

Superresolución en Imágenes captadas por Dron para la Detección de Frailejones

Joan Sebastián Pedraza Paula Uzcátegui León

Universidad Industrial de Santander

joan2210100@correo.uis.edu.co, paula2211475@correo.uis.edu.co

Abstract

La detección automática de frailejones en ecosistemas de páramo mediante imágenes captadas por dron se ve limitada por la resolución espacial (GSD) de las imágenes. Este proyecto propone el uso de técnicas de superresolución para mejorar la calidad de imágenes de baja resolución (GSD 3.3cm) antes de aplicar modelos de detección de objetos. Se implementó el modelo Real-ESRGAN y se evaluó su capacidad para reconstruir imágenes de alta resolución (GSD 1.3 cm) que fueron degradadas artificialmente mediante reducción de resolución. La calidad de la reconstrucción se midió utilizando la métrica PSNR. Se evaluó el desempeño de Real-ESRGAN en la mejora de imágenes de baja resolución para la detección automática de frailejones, utilizando un modelo YOLOv11 previamente entrenado con imágenes de alta resolución. Los resultados muestran que, si bien la superresolución mejora la calidad visual y la precisión de detección, la resolución original de las imágenes sigue siendo un factor limitante, por lo que deberían explorarse otras estrategias complementarias. Finalmente, esta estrategia ofrece una solución práctica y eficaz para maximizar el uso de vuelos con drones en entornos que son de difícil acceso, sin comprometer la precisión en las tareas de monitoreo ambiental.

1. Introduction

La detección de frailejones en imágenes captadas por drones es una herramienta fundamental para el monitoreo remoto de estas especies, permitiendo estimar su estado de salud y cuantificar su presencia en ecosistemas de páramo. Los drones resultan especialmente útiles en estos entornos debido a que los páramos suelen ubicarse en zonas remotas y de difícil acceso.

Sin embargo, la captura de imágenes en estas regiones presenta múltiples desafíos. Las condiciones climáticas adversas, como la neblina y el viento, pueden dificultar la obtención de datos de calidad. Además, la baja concentración de oxígeno en los páramos afecta el rendimiento de las baterías, limitando el tiempo de vuelo y, por ende, el

área que puede ser cubierta por misión.

Uno de los factores críticos en la adquisición de imágenes es la resolución espacial, determinada por parámetros como el tipo de sensor, la altura de vuelo, el tamaño del objeto de interés y el nivel de detalle requerido. Volar a mayor altitud permite cubrir grandes extensiones en menos tiempo, pero a costa de una menor resolución. En contraste, vuelos a baja altitud ofrecen mayor nivel de detalle, aunque con menor cobertura territorial y mayor consumo de recursos. En muchos casos, la resolución final de las imágenes está determinada por las condiciones del momento, más que por una elección técnica ideal.

En teledetección, la resolución espacial se mide comúnmente en términos de *Ground Sampling Distance* (GSD), que indica el tamaño real del terreno representado por cada píxel en la imagen (expresado en centímetros, metros o kilómetros por píxel). En la Figura 1 se observan dos imágenes de una misma área captadas con un dron Mavic 3 Multiespectral [1] a diferentes alturas. Un mismo frailejón luce distinto a diferentes GSDs: 1.3 cm/píxel y 3.3 cm/píxel. Aunque se trata del mismo individuo, la reducción de resolución implica la perdida de detalles, particularmente en la forma de roseta de la planta, lo que dificulta su detección automática.

Trabajos previos han explorado el uso de técnicas de superresolución como etapa de preprocesamiento antes de aplicar modelos de aprendizaje profundo para la detección de objetos en imágenes captadas por dron. Para la detección de rosas en imágenes captadas por dron [6], implementan diversos modelos de reconstrucción, incluyendo interpolación bicúbica, Real-ERSGAN, SECNN, RCANN, EDSR, SWINIR y MambaIR, y evalúan su desempeño con imágenes de alta resolución degradadas artificialmente. Otros se enfocan exclusivamente en el uso de Real-ERSGAN para la detección de lichies en imágenes captadas por dron [3]. Ambos estudios concluyen que el uso de superresolución permite obtener resultados de detección comparables a los obtenidos con imágenes de alta resolución.

Real-ESRGAN [4] es un modelo de superresolución diseñado para aumentar el tamaño de las imágenes de entrada, mejorando notablemente tanto sus detalles como su

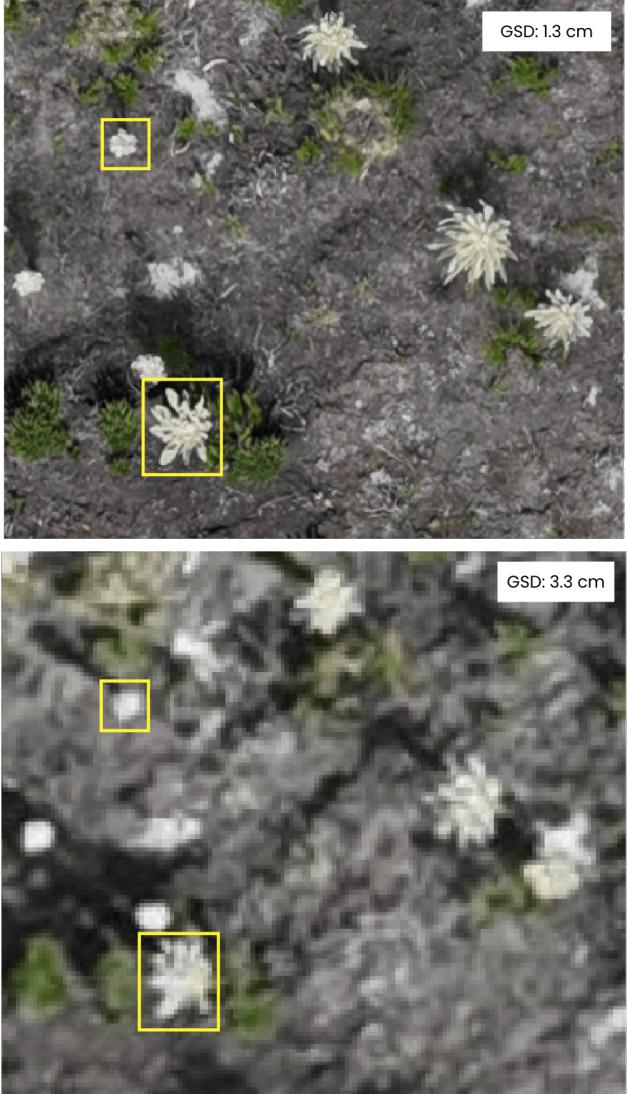


Figure 1. Imágenes con GSD de 1.3cm y 3.3cm, en las cajas amarillas se identifican dos frailejones de diferentes tamaños. Ambas imágenes fueron capturadas con un dron Mavic 3 Multiespectral en la misma ubicación pero volando a diferentes alturas. Se observa como la forma del frailejón pierde definición en la imagen de menor resolución, especialmente cuando la roseta es de menor tamaño, lo que puede dificultar su reconocimiento y llevar a confusiones con otras plantas o incluso con rocas.

calidad visual. Se basa en la arquitectura ESRGAN [5] (*Enhanced Super-Resolution Generative Adversarial Networks*), la cual a su vez se deriva de SRResNet [2], y aprovecha el enfoque de redes generativas adversariales (GANs).

En este tipo de arquitectura, se entrena conjuntamente dos redes: una generadora, que intenta reconstruir imágenes de alta resolución a partir de versiones de baja resolución, y una discriminadora, que aprende a distinguir entre imágenes

reales (de alta resolución) y las imágenes generadas. El objetivo es que, con el tiempo, la red generadora produzca imágenes tan realistas que el discriminador no pueda diferenciarlas de las reales. En la Figura ?? se muestra un esquema de la arquitectura de este tipo de modelos.

Real-ESRGAN enfoca principalmente en mejorar las versiones anteriores al incorporar simulaciones de degradaciones más realistas durante el entrenamiento (como ruido, desenfoque y compresión), lo que permite que el modelo sea más robusto ante condiciones del mundo real (de ahí el nombre *Real*). Reportan una mayor capacidad para eliminar ruido y producir imágenes más nítidas que los métodos tradicionales. Además, destaca por su facilidad de implementación, respaldada por una documentación clara y completa disponible en su repositorio oficial: <https://github.com/xinntao/Real-ESRGAN>.

2. Metodología

En la Figura 3, se presenta un esquema del flujo de datos propuesto. Principalmente dividimos nuestra metodología en tres partes: Generación y preprocesamiento de los datos, Modelo de Superresolución y Evaluación con modelo de detección.

2.1. Generación del dataset

Se dispuso de dos mosaicos ortorrectificados de una misma área de aproximadamente 4 hectáreas, mostrados en la Figura 4. Uno de ellos presenta un GSD de 1.3 cm/píxel, al que nos referiremos como alta resolución, y el otro tiene un GSD de 3.3 cm/píxel, que denominaremos baja resolución. Aunque ambos mosaicos cubren exactamente la misma zona geográfica, la identificación visual de los frailejones resulta considerablemente más difícil en el mosaico de baja resolución.

A partir de estos mosaicos se generaron imágenes de $8 \text{ m} \times 8 \text{ m}$ de extensión en terreno. En el caso del mosaico de alta resolución (1.3 cm/píxel), esto se aproxima a imágenes de 512×512 píxeles. Para el mosaico de baja resolución (3.3 cm/píxel), el tamaño resultante se aproxima a un tamaño de 256×256 píxeles. Se construyeron dos conjuntos de prueba, uno por cada resolución, ambos compuestos por 77 imágenes.

En cada conjunto, los frailejones fueron anotados mediante cajas delimitadoras (*bounding boxes*), lo que permite evaluar de forma objetiva el impacto de la superresolución en tareas de detección. El conjunto de baja resolución es el objetivo principal del proceso de reconstrucción por superresolución, ya que se pretende emular la calidad del mosaico original de alta resolución y, con ello, mejorar el desempeño del modelo de detección sobre imágenes de menor calidad.

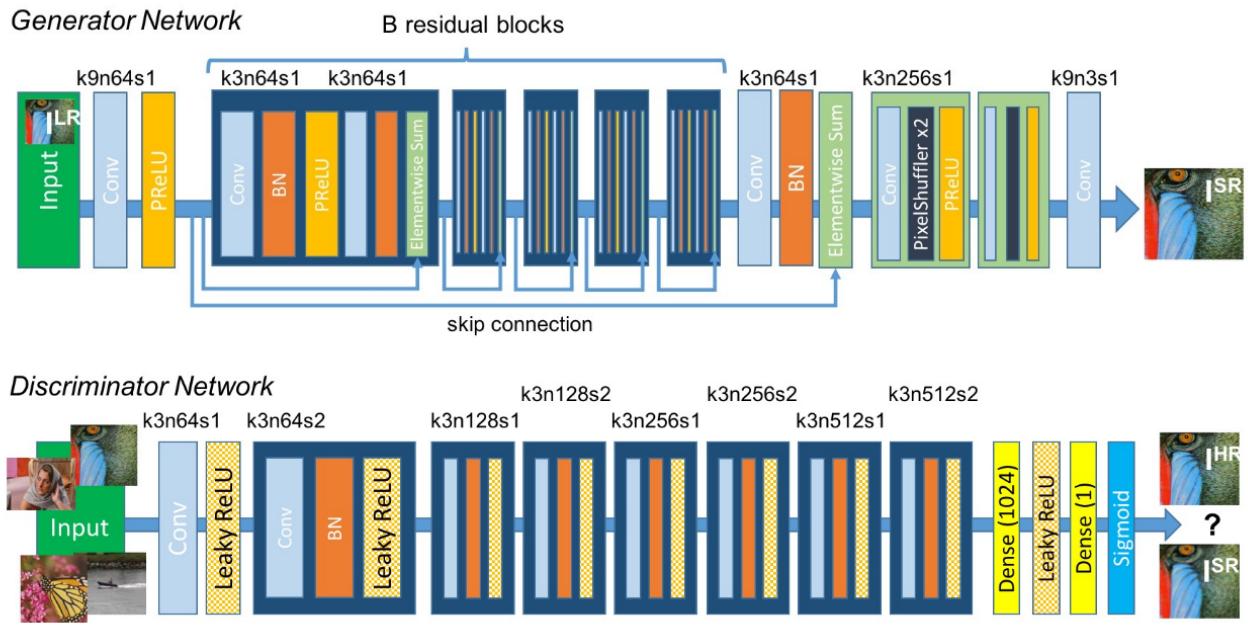


Figure 2. Arquitectura adversarial del modelo SRResNet, en el que se basan los modelos Real-ERSGAN y ERSGAN. Tomado de [2]

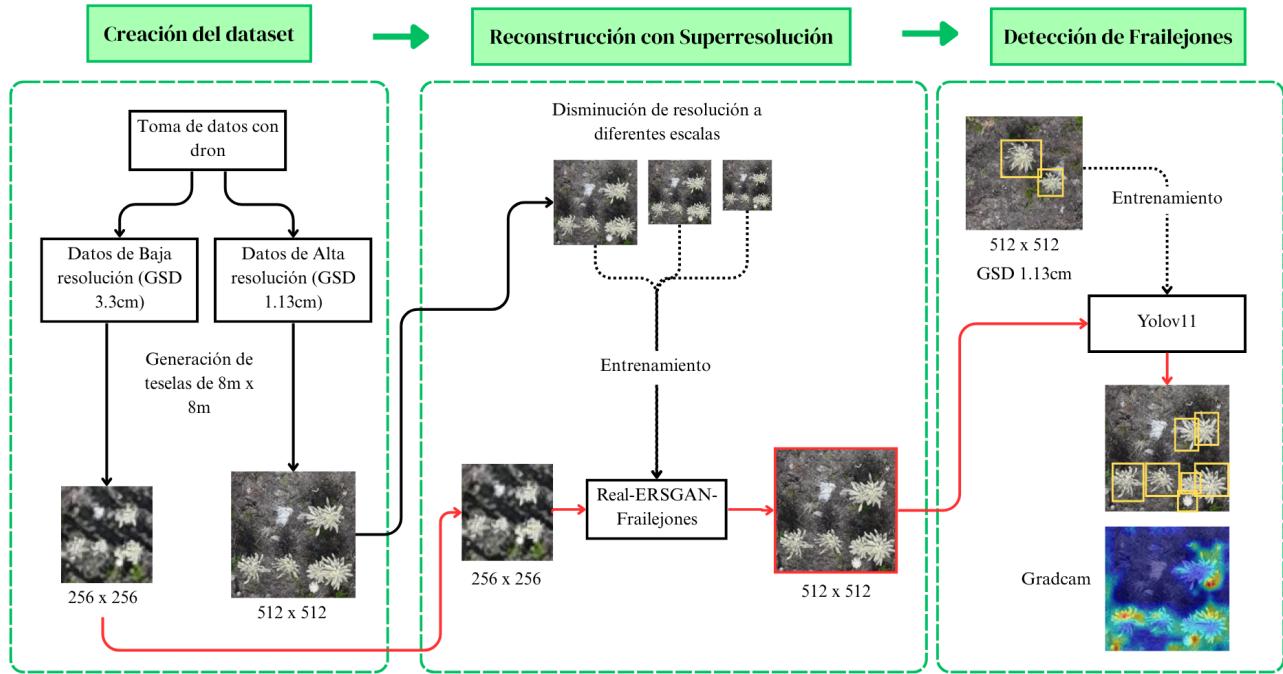


Figure 3. Metodología implementada para el flujo de datos.

2.2. Superresolución

Para implementar la reconstrucción, se utilizó el modelo Real-ERSGAN, siguiendo la implementación oficial



Figure 4. Mosaico de aproximadamente 4 hectáreas, generado con dos GSD distintos: 1.3 cm/píxel y 3.3 cm/píxel. El área delimitada por la línea roja indica la zona utilizada para el testeo.

disponible en el repositorio de GitHub de los autores <https://github.com/xinntao/Real-ERSGAN>. Se crea un conjunto de datos de entrenamiento y testeo a partir de imágenes de alta resolución (diferentes a las del conjunto de prueba mencionado anteriormente), obteniendo un total de 55 imágenes para la partición de testeo y 220 imágenes para la partición de entrenamiento, que se utilizarán para hacer ajuste fino a Real-ERSGAN. Se hace aumento de datos sobre este conjunto, degradando mediante reescalado a factores de 0.75, 0.5 y 0.33, con el fin de simular distintas condiciones de pérdida de resolución.

Primero, se utilizó el modelo base *RealESRGAN_{x4plus}* para realizar la reconstrucción de las imágenes reescaladas con un factor de escalado de 4. Luego de evaluar visualmente las imágenes reconstruidas, se decidió hacer ajuste fino al mismo modelo y así ver cómo mejoraba la reconstrucción de las imágenes para nuestro dataset.

Para realizar el ajuste fino, se usaron dos particiones, un **Ground Truth** con las imágenes originales y un **Multiscale** con las imágenes degradadas a distintas escalas, ya teniendo los datos, se entrenó el modelo con 880 imágenes reescaladas y por 2200 iteraciones, o 10 epochs, con un tamaño de batch de 4, en cada iteración se pasa un batch del dataset. Este proceso se realizó en Google Colab y tomó más de media hora. Con el modelo ya entrenado, pasamos a realizar inferencia en la partición de testeo del dataset y a comparar los resultados con el modelo base.

Para evaluar la calidad del modelo de superresolución se mide la calidad de la reconstrucción de los datos de alta resolución, mediante la métrica PSNR

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \quad (1)$$

$$\text{donde } \text{MSE} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n [I(i, j) - K(i, j)]^2 \quad (2)$$

Siendo MAX el valor máximo posible de un pixel.

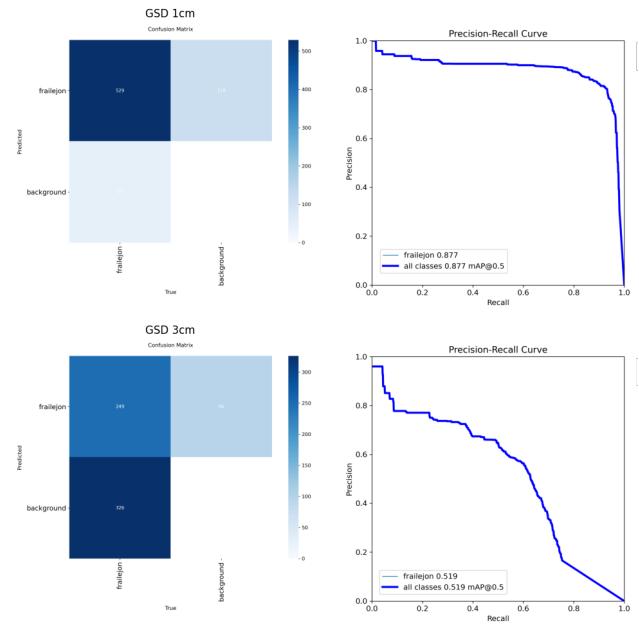


Figure 5. Matriz de confusión y curva de Precisión–Recall para detección de frailejones en imágenes con GSD de 1.3cm/pixel y 3.3cm/pixel.

2.3. Detección

Para evaluar la efectividad de la superresolución en la mejora de la identificación automática de frailejones, se utilizaron las predicciones generadas por un modelo YOLOv11 previamente entrenado con imágenes de frailejones capturadas a una resolución de 1.3 cm/píxel (alta resolución).

La métrica principal utilizada fue el *Average Precision* (AP), definida como el área bajo la curva *Precisión–Recall*. Esta métrica cuantifica el equilibrio entre la capacidad del modelo para detectar correctamente los objetos de interés (*Recall*) y su habilidad para evitar falsos positivos (Precisión). El uso del AP permite determinar si las imágenes procesadas mediante superresolución efectivamente mejoran el rendimiento del modelo detector.

En la Figura 5 se muestran los resultados de detección para el área de prueba en ambos niveles de resolución. Como se puede observar, el modelo presenta un desempeño considerablemente inferior en las imágenes de baja resolución, con un *Average Precision* de apenas 0.51, en comparación con un valor de 0.87 obtenido en las imágenes de alta resolución. En este caso, la mayoría de las detecciones corresponden a falsos positivos, lo que indica que el modelo tiende a confundir el suelo u otros elementos del entorno con frailejones reales.

3. Resultados

Luego de entrenar el modelo, medimos el desempeño del modelo base y el modelo entrenado con una partición de testeo obteniendo un valor de PSNR para una escala de 0.25, así haciendo que los modelos reconstruyeran la imagen a su estado original.

Se calculó el valor de PSNR de cada imagen de la partición de testeo, así como un PSNR total para toda la partición, los resultados de los PSNR promedios fueron los siguientes:

$$\text{PSNR (Modelo Base)} = 21.59 \text{dB} \quad (3)$$

$$\text{PSNR (Modelo Entrenado)} = 21.05 \text{dB} \quad (4)$$

Los resultados de los PSNR individuales para algunas imágenes se muestran en la figura 6

Durante la fase de inferencia, se generaron reconstrucciones de imágenes las imágenes de baja resolución (GSD de 3.3cm) utilizando el modelo Real-ERSGAN tanto en su versión original (sin reentrenar) como en una versión reentrenada con imágenes degradadas desde un GSD de 1.3cm/píxel, tal como se describió en la sección anterior. Las imágenes resultantes se presentan en la Figura 8.

Se calcularon las métricas de detección para las imágenes reconstruidas. En la Tabla 1 se reportan los valores de *Average Precision* (AP) con un umbral de intersección sobre la unión (IoU) de 0.5 para diferentes métodos de reconstrucción. Como se también se puede observar en la Figura 7, la introducción de métodos de superresolución logra mejorar la métrica de Average Precision en cerca de un 30%. Sin embargo, el uso de modelos como Real-ERSGAN no mejoran realmente la detección de forma significativa, pues tienen un desempeño tan solo un poco mejor que la interpolación bicúbica, a la vez que es mucho más costoso de implementar. Además de esto el modelo Real-ERSGAN entrenado muestra un desempeño mucho más pobre que el modelo sin entrenar usado en formato *zero-shot*. Esto también se puede ver en las imágenes de la Figura 8, el entrenamiento del modelo empeora el aspecto visual de la imagen y además dificulta la tarea de detección de frailejones. Esto podría deberse a que se necesitaba un mayor entrenamiento del modelo para obtener mejores resultados, o a que posiblemente este tipo de imágenes no obtienen un buen desempeño con esta arquitectura en específico para la tarea de superresolución.

Imagen	<i>AP@0.5</i>
GSD 1.3 cm	0.877
GSD 3.3 cm	0.519
GSD 3.3 cm – Interpolación Bicúbica	0.526
GSD 3.3 cm – Real-ERSGAN	0.541
GSD 3.3 cm – Real-ERSGAN reentrenado	0.142

Table 1. Resultados de detección de frailejones en imágenes reconstruidas. El modelo Real-ERSGAN sin reentrenar supera ligeramente a la interpolación bicúbica.

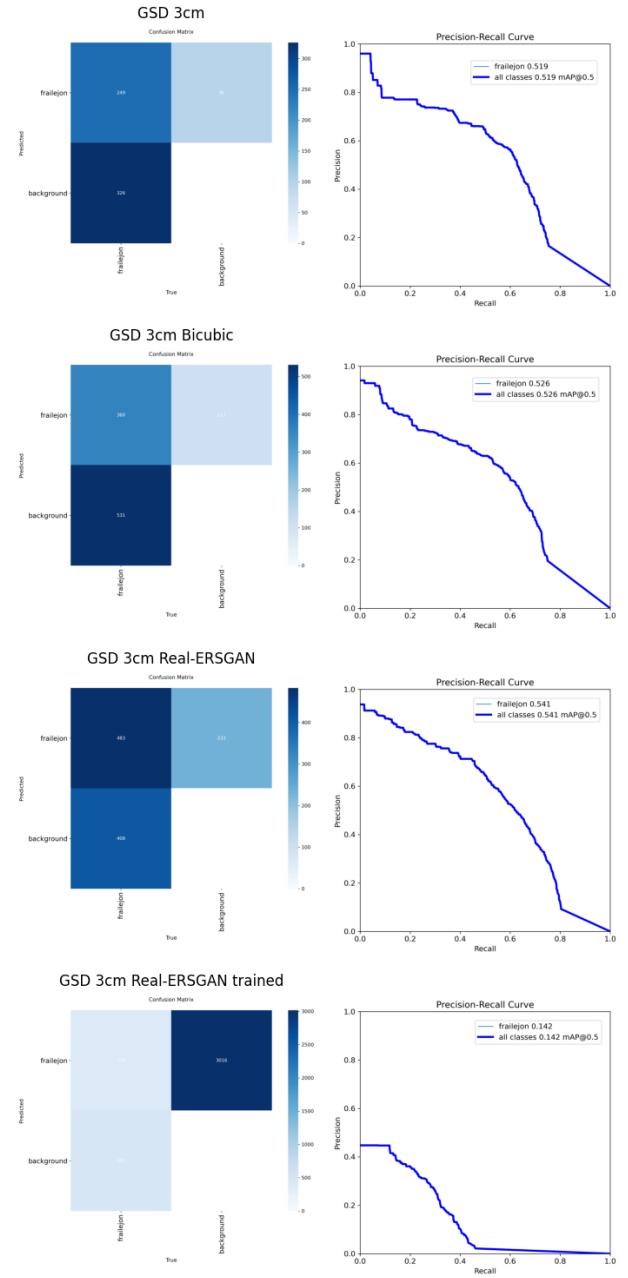


Figure 7. Matriz de confusión y curvas de precisión–recall para las reconstrucciones de imágenes con GSD de 3.3cm/píxel.

Adicionalmente, se aplicó la técnica Grad-CAM (*Gradient-weighted Class Activation Mapping*) para visualizar las regiones de las imágenes que influyen en las decisiones del modelo. Esta técnica permite interpretar qué partes de una imagen son más relevantes para la detección, revelando el foco de atención del modelo en cada predicción.

En el caso específico de YOLOv11, que emplea capas inspiradas en transformadores visuales (ViT) con mecanismos de atención, se utilizó una adaptación basada en descomposición en valores singulares (SVD). Esta técnica extrae los componentes principales de las activaciones y proyecta el vector de mayor varianza para generar un mapa de atención espacial. Se implementó a partir del repositorio disponible en <https://github.com/rigvedrs/YOLO-V11-CAM>.

La Figura 9 muestra los resultados de Grad-CAM aplicados al último bloque del backbone de YOLOv11. Las regiones más brillantes indican las zonas donde el modelo concentra su atención al predecir la presencia de frailejones. Como se puede observar las imágenes con superresolución parecen apuntar mejor a las zonas con presencia de frailejones, especialmente para las imágenes generadas con Real-ERSGAN parecen haber más detecciones, aunque comete algunos errores.

4. Conclusiones

En este trabajo entrenamos y probamos una técnica de superresolución para mejorar el desempeño de un modelo de detección de frailejones en imágenes captadas por dron. Sin embargo, los resultados muestran que, con la implementación del modelo Real-ERSGAN, no se alcanzó un nivel de detección aceptable. Esto puede explicarse por dos factores principales:

Primero, que el modelo de superresolución requiere un entrenamiento más específico con imágenes del dominio aéreo, ya que fue originalmente entrenado con imágenes naturales. El ruido y los artefactos presentados en imágenes captadas por dron son muy diferentes de la degradación que tienen normalmente las imágenes naturales. Esta diferencia de dominio provoca que el modelo introduzca artefactos irreales “alucinaciones” en las reconstrucciones, afectando negativamente la detección.

Segundo, puede que la brecha entre las resoluciones de 1.3 cm/pixel y 3.3 cm/pixel es demasiado amplia. Esta pérdida de información dificulta la reconstrucción precisa de los detalles necesarios para la detección, especialmente cuando el modelo detector ha sido entrenado exclusivamente con imágenes de alta resolución.

References

- [1] DJI. Mavic 3 multispectral. DJI Agriculture, 2025. Retrieved February 22, 2025. [1](#)
- [2] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. [2, 3](#)
- [3] Changjiang Liang, Juntao Liang, Weiguang Yang, Weiyi Ge, Jing Zhao, Zhaorong Li, Shudai Bai, Jiawen Fan, Yubin Lan, and Yongbing Long. Enhanced visual detection of litchi fruit in complex natural environments based on unmanned aerial vehicle (uav) remote sensing. *Precision Agriculture*, 26(1):23, 2025. [1](#)
- [4] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*. [1](#)
- [5] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. [2](#)
- [6] Fan Zhao, Zhiyan Ren, Jiaqi Wang, Qingyang Wu, Dianhan Xi, Xinlei Shao, Yongying Liu, Yijia Chen, and Katsunori Mizuno. Smart uav-assisted rose growth monitoring with improved yolov10 and mamba restoration techniques. *Smart Agricultural Technology*, 10:100730, 2025. [1](#)

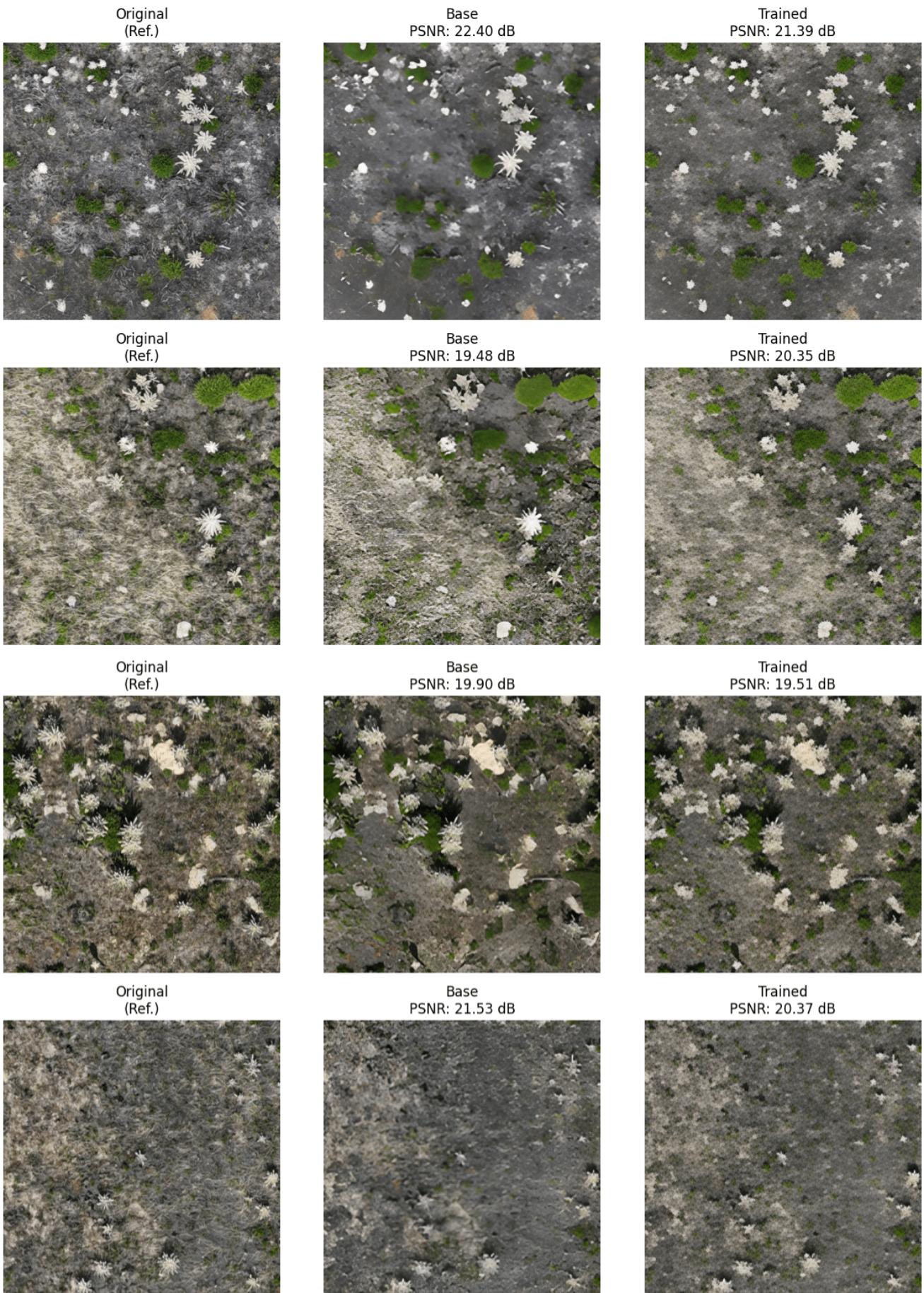


Figure 6. Resultados del PSNR para el modelo base y el modelo entrenado

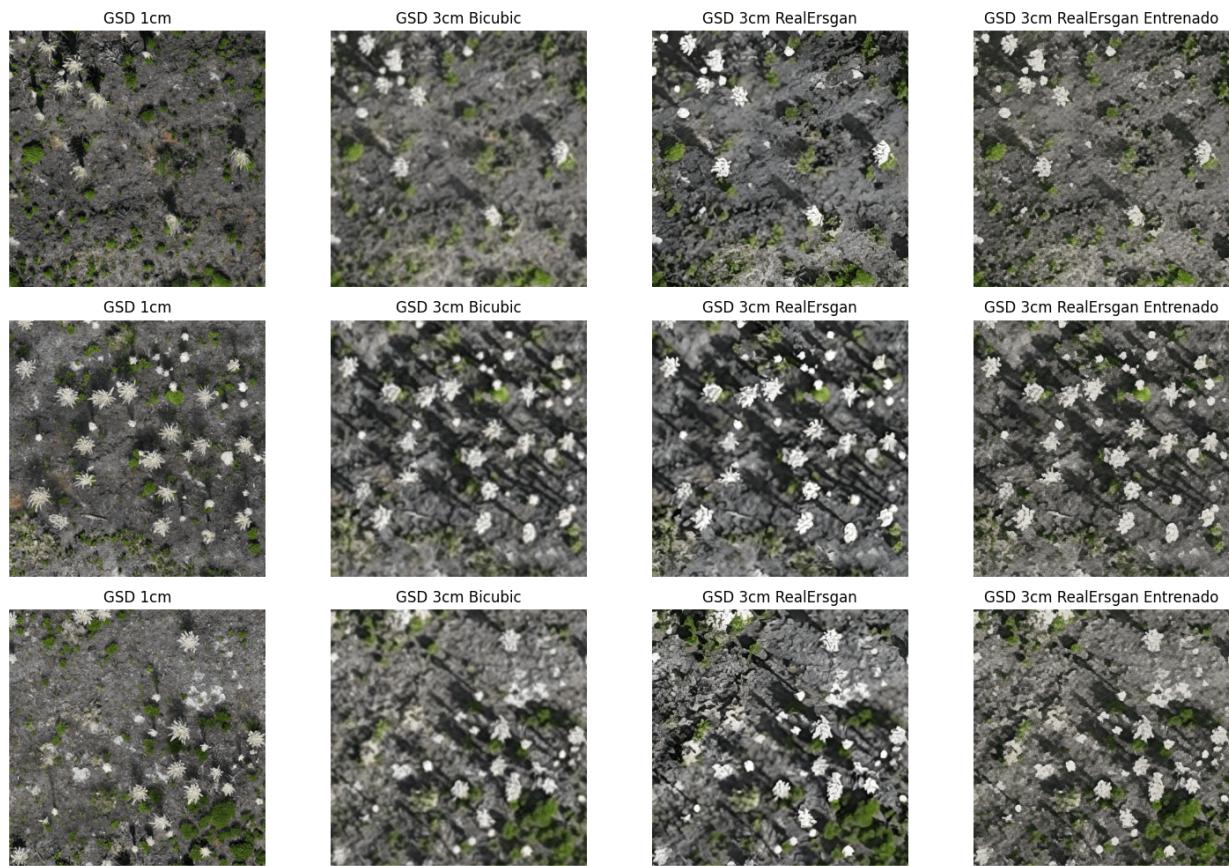


Figure 8. Imágenes reconstruidas mediante superresolución comparadas con las imágenes originales captadas por el dron a 1.3cm

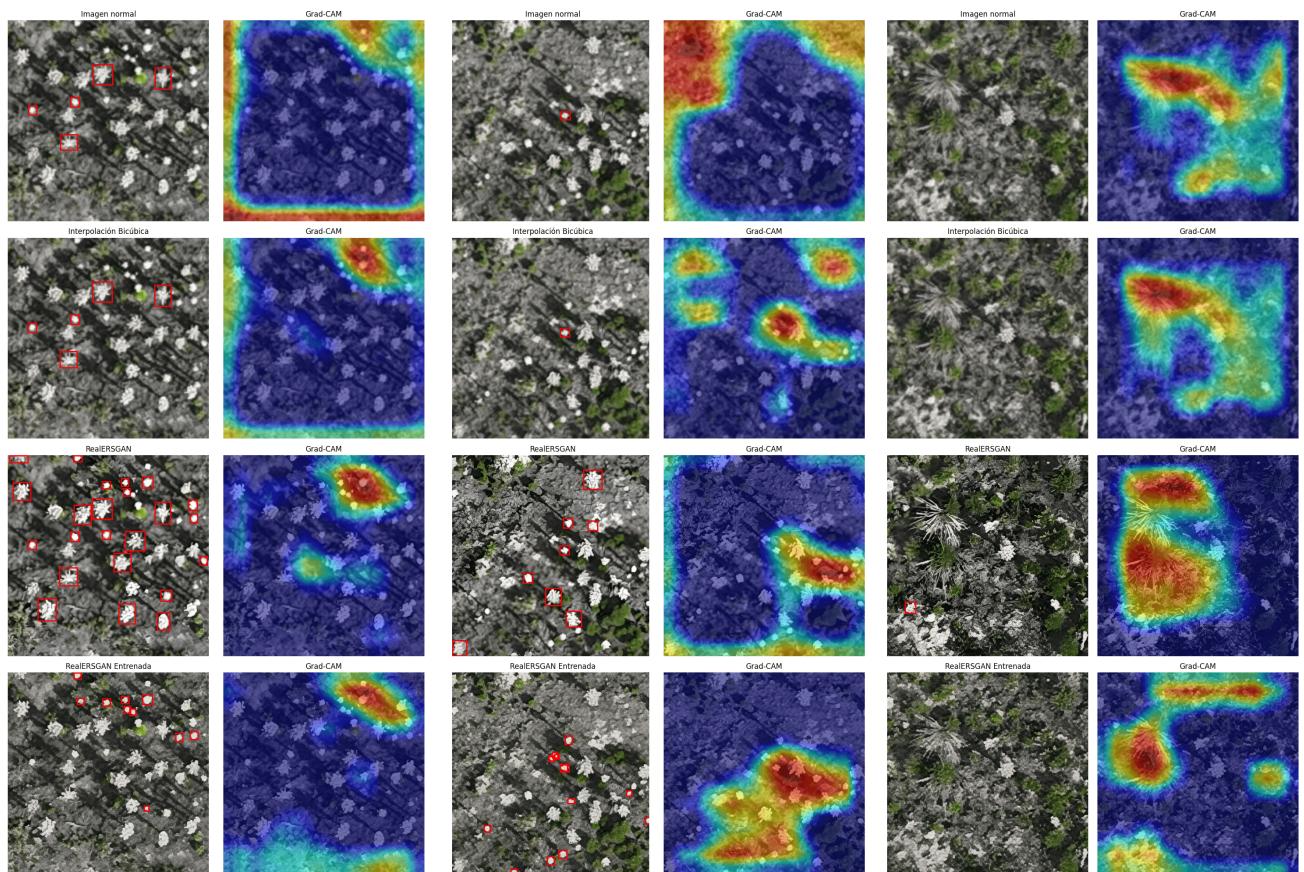


Figure 9. Visualización Grad-CAM para reconstrucciones de imágenes con GSD de 3.3cm