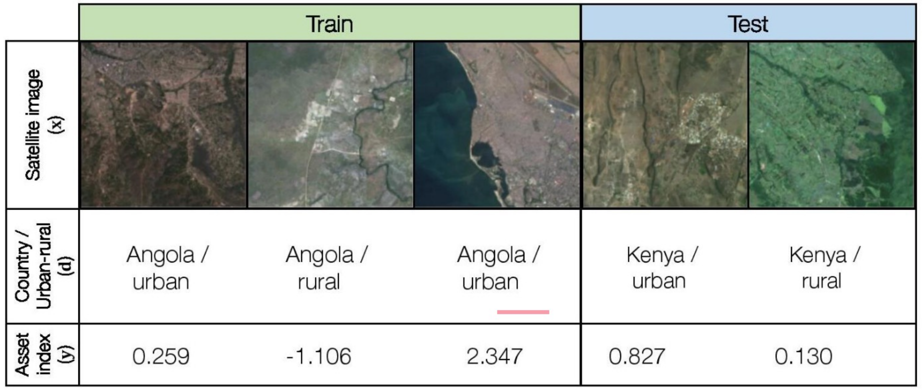



Image credit: https://wilds.stanford.edu/datasets/

# How C-Mixup Performs? Out-of-Distribution Generalization

| | | Worst Group/Domain | | |
|---|---|---|---|---|
| | | ERM | Best prior OOD method | **C-Mixup** |
| Image | RCF-MNIST (RMSE) ↓ | 0.162 | 0.153 | **0.146** |
| | PovertyMap (R) ↑ | 0.50 | 0.48 | **0.53** |
| Tabular | Crime (RMSE) ↓ | 0.173 | 0.152 | **0.146** |
| | SkillCraft (RMSE) ↓ | 10.182 | 7.444 | **7.362** |
| Drug | DTI (R) ↑ | 0.429 | 0.443 | **0.458** |

# Analysis

## Analysis I: different distance metrics

| | Feature/Repr Dist | C-Mixup |
|---|---|---|
| DTI | 0.477 | **0.498** |

## Analysis II: Batch-wise C-Mixup (Scalability)

| | Dataset | Airfoil | NO2 | Exchange-Rate | Electricity |
|---|---|---|---|---|---|
| RMSE $\downarrow$ | C-Mixup-batch | $2.792 \pm 0.135$ | $0.510 \pm 0.007$ | $0.0205 \pm 0.0017$ | $0.0576 \pm 0.0002$ |
| | C-Mixup | $2.717 \pm 0.067$ | $0.509 \pm 0.006$ | $0.0203 \pm 0.0011$ | $0.0570 \pm 0.0006$ |
| MAPE $\downarrow$ | C-Mixup-batch | $1.616 \pm 0.053\%$ | $12.894 \pm 0.180\%$ | $2.064 \pm 0.218\%$ | $13.697 \pm 0.155\%$ |
| | C-Mixup | $1.610 \pm 0.085\%$ | $12.998 \pm 0.271\%$ | $2.041 \pm 0.134\%$ | $13.372 \pm 0.106\%$ |

# Takeaways

- Simple strategy for data interpolation in Regression
- Mixing examples with closer continuous labels avoids producing arbitrary mixed labels

Code: https://github.com/huaxiuyao/C-Mixup

Thanks

Q & A