上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Lecture 19: Self-Supervised Learning

Tao Huang

John Hopcroft Center, School of Computer Science, Shanghai Jiao Tong University

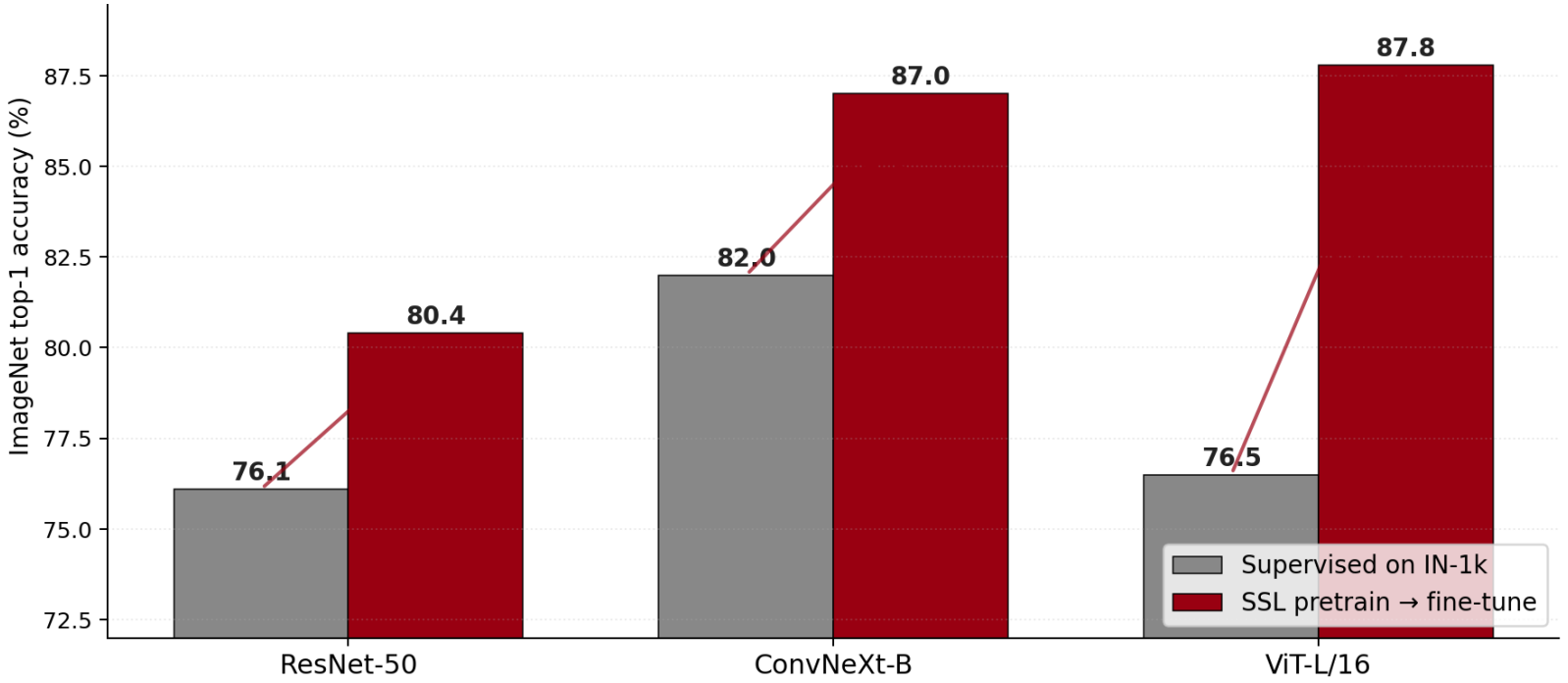
https://taohuang.info/cs3317

https://oc.sjtu.edu.cn/courses/89538

AI tools assisted in generating some figures in these slides. All such content has been reviewed, and the instructor is responsible for its accuracy.

The Promise from L18

- L18 closed with a claim: "the pretraining recipe makes both ConvNeXt and ViT actually work."
- That recipe has a name. It is self-supervised pretraining.
- Same architecture. Same fine-tuning. The 4–6 point gap is the recipe.



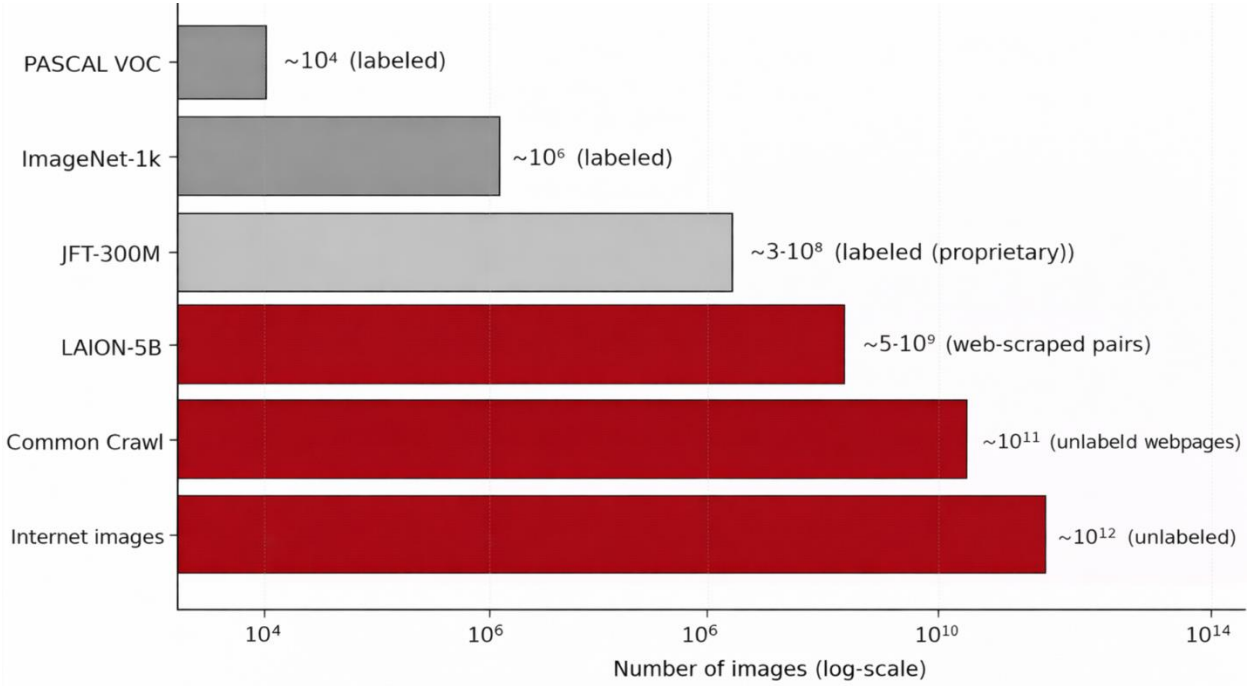
A Pile of Unlabeled Pixels

- We sit on six orders of magnitude more unlabeled data than labeled.
- But: Supervised learning ignores it.
- Question: *Can we learn useful features from raw pixels — without a single label?*

Some numbers:

- ImageNet-1k: 1.2M images
- LAION-5B: 5 billion image–text pairs, no manual labels
- YouTube: ~1 hour of video uploaded per second

Internet: ~10^{12} images, free



99.999% of the world's pixels never get a human label.

Objectives

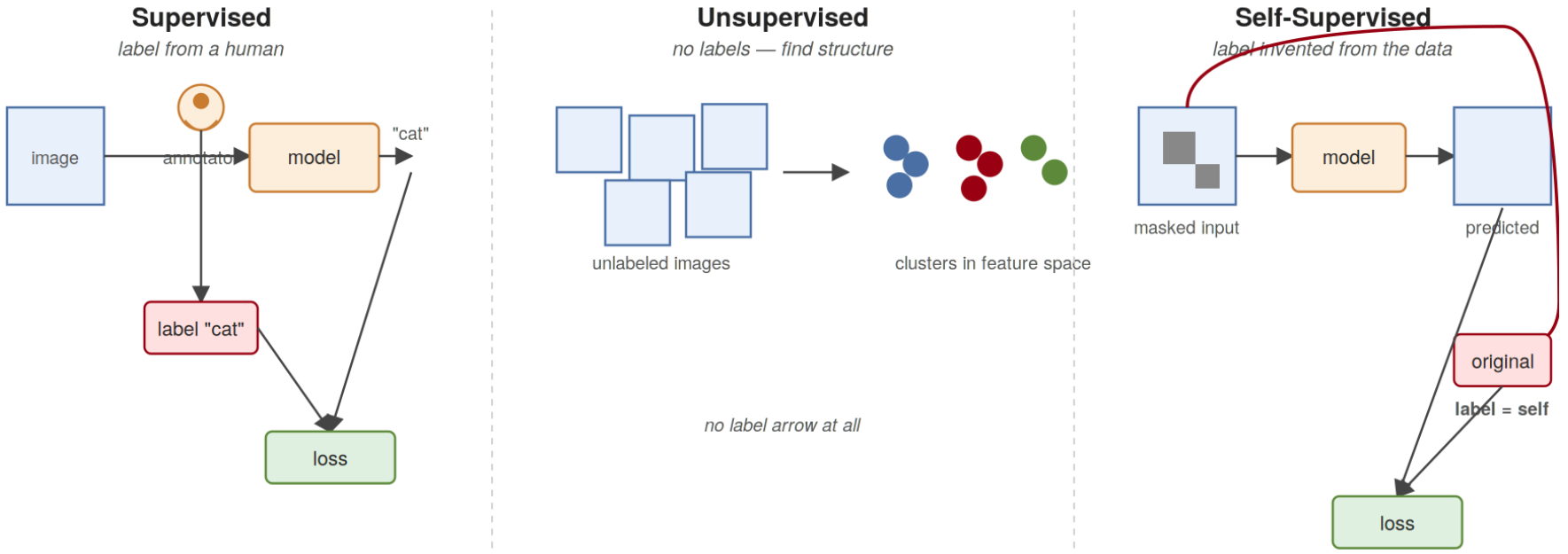
By the end of this lecture, you should be able to:

- Distinguish supervised, unsupervised, and self-supervised learning by where the "label" comes from.
- Construct a pretraining task from raw data: contrastive (SimCLR), predictive (BYOL), generative (MAE).
- Derive the InfoNCE loss and explain why temperature τ matters.
- Evaluate when to use SSL vs. supervised pretraining vs. training from scratch on a real budget.

1. What Even Is Self-Supervision?

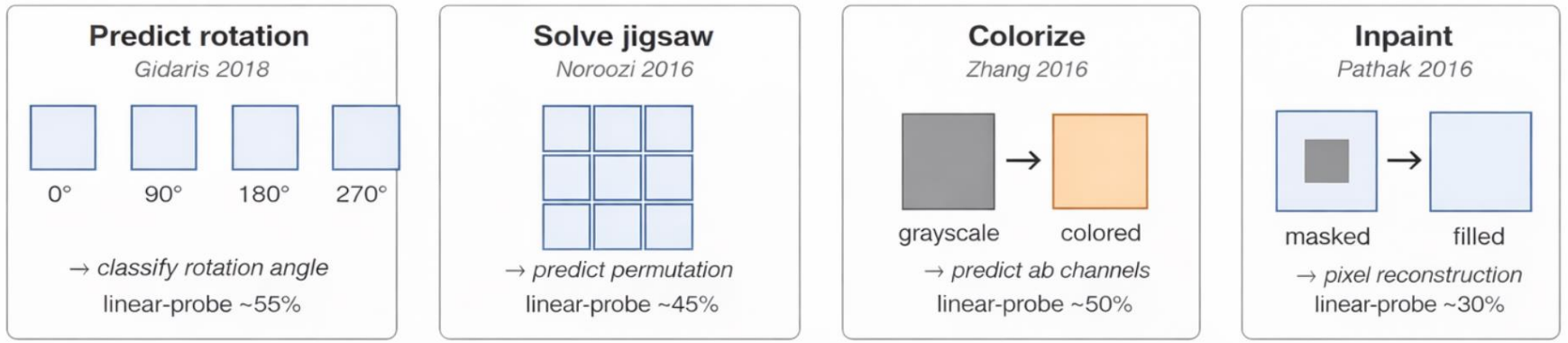
Three Paradigms — Where the Label Comes From

Paradigm	Where the label comes from	Example
Supervised	A human, externally	ImageNet-1k (cat/dog/…)
Unsupervised	No label at all — find structure	k-means, PCA
Self-supervised	**From the data itself — invented by you, free**	**Predict masked words; rotated image**



Early Pretext Tasks — Clever, but Wrong

- **The first wave (2014–2018):** invent a puzzle whose solution requires understanding.



For reference: supervised pretraining ≈ 76% linear-probe

All clever pretext tasks plateaued well below this.

Why? The pretext task was too narrow.

Models learned rotation cues (sky on top, gravity, faces) — not general visual concepts.

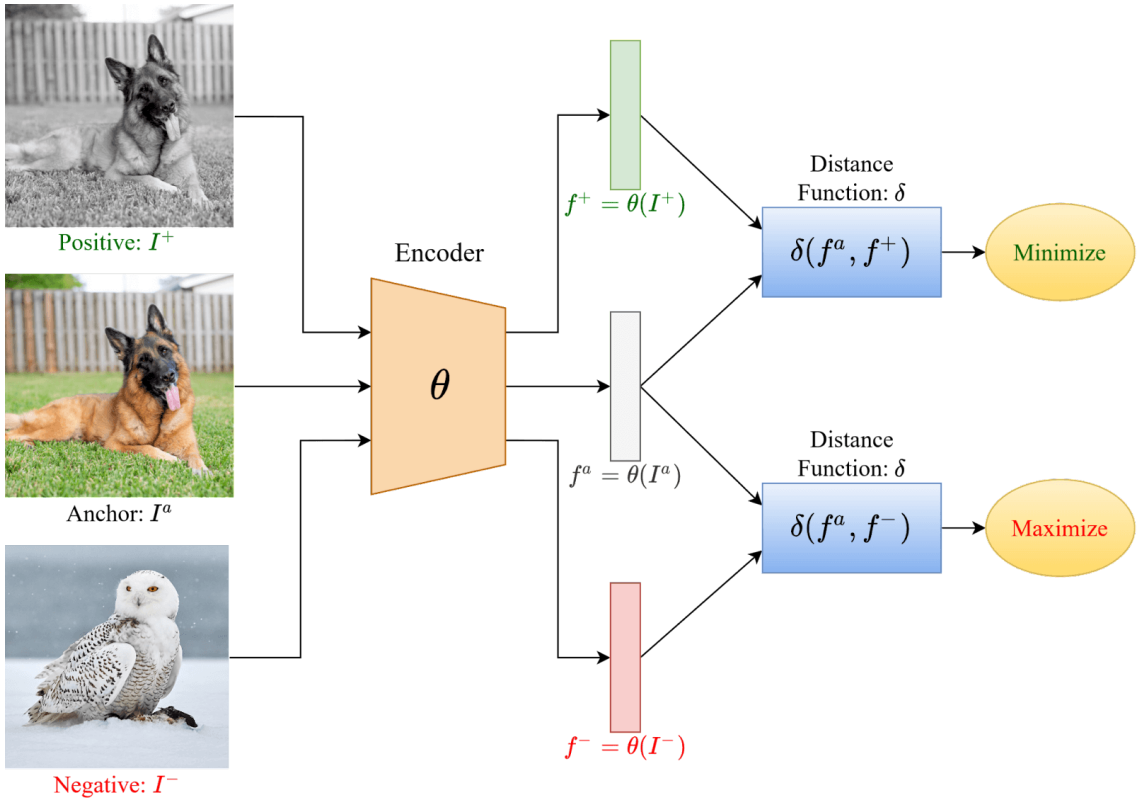
Lesson that took 5 years to learn: clever puzzles plateau.

The pretext should be as broad as "understand this image" — not a specific puzzle.

2. The Contrastive Revolution (2019–2021)

Two Crops of the Same Image Are More Alike

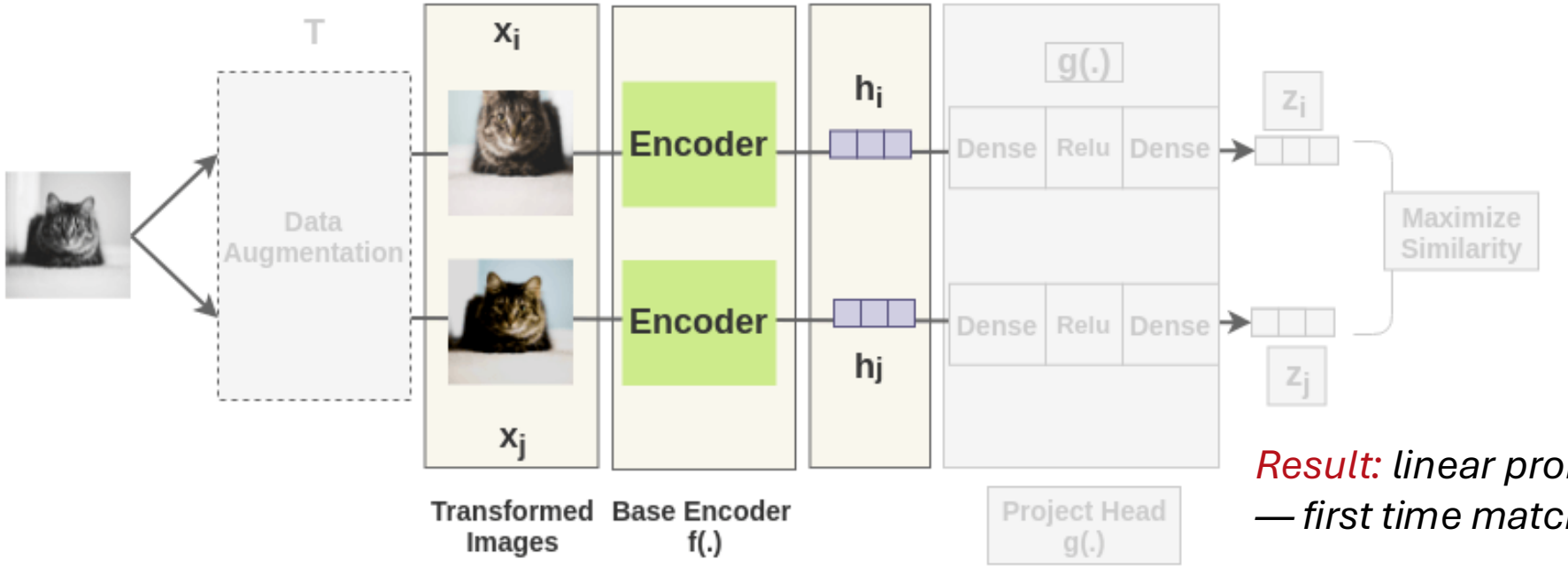
- **Contrastive insight in one sentence:** two random augmentations of the same image (with different color jitter, crops) should produce embeddings that are close — and far from any crop of a different image.
- Why this is brilliant:
  1. the "label" is free;
  2. the task is broad — solving it requires understanding content;
  3. invariance is built in by construction.



SimCLR (2020) — The Recipe That Worked

Chen et al. summarize contrastive learning to four components:

1. Strong augmentation pipeline — crop + flip + color jitter + Gaussian blur.
2. Base encoder f(·) — usually ResNet-50.
3. Projection head g(·) — small MLP, used only during pretraining, then discarded.
4. InfoNCE loss — the contrastive objective itself.



Result: *linear probe 60% → 76.5%*
— first time matching supervised.

Derive the InfoNCE Loss

Setup: batch of N images, augment each twice → 2N views.

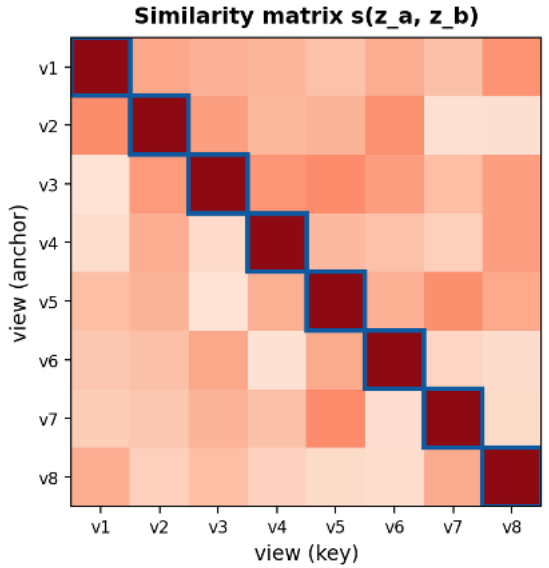
For anchor i, positive = i^+. Other 2N−2 = negatives.

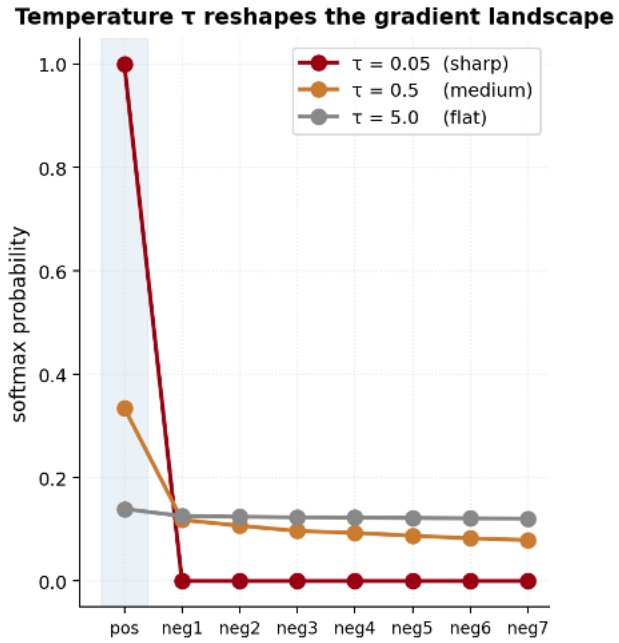
- Cosine similarity:

$$s(z_a, z_b) = \frac{z_a \cdot z_b}{\|z_a\| \, \|z_b\|}$$

- InfoNCE for anchor i:

$$L_i = -\log \left[\frac{\exp\left(s(z_i, z_i^+)/\tau\right)}{\sum_{k \neq i} \exp\left(s(z_i, z_k)/\tau\right)} \right]$$





This is just cross-entropy on a (2N−1)-way classification problem:

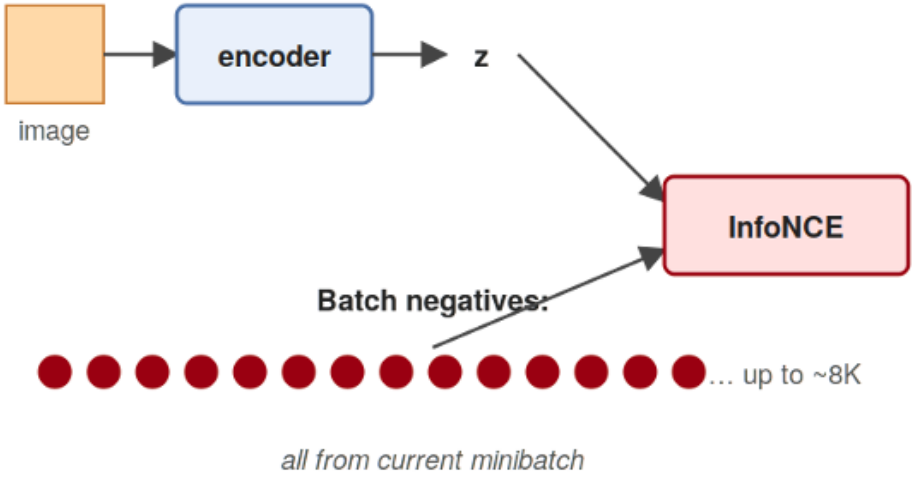
"out of all these views, which is my positive?"

MoCo (2020) — A Queue + a Slow Teacher

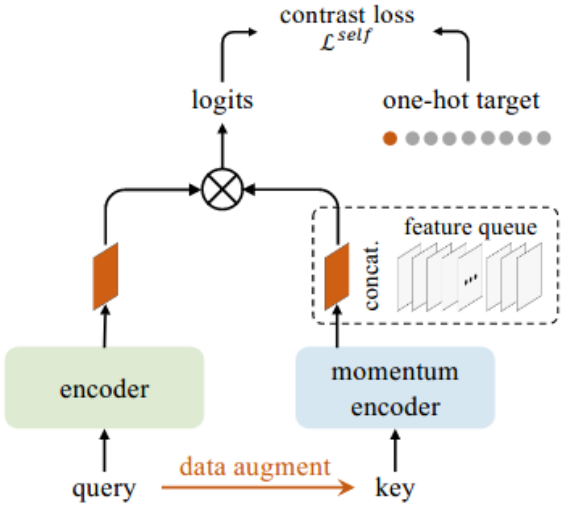
He et al. asked: why should the number of negatives be tied to batch size?

1. Memory queue: FIFO of 65,536 past embeddings. New batch enqueues; oldest dequeues.
2. Momentum encoder: key encoder = EMA of query encoder, m≈0.999 — keeps queued embeddings consistent.

SimCLR



MoCo



3. Beyond Negatives — The Predictive Methods

Do We Even Need Negatives?

- Negatives were assumed essential. Without them, the trivial solution is "output the same embedding for everything" (representation collapse).
- Then in 2020, two papers broke this assumption:

Contrastive (e.g. SimCLR, MoCo)

requires negatives in denominator

positive ↑

$$\frac{\exp\{s(z_1, z_2) / \tau\}}{\sum_k \exp(s(z_1, z_k) / \tau)}$$

negatives ↑

denominator stops collapse

(if all z's are equal, loss is large)

Predictive (BYOL, SimSiam)

no negatives at all

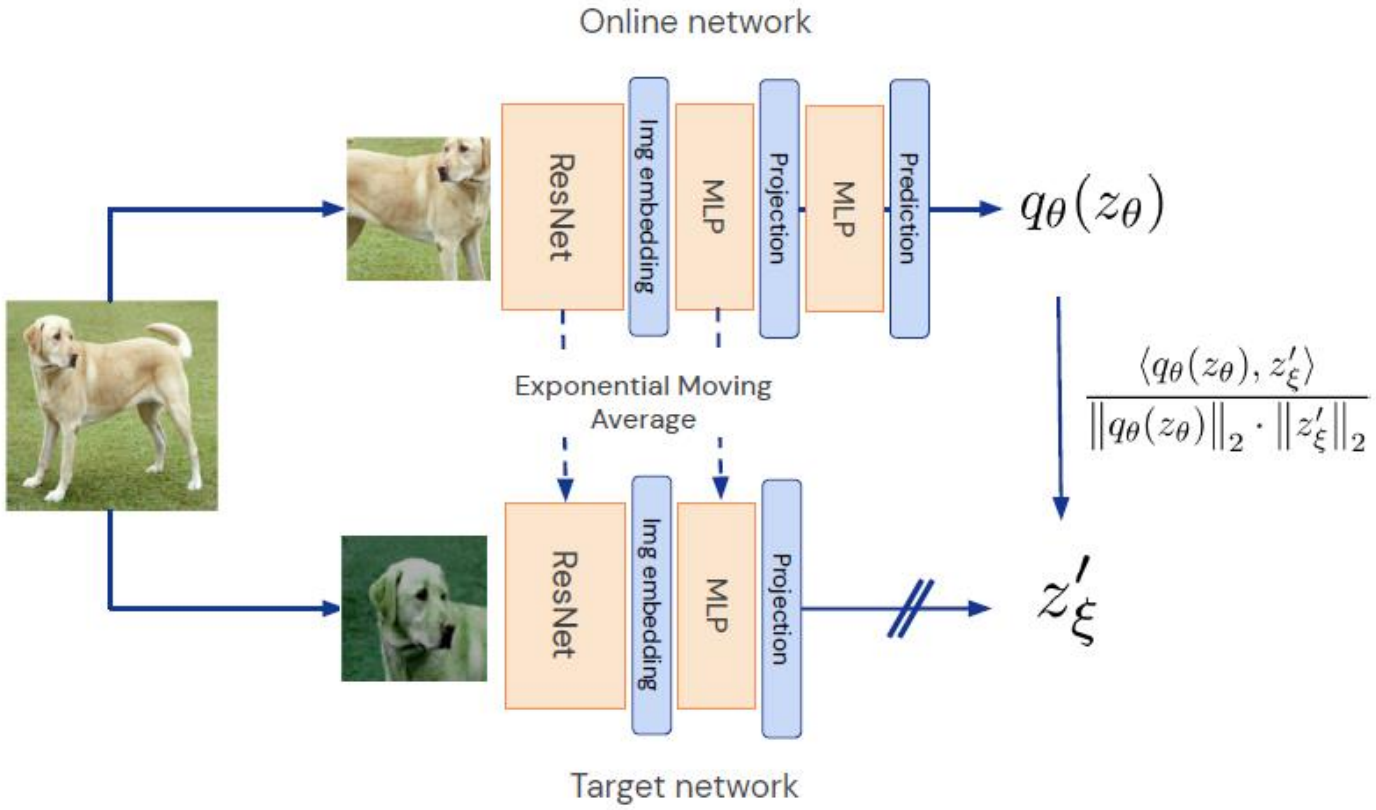
$$L = \|q(z_1) - sg(z_2)\|_2^2$$

no denominator, no negatives.

Why doesn't this collapse?

(next slide answers it — sort of)

BYOL — Stop-Gradient + Predictor = No Collapse



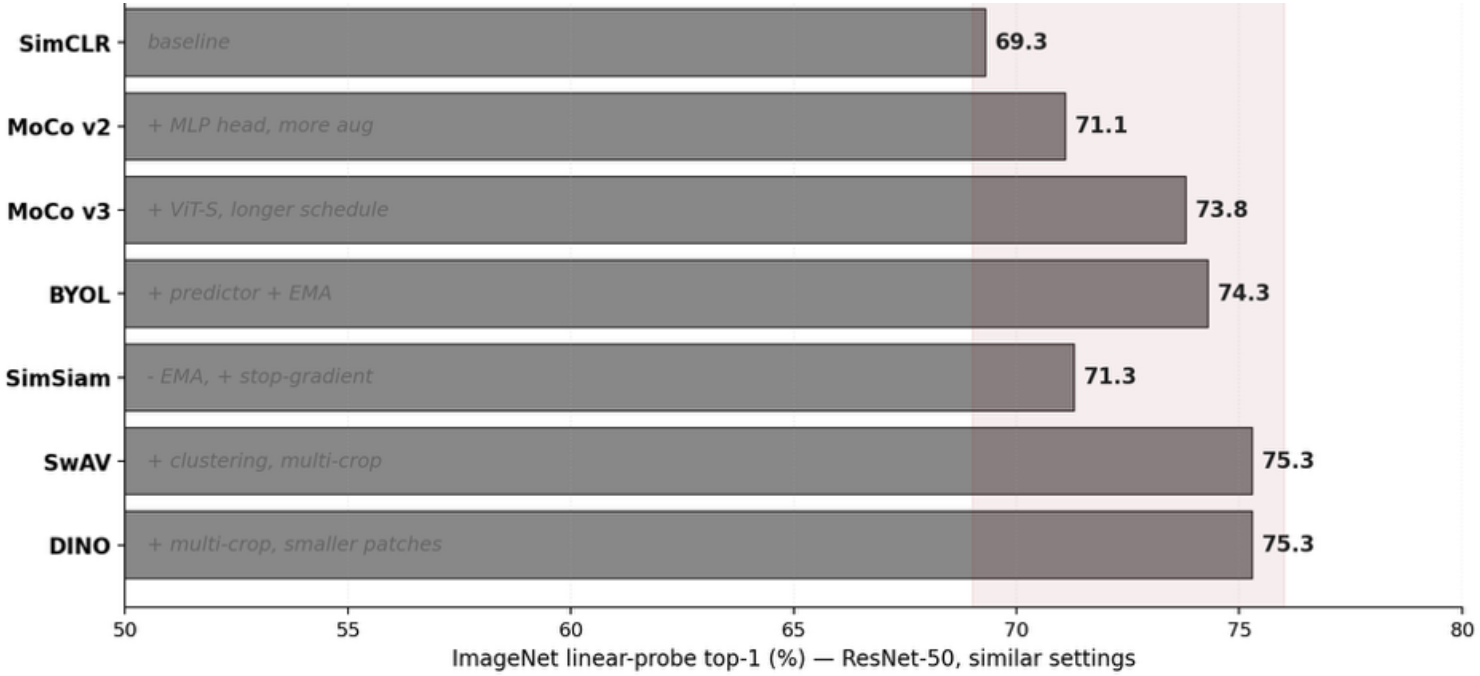
Why does it work? EMA target moves slowly; the predictor must compensate; this dance creates implicit dynamics with a non-collapsed fixed point.

A Critical Honest Look at the SSL Literature

Are we sure it's the loss, not the recipe?

When matched on architecture, augmentations, batch size, optimizer, and epochs:

- MoCo v3, BYOL, SimSiam, SwAV, DINO are all within ~1 point.
- The augmentation pipeline matters more than the loss.
- The projection head depth matters more than whether you use negatives.



4. The Generative Comeback — Masked Autoencoders

A Plot Twist from NLP

What if we just copy BERT?

NLP solved SSL years before vision did. BERT (2018):

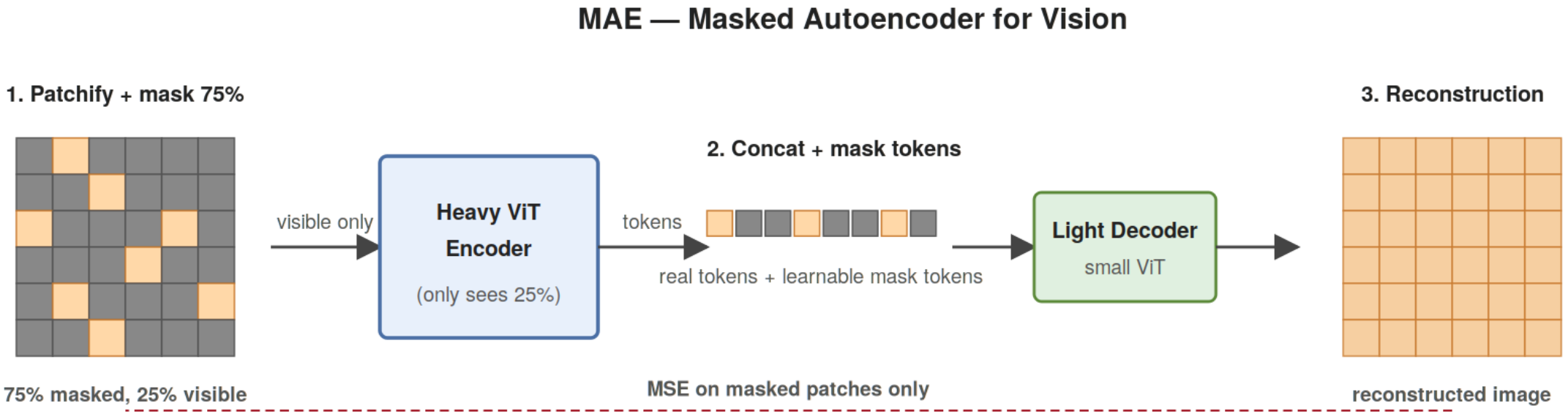
1. Take a sentence.
2. Mask 15% of the tokens.
3. Predict the masked tokens.
4. That's it. No contrastive loss, no negatives, no predictor, no temperature.

For five years, vision tried to copy this — and failed.

CNN-based "masked image modeling" never matched contrastive methods.

Then ViT arrived (L18) — and suddenly masking made sense again. Patches are tokens.

MAE (2021) — Masked Autoencoders Scale for Vision



Why this works (and 2015's "context inpainting" didn't)

75% masking: task is genuinely hard; can't solve by interpolation.

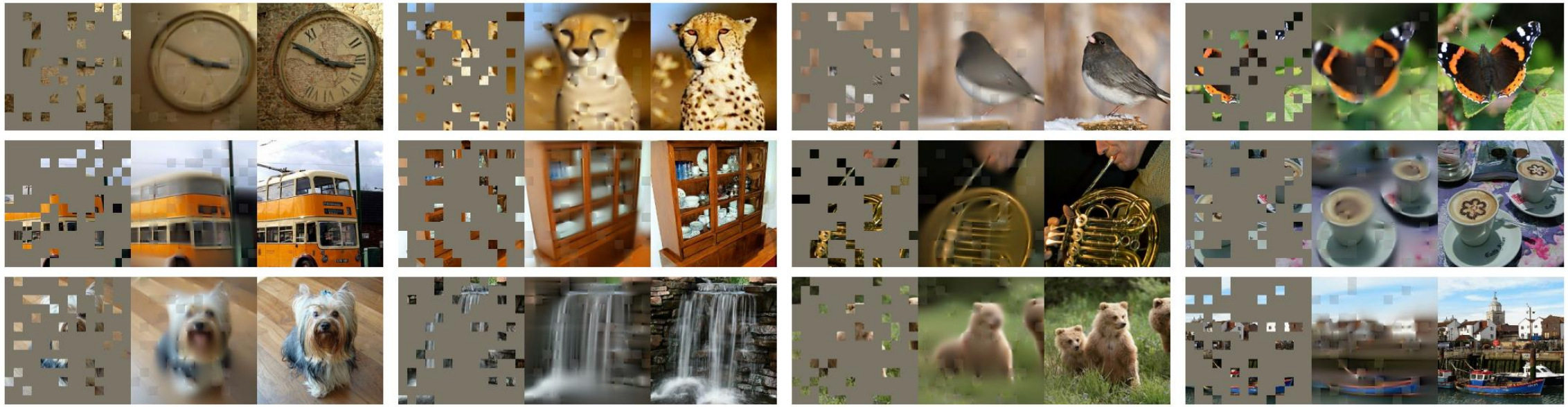
Encoder only sees 25%: 4× speedup; makes ViT-Huge trainable.

Asymmetric decoder: decoder is light; reconstruction is a means, not the goal.

→ MAE-pretrained ViT-Huge fine-tuned to 87.8% top-1 on ImageNet.

Why 75% Masking?

- BERT masks 15% of tokens. MAE masks 75% of patches. Why the 5× difference?
- => Image patches are more redundant than words.

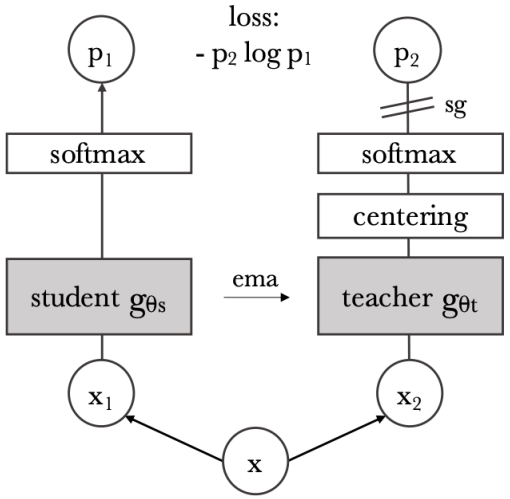


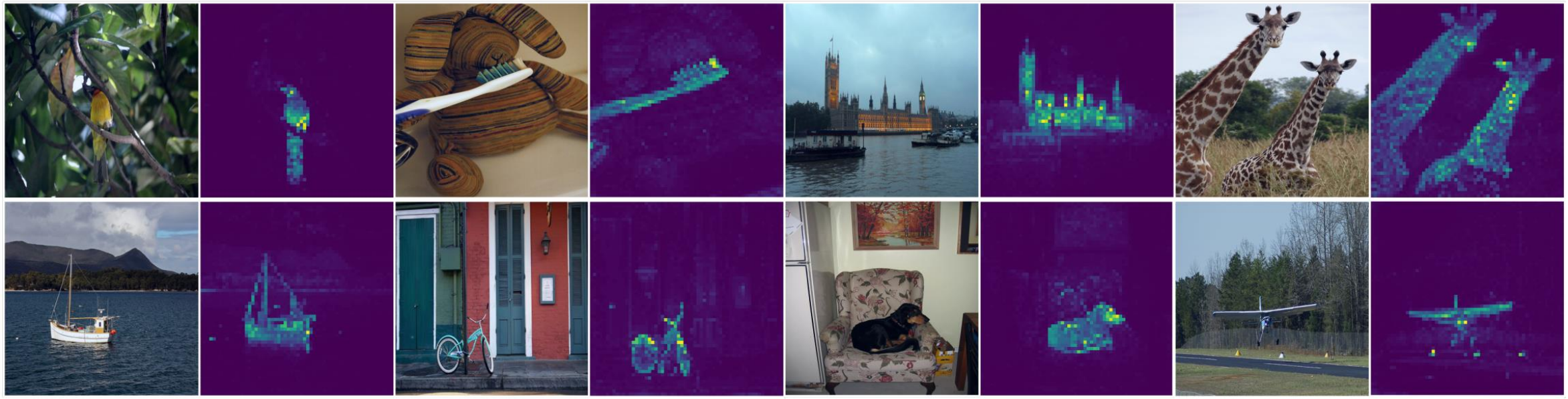
Contrastive vs. Generative — Two Philosophies

	Contrastive (SimCLR, MoCo)	Generative (MAE, BEiT)
Pretext task	Discriminate views	Reconstruct masked input
What's preserved	Invariances (crop, color)	Pixel-level structure
Best for	Classification, retrieval	Detection, segmentation, dense prediction
Hardware footprint	Big batches, many negatives	Many epochs, big model
Linear probe	Strong (DINO ≈ 80%)	Weaker (MAE ≈ 68%)
Fine-tune	Strong	Stronger (MAE ≈ 87.8%)

DINO (2021) — Self-Distillation, Emergent Object-ness

- DINO uses a student-teacher setup like BYOL, no negatives, on a ViT backbone.
- The result that shocked the community: when you visualize the attention maps of a DINO-trained ViT, they spontaneously segment objects.

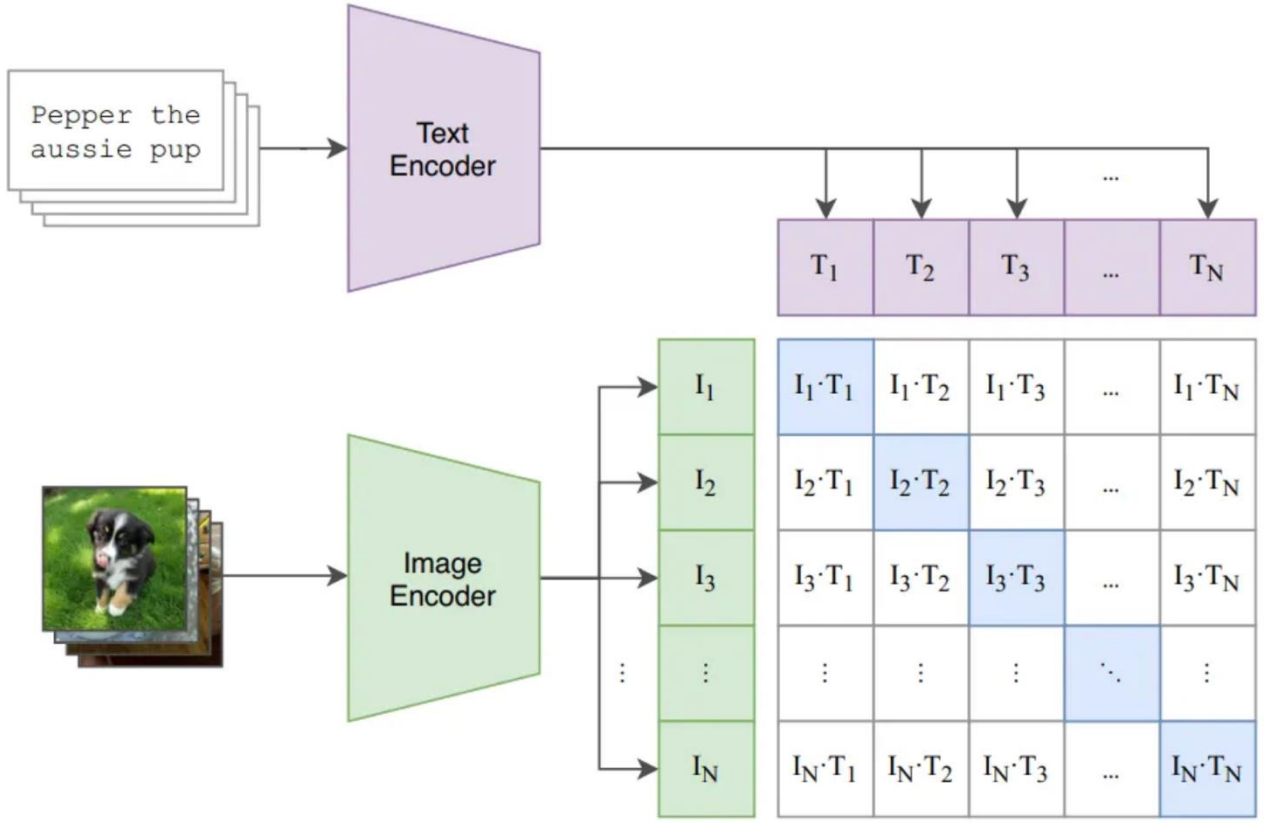




5. SSL Beyond Vision — Where the Field Went Next

CLIP (2021) — When the Internet Becomes the Label

- OpenAI's CLIP didn't invent contrastive learning. It applied it to 400M image–text pairs scraped from the web. The pretext task: match each image to its caption (vs. all other captions in the batch).



SSL in the Era of Foundation Models

Almost every foundation model you use was pretrained self-supervised:

Model family	SSL pretext task	Scale
GPT / Llama / Claude	Predict next token	trillions of tokens
BERT / T5	Masked language modeling	hundreds of billions
CLIP / SigLIP	Image–text contrastive	billions of pairs
DINOv2 / DINOv3	Self-distillation, ViT	100M+ images
MAE / VideoMAE	Masked reconstruction	millions of images / videos
Stable Diffusion / Imagen	Denoising (a generative SSL!)	billions of images
VLA models (RT-2, π0)	Imitation + SSL vision encoder	robot demos + web

Summary

1. The label can come from the data itself.
2. The pretext task should be broad and hard.
3. Two paradigms, both work.
   – *Contrastive (invariance) and generative (reconstruction)*
4. SSL is what makes scale work.
   – *Every modern foundation model — LLMs, CLIP, MAE, diffusion — is fundamentally self-supervised at pretraining.*