Training Vision Transformers to Focus: Emergence of Spatial Attention Head Specialization

Yuji Nozawa, Yu-Chieh Lin, Youyang Ng
Kioxia Corporation

The Fortieth Annual Conference on Neural Information Processing Systems (NeurIPS 2026)

KIOXIA

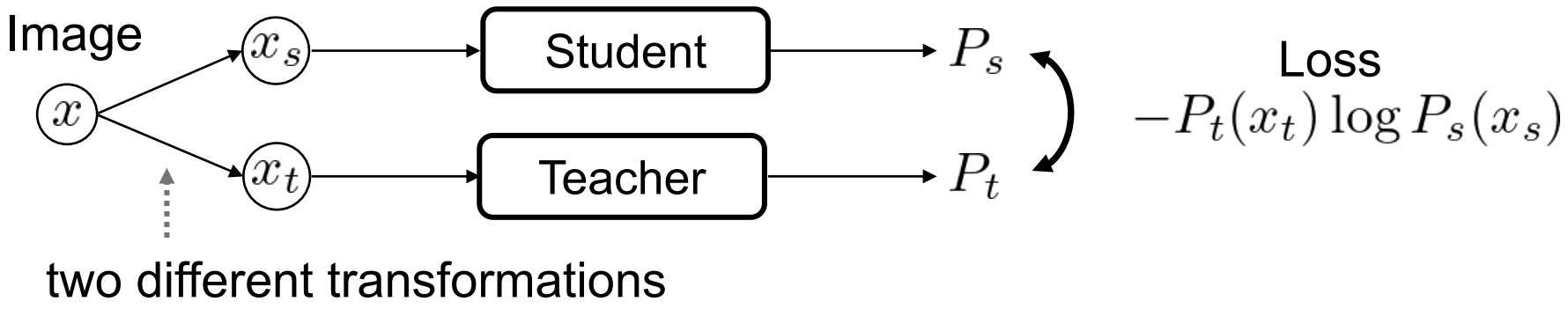
NEURAL INFORMATION PROCESSING SYSTEMS

Project Page

Introduction: SSL for ViTs

- We study self-supervised learning (SSL) for Vision Transformers (ViTs)

Standard setting (DINO loss [1]):



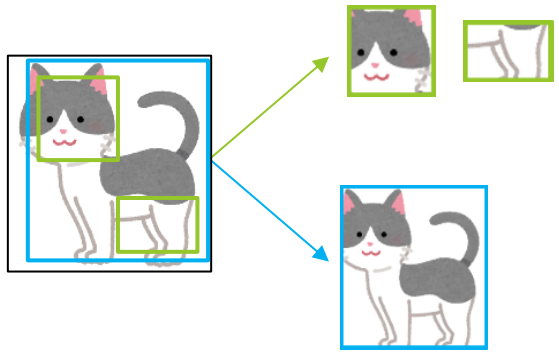
Aligning student and teacher outputs from differently transformed inputs

[1] Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021

Key transformation in DINO

Local-to-global correspondence

- Align local features from local crops with global features from global crops



local crop: small-region crop



global crop: large-region crop

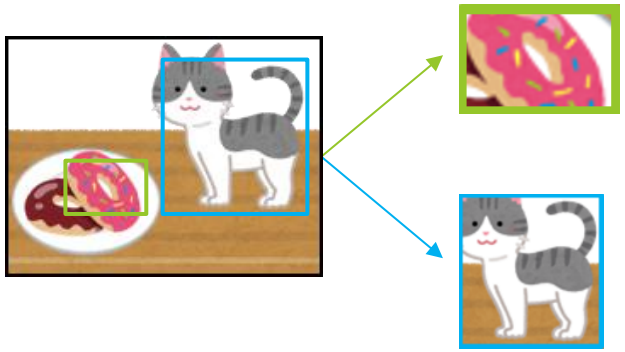
student: global & local
teacher: global only

Strength: captures correspondence between the global representation of an object and local representations of its parts

Limitation of local-to-global correspondence

Not designed for multiple objects or semantic regions

Example:



Undesirably matches features from different objects

Question: How can we establish correspondence suitable for images with multiple objects or semantic regions in SSL?

Our approach: **Region-to-Focus Correspondence (ReFoCo)**

Train ViTs to learn **how to focus** on different regions within an image

Previous work used in our training

Prompt-guided Head Selection (PHS) [2]

Method for modifying features using a visual prompt as a region of interest (ROI)

- Specify an ROI using a binary mask
- Select attention heads that strongly attend to the ROI
- Modify features using a head selection method

In the last attention layer:

$$\mathrm{head}_i = \mathrm{Softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_k}}\right)\mathbf{V}_i \longrightarrow \mathrm{head}_i^{\mathrm{HS}} = \begin{cases} \mathrm{head}_i \times h/h_{\mathrm{on}} & \text{if } i \in i_{\mathrm{on}}, \text{ selected} \\ 0 & \text{if } i \notin i_{\mathrm{on}}. \text{ not selected} \end{cases}$$

$$\mathrm{MHA} = \mathrm{Linear}\left([\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h]\right) \longrightarrow \mathrm{MHA}^{\mathrm{HS}}$$

Modified features: $P(x) \longrightarrow P^{\mathrm{HS}}(x, ROI)$

[2] Nozawa et al., Prompt-Guided Attention Head Selection for Focus-Oriented Image Retrieval, CVPRW 2025

Region-to-Focus Correspondence (ReFoCo)

We use PHS to implement ReFoCo

Teacher side:

Transform the image to emphasize the ROI

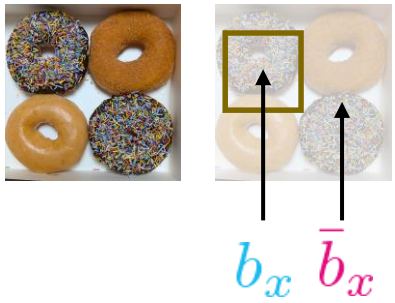


Student side:

Use PHS to modify features with the ROI

Visual focusing loss functions

Focusing on inside / outside box



We introduce **two visual focusing loss functions**

- Generate a small box at a random location in each training image
- Use the inside b_x and outside $\bar{b}_x$ of the box as ROIs for PHS
- Associate b_x and $\bar{b}_x$ with the crop and masked images, respectively

✓ No annotations required

Inside-focus loss:

$$\mathcal{L}_{\text{in}} = -\frac{1}{B} \sum_{x:\text{image}} P_{\text{t}}(x_{\text{crop}}) \log P_{\text{s}}^{\text{PHS}}(x, ROI = b_x)$$

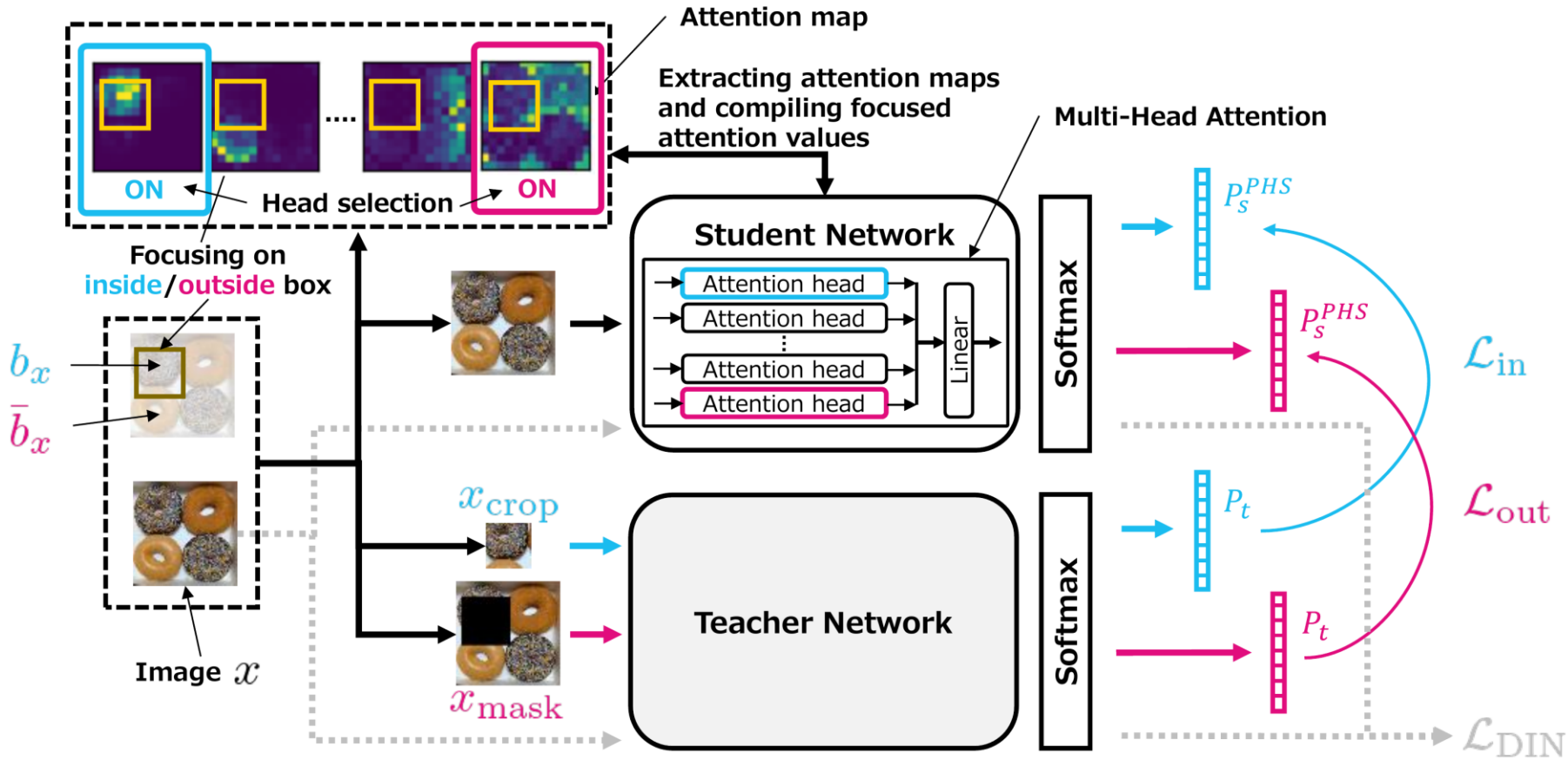
Outside-focus loss:

$$\mathcal{L}_{\text{out}} = -\frac{1}{B} \sum_{x:\text{image}} P_{\text{t}}(x_{\text{mask}}) \log P_{\text{s}}^{\text{PHS}}(x, ROI = \bar{b}_x)$$

Overview of our method

Total loss: $\mathcal{L}_{\text{focus}} = (1 - w_{\text{in}} - w_{\text{out}}) \cdot \mathcal{L}_{\text{DINO}} + w_{\text{in}} \cdot \mathcal{L}_{\text{in}} + w_{\text{out}} \cdot \mathcal{L}_{\text{out}}$

$+ \mathcal{L}_{\text{iBOT}} + 0.1 \cdot \mathcal{L}_{\text{koleo}}$: DINOv2 loss [3]



[3] Oquab et al., DINOv2: Learning Robust Visual Features without Supervision, TMLR 2024

Experiments: Zero-shot image segmentation

Training setup for main experiments:

- Training from scratch
- Dataset: ImageNet-1k
- Architecture: ViT-L/16 with registers [4]
- Loss weights: $(w_{\text{in}}, w_{\text{out}}) \in \{(0,0),\ (0.1,0),\ (0,0.1),\ (0.05,0.05)\}$

DINOv2 $\quad \mathcal{L}_{\text{in}} \quad \mathcal{L}_{\text{out}} \quad \mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$

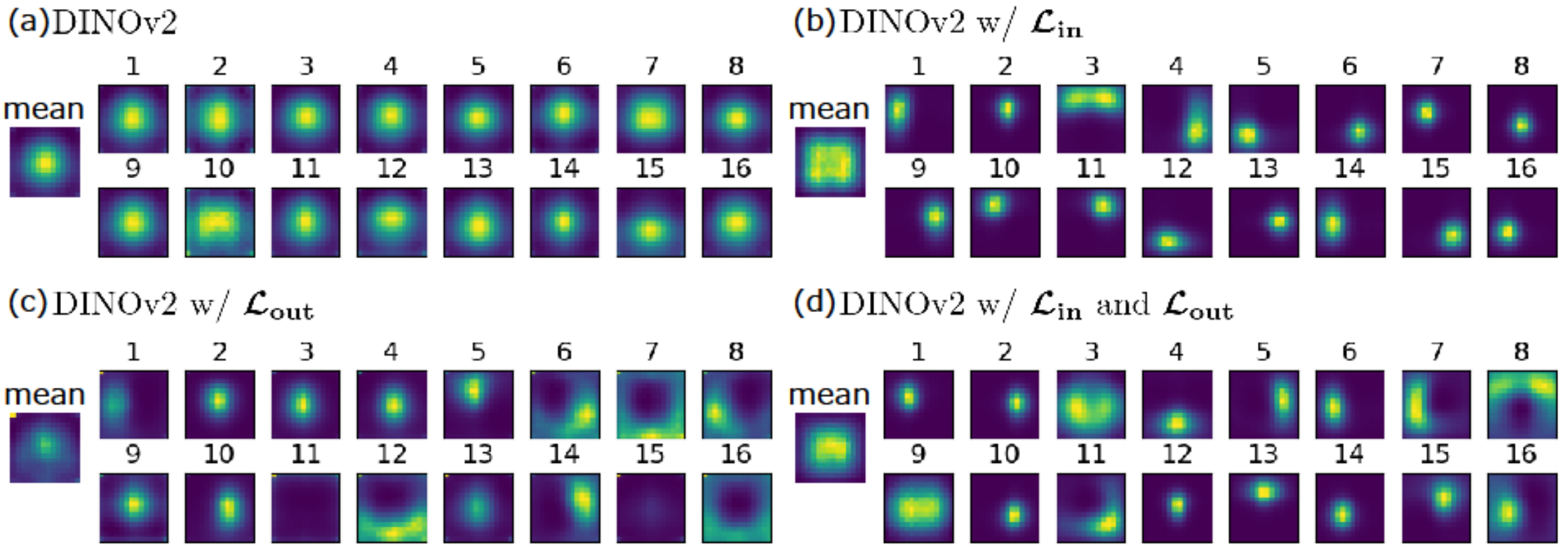
	ImageNet-Seg.			PASCAL VOC		
Loss	mIoU (%)↑	Acc (%)↑	mAP (%)↑	mIoU (%)↑	Acc (%)↑	mAP (%)↑
DINOv2	63.67	80.69	85.39	58.10	75.28	84.79
w/ $\mathcal{L}_{\text{in}}$	63.93	80.61	86.03	57.82	74.76	85.00
w/ $\mathcal{L}_{\text{out}}$	63.65	80.94	84.69	58.23	75.73	84.97
w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	**67.29**	**83.20**	**86.43**	**60.68**	**77.36**	**85.27**

Our $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$ model outperforms the original DINOv2 model in zero-shot image segmentation task

[4] Darcet et al., Vision Transformers Need Registers, ICLR 2024

Experiments: Attention maps visualizations

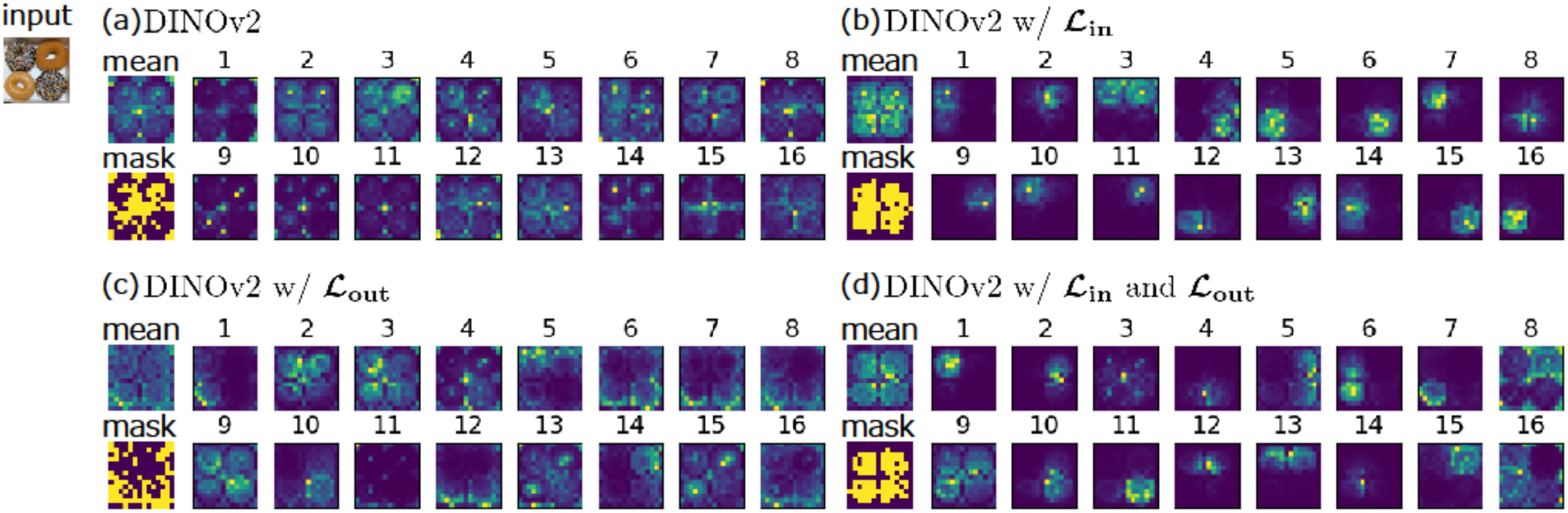
ImageNet-1k average



Emergence of Spatial Attention Head Specialization: Different attention heads focus on distinct spatial positions, integrating information across regions

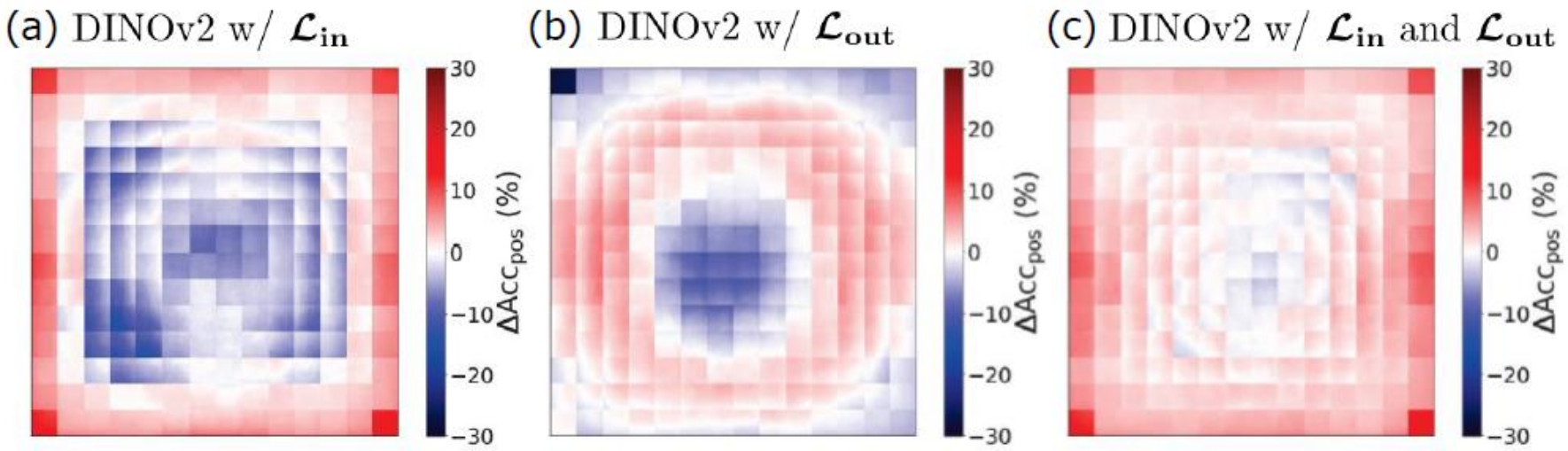
Experiments: Attention maps visualizations

Single-image example



Emergence of Spatial Attention Head Specialization: Different attention heads focus on distinct spatial positions, integrating information across regions

Experiments: Segmentation performance by position



Changes in segmentation accuracy at each position compared to the accuracy of the original DINOv2 model

$\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$ play complementary roles in improving segmentation accuracy

Experiments: Results across model sizes

Size	Loss	ImageNet-Seg.			PASCAL VOC		
		mIoU (%)↑	Acc (%)↑	mAP (%)↑	mIoU (%)↑	Acc (%)↑	mAP (%)↑
B	DINOv2	68.32	83.81	87.63	62.20	78.14	86.46
	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	**69.33**	**84.40**	**87.84**	**62.98**	**78.65**	**86.77**
	Δ	+1.01	+0.59	+0.21	+0.78	+0.51	+0.31
L	DINOv2	63.67	80.69	85.39	58.10	75.28	84.79
	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	**67.29**	**83.20**	**86.43**	**60.68**	**77.36**	**85.27**
	Δ	+3.62	+2.51	+1.04	+2.58	+2.08	+0.48
g	DINOv2	63.37	80.53	85.20	58.34	75.55	84.85
	w/ $\mathcal{L}_{\text{in}}$ & $\mathcal{L}_{\text{out}}$	**67.61**	**83.31**	**86.86**	**61.31**	**77.65**	**86.09**
	Δ	+4.24	+2.78	+1.66	+2.97	+2.10	+1.24

Larger models show greater performance gains

Summary

- Introduced two visual focusing loss functions to implement ReFoCo
- Improved zero-shot image segmentation performance
- Clearer and more interpretable attention maps
- Improvements also observed in other tasks such as object discovery
- Trade-off: slight reduction in classification performance

Project Page

GitHub

Hugging Face

