Otras tareas fundacionales de visión artificial

Bonus extra.

Estimación de pose

Identificar y localizar puntos clave del cuerpo humano o de objetos en imágenes y vídeos.

Super-resolución

Mejorar la calidad y resolución de imágenes o vídeos, a menudo a partir de datos de baja resolución.

Estimación de profundidad

Calcular la distancia de los objetos en una escena respecto a la cámara, creando mapas de profundidad.

OCR moderno

Extraer texto de imágenes y documentos, transformándolo en datos editables y buscables.

Background removal

Segmentar y eliminar automáticamente el fondo de una imagen, aislando el objeto principal.

Image matching

Encontrar correspondencias o similitudes entre dos o más imágenes para tareas como la localización o el reconocimiento.

Estimación de pose humana: de píxeles a esqueletos

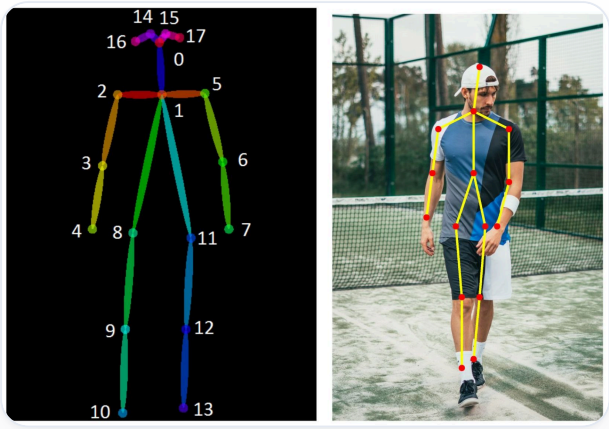
Estructurando el cuerpo humano en coordenadas digitales.

¿Qué es? Tarea de visión que localiza articulaciones clave (keypoints) para modelar la postura y el movimiento.

El estándar (COCO-17): se buscan típicamente 17 puntos.

- **Cabeza:** nariz, ojos (2), orejas (2).
- **Tren superior:** hombros (2), codos (2), muñecas (2).
- **Tren inferior:** caderas (2), rodillas (2), tobillos (2).

El output: una lista de coordenadas (x, y) más un score de confianza por punto.



> Transformamos una matriz de píxeles ininteligible en una estructura geométrica analizable (skeleton).

Enfoques de modelado: top-down vs. bottom-up

Trade-off entre precisión y velocidad en multitudes.

Top-down (alta precisión)

Proceso: primero detecta personas (cajas), luego estima la pose dentro de cada una.

Ejemplos: HRNet, ViTPose.

- Pros: máxima precisión, robustez a distintas escalas.
- Contras: lento en grandes aglomeraciones, con coste lineal en el número de personas.

Bottom-up (alta velocidad)

Proceso: detecta todos los puntos clave de la imagen a la vez, luego los agrupa en personas.

Ejemplo: OpenPose.

- Pros: velocidad constante, independiente del número de personas.
- Contras: menos preciso, el agrupamiento de puntos es complejo.

Single-stage (el equilibrio moderno)

Proceso: redes como YOLO-Pose predicen cajas y puntos simultáneamente en una sola pasada (end-to-end).

- Ventaja: velocidad en tiempo real con precisión competitiva, ideal para dispositivos edge.

Modelos destacados y evaluación

Herramientas prácticas para cada escenario.

01 **MediaPipe (Google)**

Funciona eficientemente en CPU. Ideal para aplicaciones de fitness, filtros AR o escenarios de una sola persona. El problema es que solo da los keypoints de una sola persona. También sirve para manos y cara.

02 **YOLOv8-Pose (ultralytics)**

El estándar industrial actual. Ofrece un equilibrio óptimo entre velocidad y precisión. Excelente para detección multi-persona y es fácil de entrenar y adaptar.

03 **OpenPose (CMU)**

El pionero académico que estableció las bases. Aunque influyente, hoy en día es computacionalmente pesado en comparación con opciones más modernas.

04 **HRNet**

Prioridad en precisión extrema. Cuando la exactitud es crítica (análisis médico, biomecánico) y la velocidad de inferencia es secundaria.

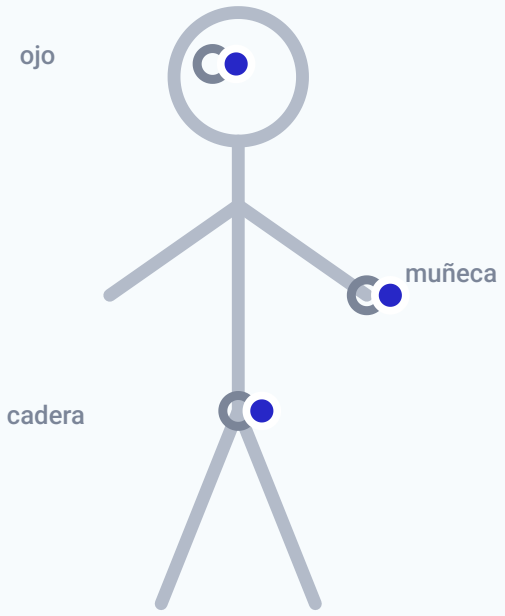
¿Cómo medimos la calidad? La métrica OKS

OKS (Object Keypoint Similarity) es el equivalente a IoU para esqueletos. No mide la distancia en píxeles de forma bruta, sino una «similitud perceptual» entre los puntos clave detectados y los reales.

El OKS **penaliza de forma inteligente**: un error de 5 píxeles en un punto rígido como el ojo penaliza mucho más que un error de 5 píxeles en una articulación flexible como la muñeca o la cadera.

El mismo error no penaliza igual

La tolerancia depende de la articulación y del tamaño de la persona. Mueve las dos cosas y mira cuándo la pose deja de contar como acierto.



El muñeco mide siempre lo mismo y representa a la persona entera: si la persona encoge, el mismo error en píxeles se ve mayor sobre ella. Las tolerancias son las constantes de COCO.

Cuenta como acierto

OKS del esqueleto

0,92

Umbral de acierto

0,50

Ojo rígido

0,80

Muñeca flexible

0,96

Cadera flexible

0,99

En una persona de 150 px, 5 px dejan el ojo en 0,80 y la cadera en 0,99. El ojo pierde medio punto a partir de 9 px.

> Esto explica el fallo más habitual en planta: el modelo va bien con la persona cerca y se desmorona con la persona al fondo, aunque el error en píxeles sea el mismo. Las dos salidas son acercar la cámara o trabajar en top-down, detectando primero a cada persona y recortándola para que llegue grande al modelo de pose.

Esta página recoge el estado con 5 px de error en una persona de 150 px. OKS aplica a cada punto $\exp(-d^2 / (2\,s^2\,k^2))$, donde d es el error en píxeles, s la escala de la persona y $k = 2\sigma$ una constante propia de cada articulación medida sobre anotaciones humanas. Los ojos tienen $\sigma = 0{,}025$ y la cadera $\sigma = 0{,}107$, así que la cadera tolera más de cuatro veces el error antes de penalizar lo mismo. La escala entra dividiendo, de modo que la tolerancia en píxeles es proporcional al tamaño de la persona: en una persona de 60 px el ojo pierde medio punto a partir de 4 px, y en una de 400 px aguanta hasta 24 px. El OKS del esqueleto es la media de los puntos visibles, y sustituye al IoU cuando se calcula AP para

pose: COCO promedia AP con umbrales de 0,50 a 0,95, así que el 0,50 de esta pantalla es el criterio más indulgente de los diez.

Cinco píxeles no son cinco píxeles

La tolerancia de cada articulación sale de cómo de acuerdo se ponen los anotadores humanos.

Dos articulaciones fallan cinco píxeles. ¿Penaliza OKS siempre igual?

- A Sí, solo cuenta la distancia en píxeles.
- B No, depende de escala y articulación.
- C No, solo cuenta cuántas personas hay.

Respuesta B. Exacto: normaliza por tamaño y usa una tolerancia específica por keypoint.

Las constantes por articulación se estimaron midiendo la dispersión entre anotadores. Donde las personas coinciden mucho, como los ojos, el modelo tiene menos margen.

Impacto real y desafíos técnicos

Dónde genera valor la estimación de pose.

Sport y fitness

- **Entrenador IA:** conteo preciso de repeticiones (sentadillas, flexiones…).
- **Análisis de técnica:** medición de ángulos corporales para prevenir lesiones.
- Aplicaciones en plataformas como Nike Training o Peloton.

Seguridad industrial y ergonomía

- **Detección de posturas peligrosas:** identificar movimientos de riesgo al levantar peso.
- **«Hombre caído»:** alertas de seguridad en tiempo real en entornos peligrosos.
- **Optimización:** mejora de la eficiencia en líneas de montaje.

Salud y rehabilitación

- **Fisioterapia remota:** validación de ejercicios en casa sin supervisión presencial.
- **Análisis de marcha:** detección temprana de patologías neurológicas o musculoesqueléticas.

El gran reto: oclusiones y reconstrucción 3D

El problema: la estimación de pose se complica cuando partes del cuerpo se ocultan entre sí o por objetos externos, perdiendo información crucial.

La solución: modelos avanzados emplean memoria temporal para rastrear y predecir oclusiones, o realizan inferencia 3D para «imaginar» las partes ocultas basándose en la anatomía y el contexto.

OCR moderno: de leer caracteres a entender documentos

La evolución de Tesseract a los transformers (TrOCR).

El pasado: enfoque clásico (Tesseract)

Metodología: basado en reglas heurísticas y detección de contornos para identificar caracteres individuales.

Limitaciones comunes:

- Falla con texto curvo, distorsionado o en perspectivas inusuales.
- Poca robustez ante fuentes manuscritas complejas o variaciones tipográficas.
- Sensible al ruido de fondo, baja resolución o iluminación deficiente.
- Requiere preprocesamiento intensivo (normalización, binarización).

El presente: transformers para OCR (TrOCR y VLM)

Arquitectura end-to-end: utiliza Vision Transformers (ViT) para codificar la imagen de forma global y un modelo de lenguaje (como BERT o GPT) decodifica la secuencia de texto resultante.

Ventaja fundamental: no «lee» letra a letra, sino que «entiende» la secuencia visual completa del texto en contexto, de forma similar a como los humanos leen.

Estado del arte (SOTA):

- **TrOCR (Microsoft):** un referente específico para tareas de OCR.
- **Qwen 2.5/3-VL:** un VLM avanzado que integra OCR con capacidades de razonamiento multimodal.

Herramientas prácticas de OCR

Implementación rápida con Hugging Face y EasyOCR.

EasyOCR: el comodín versátil

Una opción todo terreno para cualquier proyecto. Es ligera, soporta más de 80 idiomas y funciona eficientemente en CPU. Ideal para prototipos rápidos y escenarios donde no se dispone de GPU.

TrOCR (Hugging Face): precisión avanzada

Ofrece la máxima precisión, especialmente diseñado para textos manuscritos, históricos o muy específicos. Requiere una GPU para un rendimiento óptimo debido a su arquitectura basada en Transformers.

VLM (Qwen, GPT-4V): documentos estructurados

Perfectos para casos de uso complejos como el procesamiento de facturas o tablas. Permiten extraer información estructurada, por ejemplo JSON, en lugar de solo texto plano, integrando visión y lenguaje para un razonamiento avanzado.

Lo que devuelve cada generación de OCR

La misma foto de una placa, leída de tres formas. Cambia la herramienta y mira qué recibe tu programa.

Serigrafía de una placa real, con texto pequeño, girado por el brillo del cobre y mezclado con los componentes.

Qué herramienta usas

- **Tesseract: solo texto**
- EasyOCR o TrOCR: texto y cajas
- VLM: campos estructurados

```
C1  C5  R6  C4  PULSE
T1  R3  MCP6004
IC2  R4  R11  R10  R5  R9
J1  VCC  EN  VOUT  GND
R1  R2  R7  R8
Place finger  SENSOR
Easy Pulse
www.Embedded-Lab.com
```

Sale una tira de caracteres y poco más. El orden lo inventa el recorrido por líneas, así que «VCC EN VOUT GND» y las referencias de las resistencias se mezclan sin que puedas saber qué va con qué.

> La pregunta útil no es qué herramienta lee mejor, sino qué necesita tu programa. Para contar si la serigrafía está impresa basta la primera; para comprobar que cada referencia está junto a su componente hacen falta coordenadas; para rellenar un parte de recepción con el modelo y el fabricante, el tercero te ahorra el pegamento.

Las cajas de esta pantalla están medidas a mano sobre la foto para ilustrar el formato de salida, no las ha producido un OCR. Tres cosas que se ven en esta imagen y que aparecen en cualquier planta: el texto de la serigrafía es de bajo contraste sobre verde y el brillo del cobre lo rompe, hay texto a distintas escalas en la misma foto, y varias referencias quedan parcialmente tapadas por los componentes. Por eso la resolución de captura manda más que el modelo: con menos de veinte píxeles de altura por

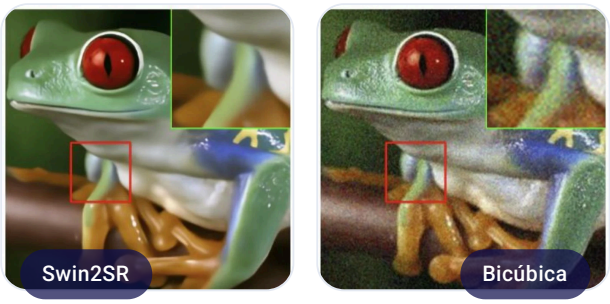
carácter, ningún OCR actual lee de forma fiable. Sobre latencia y coste, Tesseract corre en CPU en decenas de milisegundos, un modelo de detección más reconocimiento ronda las décimas de segundo en GPU, y un VLM se va a varios segundos por imagen, así que en línea suele combinarse: detección rápida siempre y VLM solo sobre las piezas dudosas.

Super-resolución: restaurando el detalle perdido

El uso de deep learning para generar detalles plausibles de alta frecuencia a partir de una imagen de baja resolución, superando los límites de la interpolación tradicional (bilineal, bicúbica…).

El modelo SOTA: Swin2SR.

- Basado en Swin Transformers V2, lo que le permite procesar imágenes de manera eficiente y global.
- Entrenado para escalar la imagen y también para eliminar artefactos de compresión JPEG y ruido.
- **Ventaja clave:** recupera y sintetiza texturas finas (como pelo, tela o detalles faciales) que los métodos clásicos dejarían borrosas o artificiales.



Mira el recuadro rojo: la textura fina de la derecha la sintetiza el modelo.

> **Nota importante:** la super-resolución no es magia. El modelo «imagina» detalles basándose en patrones aprendidos de su vasto conjunto de entrenamiento, no reconstruye información que nunca existió.

Aplicaciones reales y limitaciones

Dónde aplicar superresolución en industria.

Restauración de archivos

- Digitalización de fotos antiguas o documentos históricos dañados.
- Mejora de contenido multimedia antiguo (remasterización).

Medicina e imagen satelital

- Mejorar la legibilidad de escáneres médicos.
- Zoom en imágenes aéreas o satelitales.

E-commerce y marketing

- Mejorar fotos de productos de baja calidad enviadas por usuarios o proveedores.
- Lograr una apariencia profesional en la web para el catálogo de productos.

Riesgo crítico: «alucinaciones». El modelo puede inventar un tumor o una mancha. Nunca usar para diagnóstico crítico sin supervisión.

Detalle generado no es evidencia

Lo que aparece al ampliar puede venir del modelo, no del sensor.

Una superresolución muestra una mancha nueva en una imagen médica. ¿Qué concluyes?

- A La mancha estaba en el original.
- B Hay que validarla contra el original y el proceso clínico.
- C Una mayor resolución siempre elimina el error.

Respuesta B. Exacto: la mejora visual no convierte el detalle generado en evidencia.

La superresolución sintetiza textura plausible a partir de patrones aprendidos. En un contexto diagnóstico, cualquier estructura nueva debe comprobarse sobre la imagen original.

Eliminación de fondo profesional

El reto del «alpha matting» y el modelo RMBG.

El problema: bordes difusos

Segmentar objetos sólidos es fácil. Pero elementos como el pelo, el pelaje o el vidrio transparente son extremadamente difíciles de separar debido a sus bordes difusos y semitransparentes.

La solución: RMBG-1.4 (BRIA AI)

Un modelo de vanguardia diseñado para el alpha matting. Genera un canal alfa de transparencia gradual en los bordes, no una máscara binaria. Esto permite recortes profesionales que se integran perfectamente en cualquier fondo, sin el efecto «pegatina».

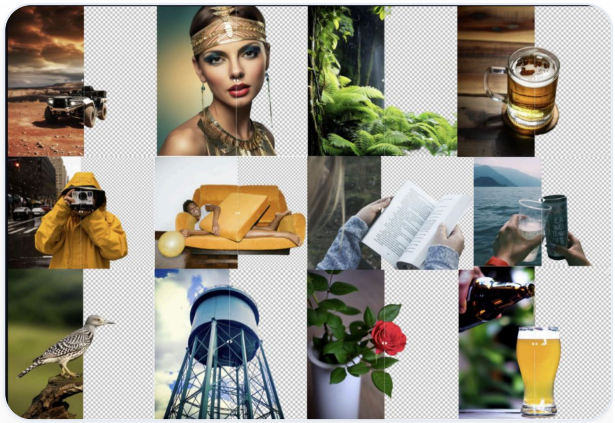


Image matching: encontrando parejas visuales

El fin de los algoritmos heurísticos (SIFT, ORB…).

¿Qué es? Encontrar puntos idénticos en dos fotos de la misma escena tomadas desde ángulos diferentes, a pesar de cambios en la perspectiva o la iluminación. Esta tarea es fundamental para aplicaciones como la reconstrucción 3D, la realidad aumentada y la localización de robots.

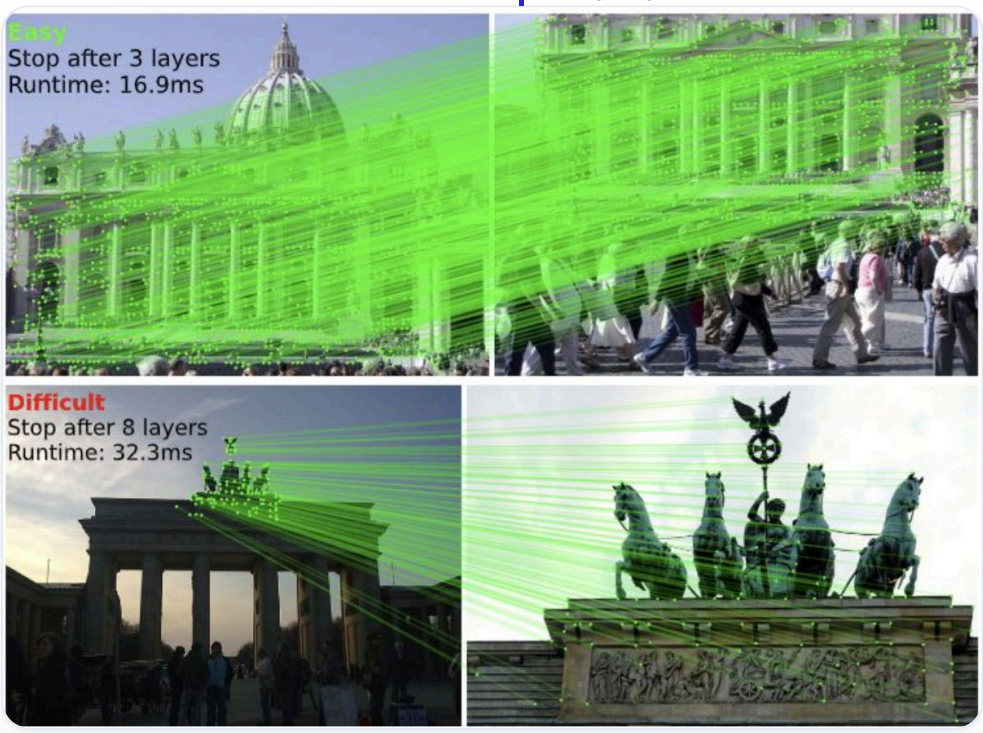
La nueva era: deep learning

Detector (SuperPoint, DISK)

Una red neuronal avanzada que identifica los puntos clave y más distintivos (esquinas, texturas complejas) en una imagen, superando con creces la precisión de los algoritmos matemáticos clásicos.

Matcher (LightGlue)

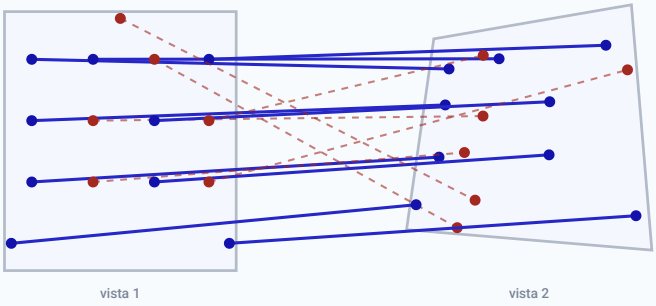
Un modelo basado en Transformer que evalúa los puntos detectados en ambas imágenes y determina cuáles corresponden entre sí, filtrando eficazmente los errores geométricos y emparejamientos incorrectos.



Ventaja clave: este enfoque moderno funciona muy bien incluso con cambios de luz (día y noche), grandes variaciones de perspectiva y oclusiones, ofreciendo una gran robustez.

La mitad de las parejas sobran

Un matcher devuelve más parejas de las que valen. El filtro que las separa no mira el aspecto, mira la geometría: mueve el umbral y comprueba cuáles encajan con la escena.



Esquema con una homografía real: la misma fachada vista desde otro sitio. Las parejas y su error se calculan en el navegador.

Estimación sólida

Parejas aceptadas

9

Descartadas

6

9 parejas coherentes y 6 descartadas. Las descartadas son las que se parecían sin estar en el mismo sitio.

Este umbral es el que se pasa a `cv2.findHomography` con RANSAC, y es el mando que más cambia el resultado en fotogrametría y en SLAM. Demasiado bajo y te quedas sin puntos para calcular; demasiado alto y una ventana confundida con otra te tuerce la pose de la cámara.

Esta página recoge el estado con 2 px de umbral. El procedimiento real es RANSAC: se eligen cuatro parejas al azar, se calcula la homografía que implican, se cuenta cuántas de las demás quedan por debajo del umbral de reproyección, y se repite unos cientos de veces quedándose con la hipótesis que más apoyo reúne. Esta pantalla enseña el paso de recuento con la homografía correcta ya fijada, que es donde se ve el compromiso. Las cuatro parejas muy alejadas son los errores típicos del matching: dos ventanas iguales en fachadas repetidas, un reflejo y un punto en el cielo, donde no hay textura estable. Las dos parejas intermedias, a 4,5 y 7,5 px, son las que deciden si el umbral está bien puesto: entran al subirlo y arrastran error a la estimación. Con imágenes de verdad, el umbral se elige en píxeles de la imagen original, así que cambiar la resolución de captura obliga a revisarlo.

Reconstruyendo el mundo en 3D

Aplicaciones críticas del matching.

Fotogrametría (SfM)

Al identificar y conectar puntos coincidentes en múltiples fotos de una escena, se puede reconstruir la geometría 3D del objeto y la posición de la cámara en cada toma.

Panoramas (stitching)

Permite fusionar sin problemas varias imágenes superpuestas en una única foto de gran formato o una vista de 360 grados, eliminando duplicados y alineando las perspectivas.

SLAM (robótica)

Un robot utiliza el matching de características para comprender su posición dentro de un entorno y construir un mapa del mismo, incluso mientras se mueve y explora.

Depth estimation: 3D desde una sola imagen

La revolución de la profundidad monocular.

Depth Anything V2/V3

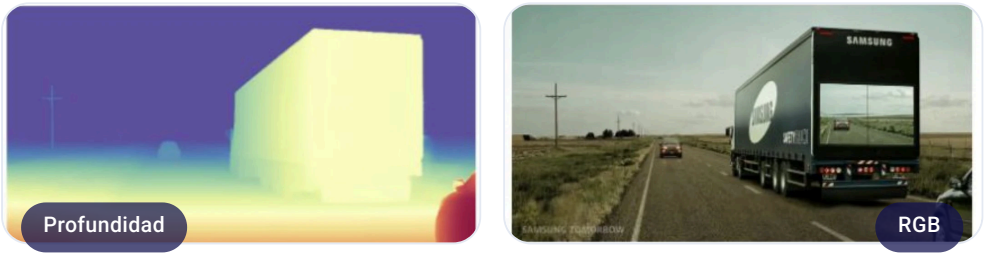
Un modelo fundacional entrenado con millones de imágenes etiquetadas y no etiquetadas, aprovechando la diversidad del mundo real.

Capacidad

Genera un mapa de profundidad (heatmap) donde el color indica la distancia relativa, revelando la estructura 3D subyacente.

Robustez

Funciona en una gran variedad de escenarios: interiores complejos, exteriores dinámicos, y con elementos desafiantes como objetos transparentes o superficies reflejantes.



Una sola cámara y una sola imagen: el modelo deduce qué está cerca y qué está lejos.

Aplicaciones depth

¿Para qué sirve un mapa de profundidad?

Efecto retrato (bokeh)

Desenfocar el fondo sintéticamente, creando imágenes con una profundidad de campo profesional y destacando el sujeto principal.

Robótica y navegación

Permitir que robots y vehículos autónomos detecten obstáculos y comprendan su entorno 3D, utilizando solo una cámara de bajo coste.

Generación de nubes de puntos

Convertir una foto plana en una estructura 3D navegable (malla o nube de puntos) al combinar la imagen RGB con su mapa de profundidad.

Filtro de imagen

Podemos eliminar partes de la imagen que no están en la escena que nos interesa aplicando un threshold.

Profundidad relativa vs. métrica. Estos modelos suelen estimar profundidad relativa (mapas de calor). Para obtener metros reales se necesita calibración de cámara o modelos específicos de metric depth estimation.

Relativa no es métrica

Un mapa normalizado ordena distancias, pero no las mide.

Necesitas saber si un obstáculo está a 1,5 metros. ¿Basta un mapa relativo normalizado?

- A Sí, porque el color ya representa metros.
- B Sí, si la imagen tiene buena resolución.
- C No, hace falta escala métrica o calibración.

Respuesta C. Exacto: necesitas un modelo métrico, calibración u otra referencia de escala.

Un mapa relativo dice qué está más cerca que qué. Para responder «a cuántos metros» hacen falta calibración de cámara, una referencia conocida o un modelo entrenado para profundidad métrica.

Las seis tareas, seis recetas

Lo mínimo para arrancar cada una. Todas siguen el mismo patrón de Transformers.

```
from transformers import pipeline
from PIL import Image

image = Image.open("mi_imagen.jpg").convert("RGB")

# 1. Pose: puntos clave del cuerpo
pose = pipeline("keypoint-detection", model="usyd-community/vitpose-base-simple")
print(pose(image))

# 2. OCR: texto de la imagen
ocr = pipeline("image-to-text", model="microsoft/trocr-base-printed")
print(ocr(image))

# 3. Superresolución: más detalle
sr = pipeline("image-to-image", model="caidas/swin2SR-classical-sr-x2-64")
sr(image).save("ampliada.png")

# 4. Fondo: recorte con canal alfa
fondo = pipeline("image-segmentation", model="briaai/RMBG-1.4")
fondo(image).save("sin_fondo.png")

# 5. Profundidad: distancia relativa
depth = pipeline("depth-estimation", model="depth-anything/Depth-Anything-V2-
Small-hf")
depth(image)["depth"].save("profundidad.png")

# 6. Matching: correspondencias entre dos fotos
match = pipeline("keypoint-matching", model="ETH-CVG/lightglue_superpoint")
print(match([[Image.open("vista_1.jpg"), Image.open("vista_2.jpg")]]))
```

- Cada tarea es independiente: ejecuta solo la que necesites y no cargues el resto.
- Revisa la licencia antes de llevarlo a producción. **RMBG-1.4 no permite uso comercial** sin acuerdo con BRIA.
- Para encadenar dos, empieza por profundidad y usa el mapa como máscara: es la combinación más directa.

Los identificadores corresponden a modelos publicados en Hugging Face a fecha de septiembre de 2026. La API de `pipeline` puede cambiar el nombre de alguna tarea entre versiones de transformers; si una falla, busca la tarea en la ficha del modelo. El cuaderno del tema ejecuta estas seis con imágenes de ejemplo y deja el encadenado como ejercicio.

Antes de usar cualquiera de estas seis

Seis comprobaciones comunes a tareas muy distintas.

- [ ] **Sabes qué es medida y qué es generación**
  Pose y profundidad estiman; superresolución sintetiza. No se usan igual como evidencia.
- [ ] **La licencia permite tu uso**
  Varias de estas familias tienen restricciones comerciales. Compruébalo por modelo.
- [ ] **La escala está resuelta**
  Si necesitas metros, un mapa relativo no sirve: hace falta calibración o un modelo métrico.
- [ ] **Probado con tus imágenes**
  Resolución, encuadre y calidad reales, no las fotos de la demo.
- [ ] **Sin diagnóstico automático**
  En medicina o seguridad, revisión humana sobre la imagen original.
- [ ] **Coste por imagen calculado**
  Si vas a procesar miles, mide tiempo y memoria antes de comprometerte.

Cinco tareas, un cuaderno

Pose, OCR, superresolución, fondo y profundidad sobre tus propias imágenes.

1. Guarda una copia en Drive.
2. Ejecuta cada tarea por separado.
3. Combina dos pipelines y comenta el resultado.

Abrir en Colab ↗

Cada tarea es independiente: puedes ejecutar solo la que te interese. La ampliación consiste en encadenar dos pipelines, por ejemplo estimar profundidad y usar un umbral para recortar el fondo, o pasar un recorte con superresolución al OCR. Anota qué modelo, qué versión y qué error observas en cada paso.