
DINOv2, DINOv3 : Self-supervised Vision Foundation Model

DMQA Open Seminar

2025.10.31

Data Mining & Quality Analytics Lab.

이혜승

발표자 소개



이혜승 (Hyeseung Lee)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S. Student (2024.09 ~ Present)

Research Interest

- Image Segmentation, Foundation Model
- Multi-Agent LLM

Contact

- hyeseunglee@korea.ac.kr

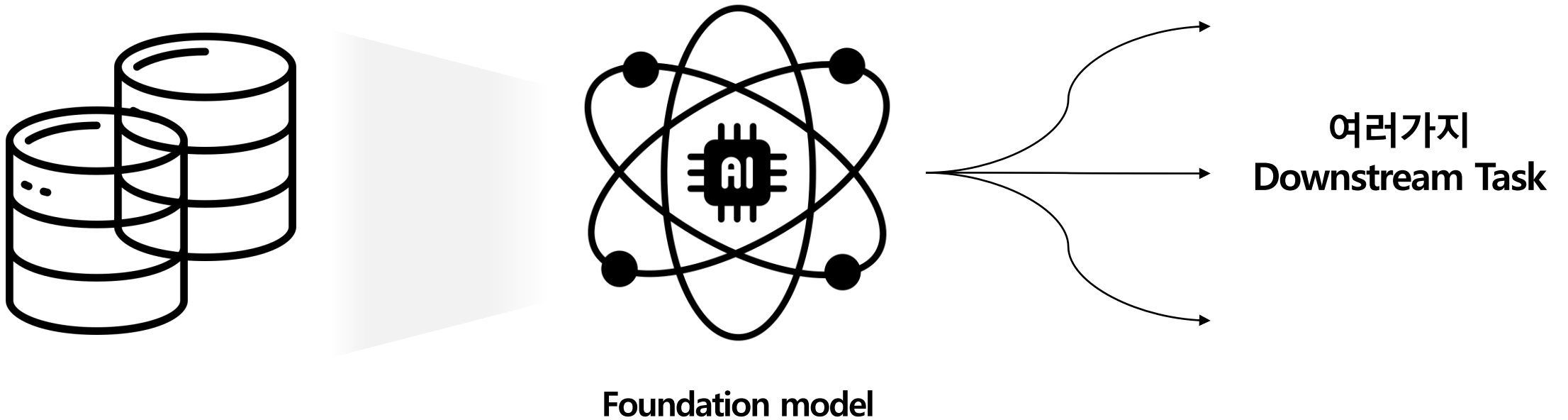
Introduction



Foundation Models

What is a foundation model?

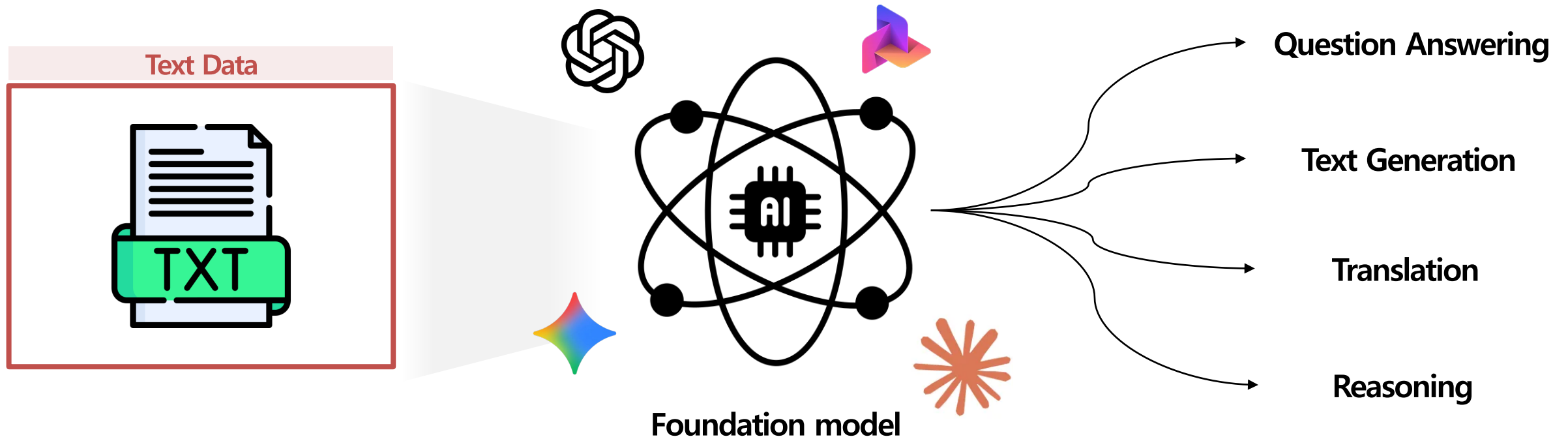
방대한 양의 데이터로 사전학습된 범용적인 모델



Foundation Models

What is a foundation model?

방대한 양의 데이터로 사전학습된 범용적인 모델

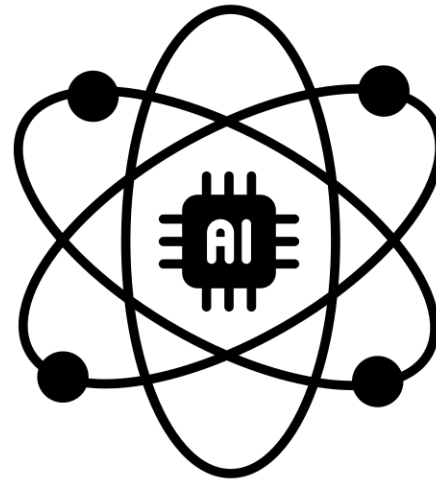


Foundation Models

What is a foundation model?

방대한 양의 데이터로 사전학습된 범용적인 모델

How about computer vision?



Foundation model

Downstream Task

Classification

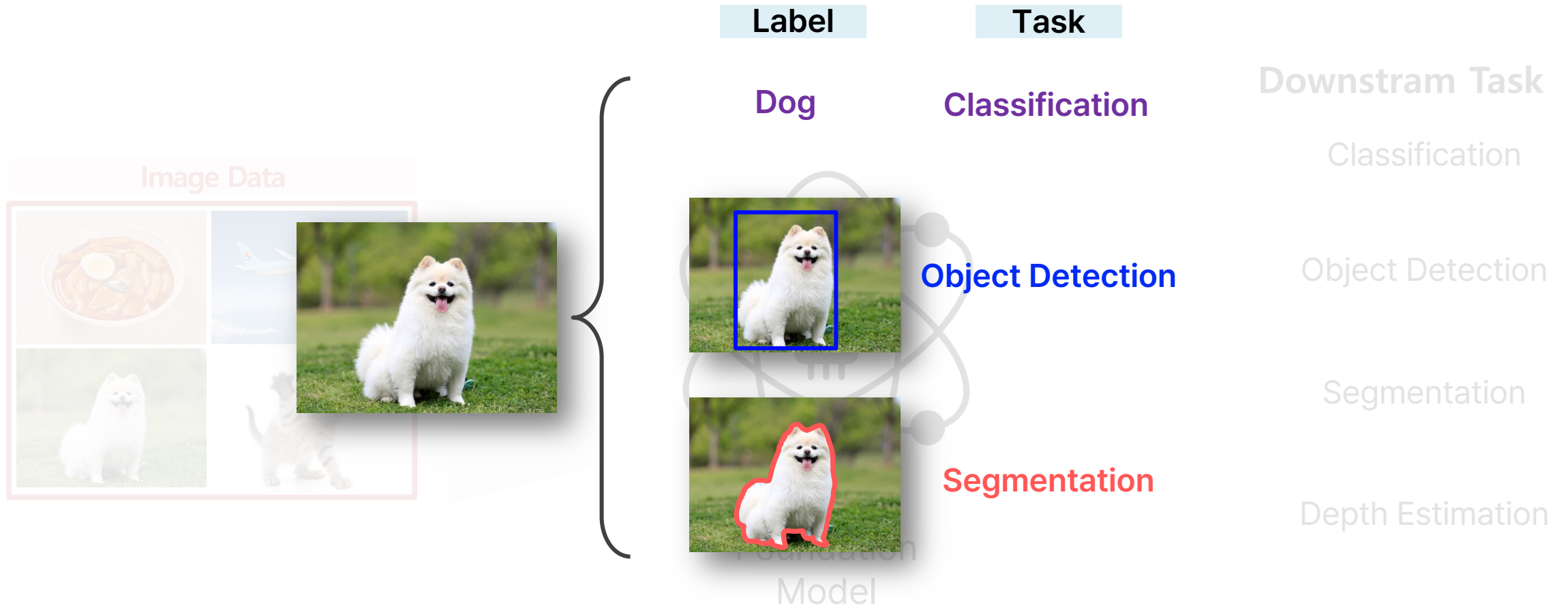
Object Detection

Segmentation

Depth Estimation

Foundation Models

Deep Learning 모델을 학습시키기 위해서는 Labeled data가 필요하다



하지만 데이터 레이블링에는 많은 시간과 노력 필요

Research Trends

DINO Series

Self-supervised learning을 활용한 Foundation Model 연구 동향



Emerging Properties in Self-Supervised Vision Transformers (2021)

Emerging Properties in Self-Supervised Vision Transformers

Mihika Cane¹, Hugo Touvron², Ilija Mitrovic³, Herve Jegou², Armand Joulin², Piotr Bojanowski²

¹Facebook AI Research ²Meta ³University of Toronto



Abstract

In this paper, we question if self-supervised learning provides any properties to Vision Transformers (ViT) that are not captured by conventional networks (convnets). Beyond the fact that ViT can be trained on large-scale datasets, we argue that ViT can learn visual concepts from self-supervised data. We show that ViT can learn visual concepts from self-supervised data, which does not emerge on convnets with supervised data. We show that ViT can learn visual concepts from self-supervised data, which does not emerge on convnets with supervised data. We show that ViT can learn visual concepts from self-supervised data, which does not emerge on convnets with supervised data.

DINOv2 (2024)

DINOv2: Learning Robust Visual Features without Supervision

Maxime Oquab¹, Timothée Darcet¹, Théo Mostakani¹, Huu V. Vo¹, Marc Saifan¹, Neil Kulkarni¹, Pierre Fernandez¹, Daniel Haindl, Prantik Maiti, Alexander El-Nor, Mohamed Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shuang Wu, Li, Ishaan Misra, Michael Rabbat, Vansh Shrivastava, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Armand Joulin², Patrick Labatut², Piotr Bojanowski²

¹Meta AI Research ²Meta

Abstract

The recent breakthroughs in natural language processing for model pretraining on large quantities of data have opened the way for similar foundation models to compare vision. These models could greatly simplify the use of images in any system by producing general-purpose visual features, i.e. features that work across image distributions and tasks without fine-tuning. This work shows that existing pretraining methods, especially self-supervised methods, can produce such features if trained on enough curated data from diverse sources. We review existing approaches and combine different techniques to scale our pretraining in terms of data and model size. Most of the technical contributions aim at accelerating and stabilizing the training at scale. In terms of data, we propose an automatic pipeline to build a dedicated, diverse, and curated image dataset instead of uncurated data, as typically done in the self-supervised literature. In terms of models, we train a ViT model (DINOv2) with 120 parameters and distill it into a series of smaller models that surpass the best available general-purpose features, OpenCLIP (Baiter et al., 2021) on most of the benchmarks at image and pixel levels.

DINOv3 (2025)

DINOv3

Orlène Siméoni¹, Huu V. Vo¹, Maximilian Seitzer¹, Federico Baldassarre¹, Maxime Oquab¹, Ojo Jose¹, Vadim Khalidov¹, Marc Saifan¹, Seungmin Yi¹, Michael Rasmussen-Jensen¹, Francisco Massa¹, Daniel Haindl¹, Luca Wolzert¹, Jueyuan Wang¹, Timothée Darcet¹, Théo Mostakani¹, Leonel Santana¹, Claire Roberts¹, Andrew Vaidell¹, Janice Telam¹, John Brann¹, Camille Couprie¹, Julien Mairal¹, Hervé Jegou¹, Patrick Labatut², Piotr Bojanowski²

¹Meta AI Research ²Meta

Abstract

Self-supervised learning holds the promise of eliminating the need for manual data annotation, enabling models to scale efficiently to massive datasets and larger architectures. By not being tailored to specific tasks or domains, this training paradigm has the potential to learn visual representations from diverse sources, ranging from natural to aerial images using a single algorithm. This technical report introduces DINOv3, a major milestone toward realizing this vision by leveraging simple yet effective strategies. First, we leverage the benefit of scaling both dataset and model size by careful data preparation, design, and optimization. Second, we introduce a new method called Grad anchoring, which effectively addresses the known, yet unsolved issue of dense feature map degradation during long training schedules. Finally, we apply post-hoc strategies that further enhance our model's flexibility with respect to resolution, model size, and alignment with text. As a result, we present a versatile vision foundation model that outperforms the specialized state of the art across a broad range of settings, without fine-tuning. DINOv3 produces high-quality dense features that achieve outstanding performance on various vision tasks, significantly surpassing previous self- and weakly-supervised foundation models. We also share the DINOv3 suite of vision models, designed to advance the state of the art on a wide spectrum of tasks and data by providing scalable solutions for diverse resource constraints and deployment scenarios.

2021 → 2025

Self-Supervised Learning (SSL)

Pretrain

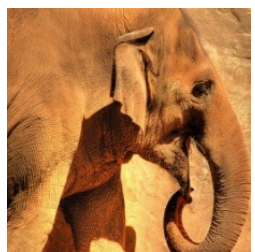
Unlabeled Data



Feature
Extractor

Finetune

Labeled Data



Feature
Extractor

Classifier

Elephant Dog Guitar Horse

Performs Well!

종료 Self-Supervised Learning (Algorithm & application)

Saekho Moon
Nov 20, 2020

Applications of Self-Supervised Learning

발표자: 이영재

2022년 3월 4일
오전 3시 ~
온라인 비디오 시청 (YouTube)
온라인 비디오 시청 (YouTube)

세미나 정보 보기 →

종료 Self-Supervised Learning (Algorithm & application)

Saekho Moon
Nov 20, 2020

Self-Supervised Learning (algorithm & ap

발표자: 분석호

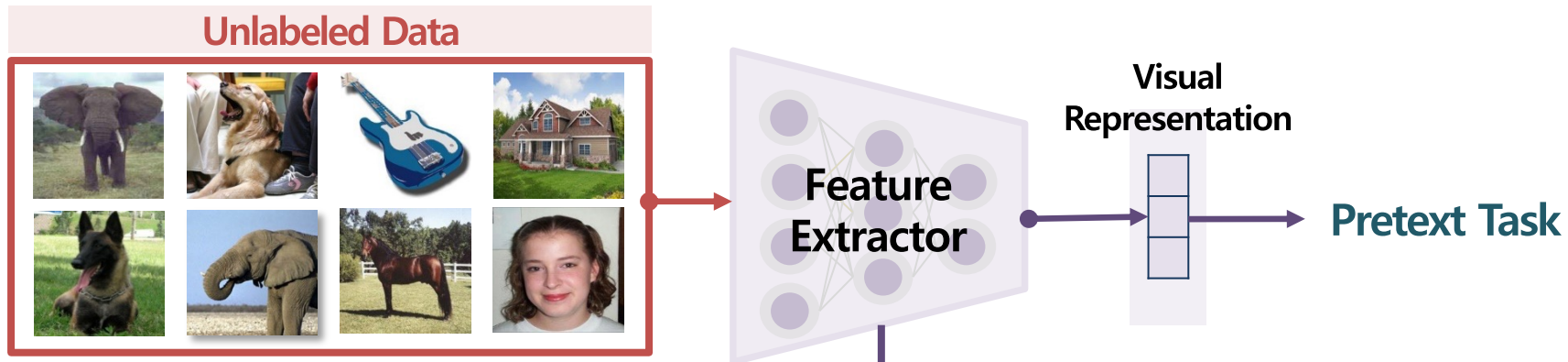
2020년 11월 20일
오후 1시 ~
온라인 비디오 시청 (YouTube)

세미나 정보 보기 →

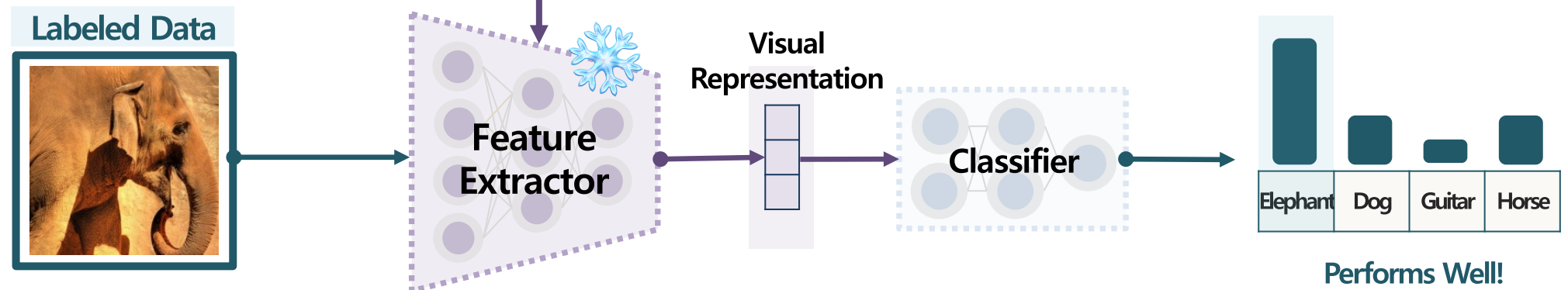
Self-Supervised Learning (SSL)

DINO Series

Pretrain



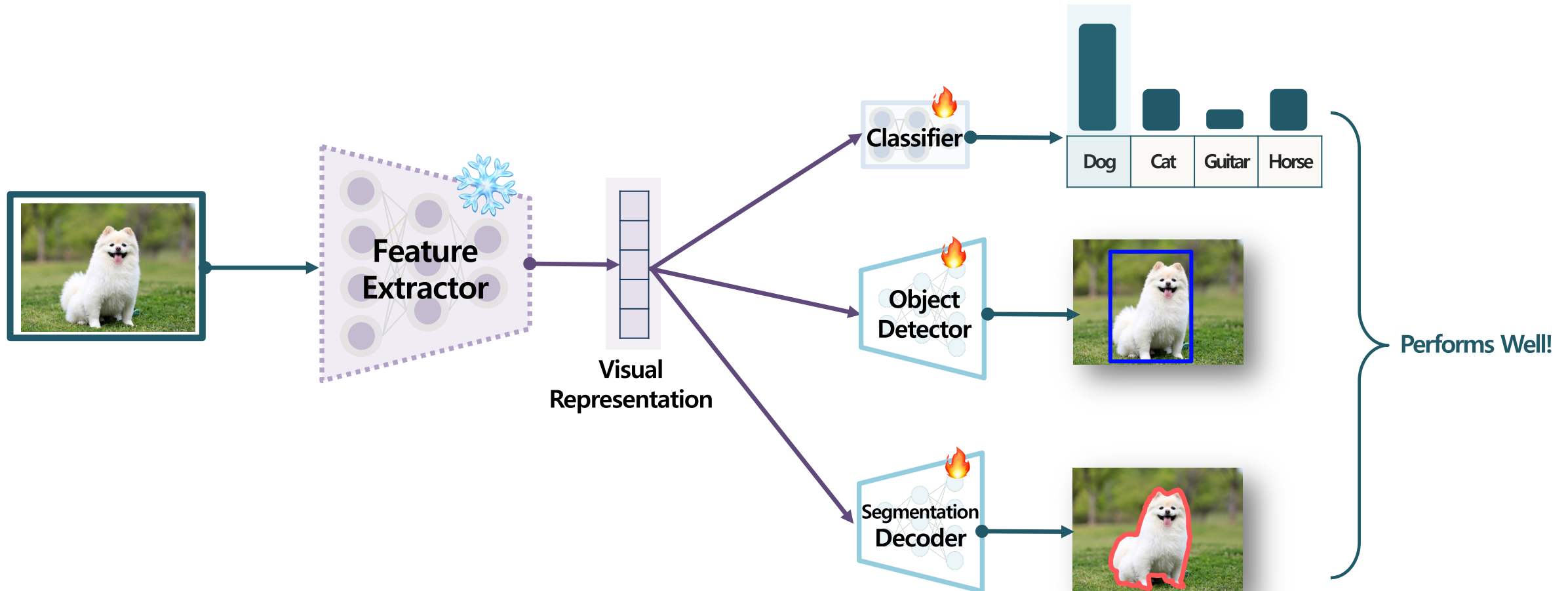
Finetune



Self-Supervised Learning

DINO Series

비교적 **소량의 labeled data**만 이용하여 **finetuning**해주어도,
각 **downstream task**에서 좋은 성능을 보임



Self-Supervised Learning

DINO Series

Pretrain

Unlabeled Data



Feature
Extractor

How does DINOv2 use unlabeled data?

Pretext Task

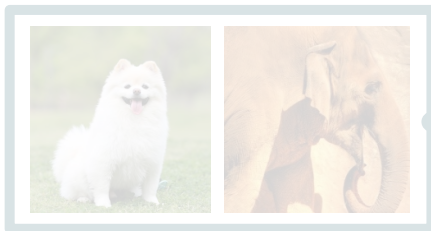


X

Y ??

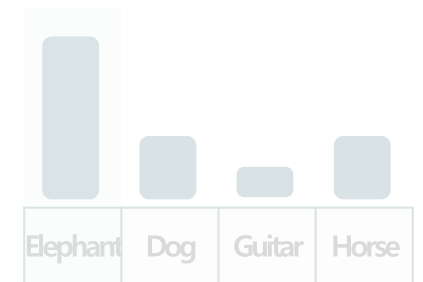
Finetune

Labeled Data



Feature
Extractor

Classifier



Performs Well!

DINOv2



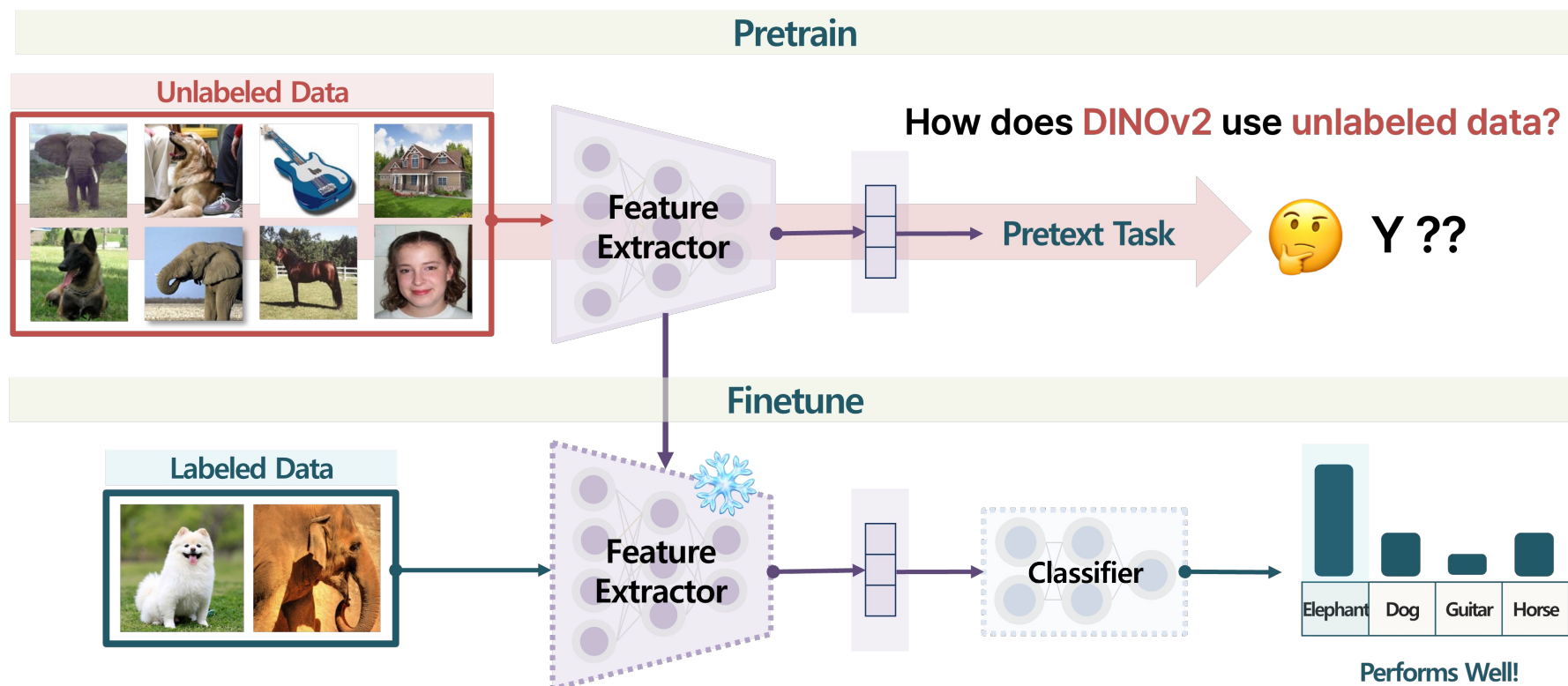
DINOv2

목적

Label 없이도 다양한 Vision Task에 적용 가능한 Feature Representation을 학습하자

방법

Data Curation, DINO loss, IBOT loss 를 이용하자!



DINOv2

Data Curation

목적

Label 없이도 다양한 Vision Task에 적용 가능한 Feature Representation을 학습하자

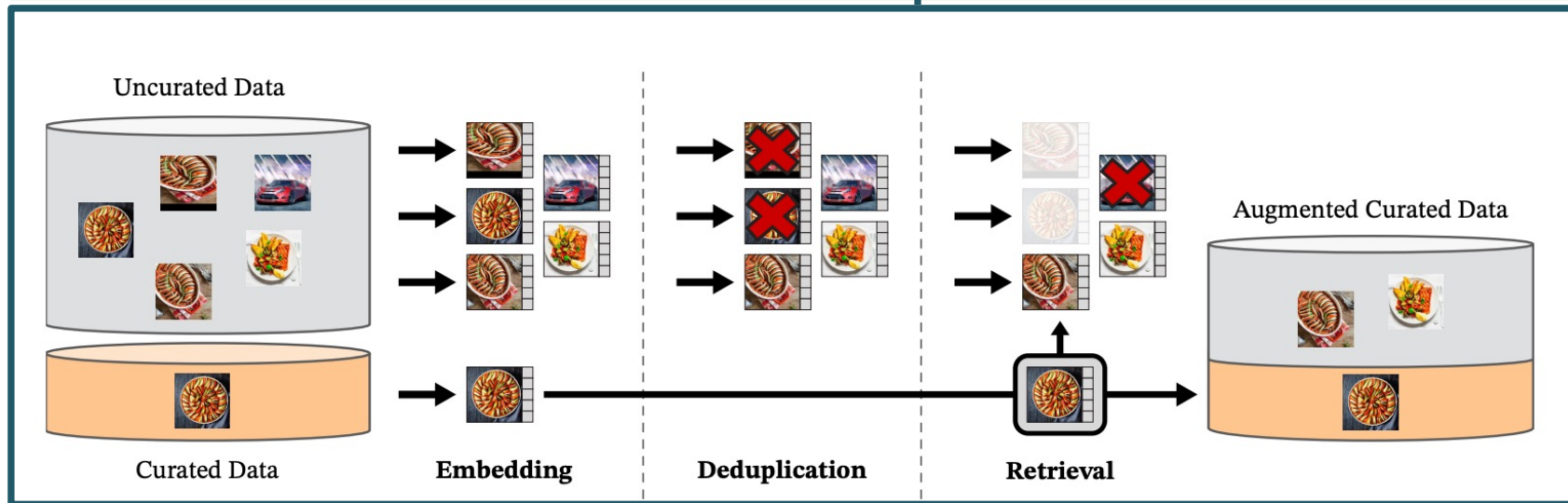
방법

Data Curation, DINO loss, IBOT loss 를 이용하자!

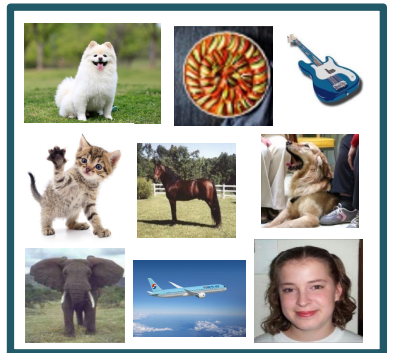
Data Curation For Large Scale SSL

데이터 **품질** 핵심 !

자동 필터링 균형화 Pipeline 구축



LVD-142M
dataset



** 이해를 돕기 위한 예시 이미지들

약 12억 장

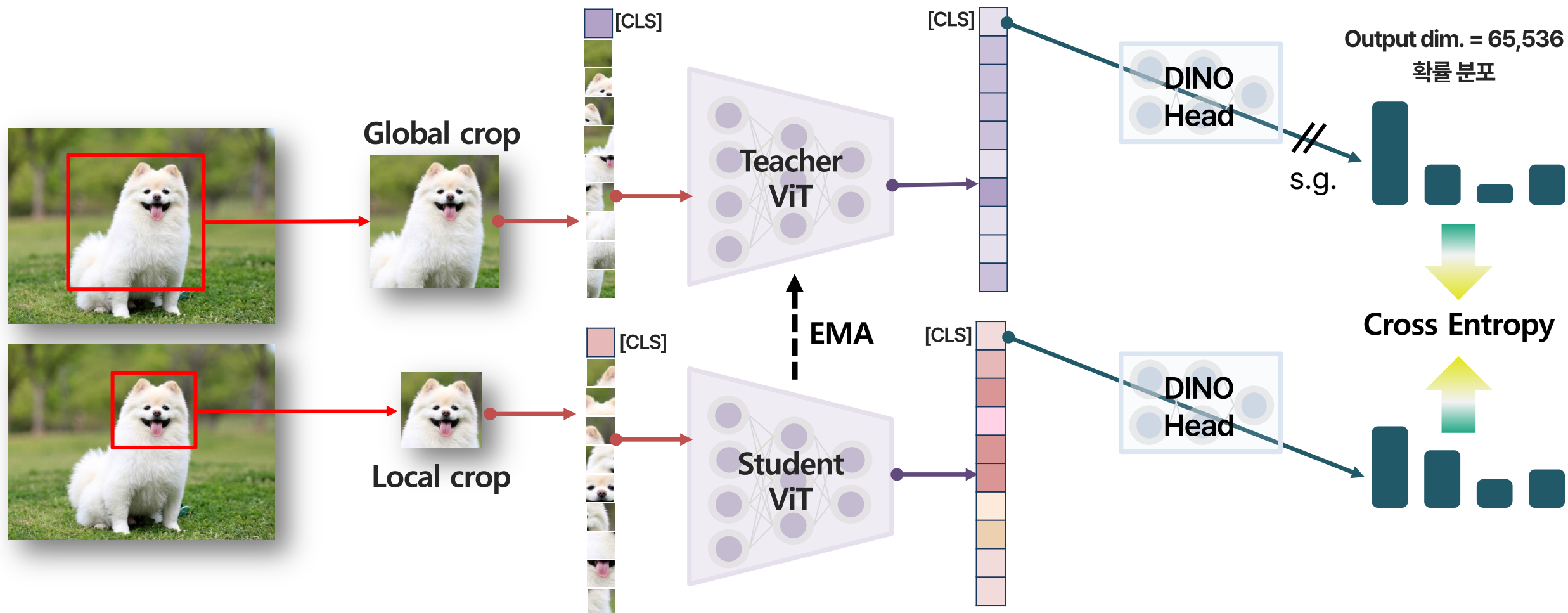
약 1.4억 장

DINO

Label 없이도 다양한 Vision Task에 적용 가능한 Feature Representation을 학습하자

Data Curation, DINO loss, IBOT loss를 이용하자!

DINO loss: Image level, Captures global image understanding



DINOv2

IBOT

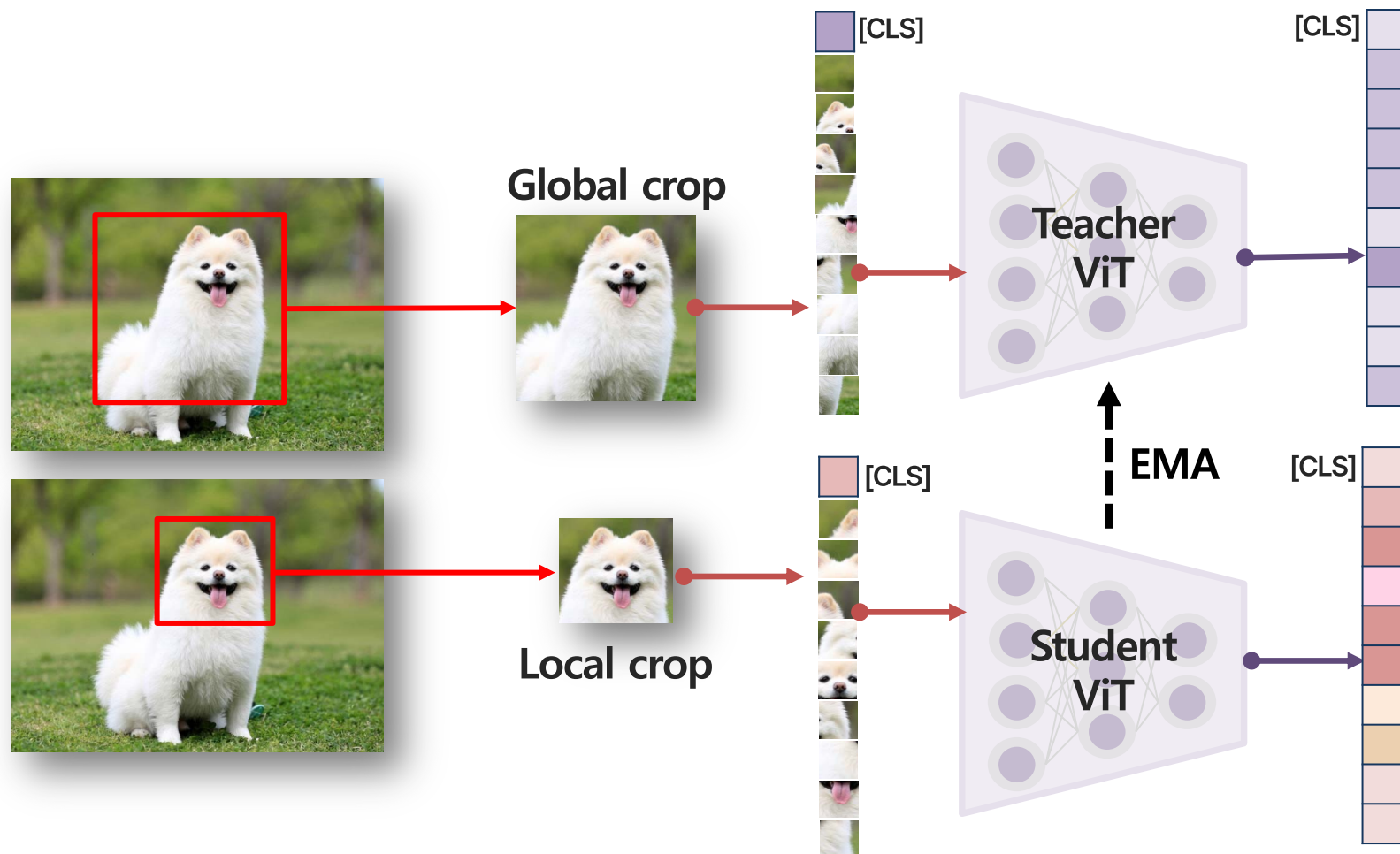
목적

Label 없이도 다양한 Vision Task에 적용 가능한 Feature Representation을 학습하자

방법

[Data Curation](#), [DINO loss](#), [IBOT loss](#) 를 이용하자!

IBOT loss: Patch level, Captures local details



DINOv2

IBOT

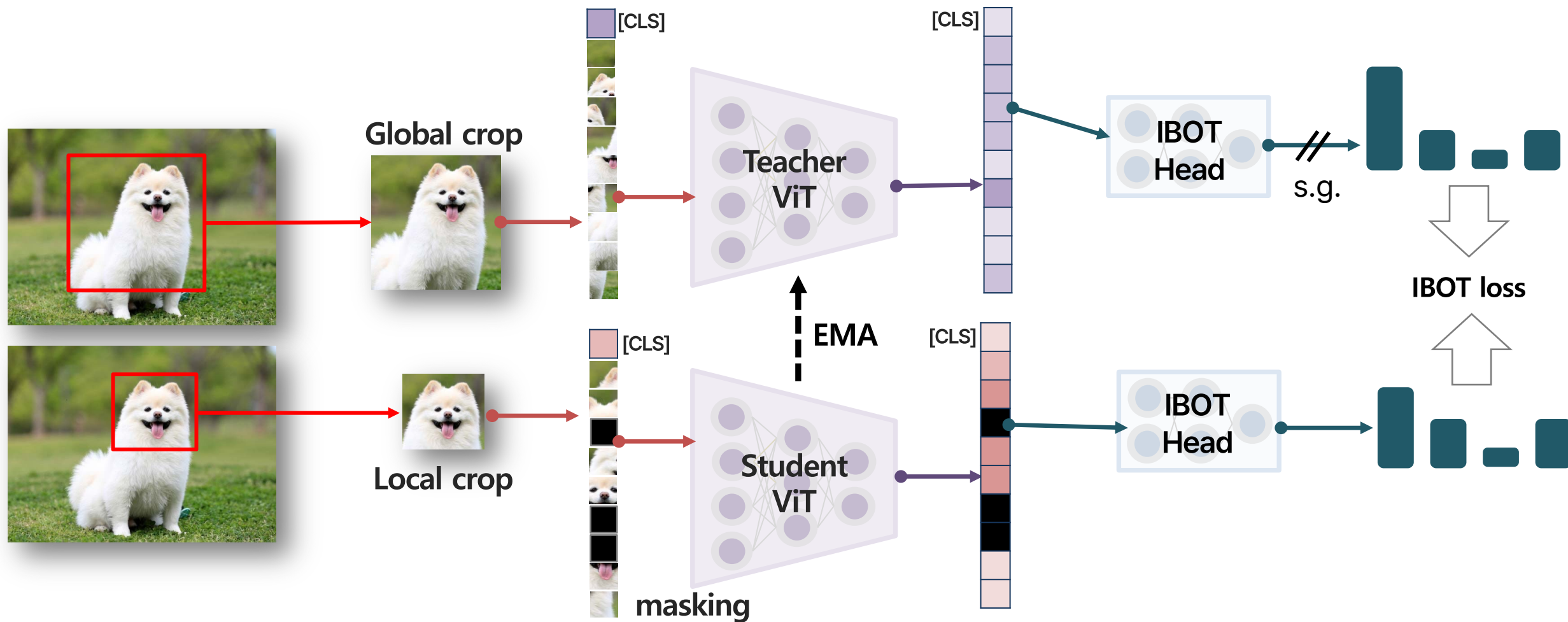
목적

Label 없이도 다양한 Vision Task에 적용 가능한 Feature Representation을 학습하자

방법

Data Curation, DINO loss, IBOT loss 를 이용하자!

IBOT loss: Patch level, Captures local details



DINOv2

IBOT

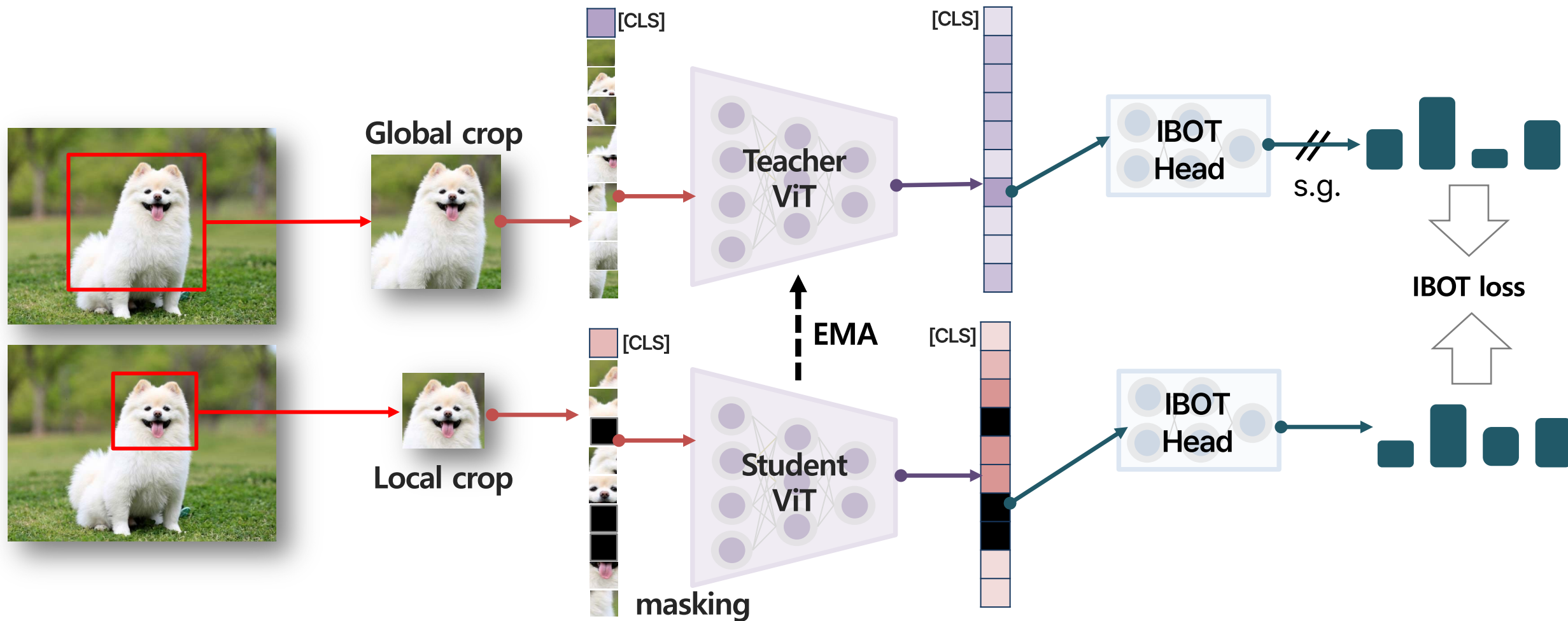
목적

Label 없이도 다양한 Vision Task에 적용 가능한 Feature Representation을 학습하자

방법

Data Curation, DINO loss, IBOT loss 를 이용하자!

IBOT loss: Patch level, Captures local details



DINOv2

IBOT

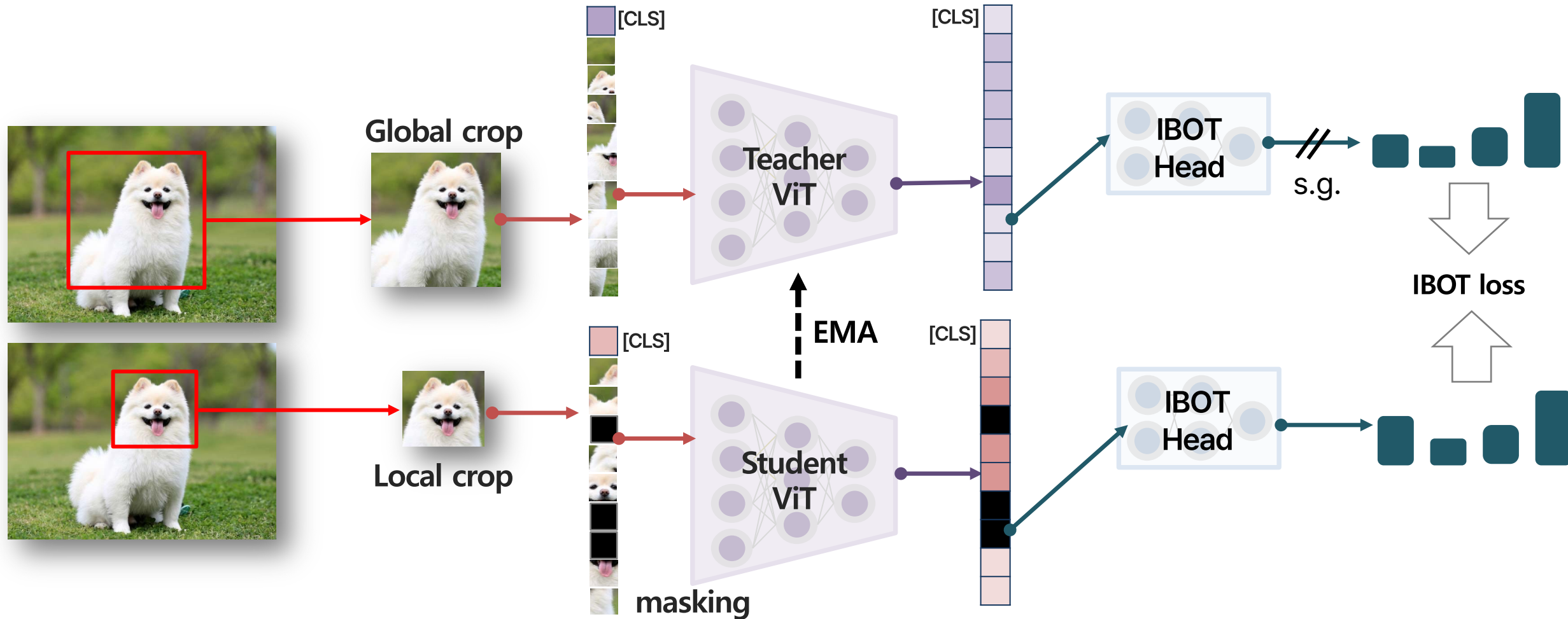
목적

Label 없이도 다양한 Vision Task에 적용 가능한 Feature Representation을 학습하자

방법

Data Curation, DINO loss, IBOT loss 를 이용하자!

IBOT loss: Patch level, Captures local details



DINOv2

Experiments

목적

Label 없이도 다양한 Vision Task에 적용 가능한 Feature Representation을 학습하자

방법

[Data Curation](#), [DINO loss](#), [IBOT loss](#) 를 이용하자!

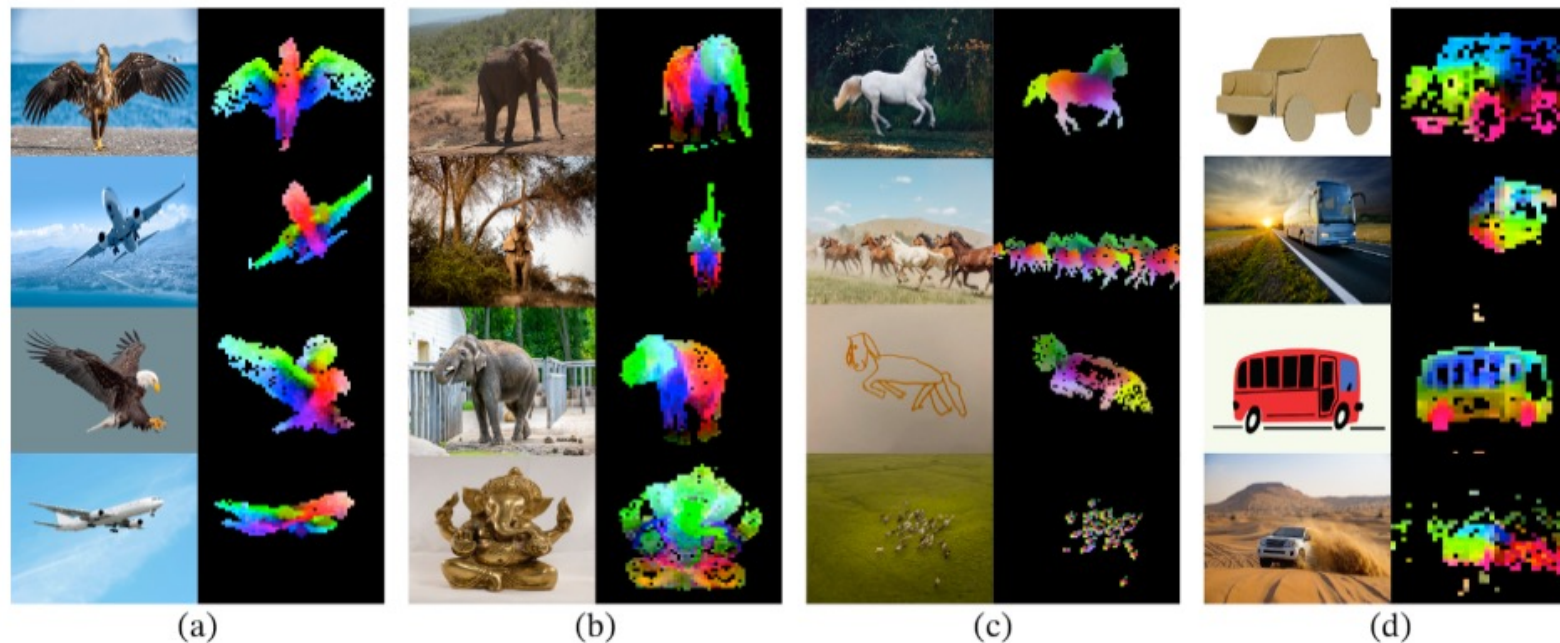


Figure 1: **Visualization of the first PCA components.** We compute a PCA between the patches of the images from the same column (a, b, c and d) and show their first 3 components. Each component is matched to a different color channel. Same parts are matched between related images despite changes of pose, style or even objects. Background is removed by thresholding the first PCA component.

DINOv2

Experiments

Linear Evaluation

Feature	Arch	Food	C10	C100	SUN	Cars	Aircr	VOC	DTD	Pets	Cal101	Flowers	CUB	Avg
OpenCLIP	ViT-G/14	94.5	98.7	91.0	84.0	96.1	80.2	89.3	86.0	95.7	98.1	99.5	89.9	91.9
MAE	ViT-H/14	78.4	96.1	83.9	63.9	56.1	63.4	84.3	75.4	89.4	95.9	92.3	57.2	78.0
DINO	ViT-B/8	85.1	97.2	86.9	70.3	76.6	70.6	86.7	79.6	93.2	95.4	97.6	81.7	85.1
iBOT	ViT-L/16	91.0	99.0	92.8	75.6	71.8	72.4	89.0	80.7	87.7	97.5	99.6	82.1	86.6
DINOv2	ViT-S/14	89.1	97.7	87.5	74.4	81.6	74.0	87.8	80.6	95.1	97.0	99.6	88.1	87.7
	ViT-B/14	92.8	98.7	91.3	77.3	88.2	79.4	88.2	83.3	96.2	96.1	99.6	89.6	90.1
	ViT-L/14	94.3	99.3	93.4	78.7	90.1	81.5	88.3	84.0	96.6	97.5	99.7	90.5	91.2
	ViT-g/14	94.7	99.5	94.4	78.7	91.4	87.2	89.0	84.5	96.7	97.6	99.7	91.6	92.1

Table 8: **Linear evaluation of frozen features on fine-grained benchmarks.** Accuracy on 12 benchmarks covering objects, scenes and textures following the evaluation protocol proposed in [Chen et al. \(2020\)](#).

목적 Label 없이도 다양한 Vision Task에 적용 가능한 Feature Representation을 학습하자

방법 [Data Curation](#), [DINO loss](#), [IBOT loss](#) 를 이용하자!

Semantic segmentation

Method	Arch.	ADE20k (62.9)		CityScapes (86.9)		Pascal VOC (89.0)	
		lin.	+ms	lin.	+ms	lin.	+ms
OpenCLIP	ViT-G/14	39.3	46.0	60.3	70.3	71.4	79.2
MAE	ViT-H/14	33.3	30.7	58.4	61.0	67.6	63.3
DINO	ViT-B/8	31.8	35.2	56.9	66.2	66.4	75.6
iBOT	ViT-L/16	44.6	47.5	64.8	74.5	82.3	84.3
DINOv2	ViT-S/14	44.3	47.2	66.6	77.1	81.1	82.6
	ViT-B/14	47.3	51.3	69.4	80.0	82.5	84.9
	ViT-L/14	47.7	53.1	70.3	80.9	82.1	86.0
	ViT-g/14	49.0	53.0	71.3	81.0	83.0	86.2

Table 10: **Semantic segmentation on ADE20K, CityScapes and Pascal VOC with frozen features**

DINOv2

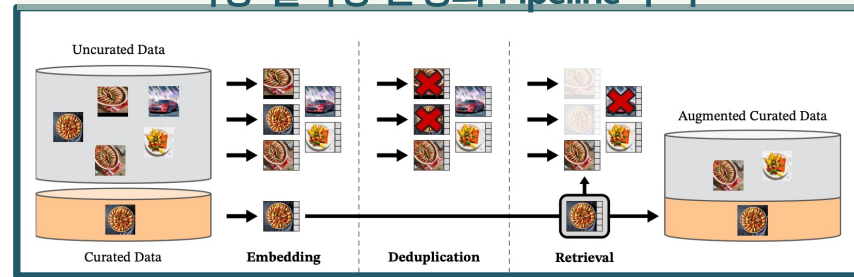
목적

Label 없이도 다양한 Vision Task에 적용 가능한 Feature Representation을 학습하자

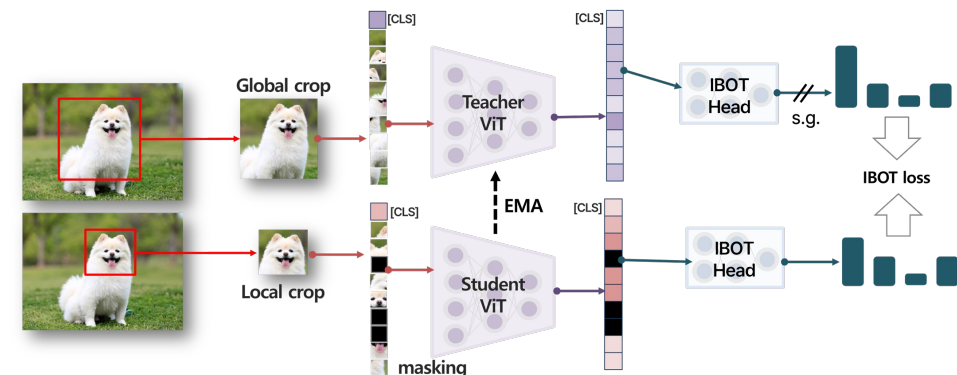
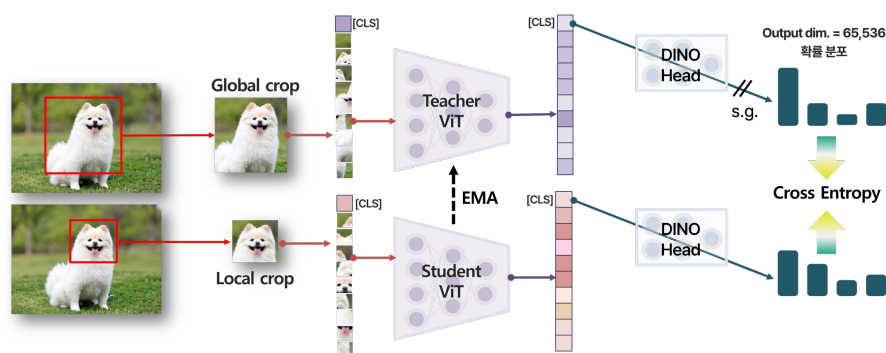
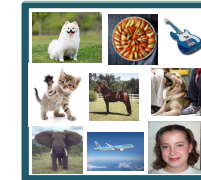
방법

Data Curation, DINO loss, IBOT loss 를 이용하자!

자동 필터링 균형화 Pipeline 구축



LVD dataset



“비지도 학습만으로도 강력한 범용 Visual Representation을 만들 수 있다”는 가능성 실증

DINOv3



DINOv3

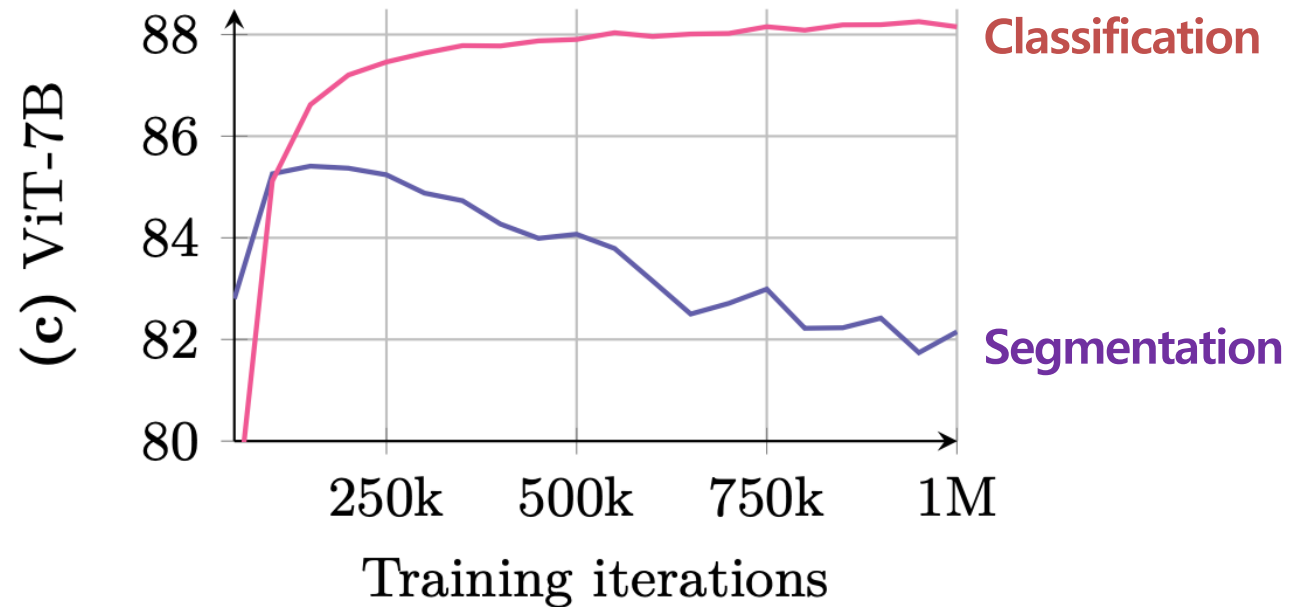
Issues with scaling Up DINOv2

배경

DINOv2에서 모델을 크게, 데이터셋을 더 많이 써도 dense feature가 오히려 망가짐

목표

문제 해결 & 더 크고 안정적인 self-supervised foundation model을 만들자!



학습이 진행될수록
dense task의 성능 저하

DINOv3

Key Components

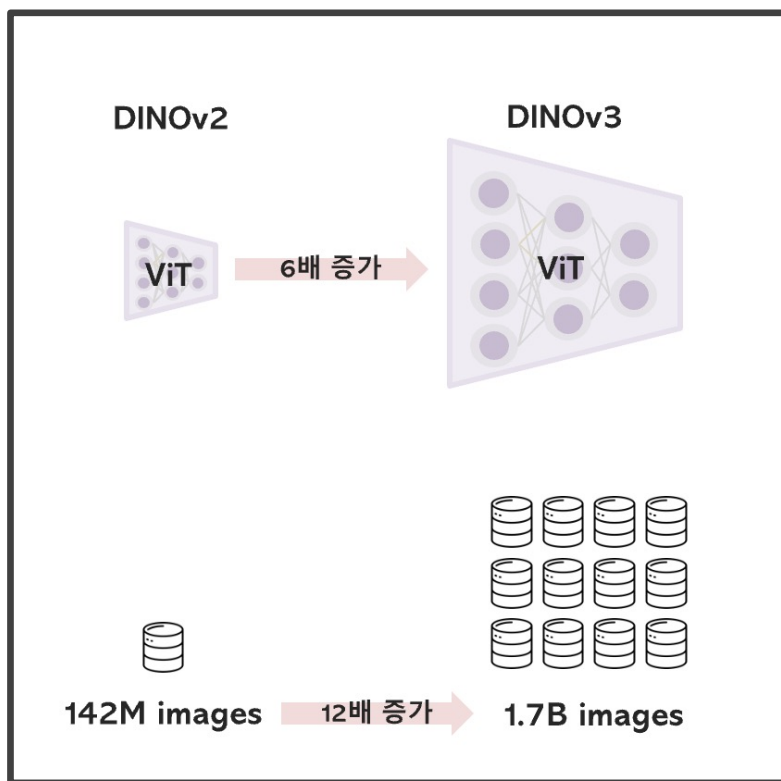
배경

DINOv2에서 모델을 크게, 데이터셋을 더 많이 써도 dense feature가 오히려 망가짐

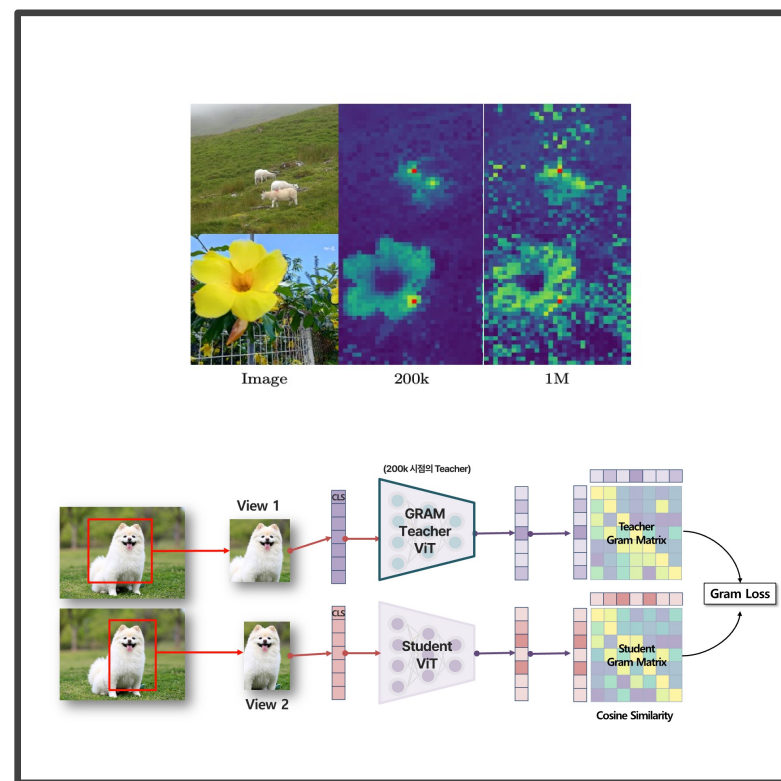
목표

문제 해결 & 더 크고 안정적인 self-supervised foundation model을 만들자!

Scaling



Gram anchoring



DINOv3

Scaling Dataset & model size

배경

DINOv2에서 모델을 크게, 데이터셋을 더 많이 써도 dense feature가 오히려 망가짐

목표

문제 해결 & 더 크고 안정적인 self-supervised foundation model을 만들자!

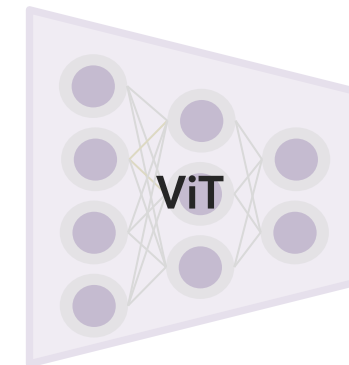
Teacher model	DINOv2	DINOv3
Backbone	ViT-giant	ViT-7B
#Params	1.1B	6.7B
#Blocks	40	40
Patch Size	14	16
Pos. Embeddings	Learnable	RoPE
Registers	4	4
Embed. Dim.	1536	4096
FFN Type	SwiGLU	SwiGLU
FFN Hidden Dim.	4096	8192
Attn. Heads	24	32
Attn. Heads Dim.	64	128
DINO Head MLP	4096-4096-256	8192-8192-512
DINO Prototypes	128k	256k
iBOT Head MLP	4096-4096-256	8192-8192-384
iBOT Prototypes	128k	96k

DINOv2



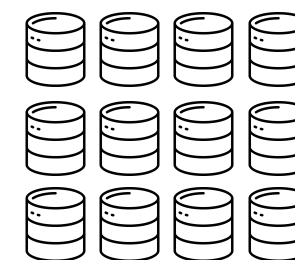
6배 증가

DINOv3



142M images

12배 증가



1.7B images

DINOv3

Gram Anchoring

배경

DINOv2에서 모델을 크게, 데이터셋을 더 많이 써도 dense feature가 오히려 망가짐

목표

문제 해결 & 더 크고 안정적인 self-supervised foundation model을 만들자!

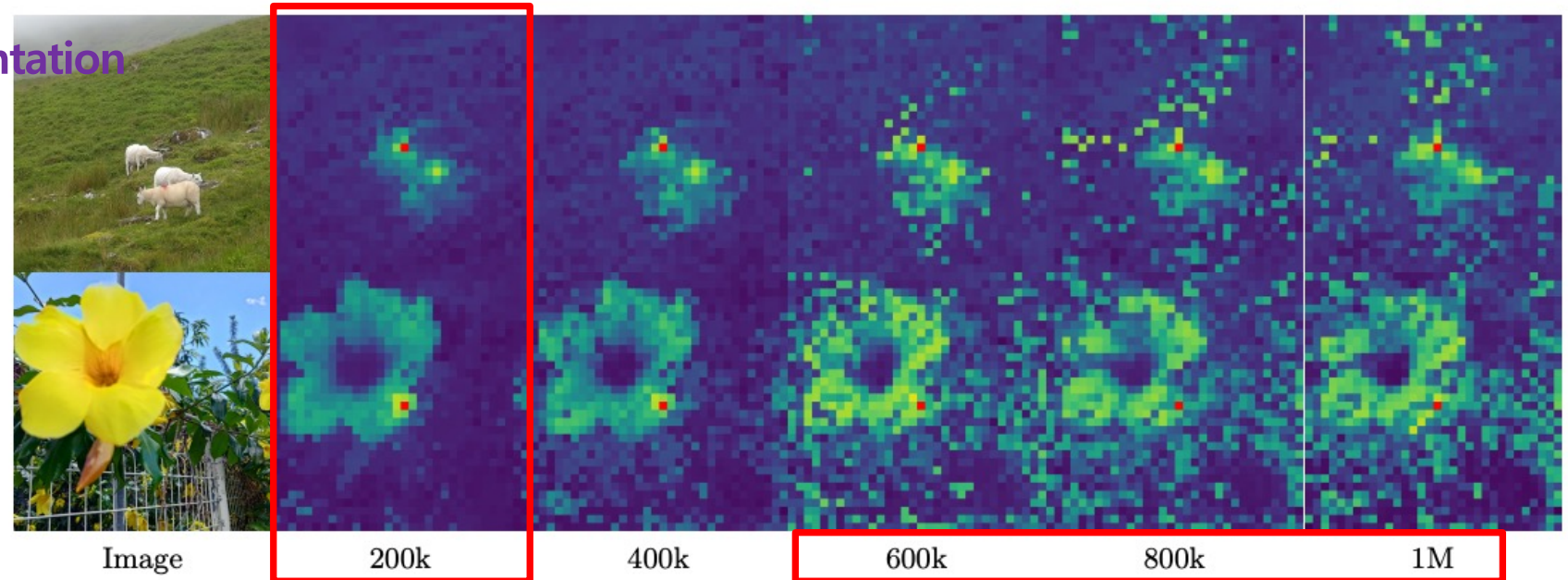
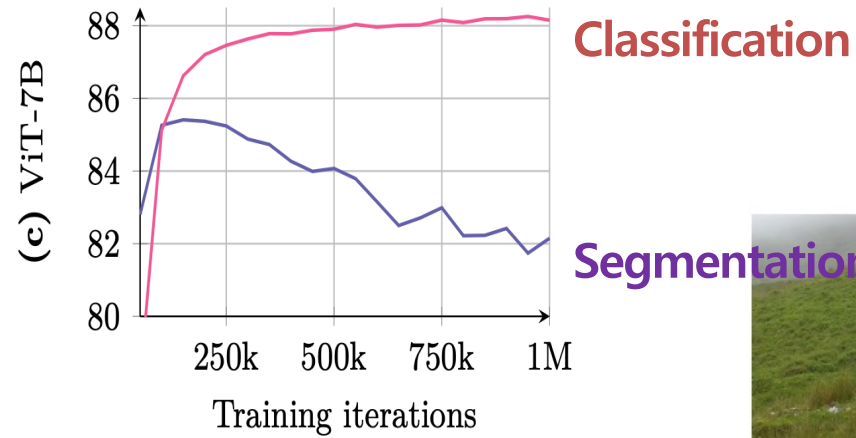
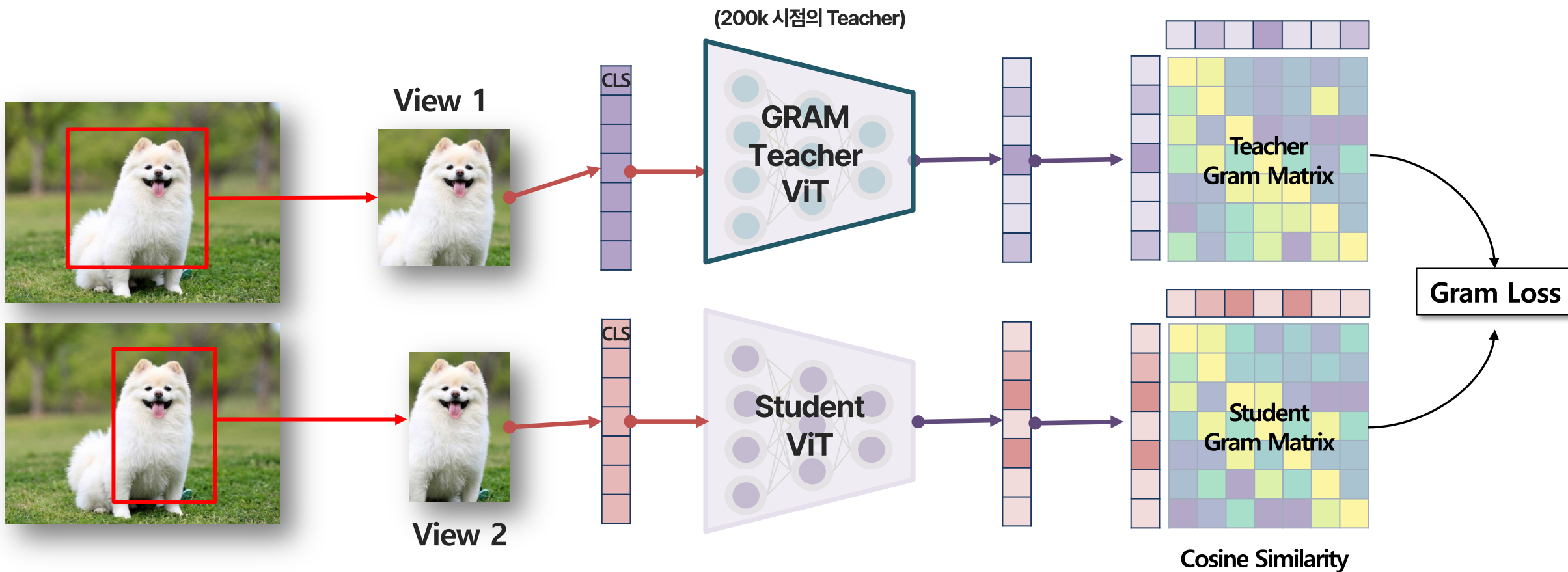
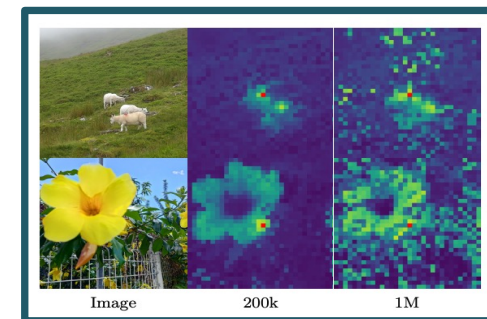


Figure 6: Evolution of the cosine similarity between the patch noted in red and all other patches. As training progresses, the features produced by the model become less localized and the similarity maps become noisier.

DINOv3

Gram Anchoring

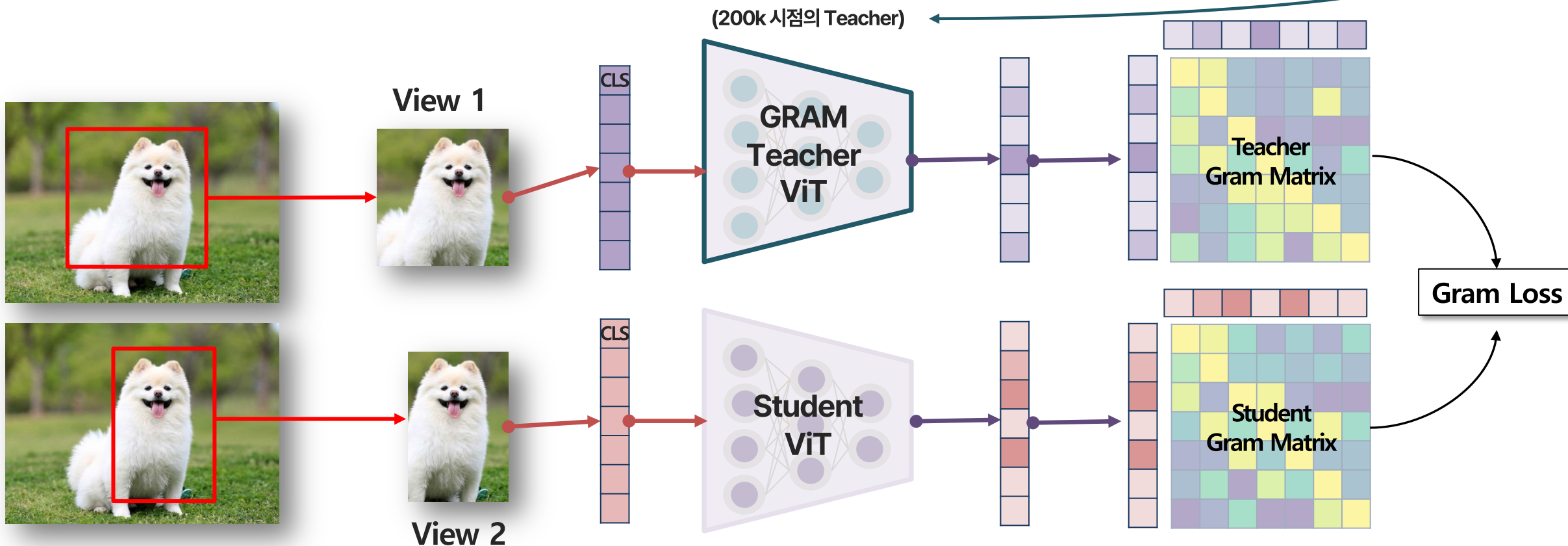
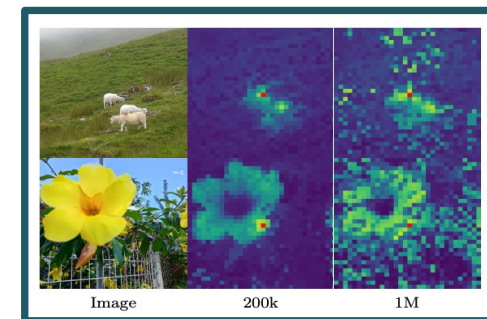
문제 상황



DINOv3

Gram Anchoring

문제 상황



DINOv3

Gram Anchoring

배경

DINOv2에서 모델을 크게, 데이터셋을 더 많이 써도 dense feature가 오히려 망가짐

목표

문제 해결 & 더 크고 안정적인 self-supervised foundation model을 만들자!

1) Quantitative effect of Gram anchoring

어느 시점의 모델을 Gram teacher로 사용?

Method	Teacher Iteration	Res.	IN1k Linear	ADE mIoU	NYU RMSE
Baseline	—	—	88.2	50.3	0.307
GRAM	200k	×1	88.0	53.6	0.285
GRAM	200k	×2	88.0	55.7	0.281
GRAM	100k	×2	87.9	55.7	0.284
GRAM	1M	×2	88.1	54.9	0.290

Semantic segmentation 성능 (↑)

(b) Ablation of Gram teachers and resolutions.

Depth estimation 성능 (↓)

→ feature가 잘 유지되어 있는 학습 초반의 teacher를 anchor로 사용하는 것이 가장 효과적이다

DINOv3

Gram Anchoring

배경

DINOv2에서 모델을 크게, 데이터셋을 더 많이 써도 dense feature가 오히려 망가짐

목표

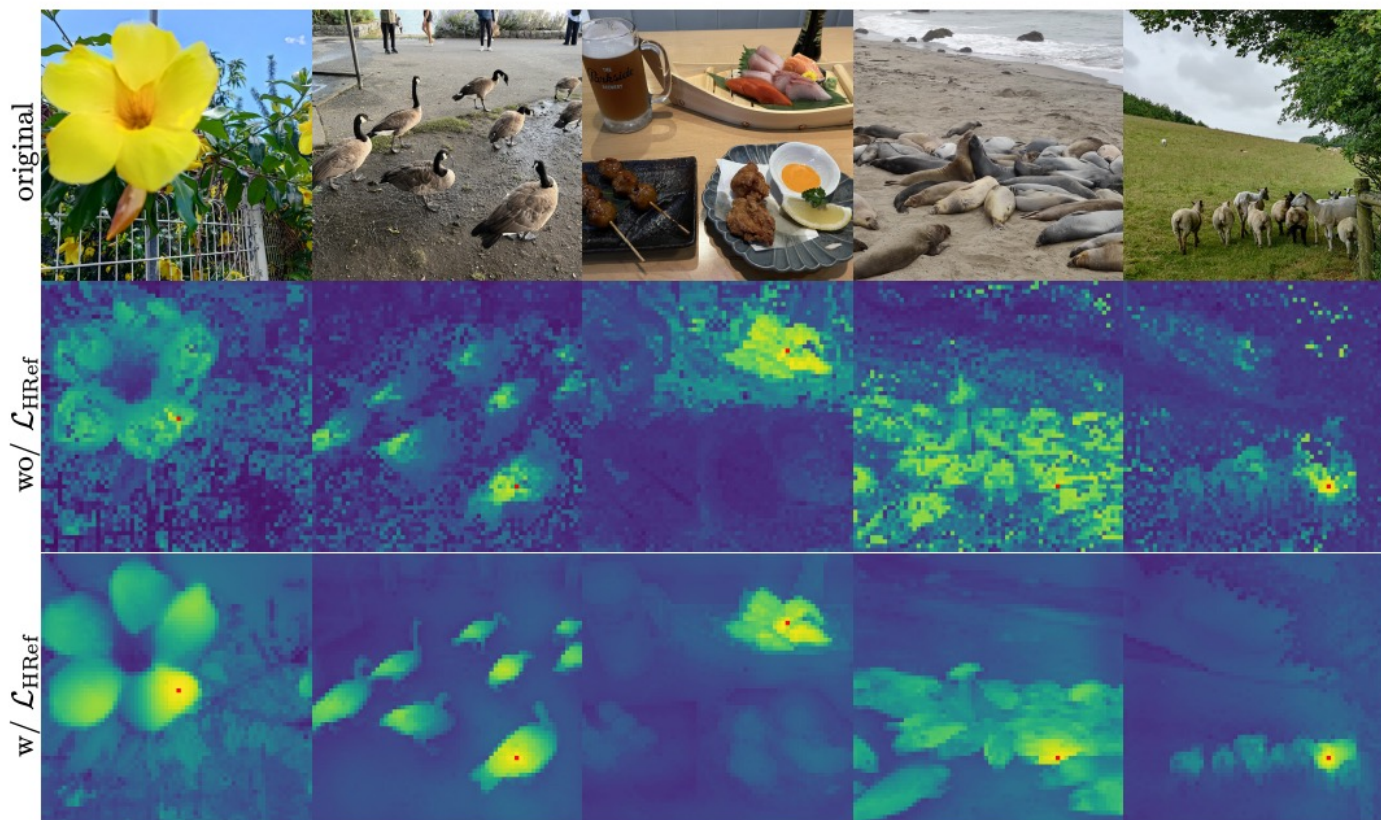
문제 해결 & 더 크고 안정적인 self-supervised foundation model을 만들자!

2) Qualitative effect of Gram anchoring

Before



After



DINOv3

Gram Anchoring

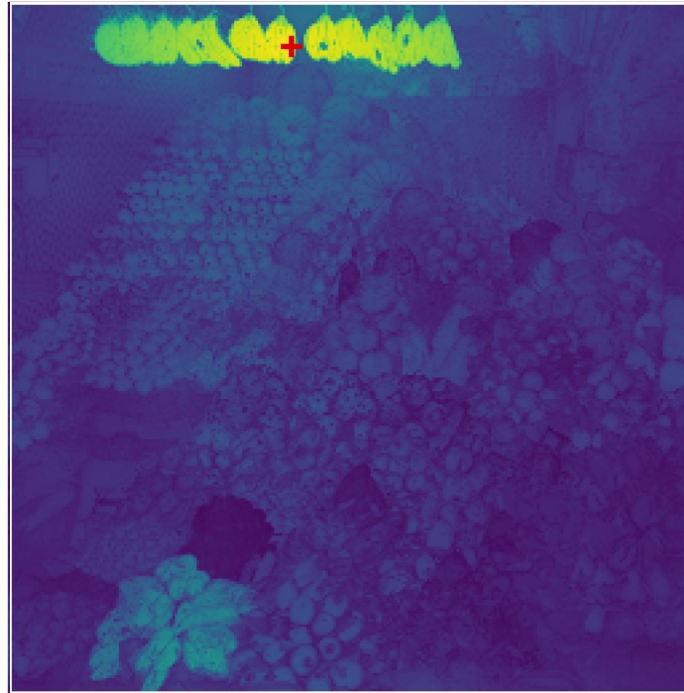
배경

DINOv2에서 모델을 크게, 데이터셋을 더 많이 써도 dense feature가 오히려 망가짐

목표

문제 해결 & 더 크고 안정적인 self-supervised foundation model을 만들자!

3) High-resolution dense features



cosine similarity maps obtained with DINOv3 output features
between the patches marked with a red cross and all other patches.

DINOv3

Experiments

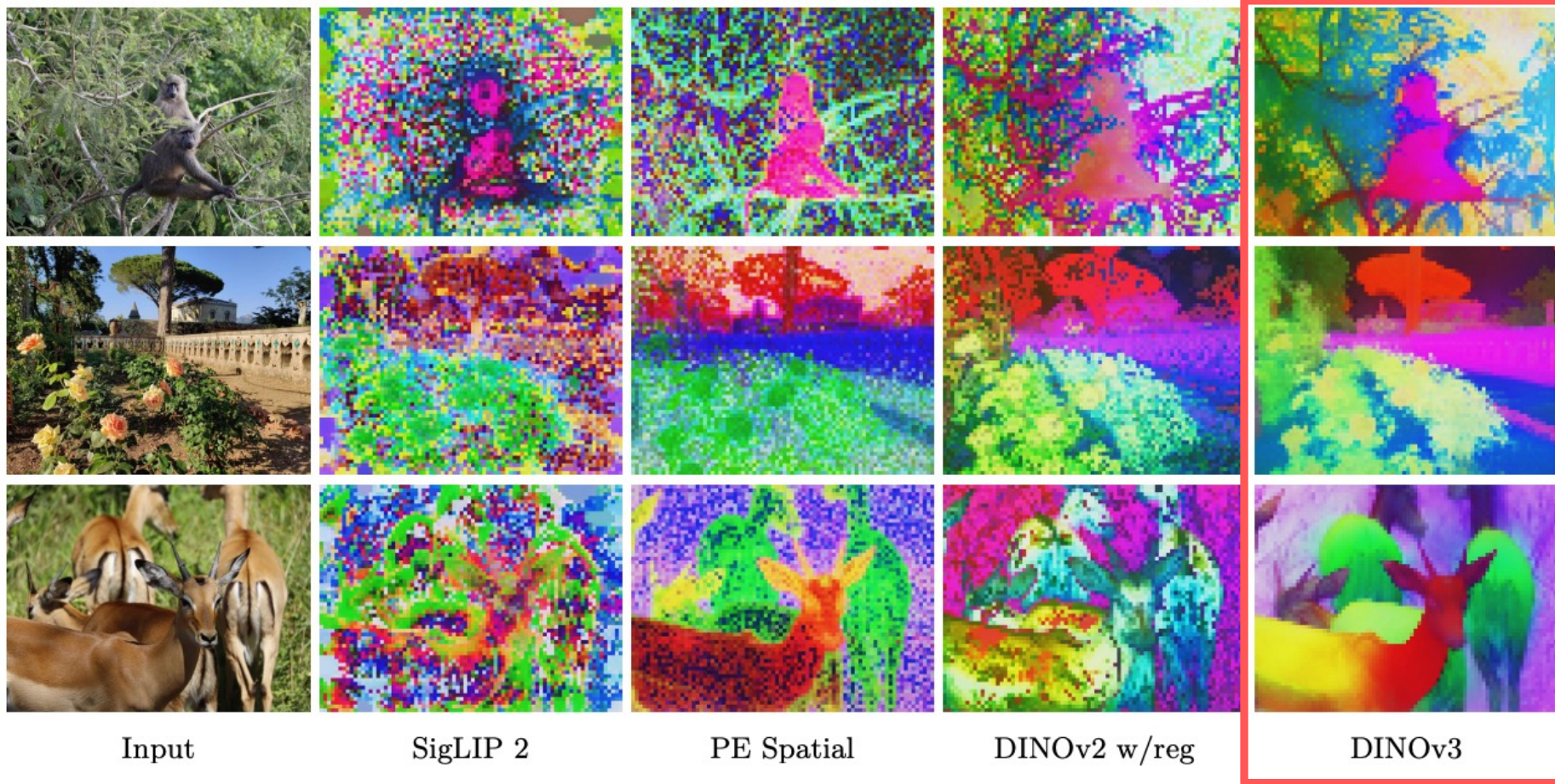
배경

DINOv2에서 모델을 크게, 데이터셋을 더 많이 써도 dense feature가 오히려 망가짐

목표

문제 해결 & 더 크고 안정적인 self-supervised foundation model을 만들자!

4) Comparison of dense features



더 깨끗하고 일관된
dense feature map 생성

5) Comparison with state-of-the-art systems on object detection

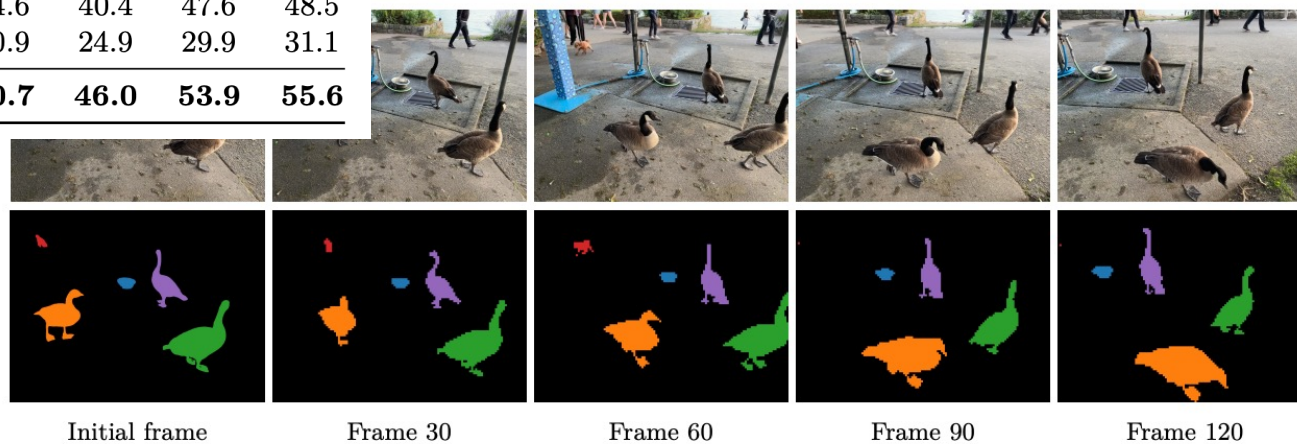
Model	Detector	FT	Parameters			COCO		COCO-O	
			Encoder	Decoder	Trainable	Simple	TTA	mAP	ER
EVA-02	Cascade	🔥	300M	—	300M	64.1	—	63.6	34.7
InternImage-G	DINO	🔥	6B	—	6B	65.1	65.3	—	—
EVA-02	Co-DETR	🔥	300M	—	300M	65.4	65.9	63.7	34.3
PEspatial	DETA	🔥	1.9B	50M	2B	65.3	66.0	64.0	34.7
DINOv3	Plain-DETR	❄️	7B	100M	100M	65.6	66.1	66.4	36.8

Backbone을 완전히 fine-tuning한 모델들에 비해서도 더 높은 성능 달성

→ DINOv3의 feature representation 품질이 강력하다는 것

6) Video segmentation tracking evaluation

		DAVIS			YouTube-VOS			MOSE		
Method	ViT	S	M	L	S	M	L	S	M	L
<i>Agglomerative backbones</i>										
AM-RADIOv2.5	g/14	66.5	77.3	81.4	70.1	78.1	79.2	44.0	52.6	54.3
PEspatial	G/14	68.4	74.5	70.5	68.5	67.5	55.6	39.3	40.2	34.0
<i>Weakly-supervised backbones</i>										
SigLIP 2	g/16	56.1	62.3	62.9	52.0	57.3	55.1	28.0	30.3	29.2
PEcore	G/14	48.2	53.1	49.8	34.7	33.0	25.3	17.8	19.0	15.4
<i>Self-supervised backbones</i>										
Franca	g/14	61.8	66.9	66.5	67.3	70.5	67.9	40.3	42.6	41.9
DINOv2	g/14	63.9	73.6	76.6	65.6	73.5	74.6	40.4	47.6	48.5
Web-DINO	7B/14	57.2	65.8	69.5	43.9	49.6	50.9	24.9	29.9	31.1
DINOv3	7B/16	71.1	79.7	83.3	74.1	80.2	80.7	46.0	53.9	55.6



7) Comparison of family of models against open-source alternatives of comparable size.

Size	Model	Global Tasks				Dense Tasks				
		IN-ReaL	IN-R	Obj.	Ox.-H	ADE20k	NYU↓	DAVIS	NAVI	SPair
S	DINOv2	87.3	54.0	47.8	39.5	45.5	0.446	73.6	53.4	51.6
S	DINOv3	87.0	60.4	50.9	49.5	47.0	0.403	72.7	56.3	50.4
S+	DINOv3	88.0	68.8	54.6	50.0	48.8	0.399	75.5	57.1	55.2
B	PEcore	87.5	68.4	57.9	20.2	37.4	0.641	44.5	41.8	13.7
B	SigLIP 2	89.3	80.6	66.9	20.2	41.6	0.512	63.2	45.4	32.8
B	DINOv2	89.0	68.4	57.3	51.0	48.4	0.416	72.9	56.9	57.1
B	DINOv3	89.3	76.7	64.1	58.5	51.8	0.373	77.2	58.8	57.2
L	PEcore	90.1	87.7	74.9	25.6	39.7	0.650	48.2	42.1	19.2
L	SigLIP 2	90.1	89.2	75.0	21.4	43.6	0.484	66.3	47.8	41.9
L	DINOv2	89.7	79.1	64.7	55.7	48.8	0.394	73.4	59.9	57.0
L	DINOv3	90.2	88.1	74.8	63.1	54.9	0.352	79.9	62.3	61.2
SO400m	SigLIP 2	90.3	90.4	76.2	23.0	44.0	0.402	64.8	48.8	38.7
H+	DINOv3	90.3	90.0	78.6	64.5	54.8	0.352	79.3	63.3	56.3

Global task와 Dense task에서 모두 우수한 성능을 보임

Conclusion

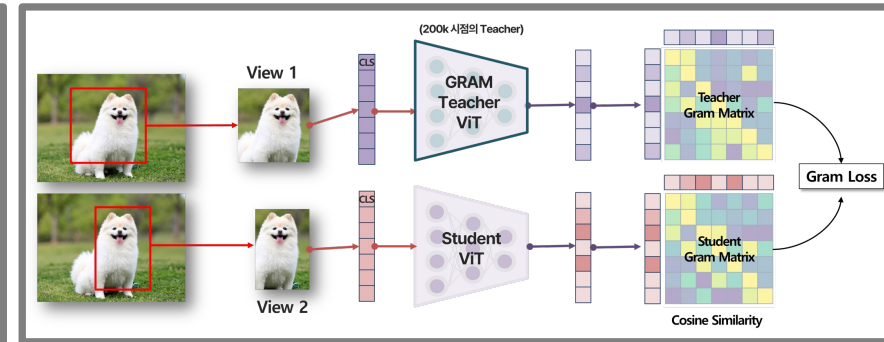
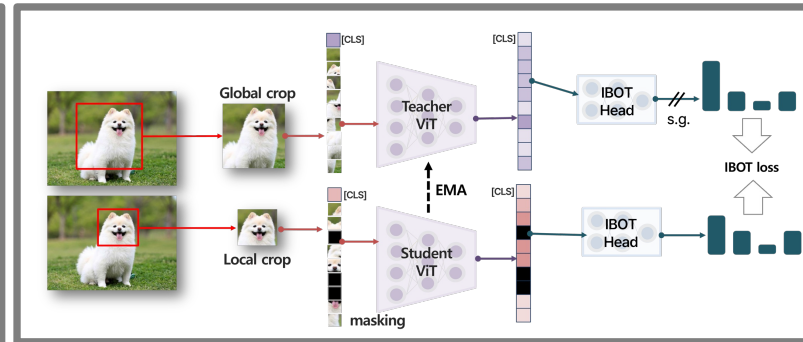
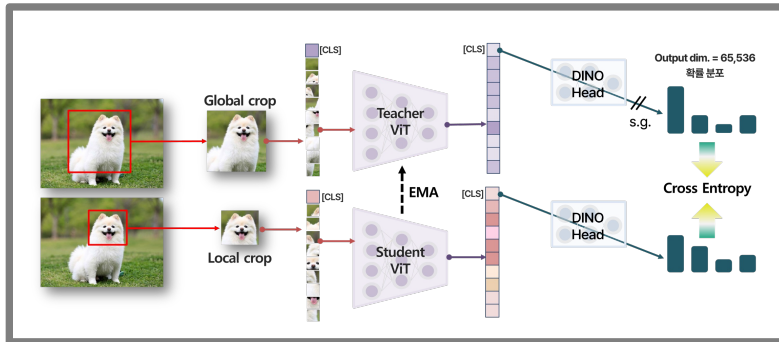
DINOv3, a Vision Foundation Model

배경

DINOv2에서 모델을 크게, 데이터셋을 더 많이 써도 dense feature가 오히려 망가짐

목표

문제 해결 & 더 크고 안정적인 self-supervised foundation model을 만들자!



No labels, yet powerful!

→ DINOv3 proves the true potential of self-supervised vision learning

Thank you

