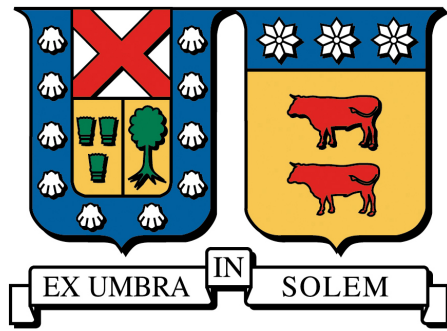




UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

Weakly supervised semantic segmentation

OPTIMIZACIÓN DE LA CLASIFICACIÓN DE MATERIALES DE RECICLAJE
EN CAMIONES USANDO WSSS CON ANOTACIONES A NIVEL IMAGEN.

Memorista:

Daniel Tapia

Profesores Guías:

Marcos Zúñiga

Nicolas Jara

Programa Memorias Multidisciplinarias

Del 14/04/2025 al 14/10/2025



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Weakly supervised semantic segmentation, Optimización de la Clasificación de Materiales de Reciclaje en Camiones usando WSSS con anotaciones a nivel imagen

Nombre del candidato(a): Daniel Iván Tapia Reyes

Carrera / Grado: Ingeniería Civil Telemática

Campus: Casa Central

Departamento: Electrónica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Marcos David Zúñiga Barraza, en mi calidad de profesor(a) guía / director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo NO contiene información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo CONTIENE información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (embargo) por:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 03/07/2026

Firma: _____

Estudiante o Candidato(a):

Fecha: 03/07/2026

Firma: _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

TABLA DE CONTENIDOS

1	Introducción	1
1.1	Contexto	1
1.1.1	Negocio	1
1.1.2	Productos	2
1.2	Problema a Resolver	3
1.3	Acercamiento a la Solución	4
1.4	Hipótesis	4
1.5	Objetivos	5
1.5.1	General	5
1.5.2	Específicos	5
1.6	Estructura de la Memoria	6
2	Estado del arte	7
2.1	Soluciones existentes	7
2.1.1	Segmentación semántica débilmente supervisada (WSSS)	7
2.1.2	Evolución histórica	8
2.2	Herramientas y Métodos	8
2.2.1	Metodologías para la generación de CAM	10
2.2.2	Marcos metodológicos que implican la mejora de las CAM	10
2.3	Discusión	11
2.3.1	Puesta a tierra del SAM	11
2.3.2	Text Region	11

2.3.3	WeCLIP	12
2.3.4	Síntesis de la discusión	13
3	Desarrollo de la solución	15
3.1	Exploración de datos	15
3.2	CAM	15
3.3	Grounding SAM	17
3.4	DINOv2 + Featup	17
3.5	Text Region	17
3.6	WeCLIP+	19
3.7	Limpieza de datos	20
3.8	Entrenamiento	21
4	Validación	22
4.1	Resultados	22
4.2	Optimización	23
4.3	Discusión de resultados	24
4.4	Modelo de segmentación	25
4.5	Validación de la hipótesis	25
5	Evaluación crítica	26
6	Conclusión	27
6.1	Principales logros	27
6.2	Limitaciones	27
6.3	Lecciones aprendidas	27

6.4 Trabajo futuro	28
6.5 Agradecimientos	28
Referencias	29
Referencias Bibliográficas	29

1 Introducción

El 14 de abril comencé mis prácticas finales con el equipo de Datos de Alta Road, una empresa innovadora especializada en la recopilación y análisis de datos relacionados con infraestructuras viales y flujos logísticos. Estas prácticas me brindaron la oportunidad de poner en práctica las habilidades adquiridas durante mis estudios, tanto a nivel técnico como interpersonal, a la vez que descubrí las exigencias y la dinámica de un entorno profesional estimulante. Las prácticas se estructuraron en dos partes: un proyecto de inmersión de un mes, diseñado para familiarizarme con las herramientas, los métodos de trabajo del equipo, así como el entorno técnico de la empresa, seguido de un proyecto principal más ambicioso y de alto impacto, centrado en la evolución del producto estrella de Alta Road.

El objetivo principal del proyecto de inmersión fue modernizar el modelo de anonimización utilizado para procesar imágenes de cámaras instaladas en obras de construcción o logísticas. Estas cámaras, integradas en la solución de Alta Road, capturan imágenes en las que pueden aparecer personas. Para cumplir con la normativa de protección de datos personales (en particular, el RGPD), es fundamental garantizar el anonimato de las personas grabadas. Por lo tanto, mi misión consistió en analizar el modelo existente, proponer mejoras e implementar una nueva versión del modelo de mayor anonimización eficiente.

El segundo proyecto, de carácter más técnico y estratégico, tuvo como objetivo desarrollar un nuevo modelo de detección de materiales basado en las imágenes ya recopiladas. El objetivo era mejorar la capacidad del sistema para distinguir y segmentar con precisión los diferentes materiales presentes, mejorando así la precisión de los análisis automatizados. Este trabajo forma parte del compromiso más amplio de la empresa de aprovechar al máximo la riqueza de sus datos visuales para generar valor añadido para sus clientes.

Más allá de los aspectos puramente técnicos, estas prácticas me permitieron desarrollar importantes habilidades transferibles: colaboración con equipos multidisciplinarios, gestión del tiempo y las prioridades, autonomía en la gestión de proyectos y adaptación a las limitaciones del mundo real. Trabajar en Alta Road también me permitió descubrir un sector en rápida evolución, en la intersección de la tecnología, el medio ambiente y la logística, lo que reforzó mi interés por la IA aplicada a casos concretos y de utilidad social.

1.1 Contexto

1.1.1 Negocio

Altaroad, una startup especializada en optimizar la trazabilidad de los flujos de residuos y materiales en la industria de la construcción, desarrolla soluciones innovadoras para optimizar la toma de decisiones empresariales. La empresa implementa sistemas autónomos equipados con cámaras de IA para el seguimiento automatizado de camiones y el reconocimiento preciso de los materiales transportados, todo ello integrado en una plataforma de monitorización dedicada al cliente.

1.1.2 Productos

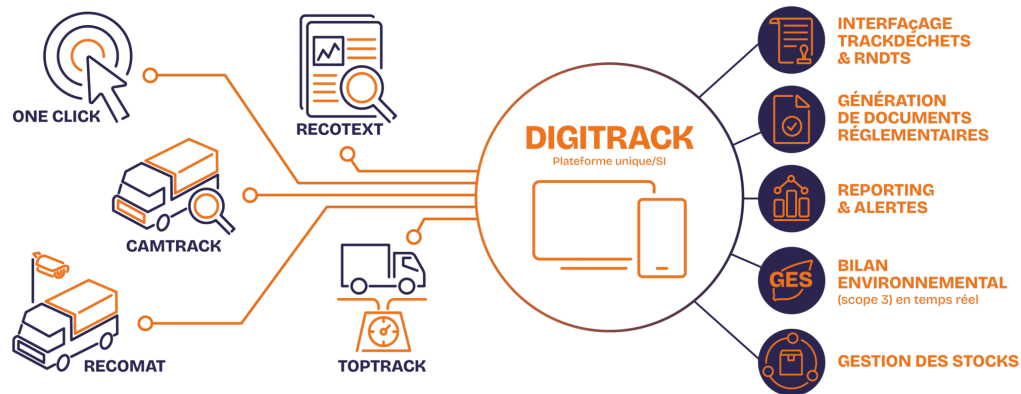


Figura 1.1: Principales productos de Alta Road

Digitrack La interfaz digital que garantiza la trazabilidad de los flujos de materiales y vehículos que entran y salen de las obras. Es accesible desde ordenadores, tabletas y teléfonos, y facilita la trazabilidad mediante el registro automatizado.

Recomat RecoMat es un sistema innovador que automatiza la identificación de materiales transportados por camiones. Utiliza cámaras y modelos de inteligencia artificial para analizar el contenido y generar automáticamente estimaciones sobre los diferentes materiales transportados, su proporción, el nivel de llenado de la caja del camión y la detección de artículos prohibidos en su interior. Esto, a su vez, permite la generación de informes de transporte y reduce significativamente la necesidad de intervención humana, mejora la precisión de los controles y optimiza la trazabilidad de las operaciones logísticas.

Camtrack La plataforma Digitrack está conectada a un sistema de cámaras inteligentes (Camtrack). Las cámaras se instalan en la entrada y/o salida del recinto, eliminando la necesidad de introducir manualmente los datos del vehículo y automatizando la recopilación de información diversa, como las matrículas, que posteriormente se pueden vincular a los vehículos. También se puede utilizar para supervisar los vehículos que entran al recinto.

Toptruck La solución Camtrack se ve reforzada por la integración de un sistema de pesaje móvil diseñado para camiones en movimiento. Este innovador sistema de pesaje es especialmente eficiente en cuanto a tiempo de instalación. Ofrece funciones en tiempo real que alertan rápidamente a los usuarios en caso de sobrecarga o carga insuficiente.

Recotext Solución basada en OCR que permite una gestión eficiente de archivos, automatizando la entrada de datos y reduciendo el tiempo dedicado a tareas administrativas. Garantiza una mayor

precisión de la información y optimiza los procesos documentales, mejorando así la productividad general.

1.2 Problema a Resolver

Hoy en día, uno de los productos estrella de Alta Road y de alto valor añadido para los clientes, es el sistema RecoMat. Esta solución es uno de los pilares tecnológicos de la compañía, donde el desarrollo de la inteligencia artificial desempeña un papel fundamental [4].

El objetivo de RecoMat es identificar los materiales transportados por los camiones. Tradicionalmente, este trabajo lo realizan manualmente los operadores, encargados de completar las hojas de seguimiento de cada vehículo que entra o sale de un área definida. Estas áreas varían según los clientes y sus necesidades específicas: por ejemplo, puede implicar el seguimiento de materiales entre un almacén y una obra, o entre una obra y un área de reciclaje.

La plataforma Digitrack permite personalizar la visualización y el seguimiento de datos según las necesidades del cliente. RecoMat se incorpora como una extensión inteligente que permite la generación automática de informes de transporte de materiales sin intervención humana.

En la práctica, el sistema se basa en la instalación de cámaras de visión superior montadas en postes fijos o móviles que apuntan sobre los camiones. Estas cámaras capturan imágenes de la mercancía transportada que se puede observar por arriba. Estas imágenes se transmiten a los sistemas en la nube de Alta Road, donde se procesan mediante modelos de IA para extraer información sobre el tipo de material detectado (Figura 9). A cada imagen se le asigna una etiqueta correspondiente al material transportado. Toda esta información se envía a la plataforma del cliente, donde se generan automáticamente los documentos asociados y los datos se almacenan para su posterior revisión o verificación en caso de errores.

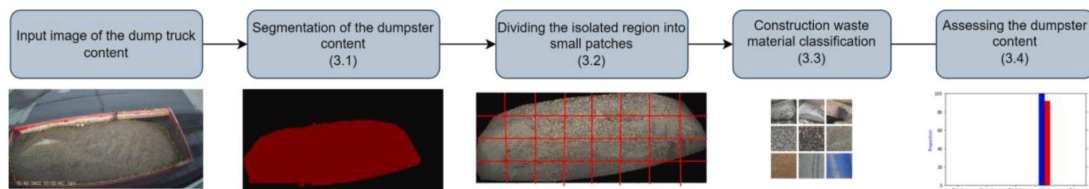


Figura 1.2: Diagrama que explica el proceso de RecoMat de forma simplificada. [4]

Como era de esperar, RecoMat se basa en un modelo de clasificación de imágenes. El algoritmo consta de varios pasos; el primero consiste en la detección automática del contenedor en la imagen de entrada. Una vez aplicado el cuadro delimitador, se extrae el área que contiene únicamente el material transportado.

A continuación, esta sección se divide en sub imágenes (parches), utilizadas para representar las diferentes partes del contenido. Cada parche se envía al modelo de clasificación, que predice el tipo de material presente en esa área. Las predicciones obtenidas se proyectan sobre la imagen del contenedor, generando un mapa de los materiales detectados y una estimación de la proporción de cada material identificado.

El entrenamiento y la optimización de los modelos están a cargo del equipo de Datos, cuyos

esfuerzos se centran en varias áreas: mejorar la precisión, explorar y enriquecer conjuntos de datos, implementar técnicas de postprocesamiento y mantener los modelos en producción.

Es importante destacar que cada cliente posee tipos específicos de materiales que deben clasificarse, lo que requiere el entrenamiento frecuente de nuevos modelos. Por ello, esta tarea forma parte de la rutina diaria del equipo, que actualmente cuenta con una sólida experiencia en este ámbito.

1.3 Acercamiento a la Solución

La estrategia seguida consistió en partir del funcionamiento actual de RecoMat y estudiar si sus anotaciones existentes podían reutilizarse para un problema más fino: pasar de saber qué material aparece en un parche a estimar dónde aparece dentro de la imagen. Para ello, el trabajo se estructuró como una exploración progresiva de métodos WSSS y modelos fundacionales, evaluando primero alternativas como CAMs, Grounding SAM, DINOv2 con FeatUp y Text Region.

A partir de esta exploración, se eligió WeCLIP como enfoque principal, ya que combina CLIP y DINO con codificadores fijos, requiere menos recursos de entrenamiento que otras alternativas y se adapta de forma natural a un escenario donde solo existen etiquetas a nivel de imagen. El proceso propuesto incluye limpieza y balanceo de datos, generación de pseudo-etiquetas, entrenamiento de un decodificador liviano, optimización de *text prompts* y evaluación mediante un conjunto monolabel anotado manualmente.

1.4 Hipótesis

Este proyecto no surgió de un fallo del modelo actual, sino de la necesidad de innovar en el desarrollo de los sistemas de inteligencia artificial de la empresa. El objetivo era responder a la siguiente pregunta: ¿es posible modificar la base de RecoMat para mejorar la identificación espacial de materiales y, al mismo tiempo, garantizar su compatibilidad conceptual con el resto de la cadena de producción?

RecoMat ha estado en producción durante más de 1,5 años, lo que ha permitido la acumulación de varios cientos de miles de imágenes. Este vasto corpus constituye una base de datos valiosa para la mejora continua de la solución. En este contexto, surgió la necesidad de una pasantía de investigación aplicada, centrada en explorar nuevas vías para aumentar la precisión en la detección de materiales.

La hipótesis principal de este trabajo es que el sistema RecoMat puede evolucionar desde una clasificación basada en parches hacia un enfoque de segmentación semántica débilmente supervisada, utilizando únicamente las anotaciones a nivel de imagen ya disponibles. Esta hipótesis se considera verificable si los modelos fundacionales y los métodos WSSS permiten generar máscaras coherentes de materiales, evaluables tanto de forma cualitativa como mediante la métrica mIoU sobre un conjunto de imágenes anotadas manualmente.

De este modo, la validación no se limita a observar si el modelo produce una salida visual, sino que busca determinar si dicha salida es suficientemente consistente para justificar una futura integración en la cadena de RecoMat. La hipótesis se evalúa, por tanto, a partir de tres criterios:

la capacidad de producir máscaras semánticas, la compatibilidad del flujo con las anotaciones existentes y la calidad de los resultados obtenidos en las clases evaluadas.

Este problema dio lugar a una fase inicial de investigación exploratoria destinada a validar la viabilidad técnica del enfoque. Los resultados obtenidos al final de esta etapa fueron prometedores, lo que justificó el lanzamiento de una segunda fase dedicada a pruebas reales con imágenes de la empresa. Asimismo, se prevé una tercera fase orientada a la implementación de este nuevo enfoque en producción. El objetivo a largo plazo es desarrollar un proceso completo, desde la entrada de imágenes hasta el cálculo automático del porcentaje de materiales presentes en un camión, basándose en las máscaras generadas mediante segmentación semántica.

		May			June			July			August			September			Octobre			
Steps	Deliverable	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
1. Revue de littérature	Report that include the SOTA method for SSL image segmentation using multilabel annotation																			
2. Définition des hypothèses à tester	Report on: What hypothesis are to be tested, why were these hypothesis selected, and how can they implemented to our usecase.																			
3. Méthodologies	Diagram to visualize the different methodologies + planning for test and implementation																			
4. Dataset disponible - EDA	A dataset(s) - report changes made to the original dataset (if necessary)																			
5. Implementation entraînement + Expérimentations	.py scripts incorporating the training loop plugged in our picsellia infrastructure																			
6. Evaluation des méthodes (Performances, comparaison, ...)	a comparative Study of the different methodologies tested as well as the current baseline																			
7. Inclusion du code d'entraînement + test pour CV-inference-pipeline	.py scripts																			
8. Refacto, et implementation des codes pour la prod + test	.py scripts																			

Tabla 1.1: Planificación del proyecto

1.5 Objetivos

1.5.1 General

Evaluar la evolución del sistema “RecoMat” desde un modelo de clasificación basado en parches hacia un enfoque de segmentación semántica débilmente supervisada (WSSS), con el fin de estimar de manera más precisa la distribución espacial de los materiales presentes en imágenes de camiones y generar un mayor valor añadido para los clientes de Alta Road.

1.5.2 Específicos

- **Marco técnico identificado:** Contar con una revisión de métodos SOTA de WSSS y modelos fundacionales aplicables al problema industrial de RecoMat.
- **Viabilidad técnica establecida:** Determinar si modelos como CLIP, DINO y SAM permiten producir máscaras iniciales a partir de etiquetas globales de imagen.
- **Datos preparados para experimentación:** Obtener un conjunto de datos depurado y más equilibrado, reduciendo duplicados, valores atípicos y desequilibrios de clase que puedan afectar el entrenamiento.

- **Prototipo adaptado:** Disponer de una adaptación funcional de WeCLIP al conjunto de datos de Alta Road, integrando generación de pseudo-etiquetas y entrenamiento de un decodificador liviano.
- **Prompts optimizados:** Identificar formulaciones textuales para CLIP que activen mejor las regiones de interés asociadas a cada material.
- **Evaluación cuantitativa y cualitativa:** Obtener resultados medibles mediante mIoU y análisis visual sobre un conjunto de evaluación anotado manualmente.

1.6 Estructura de la Memoria

El documento se organiza en cinco partes principales. En primer lugar, se presenta el contexto de Alta Road, el funcionamiento de RecoMat, el problema industrial y los objetivos del trabajo. En segundo lugar, se desarrolla el estado del arte sobre segmentación semántica débilmente supervisada y se comparan las alternativas metodológicas consideradas.

Luego, se describe el desarrollo de la solución, incluyendo la exploración de datos, las pruebas con distintos métodos y la adaptación de WeCLIP al caso de estudio. Posteriormente, el capítulo de validación analiza los resultados obtenidos, relacionándolos con la hipótesis y los objetivos planteados. Finalmente, se presenta una evaluación crítica del trabajo realizado y una conclusión que sintetiza logros, limitaciones, aprendizajes y líneas futuras.

2 Estado del arte

Este proyecto busca mejorar el reconocimiento visual de residuos de construcción en contenedores de camiones utilizando datos existentes anotados a nivel de imagen. El objetivo es aprovechar los datos ya recopilados por el sistema, extrayendo automáticamente máscaras semánticas a nivel de píxel para cada tipo de residuo, sin necesidad de realizar una anotación manual completa.

La anotación manual píxel a píxel es una tarea costosa y que requiere mucho tiempo, por lo que suele representar un cuello de botella en los proyectos de inteligencia artificial. Actualmente, se están explorando numerosas técnicas para optimizar este proceso. Entre ellas se encuentra el aprendizaje semi supervisado (SSL), que permite utilizar un pequeño conjunto de datos anotados junto con una gran cantidad de datos sin anotar para realizar tareas como clasificación o segmentación.

En este contexto, a partir de los parches de imagen ya anotados a nivel de imagen, el objetivo es entrenar un modelo capaz de predecir máscaras de segmentación a nivel de píxel mediante la combinación de técnicas de segmentación semántica débilmente supervisada (WSSS).

2.1 Soluciones existentes

Dado un conjunto de datos de imágenes de residuos de construcción con:

- Anotación de imágenes con múltiples etiquetas.
- Sin anotaciones a nivel de píxel.
- Parches extraídos de imágenes de cajas de camiones.

El objetivo es desarrollar un enfoque WSSS que genere máscaras de segmentación semántica precisas para diferentes tipos de residuos de construcción. Esto implica implementar el modelo en la canalización existente tras probar y demostrar su eficacia.

2.1.1 Segmentación semántica débilmente supervisada (WSSS)

WSSS [1] [9] [16] [17] es una técnica que busca entrenar modelos de segmentación únicamente a partir de etiquetas de imagen, sin anotaciones precisas a nivel de píxel. En otras palabras, en lugar de usar máscaras que indiquen con precisión la posición de cada objeto en una imagen, el modelo se entrena con información que simplemente indica qué clases están presentes en la imagen, sin una ubicación precisa.

La idea principal de WSSS es generar pseudo etiquetas a partir de índices derivados de la clasificación de imágenes, como los Mapas de Activación de Clases (CAM). Estos CAM resaltan las regiones de la imagen que más influyen en la decisión de clasificación. Posteriormente, se refinan y se utilizan como anotaciones aproximadas para entrenar un modelo de segmentación semántica.

2.1.2 Evolución histórica

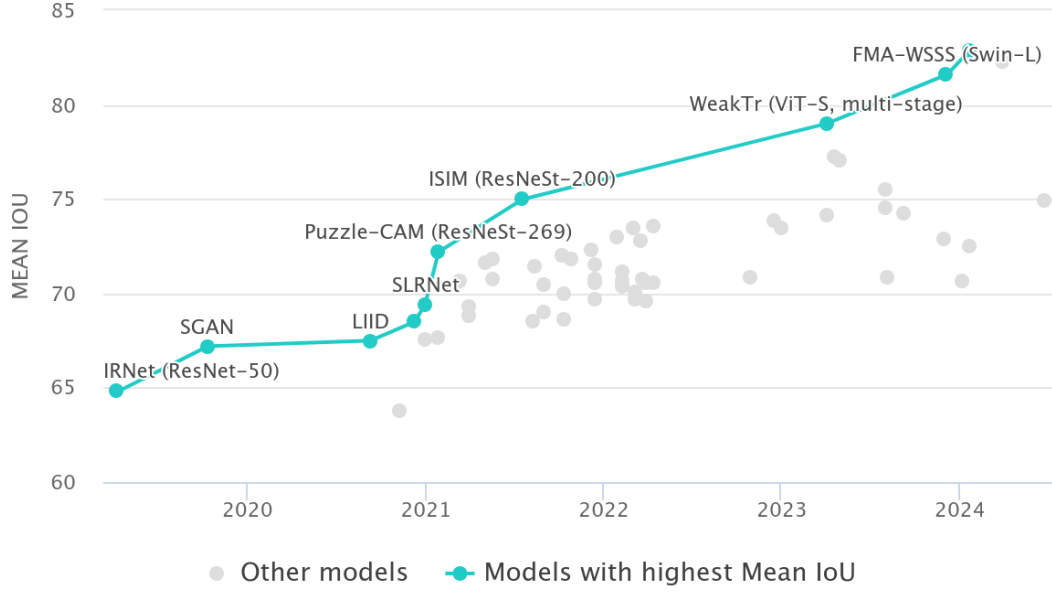


Figura 2.1: Clasificación del conjunto de datos de prueba PASCAL VOC 2012. Del papers with code sobre WSSS.

Los primeros enfoques de segmentación semántica débilmente supervisada (WSSS) surgieron alrededor de 2014-2015, con la idea de utilizar etiquetas de imagen para guiar el aprendizaje de modelos de segmentación [16]. Estos métodos empleaban principalmente redes neuronales convolucionales combinadas con técnicas de agregación, relacionando la presencia de clases en una imagen con regiones localizadas.

Los mapas de activación de clases (CAM) se consolidaron rápidamente como un método de referencia para generar mapas aproximados de localización. Entre 2016 y 2019, gran parte de la investigación se centró en superar las limitaciones de estos mapas mediante el refinamiento de la calidad de los contornos y la mejora de la cobertura de los objetos [1]. Para ello, se introdujeron estrategias de regularización, aprendizaje progresivo e integración con el fin de aumentar la precisión de los mapas generados.

A partir de 2020, los métodos se volvieron más sofisticados, combinando con frecuencia varias tareas o módulos, como redes codificador-decodificador, transformadores [8] y modelos fundacionales. Estos enfoques han permitido mejorar progresivamente la capacidad de los modelos para producir pseudo máscaras de mayor calidad, a pesar de mantener un esquema de supervisión débil.

2.2 Herramientas y Métodos

A pesar de estos rápidos avances, el campo aún enfrenta varios desafíos, incluyendo limitaciones relacionadas con los escenarios de los conjuntos de datos, el gran tamaño de algunos modelos y tiempos de entrenamiento e inferencia lentos, que dificultan las aplicaciones en tiempo real. No

obstante, la brecha de rendimiento entre los métodos con supervisión flexible y los totalmente supervisados se ha reducido considerablemente.

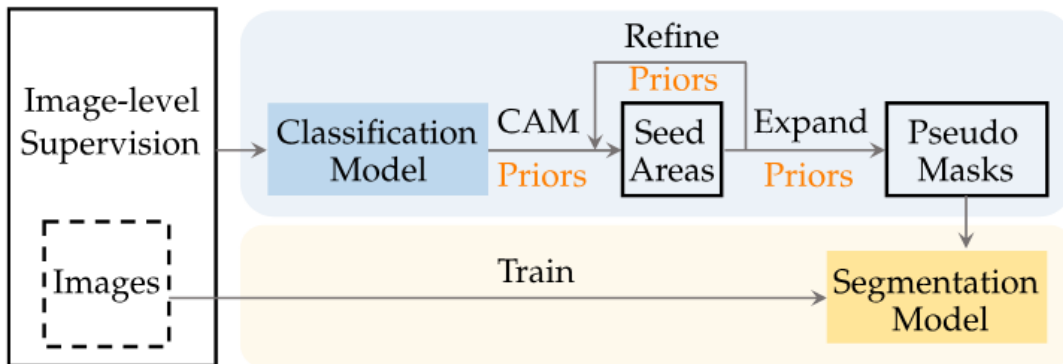


Figura 2.2: Flujo de trabajo típico para publicaciones sobre segmentación semántica débilmente supervisada (WSSS). [1]

La mayoría de los métodos WSSS basados en anotaciones a nivel de imagen siguen un proceso de tres pasos [7], diseñado para acortar la distancia entre las anotaciones débiles y las predicciones densas a nivel de píxel. Este proceso suele constar de tres fases principales, como se ilustra en la Figura 11:

- **Modelo de clasificación e índices iniciales:** Una red de clasificación de imágenes se entrena utilizando etiquetas a nivel de imagen. Este modelo se utiliza para generar índices de ubicación iniciales, generalmente CAM. Estos CAM resaltan las regiones relevantes de la imagen y sirven como punto de partida para generar pseudo etiquetas.
- **Refinamiento de pseudo etiquetas:** Las pseudo etiquetas son, por naturaleza, ruidosas e imprecisas. Por lo tanto, se aplican diversas técnicas de refinamiento para mejorar su calidad, con el fin de producir máscaras de pseudo píxeles más completas y precisas.
- **Entrenamiento de redes de segmentación:** Finalmente, se entrena una red de segmentación semántica totalmente supervisada (por ejemplo, DeepLabv3 o U-Net) utilizando estas máscaras refinadas como si fueran anotaciones pixeladas reales.

Algunas investigaciones también exploran métodos de un solo paso diseñados para reducir la complejidad del modelo y el número de parámetros. Sin embargo, actualmente, el mejor rendimiento se logra generalmente utilizando pipelines progresivos, a pesar del problema inherente de la propagación de errores entre pasos.

Debido a este riesgo de errores en cascada, los esfuerzos de investigación y documentación se centran principalmente en las primeras etapas del proceso (generación y refinamiento del índice inicial) en lugar de en el entrenamiento del modelo de segmentación en sí.

2.2.1 Metodologías para la generación de CAM

Las CAM [3] desempeñan un papel esencial en la segmentación semántica débilmente supervisada (WSSS), ya que proporcionan una localización inicial de los objetos a partir de las etiquetas de las imágenes. Derivadas de redes de clasificación, las CAM se obtienen ponderando los mapas de características de la última capa convolucional con pesos específicos de clase provenientes de la capa de clasificación, generalmente después de una operación de Global Average Pooling (GAP).

El mapa de calor resultante resalta las regiones de la imagen más relevantes para la decisión de clasificación de la red, funcionando así como una máscara de segmentación rudimentaria que permite relacionar las etiquetas globales de la imagen con detalles a nivel de píxel.

Un concepto importante a tener en cuenta con respecto a las CAM es la “maldición de la discriminabilidad”. Las CNN tienden a buscar la evidencia más óptima para clasificar una imagen (por ejemplo, el pico de un ave en lugar de todo su cuerpo), lo que lleva a las CAM a resaltar solo ciertas partes discriminantes. La segmentación semántica, por otro lado, requiere una cobertura completa del objeto, lo cual plantea un problema.

Esto implica que mejorar los modelos de clasificación no garantiza necesariamente mejores CAM. Por ello, se están realizando esfuerzos para ampliar los CAM para que cubran una mayor porción de la imagen.

2.2.2 Marcos metodológicos que implican la mejora de las CAM

Se han propuesto varios métodos para mejorar la calidad de los mapas de activación en el contexto de la segmentación débilmente supervisada (WSSS). IRNet [2] introduce una red de relaciones entre píxeles para modelar dependencias locales entre píxeles, lo que resulta en mapas más consistentes y conectados. SEAM [12] adopta una estrategia auto supervisada basada en un mecanismo de atención equivariante, lo que garantiza la estabilidad de mapas ante transformaciones geométricas. Puzzle-CAM [6] incorpora una tarea auxiliar de reconstrucción de rompecabezas, que obliga al modelo a aprender relaciones espaciales globales. Otros enfoques, como ISIM [15] adopta un marco iterativo que combina clasificación y segmentación dentro de un modelo codificador-decodificador.

Recientemente, han surgido métodos que aprovechan los modelos fundacionales. Por ejemplo, SAM-WSSS [18] y S2C [11] utilizan SAM para mejorar las pseudo etiquetas generadas a partir de CAM, aprovechando la precisión de borde proporcionada por SAM. WSSS-SAM [10], diseñado exclusivamente para inferencia, genera pseudo etiquetas de alta calidad combinando SAM con DINO, como Grounding-DINO. WeakTr [19] basado en una arquitectura transformers, ofrece una estimación ponderada de la importancia de los cabezales de atención ViT y fusiona de forma adaptativa sus mapas para generar CAM más completos. FMA-WSSS [14] y SemPLes [13] explotan respectivamente CLIP [23] y SAM, así como prompts contrastivos y estrategias de regularización semántica, para alinear con mayor precisión las regiones segmentadas con las categorías objetivas, mejorando así la calidad de las pseudo máscaras. Finalmente, WeCLIP [20] utiliza un modelo CLIP y DINO fijó como base, acoplado con un decodificador liviano, para segmentar directamente objetos usando etiquetas a nivel de imagen.

2.3 Discusión

Tras revisar varios artículos científicos y repositorios de GitHub, estos candidatos fueron seleccionados en función de su viabilidad, accesibilidad del código, metodología de evaluación, retroalimentación de la comunidad y complejidad técnica. El objetivo es adaptar las metodologías existentes, no desarrollar nuevas.

2.3.1 Puesta a tierra del SAM

El uso de modelos fundacionales parece ser uno de los marcos más eficaces para esta tarea, especialmente combinando dos modelos fundamentales: DINO y SAM.

SAM puede ofrecer un rendimiento competitivo sin necesidad de ajustes, mientras que los métodos WSSS existentes suelen requerir engorrosos pasos de reentrenamiento para la clasificación y la generación de pseudo etiquetas. Esto representa una ventaja significativa sobre otros enfoques, que implican ciclos de entrenamiento complejos.

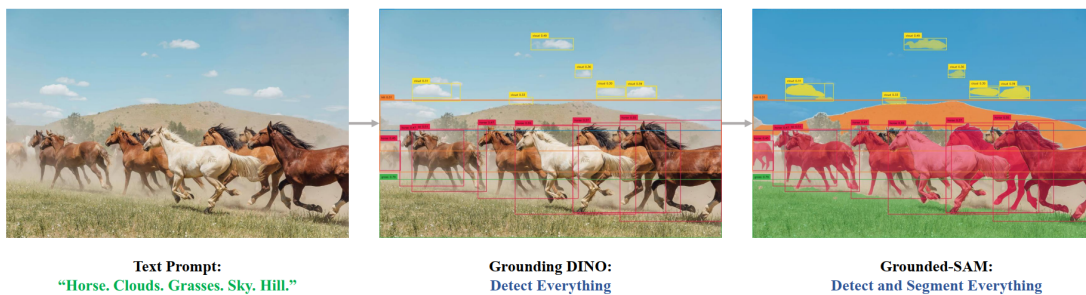


Figura 2.3: Flujo de trabajo de GroundingSAM

Como se ilustra en la Figura 12, Grounding SAM permite usar DINO como generador de bounding boxes simplemente proporcionando las clases presentes en la imagen como texto. Una vez generados estos cuadros delimitadores con sus etiquetas, se pasan a SAM como input para generar una pseudo máscara junto con la etiqueta del cuadro. Esto genera directamente una máscara de segmentación anotada.

Este método logra excelentes resultados en el conjunto de datos PASCAL VOC 2012, pero debido a su naturaleza como modelo fundacional pre entrenado en volúmenes muy grandes de datos, no está incluido en los puntos de referencia oficiales de PASCAL para WSSS.

A modo de indicación, este método alcanza el 77,1 % mIoU en el conjunto de prueba de máscara PASCAL, superando el rendimiento de los métodos anteriores.

2.3.2 Text Region

Text Region utiliza CLIP y SAM para conectar la comprensión espacial y semántica. Esto es posible gracias a la combinación de las capacidades de segmentación de SAM con la potente

correspondencia de texto a imagen de CLIP, a menudo sin necesidad de entrenamiento o con plug-and-play, es decir, sin reentrenar los modelos subyacentes.

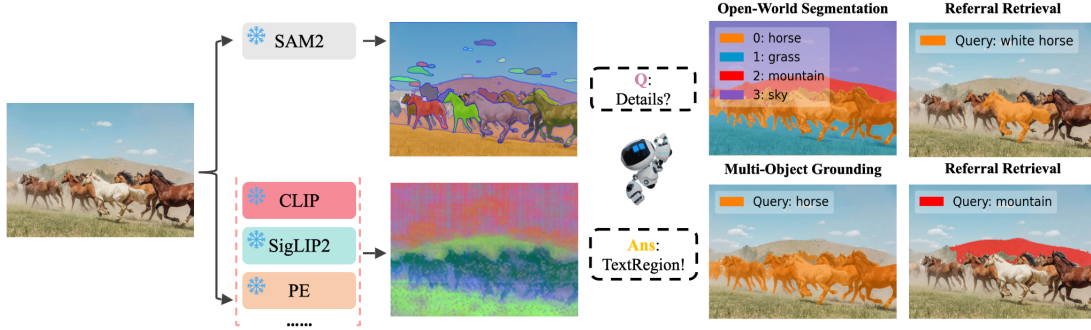


Figura 2.4: Flujo de trabajo de la región de texto

2.3.3 WeCLIP

Este modelo busca realizar la segmentación utilizando los codificadores fijos de imagen y texto de CLIP como base de un flujo de trabajo de una sola etapa. Las características semánticas se extraen mediante el codificador de imagen, mientras que las clases son procesadas por el codificador de texto para generar representaciones textuales. Solo se entrenan el decodificador y el módulo de refinamiento.

La pérdida total del modelo se define en la Ecuación (1). Equilibra la tarea principal de segmentación con la tarea de refinamiento de la atención. El segundo término es la pérdida de atención, que garantiza que el mapa de atención refinado A_{cd} corresponde a un objetivo $\hat{A}^+$. λ es un parámetro de ponderación.

$$L = L_{\text{Seg}} + \lambda \cdot L_{\text{ce}}(A_{cd}, \hat{A}^+) \quad (1)$$

La pérdida de segmentación principal se define en la Ecuación (2). Combina las pérdidas de segmentación clásicas con un componente de supervisión cruzada entre las ramas CLIP y DINO. Los dos primeros términos comparan la predicción final con la pseudo-etiqueta final usando entropía cruzada y el dice score. El tercer término es la pérdida de supervisión cruzada, donde la predicción de la rama CLIP está supervisada por la pseudo etiqueta de la rama DINO y viceversa. β es un parámetro de ponderación.

$$L_{\text{seg}} = L_{\text{ce}}(P, M_p) + L_{\text{dice}}(P, M_p) + \beta (L_{\text{ce}}(P_{\text{clip}}, M_{\text{dino}}) + L_{\text{ce}}(P_{\text{dino}}, M_{\text{clip}})) \quad (2)$$

Lo más destacable es la velocidad del modelo en comparación con otros enfoques. El marco WeCLIP requiere un máximo de 6,2 GB de memoria GPU durante el entrenamiento, significativamente menos que otros métodos, que requieren al menos 12 GB. El entrenamiento duró 4,5 horas, lo que equivale a 30.000 iteraciones, según el artículo.

El código también está bien documentado y los problemas se abordan de forma proactiva en el

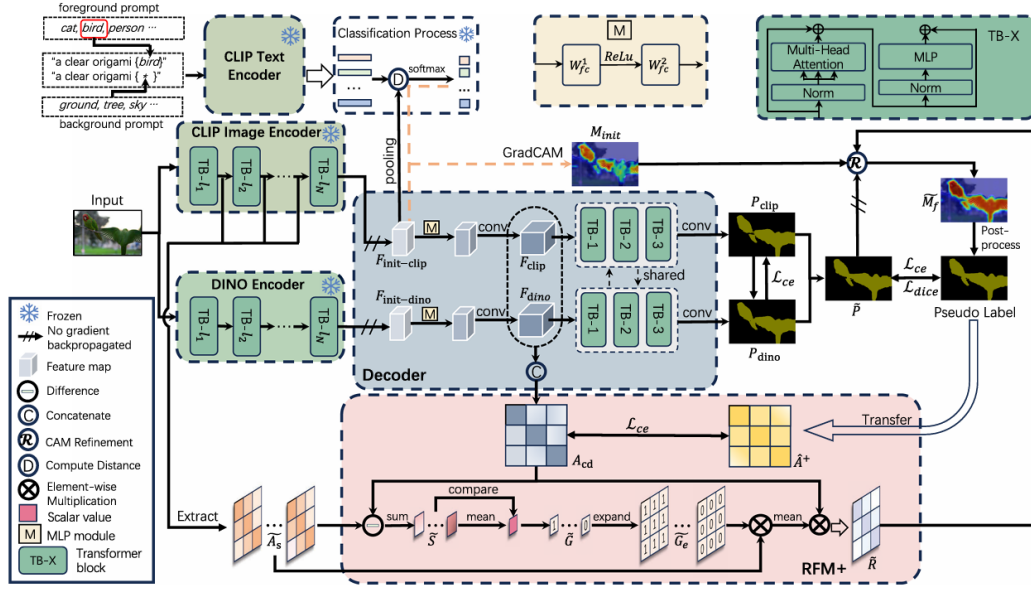


Figura 2.5: El marco WebCLIP. La imagen se introduce en el codificador de imágenes CLIP y en el codificador DINO para generar dos mapas de características de imagen independientes. Las etiquetas de clase se utilizan para construir indicaciones de texto, que luego se introducen en el codificador de texto CLIP para generar las características del texto. Las puntuaciones de clasificación se generan en función de la distancia entre las características de la imagen y las características del texto. Con GradCAM, se genera el CAM inicial M_{init} . A continuación, los mapas de características de la imagen se introducen en el decodificador compartido para generar dos predicciones de segmentación semántica individuales, P_{clip} y P_{dino} , respectivamente. Tras la fusión se genera la predicción final P . Simultáneamente, el mapa de afinidad A_f de salida del decodificador y los mapas de atención de múltiples cabezas A_s son ingresados en el módulo RFM+ para establecer mapas de refinación R , que luego combinan la predicción final P . Después del postprocesamiento, se utilizará como supervisor para entrenar al decodificador.

repositorio de GitHub, lo que contrasta con otros proyectos donde muchas preguntas siguen sin respuesta. Finalmente, el modelo alcanza un rendimiento excepcional, con una puntuación mIoU del 75,2 % en el conjunto de pruebas VOC 2012.

2.3.4 Síntesis de la discusión

Las soluciones revisadas muestran que los modelos fundacionales ofrecen una oportunidad real para abordar la segmentación semántica con supervisión débil, pero también presentan limitaciones importantes cuando se trasladan a un contexto industrial específico. Grounding SAM y Text Region son atractivos por su capacidad de inferencia sin reentrenamiento, aunque dependen fuertemente de la calidad de los *prompts* y pueden producir detecciones demasiado generales cuando los materiales no corresponden a objetos bien delimitados. DINOv2 y FeatUp permiten obtener particiones visuales detalladas, pero requieren una etapa adicional para asignar significado semántico a cada clúster, lo que complica su integración directa.

En este contexto, WeCLIP fue seleccionado como enfoque principal porque ofrece un equilibrio

adecuado entre viabilidad técnica, rendimiento reportado, consumo de recursos y compatibilidad con las etiquetas disponibles en RecoMat. A diferencia de métodos que exigen máscaras manuales o una cadena de reentrenamiento más pesada, WeCLIP permite aprovechar anotaciones a nivel de imagen y combinar información visual y textual mediante CLIP y DINO.

La contribución de esta memoria consiste en estudiar cómo estos métodos pueden adaptarse a un problema industrial concreto: la identificación de materiales transportados en camiones a partir de datos reales, desbalanceados y con anotaciones débiles. Por ello, el aporte no se limita a reproducir un método académico, sino que incluye la preparación del conjunto de datos, la evaluación de alternativas, la adaptación del flujo de entrenamiento, la construcción de un conjunto de evaluación manual y el análisis de las condiciones necesarias para una futura integración en RecoMat.

3 Desarrollo de la solución

3.1 Exploración de datos

Fue necesario un análisis inicial de datos para comprender plenamente el proyecto. Para este trabajo, se proporcionó un conjunto de datos de imágenes anotadas. Un estudio preliminar reveló un desequilibrio entre las clases y el número de clases por imagen. La Figura 15 ilustra estos desequilibrios.

La gran mayoría de imágenes corresponden a imágenes que contienen un solo material, y las clases mayoritarias son generalmente Tarp y Empty. Excluyendo estas dos clases, que corresponden principalmente a imágenes de una sola etiqueta, observamos casi la misma distribución del número de clases por imagen.

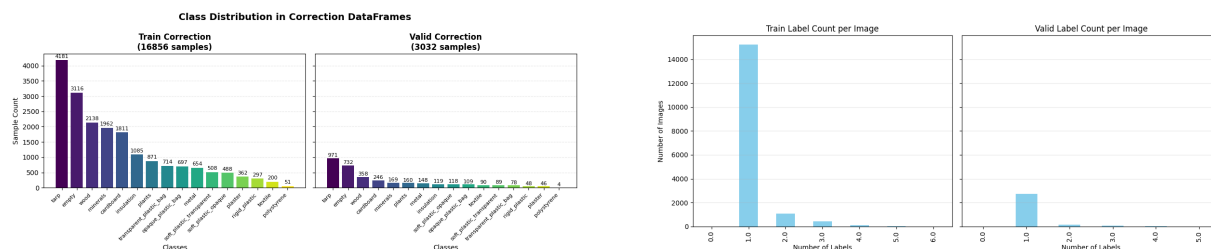


Figura 3.1: Distribución de clases y número de etiquetas por imagen

Se realizó una revisión imagen por imagen para identificar los diferentes tipos de imágenes presentes en el conjunto de datos. Este análisis proporciona una indicación inicial de la necesidad de equilibrar y depurar los datos. La Tabla 4 presenta algunos ejemplos representativos identificados.

Para la fase de limpieza, primero se utilizó CleanLab para detectar y filtrar imágenes duplicadas o similares, problemas de anotación y valores atípicos.

Para la identificación de errores de anotación, se explotaron las características (*features*) y las probabilidades de predicción (*pred proba*) procedentes del modelo de clasificación presentado anteriormente. Sin embargo, tras el análisis, se observó que la mayoría de los casos señalados no corresponden realmente a errores de anotación. La confusión provenía principalmente de la sustitución de la etiqueta *textile* por *soft_plastic_transparent* en situaciones donde ello no estaba justificado.

Paralelamente, se aplicó un *hashing* a las imágenes con el fin de garantizar la eliminación de duplicados. También se implementó un detector de imágenes borrosas (*blurry*), pero los resultados no aportaron información significativa y esta línea de trabajo no se profundizó.

3.2 CAM

La extracción de las CAMs se realizó principalmente para analizar qué es lo que el modelo de clasificación “observa” y para probar diferentes métodos sobre nuestras imágenes, con el fin de evaluar la viabilidad de una aplicación de PuzzleCAM o IRNet.

Tabla 3.1: Ejemplos de imágenes representativas del conjunto de datos

	Parche completo La imagen completa debe considerarse como una máscara.		Parche fronterizo Debe estar presente una máscara de fondo.
rigid_plastic 	Etiqueta mono falsa Imagen etiquetada incorrectamente como etiqueta única.	plants + plaster 	Falso Bilabel La segunda clase no coincide.
plaster + wood 	Parche compartido Imagen etiquetada correctamente como multiclase.	insulation rigid_plastic soft_plastic_opaque soft_plastic_transparent textile wood 	Sobre Etiquetado Las clases son difíciles de identificar, incluso para la anotación humana.

Para las imágenes monolabel, LayerCAM [5] parece ser el método más adecuado. En efecto, permite una expansión más amplia sobre la imagen, capturando una mayor cantidad de elementos y diferenciando de manera más clara los objetos del fondo.

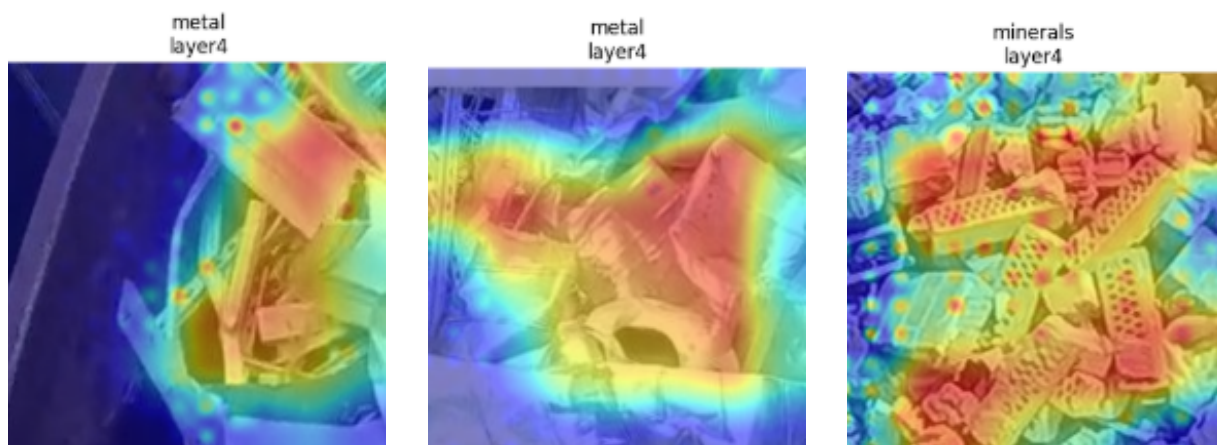


Figura 3.2: Ejemplos de LayerCAM aplicados a imágenes de etiqueta única

Para imágenes bi-etiquetadas, la situación es más compleja. Los métodos ScoreCAM y GradCAM parecen más adecuados, ya que ofrecen mayor precisión al diferenciar componentes. Sin embargo, GradCAM, si bien permite separar clases, a veces presenta dificultades: la activación del CAM para la segunda clase puede solapar parcialmente la primera, lo que limita significativamente la distinción entre las clases.

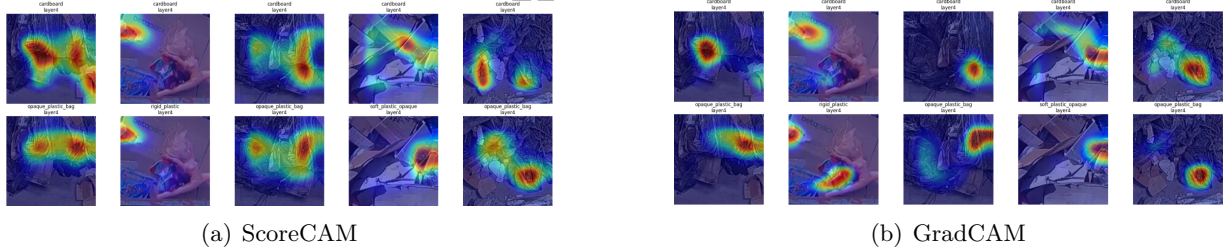


Figura 3.3: Ejemplos de CAM aplicados a imágenes bi-etiquetadas

3.3 Grounding SAM

Los resultados siguen siendo mitigados. Las bounding boxes tienden a ser amplias y a abarcar la totalidad del patch, lo que no permite distinguir claramente la máscara de segmentación producida por SAM. Además, la detección de materiales sigue siendo un desafío importante, especialmente durante las primeras etapas del procesamiento.

Por el contrario, para las imágenes situadas en los bordes, la detección de un objeto monolabel frecuente resulta fiable. Asimismo, también se garantiza la localización precisa de un objeto común dentro de la imagen.

3.4 DINOv2 + Featup

La creación de clústeres de *features* a partir de DINO es similar a la segmentación propuesta por SAM. El problema es que el número de clústeres no es conocido (entre 2 y n) y que la segmentación está fuertemente influenciada por las formas o los colores, como se muestra en la Figura 19. Utilizando FeatUp [22], es posible obtener finalmente máscaras de segmentación detalladas.

Este método genera máscaras candidatas que posteriormente se clasifican. Una técnica para extraer las máscaras obtenidas mediante este proceso, seguida de su clasificación, podría constituir una solución al problema. Sin embargo, esta es una vía a explorar en el futuro. En el marco de este proyecto, este enfoque no se exploró por limitaciones de tiempo.

3.5 Text Region

Los resultados son prometedores, pero todavía se producen fallos cuando las indicaciones son inadecuadas o las imágenes son demasiado complejas.

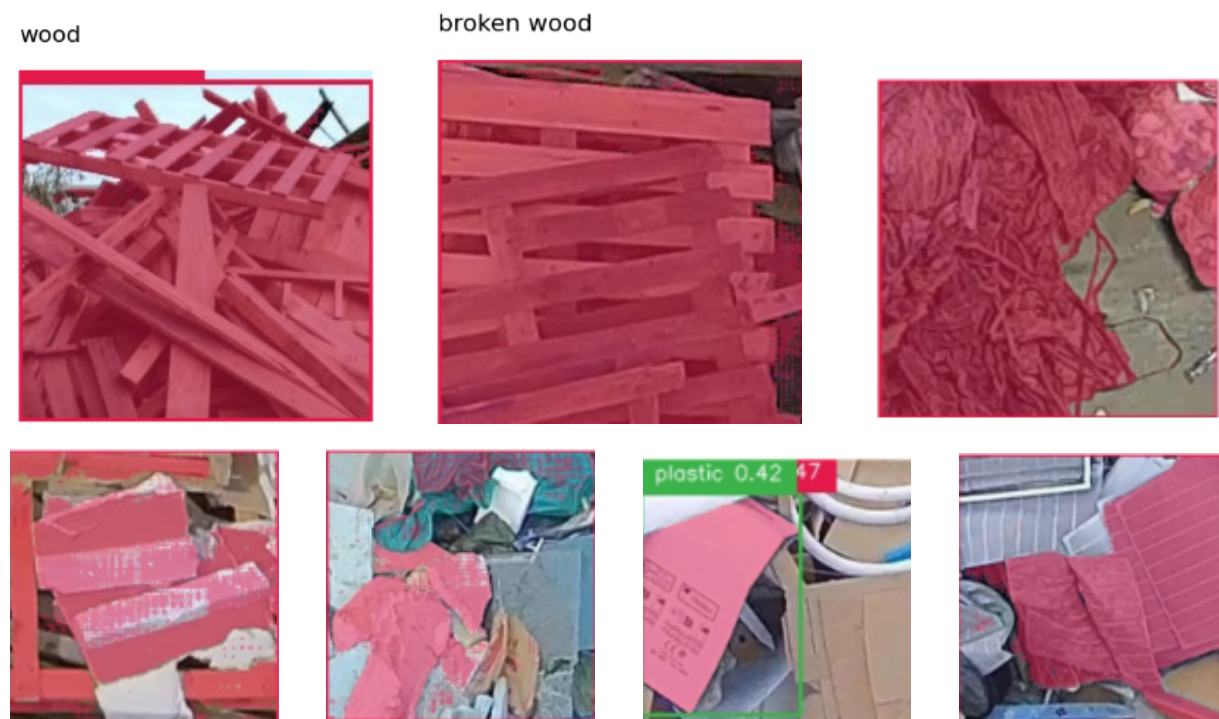


Figura 3.4: En la parte superior, ejemplos de detecciones correctas; en la parte inferior, fallos de detección por parte del modelo.



Figura 3.5: Imágenes superiores de una sola etiqueta; imágenes inferiores de dos etiquetas.

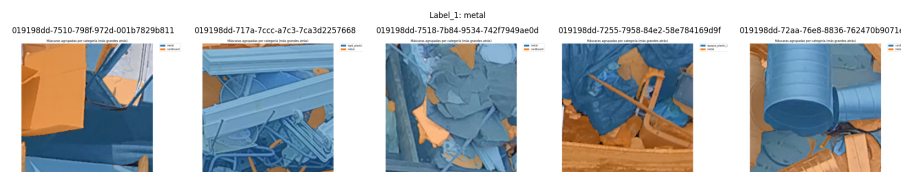


Figura 3.6: Resultados donde se confunden algunos materiales

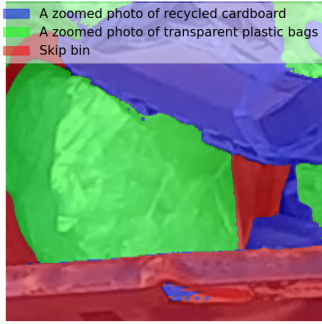
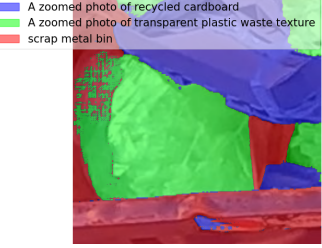
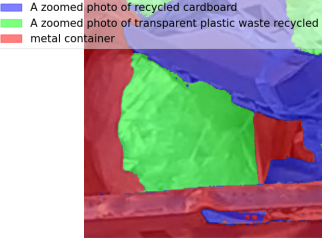
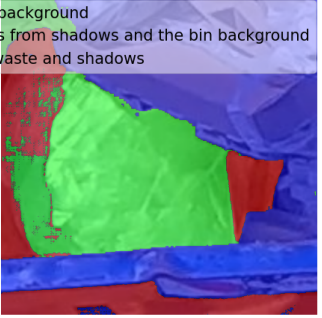
Sin embargo, cuando los elementos son claramente distintos o frecuentes, el método funciona con gran precisión.



Figura 3.7: Caso en el que la detección funciona bien

El texto de indicación sí indujo variaciones en la detección. La Tabla 5 presenta la misma imagen analizada con diferentes textos de indicación.

Tabla 3.2: Diferentes resultados según los textos de indicación utilizados

	label: ['A zoomed photo of recycled cardboard', 'A zoomed photo of transparent plastic waste recycled', 'metal container']		label: ['A zoomed photo of recycled cardboard', 'A zoomed photo of transparent plastic waste texture', 'scrap metal bin']
	etiqueta: ['Una foto ampliada de cartón reciclado', 'Una fotografía ampliada de residuos plásticos transparentes reciclados', 'contenedor de metal']		label: ['A photo taken from above of a metal bin in a truck, showing recycled cardboard patches with clear edges, distinguishing shadows and background', 'An overhead view of a metal bin in a truck...']

Este método permanece en espera de ser utilizado como segunda opinión; su precisión es bastante alta, como se aprecia en algunas imágenes. Sin embargo, aún queda trabajo por hacer en el texto de la propuesta.

3.6 WeCLIP+

Tras esta exploración y teniendo en cuenta el conjunto de observaciones realizadas, el enfoque elegido fue apoyarse en modelos fundamentales como CLIP para establecer un vínculo entre las etiquetas (*labels*) y las CAMs, en lugar de utilizar exclusivamente el modelo de clasificación previamente entrenado. Esta decisión se explica principalmente por la capacidad de CLIP para adaptar dinámicamente los *text prompts* y alinear mejor la definición de las clases con el contenido de las imágenes. En comparación con las CAMs del modelo de clasificación, que lograba marcar las zonas correspondientes, pero tendía a mezclar las regiones cuando se trataba de imágenes bilabel o con más etiquetas. Esto se debe a la naturaleza de la activación por competencia: “si hay cartón, es probable que también haya bolsas de plástico”.

La solución propuesta consiste en utilizar WebCLIP, que combina la generación de CAMs mediante CLIP con la producción de pseudo etiquetas (*pseudo-labels*) a través de DINOv2. Este proceso iterativo mejora progresivamente las CAMs gracias a los *attention heads* de ambos modelos fundamentales, que posteriormente se utilizan como verdad de terreno (*ground truth*) para entrenar un decodificador encargado de producir las máscaras finales. Como se mencionó

anteriormente, este modelo presenta la ventaja de requerir menos recursos GPU y, al mismo tiempo, adaptarse bien a nuestro contexto.

El trabajo actual consiste en adaptar el código a un conjunto de datos personalizado (*custom dataset*) y obtener los primeros resultados.

3.7 Limpieza de datos

Se creó una versión limpia y aumentada del conjunto de datos para WebCLIP. El proceso implicó el uso de CleanLab para una limpieza rápida de imágenes. Esto permitió identificar valores atípicos e imágenes similares mediante el modelo de clasificación para extraer características. Posteriormente, se implementó el submuestreo y el sobremuestreo de clases en el primer paso para equilibrar las clases. La distribución resultante se muestra en la Figura 23.

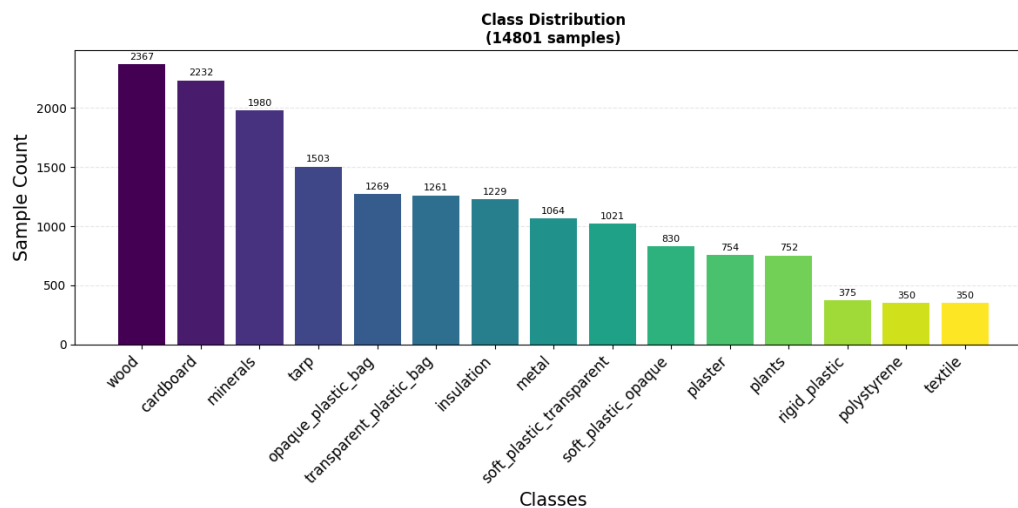


Figura 3.8: Distribución de clases para el conjunto de datos de entrenamiento limpio

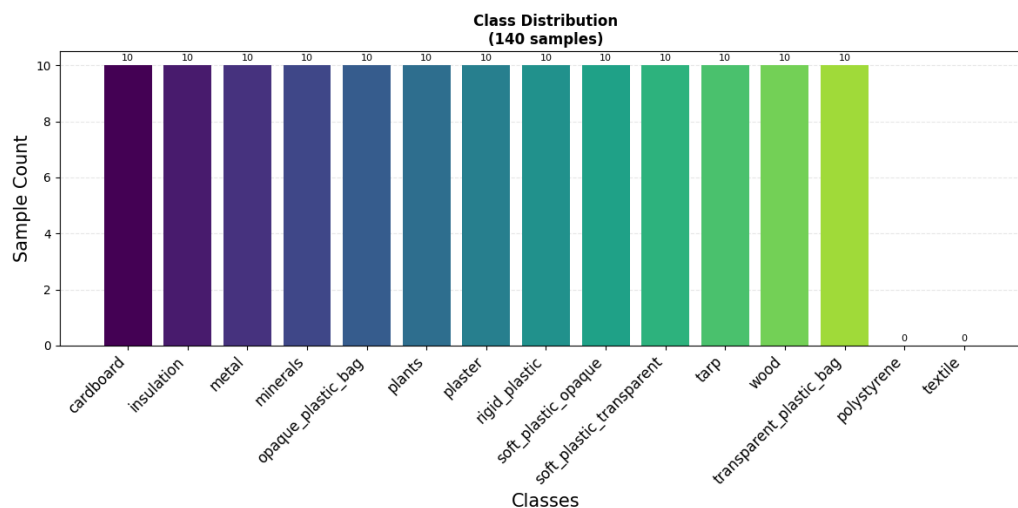


Figura 3.9: Distribución de clases para el conjunto de datos de prueba de etiqueta única.

Paralelamente, para obtener resultados, se decidió crear un conjunto de datos anotados manualmente. Para ello, se seleccionaron 10 imágenes de cada clase de etiqueta única y se generaron sus respectivas máscaras (excepto para el poliestireno y los textiles, ya que no tenían etiquetas únicas). Este conjunto de datos se denominará *monolabel_evaluation*.

3.8 Entrenamiento

Una vez adaptado el código fuente a nuestro proyecto, se realizaron cuatro entrenamientos diferentes. Una primera versión utilizando el conjunto completo de datos de entrenamiento. Una segunda versión sin duplicados y fusionando la clase *empty* con *background*. Una tercera versión añadía a la segunda la limpieza realizada mediante CleanLab, así como el *upsampling* y el *downsampling* de los datos (Figura 25). Finalmente, una cuarta versión en la que se modificaron los *text prompts*.

El modelo permite ajustar ciertos hiperparámetros. Una parte importante de estos corresponde a las aumentaciones (*augmentations*) *on-the-fly*. Otros hiperparámetros, como el optimizador, el *scheduler* y el número de épocas, se mantienen con los valores predeterminados.

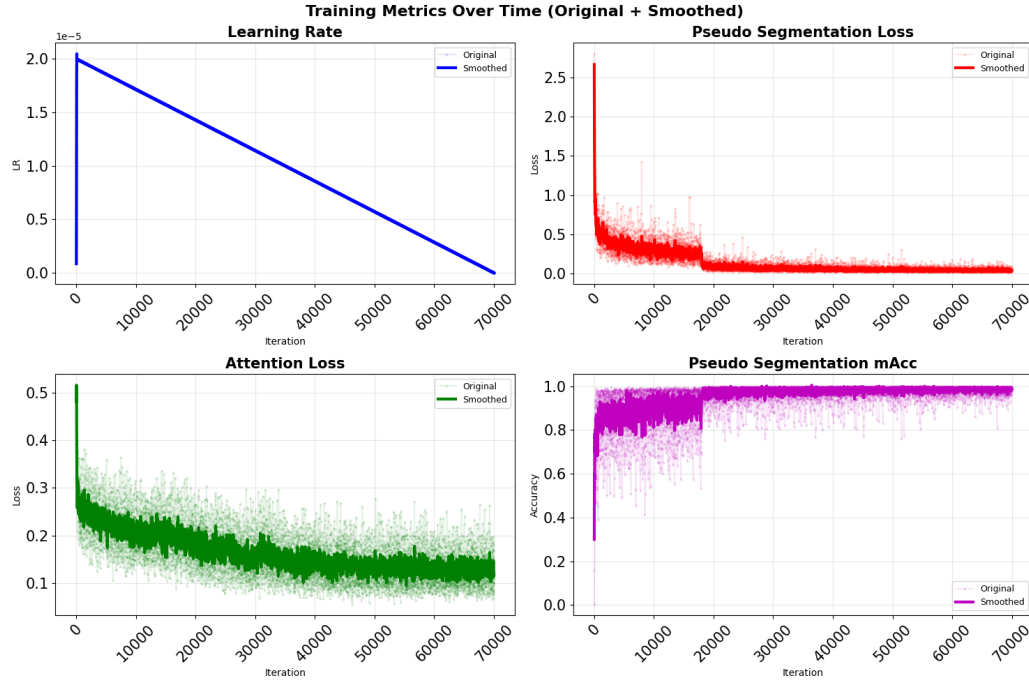


Figura 3.10: Curvas de pérdida del tren versión 3

Pérdida de pseudo segmentación: Esta es la función de pérdida principal utilizada para entrenar el decodificador de segmentación. Mide la diferencia entre la predicción final del decodificador y la pseudo etiqueta en línea generada dinámicamente.

Pérdida de atención: El RFM funciona creando A_f , que ayuda a refinar la CAM inicial bruta. La pérdida de atención (*attention loss*) compara este mapa de afinidad generado con el mapa de afinidad de referencia $\hat{A}^+$ derivado de las pseudo etiquetas (*pseudo-labels*).

Pseudo Segmentación mAcc: Durante el entrenamiento, esta métrica calcula la precisión promedio de las predicciones del decodificador en relación con las pseudo etiquetas en línea para esa misma iteración.

4 Validación

La validación de este trabajo se organiza alrededor de la hipótesis planteada: comprobar si es posible reutilizar anotaciones a nivel de imagen para obtener máscaras semánticas útiles en el contexto de RecoMat. Para ello, se conectaron los objetivos específicos con tres tipos de evidencia. En primer lugar, la evidencia técnica, correspondiente a la adaptación funcional de WeCLIP y a la generación de predicciones de segmentación. En segundo lugar, la evidencia cuantitativa, medida mediante mIoU sobre el conjunto *monolabel_evaluation*. En tercer lugar, la evidencia cualitativa, basada en la inspección visual de máscaras correctas, errores de segmentación y casos de sobreajuste.

La métrica principal utilizada fue mIoU, ya que permite comparar la superposición entre una máscara predicha y una máscara de referencia anotada manualmente. Esta métrica es adecuada para este trabajo porque evalúa directamente el objetivo de pasar de clasificación a segmentación: no basta con identificar la clase presente, sino que también se debe localizar espacialmente la región correspondiente.

Los experimentos se articularon de la siguiente manera: primero se verificó la generación de máscaras mediante WebCLIP; luego se compararon distintos *checkpoints* y versiones del entrenamiento; finalmente, se evaluó el efecto de la optimización de *text prompts*. Esta organización permite discutir no solo si el modelo funciona, sino también bajo qué condiciones mejora o empeora.

4.1 Resultados

Estos primeros resultados son alentadores, aunque aún pueden mejorarse. La optimización de los hiperparámetros todavía no se ha realizado, principalmente debido a una carencia crítica: la ausencia de un conjunto de datos de validación completo. Aunque el modelo permite configurar una métrica de validación, esta etapa se omite actualmente por falta de datos anotados disponibles.

Como puede observarse, en algunos casos WeCLIP incluso detectó materiales que no habían sido anotados durante el entrenamiento. Esto se explica porque, en modo de inferencia, la parte CAM y RFM está desacoplada del módulo principal, lo que equivale a realizar un simple *forwarding* de la imagen, cuyas *features* son extraídas mediante DINO y CLIP antes de ser transmitidas al decodificador, que es el encargado de realizar la segmentación.

Sin embargo, en algunos casos se generaron máscaras incorrectas. La Figura 27 ilustra, a la izquierda, la predicción del modelo después de 30 000 iteraciones y, a la derecha, la obtenida tras 70 000 iteraciones. Se observa que algunas imágenes procedentes del *checkpoint* de 70 000 iteraciones no están completas, a diferencia de las del *checkpoint* de 30 000 iteraciones.

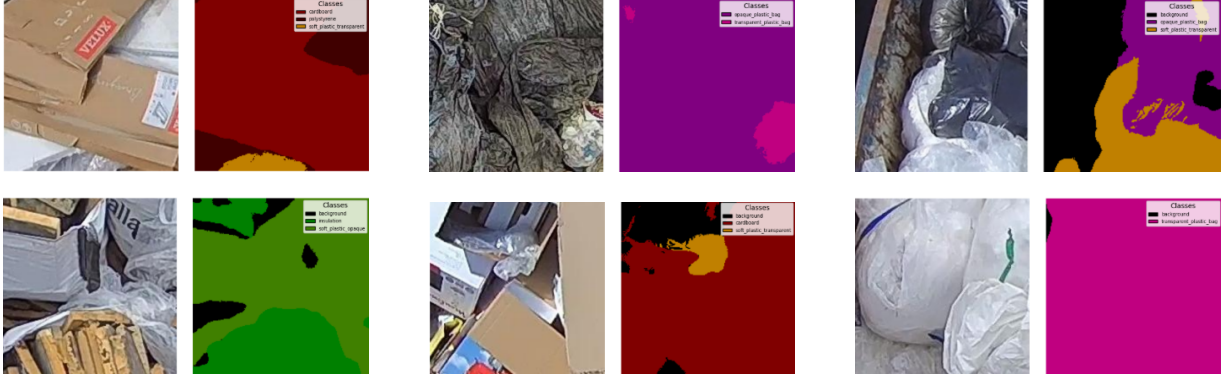


Figura 4.1: Predicciones exitosas de WebCLIP



Figura 4.2: Sobreajuste del modelo que resulta en una reducción de la clase PLASTER; (izquierda) Imagen original, (centro) Predicción de 30k, (derecha) Predicción de 70k

Por otra parte, un factor controlable importante en WeCLIP es la elección de los *text prompts*, que influye fuertemente en los resultados. Una buena selección de estos, adaptada a nuestras imágenes, puede resultar decisiva, tal como se demostró en el experimento sobre *Text Region*. La forma en que los materiales son definidos y aparecen en las imágenes impactan directamente el rendimiento del modelo.

4.2 Optimización

La optimización de los *text prompts* se basa en la búsqueda de la mejor correspondencia entre el texto y la imagen que el modelo CLIP puede interpretar. Para ello, se desarrolló un algoritmo con el fin de evaluar diferentes candidatos de *prompts*. Cada *text prompt* es utilizado inicialmente por el modelo para extraer CAMs iniciales. Estos mapas se comparan posteriormente con las máscaras de verdad de terreno (*GT masks*), lo que permite medir en qué medida el *prompt* activa correctamente los píxeles situados en la región de interés.

A partir de un LLM, se generan 30 candidatos de *text prompts* para cada clase. Posteriormente, para cada uno de estos *prompts*, se extrae la CAM sobre 10 imágenes de la clase correspondiente y se evalúa utilizando la métrica mIoU.

De esta manera, cada *text prompt* obtiene una puntuación, lo que permite determinar cuáles activan mejor los píxeles correspondientes a la región de interés e identificar las formulaciones textuales más eficaces para el modelo.

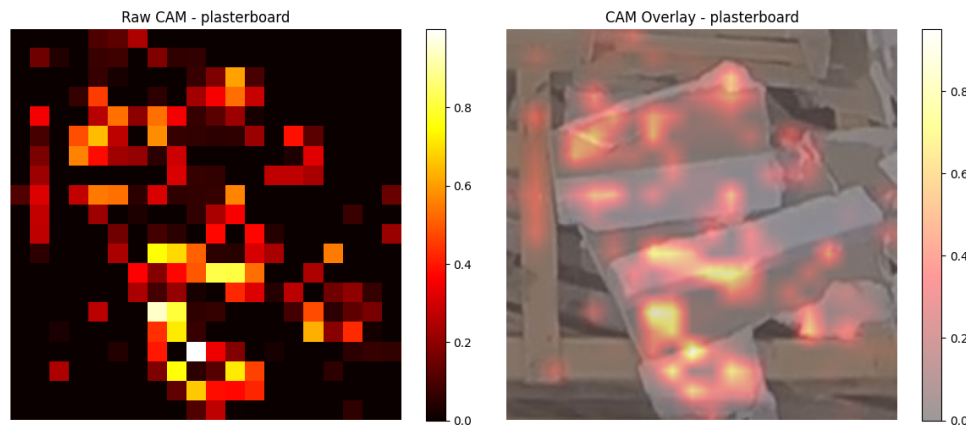


Figura 4.3: Ejemplo de CAM para la clase de PLASTER

Con la ayuda del *dataset* de prueba monolabel, se obtuvo una clasificación de los *text prompts* que debían utilizarse en el entrenamiento de la versión 4. La Tabla 5 presenta los resultados de mIoU para cada clase correspondientes a los dos *checkpoints* v3 y v4.

Tabla 4.1: Resultados de mIoU para cada clase.

Clase (ID)	versión 3 30k	versión 3 70k	versión 4 30k	versión 4 70k
0 (background)	0.4011	0.3989	0.3267	0.3366
1 (cardboard)	0.9214	0.8356	0.6542	0,7511
2 (insulation)	0.6933	0.6363	0.6496	0.4436
3 (metal)	0,7606	0.6542	0.6367	0,7751
4 (minerals)	0.9802	0.9090	0,7177	0.6597
5 (opaque_plastic_bag)	0.8127	0.9664	0.5174	0.8032
6 (plants)	0.4876	0.6091	0.4665	0.5549
7 (plaster)	0,7321	0.6648	0,7203	0.6237
8 (polystyrene)	0.0000	0.0000	0.0000	0.0000
9 (rigid_plastic)	0.3897	0.3514	0.4160	0,1373
10 (soft_plastic_opaque)	0.8042	0.6790	0.5640	0,5682
11 (soft_plastic_transparent)	0,7219	0.6374	0.6043	0,7143
12 (tarp)	0,1648	0,2352	0,2336	0.1340
13 (transparent_plastic_bag)	0,7666	0.9355	0.8305	0.9374
14 (textile)	0.0000	0.0000	0.0000	0.0000
15 (wood)	0.8404	0,7784	0.6429	0.6288

Entre la versión 3 y la versión 4, los cambios más significativos corresponden a una disminución del rendimiento para la mayoría de las clases, con la excepción de las clases *transparent_plastic_bag* y *metal*, que son las únicas que mostraron una mejora.

En resumen, solo algunos casos mejoraron, lo que invita a replantear la búsqueda de mejores *text prompts*, a cubrir un espacio de descripciones textuales más diverso, a combinar espacios vectoriales o a ampliar el conjunto de datos utilizado para calcular las métricas mIoU.

4.3 Discusión de resultados

Los resultados muestran que la hipótesis es plausible, pero todavía no puede considerarse completamente validada para un uso en producción. En varias clases, como *cardboard*, *minerals*,

metal, *transparent_plastic_bag* y *wood*, se obtuvieron valores de mIoU altos en al menos una configuración, lo que indica que el modelo puede aprender regiones semánticas coherentes a partir de supervisión débil. Además, las predicciones visuales confirman que WeCLIP es capaz de producir máscaras interpretables en imágenes reales del dominio industrial.

Sin embargo, el comportamiento no fue homogéneo entre clases. Algunas clases obtuvieron resultados nulos, como *polystyrene* y *textile*, debido principalmente a la ausencia de ejemplos monolabel disponibles para la evaluación. Otras clases mostraron sensibilidad al *checkpoint* utilizado: el entrenamiento de 70 000 iteraciones no siempre mejoró los resultados respecto al de 30 000 iteraciones, lo que sugiere posibles efectos de sobreajuste o de degradación de ciertas clases durante el entrenamiento prolongado.

La optimización de *text prompts* produjo mejoras puntuales, especialmente en *metal* y *transparent_plastic_bag*, pero también redujo el rendimiento en otras clases. Esto confirma que los *prompts* influyen fuertemente en la calidad de las CAMs y, por extensión, en las pseudo-etiquetas utilizadas durante el entrenamiento. No obstante, también muestra que una búsqueda de *prompts* basada en un conjunto monolabel reducido puede no generalizar de forma estable a todas las clases.

4.4 Modelo de segmentación

Siguiendo el proceso clásico de WSSS, en una primera etapa se entrenó un modelo de segmentación sobre estos datos. Sin embargo, dado que esta fase de optimización y análisis de resultados estaba prevista para este período, todavía no se ha llevado a cabo. Continúa prevista como la siguiente etapa, incluyendo una verificación exhaustiva de los datos seguida de una optimización de los hiperparámetros con el objetivo de mejorar el rendimiento del modelo.

4.5 Validación de la hipótesis

Con base en los experimentos realizados, la hipótesis se considera parcialmente validada. El trabajo demuestra que es técnicamente posible adaptar un enfoque WSSS basado en CLIP y DINO al contexto de RecoMat, reutilizando anotaciones a nivel de imagen para generar máscaras semánticas. También se demuestra que estas máscaras pueden alcanzar resultados cuantitativos relevantes en varias clases y que el flujo propuesto es compatible con la lógica general de la cadena actual, ya que parte de imágenes y etiquetas ya disponibles.

No obstante, la validación sigue siendo parcial porque el conjunto de evaluación manual es limitado, algunas clases no pudieron medirse adecuadamente y el rendimiento depende de factores todavía no estabilizados, como el balanceo de datos, la selección de *prompts*, el número de iteraciones y la calidad de las pseudo-etiquetas. Por lo tanto, el enfoque es prometedor para una futura evolución de RecoMat, pero requiere una etapa adicional de consolidación antes de ser considerado una solución lista para producción.

5 Evaluación crítica

Este proyecto fue particularmente exigente, principalmente debido a su componente técnico, situado en la vanguardia de la investigación en visión por computador, pero también por la autonomía de la que disfruté durante todo el trabajo. Partir de un estado del arte muy teórico para después probar un modelo y comenzar a diseñar un prototipo representó un verdadero desafío.

Aunque conté con el acompañamiento de mi tutor y del equipo, el trabajo se desarrolló con frecuencia de manera colaborativa, compartiendo mis ideas para debatir conjuntamente qué era conveniente hacer, modificar o mejorar. Fue un trabajo parcialmente autónomo, pero acompañado de momentos de intercambio y asesoramiento colectivo. Las mejores ideas siempre nacen de las discusiones.

Por otra parte, la cantidad de ideas que surgen en un proyecto tan complejo puede resultar algo abrumadora. Elegir cuáles explorar implica, en ocasiones, tomar decisiones sin disponer de toda la información necesaria, lo cual es completamente normal. Esta experiencia me ayudó mucho a estructurar mi trabajo, avanzar de forma progresiva y obtener resultados concretos.

Además, desarrollé una visión más responsable al utilizar todos estos modelos, como CLIP y DINO, así como ChatGPT, Gemini, Claude y otros. Me ayudaron enormemente, no solo con el código, sino también con la redacción de este informe, las traducciones, la búsqueda de artículos científicos, el análisis, la generación de ideas e incluso las sesiones de *brainstorming*. Realmente representan un cambio de paradigma (*game changer*), y creo que hoy es importante tomar distancia y reflexionar sobre su impacto [21]. No dispongo de una estimación precisa de la cantidad de código escrito manualmente, pero estoy convencido de que, sin estas herramientas, nunca habría podido avanzar tan rápido.

Reconozco, sin embargo, que no he dado un lugar muy importante a estas consideraciones medioambientales en mi proyecto actual; no obstante, considero que uno de los desafíos futuros será precisamente construir, alrededor de estas tecnologías, soluciones que integren en mayor medida la dimensión ecológica y ética, con el fin de reducir su impacto mientras se maximiza su potencial.

Por último, estos modelos evolucionan a un ritmo extremadamente rápido. Durante la realización de este proyecto, por ejemplo, DINOv3 acababa de publicarse y SAM3 ya era anunciado como un modelo capaz de realizar segmentación semántica a partir de *text prompts*. Incluso GPT-5 fue lanzado durante este período. Esta era de crecimiento exponencial no parece mostrar límites a corto plazo, y es precisamente por ello que resulta esencial reflexionar sobre su utilización de manera responsable y eficiente, sin generar gastos excesivos. Este tema, además, está comenzando a cobrar importancia, especialmente dentro de la comunidad de la inteligencia artificial.

6 Conclusión

6.1 Principales logros

Esta práctica dentro del equipo de Data me aportó una gran cantidad de nuevas experiencias, no solo desde el punto de vista técnico, sino también en los ámbitos organizativo y humano. Desde el punto de vista del proyecto, el principal logro fue demostrar que es posible adaptar un enfoque de segmentación semántica débilmente supervisada al contexto de RecoMat, utilizando datos reales de la empresa y anotaciones disponibles a nivel de imagen.

El trabajo permitió construir un flujo experimental completo: análisis y limpieza de datos, revisión de métodos WSSS, comparación de alternativas basadas en modelos fundacionales, adaptación de WebCLIP, creación de un conjunto de evaluación monolabel y medición de resultados mediante mIoU. Los resultados obtenidos no constituyen todavía una solución final, pero sí muestran que el enfoque es técnicamente viable y que algunas clases pueden alcanzar máscaras coherentes y resultados cuantitativos prometedores.

6.2 Limitaciones

La principal limitación del trabajo fue la ausencia de un conjunto de validación amplio y completamente anotado a nivel de píxel. Esto restringió la evaluación cuantitativa y obligó a trabajar con un conjunto monolabel reducido que tomó tiempo y energía en crear. Además, el desbalance entre clases, la presencia de etiquetas imperfectas y la falta de ejemplos para algunas categorías, como *polystyrene* y *textile*, afectaron directamente la interpretación de los resultados.

Otra limitación importante fue la sensibilidad del modelo a los *text prompts*. Aunque la optimización de prompts produjo mejoras en algunas clases, también redujo el rendimiento en otras, lo que muestra que esta etapa requiere una metodología más robusta. Asimismo, los resultados sugieren que el entrenamiento prolongado no siempre mejora el desempeño, por lo que la optimización de hiperparámetros y la selección de *checkpoints* deben estudiarse con mayor profundidad.

6.3 Lecciones aprendidas

Durante estos seis meses, tuve la oportunidad de involucrarme más profundamente en el universo de las *start-ups* tecnológicas y de descubrir de primera mano cómo se toman e implementan las decisiones para alcanzar resultados concretos. Esta visión más amplia, que consiste no sólo en obtener máscaras de segmentación sino también en aportarles valor y presentar ese valor al equipo, representó para mí un nuevo desafío. Esto me llevó a ponerme a prueba en dimensiones distintas a la técnica.

Dar una utilidad concreta al proyecto una vez obtenidos los resultados es una etapa esencial que requiere tiempo y que debe desarrollarse con el mismo cuidado que la parte técnica. Sin este trabajo de valorización y comunicación, los avances no pueden aprovecharse plenamente si la visión y el alcance de una nueva solución no son comprendidos por todos.

Las competencias adquiridas a lo largo de mis estudios, y en particular durante mi último año en IMT Atlantique, me permiten adaptarme a un entorno de trabajo dinámico, donde la gestión de proyectos, la redacción, la generación de ideas y los conocimientos adquiridos a través de mis trabajos académicos fueron valiosos para llevar a cabo esta parte menos técnica.

6.4 Trabajo futuro

El trabajo futuro debería centrarse en ampliar el conjunto de evaluación con máscaras manuales más representativas, especialmente para clases poco frecuentes o ausentes en el conjunto monolabel. También sería necesario profundizar en la limpieza de datos, estudiar estrategias de balanceo más estables y realizar una optimización sistemática de hiperparámetros.

Otra línea importante consiste en mejorar la selección de *text prompts*. Una búsqueda más amplia, apoyada en más imágenes anotadas y posiblemente en combinaciones de descripciones por clase, podría permitir una activación más estable de las regiones de interés. Finalmente, una etapa posterior debería estudiar la integración del modelo de segmentación ligero en la cadena de RecoMat, conectando las máscaras generadas con el cálculo automático de proporciones de materiales en camiones.

6.5 Agradecimientos

El proyecto que me fue confiado me permitió apreciar mejor el valor de mi trabajo. Tuve la suerte de estar supervisado por Adel, quien me guió durante toda la práctica con una actitud de colega y de respeto, mostrando un interés constante por mis logros y brindándome apoyo tanto en los aspectos técnicos como en los organizativos y estratégicos. Su acompañamiento fue determinante para el éxito de mi práctica.

También deseo agradecer a Younes, *Senior Data Scientist* del equipo, quien me acompañó durante todo el proyecto. Su experiencia y su dominio de la inteligencia artificial fueron especialmente valiosos en el diseño de los *pipelines*.

En resumen, esta práctica no ha sido para mí más que una fuente de experiencias positivas, reforzando mi motivación para trabajar en el ámbito del desarrollo de la inteligencia artificial, con la esperanza de contribuir a proyectos con una vocación sostenible y medioambiental, como fue el caso de esta *start-up*, comprometida con la reducción de la huella de carbono.

Quisiera también añadir, para finalizar, mi agradecimiento a mi familia, a mis amigos y a todas las personas que me han ayudado a llegar hasta aquí. Aprovecho estas últimas líneas para darles las gracias, porque sin ellos nada de esto habría sido posible. El éxito siempre depende, en cierta medida, de la suerte, y la mía ha sido tenerlos en mi vida.

Referencias

- [1] Shen, Wei et al. (2023). Un estudio sobre la segmentación profunda de imágenes con eficiencia de etiquetas: superando la brecha entre la supervisión débil y la predicción densa, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, n.º 8.
- [2] Ahn, J., Kwak, S. y Han, B. (2019). Aprendizaje débilmente supervisado de la segmentación de instancias con relaciones entre píxeles. En *Actas de la Conferencia IEEE/CVF sobre Visión por Computadora y Reconocimiento de Patrones* (págs. 2199–2208).
- [3] Z. Chen et al. (2022). Mapas de reactivación de clases para segmentación semántica con supervisión débil. En la Conferencia IEEE/CVF de 2022 sobre visión artificial y reconocimiento de patrones (CVPR), Nueva Orleans, LA, EE. UU., págs. 959–968, doi: 10.1109/CVPR52688.2022.00104.
- [4] Driouache, Y., Milpied, J., Motamedi, A. (2024). Método basado en la visión para identificar materiales transportados por camiones volquete. *Eng. Appl. Artif. Intel.*, 135, 108768.
- [5] Jiang, P. T., Zhang, Y., Hou, Q., Cheng, M. M. y Wei, Y. (2021). LayerCAM: Exploración de mapas de activación de clases jerárquicas para la localización. *Transacciones IEEE sobre procesamiento de imágenes*, 30, 5875–5888.
- [6] Jo, Y. y Lee, J. (2022). Puzzle-CAM: Localización mejorada mediante la coincidencia de características parciales y completas. En *Conferencia internacional sobre aprendizaje automático* (págs. 10168–10182). PMLR.
- [7] Kolesnikov, A. y Lampert, C. H. (2016). Sembrar, expandir y restringir: Tres principios para la segmentación de imágenes con supervisión débil. En *Conferencia Europea sobre Visión por Computador* (págs. 695–711). Springer, Cham.
- [8] Lee, J., Jo, Y. y Kim, S. (2022). Afinidad de aprendizaje a partir de la atención: Segmentación semántica de extremo a extremo débilmente supervisada con transformadores. En *Actas de la Conferencia IEEE/CVF sobre Visión por Computadora y Reconocimiento de Patrones* (págs. 11574–11584).
- [9] Lheritier, A. y Boremann, F. (2023). Un estudio de la segmentación semántica de imágenes semisupervisada y débilmente supervisada. *Revista de Inteligencia Artificial*, 56(9), 9721–9769.
- [10] Pan, Z. & Hay, N. (2023). *¿Una alternativa al WSSS? Un estudio empírico del modelo de segmentación de cualquier cosa (SAM) en problemas de segmentación semántica con supervisión débil*. Preimpresión de arXiv arXiv:2304.14813.
- [11] Su, Y., Zhou, Y., Chen, J., Zhao, M. y Yang, J. (2024). De SAM a CAM: Explorando el modelo de segmentación semántica débilmente supervisada. En *Actas de la Conferencia de Invierno IEEE/CVF sobre Aplicaciones de la Visión por Computadora* (págs. 539–548).
- [12] Wang, Y., Zhang, J., Kan, M., Shan, S. y Chen, X. (2020). Mecanismo de atención equivariante autosupervisado para la segmentación semántica débilmente supervisada. En *Actas de la Conferencia IEEE/CVF sobre Visión por Computadora y Reconocimiento de Patrones* (págs. 12275–12284).

- [13] Wu, R., Ma, C., Liu, S., Li, G. y Lin, L. (2024). Aprendizaje de indicaciones semánticas para la segmentación semántica débilmente supervisada. En *Actas de la Conferencia AAAI sobre Inteligencia Artificial*, 38(5), 5625–5633.
- [14] Wu, S., Xie, E., Gu, J., Liu, J., Wang, W. y Luo, P. (2023). Segmentación semántica débilmente supervisada asistida por el modelo de base. En *Actas de la Conferencia Internacional IEEE/CVF sobre Visión por Computador* (págs. 574–584).
- [15] Xu, C., Yao, T., Lu, H., He, Z. y Fan, X. (2023). ISIM: Modelo iterativo automejorado para segmentación débilmente supervisada. En *Actas de la Conferencia IEEE/CVF sobre Visión por Computadora y Reconocimiento de Patrones* (págs. 3584–3593).
- [16] Zhang, D. y Zhang, H. (2020). *Un estudio sobre segmentación semántica débilmente supervisada*. Preimpresión de arXiv arXiv:2010.00950.
- [17] Zhou, Z. H. (2018). Una breve introducción al aprendizaje débilmente supervisado. *Revista Nacional de Ciencias*, 5(1), 44–53.
- [18] Chen, Tianle, Mai, Zheda, Li, Ruiwen y Chao, Wei-Lun. (2023). Modelo Segment Anything (SAM): Pseudoetiquetas mejoradas para segmentación semántica débilmente supervisada. doi: 10.48550/arXiv.2305.05803.
- [19] Zhu, L., Li, Y., Fang, J., Liu, Y., Xin, H., Liu, W. y Wang, X. (2023). WeakTr: Explorando el transformador de visión simple para la segmentación semántica débilmente supervisada. *ArXiv*, abs/2304.01184.
- [20] B. Zhang, S. Yu, J. Xiao, Y. Wei y Y. Zhao. Frozen CLIP-DINO: una sólida estructura para la segmentación semántica débilmente supervisada. *Transacciones IEEE sobre análisis de patrones e inteligencia de máquinas*, vol. 47, núm. 5, págs. 4198–4214, mayo de 2025, doi: 10.1109/TPAMI.2025.3543191.
- [21] Patterson, David, González, Joseph, Le, Quoc, Liang, Chen, Munguía, Lluís-Miquel, Rothchild, Daniel, So, David, Texier, Maud y Dean, Jeff. (2021). Emisiones de carbono y formación de grandes redes neuronales. doi: 10.48550/arXiv.2104.10350.
- [22] Brandt, Laura, Fu, Stephanie, Hamilton, Mark, Feldmann, Axel, Zhang, Zhoutong y Freeman, William. (2024). FeatUp: Un marco agnóstico del modelo para características en cualquier resolución.
- [23] Lin, S., Su, Y., Zhou, Y., Chen, J. y Yang, J. (2024). Exploración del conocimiento denso de CLIP para la segmentación semántica débilmente supervisada. *Transacciones IEEE sobre análisis de patrones e inteligencia de máquinas*, 46(7), 4301–4315.