

DiffCut: Catalyzing Zero-Shot Semantic Segmentation with Diffusion Features and Recursive Normalized Cut

Paul Couairon, Mustafa Shukor, Jean-Emmanuel Haugeard,
Matthieu Cord, Nicolas Thome

NeurIPS 2024

19 de marzo de 2025

Introducción

- ▶ **Segmentación semántica:** Proceso de dividir una imagen en regiones donde cada región corresponde a un concepto.
- ▶ **Zero-shot:** Segmentación sin ejemplos etiquetados previamente.
- ▶ **Modelos de difusión:** Modelos generativos que transforman ruido en imágenes reales mediante un proceso iterativo de denoising.

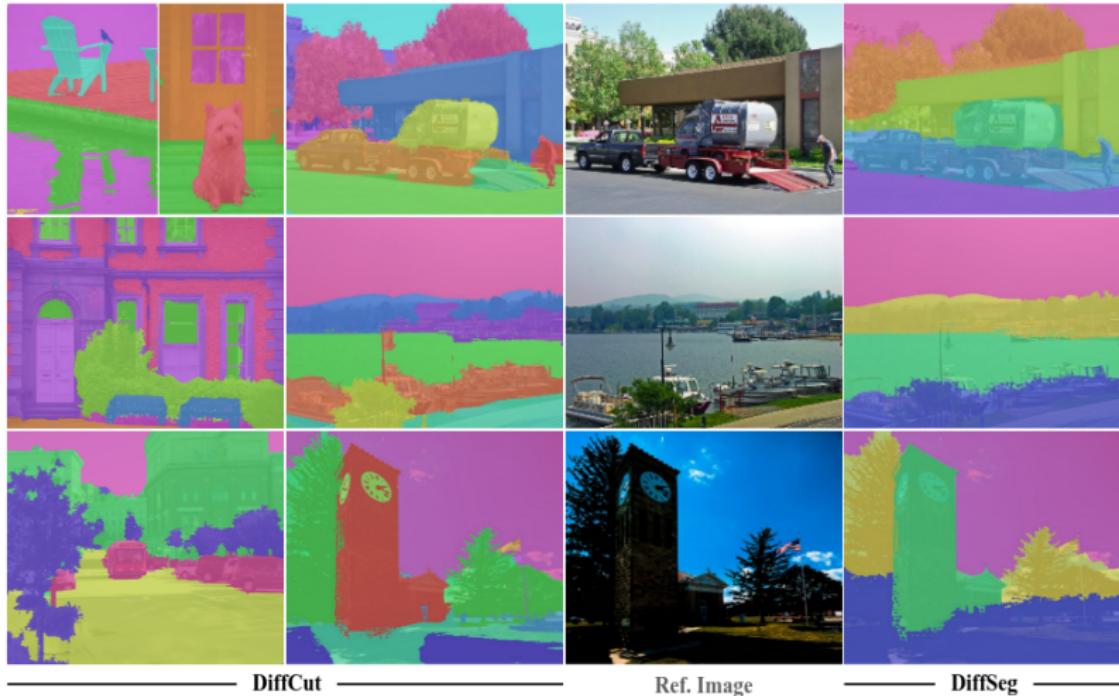


Figura: DiffCut Example

Modelos de Difusión y Extracción de Características

► Modelos de difusión:

- Generan imágenes a partir de ruido.
- **Latent diffusion:** Se codifica la imagen en un espacio latente utilizando un VQ-encoder para mejorar la eficiencia computacional.

► Encoder UNet y Self-Attention:

- **UNet:** Arquitectura con un encoder y un decoder conectados por skip connections.
- **Self-Attention:** Mecanismo que permite a cada posición atender a todas las demás para capturar relaciones globales.
- Se extraen las características de la *última capa de self-attention*, denotadas como \hat{z} , por su alta calidad semántica.

Diffusion models



Figura: Diffusion Models

Construcción de la Matriz de Afinidad

- ▶ **Similitud Coseno:** Para dos vectores \hat{z}_i y \hat{z}_j , se define:

$$\text{sim}(i, j) = \frac{\langle \hat{z}_i, \hat{z}_j \rangle}{\|\hat{z}_i\|_2 \|\hat{z}_j\|_2}$$

- ▶ Los valores se normalizan en el rango $[0, 1]$.
- ▶ **Aplicación del exponente α :**

$$W_{ij} = \left(\frac{\langle \hat{z}_i, \hat{z}_j \rangle}{\|\hat{z}_i\|_2 \|\hat{z}_j\|_2} \right)^\alpha$$

Este exponente actúa como un umbral suave, enfatizando las similitudes fuertes y suprimiendo las débiles.

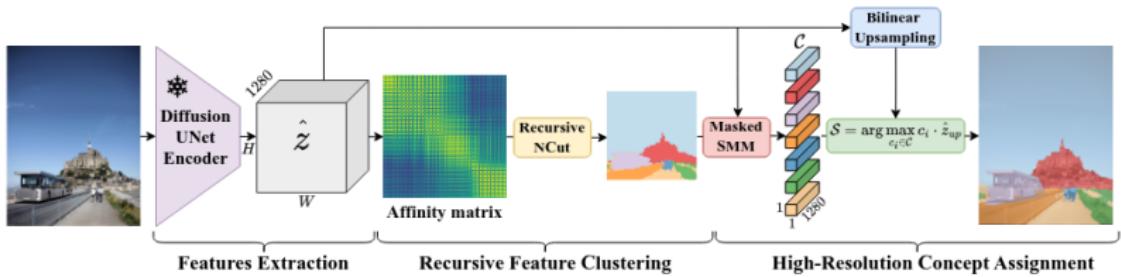


Figura: Pipeline

Normalized Cut y Problema Eigen

- ▶ **Normalized Cut (NCut):** Objetivo: Minimizar la disimilitud entre clusters y maximizar la similitud interna.
- ▶ Definición del corte:

$$\text{NCut}(A, B) = \frac{\text{cut}(A, B)}{\text{assoc}(A, V)} + \frac{\text{cut}(A, B)}{\text{assoc}(B, V)}$$

donde **cut** es la suma de los pesos entre A y B, y **assoc** es la suma de los pesos de A con todos los nodos.

- ▶ **Matriz diagonal D:** Con elementos

$$d(i) = \sum_j W_{ij}$$

- ▶ Se formula el problema como:

$$(D - W)x = \lambda Dx$$

Buscamos el eigenvector asociado al segundo menor eigenvalor, conocido como el *Fiedler vector*.

- ▶ **Bipartición:** Se evalúan múltiples puntos de corte en el Fiedler vector y se elige aquel que minimice el NCut.

Partición Recursiva y Control de Granularidad

- ▶ En lugar de una única bipartición, se aplica el NCut de forma recursiva en cada subgrafo.
- ▶ **Criterio de parada:** Se detiene la partición cuando el costo NCut de un subgrafo supera el umbral τ .
- ▶ τ : Parámetro que regula la granularidad: valores bajos generan menos segmentos y valores altos permiten segmentaciones más finas.

Comparativa τ

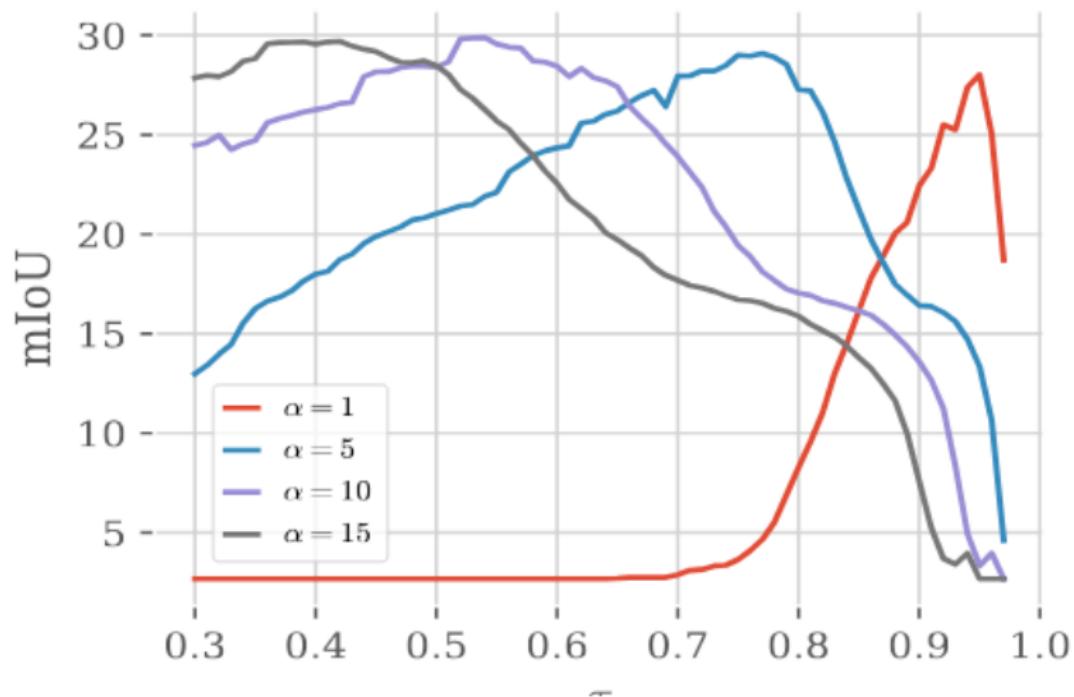


Figura: Sensitivity of DiffCut.

Comparativa τ

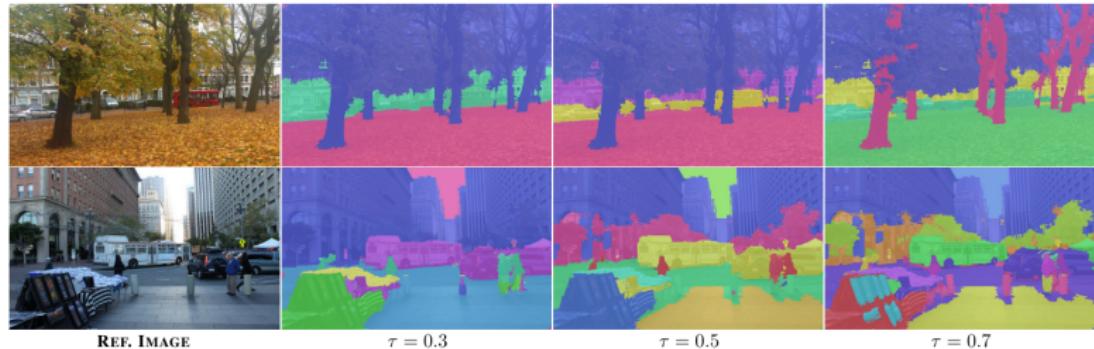


Figura: τ Inference

Asignación de Conceptos en Alta Resolución

- ▶ **Masked Spatial Marginal Mean (SMM):** Se colapsa la dimensión espacial de las características de cada segmento para obtener un *embedding semántico*.
- ▶ **Upsampling Bilineal:** Se reescalas el mapa de segmentación a la resolución original de la imagen.
- ▶ **Asignación de píxeles:** Para cada píxel, se calcula la similitud con los embeddings y se asigna el segmento con mayor similitud (operación de argmax).
- ▶ **Refinamiento con PAMR:** Se utiliza un módulo pixel-adaptativo (PAMR) para afinar los contornos y suavizar la segmentación.

Resumen del Método DiffCut

- ▶ **Entrada:** Imagen original.
- ▶ **Extracción de características:** Codificación con VQ-encoder y procesamiento en el encoder UNet para obtener \hat{z} .
- ▶ **Matriz de Afinidad:** Cálculo de similitud coseno y aplicación del exponente α para formar W .
- ▶ **NCut Recursivo:** Resolución del problema eigen y partición recursiva hasta que el costo NCut supere τ .
- ▶ **Asignación en Alta Resolución:** Uso de Masked SMM, upsampling bilineal y refinamiento con PAMR para obtener el mapa final.

Resultados Experimentales y Comparaciones

- ▶ **Benchmarks:** Pascal VOC, Pascal Context, COCO-Object, COCO-Stuff-27, Cityscapes, ADE20K.
- ▶ **Métrica:** mIoU (mean Intersection over Union) usando asignación mediante Hungarian matching.
- ▶ **Comparaciones:**
 - ▶ DiffCut vs DiffSeg, MaskCut, etc.
 - ▶ Resultados muestran mejoras significativas en mIoU.

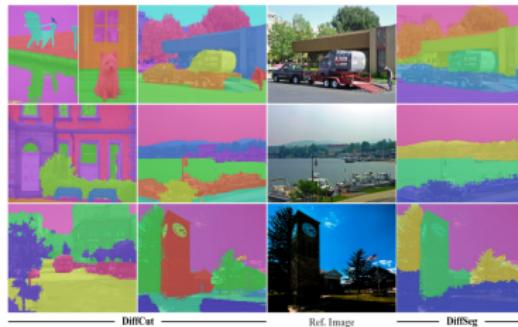
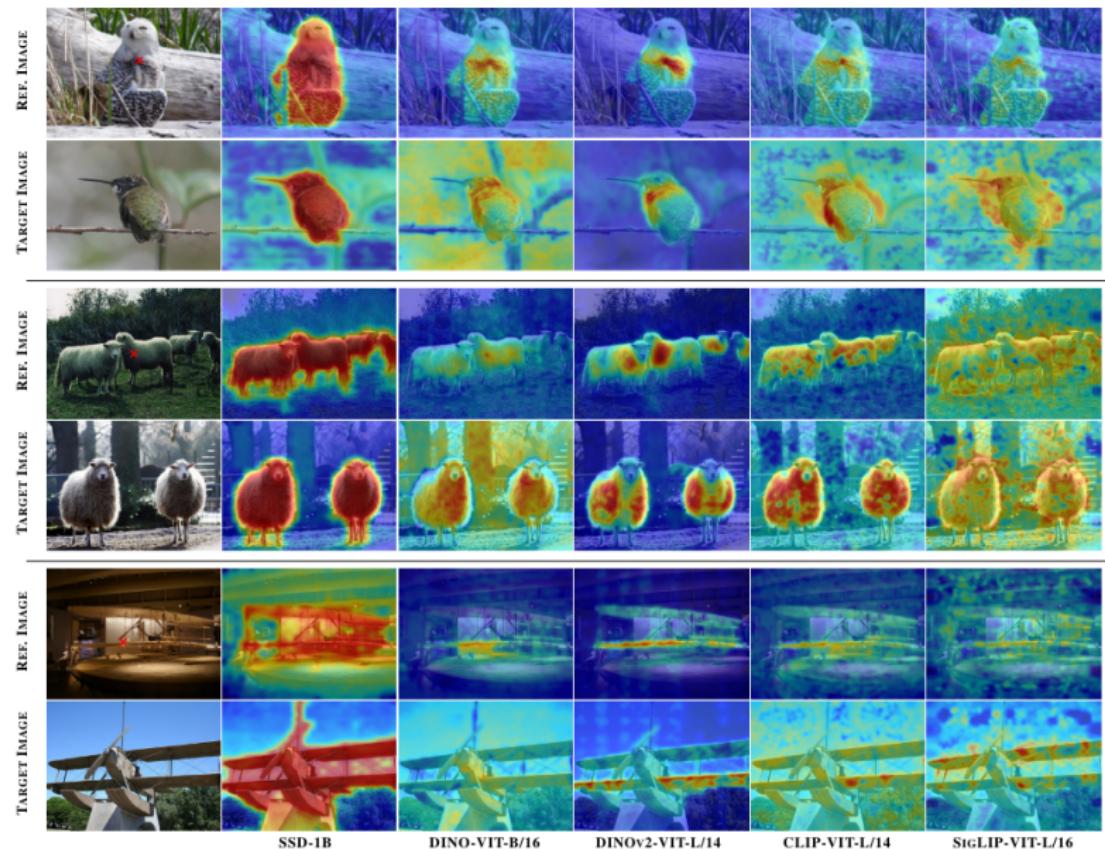


Figura: Comparativa

Comparativa



Estudios Ablativos y Extensión a Open-Vocabulary

► **Estudios Ablativos:**

- ▶ Comparación del uso exclusivo de características del encoder versus combinaciones jerárquicas.
- ▶ Impacto de la partición recursiva frente a métodos de clustering clásico (e.g., AutoSC).
- ▶ Análisis de los hiperparámetros α y τ .

► **Extensión a Open-Vocabulary:**

- ▶ Se incorpora un encoder visual basado en CLIP para asignar etiquetas semánticas.
- ▶ Se realiza mask-pooling y se compara la similitud entre embeddings visuales y textuales.

Conclusiones y Perspectivas Futuras

▶ Contribuciones Principales:

- ▶ Segmentación zero-shot sin necesidad de anotaciones mediante modelos de difusión.
- ▶ Partición recursiva adaptativa con Normalized Cut.
- ▶ Eficiencia computacional al utilizar únicamente el encoder.

▶ Desafíos:

- ▶ Ajuste fino de los hiperparámetros α y τ .
- ▶ Aplicación en dominios específicos (e.g., imágenes biomédicas).
- ▶ Mejor integración con técnicas de open-vocabulary.

Sesión de Preguntas

¿Preguntas?