Arxiv: https://arxiv.org/abs/2212.14504
Code: https://github.com/tractableai/perceptual-mae

Improving Visual Representation Learning through Perceptual Understanding

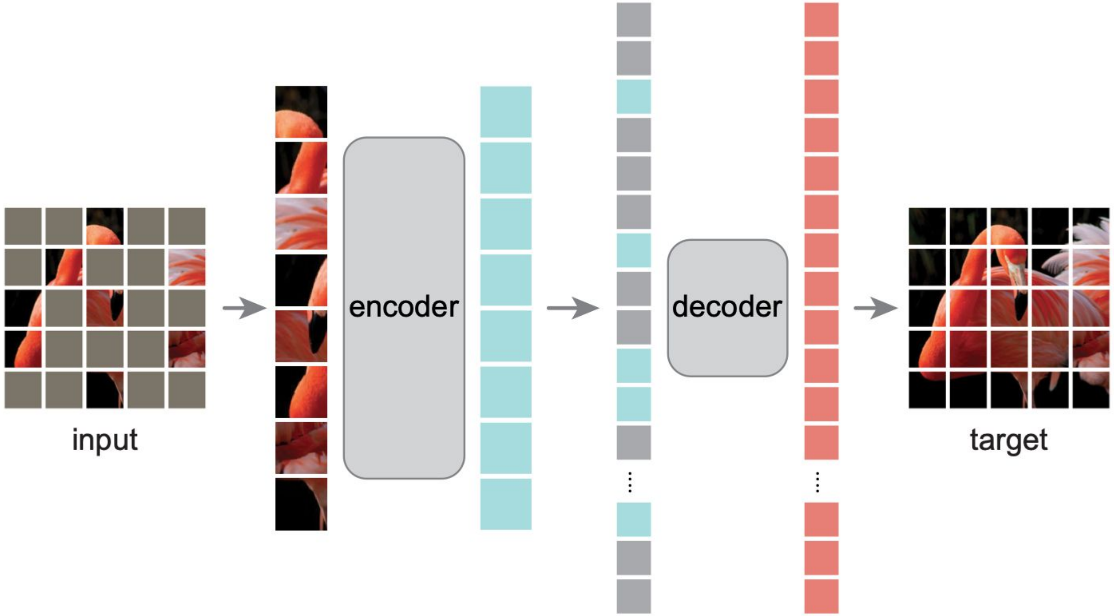
  

Samyakh Tukra Fred Hoffman Ken Chatfield

WED-PM-204

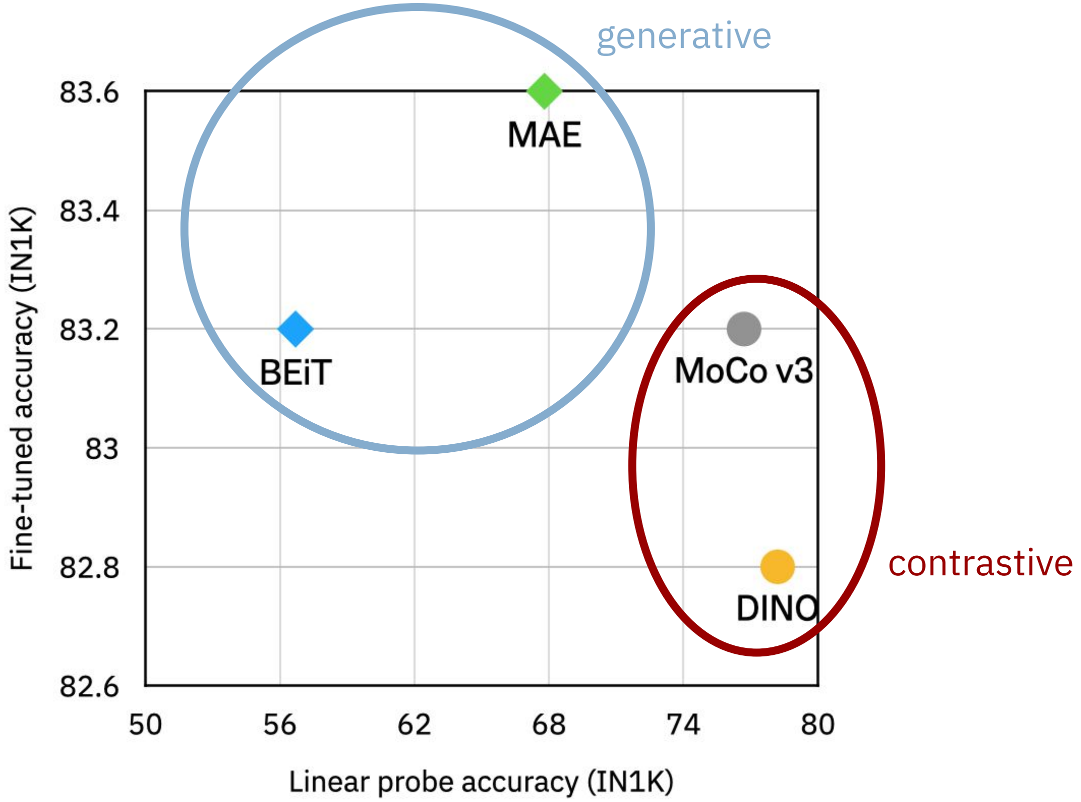
Tractable

The self-supervised revolution in NLP has made it to vision

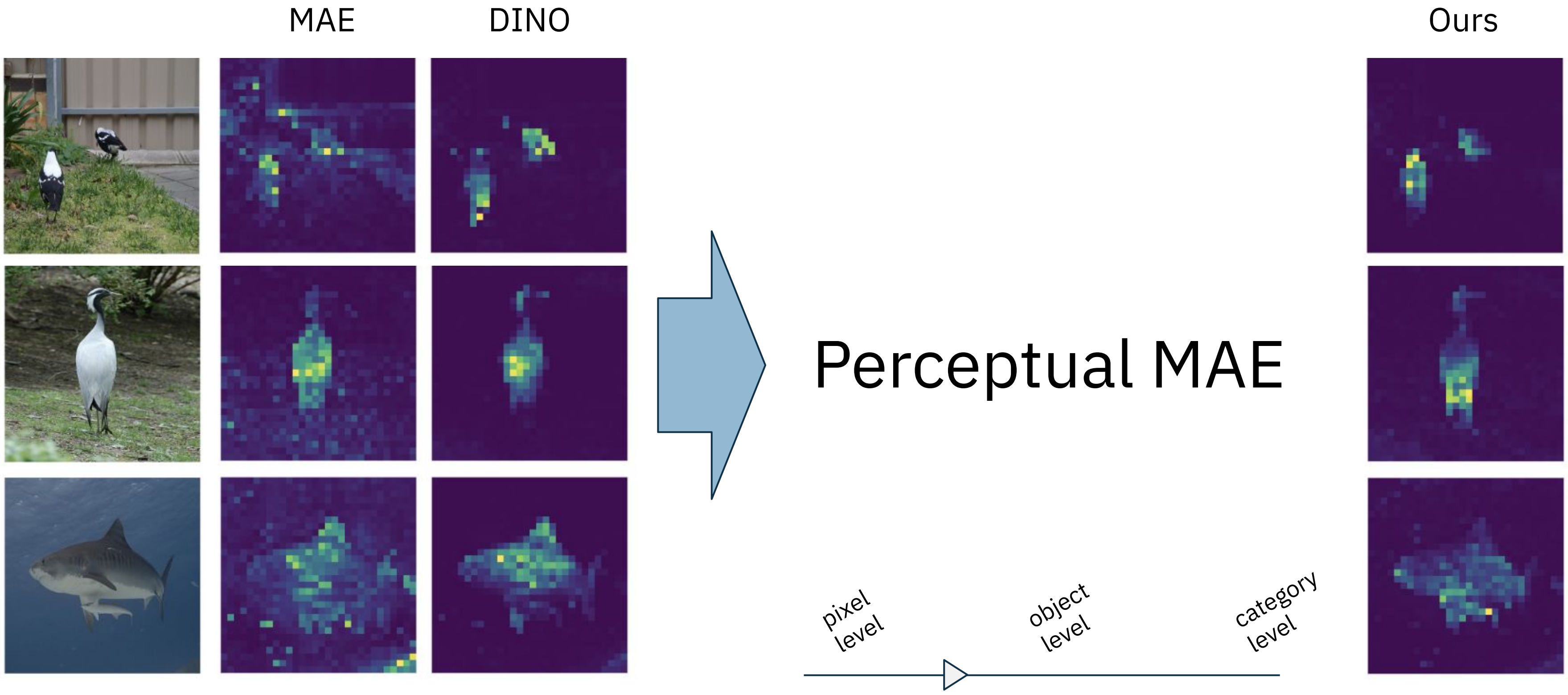


Masked Autoencoders are Scalable Visual Learners
Kaimeng He et al. (CVPR 2022)

Problem: generative SSL still underperforms when not fine-tuning



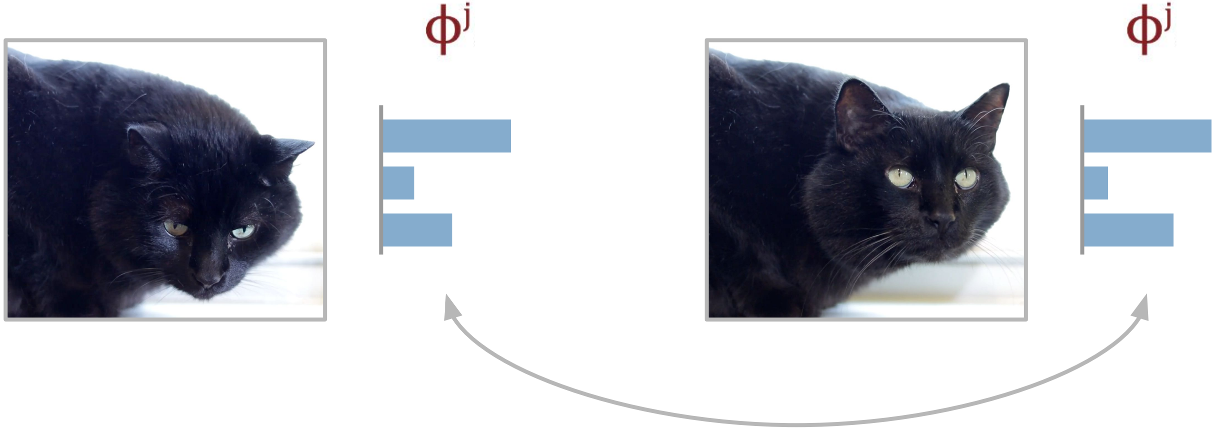
Our aim: incorporate learning higher-level features into masked autoencoders



How? We add an explicit image-level perceptual penalty to the loss

$$L^G = || G(I_m) - I ||_1 + L^G_{feat}$$

Perceptual loss by **feature matching:**



How? We add an explicit image-level perceptual penalty to the loss

$$L^G = || G(I_m) - I ||_1 + L^G_{feat}$$

Perceptual loss by **feature matching:**

$$L^G_{feat} = \delta_f \sum_j \frac{1}{N_j} || \phi^j(G(I_m)) - \phi^j(I) ||_1$$

How? We add an explicit image-level perceptual penalty to the loss

$$L^G = || G(I_m) - I ||_1 + L^G_{feat}$$

Perceptual loss by **feature matching:**

$$L^G_{feat} = \delta_f \sum_j \frac{1}{N_j} || \phi^j(G(I_m)) - \phi^j(I) ||_1 +$$

$$\delta_s \sum_j \frac{1}{N_j} || \psi(\phi^j(G(I_m))) - \psi(\phi^j(I)) ||_1$$

How? We add an explicit image-level perceptual penalty to the loss

$$L^G = || G(I_m) - I ||_1 + L^G_{feat} + L^G_{adv}$$

Perceptual loss by **feature matching:**

$$L^G_{feat} = \delta_f \sum_j \frac{1}{N_j} || \phi^j(G(I_m)) - \phi^j(I) ||_1 + \delta_s \sum_j \frac{1}{N_j} || \psi(\phi^j(G(I_m))) - \psi(\phi^j(I)) ||_1$$

Trick: *where* ϕ *is an adversarial discriminator*

Image-level adversarial term

Penalise reconstruction which can be distinguished from real image

How? We add an explicit image-level perceptual penalty to the loss

$$L^G = || G(I_m) - I ||_1 + L^G_{feat} + L^G_{adv}$$

Perceptual loss by **feature matching:**

$$L^G_{feat} = \delta_f \sum_j \frac{1}{N_j} || \phi^j(G(I_m)) - \phi^j(I) ||_1 +$$

$$\delta_s \sum_j \frac{1}{N_j} || \psi\, (\phi^j(G(I_m))) - \psi\, (\phi^j(I)) ||_1$$

Trick: *where* ϕ *is an adversarial discriminator*

Image-level adversarial term

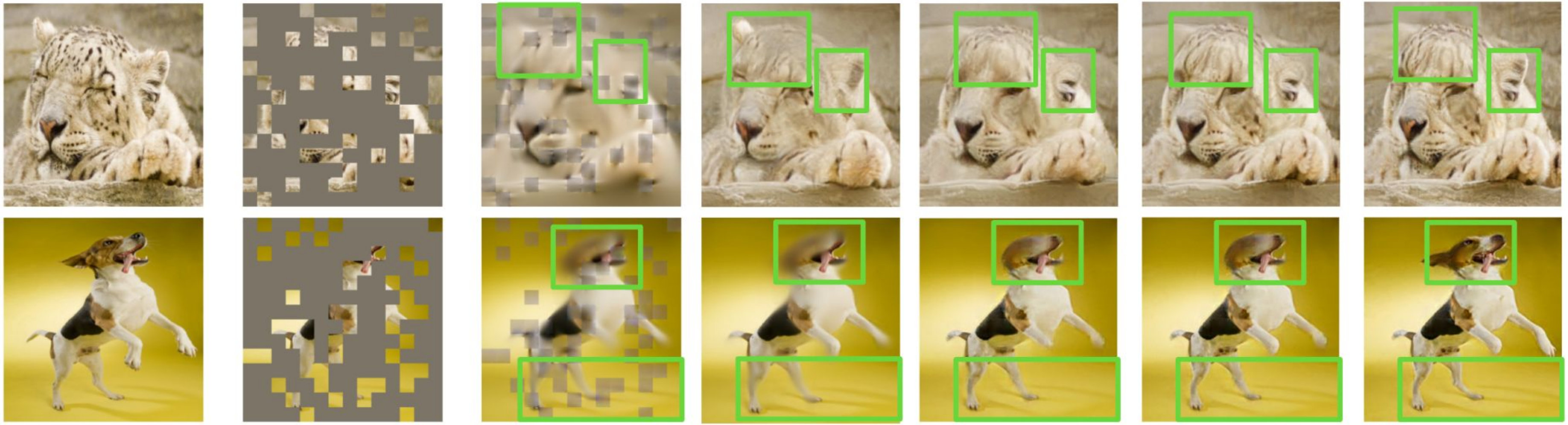
Penalise reconstruction which can be distinguished from real image

Plus from the generative adversarial toolbox:

- multi-scale gradients
- perceptual path reg.
- adaptive discriminator augmentation (ADA)

Results

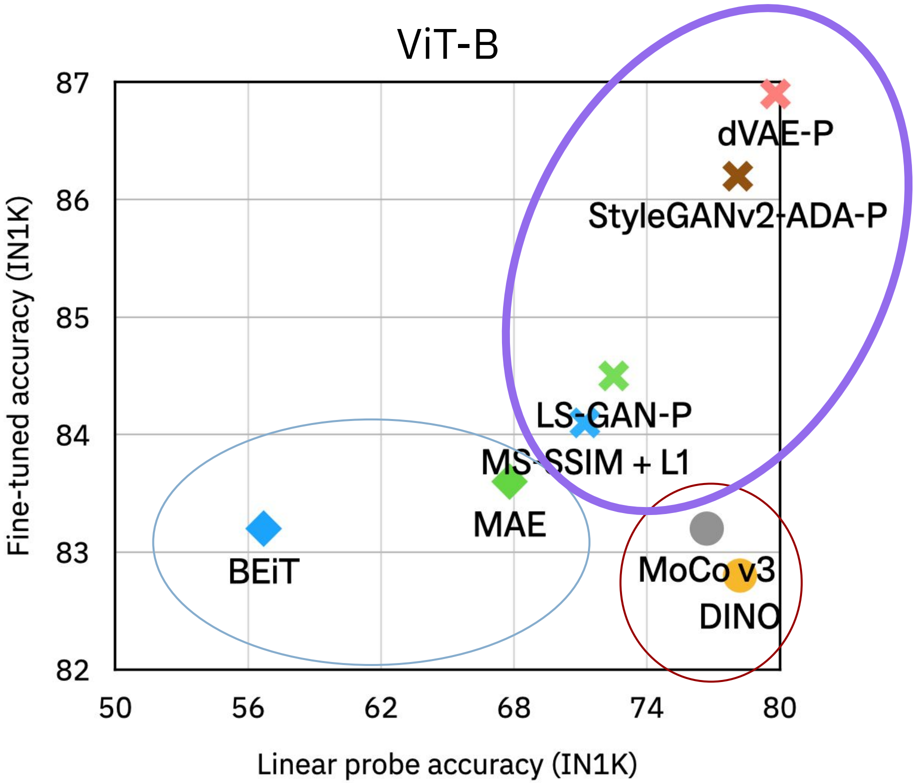
Not only does this improve decoder reconstruction



	MAE	MS-SSIM	LS-GAN-P	MSG-GAN-P	StyleGANv2-ADA-P
L1	0.25	0.21	0.16	0.11	0.06
IS	6.33	8.01	16.2	32.1	36.8

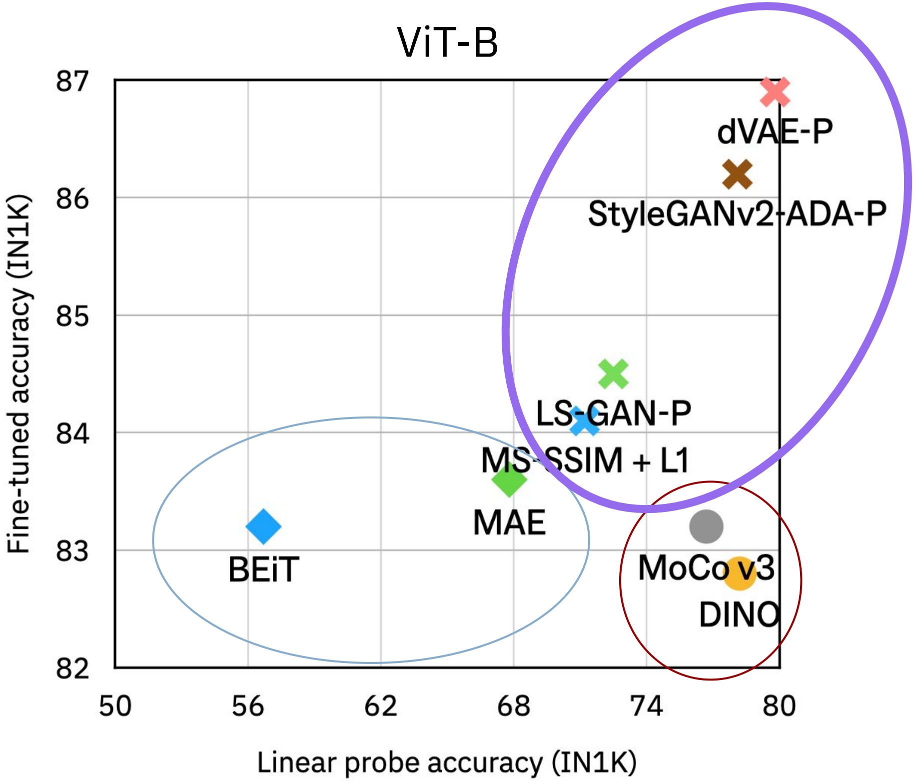
Results

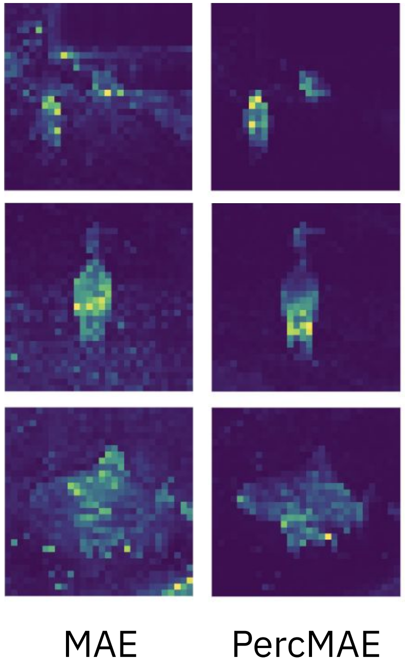
But also boosts both fine-tuned and few-shot settings for classification



Results

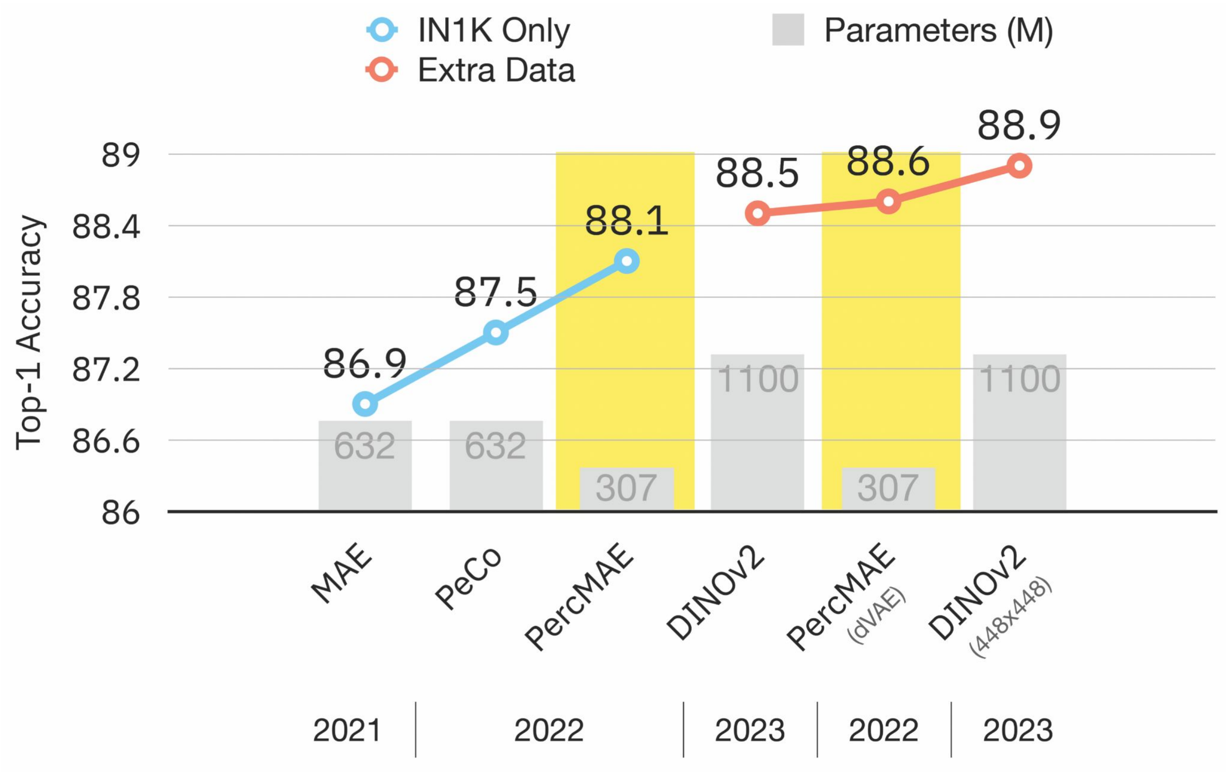
But also boosts both fine-tuned and few-shot settings for classification





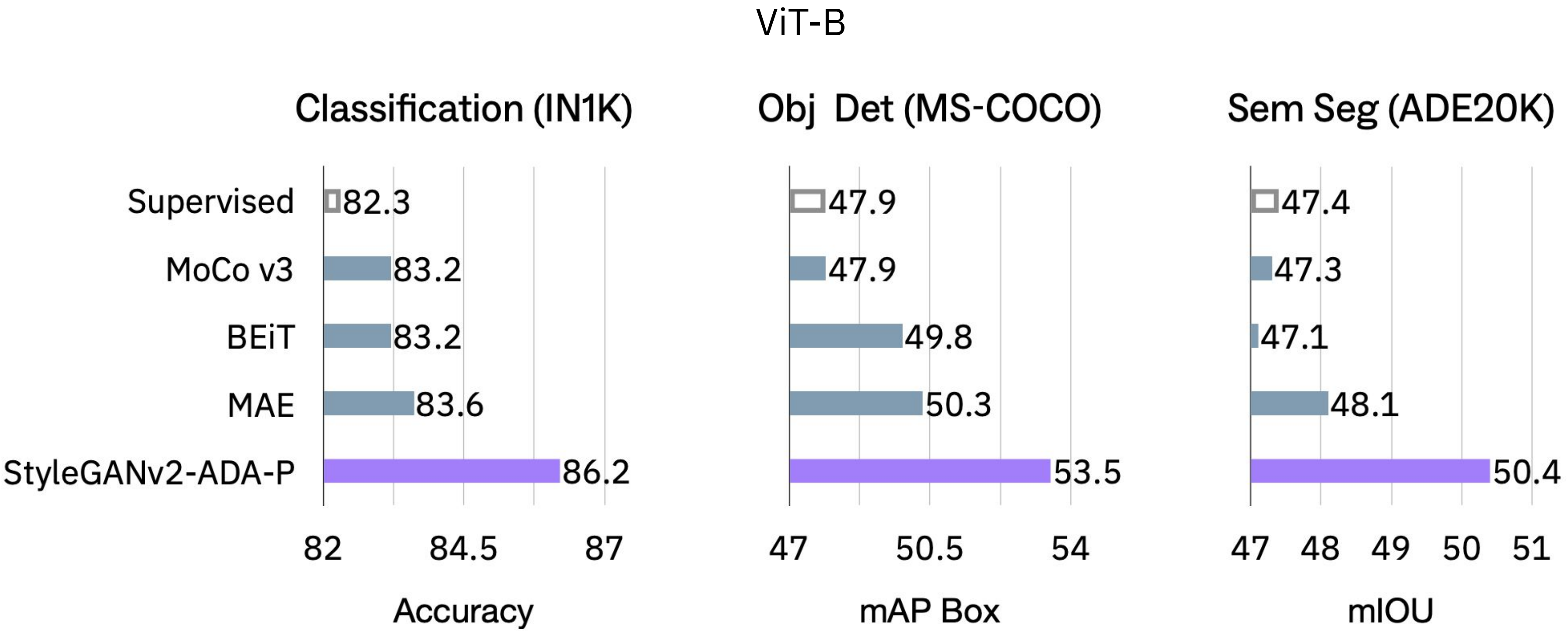
Results

All whilst being much more data and compute efficient than alternate methods



Results

And generalises across tasks



Poster session:
WED-PM-204

https://github.com/tractableai/perceptual-mae