Self-Supervised Learning

Labels hidden in plain sight — representation learning

Prof. Nipun Batra

ES 667 · Deep Learning · IIT Gandhinagar

Representation learning is the thread

A network is also a representation map

D DERIVATION

Every classifier you have built already splits into two parts:

$$x \xrightarrow{f_\theta} h \xrightarrow{g_\varphi} \hat{y}$$

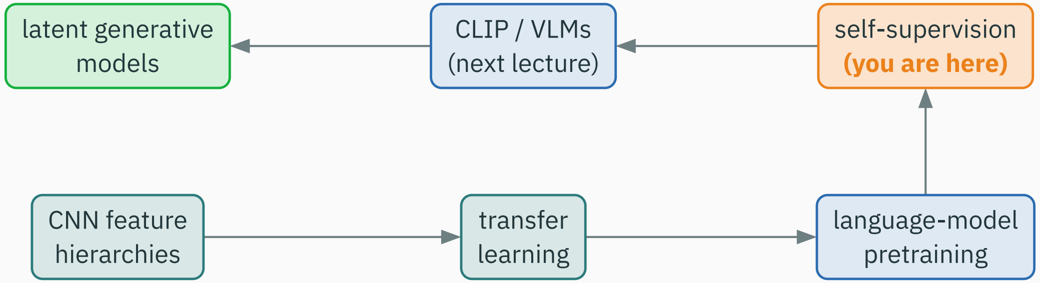
the **head** g_φ solves **today's** task — one label set, one loss.

the **representation** $h = f_\theta(x)$ is what we hope survives to **tomorrow's** task.

the whole game of representation learning: make h good **before** we know the downstream task

The course's representation map

One thread runs through the whole second half of the course:



everything here is about **learning h from structure already in the data** — today's stop is self-supervision

Supervised learning's bargain, and the label bottleneck

D DERIVATION

Supervised learning strikes a clean bargain:

$$(x_i, y_i) \quad \rightarrow \quad \text{learn} \quad f_\theta(x_i) \quad \text{useful for} \quad y_i$$

Labels make the objective crisp — but they **cost**. Contrast three data pools:

Pool	Scale	Cost
curated labeled images	10^6–10^7	expensive human annotation
web images / text / audio / sensors	10^9+	no manual labels; noisy and costly to curate
what a student / lab can collect	modest	cheap, but tiny

expensive labels **bottleneck** the very scale that makes deep learning work

The central pivot

D DERIVATION

The whole field turns on one move:

instead of receiving a human label, manufacture a target from the input itself

The data becomes its own teacher. Hide part of it, transform it, or pair it with a copy — and predict what you hid. Self-supervision turns **raw data into its own training signal**.

Three task families

D DERIVATION

Three recurring recipes — memorize the **pattern**, not a list of method names:

Family	**Constructed target**	**Main intuition**
contrastive / predictive	another **related view**	keep related things close
masked reconstruction	a **hidden part**	infer context and structure
teacher–student	a **teacher representation**	stable targets without labels

match related views | reconstruct hidden parts | match a slowly-changing teacher

Contrastive learning: make two views agree

Hook: one object, two views

Q QUESTION

Two hard crops of the **same** dog, and a crop of a **different** dog.

what should the representation treat as the **same**? (the two views of one dog)

what should it treat as **different**? (the other dog)

which differences should the representation learn to **ignore**?

Augmentations define the invariance

D DERIVATION

Two random transforms make two views — each **declares** what does **not** change identity:

$$x \quad \xrightarrow{t_1} \quad v_1, \qquad\qquad x \quad \xrightarrow{t_2} \quad v_2$$

- **random crop** $\to$ ignore exact location and framing;
- **colour jitter** $\to$ ignore global colour statistics;
- **blur** $\to$ ignore fine texture; **small geometric** $\to$ ignore small pose changes.

the set of augmentations **is** the set of invariances you ask for

A caution: an augmentation is a modeling assumption

D DERIVATION

The invariances are a choice — and a **wrong** choice ruins the representation:

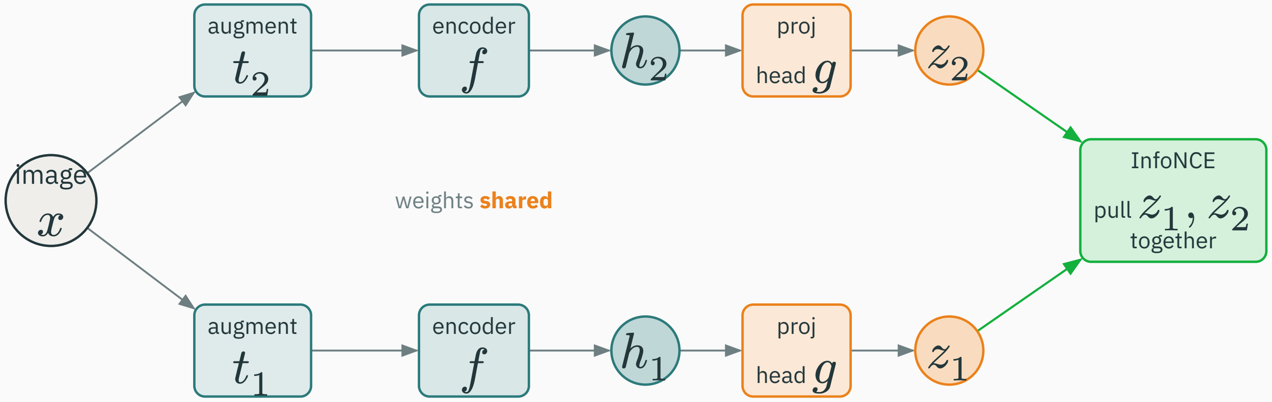
rotation is harmless for natural photos, but **disastrous** for digit recognition (6 vs 9). **time-reversal** is invalid for causal signals. **colour** jitter is fine for species, unacceptable for medical staining.

wrong invariance $\Rightarrow$ wrong representation — SSL moves domain expertise into **objective design**

The SimCLR pipeline

V VISUAL

SimCLR (Chen et al. 2020): two views, a **shared** encoder, a projection head, a contrastive loss.



distinguish the encoder output $h = f(v)$ from the contrastive vector $z = g(h)$

Why a projection head?

D DERIVATION

The contrastive loss **discards** nuisance information to make views agree. Let g absorb that pressure:

$z = g(h)$ — trained hard for the contrastive task; may throw away detail. **Discarded** after pretraining.

$h = f(v)$ — kept **richer**, less over-specialized to the pretext task. This is what transfers.

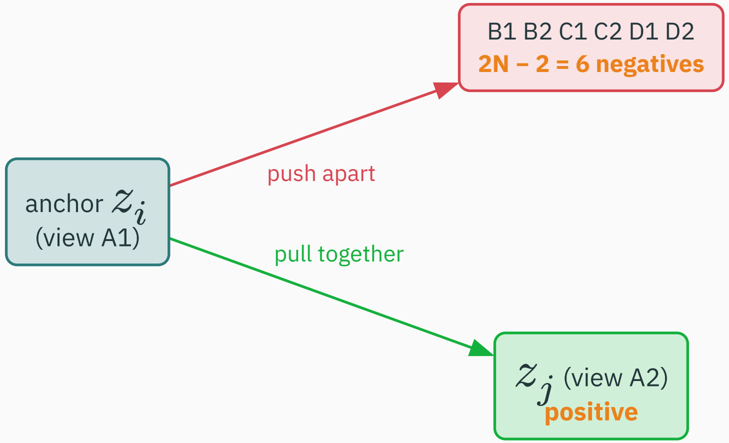
measure and **use** h; optimize **through** z

A contrastive batch

D DERIVATION

From N source images make $2N$ views (two per image):

- pick any view as the **anchor** z_i;
- exactly **one** other view is its **positive** (the sibling of the same source);
- the remaining $2N - 2$ views act as **negatives**.



Similarity: compare by angle, not length

D DERIVATION

Normalize the vectors, then take a cosine similarity:

$$s(z_i, z_j) = \frac{z_i^\top z_j}{\|z_i\| \, \|z_j\|}$$

dividing by the norms means similarity is the **angle** between embeddings, not their magnitude

InfoNCE: identify the positive among candidates

D DERIVATION

For anchor i with positive j, the loss (van den Oord et al. — InfoNCE):

$$\ell_{i,j} = -\log \frac{\exp\big(s\big(z_i, z_j\big)/\tau\big)}{\sum_{k \neq i} \exp(s(z_i, z_k)/\tau)}$$

this is just a $(2N-1)$-way softmax classification: "which of the other views is my positive?"

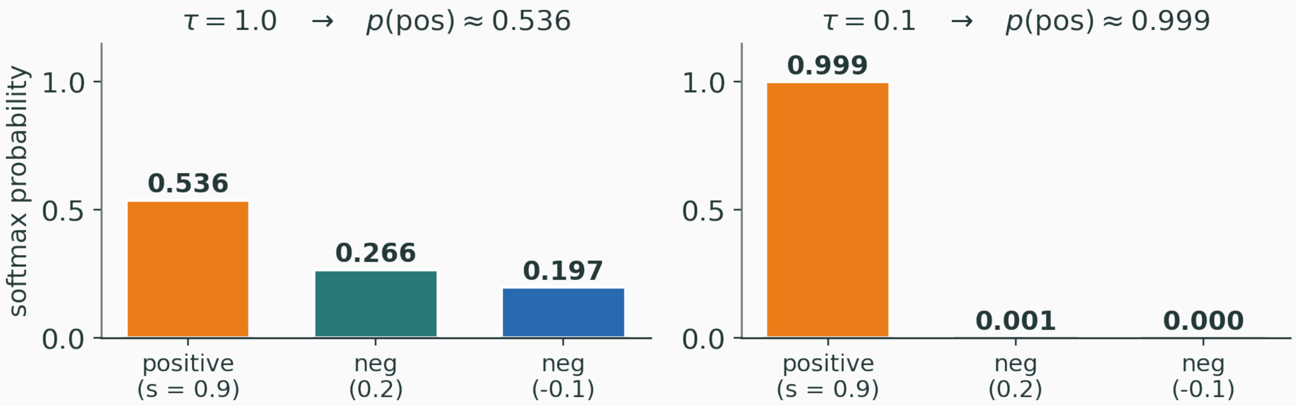
not a mysterious new loss — cross-entropy where the "classes" are the other views in the batch

Board: temperature is a contrast amplifier

D DERIVATION

One anchor, three similarities: 0.9 (positive), 0.2, -0.1. Softmax the scaled scores s/τ:

same similarities, two temperatures — smaller τ sharpens the competition



smaller τ **sharpens the competition** — the model is forced to beat the **hardest** negatives

What prevents trivial constant features?

D DERIVATION

Suppose the encoder collapses — every input maps to the **same** embedding.

- then $s(z_i, z_{\text{positive}}) = s(z_i, z_{\text{negative}})$ for all pairs;
- the positive can **never** win the softmax — the loss stays high.

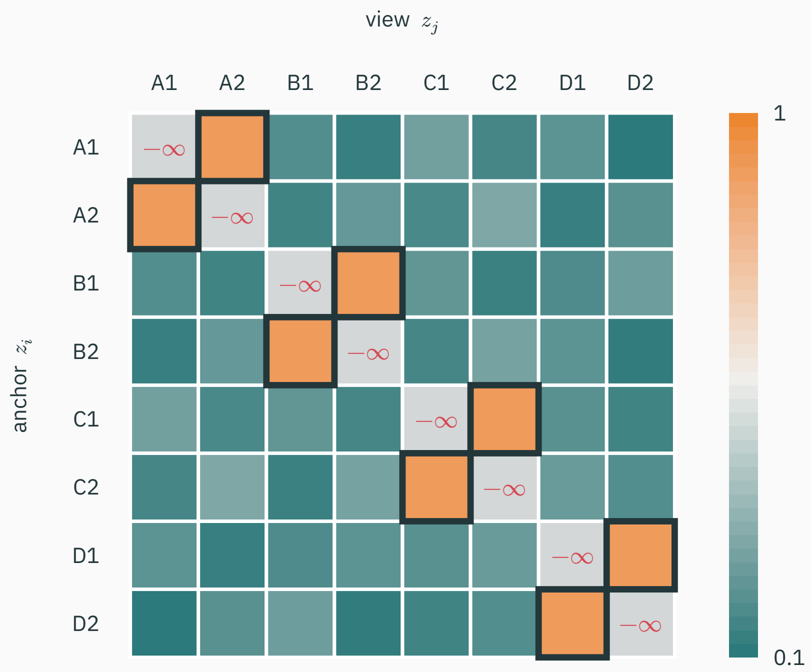
the denominator (the negatives) makes collapse a bad solution — competition is built in

contrast between positive and negatives is exactly what rules out the constant shortcut

The similarity-matrix view

V VISUAL

Stack all $2N$ views: entry (i, j) is $s(z_i, z_j)/\tau$. Each row is one softmax problem.



one boxed positive per row, the diagonal ignored — the **same** “correct match in a batch” geometry that powers CLIP

Evaluation: feature quality is not head capacity

D DERIVATION

A powerful head can mask a weak representation. Separate the two:

Protocol	**Encoder**	**What it tests**
linear probe	frozen	representation quality (linearly readable?)
fine-tuning	updated	downstream ceiling
retrieval / nearest-neighbour	frozen	semantic geometry, no new head
dense-task transfer	often updated	locality and spatial detail

a strong **frozen linear probe** is the real signal that pretraining worked

Interactive: contrastive learning

I INTERACTIVE

INTERACTIVE

↗ nipunbatra.github.io/interactive-articles/vision-ssl

Build two augmented views, watch them pass through the shared encoder and projection head, and see the InfoNCE softmax pick the positive out of the batch as you drag the temperature τ.

Masked reconstruction: infer what is missing

A different question

DERIVATION

Contrastive and masked modeling ask **different** questions of the data:

contrastive — “which view **matches** this one?” Pressure toward **invariance**.

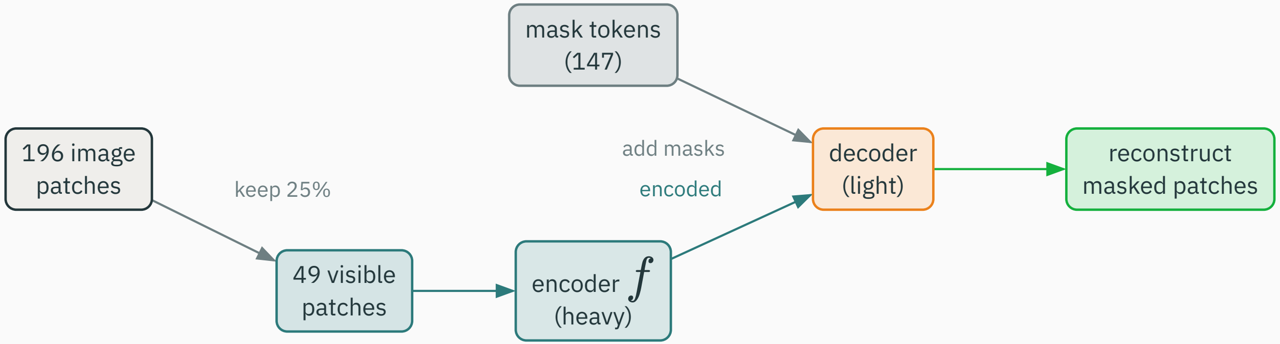
masked — “what must be **true** about the world for the hidden part to make sense?” Pressure toward **context**.

no augmentations, no negatives — the supervision is the **hidden content itself**

The masked-autoencoder setup

V VISUAL

MAE (He et al. 2022): hide most of the image, reconstruct it from the little that remains.



patches ⟶ hide a large subset ⟶ encoder sees visible only ⟶ light decoder fills the rest

MAE is asymmetric

D DERIVATION

The two halves do very different amounts of work:

- the **expensive encoder** processes **only the visible** patches — cheap because most are gone;
- the **lightweight decoder** gets encoded-visible tokens **plus** learned **mask tokens**;
- the decoder is **discarded** after pretraining — only f transfers.

For a 224×224 image with 16×16 patches: $N = (224/16)^2 = 196$. At 75% masking the heavy encoder sees only $196 \times 0.25 = \mathbf{49}$ tokens — a learning **and** an efficiency choice.

A simple reconstruction objective

D DERIVATION

Mean-squared pixel error, computed **only on the masked patches M**:

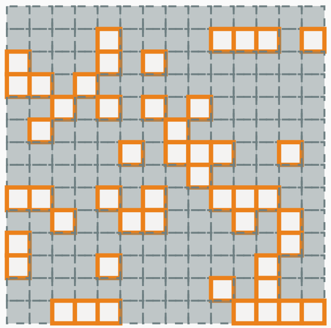
$$L_{\text{mask}} = \frac{1}{|M|} \sum_{p \in M} \left\| x_p - \hat{x}_p \right\|^2$$

no label anywhere — the target x_p is just the pixels we chose to hide

Why mask heavily?

V VISUAL

The mask ratio decides whether the task is **trivial** or **forces semantics**:



- a **few** hidden patches → copy local texture from neighbours — **trivial**;
- **≈75%** hidden → local copying fails → the model must use **global context**.

No universal number: BERT masks ≈15% (text is dense), MAE ≈75% (images are redundant) — mask ratio is a **modality** choice.

Contrastive versus masked

V VISUAL

	Contrastive	**Masked reconstruction**
supervision	a related view	hidden content
main pressure	invariance	contextual prediction
common strength	global classification / retrieval	spatial / dense transfer
key design choice	augmentations, negatives, τ	mask pattern, reconstruction target

two different assumptions about **what makes two inputs related** — often complementary

Bridge to language models

D DERIVATION

The masked idea is exactly what pretrains modern language models:

Model	**Predict...**
BERT	masked **tokens** from surrounding context
GPT	the **next token** from preceding context
MAE	masked image **patches** from visible patches

language-model pretraining is **not** separate from SSL — it is its most familiar instance

One unifying view

D DERIVATION

Strip away the modality and every recipe here is the same objective:

$$\boxed{\textbf{learn} \quad p(\text{hidden or future part} \mid \text{observed context})}$$

the **output spaces** differ — tokens, pixels, audio frames — but the self-supervised **logic** is shared

Teacher–student consistency: learn without negatives

The apparent problem: collapse

D DERIVATION

Drop the negatives. Train two networks **only to agree** on two views — what breaks?

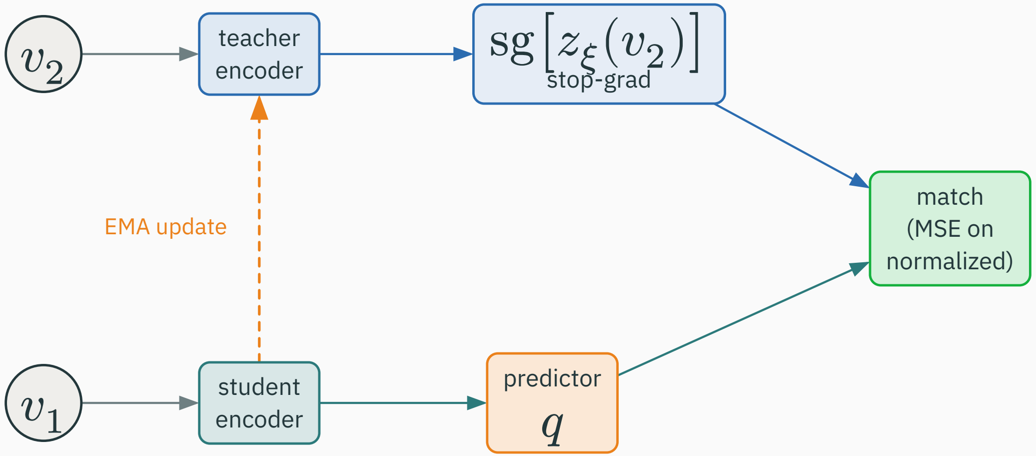
Both can output the **same constant** vector $[0, 0, ..., 0]$ for every input. The matching loss is **perfect**, yet the representation says **nothing**. This is **collapse**.

agreement alone is **not** enough — we need an asymmetry that makes the constant unreachable

BYOL at a glance

V VISUAL

BYOL (Grill et al. 2020): a student **predicts** a slowly-moving **teacher** — a design that can avoid collapse in practice.



student path has an extra **predictor**; the teacher target is **stop-gradient**; teacher weights are an **EMA** of the student

The three asymmetries to teach

D DERIVATION

BYOL avoids the trivial solution by breaking the symmetry three ways:

1. a **predictor** q exists **only** on the student path;
2. the teacher output is **stop-gradient** — no gradient flows into the teacher;
3. the teacher weights are an **exponential moving average** of the student.

the two branches are **not** interchangeable — so copying a constant is not an easy fixed point

The EMA update

DERIVATION

The teacher is never trained directly — it **trails** the student:

$$\theta_{\text{teacher}} \quad \leftarrow \quad m\,\theta_{\text{teacher}} + (1-m)\,\theta_{\text{student}}$$

Worked step, $m = 0.99$: teacher weight 0.40, student now $0.70 \rightarrow \quad 0.99(0.40) + 0.01(0.70) = \mathbf{0.403}$. The teacher barely moves — a **stable target** to chase.

Why this does not collapse — a careful answer

D DERIVATION

Do not say "the EMA alone prevents collapse." That is the common exam mistake.

In the working recipe, several ingredients act **together**:

- **stop-gradient** + **predictor asymmetry** break the student↔teacher symmetry;
- **normalization** keeps outputs from shrinking to zero;
- the **slowly-moving target** gives a stable, non-trivial thing to predict.

remove any one and the argument weakens — it is the **combination** that works

DINO and multi-crop

D DERIVATION

DINO (Caron et al.) / DINOv2 (Oquab et al.) scale the teacher–student idea:

- student and teacher predict **distributions over features** from **differently cropped** views;
- **multi-crop**: two **global** views + several small **local** views — the student must match a global target from local content;
- result: **emergent** semantic regions and strong **frozen** features, with **no labels**.

we keep this high-level — centering / sharpening details are beyond today

One-slide landscape, not a zoo

VISUAL

Do not memorize methods — file each under a **recipe**:

If you meet...	**It is a variation on...**
MoCo, SwAV, SimSiam, Barlow Twins	contrast / anti-collapse design
BYOL, DINO	teacher–student consistency
MAE, masked language modeling	reconstruct / predict hidden content

three recipes explain a decade of “new” methods

The same question across modalities

Cross-modal comparison

V VISUAL

Data	Contrastive / predictive	Masked	Consistency
images	augmented views; image–caption	masked patches	teacher / student crops
text	adjacent spans; sentence pairs	masked / next token	teacher / student encodings
audio / time series	nearby or future windows	masked spans / channels	noisy-view agreement

same three recipes, different output spaces — the objective **family** transfers, not the architecture

The transfer question — decide before you train

DERIVATION

State the **invariance** and the **downstream task first**, then choose the objective:

"ignore colour" — ideal for **species** recognition (a robin is a robin in any light).

"ignore colour" — **unacceptable** for **medical** imaging, where stain colour is diagnostic.

the pretext task must make the **knowledge you want necessary** to solve it

Common misconceptions

V VISUAL

Misconception	**Correction**
"SSL has no labels."	it has **targets constructed** from the data
"Contrastive means semantic negatives."	negatives are often **just batch alternatives**
"BYOL works because of EMA."	**several asymmetries** work together
"Good reconstruction $\Rightarrow$ good classifier."	the **target** decides what features are kept

Exit ticket: design an SSL recipe

Q QUESTION

Wearable-sensor stream, no labels, later used for activity recognition. In pairs, propose:

1. **two valid** augmentations (e.g. small time-warp, additive sensor noise);
2. **one invalid** augmentation (e.g. time-reversal — destroys causal order);
3. **one masked-prediction** target (e.g. reconstruct a masked time span);
4. **one evaluation** protocol (e.g. linear probe on a small labeled set).

One mental model: choose a transformation or hidden part that makes the **desired knowledge necessary**. Everything else in SSL is an implementation of that decision.

Closing bridge to CLIP

D DERIVATION

We built a geometry: **two augmented views of one image** pulled together against a batch of negatives.

$$\text{two augmented views of an image} \quad \Longrightarrow \quad \text{an image and its caption}$$

swap the second view for a caption — the same contrastive geometry becomes a shared image–text space

INTERACTIVE

↗ nipunbatra.github.io/interactive-articles/clip-zero-shot

Next lecture: watch the batch similarity matrix and contrastive softmax reappear — now between images and text — to drive zero-shot classification.

Self-supervision turns **raw data into its own training signal**.

Three recipes: match related views · reconstruct hidden parts · match a slow teacher.

Next: **Vision-Language Models** — image views → image–caption pairs → a shared multimodal space.