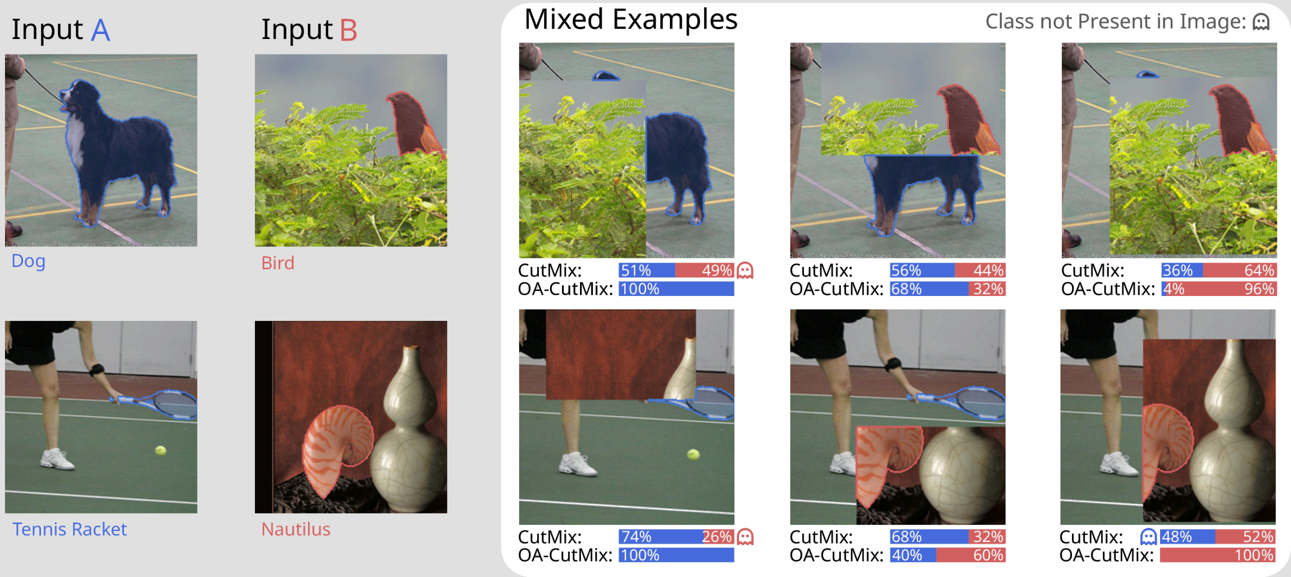
OA-CutMix: Correcting the Label Bias of CutMix



Tobias Nauen · Stanislav Frolov · Federico Raue · Brian Moser · Andreas Dengel

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau
dfki Deutsches Forschungszentrum für Künstliche Intelligenz German Research Center for Artificial Intelligence
Carl Zeiss Stiftung
SUSTAINABLE EMBEDDED AI

ENNS
ICANN26

Data augmentation creates new training data from existing images.

> Data Augmentation

→

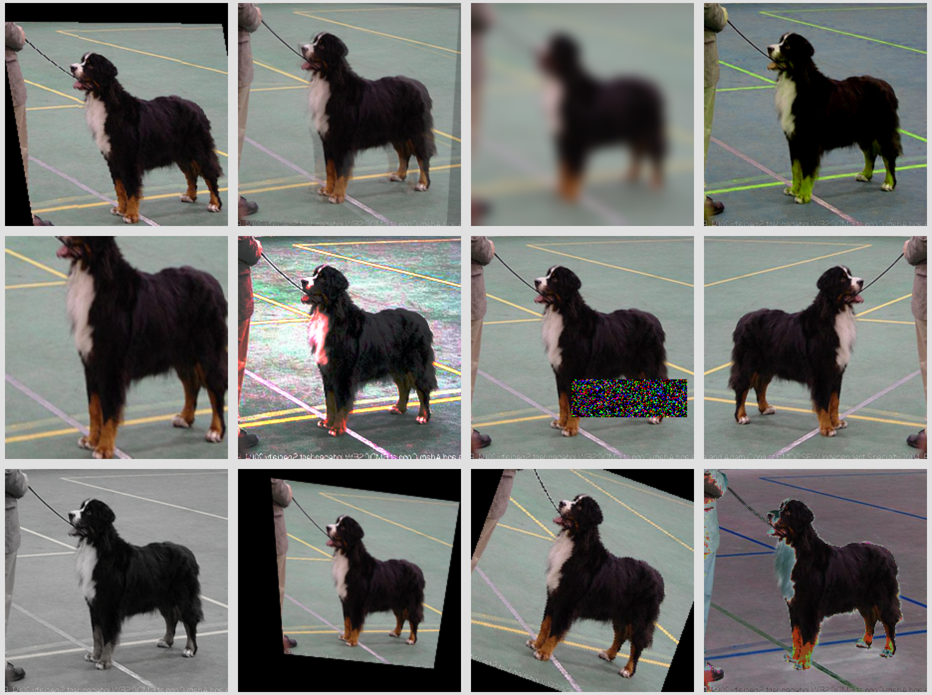
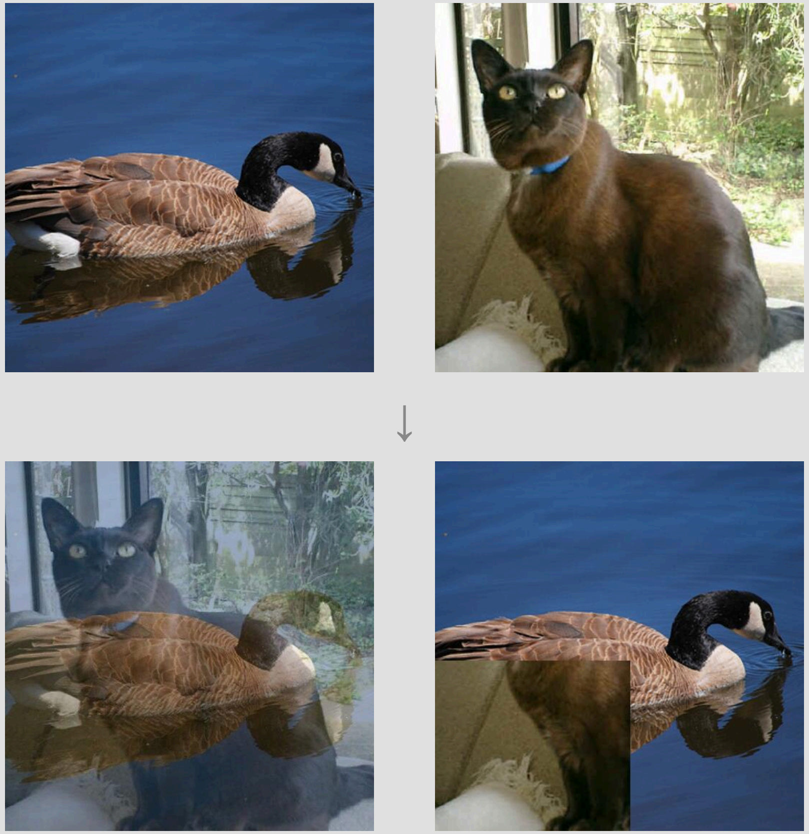


Image datasets are limited
⇒ Prevent overfitting

Encode task's inductive biases

Label stay the same

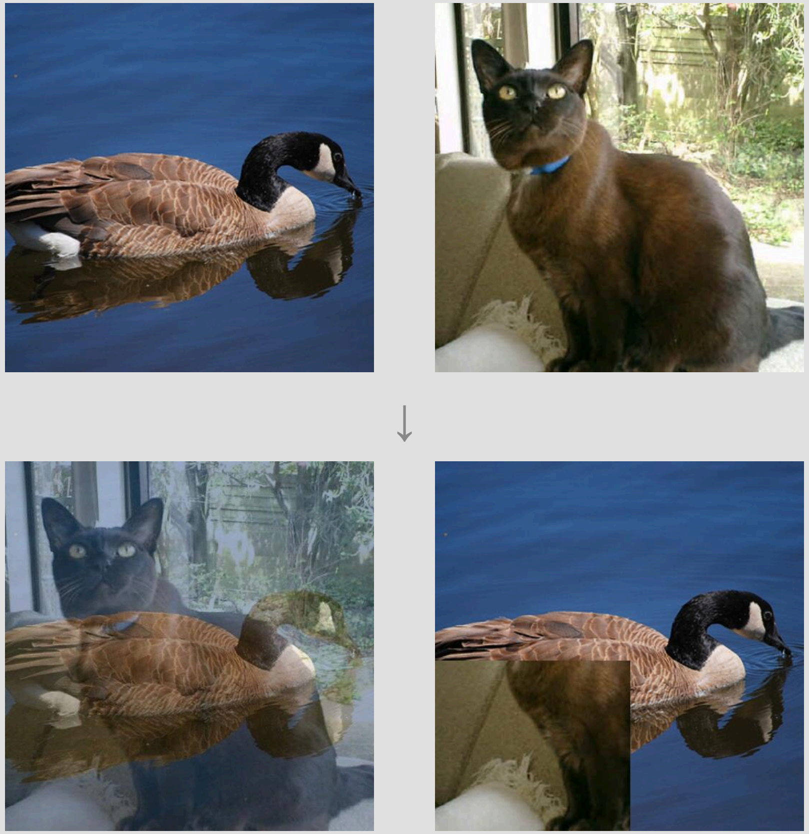
Mixing augmentations combines the semantic content of two images.



MixUp CutMix

- Mix two different images

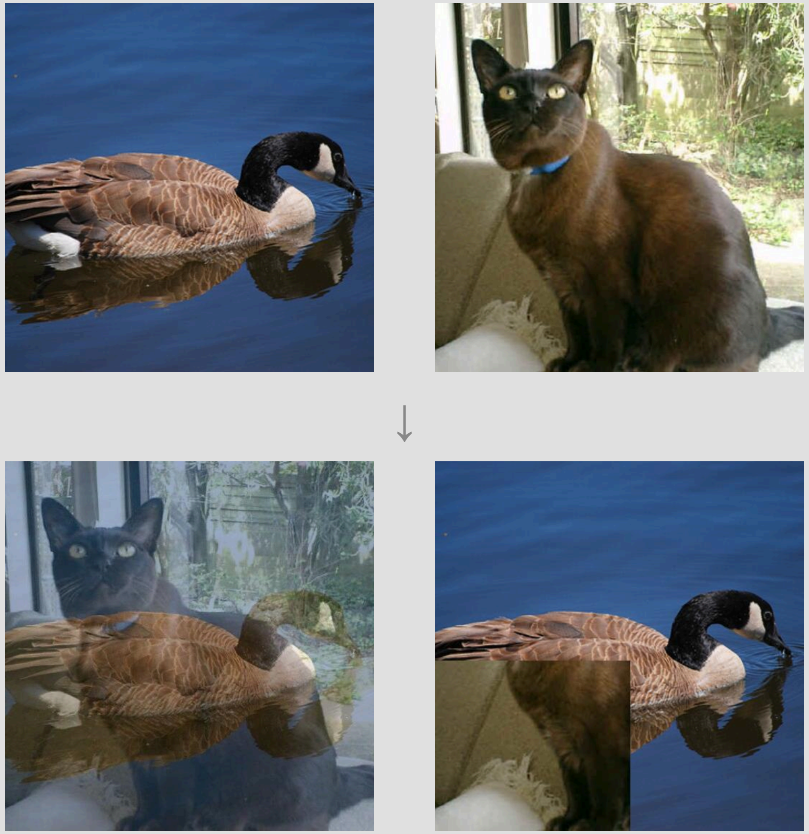
Mixing augmentations combines the semantic content of two images.



MixUp CutMix

- Mix two different images
- Label is adjusted to combined image

Mixing augmentations combines the semantic content of two images.



MixUp CutMix

Mix two different images

Label is adjusted to combined image

Static methods mix independent of content.

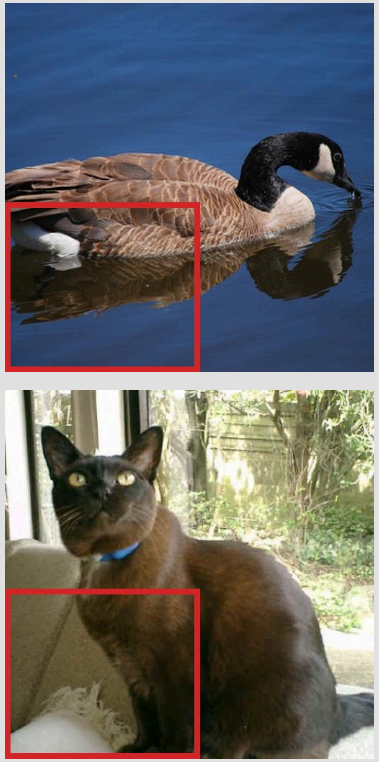
Examples: CutMix, MixUp, Fmix, ResizeMix

Dynamic methods adjust the mixing based on the image contents.

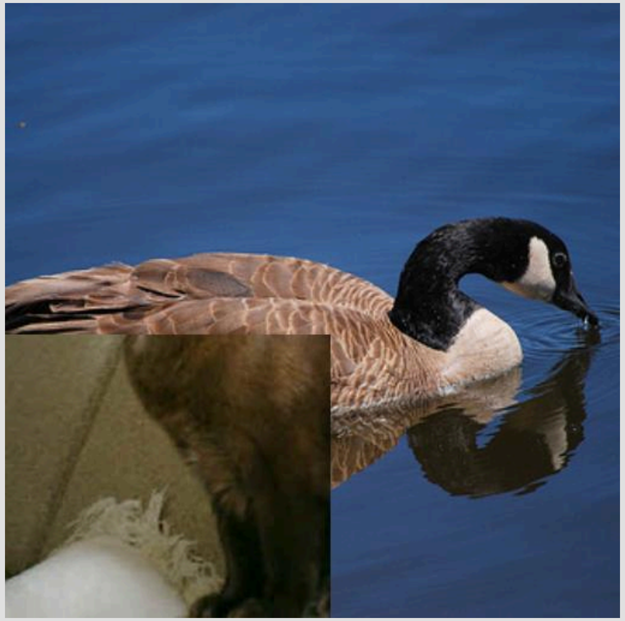
Examples: Attentive CutMix, SAMix, PuzzleMix

CutMix uses image area to set the combined label.

> Data Augmentation



→

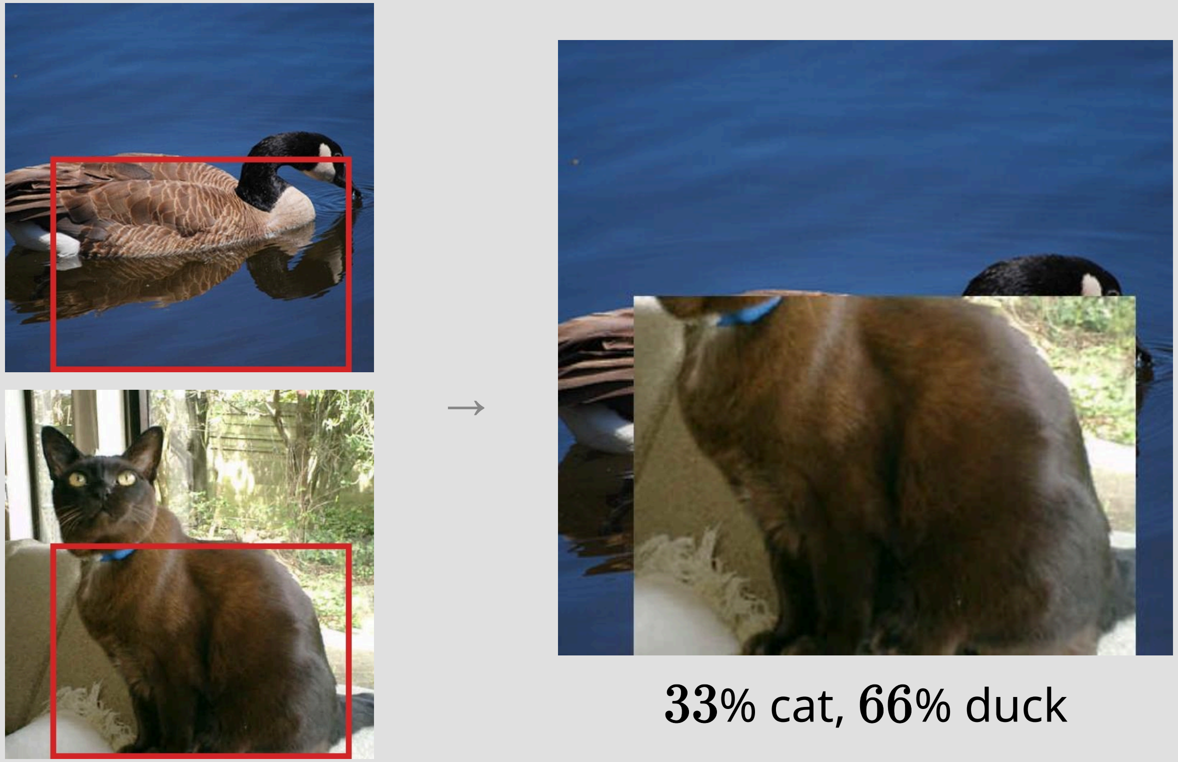


20% cat, 80% duck

$$\hat{y} = (1-\lambda)y_A + \lambda y_B, \qquad \lambda = \text{patch area}/\text{image area}$$

The CutMix label counts background as relevant area.

> Label Misalignment

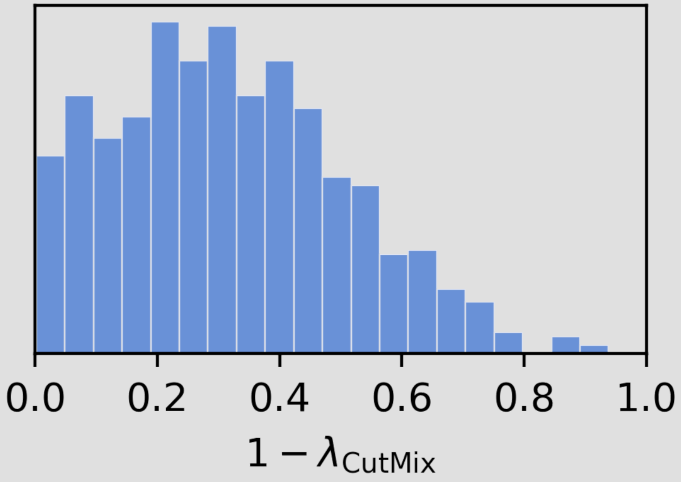


33% cat, 66% duck

CutMix considers background as relevant area even when objects are cropped out.

Labels are misaligned by 21.5 percentage points on average.

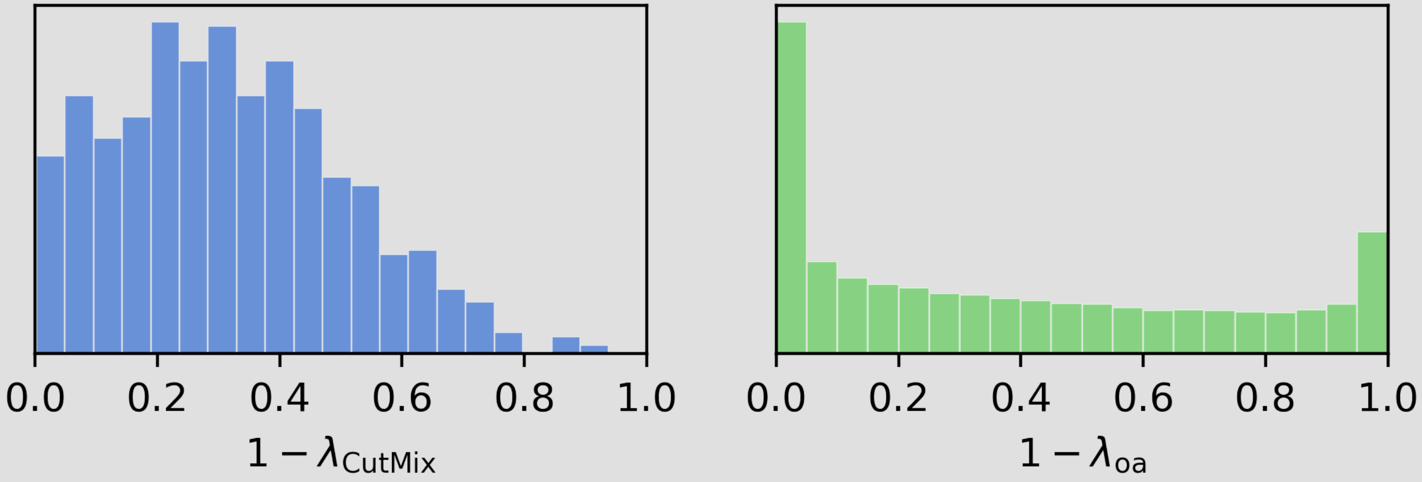
> Label Misalignment



λ_{oa} is the relative number of object pixels, i.e. an object-aware label.

Labels are misaligned by 21.5 percentage points on average.

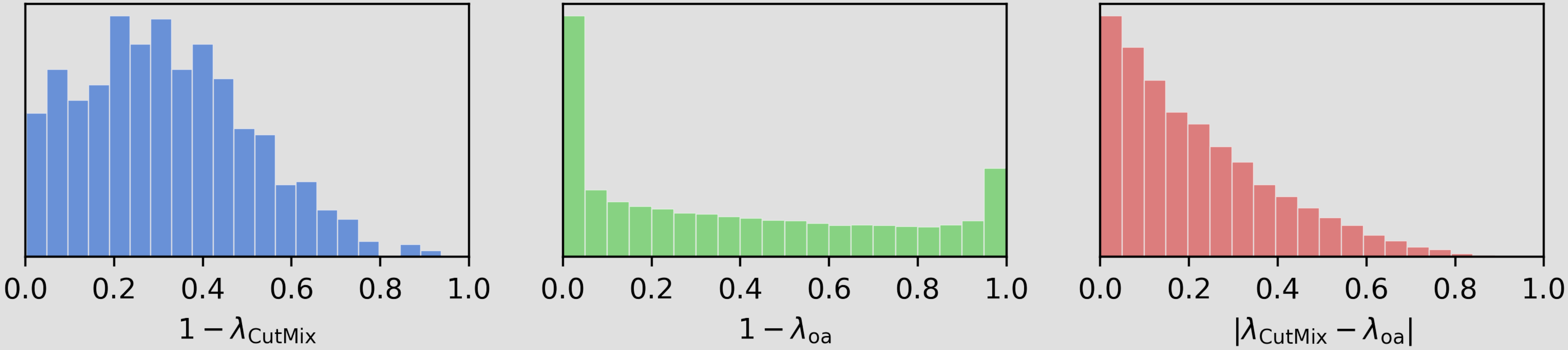
> Label Misalignment



λ_{oa} is the relative number of object pixels, i.e. an object-aware label.

Labels are misaligned by 21.5 percentage points on average.

> Label Misalignment



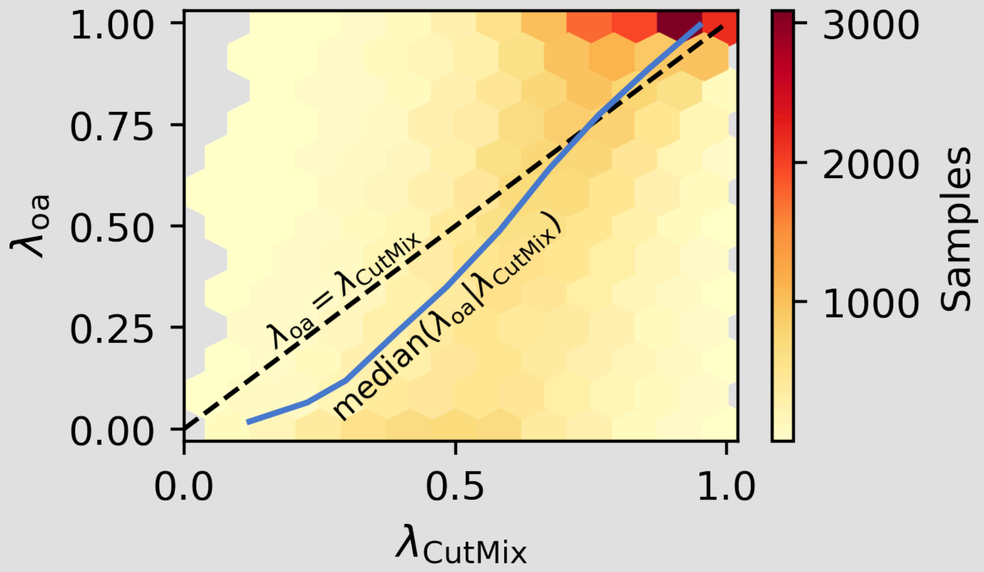
λ_{oa} is the relative number of object pixels, i.e. an object-aware label.

Object pixels give more extreme labels than image area.

Mean absolute difference $|\lambda_{\mathrm{CutMix}} - \lambda_{\mathrm{oa}}| = 21.5$ percentage points.

Label misalignment is largest for small objects.

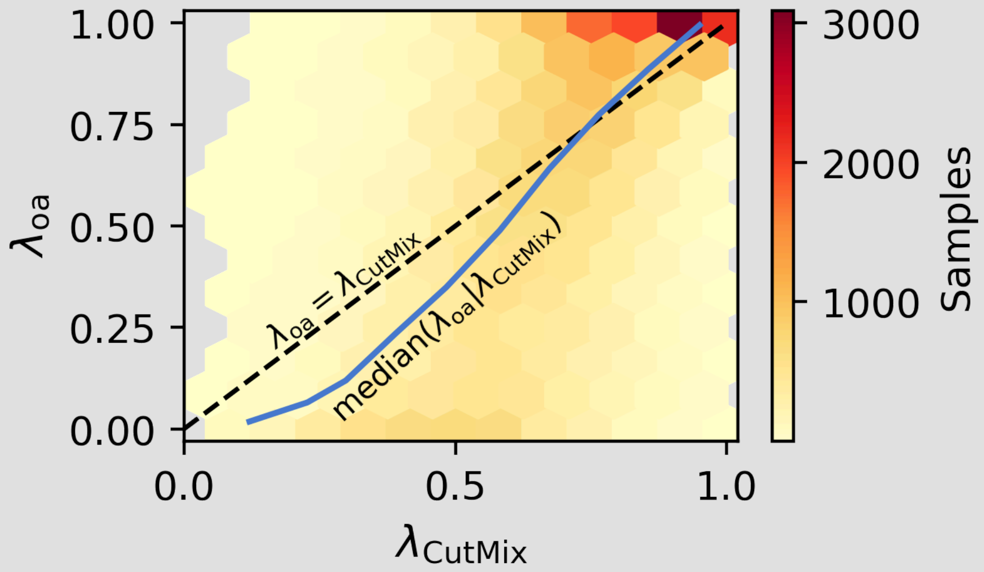
> Label Misalignment

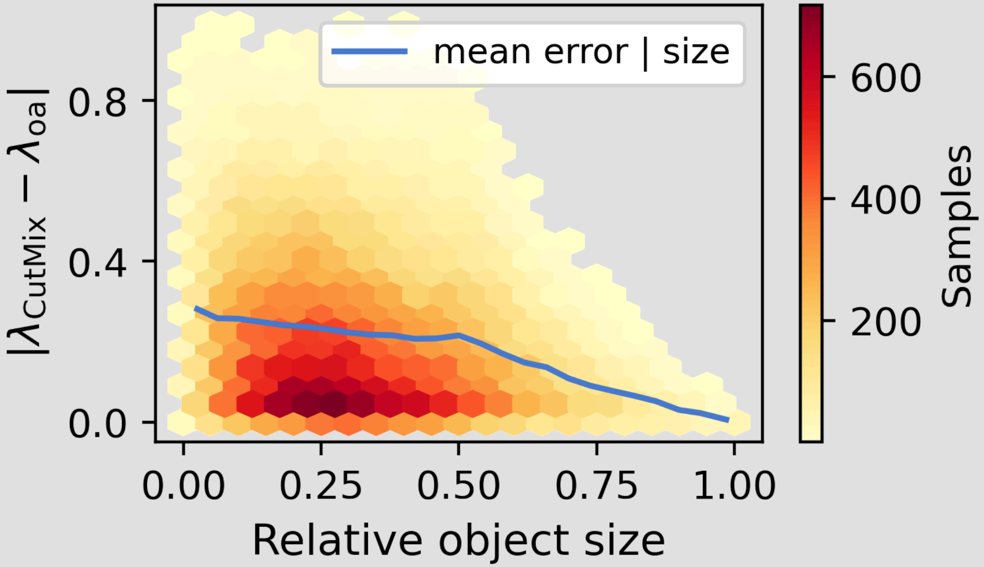


CutMix has systematic error towards less extreme labels for small and large patches.

Label misalignment is largest for small objects.

> Label Misalignment



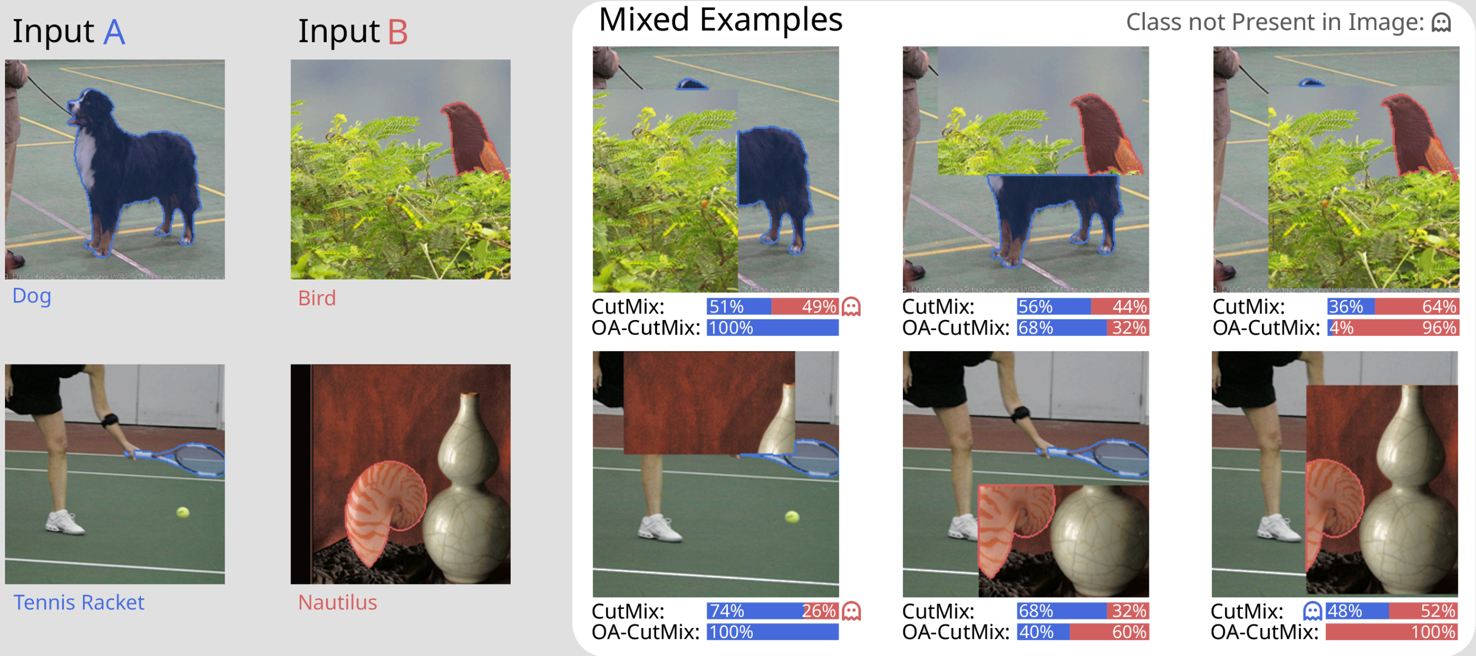


CutMix has systematic error towards less extreme labels for small and large patches.

Small objects are easy to miss/cover.
⇒ Larger error for smaller objects.

A label naming an absent class is a ghost label.

> Label Misalignment

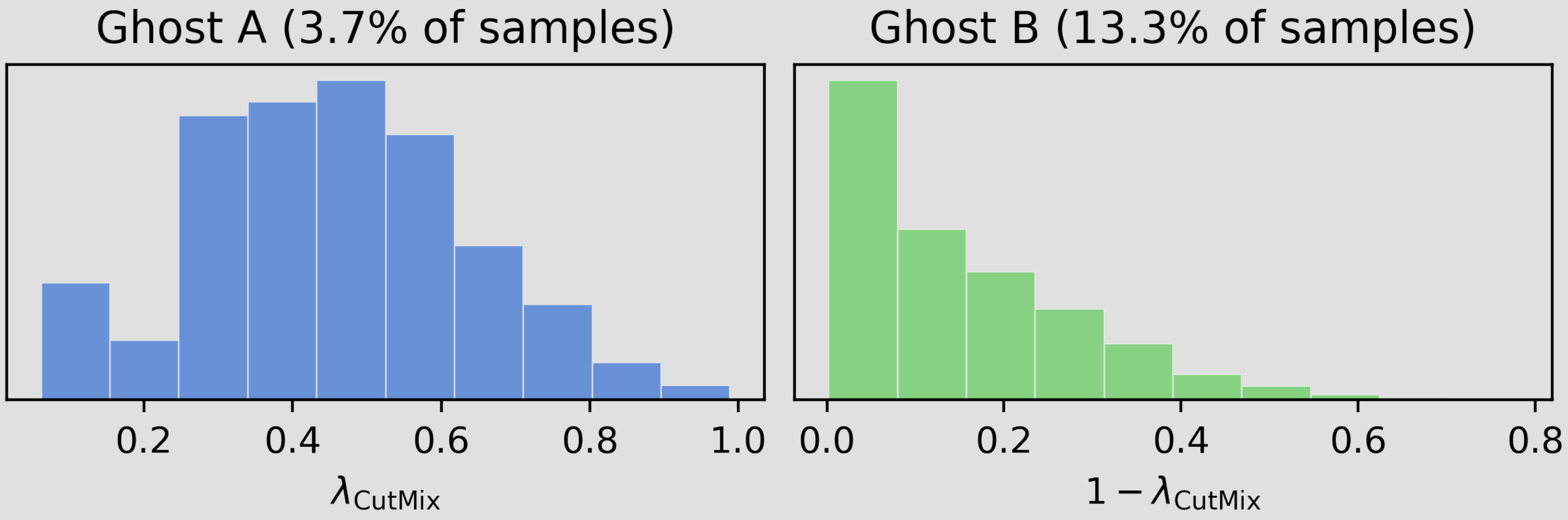


Definition (Ghost Label):

Let x be an image with label y. We call y a **ghost label** for x if and only if y labels a class with weight > 0 that has **no visual evidence** in x.

Ghost labels occur in 17% of all samples.

> Label Misalignment



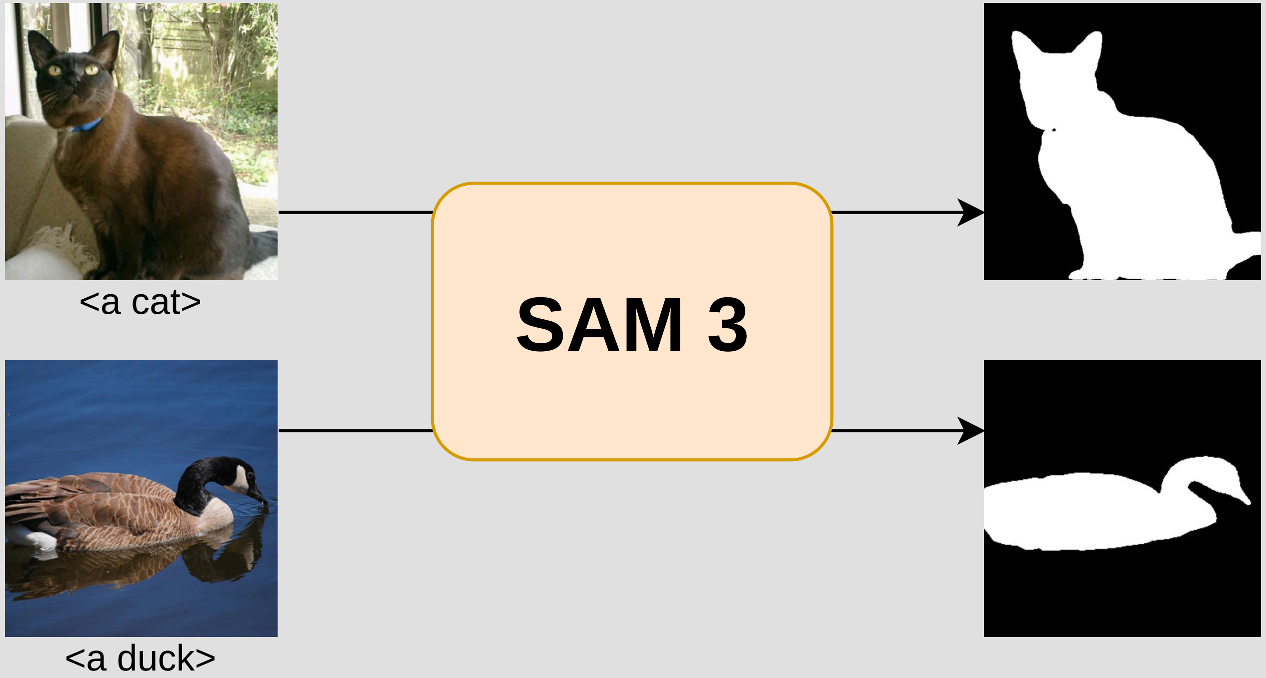
Mean error for **Ghost A** (object covered by crop) 45.4 p.p.

Mean error for **Ghost B** (object missed by crop) 15.3 p.p.

Ghost labels occur for $\approx 17\%$ of all samples (ImageNet + RandAugment).
For the smallest 10% of objects, the ghost label rate is $> 40\%$.

A segmentation model tells us what’s actually visible.

> OA-CutMix



SAM3 for detecting object boundaries with prompt:
`” a <label>, <hypernym> ``.`

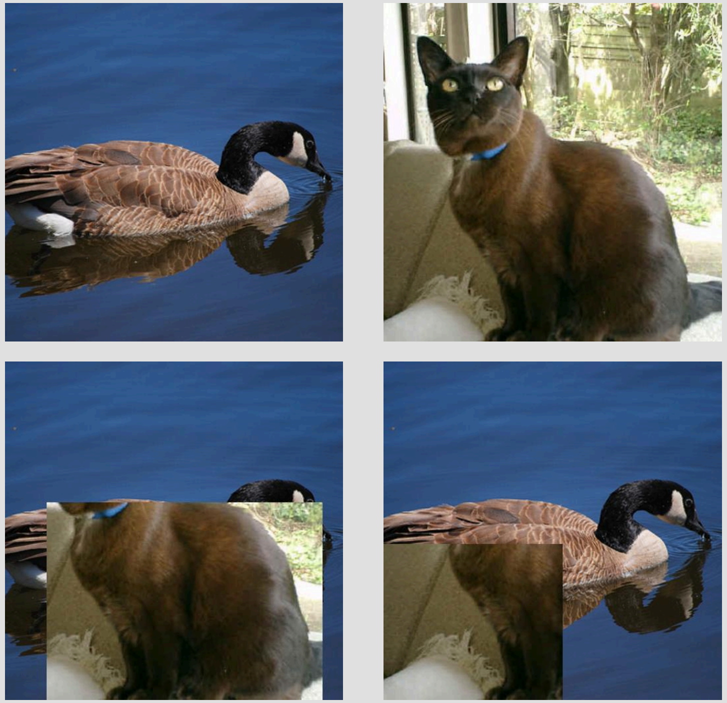
Include all detections above threshold t^*.

$$t^*(x) = \max(0.01, \min(t_{\max}(x), 0.5))$$

OA-CutMix labels each image by visible object area.

> OA-CutMix

Let $M_{\mathcal{B}} \in \{0,1\}^{H \times W}$ be the mask for the CutMix patch and $\mathbf{m}_{A/B}$ be the object masks for Image A/B from SAM3.



CutMix:

33% cat, 66% duck   20% cat, 80% duck

OA-CutMix labels each image by visible object area.

> OA-CutMix

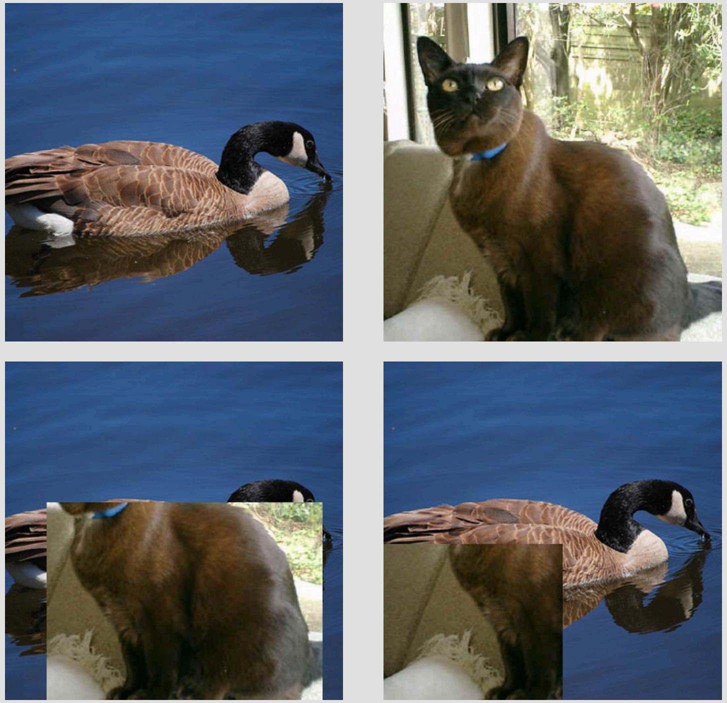
Let $M_{\mathcal{B}} \in \{0,1\}^{H \times W}$ be the mask for the CutMix patch and $\mathbf{m}_{A/B}$ be the object masks for Image A/B from SAM3.

Absolute Mode counts the number of object pixels from each class.

$$s_A := ||(1 - M_{\mathcal{B}}) \odot \mathbf{m}_A||_1$$

$$s_B := ||M_{\mathcal{B}} \odot \mathbf{m}_B||_1$$

$$\hat{y}_{oa} = \frac{s_A}{s_A + s_B} y_A + \frac{s_B}{s_A + s_B} y_B$$



CutMix:
33% cat, 66% duck 20% cat, 80% duck

OA-CutMix/abs:
89% cat, 11% duck 43% cat, 57% duck

OA-CutMix labels each image by visible object area.

> OA-CutMix

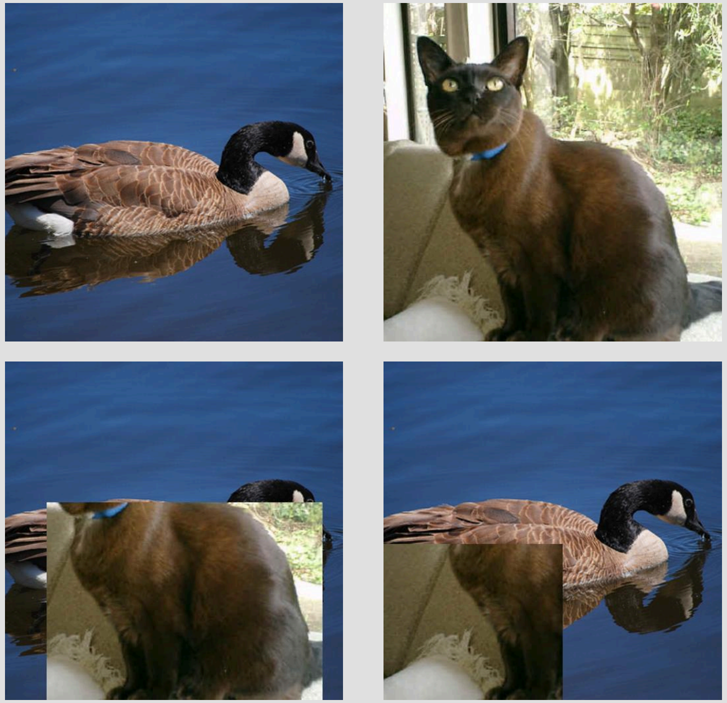
Let $M_{\mathcal{B}} \in \{0,1\}^{H\times W}$ be the mask for the CutMix patch and $\mathbf{m}_{A/B}$ be the object masks for Image A/B from SAM3.

Relative Mode normalizes the object pixels per image.

$$s_A := \frac{||(1 - M_{\mathcal{B}}) \odot \mathbf{m}_A||_1}{||\mathbf{m}_A||_1}$$

$$s_B := \frac{||M_{\mathcal{B}} \odot \mathbf{m}_B||_1}{||\mathbf{m}_B||_1}$$

$$\hat{y}_{oa} = \frac{s_A}{s_A + s_B} y_A + \frac{s_B}{s_A + s_B} y_B$$



CutMix:
33% cat, 66% duck   20% cat, 80% duck

OA-CutMix/abs:
89% cat, 11% duck   43% cat, 57% duck

OA-CutMix/rel:
80% cat, 20% duck   27% cat, 72% duck

OA-CutMix reaches dynamic-method accuracy at static-method speed.

Method	CIFAR-100		ImageNet200		TinyImageNet		Mean	latency
	DeiT-Ti	DeiT-S	DeiT-Ti	DeiT-S	DeiT-Ti	DeiT-S		[ms]
SnapMix [7]	76.17	76.16	81.09	82.86	54.39	56.09	71.13	39.98
SAMix [11]	79.72	79.48	81.24	82.12	55.46	57.91	72.65	88.27
PuzzleMix [8]	78.32	77.27	81.22	82.75	56.96	60.81	72.89	273.28
TokenMix [14]	78.42	78.82	80.03	82.66	60.10	60.49	73.42	39.91
Attentive CutMix [28]	79.49	79.92	81.76	83.70	59.95	56.57	73.56	56.10
Vanilla	66.46	66.91	73.32	74.56	42.83	46.18	61.71	
ResizeMix [19]	71.85	70.80	79.72	79.68	50.50	51.45	67.33	1.05
Mixup [31]	72.58	73.69	77.69	79.18	49.62	53.71	67.74	0.73
FMix [5]	76.30	73.76	80.56	81.35	54.41	52.83	69.87	6.80
Mixup [31] + CutMix [30]	78.75	77.99	81.24	82.51	56.74	60.36	72.93	0.70
CutMix [30]	79.83	79.14	81.78	83.43	61.08	58.86	74.02	0.67
CutMix + decoupled [15]	<u>79.93</u>	<u>80.32</u>	<u>81.90</u>	<u>83.84</u>	<u>61.65</u>	<u>61.53</u>	<u>74.86</u>	0.67
OA-CutMix/rel	**80.58**	**80.42**	**82.56**	**84.18**	**62.82**	**64.94**	**75.92**	1.03

OA-CutMix clearly outperforms all other mixing methods. Even dynamic methods are outperformed at a fraction of the training-time cost.

OA-CutMix reaches dynamic-method accuracy at static-method speed.

> Results

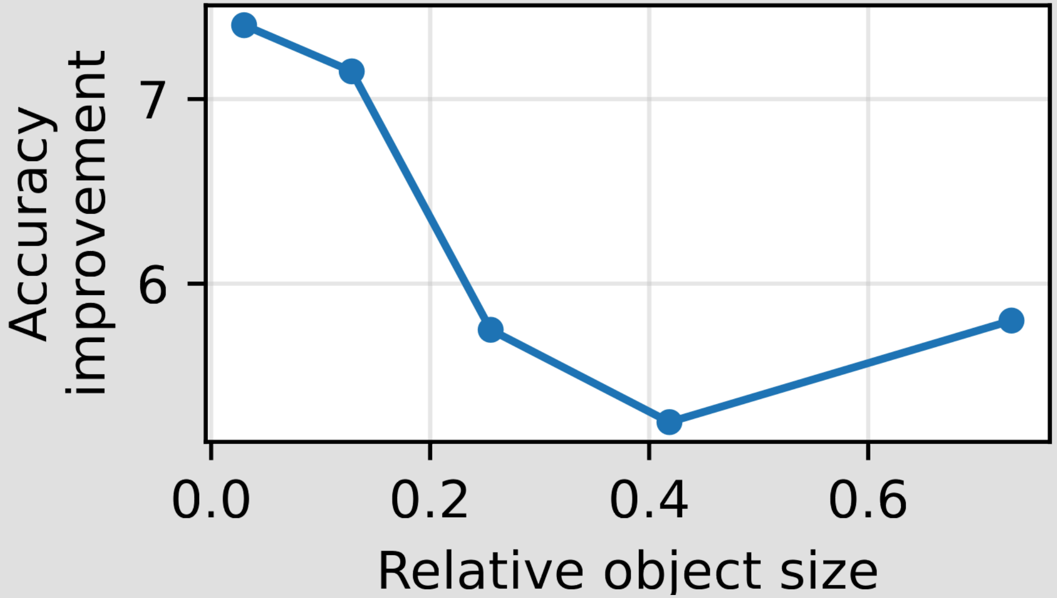
Method	CIFAR-100 ResNet18	CIFAR-100 ResNet50	ImageNet200 ResNet18	ImageNet200 ResNet50	TinyImageNet ResNet18	TinyImageNet ResNet50	Mean	latency [ms]
PuzzleMix [8]	**80.90**	81.20	84.29	87.91	57.71	65.37	76.23	301.00
SAMix [11]	80.89	**84.20**	55.23	59.96	57.06	60.10	66.24	16.68
SnapMix [7]	79.81	79.18	84.07	87.57	59.25	64.64	75.75	51.49
Attentive CutMix [28]	79.34	79.13	84.07	87.54	**60.75**	**66.14**	76.16	56.36
Vanilla	77.27	78.09	83.17	86.81	50.11	57.37	72.14	
Mixup [31] + CutMix [30]	78.64	79.08	83.80	87.52	53.71	59.92	73.78	0.67
Mixup [31]	79.09	79.76	83.36	87.74	54.08	60.24	74.05	0.73
alignmix	80.52	81.70	83.73	87.56	54.34	59.56	74.57	1.98
FMix [5]	78.92	79.11	83.82	87.47	57.75	62.14	74.87	6.68
ResizeMix [19]	80.13	79.81	**84.41**	88.01	56.04	62.28	75.11	0.62
CutMix [30]	79.26	79.34	84.08	87.74	59.43	61.79	75.27	0.61
CutMix + decoupled [15]	79.09	79.27	84.29	87.64	59.88	62.41	75.43	0.60
OA-CutMix/rel	79.79	81.41	84.31	**88.19**	59.49	64.48	**76.28**	1.02

OA-CutMix is the best overall method with some dynamic methods winning for certain setups but at way more training-time compute.

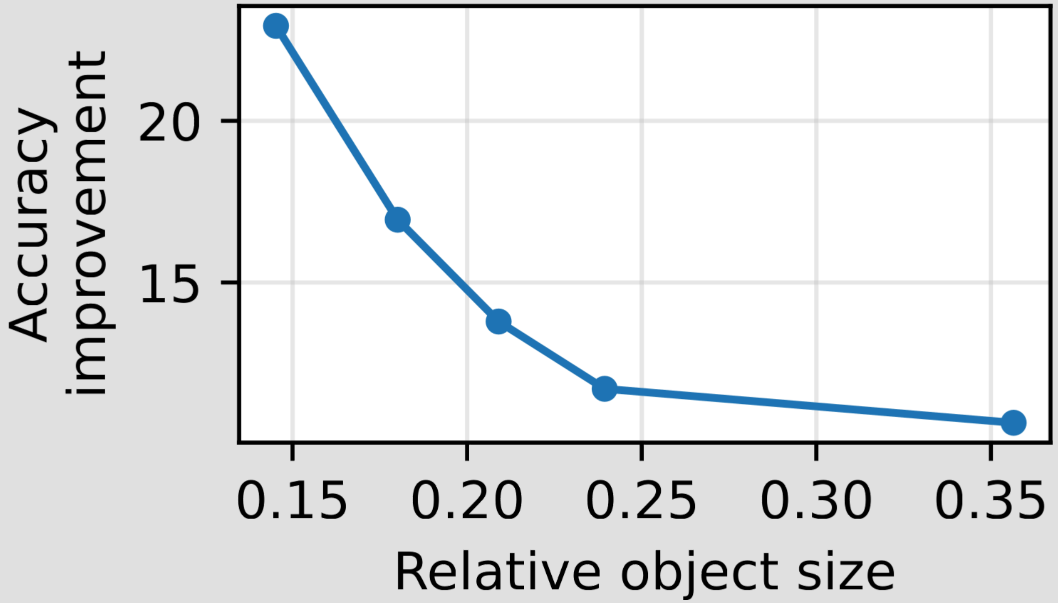
The accuracy gain is largest for small objects.

> Results

DeiT-S on TinyImageNet:



ResNet50 on FGVC-Aircraft:



OA-CutMix improves accuracy more for small objects. ⇒ Exactly where the label-error was largest.

Even coarse boxes recover most of the accuracy gain.

> Results

Method	Masks	Pos.	Size	absolute mode				relative mode			
				CUB200		TinyImageNet		CUB200		TinyImageNet	
				DeiT-S	ResNet50	DeiT-S	ResNet50	DeiT-S	ResNet50	DeiT-S	ResNet50
CutMix [30]	✗			79.69	49.62	58.86	61.79	79.69	49.62	58.86	61.79
CutMix + decoupled [15]	✗			80.00	81.84	61.53	62.41	80.00	81.84	61.53	62.41
OA-CutMix	Inter-Class Mix	✗	✗	80.48	81.12	61.16	62.50	80.79	81.34	62.50	63.39
	Intra-Class Mix	✗	~	81.14	81.08	61.32	62.64	<u>81.24</u>	81.41	62.05	63.39
	bounding box	✓	✗	81.01	81.98	63.30	64.27	81.00	**82.10**	<u>64.04</u>	<u>63.96</u>
	SAM3	✓	✓	**81.45**	**82.71**	**63.69**	**65.53**	**82.00**	<u>82.00</u>	**64.94**	**64.48**
	annotations	✓	✓	<u>81.19</u>	<u>82.64</u>			80.86	81.98		

The exact mask is not necessary; even bounding boxes are a large help.

Even coarse boxes recover most of the accuracy gain.

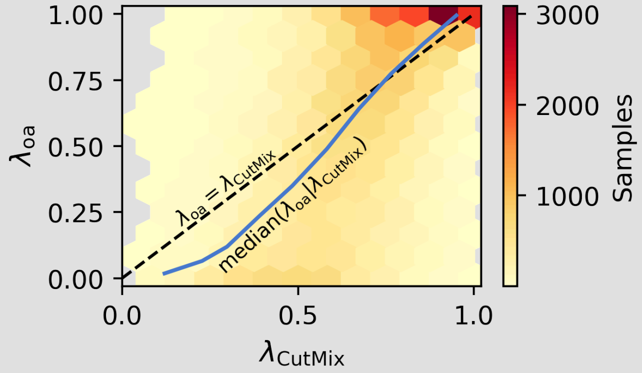
> Results

Method	Masks	Pos.	Size	absolute mode				relative mode			
				CUB200		TinyImageNet		CUB200		TinyImageNet	
				DeiT-S	ResNet50	DeiT-S	ResNet50	DeiT-S	ResNet50	DeiT-S	ResNet50
CutMix [30]				79.69	49.62	58.86	61.79	79.69	49.62	58.86	61.79
CutMix + decoupled [15]				80.00	81.84	61.53	62.41	80.00	81.84	61.53	62.41
OA-CutMix	Inter-Class Mix	✗	✗	80.48	81.12	61.16	62.50	80.79	81.34	62.50	63.39
OA-CutMix	Intra-Class Mix	✗	~	81.14	81.08	61.32	62.64	81.24	81.41	62.05	63.39
OA-CutMix	bounding box	✓	✗	81.01	81.98	63.30	64.27	81.00	**82.10**	64.04	63.96
OA-CutMix	SAM3	✓	✓	**81.45**	**82.71**	**63.69**	**65.53**	**82.00**	82.00	**64.94**	**64.48**
OA-CutMix	annotations	✓	✓	81.19	82.64			80.86	81.98		

The exact mask is not necessary; even bounding boxes are a large help.

OA-CutMix with mixed masks is still helpful, because the overall label-distribution is closer to the images.

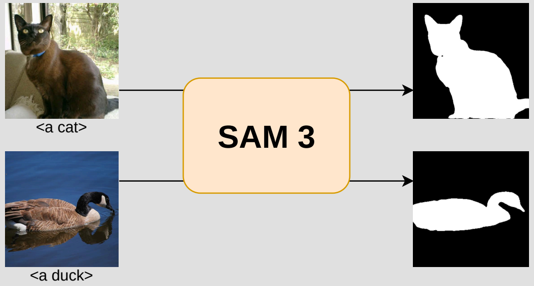
Correct the label of CutMix, not the combined image!



CutMix has a large and systematic label bias.

↓

OA-CutMix fixes the label by using additional object annotations.



↓

Improved accuracy at fast training speed.

PDF, Code, Blogpost

Generating masks is a one-time cost that amortizes over multiple runs.

Generating masks for the TinyImageNet training set takes 152.1 minutes on a single H100 GPU (657.4 img/s).

One-time cost per image $\Rightarrow$ scales linearly.

One training run with CutMix for 400 epochs takes 223.1 min.

$\Rightarrow$ time of $\approx \mathbf{270}$ epochs as overhead for detection.

Amortizes quickly over multiple training runs.

Naming the exact class requires fine-grained expertise.

What is the correct class for this image?

A: Greater Swiss Mountain Dog **B:** Tibetan Mastiff **C:** Bernese Mountain Dog **D:** Leonberger

Finding the object is easy.

Where in the image is a `Bernese Mountain Dog, Dog` ?

Finding the object is easy.

Where in the image is a `Bernese Mountain Dog, Dog`?

OA-CutMix in relative mode outperforms absolute mode.

Method	DeiT-S				ResNet50			
	CIFAR-100	ImageNet200	Aircraft	CUB200	CIFAR-100	ImageNet200	Aircraft	CUB200
CutMix [30]	79.14	83.43	79.60	79.69	79.34	87.74	72.07	49.62
CutMix + decoupled [15]	80.32	83.84	80.53	80.00	79.27	87.64	86.23	81.84
OA-CutMix/abs + decoupled	80.09	83.96	82.12	81.45	80.54	87.90	86.86	82.67
OA-CutMix/abs	80.78	83.68	81.82	81.45	80.35	87.47	**87.28**	**82.71**
OA-CutMix/rel + decoupled	**80.97**	84.06	81.58	81.71	79.05	88.11	86.44	82.36
OA-CutMix/rel	80.42	**84.18**	**82.42**	**82.00**	**81.41**	**88.19**	85.96	82.00

Relative mode outperforms absolute mode in most settings.

OA-CutMix improves calibration, but /abs beats /rel here.

Method	CIFAR-100		ImageNet200		TinyImageNet		Mean	latency
	DeiT-Ti	DeiT-S	DeiT-Ti	DeiT-S	DeiT-Ti	DeiT-S		[ms]
SnapMix [3]	9.07	7.92	18.05	14.44	4.69	6.72	10.15	39.98
Attentive CutMix [10]	5.96	5.61	9.83	6.84	5.87	11.74	7.64	56.10
TokenMix [6]	5.74	5.18	13.41	8.36	8.01	<u>4.45</u>	7.52	39.91
PuzzleMix [4]	9.69	4.38	12.17	2.42	4.13	**3.81**	6.10	273.28
Vanilla	7.56	19.07	3.47	11.19	17.01	30.36	14.78	
CutMix [11]	7.77	8.92	13.15	11.06	5.61	5.98	8.75	0.67
ResizeMix [8]	7.49	4.64	6.55	5.19	11.18	9.57	7.44	1.05
Mixup [12]+CutMix [11]	8.86	7.33	11.73	7.16	4.08	5.31	7.41	0.70
FMix [2]	3.98	4.89	7.38	5.66	8.91	10.63	6.91	6.80
Mixup [12]	6.49	6.36	7.88	6.02	4.02	5.42	6.03	0.73
CutMix + decoupled [7]	<u>3.04</u>	**2.43**	<u>2.40</u>	2.57	3.88	13.96	4.71	0.67
OA-CutMix/rel + decoupled	4.76	3.83	2.80	1.69	3.99	10.53	4.60	1.11
OA-CutMix/rel	3.77	<u>3.39</u>	**2.21**	1.66	3.60	9.15	3.96	1.03
OA-CutMix/abs + decoupled	**3.01**	4.35	3.26	<u>0.77</u>	<u>2.84</u>	7.75	<u>3.66</u>	1.01
OA-CutMix/abs	3.18	3.78	2.73	**0.68**	**2.31**	6.94	**3.27**	1.01

OA-CutMix improves calibration, but here /abs is better than /rel.
⇒ accuracy–calibration trade-off

OA-CutMix improves calibration, but /abs beats /rel here.

Method	CIFAR-100		ImageNet200		TinyImageNet		Mean	latency
	ResNet18	ResNet50	ResNet18	ResNet50	ResNet18	ResNet50		[ms]
PuzzleMix [4]	**1.10**	4.05	17.09	15.86	9.70	5.11	8.82	301.00
Attentive CutMix [10]	8.53	12.26	9.13	5.86	7.51	8.84	8.69	56.36
SnapMix [3]	3.10	5.73	7.71	5.79	8.90	5.82	6.17	51.49
Mixup [12]+CutMix [11]	11.54	6.46	17.79	17.99	16.30	**3.44**	12.25	0.67
Mixup [12]	11.64	6.03	15.21	17.58	16.62	5.99	12.18	0.73
Vanilla	5.60	8.20	**4.06**	4.39	18.26	15.46	9.33	
CutMix [11]	6.14	11.08	13.26	11.84	6.93	5.26	9.08	0.61
FMix [2]	4.15	4.12	10.60	14.47	7.10	5.44	7.65	6.68
CutMix + decoupled [7]	2.88	7.28	11.68	10.76	6.47	5.48	7.42	0.60
Alignmixup [9]	4.00	<u>3.61</u>	<u>4.33</u>	**1.85**	**2.18**	20.81	6.13	1.98
ResizeMix [8]	1.93	5.17	6.61	<u>3.94</u>	5.40	7.18	5.04	0.62
OA-CutMix/abs + decoupled	<u>1.74</u>	5.63	7.83	5.19	3.14	5.58	4.85	1.01
OA-CutMix/rel	2.56	3.69	7.00	4.97	3.49	6.87	4.76	1.02
OA-CutMix/rel + decoupled	2.77	4.53	5.80	4.67	3.18	7.02	<u>4.66</u>	1.00
OA-CutMix/abs	1.79	**3.37**	7.26	5.88	<u>3.11</u>	<u>4.19</u>	**4.27**	1.05

OA-CutMix improves calibration, but here /abs is better than /rel.
⇒ accuracy–calibration trade-off

OA-CutMix helps even when finetuning from non-OA-CutMix weights.

Method	Aircraft		Cars		CUB200		Mean	latency
	DeiT-S	ResNet50	DeiT-S	ResNet50	DeiT-S	ResNet50		[ms]
TokenMix [14]	78.70		89.83		80.69			39.91
Attentive CutMix [28]	78.46	71.98	89.44	76.78	82.48	51.38	75.09	56.10
PuzzleMix [8]	**82.93**	76.45	**91.21**	80.62	**84.05**	59.08	79.06	273.28
Vanilla	79.99	86.02	88.92	90.16	81.69	9.44	72.70	
FMix [5]	76.93	70.87	89.69	77.04	79.89	49.90	74.05	6.80
CutMix [30]	79.60	72.07	89.14	76.58	79.69	49.62	74.45	0.67
ResizeMix [19]	78.43	74.89	90.25	80.36	81.64	55.82	76.90	1.05
CutMix + decoupled [15]	80.53	86.23	89.67	90.98	80.00	81.84	84.87	0.67
Mixup [31]	79.84	86.80	89.47	91.00	82.52	**83.00**	85.44	0.73
Mixup [31] + CutMix [30]	80.65	86.65	90.40	90.51	82.50	82.24	85.49	0.70
OA-CutMix/rel	82.42	85.96	90.66	91.00	82.00	82.00	85.67	1.03
OA-CutMix/abs	81.82	**87.28**	90.61	**91.42**	81.45	82.71	**85.88**	1.01

PuzzleMix outperforms OA-CutMix for DeiT-S, but at $270\times$ the training compute.

OA-CutMix /abs and MixUp work very well for ResNet. This might be an artifact of how this particular ResNet is trained (MixUp).

These are the hyperparameters we trained with.

Dataset	Training Samples	Resolution	Pretrained	Epochs	Batch Size
CIFAR-100	50 000	32 ResNet, 224 DeiT	✗	400	100
TinyImageNet	100 000	64	✗	400	256
ImageNet200	258 758	224	✗	200	256
FGVC-Aircraft	6 667	224	✓	200	32
Stanford Cars	8 144	224	✓	200	32
CUB200	5 994	224	✓	200	16

Training from Scratch	Finetuning
	Resize
RandomResizedCrop	RandomResizedCrop
RandomHorizontalFlip	RandomHorizontalFlip
RandAugment	ColorJitter