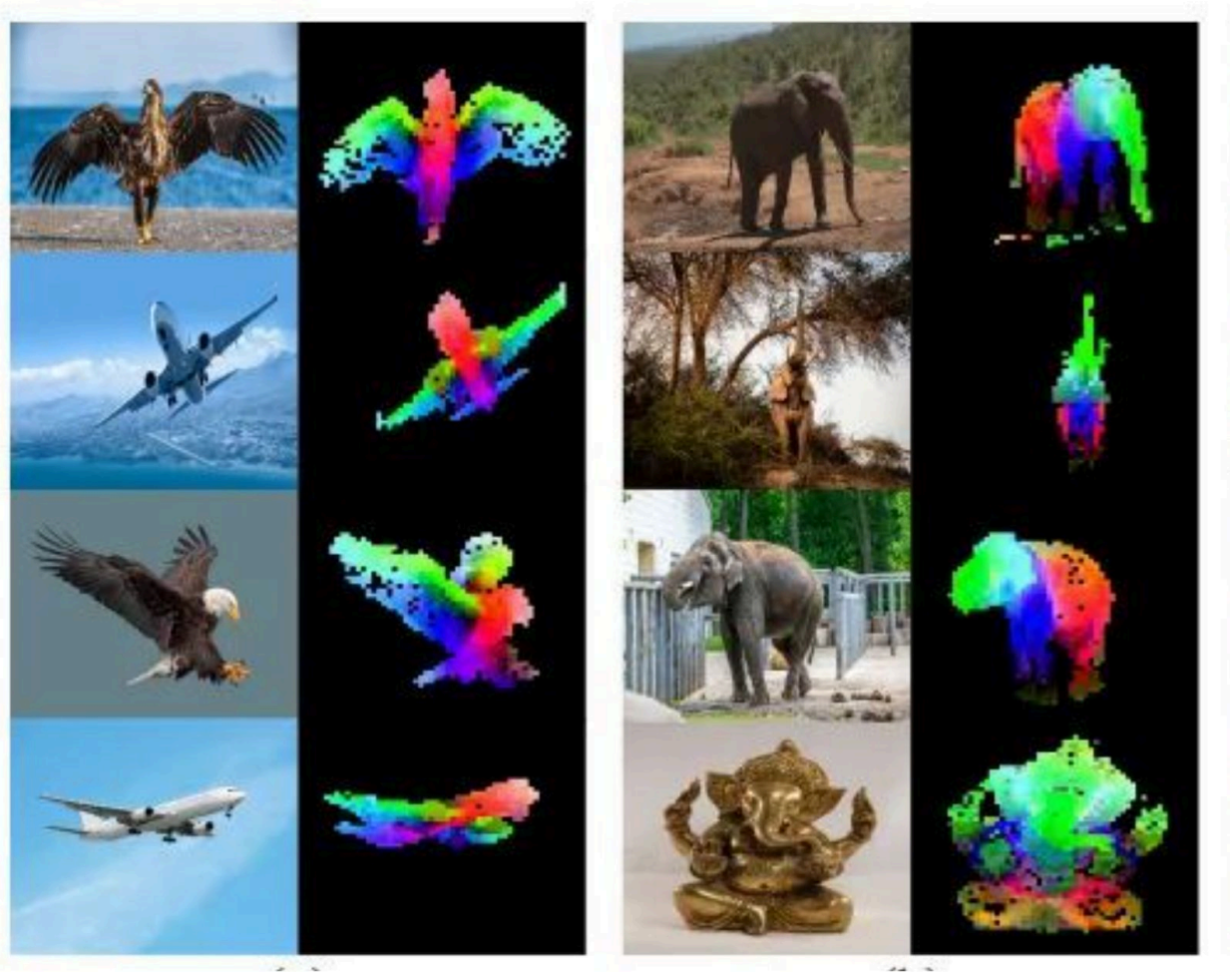
Ver sin etiquetas

DINO aprende representaciones visuales generales directamente de las imágenes.



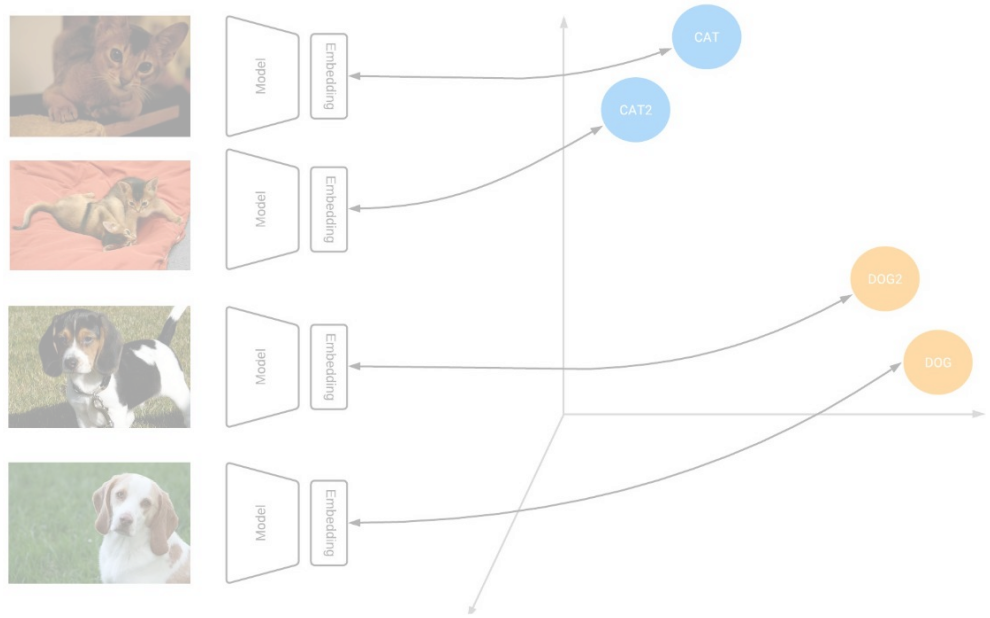
La representación puede reutilizarse para buscar, agrupar, localizar o comparar.

DINO significa self-distillation with no labels. Su objetivo principal es aprender un backbone visual transferible, no resolver por sí solo una única tarea final.

La señal está en los datos

La auto-supervisión crea objetivos de entrenamiento sin categorías anotadas por personas.

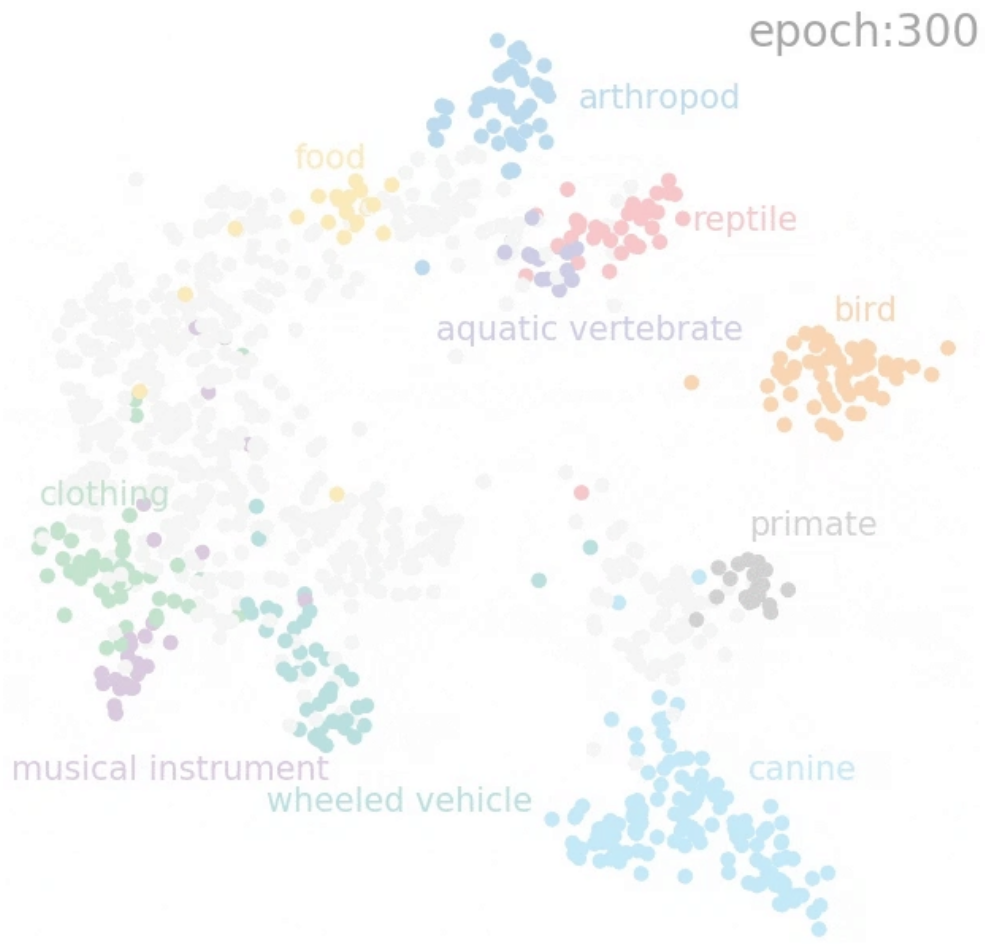
Supervisado	Auto-supervisado
Predice etiquetas humanas y queda ligado a su taxonomía.	Explota estructura, vistas y relaciones presentes en las imágenes.



La ventaja principal es escalar el preentrenamiento y reutilizar las características en dominios donde anotar resulta caro. La tarea final aún puede necesitar etiquetas para entrenar o evaluar una cabeza específica.

El modelo fabrica su ejercicio

Transformaciones, partes ocultas o agrupaciones proporcionan una señal de aprendizaje.



En aprendizaje contrastivo se relacionan vistas compatibles; en masked prediction se reconstruyen regiones ocultas; otras tareas históricas incluyen predecir rotaciones o agrupar representaciones parecidas.

La estudiante sigue una referencia

La red maestra es una versión suavizada de la estudiante y produce objetivos estables.

¿Qué función cumple la red profesora en DINO?

- A Ofrecer objetivos para que aprenda la estudiante.
- B Etiquetar manualmente cada imagen.
- C Generar texto sobre la imagen.

Respuesta A. Exacto: la profesora proporciona una referencia durante el entrenamiento.

La estudiante se actualiza por gradiente. La maestra no recibe gradiente directo: sus pesos siguen un promedio móvil de los pesos de la estudiante. Ambas procesan vistas distintas de la misma imagen y se fuerza la consistencia de sus salidas.

Emergen regiones coherentes

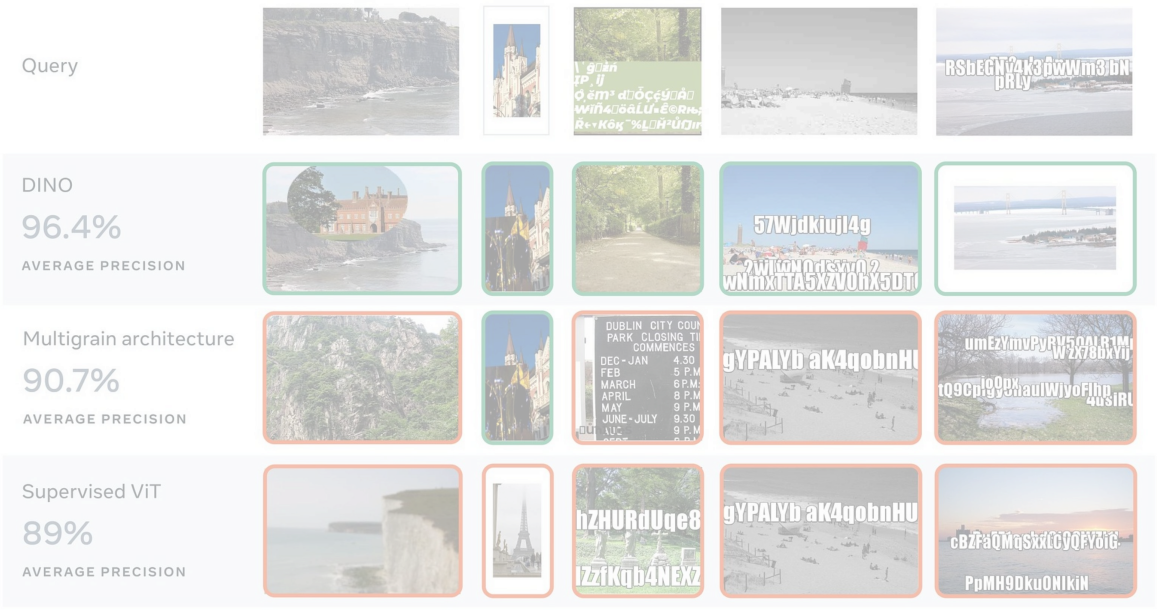
Los mapas de atención pueden alinearse con objetos aunque no haya máscaras de entrenamiento.

La atención muestra dónde se concentra el modelo; no es una máscara validada por sí sola.

En DINO v1 se observó que algunos heads de auto-atención de un Vision Transformer separaban objetos salientes del fondo. Es una capacidad emergente y una herramienta de interpretación, no una garantía de segmentación perfecta.

Reconocer la misma imagen

Embeddings robustos permiten recuperar copias incluso después de varias transformaciones.



La similitud semántica puede resistir recortes, cambios de color o texto superpuesto.

La identificación de copias es una aplicación de recuperación: se indexan embeddings y se buscan vecinos cercanos. El umbral debe evaluarse con transformaciones y negativos representativos.

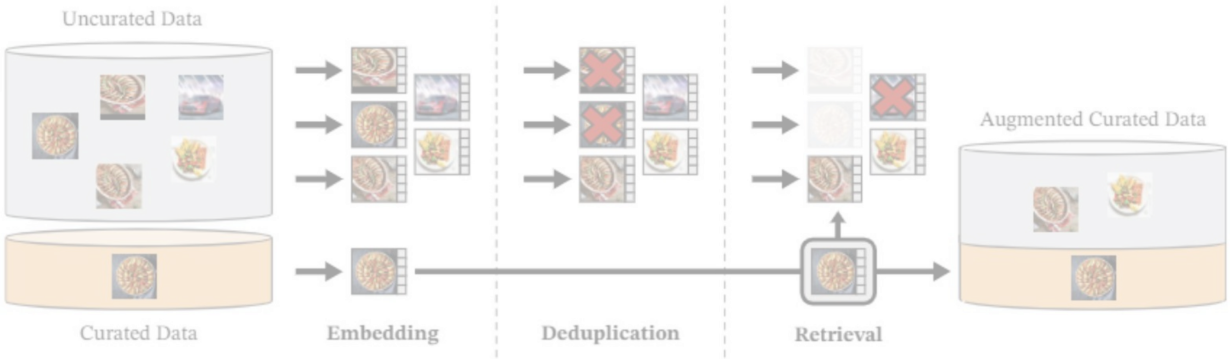
De DINO a DINOv3

La familia escala datos y modelos mientras mejora la calidad de las características densas.

DINO apareció en 2021; DINOv2 se publicó en 2023 con un pipeline de datos curados; DINOv3 se presentó en 2025 y escaló el entrenamiento auto-supervisado, incorporando técnicas para conservar la calidad de características densas durante entrenamientos largos.

La curación también aprende

DINOv2 usa embeddings para deduplicar y recuperar imágenes parecidas a un conjunto curado.



El modelo ayuda a construir un conjunto diverso sin etiquetar cada imagen.

El flujo combina datos curados y no curados, calcula embeddings, elimina duplicados y recupera ejemplos visualmente próximos al conjunto de referencia. La calidad del backbone depende también de esta selección de datos.

Dos vistas de una imagen

La estudiante observa detalles y aprende a concordar con el contexto de la maestra.

Aumentar

Crops locales y globales.

Estudiante

Procesa varias vistas.

Maestra

Procesa vistas globales.

Alinear

Se comparan salidas.

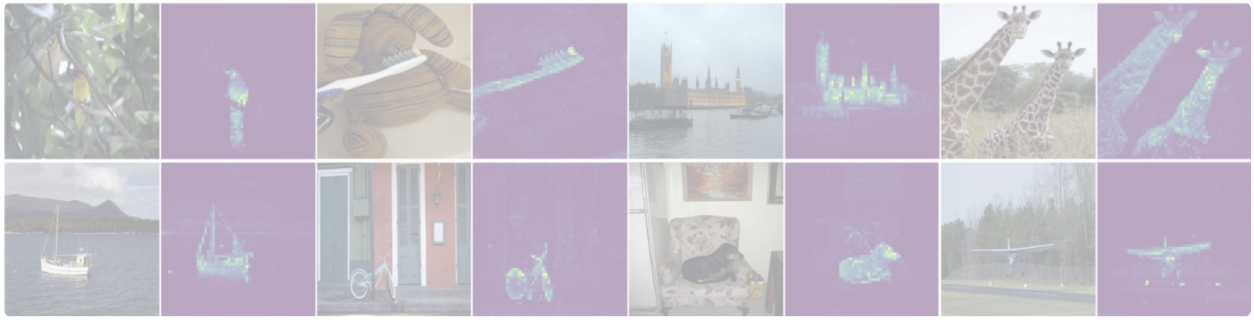
Actualizar

Promedio móvil.

Las vistas preservan la identidad de la escena y cambian recorte, color u otras propiedades. El objetivo fuerza una representación consistente ante esas transformaciones sin usar una etiqueta de clase.

Una base para muchas tareas

Las mismas características sirven para regiones, correspondencias, recuperación y clasificación.



La tarea final añade una regla, un índice o una cabeza sobre el backbone.

Los mapas densos apoyan segmentación y correspondencias; el token global resume la imagen para recuperación o clasificación con k-NN. Estas capacidades se evalúan por separado: una buena representación no sustituye la validación de la tarea final.

Global y por regiones

DINO produce un resumen de imagen y una cuadrícula de embeddings por patch.

El token global compara imágenes; los patches conservan información espacial.

El notebook usa DINOv2 como extractor congelado. El token CLS se emplea para similitud global y los tokens de patch para visualizar o comparar regiones.

Cercanía según el modelo

La similitud coseno compara la dirección de dos embeddings normalizados.

Extrae un vector por imagen.

Normaliza antes de comparar.

Ordena vecinos y revisa errores.

Dos imágenes tienen embeddings cercanos. ¿Qué puedes concluir?

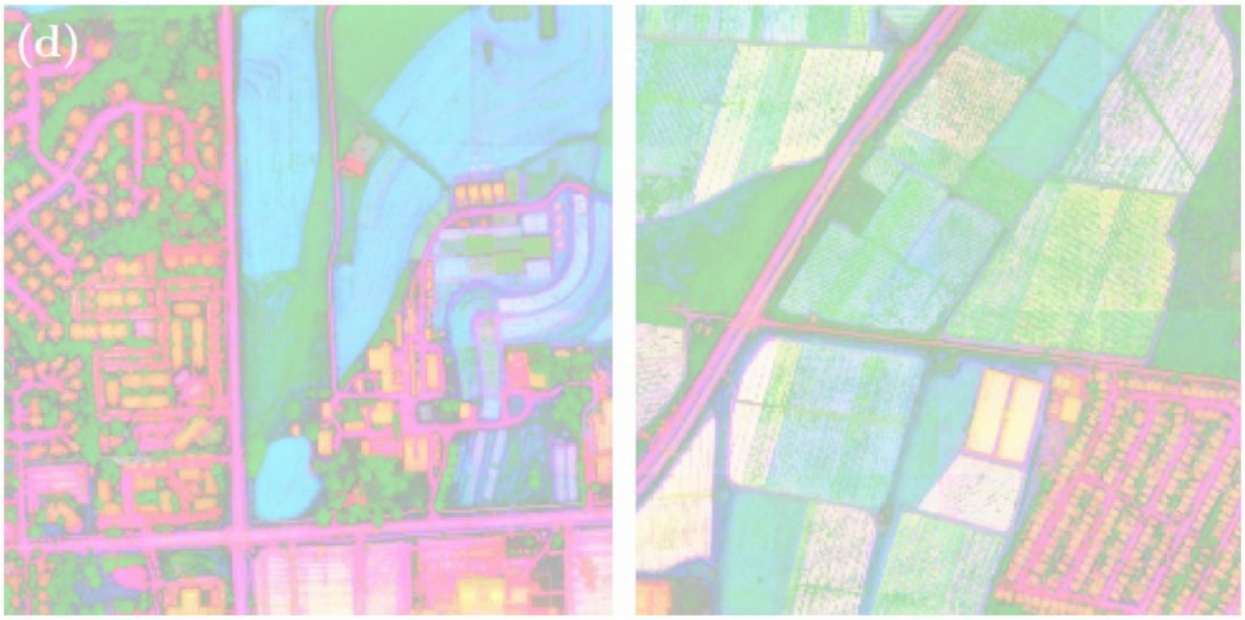
- A El modelo las representa como similares.
- B Son idénticas píxel a píxel.
- C Pertenecen a una clase legalmente autorizada.

Respuesta A. Exacto: la cercanía expresa similitud según esa representación.

Una similitud alta significa que el modelo representa las imágenes de forma próxima. No demuestra identidad píxel a píxel ni una categoría concreta.

De features a un mapa visible

PCA colorea patches; t-SNE y UMAP proyectan grupos de embeddings en dos dimensiones.

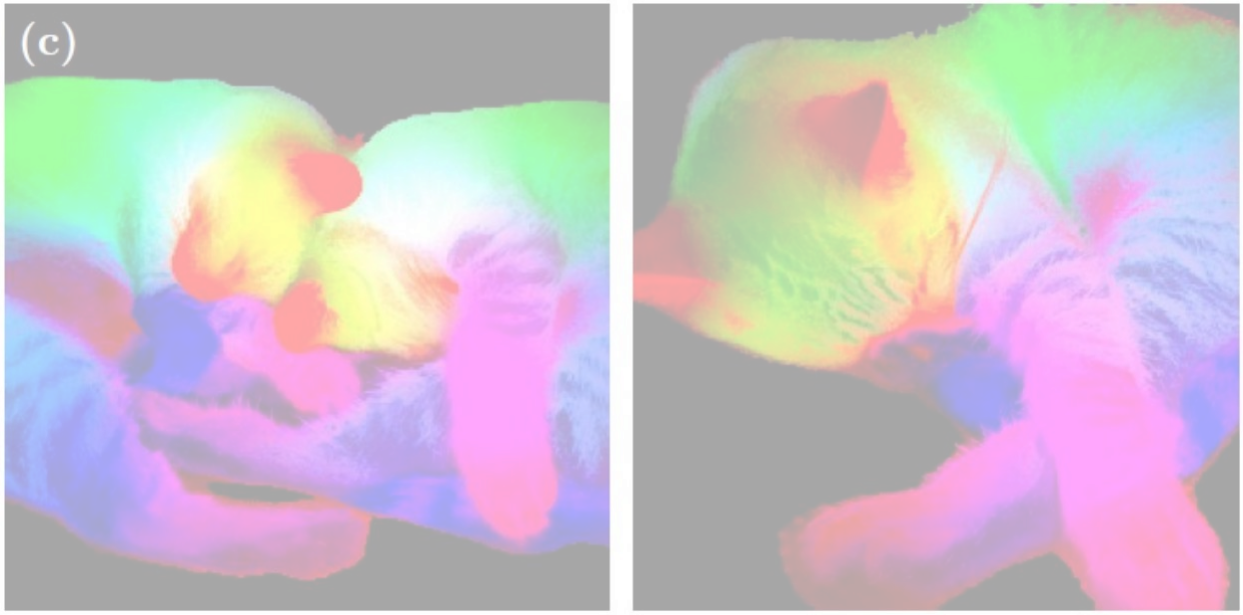


Los tres primeros componentes pueden mapearse a RGB para inspeccionar regiones.

PCA es una proyección lineal útil para colorear características densas y observar variación espacial. t-SNE y UMAP son técnicas no lineales que proyectan embeddings globales o de patches a 2D para explorar vecindarios y grupos. Ninguna de ellas crea etiquetas ni demuestra por sí sola que un cluster sea una clase real.

Buscar correspondencias visuales

Un patch consulta qué regiones tienen una representación parecida.

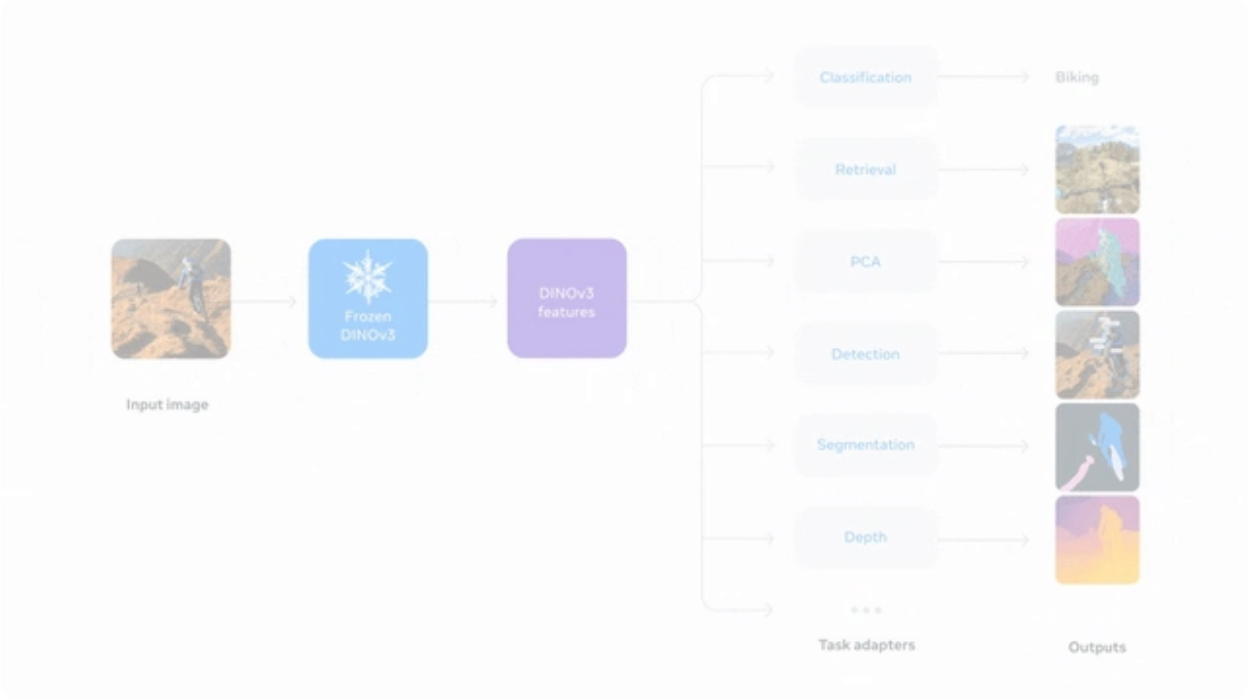


Los vecinos permiten comprobar si la representación conserva partes equivalentes.

La búsqueda de vecinos puede realizarse entre patches de la misma imagen, de imágenes distintas o de una base indexada. Sirve para correspondencias, recuperación y análisis de consistencia.

Características densas estables

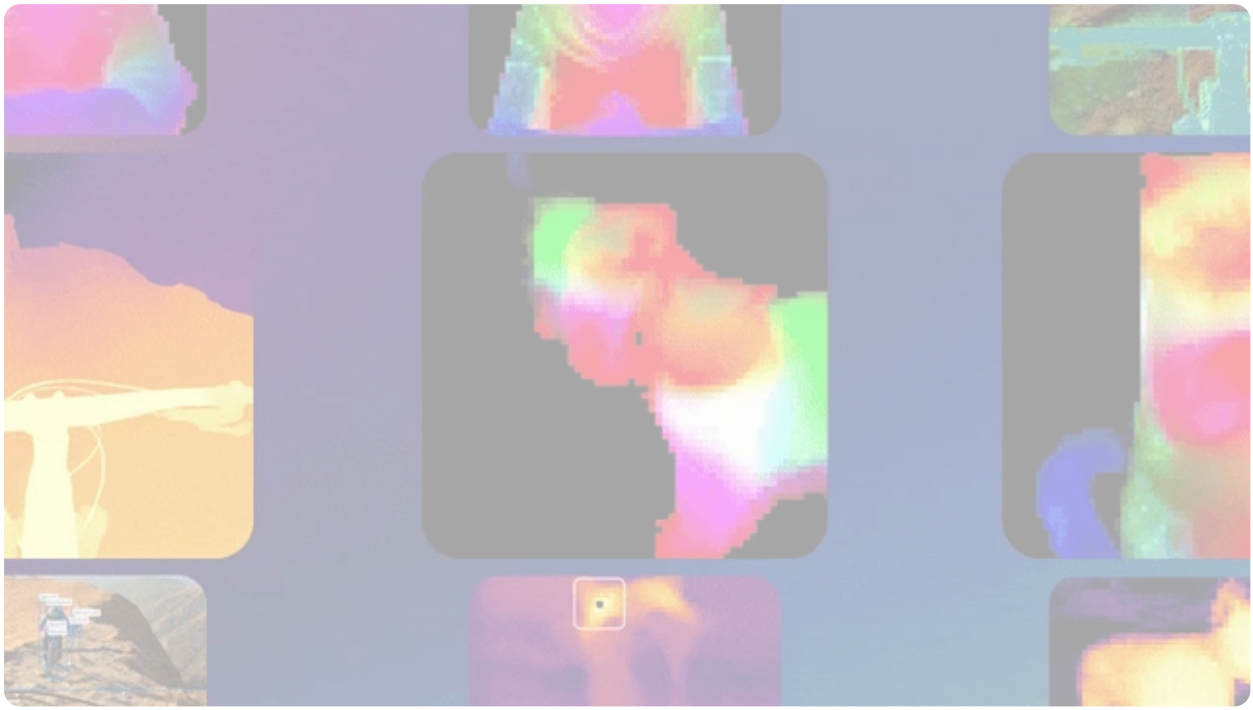
DINOv3 está diseñado para conservar detalle espacial al escalar el entrenamiento.



DINOv3 introduce gram anchoring para combatir la degradación de mapas densos en entrenamientos prolongados. El resultado es un backbone congelado útil en tareas de imagen natural, aérea y satelital.

El mismo modelo, varias escalas

Las representaciones densas separan regiones sin recibir una taxonomía de píxeles.



Estas visualizaciones muestran estructura emergente, pero una aplicación necesita una cabeza o regla downstream, datos de validación y una métrica acorde al objetivo.

Representar no es decidir

DINO encaja cuando faltan etiquetas y necesitas características reutilizables.

Úsalo para similitud, clustering o rasgos densos.

Compara con supervisado si hay muchas etiquetas fiables.

Mide coste y latencia en el hardware real.

Quieres detectar una anomalía industrial con DINO. ¿Qué falta además del embedding?

- A Ejemplos normales y evaluación del umbral.
- B Una etiqueta textual para CLIP.
- C Solo más memoria GPU.

Respuesta A. Exacto: la distancia requiere referencias y una decisión validada.

Un modelo supervisado pequeño puede ser mejor en una tarea fija y un entorno controlado. DINO aporta valor cuando las clases cambian, la anotación es costosa o el mismo backbone debe alimentar varias tareas.

Aprender cómo se ve lo correcto

La detección de anomalías parte de ejemplos normales, no de una lista completa de fallos.

Referencias	Patches	CoreSet
Imágenes sin defecto y bien alineadas.	Embeddings DINO normalizados.	Subconjunto que resume la variación normal.

Un enfoque tipo PatchCore almacena representaciones de regiones normales. La cobertura de variaciones válidas —iluminación, postura, lote o cámara— es decisiva para evitar falsas alarmas.

La distancia localiza la desviación

Cada patch nuevo se compara con su vecino normal más cercano.

Imagen

Extrae patches de test.

kNN

Busca vecinos normales.

Mapa

La distancia marca regiones.

Umbral

Convierte score en decisión.

La distancia puede ser coseno o L2. El heatmap se construye con scores por patch y el score global resume la imagen. Umbral, sensibilidad y localización deben validarse con fallos reales o simulados.

Explora los embeddings

Extrae representaciones globales y densas con DINOv2.

1. Guarda una copia en Drive.
2. Compara dos imágenes.
3. Interpreta patches y PCA.

Abrir en Colab ↗

Compara el token global para similitud entre imágenes y los tokens de patch para estructura espacial. Observa si PCA, vecinos cercanos y distancia coseno cuentan una historia coherente; una visualización atractiva no sustituye la validación de la tarea final.