ICLR
International Conference On
Learning Representations

MixPro: Data Augmentation with MaskMix and Progressive Attention Labeling for Vision Transformer


Qihao Zhao


Yangyu Huang


Wei Hu


Fan Zhang

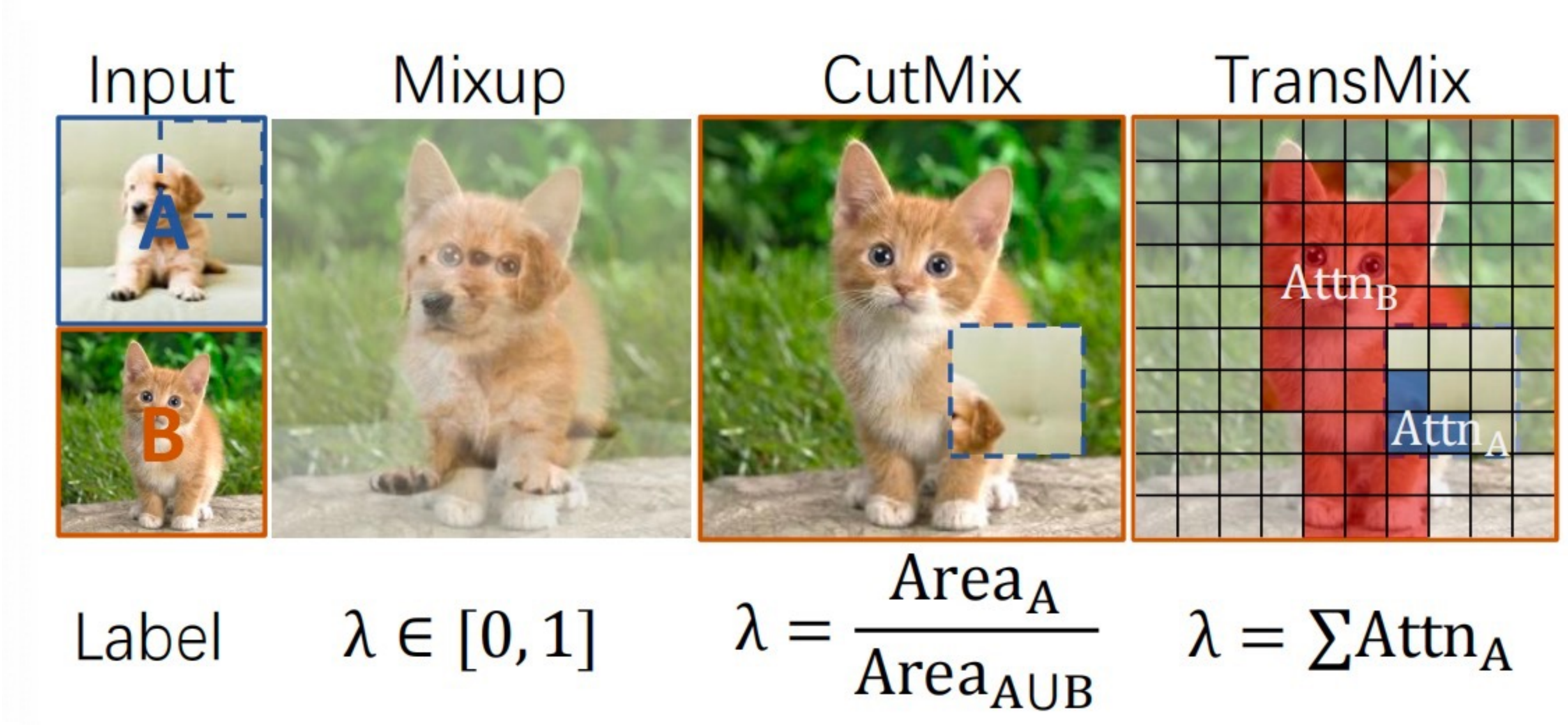

Jun Liu

Beijing University of Chemical Technology

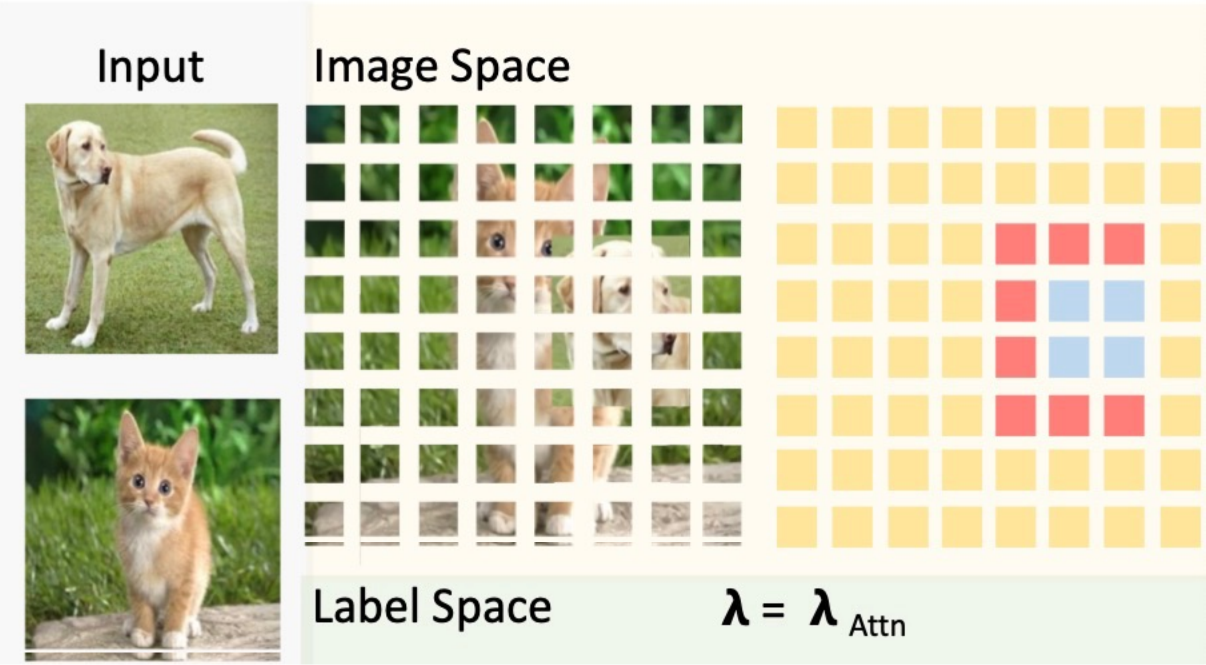
Microsoft Research

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Previous Mixed augmentation methods for Vision Transformer



Shortcomings of TransMix



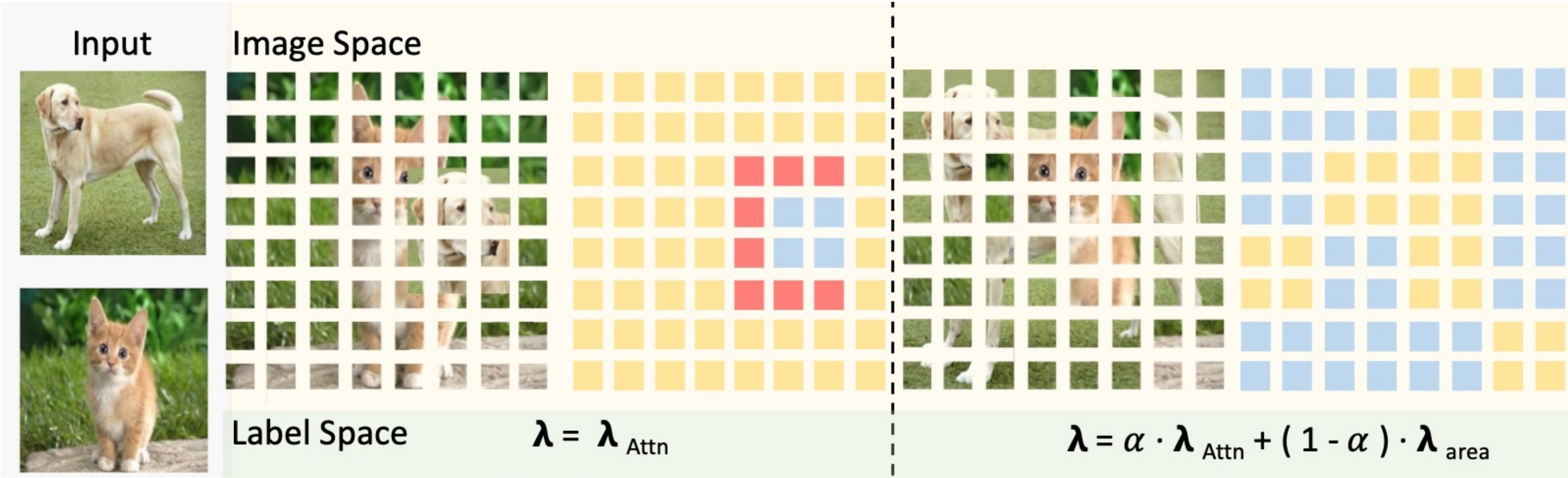
- ViTs has long-range dependence, region-based mixed images may provide insufficient regularization.

- Cropped patches with sharp rectangular borders are clearly distinguishable from the background (viewed as red patches), resulting in a basis weight of attention regardless of whether the patch contains useful information.

Attention maps may not always be reliable during the training process.

- At the beginning of the training, the model has no representation capability, and the attention maps gained are unreliable.
- it is possible to obtain difficult samples using massive data augmentation strategies, and the attention map is also unreliable.

MixPro



(a) TransMix (b) MixPro

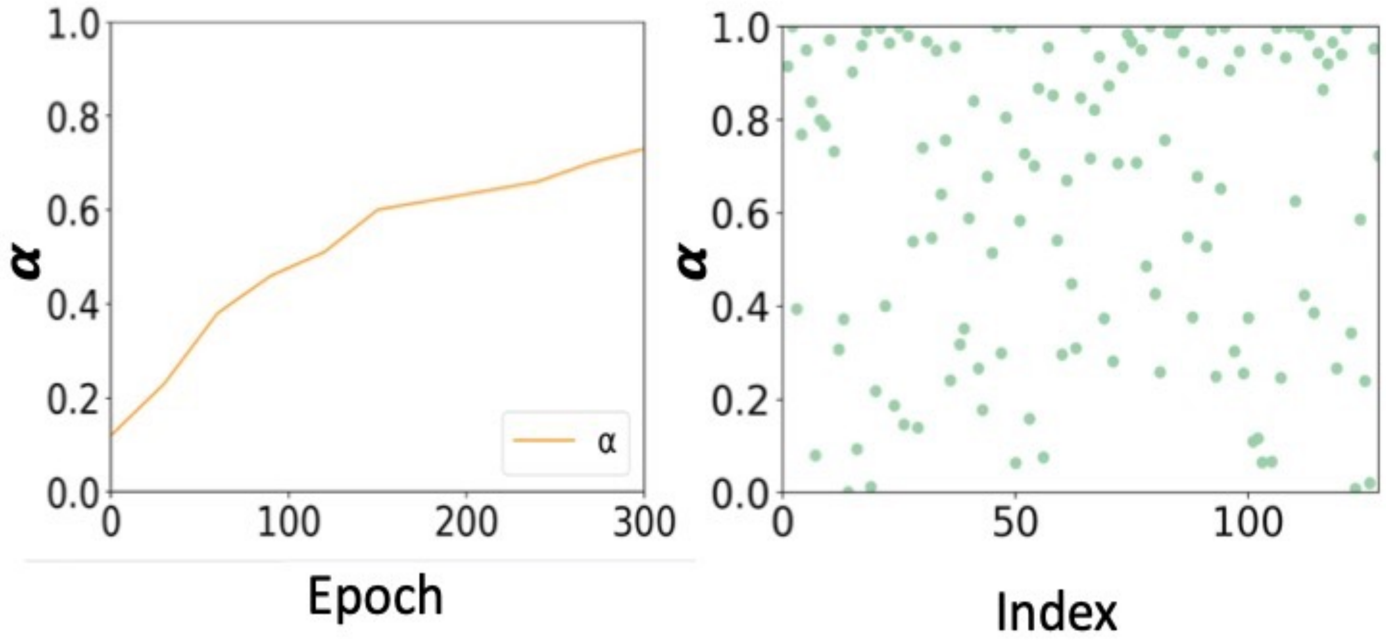
How to get progressive factor $\boldsymbol{\alpha}$?

Setp 1

$$\widetilde{y} = \lambda_{area} \odot y_i + (\mathbf{1} - \lambda_{area}) \odot y_j$$

Setp 2

$$\boldsymbol{\alpha} = \mathbf{d}(\mathbf{p}, \widetilde{\mathbf{y}}) = \frac{\mathbf{p} \cdot \widetilde{\mathbf{y}}^{\top}}{||\mathbf{p}|| \cdot ||\widetilde{\mathbf{y}}||},$$



(a) Trends of $\boldsymbol{\alpha}$

(b) $\boldsymbol{\alpha}$ in a mini-batch

Experiments

ICLR International Conference On Learning Representations

Table 1: Compared to TransMix, MixPro provides better performance on a wide range of model variants, e.g. DeiT, PVT, CaiT, XCiT , Swin on ImageNet-1k classification. All the baselines are reported in TransMix (Chen et al., 2021).

Models	Params	#FLOPs	Top-1 Acc(%)	Top-1 Acc(%) +TransMix	Top-1 Acc(%) +MixPro
DeiT-T (Touvron et al., 2021a)	5.7M	1.6G	72.2	72.6	73.8+(+1.2)
PVT-T (Wang et al., 2021)	13.2M	1.9G	75.1	75.5	76.7+(+1.2)
XCiT-T (Ali et al., 2021)	12M	2.3G	79.4	80.1	81.2+(+1.1)
CA-Swin-T (Liu et al., 2021)	28.3M	4.2G	81.6	81.8	82.8+(+1.0)
CaiT-XXS	17.3M	3.8G	79.1	79.8	80.6+(+0.8)
DeiT-S (Touvron et al., 2021a)	22.1M	4.7G	79.8	80.7	81.3+(+0.6)
PVT-S (Wang et al., 2021)	24.5M	3.8G	79.8	80.5	81.2+(+0.7)
XCiT-S (Ali et al., 2021)	26M	4.8G	82.0	82.3	82.9+(+0.6)
CA-Swin-S (Liu et al., 2021)	49.6M	8.5G	82.8	83.2	83.7+(+0.5)
PVT-M (Wang et al., 2021)	44.2M	6.7G	81.2	82.1	82.7+(+0.6)
PVT-L (Wang et al., 2021)	61.4M	9.8G	81.7	82.4	82.9+(+0.5)
XCiT-M (Ali et al., 2021)	84M	16.2G	82.7	83.4	84.1+(+0.7)
DeiT-B (Touvron et al., 2021a)	86.6M	17.6G	81.8	82.4	82.9+(+0.5)
XCiT-L (Ali et al., 2021)	189M	36.1G	82.9	83.8	84.7+(+0.9)

Experiments

pretrained	Backbone	Decoder	mIoU	+MS
ResNet101	ResNet101	Deeplabv3+	47.3	48.5
DeiT-S			49.1	49.6
+TransMix	DeiT-S	Linear	49.7	50.3
+MixPro			**50.3**	**50.9**
DeiT-S			49.7	50.5
+TransMix	DeiT-S	Segmenter	50.6	51.2
+MixPro			**51.1**	**51.6**

Semantic Segmentation

Backbone	#Params	Object detection			Instance segmentation		
		AP^b	AP^b_{50}	AP^b_{75}	AP_m	AP^m_{50}	AP^m_{75}
ResNet50	44.2M	38.0	58.6	41.4	34.4	57.1	36.7
ResNet101	63.2M	40.4	61.1	44.2	36.4	57.7	38.8
PVT-S	44.1M	40.4	62.9	43.8	37.8	60.1	40.3
TransMix-PVT-S	44.1M	40.9	63.8	44.0	38.4	60.7	41.3
MixPro-PVT-S	44.1M	**41.4**	**64.2**	**44.4**	**38.9**	**61.1**	**41.7**

Objection detection and Instance Segmentation

Summary

- We propose a new data augmentation method, MixPro, to address the shortcomings of TransMix from the perspective of image space and label space, respectively.
- From the perspective of image space, MixPro ensures that each image patch comes from only one image and uses a global mixed mask to provide more regularization. From the perspective of label space, MixPro utilizes a progressive factor to dynamically re-weight the attention weight of the mixed attention label.
- In experiments, we demonstrate extensive evaluations of MixPro on various ViT-based models and downstream tasks. It boosts Deit-T achieving 73.8% on ImageNet-1K. Furthermore, compared to TransMix, MixPro also shows stronger robustness on three different benchmarks.

Code link: https://github/fistyee/MixPro

Thank you