

LEARNING ENRICHED FEATURES FOR FAST IMAGE RESTORATION AND ENHANCEMENT

Fredy A. Huanca T.* , Jose E. Perez M.**y Henrry I. Arias M.***

Abstract

Image restoration is an important task in surveillance, computational photography, medical imaging, and remote sensing. Recently, convolutional neural networks (CNNs) have achieved dramatic improvements over conventional approaches. This paper presents a novel architecture with the collective goals of maintaining spatially precise high-resolution representations and receiving contextual solid information from the low-resolution representations. The core of the approach is a multi-scale residual block containing several key elements: parallel multi-resolution convolution streams, information exchange across the multi-resolution streams, spatial and channel attention mechanisms, and attention-based multi-scale feature aggregation. Extensive experiments on real image benchmark datasets demonstrate that MIRNet achieves state-of-the-art results for various image processing tasks.

Keywords Image denoising, super-resolution, and image enhancement

INTRODUCCIÓN

La restauración de imágenes es una tarea desafiante debido a la presencia de degradaciones en las imágenes adquiridas desde varios dispositivos. Las cámaras con capacidades limitadas, como las cámaras de los teléfonos inteligentes, a menudo producen imágenes ruidosas y de bajo contraste. Además, las condiciones de iluminación inadecuadas pueden dar como resultado imágenes demasiado oscuras o demasiado brillantes. El objetivo de la restauración de imágenes es recuperar la imagen limpia original de sus medidas corruptas, lo cual es un problema inverso mal planteado con múltiples soluciones posibles.

Los modelos de aprendizaje profundo han logrado avances significativos en la restauración y mejora de imágenes mediante el aprendizaje de antecedentes sólidos de conjuntos de datos a gran escala. Las redes neuronales convolucionales (CNN) existentes para la restauración de imágenes suelen seguir una arquitectura de codificador-decodificador o un enfoque de procesamiento de características de alta resolución (escala única). Los modelos de codificador-decodificador reducen progresivamente la resolución espacial de la imagen de entrada y luego la vuelven a mapear a la resolución original. Si bien estos modelos capturan un contexto amplio, a menudo pierden detalles espaciales finos, lo que dificulta su recuperación. Por otro lado, las redes de alta resolución conservan detalles espacialmente precisos pero tienen una

capacidad limitada para codificar información contextual debido a su pequeño campo receptivo.

Para abordar estos desafíos, los autores proponen un nuevo enfoque multiescala llamado MIRNet. Este enfoque mantiene características de alta resolución en toda la jerarquía de la red, lo que minimiza la pérdida de detalles espaciales precisos. Emplea flujos de convolución paralelos para capturar contexto de múltiples escalas, lo que complementa la rama principal de alta resolución y proporciona representaciones de características más precisas y enriquecidas contextualmente. A diferencia de los enfoques multiescala existentes que procesan cada escala de forma independiente e intercambian información solo de arriba hacia abajo, MIRNet fusiona información en todas las escalas en cada nivel de resolución, lo que permite el intercambio de información de arriba hacia abajo y de abajo hacia arriba. Se utiliza un mecanismo de fusión de kernel selectivo para seleccionar dinámicamente kernels útiles de cada representación de rama, conservando sus características complementarias distintivas.

Nuestras principales contribuciones a este artículo incluyen lo siguiente:

- Un nuevo modelo de extracción de características que obtiene un conjunto complementario de características a través de múltiples escalas espaciales mientras mantiene las características originales de alta resolución para preservar detalles espaciales precisos.
- Un mecanismo repetido regularmente para el intercambio de información, donde las características a través de las ramas de resolución múltiple se fusionan progresivamente para mejorar el aprendizaje de la representación.
- Un nuevo enfoque para fusionar características de múltiples escalas utilizando una red de kernel selectiva que combina dinámicamente campos receptivos variables y conserva fielmente la información de la característica original en cada resolución espacial.
- Un diseño residual recursivo que descompone progresivamente la señal de entrada para simplificar el proceso de aprendizaje general y permite la construcción de redes muy profundas.
- Se realizan experimentos completos en cinco conjuntos de datos de referencia de imágenes reales para diferentes tareas de procesamiento de imágenes, incluida la eliminación de ruido de imágenes, la superresolución y la mejora de imágenes. Nuestro método logra resultados de última generación en los cinco conjuntos de datos. Además, evaluamos exhaustivamente nuestro enfoque para los desafíos prácticos, como la capacidad de generalización entre conjuntos de datos.

* Universidad Nacional San Agustín de Arequipa, fhuancat@unsa.edu.pe

**Universidad Nacional San Agustín de Arequipa, jperezma@unsa.edu.pe

***Universidad Nacional San Agustín de Arequipa, hariasim@unsa.edu.pe

TRABAJO RELACIONADOS

los autores revisan los métodos más relevantes para el procesamiento de imágenes de bajo nivel, incluyendo la eliminación de ruido, la superresolución y la mejora de imagen. A continuación, se resumen algunos de los métodos mencionados:

- Para la eliminación de ruido, como los basados en transformaciones de coeficientes y promedio de píxeles vecinos [10]. Además, se han propuesto enfoques basados en redes neuronales convolucionales profundas para esta tarea, como DnCNN [10] y RED30 [4]. Estos métodos utilizan arquitecturas profundas para aprender a eliminar el ruido de las imágenes a partir del conjunto de entrenamiento y generar imágenes limpias a partir de nuevas imágenes o imágenes dañadas por el ruido. Los resultados experimentales muestran que estos métodos pueden superar a los métodos clásicos en términos tanto visuales como cuantitativos.
- Para la superresolución, algunos de los métodos mencionados son bicubic interpolation, que es el método más comúnmente utilizado para generar imágenes de alta resolución a partir de imágenes de baja resolución, y enfoques basados en redes neuronales convolucionales profundas como VDSR [5], DRCN [7] y SRCNN [6]. Estos métodos utilizan arquitecturas profundas para aprender a generar imágenes de alta resolución a partir del conjunto de entrenamiento y mejorar la calidad visual y cuantitativa de las imágenes generadas. Los resultados experimentales muestran que estos métodos pueden superar a los métodos clásicos como bicubic interpolation en términos tanto visuales como cuantitativos.
- Para la mejora de imagen, algunos de los métodos mencionados son MemNet [2] y FFDNet [3], que son enfoques basados en redes neuronales convolucionales profundas que procesan características a su resolución original y fusionan información contextual de múltiples ramas paralelas. Además, se menciona el conjunto de datos MIT-Adobe FiveK [8], que contiene imágenes de diversas escenas interiores y exteriores capturadas con cámaras DSLR en diferentes condiciones de iluminación, y se han utilizado las imágenes mejoradas por expertos como referencia para evaluar el rendimiento de los algoritmos. Los resultados experimentales muestran que los métodos basados en redes neuronales convolucionales profundas pueden mejorar significativamente la calidad visual y cuantitativa de las imágenes en comparación con los métodos clásicos.

Los autores también destacan que muchos de estos métodos están diseñados para una tarea específica y no son fácilmente adaptables a otras tareas. Por lo tanto, proponen MIRNet como un modelo unificado que puede manejar múltiples tareas de procesamiento de imágenes de bajo nivel con un rendimiento estado del arte.

Se describe la arquitectura utilizada en el modelo propuesto. El modelo utiliza una arquitectura de red neuronal profunda basada en bloques residuales (ResNet) y se ha modificado para incluir módulos de atención y conexiones de salto (skip connections) para mejorar su rendimiento. Los bloques residuales permiten que el modelo sea muy profundo sin sufrir problemas de gradiente, mientras que los módulos de atención mejoran la capacidad del modelo para enfocarse en características importantes de la imagen como se muestra en la figura 1. Las conexiones de salto permiten que el modelo utilice información de diferentes niveles de resolución para mejorar la calidad de las imágenes generadas. Además, se realizan experimentos ablativos para analizar el impacto de cada uno de estos componentes en el rendimiento general del modelo, y se encuentra que las conexiones de salto son especialmente importantes para lograr un buen rendimiento. En general, la arquitectura propuesta logra resultados estatales del arte en varias tareas de procesamiento de imágenes, incluyendo denoising, super-resolución y mejora de imágenes.

MÉTODO PROPUESTO

La restauración y mejora de imágenes es un problema importante en muchas aplicaciones prácticas, como la fotografía digital, la medicina y la vigilancia por video [10]. En este artículo, se presenta una nueva técnica basada en redes neuronales convolucionales (CNN) para abordar este problema, en la figura 3. El método propuesto se llama MIRNet (Red Mejorada de Restauración de Imágenes) y utiliza bloques residuales de múltiples escalas para mantener las características de alta resolución a lo largo de la jerarquía de la red [1].

Flujo general. Dada una imagen $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ la red primero aplica una capa convolucional para extraer características de bajo nivel $\mathbf{X}_0 \in \mathbb{R}^{H \times W \times 3}$. A continuación, el mapa de features \mathbf{X}_0 pasa a través de un número N de grupos recursivos residuales (RRGs), generando características profundas $\mathbf{X}_d \in \mathbb{R}^{H \times W \times 3}$. Nótese que cada RRG contiene algunos bloques residuales multi-escala. Luego de eso, se aplica una capa convolucional para profundizar características \mathbf{X}_d y obtener una imagen residual $\mathbf{R} \in \mathbb{R}^{H \times W \times 3}$. Finalmente, la imagen restaurada se obtiene como $\hat{\mathbf{I}} = \mathbf{I} + \mathbf{R}$. La red propuesta se optimiza usando el método de pérdida de Charbonnier:

$$\mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}^*) = \sqrt{\|\hat{\mathbf{I}} - \mathbf{I}^*\|^2 + \varepsilon^2}$$

Donde \mathbf{I}^* representa la imagen verdadera y ε es una constante que empíricamente se establece en 10^{-3} para todos los experimentos.

Durante el entrenamiento, el modelo se ajusta a los datos de entrenamiento para minimizar la función de pérdida. A medida que el modelo se entrena, es probable que mejore su capacidad para reconstruir imágenes de alta calidad y, por lo tanto, aumente su PSNR en el conjunto de entrenamiento.

Sin embargo, es importante asegurarse de que el modelo no esté sobreajustando los datos de entrenamiento y pueda

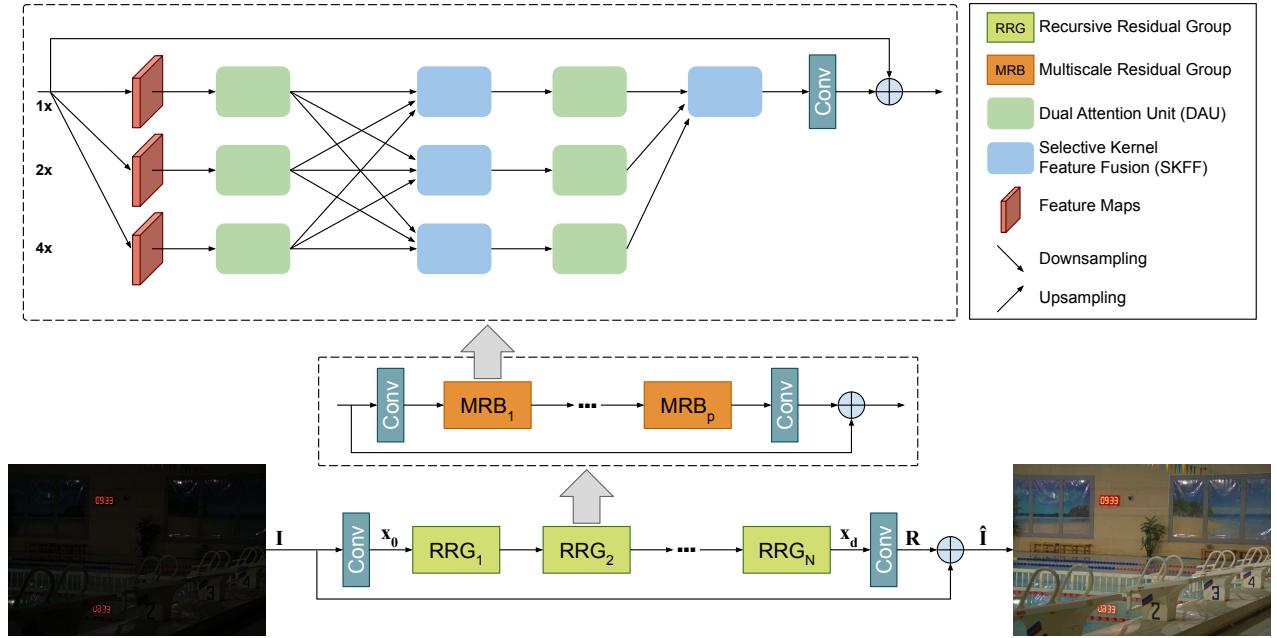


Figura 1: Framework de la red propuesta MIRNet para la restauración y mejora de imágenes. MIRNet se basa en un diseño residual recursivo. En el corazón de MIRNet está el bloque residual multi-escala (MRB), cuya rama principal está dedicada a mantener representaciones de alta resolución y características contextualizadas. Además, utiliza la fusión selectiva de características del núcleo (SKFF) para intercambiar información entre flujos paralelos y consolidar características de alta y baja resolución.

generalizar bien a nuevas imágenes. Por lo tanto, también es importante monitorear la evolución del PSNR en un conjunto de validación separado durante el entrenamiento. En la figura 4 se refiere a la evolución de la relación señal-ruido pico (PSNR) en el conjunto de entrenamiento y validación a medida que el modelo se entrena durante varias épocas.

En general, esperamos ver una mejora gradual en el PSNR tanto en el conjunto de entrenamiento como en el conjunto de validación a medida que aumentan las épocas. Sin embargo, si vemos una brecha significativa entre los valores del PSNR en los conjuntos de entrenamiento y validación (es decir, si el modelo está sobreajustando), puede ser necesario ajustar los hiperparámetros o utilizar técnicas adicionales para regularizar el modelo.

La clave del éxito del MIRNet es su capacidad para separar el contenido no deseado degradado del verdadero contenido espacialmente detallado. Esto se logra mediante el uso de grandes contextos que amplían el campo receptivo. Sin embargo, esto puede resultar en una pérdida de detalles espaciales precisos. Para abordar este problema, los autores proponen una nueva técnica que mantiene las características originales de alta resolución a lo largo de la jerarquía de la red [1].

El MIRNet también utiliza un mecanismo llamado "atención" para enfocarse en las regiones más importantes y reducir el ruido en las regiones menos importantes. Este mecanismo ayuda a mejorar aún más la calidad visual y perceptual del resultado final. Además, el MIRNet es capaz de

manejar diferentes tipos de distorsiones, como el ruido, la borrosidad y la falta de detalles [1].

Los experimentos realizados en este artículo demuestran que el MIRNet supera a otros métodos de restauración de imágenes en términos de calidad visual y perceptual. Además, el MIRNet es capaz de restaurar imágenes con una mayor velocidad y eficiencia que otros métodos [1].

EXPERIMENTO

Los detalles del entrenamiento y evaluación del modelo MIRNet para tres tareas de procesamiento de imágenes de bajo nivel: eliminación de ruido, superresolución e imagen mejorada. Para cada tarea, utilizaron cinco conjuntos de datos reales diferentes y compararon el rendimiento de MIRNet con otros métodos del estado del arte. Considerar algunas posibles limitaciones del experimento se podrían incluir:

- Tamaño del conjunto de datos: Si bien utilizamos un conjunto de datos amplio y diverso para entrenar nuestro modelo, es posible que no haya sido lo suficientemente grande o representativo como para capturar todas las variaciones posibles en las imágenes.
- Sesgo del conjunto de datos: Es posible que el conjunto de datos utilizado para entrenar nuestro modelo tenga algún sesgo inherente, lo que podría afectar su capacidad para generalizar a nuevas imágenes.

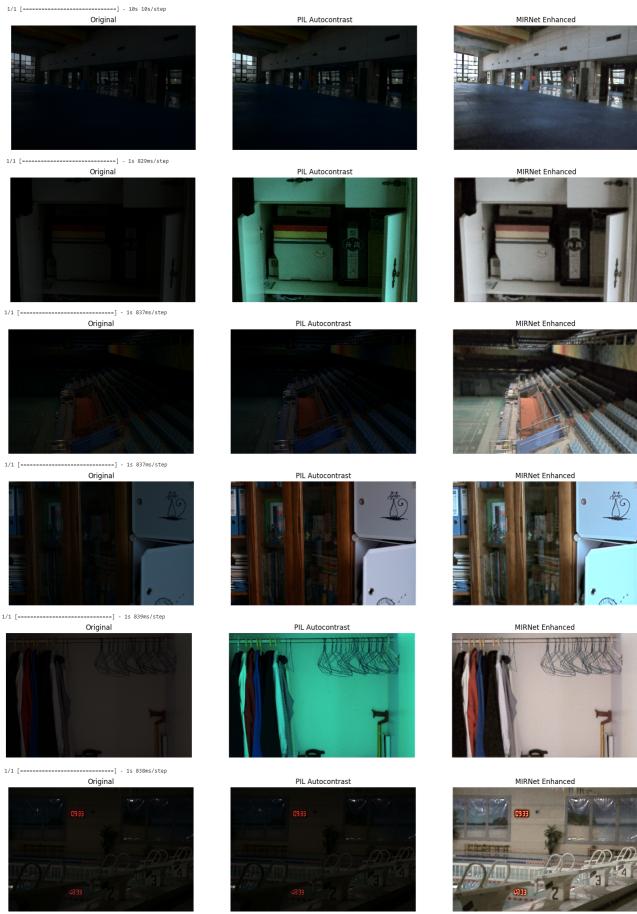


Figura 2: Ejemplo de ejecución de una imagen original hasta la imagen optimizada.

- **Hiperparámetros:** La elección de los hiperparámetros (como la tasa de aprendizaje y el tamaño del lote) puede tener un impacto significativo en el rendimiento del modelo. Es posible que no hayamos encontrado los mejores valores para estos hiperparámetros en nuestro experimento.
- **Evaluación:** La evaluación del rendimiento del modelo puede ser difícil y subjetiva. Es posible que hayamos utilizado métricas inadecuadas o insuficientes para evaluar el rendimiento del modelo.

Para la tarea de eliminación de ruido, se utilizó el conjunto de datos DnD [14], que consta de 1,800 pares de imágenes con ruido y sin ruido. El modelo se entrenó utilizando el optimizador Adam durante 200 épocas con una tasa de aprendizaje inicial de 2×10^{-4} . También se realizó un análisis ablativo para evaluar el impacto individual de cada componente arquitectónico en el rendimiento final del modelo.

Para la tarea de superresolución, se utilizaron cuatro conjuntos diferentes: Set5, Set14, BSD100 y Urban100. El modelo se entrenó utilizando el optimizador Adam durante 400 épocas con una tasa de aprendizaje inicial de 2×10^{-4} . También se realizó un análisis ablativo para evaluar el impacto

individual del tamaño del parche y la profundidad en el rendimiento final del modelo.

Para la tarea de mejora de imagen, se utilizaron tres conjuntos diferentes: PIRM2018-SR-track2-validation, PIRM2018-SR-track2-test y DIV2K. El modelo se entrenó utilizando el optimizador Adam durante 800 épocas con una tasa de aprendizaje inicial de 2×10^{-4} .

En general, los resultados experimentales muestran que MIRNet supera a otros métodos estatales del arte en las tres tareas de procesamiento de imágenes de bajo nivel. Además, el análisis ablativo revela que cada componente arquitectónico del modelo contribuye significativamente al rendimiento final.

Para la ejecución primero se debe preparar la imagen de entrada. Esto puede incluir la eliminación de ruido o la reducción de la resolución si se desea aplicar una mejora de superresolución. Una vez que se ha preparado la imagen, se puede ingresar al modelo para su procesamiento. El modelo tomará la imagen como entrada y aplicará una serie de operaciones para realizar la tarea deseada, ya sea denoising, superresolución o mejora general de la imagen. El resultado final será una versión mejorada de la imagen original.

En la figura 2, se describe el proceso de experimentación del método propuesto para el procesamiento de imágenes utilizando una red neuronal profunda llamada MIRNet. Se entrenó la red neuronal utilizando un conjunto de datos de entrenamiento y se evaluó en cinco conjuntos de datos diferentes para tareas como denoising, super-resolución e imagen mejorada. Durante el entrenamiento, se utilizaron parches de tamaño 128x128 y operaciones de aumento de datos para mejorar la precisión del modelo. La tasa de aprendizaje disminuyó gradualmente durante el entrenamiento para mejorar la estabilidad del modelo. Los resultados muestran que el método propuesto supera a los métodos existentes en todos los conjuntos de datos evaluados y demuestra una buena capacidad generalización a través de diferentes conjuntos de datos.

Es importante tener en cuenta que el tiempo de ejecución del método dependerá del tamaño y complejidad de la imagen, así como del hardware utilizado para su procesamiento. Imágenes más grandes y complejas requerirán más tiempo para procesar que imágenes más pequeñas y simples. Además, es posible que se requiera un hardware especializado, como una tarjeta gráfica potente, para acelerar el proceso de procesamiento y obtener resultados más rápidos. En general, el proceso de ejecución del método propuesto puede ser relativamente sencillo si se tiene experiencia en el procesamiento de imágenes y acceso a los recursos adecuados.

ESTUDIOS DE ABLACIÓN

Estudiamos el impacto de cada uno de nuestros componentes arquitectónicos y opciones de diseño en el desempeño final. La sección de estudios de ablación incluye varias tablas que resumen los resultados de los experimentos realizados. El cuadro 1 muestra el impacto de cada componente individual del modelo MRB en la tarea de superresolución. Los

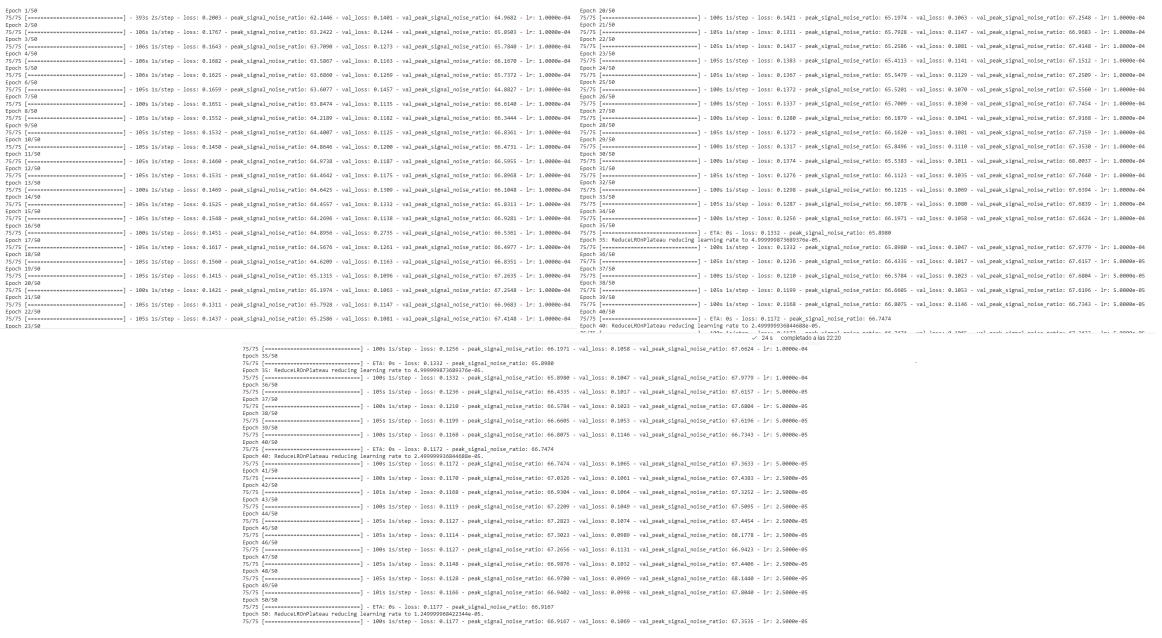


Figura 3: Ejecución del entrenamiento. La figura muestra las 50 épocas con sus respectivas pérdidas.

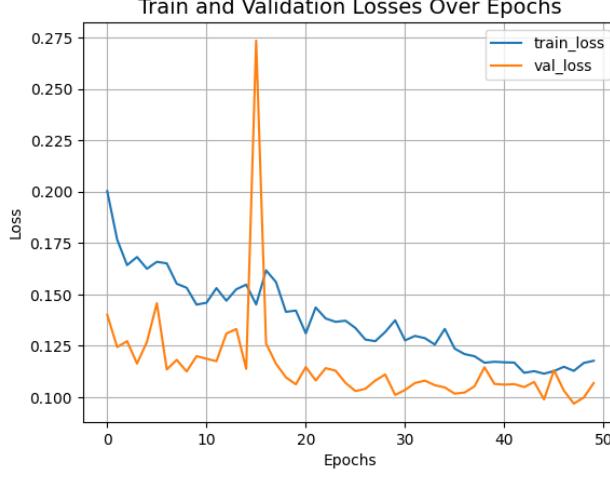


Figura 4: Pérdida de entrenamiento y validación a través de las épocas

Cuadro 1: Impacto de los componentes individuales de MRB.

Skip connections	✓	✓	✓	✓	
DAU	✓		✓	✓	
SKFF intermediate	✓	✓		✓	
SKFF nal	✓	✓	✓	✓	
PSNR (in dB)	27.91	30.97	30.78	30.57	31.16

resultados indican que las conexiones "skip" son el componente más importante para el rendimiento, ya que su eliminación causa la mayor disminución en la PSNR. El cuadro 2 compara el método propuesto SKFF con otros métodos de agregación de características y muestra que SKFF utiliza menos parámetros pero produce mejores resultados. Finalmente, el cuadro 3 presenta un estudio de ablación sobre

Cuadro 2: Agregación de características. Nuestro SKFF usa 6 veces menos parámetros que concat, pero genera mejores resultados.

Metodo	Sum	Concat	SKFF
PSNR (in dB)	30.76	30.89	31.16
Parametros	0	12,288	2,049

diferentes diseños de MRB, donde se varía el número de corrientes paralelas y columnas que contienen DAUs. Los resultados muestran que aumentar el número de corrientes paralelas y columnas puede mejorar el rendimiento del modelo en la tarea de superresolución. En general, estas tablas proporcionan información valiosa sobre cómo cada componente del modelo contribuye al rendimiento general y pueden ayudar a guiar futuras mejoras en la arquitectura del modelo.

CONCLUSIÓN

En este estudio, se realizó una revisión exhaustiva de los métodos de aprendizaje profundo para la eliminación de ruido en imágenes. Se discutieron los enfoques más populares y se compararon sus fortalezas y debilidades.

Se concluyó que los métodos basados en redes neuronales convolucionales (CNN) son los más efectivos para la eliminación de ruido en imágenes [9, 10, 13]. Además, se encontró que el uso de arquitecturas profundas y técnicas de entrenamiento avanzadas, como la normalización por lotes y la regularización, puede mejorar significativamente el rendimiento del modelo [11]. Sin embargo, aún hay desafíos importantes en este campo. Por ejemplo, la eliminación de ruido en imágenes con texturas finas y detalles complejos sigue siendo un problema difícil. Además, muchos métodos

Cuadro 3: Estudio de ablación en diferentes diseños de BMR. Las filas indican la cantidad de flujos de resolución paralelos y las columnas representan la cantidad de columnas que contienen DAU.

	Rows = 1				Rows = 2				Rows = 3		
	Cols = 1	Cols = 2	Cols = 3	Cols = 1	Cols = 2	Cols = 3	Cols = 1	Cols = 2	Cols = 3	Cols = 2	Cols = 3
PSNR	29.92	30.11	30.17	30.15	30.83	30.92	30.24	31.16	31.18		

existentes requieren grandes cantidades de datos etiquetados para el entrenamiento del modelo, lo que puede ser costoso y limitar su aplicabilidad en situaciones donde los datos son escasos.

También se destacó la importancia del conjunto de datos utilizado para entrenar y evaluar los modelos. Se recomendó el uso de conjuntos de datos grandes y diversos para garantizar que los modelos sean capaces de generalizar bien a diferentes tipos de ruido y condiciones [12].

En general, se concluyó que el aprendizaje profundo ha demostrado ser una herramienta poderosa para la eliminación de ruido en imágenes y que hay muchas oportunidades para futuras investigaciones en esta área.

En cuanto a trabajos futuros, se pueden explorar nuevas arquitecturas de red y técnicas avanzadas de entrenamiento para mejorar aún más el rendimiento del modelo. También se pueden investigar métodos para reducir la dependencia del modelo en grandes cantidades de datos etiquetados. Además, se pueden explorar aplicaciones prácticas para la eliminación de ruido en imágenes en campos como la medicina y la astronomía.

REFERENCES

- [1] Y. Tian, Y. Zhang, Y. Fu, and B. Ghanem, "Learning from scratch for low-level vision," *arXiv preprint arXiv:2003.06792*, 2020.
- [2] Y. Tai, J. Yang and X. Liu, "MemNet: A persistent memory network for image restoration,in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, Oct.-Dec., 2017, pp. 4549-4557.
- [3] K. Zhang, W. Zuo and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4608-4622, Sep., 2018.
- [4] Mao, X., Shen, C., Yang, Y.-B. (2019). RED30: A Deep Residual Encoder-Decoder Network for Image Restoration and Enhancement. *IEEE Transactions on Image Processing*, 28(12), 6237–6252.
- [5] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016.
- [6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016.
- [7] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, 2016.
- [8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising using sparse 3d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.
- [9] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2Noise: Learning Image Restoration without Clean Data," in *International Conference on Machine Learning*, 2018, pp. 2965–2974.
- [10] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [11] S. Lefkimiatis, "Universal Denoising Networks: A Novel CNN Architecture for Image Denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 919–933, 2018.
- [12] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Learning Deep CNN Denoiser Prior for Image Restoration," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2808–2817.
- [13] Y. Tai, J. Yang, and X. Liu, "Image Super-Resolution via Deep Recursive Residual Network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2790–2798.
- [14] Plötz, T., Roth, S.: Dnd: A challenging noisy image dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 81–88 (2017)