DISEÑO DE UN LENTE PARA GENERAR DESCRIPCIONES DE IMÁGENES PRESERVANDO SU PRIVACIDAD

PAULA ANDREA ARGUELLO GUTIÉRREZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2024

DISEÑO DE UN LENTE PARA GENERAR DESCRIPCIONES DE IMÁGENES PRESERVANDO SU PRIVACIDAD

PAULA ANDREA ARGUELLO GUTIÉRREZ

Trabajo de Grado para optar al título de
Ingeniera de Sistemas

Director:
Hoover Fabián Rueda Chacón
Ph.D. en Ingeniería Eléctrica y Computación

Codirectora:
Karen Yaneth Sánchez Quiroga
Ph.D. en Ingeniería

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2024

DEDICATORIA

A mi papá Henry, mi ejemplo; por su amor y por la educación que me brindó,
A mis mamás Lany y Ofelia, por su cariño y soporte,
A Roma por su gran afecto, colaboración y compañía,
A mi Coco por su acompañamiento bajo el escritorio cada día,
Este trabajo de grado es resultado del amor, respaldo, paciencia y soporte emocional que ellos me han ofrecido.

Y al grupo HDSP, mi segunda familia desde hace mucho tiempo.
Hicieron mi formación académica y personal posible.

AGRADECIMIENTOS

Agradezco a mi papá Henry y a mis mamás Ofelia y Lany, mi hermosa familia, por su soporte incondicional.

A Roma, por su cariño, gran amor y apoyo, durante estos años académicos.

A mi directora Karen Sanchez, que durante estos últimos años se convirtió en un ejemplo a seguir y le agradezco todo el tiempo dedicado.

A mi director Hoover Rueda-Chacón, por su compromiso, tiempo, paciencia, y formación tanto profesional como personal.

A el grupo de investigación HDSP, y a quienes fueron parte, por permitirme aprender de ellos diariamente.

Finalmente a mis amigos cercanos por los buenos momentos, a las secretarias de la Escuela de Sistemas por su colaboración y paciencia, y a los profesores que dejaron una huella en mi formación académica a lo largo de estos años.

CONTENIDO

LISTA DE FIGURAS

pág.

LISTA DE CUADROS

pág.

RESUMEN

TÍTULO: DISEÑO DE UN LENTE PARA GENERAR DESCRIPCIONES DE IMÁGENES PRESERVANDO SU PRIVACIDAD [*]

AUTOR: PAULA ANDREA ARGUELLO GUTIÉRREZ [**]

PALABRAS CLAVE: Preservación de la privacidad, Visión por computadora, Procesamiento de lenguaje natural, Descripción de imágenes, Diseño de óptica.

DESCRIPCIÓN:

La generación de descripciones de imágenes consiste en resumir textualmente el contenido visual de una imagen. Esta tarea ha ganado popularidad en el cruce de dos áreas de la inteligencia artificial: visión por computadora y procesamiento de lenguaje natural. No obstante, en el enfoque convencional de descripción de imágenes, se utilizan imágenes de alta resolución para entrenar los modelos. Estas imágenes pueden incluir datos sensibles que deberían ser confidenciales, tales como rostros, características personales, documentos, menores de edad, etc., los cuales podrían estar sujetos a riesgos de privacidad. Este trabajo de grado se centra en proteger la privacidad en el proceso de descripción de imágenes, enfocándose directamente en la óptica antes de la adquisición de las imágenes. Dado la tendencia emergente de integrar el diseño óptico con la inteligencia artificial, se diseñó un lente refractivo para garantizar la privacidad. El lente optimizado oculta atributos visuales sensibles en la imagen adquirida, al tiempo que extrae características esenciales para las descripciones incluso a partir de imágenes muy distorsionadas. Con un enfoque de extremo a extremo, se logró un sistema capaz de crear descripciones directamente de imágenes distorsionadas mediante la optimización de este lente, junto con el desarrollo de una arquitectura de redes neuronales profundas para la generación de descripciones de imágenes. Este método fue probado y validado a través de simulaciones y experimentos en el laboratorio. Los resultados demostraron un mejor equilibrio entre privacidad y utilidad comparado con métodos tradicionales que no consideran la privacidad en diversos conjuntos de datos.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Hoover Fabián Rueda Chacón. Codirector: Karen Yaneth Sánchez Quiroga.

ABSTRACT

TITLE: LENS DESIGN FOR ENHANCING PRIVACY IN IMAGE CAPTIONING [*]

AUTHOR: PAULA ANDREA ARGUELLO GUTIÉRREZ [**]

KEYWORDS: Privacy-preservation, Computer Vision, Natural Language Processing, Image Captioning, Deep Optics.

DESCRIPTION:

Image caption generation consists of textually summarizing the visual content of an image. This task has gained popularity at the turning point of two areas of artificial intelligence: computer vision and natural language processing. However, high-resolution images are used to train the models in the conventional image captioning approach. These images may include sensitive data that should be confidential, such as faces, personal characteristics, documents, children, etc., which could be subject to privacy risks. This work focuses on protecting privacy in the image captioning process, approaching it directly from the image acquisition. Given the emerging trend of integrating optical design with artificial intelligence, a refractive lens was designed to ensure privacy. The optimized lens hides sensitive visual attributes in the acquired image while extracting essential features for captions even from highly distorted images. With an end-to-end approach, a system capable of creating captions directly from distorted images was achieved by optimizing this lens, together with the development of a deep neural network architecture to generate image captions. This method was tested and validated through simulations and real laboratory experiments. Results showed a better balance between privacy and usability compared to traditional methods that do not consider privacy in various datasets.

---

[*] Bachelor's Thesis

[**] Faculty of Physical-Mechanical Engineering. School of Systems Engineering & Informatics. Advisor: Hoover Fabián Rueda Chacón. Co-advisor: Karen Yaneth Sánchez Quiroga

INTRODUCCIÓN

La generación de descripciones es el proceso de crear textos cortos informativos para imágenes, usando lenguaje natural, que relacionan el contenido visual y el contexto de una imagen, para informar el contenido a los espectadores. Este proceso mejora la accesibilidad para personas con discapacidad visual permitiendo la conversión de dichas descripciones en audios[1], también facilita la búsqueda de imágenes[2] y simplifica el resumen de contenido. Además, potencia a los asistentes virtuales y sistemas de inteligencia artificial, enriqueciendo los materiales educativos[3] y mejorando la comunicación en redes sociales[4]. Sin embargo, la generación de descripciones para imágenes es significativamente más compleja que tareas como la clasificación de imágenes o el reconocimiento de objetos[5], ya que se requiere describir no solo los objetos presentes, sino también cómo se relacionan entre ellos, sus atributos y acciones asociadas.

---

[1] Madhuri Bhalekar y Mangesh Bedekar. «D-CNN: A new model for generating image captions with text extraction using deep learning for visually challenged individuals». En: *Engineering, Technology & Applied Science Research* (2022).

[2] Maurizio Leotta, Fabrizio Mori y Marina Ribaudo. «Evaluating the effectiveness of automatic image captioning for web accessibility». En: *Universal access in the information society* (2023).

[3] Mohammad Nehal Hasnine et al. «Vocabulary learning support system based on automatic image captioning technology». En: *Distributed, Ambient and Pervasive Interactions: 7th International Conference, DAPI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*. Springer. 2019.

[4] Kurt Shuster et al. «Engaging Image Captioning via Personality». En: *Proceedings of the IEEE/CVF Conference on CVPR*. Jun. de 2019; Haley MacLeod et al. «Understanding blind people's experiences with computer-generated captions of social media images». En: *proceedings of the 2017 CHI CAFCS*. 2017.

[5] Qingzhong Wang, Jia Wan y Antoni B Chan. «On diversity in image captioning: Metrics and methods». En: *IEEE TPAMI* (2020).

El problema de generación de descripciones para imágenes ha sido abordado usando varias técnicas. Modelos actuales aprovechan redes neuronales convolucionales (CNN, por sus siglas en inglés) para extraer características de imágenes y redes de memoria a corto plazo de larga duración (LSTM, por sus siglas en inglés) que son capaces de procesar secuencias completas de datos para generar descripciones palabra por palabra[6]. Este enfoque ha mejorado sustancialmente la relevancia contextual y la coherencia de las descripciones generadas[7],[8]. Además, se han integrado mecanismos de atención en estos modelos, permitiéndoles enfocarse en regiones específicas de la imagen[9]. Esto mejora la calidad de las descripciones haciéndolas más estrechamente vinculadas al contenido visual, cerrando la brecha entre la imagen y las descripciones de lenguaje generadas.

En un proceso tradicional de visión por computadora, una cámara se utiliza para adquirir múltiples imágenes que son usadas para entrenar redes neuronales profundas. Sin embargo, las imágenes pueden contener información privada sensible, generando preocupaciones sobre los riesgos asociados de explotación indebida[10]. Esto ha incrementado la demanda de métodos de protección de privacidad visual en

---

6 Lirong Yao y Yazhuo Guan. «An Improved LSTM Structure for Natural Language Processing». En: *2018 IEEE IICSPI*. 2018.

7 Kelvin Xu et al. «Show, attend and tell: Neural image caption generation with visual attention». En: *ICML*. PMLR. 2015.

8 Oriol Vinyals et al. «Show and tell: A neural image caption generator». En: *IEEE/CVF Conference on CVPR*. 2015.

9 Lun Huang et al. «Attention on attention for image captioning». En: *Proceedings of the IEEE/CVF ICCV*. 2019.

10 Karl Manheim y Lyric Kaplan. «Artificial intelligence: Risks to privacy and democracy». En: *Yale JL & Tech.* (2019).

diferentes campos, como la salud[11], multimedia[12], redes sociales[13] y vigilancia por video[14]. También es crucial en entidades gubernamentales, de entretenimiento y manufactura para la seguridad de datos y protección de propiedad intelectual[15]. Los avances en óptica y algoritmos[16] han propiciado el desarrollo de sistemas integrales que permiten preservar la privacidad mientras se desarrollan aplicaciones como la estimación de pose humana[17] y el reconocimiento de acciones[18]. Asimismo, existen métodos tanto de software como de hardware específicamente diseñados para preservar la privacidad.

Para generar descripciones de imágenes manteniendo la privacidad, los autores en[19] utilizan un enfoque de encriptación para las imágenes. Aunque este método es

---

[11] Asokan Sivaprakash, Samuel NE Rajan y Sundaramoorthy Selvaperumal. «Privacy protection of patient medical images using digital watermarking technique for E-healthcare system». En: *Current Medical Imaging* (2019).

[12] Jinao Yu et al. «Gan-based differential private image privacy protection framework for the internet of multimedia things». En: *Sensors* (2020).

[13] Chi Liu et al. «Privacy intelligence: A survey on image privacy in online social networks». En: *ACM Computing Surveys* (2022).

[14] Ling Du et al. «An efficient privacy protection scheme for data security in video surveillance». En: *Journal of VCIR* (2019).

[15] Siddharth Ravi, Pau Climent-Pérez y Francisco Florez-Revuelta. «A review on visual privacy preservation techniques for active and assisted living». En: *Multimedia Tools and Applications* (2023).

[16] Vincent Sitzmann et al. «End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging». En: *ACM Transactions on Graphics (TOG)* (2018).

[17] Carlos Hinojosa, Juan Carlos Niebles y Henry Arguello. «Learning Privacy-preserving Optics for Human Pose Estimation». En: *IEEE/CVF ICCV*. 2021.

[18] Carlos Hinojosa et al. «Privhar: Recognizing human actions from privacy-preserving lens». En: *ECCV*. Springer. 2022.

[19] Antoinette Deborah Martin, Ezat Ahmadzadeh e Inkyu Moon. «Privacy-Preserving Image Captioning with Deep Learning and Double Random Phase Encoding». En: *Mathematics* (2022).

efectivo en la privacidad, el desempeño y precisión de la generación de descripciones disminuye significativamente en comparación con enfoques que no preservan la privacidad. En[20] se propone un enfoque alternativo para generar descripciones de imágenes preservando la privacidad, centrado en imágenes de ingesta dietética con rostros enmascarados manualmente, evitando que otras personas accedan a la identidad de los pacientes. Por lo tanto, la preservación de la privacidad se enfoca en evitar el uso de imágenes y utilizar únicamente descripciones para los análisis médicos. Para ello, entrenan una red neuronal profunda usando imágenes en las que las caras han sido enmascaradas previamente. Se destaca que, en los enfoques basados en software, todavía pueden existir vulnerabilidades durante la etapa de adquisición de datos.

Actualmente, existen diversas estrategias para la protección de la privacidad implementadas en hardware. La mayoría de estos métodos incluyen la adición de un componente óptico durante la adquisición de imágenes. Trabajos previos utilizan técnicas de desenfoque que ofrecen un grado de privacidad para una región restringida por el tamaño del sensor de la cámara[21],[22]. En trabajos más actuales, desarrollaron una solución de privacidad en el ámbito del hardware para sistemas de visión por computadora, utilizando técnicas de aprendizaje profundo aplicadas a la óptica[23]. Este método hace uso de un enfoque de extremo a extremo (End-to-End) que implica el diseño de un lente refractivo que difumina selectivamente información sensible,

---

[20] Jianing Qiu et al. «Egocentric image captioning for privacy-preserved passive dietary intake monitoring». En: *IEEE Transactions on Cybernetics* (2023).

[21] Francesco Pittaluga y Sanjeev J Koppal. «Privacy preserving optics for miniature vision sensors». En: *Proceedings of the IEEE Conference on CVPR*. 2015.

[22] Francesco Pittaluga y Sanjeev Jagannatha Koppal. «Pre-capture privacy for small vision sensors». En: *IEEE TPAMI* (2016).

[23] Sitzmann et al., ver n. 16; Hinojosa, Niebles y Arguello, ver n. 17; Hinojosa et al., ver n. 18.

al mismo tiempo que preserva la funcionalidad necesaria para aplicaciones de visión por computadora, como el reconocimiento de gestos[24] y la estimación de posturas humanas[25].

Por lo tanto, el objetivo de esta investigación es diseñar un lente refractivo especializado que permite generar descripciones de imágenes que han sido protegidas directamente desde su adquisición. Este enfoque integra el aprendizaje de extremo a extremo (End-to-End) del sistema óptico para optimizar la captura de imágenes, preservando la privacidad mientras mantiene la utilidad práctica de las descripciones generadas. El lente refractivo realiza la distorsión de la escena y permite la extracción adecuada de características para la descripción de imágenes.

En este trabajo se realizó una implementación en el laboratorio de la arquitectura propuesta, demostrando su viabilidad para aplicaciones en el mundo real. Se llevaron a cabo estudios de ablación extensos y se validó el enfoque en dos conjuntos de datos, evaluando también la robustez del método ante ataques adversarios como la deconvolución. Esta evaluación mostró la capacidad del método para resistir intentos adversarios de recuperar información sensible, asegurando un desempeño confiable en condiciones realistas. Finalmente, se realizó una comparación exhaustiva entre el método propuesto y otros enfoques existentes en la literatura, destacando los beneficios y la competitividad en la preservación de la privacidad.

---

[24] Hinojosa et al., ver n. 18.

[25] Hinojosa, Niebles y Arguello, ver n. 17.

1. OBJETIVOS

Objetivo general

Diseñar un lente refractivo para su implementación en un montaje óptico que preserve la privacidad en imágenes durante su adquisición, mientras que se conservan las características esenciales para la generación de descripciones en imágenes distorsionadas.

Objetivos específicos

1. Diseñar un esquema para la optimización de un lente refractivo que preserve la privacidad en conjunto con la tarea de descripción de imágenes usando un enfoque de redes neuronales extremo a extremo.

2. Implementar en software la red neuronal propuesta para obtener un diseño del lente que distorsione las imágenes y permita una descripción de la imagen.

3. Evaluar la robustez del método propuesto frente a ataques que intenten recuperar la información sensible, demostrando su capacidad para prevenir su recuperación no autorizada.

4. Realizar un análisis comparativo entre el método propuesto y técnicas existentes a nivel de software para la generación de descripciones de las imágenes, mediante simulaciones en diferentes conjuntos de datos.

5. Implementar en el laboratorio la arquitectura óptica con el lente optimizado para validar su uso en un entorno real.

2. MARCO DE REFERENCIA

2.1. Descripción de imágenes

La tarea de generación de descripciones para imágenes implica la creación de textos cortos informativos usando lenguaje natural con el propósito de relacionar el contenido visual y el contexto de una imagen[26], como se observa en la Figura 1[27]. Este proceso tiene como objetivo facilitar la comprensión de la imagen tanto para humanos como para sistemas de inteligencia artificial[28]. La generación de descripciones para imágenes abarca diversas aplicaciones, entre las que se incluyen mejorar la accesibilidad para personas con discapacidad visual[29], enriquecer la experiencia del usuario[30], facilitar la búsqueda de imágenes y resumir contenido visual[31]. A lo largo de los años, la tarea de descripción de imágenes se ha abordado de diversas formas. Algunos trabajos utilizan modelos de inteligencia artificial para identificar regiones clave en las imágenes y extraer atributos que describir.[32], mientras que otros emplean arquitecturas que combinan logros recientes en visión por computadora y

---

[26] Raimonda Staniūtė y Dmitrij Šešok. «A Systematic Literature Review on Image Captioning». En: *Applied Sciences* (2019).

[27] Tsung-Yi Lin et al. «Microsoft coco: Common objects in context». En: *ECCV*. Springer. 2014.

[28] Leotta, Mori y Ribaudo, ver n. 2.

[29] Bhalekar y Bedekar, ver n. 1.

[30] MacLeod et al., ver n. 4.

[31] Qi Wu et al. «Image captioning and visual question answering based on attributes and external knowledge». En: *IEEE TPAMI* (2017).

[32] Huang et al., ver n. 9.

traducción automática para crear oraciones naturales que describen la imagen[33].

a small girl sitting on a chair holding a white bear

a gray and white cat drinking milk from a bottle

a man helps a disabled baseball player on the mound

Figura 1. Ejemplos de descripciones cortas informativas en inglés de imágenes del conjunto de datos Common Objects in Context (COCO)[27].

2.2. Aprendizaje profundo

La tarea de descripción de imágenes se ha abordado principalmente mediante el aprendizaje profundo, cuyo impacto significativo se evidencia en la solución de diversos problemas. Los modelos de aprendizaje profundo se componen de capas interconectadas que consisten en unidades denominadas neuronas[34]. Estas unidades, mediante combinaciones lineales de datos de entrada, ajustan iterativamente sus pesos para minimizar el error entre las etiquetas reales y las predicciones del modelo, construyendo así una representación abstracta del conjunto de datos. Estos modelos demuestran su eficacia en diversas tareas, tales como clasificación[35],

---

[33] Vinyals et al., ver n. 8.

[34] Ian Goodfellow, Yoshua Bengio y Aaron Courville. *Deep learning*. MIT press, 2016.

[35] Dan Ciregan, Ueli Meier y Jürgen Schmidhuber. «Multi-column deep neural networks for image classification». En: *2012 IEEE Conference on CVPR*. IEEE. 2012.

detección de objetos[36], y generación de descripciones para imágenes[37], abarcando un amplio espectro de aplicaciones en el procesamiento de imágenes.

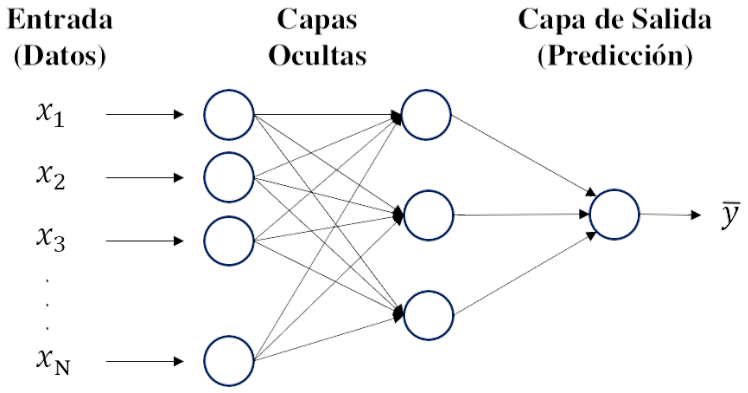

Figura 2. Ejemplo de red neuronal con capas densas. Imagen adaptada de[35].

En el contexto de un aprendizaje supervisado con un conjunto de datos $\{(x_i, y_i)\}_{i=1}^{N}$, donde x_i es un dato o imágen, y_i es su correspondiente etiqueta, y N es el número de ejemplos de entrenamiento, se introduce la función de costo:

$$J(h_\theta(x_i), y_i), \tag{1}$$

donde $h_\theta(x_i)$ es la predicción del modelo para el i-ésimo ejemplo y y_i es la etiqueta real correspondiente. La arquitectura de red neuronal profunda, tal como se puede observar en la Figura 2, es una concatenación de capas y neuronas que emplea combinaciones lineales y funciones no lineales para procesar las entradas, estableciendo así una estructura efectiva para el aprendizaje profundo. En el ámbito de la descripción de imágenes, las etiquetas corresponden a descripciones breves, textos

---

[36] Christian Szegedy, Alexander Toshev y Dumitru Erhan. «Deep neural networks for object detection». En: *Advances in neural information processing systems* (2013).

[37] Xiaoxiao Liu, Qingyang Xu y Ning Wang. «A survey on deep neural network-based image captioning». En: *The Visual Computer* (2019).

de menos de 10 palabras, pudiendo haber más de una etiqueta por imagen[38].

2.2.1. Red neuronal convolucional (CNN)

A medida que el aprendizaje profundo y las redes neuronales han avanzado, emerge una arquitectura centrada en el procesamiento de imágenes, conocida como red neuronal convolucional o CNN por sus siglas en inglés. Esta arquitectura permite el análisis avanzado y la interpretación automática de contenido visual, simplificando la manera en que las máquinas comprenden las imágenes[39]. Para esto, se emplean las capas convolucionales, que utilizan operaciones de convolución para procesar información espacial.

La convolución consiste en deslizar un filtro o "kernel" para abordar simultáneamente un conjunto de píxeles. Esto permite que la red aprenda automáticamente patrones y características locales, como bordes, texturas y detalles, en lugar de depender de características globales[40]; un ejemplo de una CNN se puede ver en la Figura 3.

La optimización y actualización de estos parámetros se ajustan de manera específica según la tarea que se desea realizar. La esencia de las redes convolucionales radica en su capacidad para extraer características clave de las imágenes, lo que las hace fundamentales en la tarea de revelar patrones visuales complejos[41].

2.2.2. Red neuronal residual (ResNet)

Las redes neuronales residuales (ResNet, por sus siglas en inglés de *Res*idual

---

[38] Staniūtė y Šešok, ver n. 26.

[39] Zewen Li et al. «A survey of convolutional neural networks: analysis, applications, and prospects». En: *IEEE TNNLS* (2021).

[40] Salman Khan et al. *A guide to convolutional neural networks for computer vision*. Springer, 2018.

[41] Thomas Wiatowski y Helmut Bölcskei. «A mathematical theory of deep convolutional neural networks for feature extraction». En: *IEEE TIT* (2017).

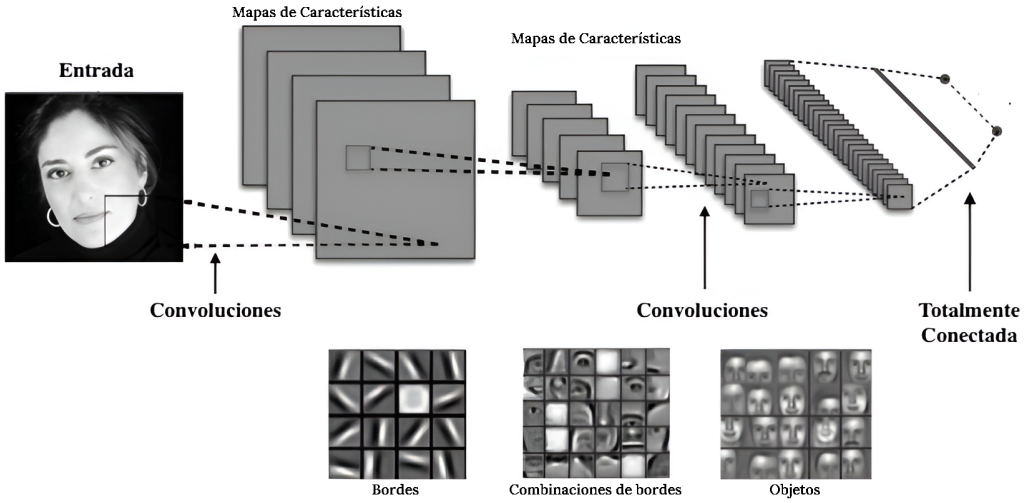

Figura 3. Ejemplo de una red neuronal convolucional. Imagen adaptada de[38].

*Net*works) son arquitecturas de CNN en las que se emplean funciones residuales en vez de las funciones directas tradicionales[42]. Además, incorporan conexiones residuales que permiten que la señal de entrada de una capa en la red neuronal se sume directamente a su salida. Así, frente a un mapeo tradicional de una CNN $\mathcal{H}(x_i)$, se propone una función residual de la forma $\mathcal{F}(x_i) = \mathcal{H}(x_i) - x_i$, y esta se mapea a una conexión residual $\mathcal{F}(x_i) + x_i$. Este enfoque es óptimo para redes con decenas o centenares de capas, permitiendo un entrenamiento más preciso al escalar en profundidad, aliviando así el problema del desvanecimiento del gradiente[43].

2.2.3. Red de memoria a corto plazo de larga duración (LSTM)

Una red de memoria a corto plazo de larga duración (LSTM por sus siglas en inglés, *Long Short Term Memory*), representa un tipo de Red Neuronal Recurrente

---

[42] Kaiming He et al. «Deep residual learning for image recognition». En: *Proceedings of the IEEE Conference on CVPR*. 2016.

[43] Sepp Hochreiter et al. *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. 2001.

(RNN) que, a diferencia de las redes neuronales profundas, tienen conexiones cíclicas en su arquitectura, lo que les permite mantener y utilizar información previa en la secuencia para tomar decisiones en el presente[44]. Esto las hace especialmente adecuadas para tareas que involucran datos secuenciales, como traducción automática, procesamiento de lenguaje natural (*NLP* por sus siglas en inglés de *Natural Language Processing*) y análisis de series temporales.

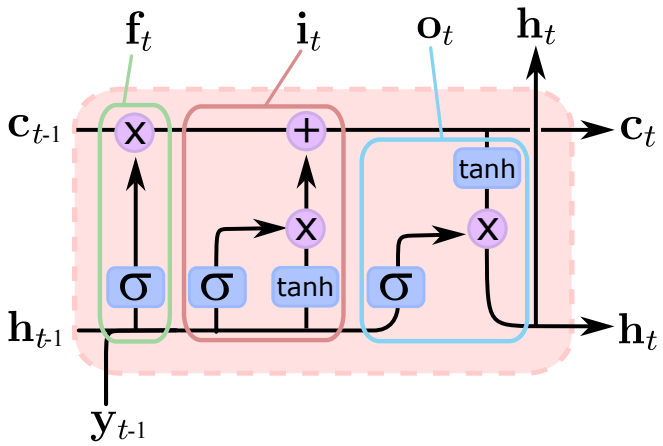

Figura 4. Ejemplo de una red LSTM.

Las LSTMs abordan eficazmente el desafío inherente a las secuencias temporales extensas mediante una estructura compleja que se basa en celdas de memoria capaces de retener y olvidar información a lo largo del tiempo. La estructura de una red LSTM se observa en la Figura 4. La descripción matemática de estas celdas se puede expresar de la siguiente manera:

---

[44] Larry R Medsker y LC Jain. «Recurrent neural networks». En: *Design and Applications* (2001).

$$
\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{T}_i \mathbf{y}_{t-1} + \mathbf{K}_i \mathbf{h}_{t-1} + \mathbf{b}_i\right), \\
\mathbf{f}_t &= \sigma\left(\mathbf{T}_f \mathbf{y}_{t-1} + \mathbf{K}_f \mathbf{h}_{t-1} + \mathbf{b}_f\right), \\
\mathbf{c}_t &= \mathbf{f}_t \mathbf{c}_{t-1} + \mathbf{i}_t \tanh\left(\mathbf{T}_c \mathbf{y}_{t-1} + \mathbf{K}_c \mathbf{h}_{t-1} + \mathbf{b}_c\right), \\
\mathbf{o}_t &= \sigma\left(\mathbf{T}_o \mathbf{y}_{t-1} + \mathbf{K}_o \mathbf{h}_{t-1} + \mathbf{b}_o\right), \\
\mathbf{h}_t &= \tanh\left(\mathbf{c}_t\right) \mathbf{o}_t.
\end{aligned}
\tag{2}
$$

En esta formulación, los datos de entrada $\mathbf{y}_t \in \mathbb{R}^n$ son un vector de características en el tiempo t, que recibe la LSTM. La variable $\mathbf{i}_t \in \mathbb{R}^n$ que corresponde al estado de entrada, $\mathbf{f}_t \in \mathbb{R}^n$ estado de olvido, $\mathbf{c}_t \in \mathbb{R}^n$ celda, $\mathbf{o}_t \in \mathbb{R}^n$ salida y $\mathbf{h}_t \in \mathbb{R}^n$ estado oculto de la red LSTM. El subíndice t es equivalente al tiempo actual en la secuencia de datos que la red está procesando. Las matrices $\mathbf{T}_s, \mathbf{K}_s \in \mathbb{R}^{n \times m}$ representan las matrices de pesos aprendidas, y los vectores $\mathbf{b}_s \in \mathbb{R}^n$ denotan los vectores de sesgo. Aquí, el subíndice $s = \{i, f, c, o\}$ indica qué variable se calcula a partir de estas matrices y sesgos aprendidos (por ejemplo, $\mathbf{K}_i$ se utiliza para calcular la entrada).

Las funciones de activación empleadas incluyen la función sigmoide denotada como $\sigma(x) = \frac{1}{1+e^{-x}}$, la cual es útil para convertir valores de entrada arbitrarios en probabilidades dado que su salida esta en el rango $(0, 1)$. En este contexto, un valor de 0 indica que la información puede ser olvidada, y un valor de 1 sugiere que debe retenerse, actuando como un mecanismo de regulación de la importancia de la información de entrada. Del mismo modo, se hace uso de la función tangente hiperbólica expresada como $\tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$, cuyo rango es de $(-1, 1)$. Esto permite que la red maneje tanto valores positivos como negativos, lo cual es útil para regular el flujo de información[45].

---

[45] Shiv Ram Dubey, Satish Kumar Singh y Bidyut Baran Chaudhuri. «Activation functions in deep learning: A comprehensive survey and benchmark». En: *Neurocomputing* (2022).

El estado de olvido decide qué información es irrelevante y puede ser descartada, permitiendo a la red olvidarse de los datos antiguos que ya no son útiles. El estado de entrada actualiza el estado de la celda con nueva información, y el estado de salida determina qué parte de la información actual de la celda debe pasar a la salida.

Esta habilidad para gestionar la memoria es crucial en tareas como la generación de descripciones de imágenes, donde la red debe concentrarse en los detalles relevantes de la imagen mientras ignora los irrelevantes. Por ejemplo, al describir una imagen, la LSTM puede retener información sobre los sujetos principales en una imagen y cómo se relacionan entre sí, mientras olvida detalles de fondo menos importantes, como el color exacto del cielo o la presencia de objetos que no influyen en la imagen principal.

2.2.4. Red de transformadores

Las redes de transformadores (más conocidas como *Transformers* en inglés) son una clase de modelos de aprendizaje automático diseñados principalmente para tareas de procesamiento de secuencias, como el NLP y la traducción automática. Los transformadores[46] representan una mejora sobre los modelos basados en RNNs y CNNs, especialmente en tareas que requieren el manejo de secuencias largas de datos.

Los transformadores se basan en un mecanismo central llamado *atención*[47], que permite al modelo enfocarse en diferentes partes de la entrada al generar cada parte de la salida. Este mecanismo habilita que todas las partes de la entrada se

---

[46] Salman Khan et al. «Transformers in vision: A survey». En: *ACM Computing Surveys (CSUR)* (2022).

[47] Ashish Vaswani et al. «Attention is all you need». En: *Advances in neural information processing systems* (2017).

procesen en paralelo, lo que permite la capacidad de captar dependencias a largo plazo en los datos.

Dentro del mecanismo de atención, el módulo más conocido es el de atención suave (*soft attention* en inglés), que calcula un peso para cada parte de la secuencia de entrada, determinando cuanta relevancia tiene cada fragmento al generar una porción específica de la secuencia de salida. A diferencia de la atención fuerte (*hard attention* en inglés) que selecciona una única parte de la entrada, la atención suave es diferenciable, lo que significa que los pesos pueden aprenderse y ajustarse mediante retropropagación durante el entrenamiento del modelo. El mecanismo de atención suave se calcula de la siguiente manera,

$$a_{t,i} = f\left(\boldsymbol{h}_t, \boldsymbol{x}_i\right). \tag{3}$$

Esta función calcula la puntuación de atención entre el estado oculto en el tiempo t y algún otro estado oculto o entrada $\boldsymbol{x}_i$. Esta función puede variar desde un producto punto hasta una pequeña red neuronal. Una vez calculada la puntuación de atención se calcula la función *softmax* para obtener los pesos de atención normalizados, de la siguiente manera

$$\begin{aligned} \alpha_{t,i} &= \mathrm{softmax}(\alpha_{t,i}), \\ &= \frac{\exp\left(a_{t,i}\right)}{\sum_{k=1}^{K} \exp\left(a_{t,k}\right)}. \end{aligned} \tag{4}$$

Estos pesos se utilizan para crear un vector de contexto ponderado que se pasa a la siguiente capa del modelo o se usa para generar la salida dependiendo de la tarea.

2.3. Protección de la privacidad en imágenes

Aunque la inteligencia artificial ha proporcionado diversas estrategias para abordar tareas en campos como la visión por computadora, el NLP y muchas otras áreas,

existen problemas de privacidad asociados a los usuarios presentes en las imágenes de entrada proporcionadas. Además, la mayoría de los métodos dependen de imágenes de alta resolución[48].

La integración de mecanismos de preservación de la privacidad en el ámbito de la visión por computadora es un área de estudio relativamente reciente pero de gran importancia. Esta atención se debe a la necesidad de respetar la privacidad de las personas representadas en las imágenes que se procesan.

2.3.1. Protección de la privacidad a nivel de software

La mayoría de los métodos de preservación de la privacidad en visión por computadora solo ofrecen protección de privacidad a nivel de software, los cuales dependen de conocimientos específicos del dominio y enfoques hechos a medida, incluyendo pixelación, desenfoque y reemplazo de rostros/objetos, para proteger información sensible[49,50]. Esto puede ser útil en entornos prácticos cuando sabemos de antemano qué proteger en la imagen. Trabajos recientes proponen un enfoque más general que aprende codificaciones preservadoras de privacidad[51,52]. Estos métodos aprenden activamente a degradar o inhibir atributos privados mientras mantienen características importantes para realizar tareas de inferencia. Aunque estos enfoques a nivel de software preservan la privacidad en la aplicación final, las imá-

---

48 Manheim y Kaplan, ver n. 10.

49 Prachi Agrawal y PJ Narayanan. «Person de-identification in videos». En: *IEEE TCSVT* (2011).

50 José Ramón Padilla-López, Alexandros Andre Chaaraoui y Francisco Flórez-Revuelta. «Visual privacy protection methods: A survey». En: *Expert Systems with Applications* (2015).

51 Francesco Pittaluga, Sanjeev Koppal y Ayan Chakrabarti. «Learning privacy preserving encodings through adversarial training». En: *2019 IEEE WACV*. IEEE. 2019.

52 Zhenyu Wu et al. «Privacy-Preserving Deep Action Recognition: An Adversarial Learning Framework and A New Dataset». En: *IEEE TPAMI* (2020).

genes adquiridas aún no protegen la privacidad. Por lo que los datos adquiridos, imágenes o videos, pueden contener datos sensibles que podrían ser expuestos mediante ataques de restauración.

2.3.2. Protección de la privacidad a nivel de hardware

Los enfoques de protección de privacidad a nivel de hardware consisten en la parametrización y configuración del sistema óptico de adquisición de imágenes de manera que sea diferenciable y adaptable mediante técnicas de aprendizaje profundo, protegiendo al mismo tiempo la información sensible durante la adquisición de las imágenes. Trabajos previos han utilizado cámaras de baja resolución para capturar videos y evitar la filtración no deseada de información de las personas involucradas[53,54]. Del mismo modo, los métodos de desenfoque proporcionan un nivel de privacidad sobre una región dentro de los límites del tamaño del sensor[55,56]; sin embargo, el uso exclusivo de desenfoque óptico para la privacidad es susceptible a métodos de inversión. Por otro lado, los autores en[57] utilizan una cámara, que incluye una apertura codificada, para realizar el reconocimiento de acciones humanas a partir de medidas codificadas sin requerir la restauración de imágenes como paso

---

[53] Michael Ryoo, Kiyoon Kim y Hyun Yang. «Extreme Low Resolution Activity Recognition With Multi-Siamese Embedding Learning». En: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018.

[54] Michael S Ryoo et al. «Privacy-preserving human activity recognition from extreme low resolution». En: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[55] Pittaluga y Koppal, «Privacy preserving optics for miniature vision sensors», ver n. 21.

[56] Pittaluga y Koppal, «Pre-capture privacy for small vision sensors», ver n. 22.

[57] Zihao W Wang et al. «Privacy-preserving action recognition using coded aperture videos». En: *Proceedings of the IEEE Conference on CVPR Workshops*. 2019.

intermedio. Más recientemente, Hinojosa et al.[58,59] propusieron un enfoque de extremo a extremo, que protege la privacidad de las imágenes en visión por computadora mediante el diseño de elementos ópticos usando aprendizaje profundo[60], diseñando un lente refractivo que oscurece información privada mientras permite realizar tareas de visión por computadora como el reconocimiento de acciones y la estimación de la pose humana. Un enfoque de extremo a extremo asegura que todos los componentes del sistema están alineados con el objetivo final desde el principio. Esto permite que el sistema ejecute tareas complejas utilizando directamente las imágenes que ya han sido procesadas para proteger la privacidad mediante la manipulación óptica.

2.4. Distorsión Óptica como Método de Protección de Privacidad

Para asegurar la protección de la privacidad desde el nivel de hardware, es esencial manipular las imágenes directamente en su etapa de adquisición, es decir, antes de la formación de la imagen, garantizando así que las imágenes permanezcan protegidas desde el principio. Esto se puede conseguir mediante técnicas de distorsión óptica aplicadas en la lente de la cámara, lo que requiere un entendimiento del proceso de formación de la imagen.

2.4.1. Formación de imagen

En el proceso de adquisición de una imagen, descrito en la Figura 5, la lente delgada con distancia focal f y a una distancia d_2 del sensor sigue la ecuación de la lente delgada: $1/f = 1/d_1 + 1/d_2$. Se representa el campo complejo inmediatamente

---

58 Hinojosa et al., ver n. 18.

59 Hinojosa, Niebles y Arguello, ver n. 17.

60 Sitzmann et al., ver n. 16.

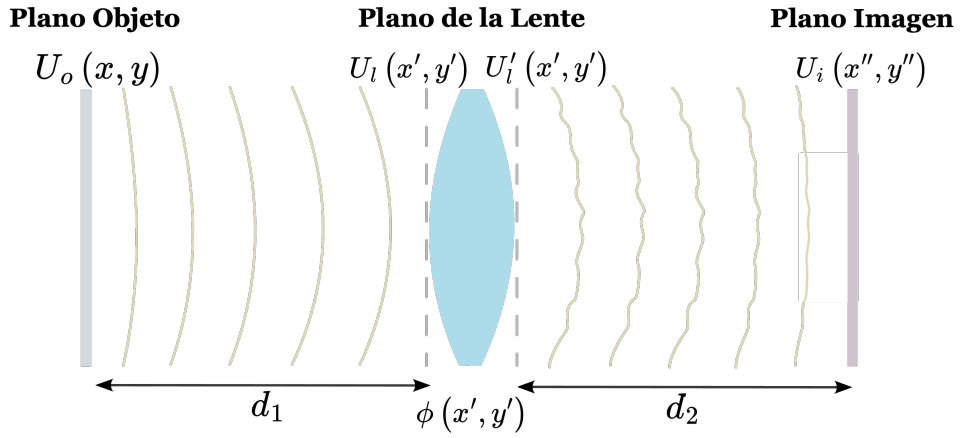

Figura 5. Proceso de formación de una imagen.

detrás del objeto como $U_o(x, y)$ el cual incide en la lente como una onda esférica que diverge desde (x, y). La aproximación paraxial de esa onda se describe como:

$$U_l(x', y', \lambda) = \frac{1}{i\lambda d_1} \exp\left\{ i\frac{\pi}{\lambda d_1}\left[(x'-x)^2 + (y'-y)^2\right]\right\}, \tag{5}$$

donde las coordenadas en el plano de la lente están representadas por (x', y'), y la longitud de onda por λ[61].

La modulación de fase se define como $t_\phi(\cdot)$, resultado tanto de la lente como de la propagación de ondas, y se representa mediante la función

$$t_\phi(x', y', \lambda) = \exp\left(i\frac{\pi}{\lambda f}\phi(x', y')\right). \tag{6}$$

El campo eléctrico inmediatamente después de la lente U_l', está definido como

$$U_l'(x', y', \lambda) = U_l(x', y', \lambda)P(x', y')t_\phi(x', y', \lambda), \tag{7}$$

donde $P(x', y')$ representa la apertura y se define como sigue

---

[61] Joseph W Goodman. *Introduction to Fourier optics*. Macmillan Learning, 4 edition, 2017.

$$P(x', y') = \begin{cases} 1, & x'^2 + y'^2 \leq r^2 \\ 0, & \textit{de otra manera}. \end{cases} \tag{8}$$

donde r es el radio de la apertura.

Finalmente a una distancia d_2 detrás de la lente aparece una distribución de campo que se representa por U_i. En vista de la linealidad del fenómeno de propagación de ondas, en todos los casos podemos expresar el campo U_i mediante la siguiente integral de superposición:

$$U_i(x'', y'', \lambda) = \iint_{-\infty}^{\infty} h(x'', y'', \lambda) U_o(x, y) dx dy, \tag{9}$$

donde $h(x'', y'', \lambda)$ es la respuesta del campo dado un punto, o PSF (por sus siglas en inglés *Point Spread Function*) producido en las coordenadas (x'', y'') por una fuente puntual de amplitud unitaria aplicada en las coordenadas del objeto (x, y), descrito como:

$$h(x'', y'', \lambda) = \frac{1}{i\lambda d_2} \iint_{-\infty}^{\infty} U_l'(x', y') \exp\left\{ i \frac{\pi}{\lambda d_2} \left[(x'' - x')^2 + (y'' - y')^2 \right] \right\} dx' dy', \tag{10}$$

donde las coordenadas espaciales en el plano del sensor son (x'', y'').

2.4.2. Parametrización de elementos ópticos difractivos

La función de modulación de fase $t_\phi(x', y', \lambda)$ en la Ecuación (6) se obtiene a partir del perfil de la superficie de la lente ϕ como se observa en la Figura 6. Esta se puede parametrizar de distintas maneras para su diseño.

En el estado del arte se pueden encontrar varias parametrizaciones de la superficie de la lente $\phi(x, y)$, entre estas están:

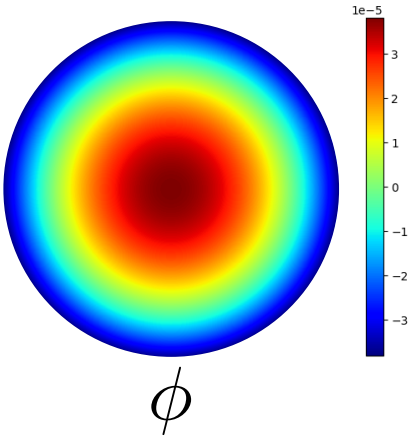

Figura 6. Superficie de la lente ϕ de una lente tradicional en unidades de micrometros.

- **Parametrización píxel a píxel:**[62] Se discretiza $\phi(x, y)$ como un arreglo bidimensional de la siguiente manera:

$$\phi(x,y) = \sum_{i=0}^{N_y} \sum_{j=0}^{N_x} M_{ij} \cdot \text{rect}\left(\frac{x}{d_p} - j, \frac{y}{d_p} - i\right), \tag{11}$$

  la función rect$(\cdot)$ es la función rectangular que define la discretización con un tamaño de píxel d_p, siendo N_x, N_y la cantidad de píxeles en los ejes horizontal y vertical, respectivamente.

- **Parametrización de anillos concéntricos:**[63] Mediante la suma de anillos concéntricos se puede parametrizar ϕ con la siguiente ecuación:

---

[62] Elias Nehme et al. «DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning». En: *Nature Methods* (2020).

[63] Xiong Dun et al. «Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging». En: *Optica* (2020).

$$\phi(x,y) = \sum_{r=0}^{N_r} p_r[\text{circ}_{r+1}(x,y) - \text{circ}_r(x,y)], \tag{12}$$

la función **circ**$_r(x,y)$ es igual a 1 si $x^2 + y^2 \leq R_r$, lo cual es conocido como la función círculo donde $R_r = rd$ para $r = 1, ..., N_r$, siendo N_r, el número total de radios y d la distancia entre ellos, y p_r, representa la altura de los anillos correspondientes a cada radio.

- **Parametrización de coeficientes de Zernike:**[64] El perfil de la superficie de la lente ϕ se define utilizando polinomios de Zernike de la siguiente manera:

$$\phi(x,y) = \sum_{j=1}^{q} \alpha_j Z_j(x,y), \tag{13}$$

donde $Z_j(x,y)$ representa el j-ésimo polinomio de Zernike en la notación de Noll, y α_j son los coeficientes correspondientes. Cada polinomio de Zernike representa una aberración específica del frente de onda, y q denota el número total de polinomios utilizados en la combinación lineal. La combinación de estas aberraciones forma el perfil de la superficie resultante, como se ilustra en la Figura 7[65].

---

[64] Yoav Shechtman et al. «Optimal point spread function design for 3D imaging». En: *Physical Review Letters* (2014).

[65] Paula Arguello et al. «Optics Lens Design for Privacy-Preserving Scene Captioning». En: *2022 IEEE ICIP*. IEEE. 2022.

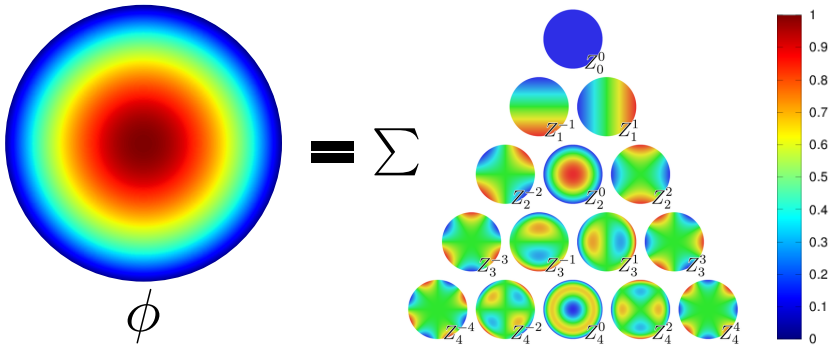

Figura 7. Representación del perfil de la superficie de la lente dada una combinación lineal de polinomios de Zernike. Imagen adaptada de[59].

2.5. Diseño de óptica basado en aprendizaje profundo

Estrategias recientes[66] utilizan la parametrización de polinomios de Zernike de la Ecuación (13), y proponen un módulo diferenciable que tiene en cuenta la propagación de la onda y el proceso de modulación de fase que ocurren dentro de la cámara. Para esto, el problema de optimización se formula de la siguiente manera

$$\min_{\phi} J(S_\phi(\boldsymbol{X}_i), \boldsymbol{X}_i) \tag{14}$$

que en el ámbito de la privacidad busca maximizar la diferencia entre la imagen obtenida $S_\phi(\boldsymbol{X}_i)$ y la original $\boldsymbol{X}_i$. Para resolver el problema de optimización, $J(\cdot)$ debe ser diferenciable con respecto a ϕ, esto se puede lograr mediante la regla de la cadena como se denota a continuación:

$$\frac{\partial J}{\partial \phi} = \frac{\partial J}{\partial S_\phi}\frac{\partial S_\phi}{\partial \phi}. \tag{15}$$

Esta formulación permite parametrizar y optimizar el diseño óptico de la lente de la cámara ϕ que finalmente se refleja en el aprendizaje de un conjunto de coeficientes

---

[66] Hinojosa, Niebles y Arguello, ver n. 17; Hinojosa et al., ver n. 18; Sitzmann et al., ver n. 16.

para los polinomios de Zernike, α_j. Esto permite obtener un efecto de distorsión[67] o desenfoque en el conjunto de datos utilizado, protegiendo partes o atributos sensibles dentro de las imágenes, como objetos personales, documentos y los rostros de las personas presentes en las escenas adquiridas.

2.6. Descripción de escenas preservando la privacidad

La protección de la privacidad es un campo de investigación en expansión, con escasos trabajos dedicados a preservar la privacidad en la tarea de descripción de escenas. Por ejemplo, Qiu et al.[68] proponen un sistema que emplea un modelo basado en *transformers* para generar descripciones para imágenes en una población específica. En lugar de almacenar las imágenes, que podrían contener información privada, el sistema retiene solo las descripciones generadas. Para ello, los autores construyeron un conjunto de datos de descripciones de imágenes donde las caras están enmascaradas manualmente antes del entrenamiento. Por otro lado, Fan et al.[69] introdujeron un modelo para la descripción de la vida cotidiana que utiliza señales de radio y el plano del piso de la casa como contexto, preservando así la privacidad. No obstante, la falta de técnicas de visión por computadora compromete su precisión, y se requiere un plano preciso de la casa, lo que podría limitar su aplicación. Recientemente se propuso un enfoque que preserva la privacidad de las imágenes antes de subirlas a la nube[70]. Sin embargo, como los enfoques vistos de privacidad en visión por computadora, este método opera a nivel de software y

---

67 Vasudevan Lakshminarayanan y Andre Fleck. «Zernike polynomials: a guide». En: *Journal of Modern Optics* 58.7 (2011), págs. 545-561.

68 Qiu et al., ver n. 20.

69 Lijie Fan et al. «In-Home daily-life captioning using radio signals». En: *ECCV*. Springer. 2020.

70 Martin, Ahmadzadeh y Moon, ver n. 19.

no protege intrínsecamente la privacidad desde la adquisición de la imagen, lo que plantea la posible exposición de datos mediante ataques de ingeniería inversa[71].

2.7. Técnicas de deconvolución como ataque a la privacidad

La deconvolución es una técnica utilizada para revertir los efectos de la convolución en una señal, que en el contexto de imágenes se relaciona con el proceso de revertir el desenfoque o distorsión[72]. La deconvolución es un problema fundamental en el procesamiento de señales e imágenes y tiene aplicaciones en diversas áreas como la visión por computadora. Existen dos tipos de técnicas de deconvolución:

- *Deconvolución Ciega (Blind Deconvolution):* La imagen original y la PSF que la distorsionó son desconocidas. La imagen original debe ser estimada solo a partir de la imagen observada, para esto se utlizan algoritmos iterativos que intentan estimar la PSF y/o la imagen desenfocada.
- *Deconvolución No Ciega (Non-Blind Deconvolution):* Se asume que la PSF es conocida o se puede estimar con precisión, lo cual es crucial para el proceso de deconvolución, y es menos compleja que la deconvolución ciega ya que reduce la cantidad de soluciones posibles.

2.7.1. Filtro Wiener

El filtro de Wiener es uno de los enfoques clásicos de Deconvolución No Ciega[73].

---

[71] David Colton. *Surveys on solution methods for inverse problems*. Springer Science & Business Media, 2000.

[72] Pooja Satish, Mallikarjunaswamy Srikantaswamy y Nataraj Kanathur Ramaswamy. «A Comprehensive Review of Blind Deconvolution Techniques for Image Deblurring.» En: *Traitement du Signal* (2020).

[73] H Poor. «On robust Wiener filtering». En: *IEEE TAC* (1980).

Figura 8. Ejemplo de deconvolución con filtro Wiener, adaptada de[68]

Se utiliza para producir una estimación del estado original de una señal que ha sido distorsionada, usualmente agregando también ruido, véase Figura 8[74]. Para recuperar una imagen $\boldsymbol{X}$ a partir de una imagen observada $\hat{\boldsymbol{X}}$ se realiza el siguiente proceso de deconvolución:

$$\bar{\boldsymbol{X}} = \frac{H^*(f)S(f)}{|H(f)|^2 S(f) + \boldsymbol{N}(f)} \cdot \hat{\boldsymbol{X}}(f), \tag{16}$$

donde $\bar{\boldsymbol{X}}$ es la imagen estimada, $H(f)$ es la PSF de la distorsión, en el dominio de la frecuencia, $S(f)$ representa el espectro de potencia de la señal original antes de ser distorsionada, y $\boldsymbol{N}(f)$ es el espectro de potencia del ruido agregado.

2.7.2. Red neuronal de desenfoque

Con el avance de las técnicas de aprendizaje profundo, las redes neuronales se han utilizado para tareas de deconvolución de imágenes, las cuales son entrenadas para aprender la relación entre imágenes desenfocadas y nítidas utilizando conjuntos de datos de pares de imágenes $\left\{\left(\hat{\boldsymbol{X}}_i, \boldsymbol{X}_i\right)\right\}_{i=1}^{N}$. Existen diferentes arquitecturas de redes neuronales para esta tarea, incluyendo:

- **Redes Neuronales Convolucionales (CNNs):** Aprenden características de

[74] Wikipedia. *Wiener-Filter*. `https://de.wikipedia.org/wiki/Wiener-Filter`. Accedido 17-02-2024. 2023.

las imágenes y han demostrado ser efectivas en el desenfoque. Las CNNs pueden ser entrenadas para mapear directamente de una imagen desenfocada a una imagen nítida.

- **Redes Generativas Adversarias (GANs):** Este enfoque de deconvolución ciega consiste de dos redes, la primera red es generadora e intenta crear imágenes nítidas a partir de imágenes desenfocadas, y la segunda red es discriminadora y promueve distinguir entre imágenes nítidas reales e imágenes generadas por la primera red. La competencia entre estas redes mejora la calidad de las imágenes distorsionadas[75].

En el contexto de la privacidad, es crucial emplear estas técnicas en las imágenes alteradas con el propósito de protección, para garantizar que estas imágenes no puedan ser restauradas. Por consiguiente, la ineficacia de estas técnicas de deconvolución en imágenes adquiridas distorsionadas por un modelo de privacidad es un punto a favor de la robustez del método.

[75] Orest Kupyn et al. «Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better». En: *Proceedings of the IEEE/CVF ICCV*. 2019.

3. MÉTODO PROPUESTO

El modelo propuesto en este trabajo de grado para la descripción de imágenes preservando la privacidad consta de dos componentes principales. En primer lugar, una cámara equipada con un elemento óptico refractivo específicamente entrenado para la distorsión de imágenes. Se diseñó un elemento óptico refractivo mediante el aprendizaje de una combinación lineal de polinomios de Zernike, como se ilustra en la Figura 9. En consecuencia, se obtienen imágenes privadas mediante el codificador óptico, distorsionando eficazmente atributos privados sensibles como objetos personales, documentos y los rostros de las personas dentro de la imagen.

El segundo componente del modelo propuesto incluye un módulo decodificador que aprende a generar descripciones directamente de las imágenes distorsionadas adquiridas por la cámara del primer componente. El decodificador emplea una red neuronal convolucional (CNN) para extraer características esenciales de las imágenes distorsionadas. Posteriormente, se implementan dos redes de memoria a corto plazo de larga duración (LSTM) con un módulo de atención entre ellas para finalmente generar descripciones para las imágenes.

Los dos componentes mencionados se entrenan en conjunto utilizando una estrategia End-to-End que permite el aprendizaje de las propiedades ópticas retropropagándose desde el decodificador de la red de descripción de imágenes hasta la capa óptica.

3.1. Codificador óptico para la preservación de privacidad

El codificador óptico de la Figura 10 realiza el proceso de adquisición de imágenes. Como se ha descrito anteriormente, la estrategia para promover la protección de la privacidad consiste en modificar el sistema óptico de la cámara mediante el

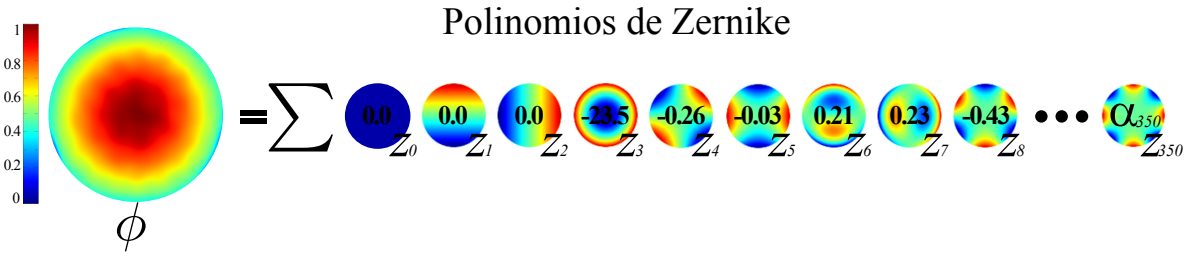

Figura 9. Mapa de alturas de 350 polinomios de Zernike utilizados para la optimización de la lente.

aprendizaje de un elemento óptico refractivo durante el entrenamiento. El objetivo es producir imágenes que distorsionen visualmente atributos sensibles y preservando características importantes para realizar las descripciones de las imágenes. Para facilitar la optimización de la lente de la cámara para la descripción de imágenes preservando la privacidad, se implementó una estrategia similar de estudios anteriores[76]. Se desarrolló un módulo diferenciable que tiene en cuenta el proceso de propagación de ondas y modulación de fase inherente a la cámara.

Se formuló el modelo de formación de imágenes basado en ondas para definir la PSF en términos del perfil de la superficie de la lente para emular la propagación del frente de onda. Se puede describir matemáticamente la PSF de la siguiente manera:

$$h(x'', y'', \lambda) = \frac{1}{i\lambda d_2} \iint_{-\infty}^{\infty} U_l'(x', y') \exp\left\{ i \frac{\pi}{\lambda d_2} \left[(x'' - x')^2 + (y'' - y')^2 \right] \right\} dx' dy'. \quad (17)$$

Del mismo modo el campo eléctrico inmediatamente después de la lente se define como:

$$U_l'(x', y', \lambda) = U_l(x', y', \lambda) P(x', y') t_\phi(x', y', \lambda). \quad (18)$$

---

[76] Hinojosa, Niebles y Arguello, ver n. 17; Hinojosa et al., ver n. 18; Sitzmann et al., ver n. 16.

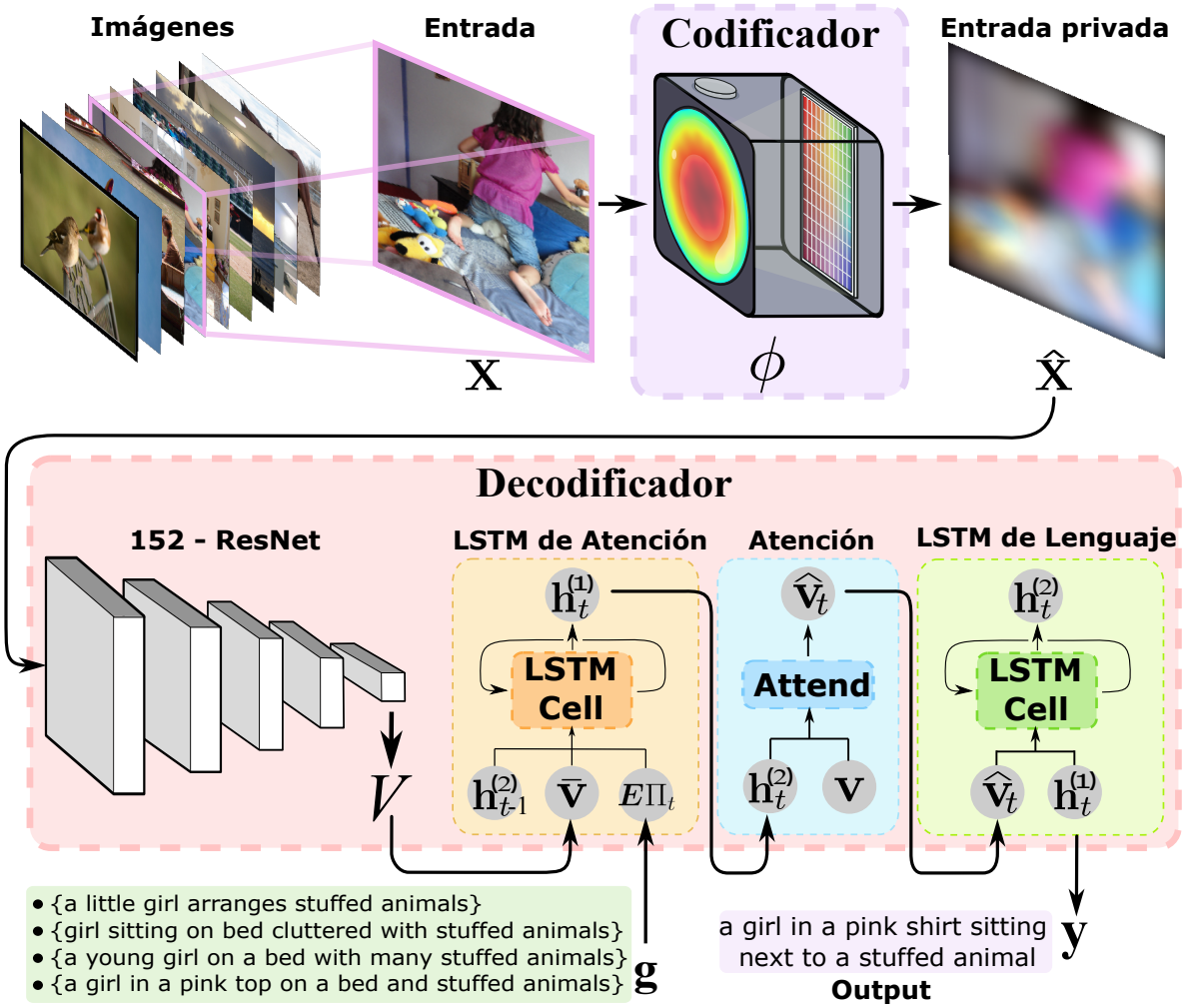

Figura 10. Modelo propuesto con estrategia de extremo a extremo. El codificador óptico consta de una cámara con un lente refractivo que está parametrizado con polinomios de Zernike. El decodificador incorpora una extracción de características mediante una CNN seguido de una LSTM de atención, un módulo de atención y una LSTM de lenguaje que genera una descripción de la imagen privada.

La función de modulación de fase $t_\phi(x', y', \lambda)$ de la Ecuación (18), definida en la Ecuación (6), se obtiene a partir del perfil de la superficie de la lente ϕ, que se define como

$$\phi(x', y') = \sum_{j=1}^{q} \alpha_j Z_j(x', y'), \tag{19}$$

donde Z es la serie de polinomios de Zernike, cada uno con su coeficiente correspondiente. Cada polinomio modela una aberración diferente. La combinación de todos los polinomios, genera el perfil de superficie resultante, como se ilustra en la Figura 9.

Posteriormente las imágenes privadas adquiridas para los tres canales RGB pueden modelarse, dado el proceso de formación de la imagen, de la siguiente manera:

$$U_i(x'', y'', \lambda) = \iint_{-\infty}^{\infty} h(x'', y'', \lambda) U_o(x, y) dx dy, \tag{20}$$

donde se produce una convolución entre la intensidad de la luz en la escena original o plano objeto $U_o(x, y)$ y la PSF del sistema, generando una distribución de intensidad en el plano de la imagen $U_i(x'', y'', \lambda)$ para una longitud de onda específica λ. Posteriormente, esta imagen resultante es captada por el sensor de la cámara de la siguiente manera:

$$x_\lambda = \mathcal{S}_\lambda(U_i) + \eta, \tag{21}$$

donde la variable $x_\lambda \in \mathbb{R}+^{w \times h}$ representa la imagen nítida con $w \times h$ píxeles, el término η corresponde al ruido gaussiano presente en el sensor, y la función $\mathcal{S}_\lambda(\cdot) : \mathbb{R}^{w \times h} \to \mathbb{R}^{w \times h}$ denota la función de respuesta de la cámara, que se modela como un operador lineal. La optimización que se realiza se centra en aprender el conjunto de coeficientes α_j en la Ecuación (19) que distorsionen la imagen x lo suficiente para preservar su privacidad.

Por otra parte, para la optimización del modelo de extremo a extremo, se debe considerar el caso en el que la imagen se adquiera sin privacidad. En este contexto, el perfil de la superficie de la lente, $\hat{\phi}$, se define de la siguiente manera:

$$\hat{\phi}(x', y') = x'^2 + y'^2. \tag{22}$$

Esta configuración de la superficie de la lente garantiza que las imágenes adquiridas sean nítidas. Finalmente, la imagen nítida se define como:

$$\hat{x}_\lambda = \mathcal{S}_\lambda(\hat{U}_i) + \eta, \tag{23}$$

donde $\hat{U}_i$ representa una versión de la Ecuación (20) que utiliza el perfil de la lente descrito en la Ecuación (22).

3.2. Decodificador para la descripción de imágenes

3.2.1. Extracción de características

Para preservar la privacidad de extremo a extremo, se entrenó una red decodificadora para que aprenda características directamente de las imágenes codificadas ópticamente (privadas) adquiridas de nuestra cámara, véase la Ecuación (21). Se implementó la ResNet152 (CNN) por su profundidad y capacidad para capturar características complejas y detalladas de las imágenes[77]. A partir esto, se generó un conjunto de vectores de características $\mathcal{V}$, que se puede denotar como:

$$\mathcal{V} = \{\boldsymbol{v}_i\}_{i=1}^{L} \in \mathbb{R}^D, \tag{24}$$

donde $\boldsymbol{v}_i \in \mathbb{R}^D$ representa un vector de características extraído de una imagen, de

---

[77] He et al., ver n. 42.

hasta de L imágenes.

3.2.2. Arquitectura de descripción de imágenes preservando la privacidad

Para la generación de descripciones se utilizó un método[78] que combina una *LSTM de atención* con una *LSTM de lenguaje*. La *LSTM de atención* utiliza características visuales para enfocarse en áreas relevantes de la imagen. Luego, esta información atendida se pasa a la *LSTM de lenguaje*, que genera palabras secuencialmente.

- **Red LSTM de atención**

  La primera red LSTM procesa una única característica vectorial $\overline{\boldsymbol{v}}$ que representa a toda la imagen, la cual es calculada como el promedio de los vectores de las características:

  $$\overline{\boldsymbol{v}} = \frac{1}{L} \sum_{i} \boldsymbol{v}_i. \tag{25}$$

  La entrada de la *LSTM de atención* consiste en la salida previa de la *LSTM de lenguaje* $\boldsymbol{h}_{t-1}^{(2)}$ concatenada con el vector de características $\overline{\boldsymbol{v}}$ y una codificación *one-hot* de la palabra de entrada en el paso de tiempo t. Estas entradas proporcionan a la *LSTM de atención* con el máximo contexto en relación con el estado de la *LSTM de lenguaje*, el contenido general de la imagen y la salida parcial de descripciones generada hasta el momento. La *LSTM de atención* se puede identificar como la primera LSTM del decodificador de la Figura 10 seguido de la ResNet-152. Sus correspondientes ecuaciones de la se presentan a continuación:

---

[78] Peter Anderson et al. «Bottom-up and top-down attention for image captioning and visual question answering». En: *Proceedings of the IEEE Conference on CVPR*. 2018.

$$
\begin{aligned}
\boldsymbol{i}_t^{(1)} &= \sigma\left(\boldsymbol{W}_{\mathbf{i}}^{(1)}\left[\boldsymbol{h}_{t-1}^{(1)}, \boldsymbol{h}_{t-1}^{(2)}, \bar{\boldsymbol{v}}, \boldsymbol{E}\Pi_t\right] + \boldsymbol{b}_{\mathbf{i}}^{(1)}\right), \\
\boldsymbol{f}_t^{(1)} &= \sigma\left(\boldsymbol{W}_{\mathbf{f}}^{(1)}\left[\boldsymbol{h}_{t-1}^{(1)}, \boldsymbol{h}_{t-1}^{(2)}, \bar{\boldsymbol{v}}, \boldsymbol{E}\Pi_t\right] + \boldsymbol{b}_{\mathbf{f}}^{(1)}\right), \\
\bar{\boldsymbol{c}}_t^{(1)} &= \tanh\left(\boldsymbol{W}_{\mathbf{c}}^{(1)}\left[\boldsymbol{h}_{t-1}^{(1)}, \boldsymbol{h}_{t-1}^{(2)}, \bar{\boldsymbol{v}}, \boldsymbol{E}\Pi_t\right] + \boldsymbol{b}_{\mathbf{c}}^{(1)}\right), \\
\boldsymbol{c}_t^{(1)} &= \boldsymbol{f}_t^{(1)}\boldsymbol{c}_{t-1}^{(1)} + \boldsymbol{i}_t^{(1)}\bar{\boldsymbol{c}}_t^{(1)}, \\
\boldsymbol{o}_t^{(1)} &= \sigma\left(\boldsymbol{W}_{\mathbf{o}}^{(1)}\left[\boldsymbol{h}_{t-1}^{(1)}, \boldsymbol{h}_{t-1}^{(2)}, \bar{\boldsymbol{v}}, \boldsymbol{E}\Pi_t\right] + \boldsymbol{b}_{\mathbf{o}}^{(1)}\right), \\
\boldsymbol{h}_t^{(1)} &= \boldsymbol{o}_t^{(1)}\tanh\left(\boldsymbol{c}_t^{(1)}\right).
\end{aligned}
\tag{26}
$$

Las variables $\boldsymbol{i}_t^{(1)}$, $\boldsymbol{f}_t^{(1)}$, $\boldsymbol{c}_t^{(1)}$, $\boldsymbol{o}_t^{(1)}$ y $\boldsymbol{h}_t^{(1)} \in \mathbb{R}^n$ corresponden al estado de entrada, estado de olvido, celda, salida y estado oculto de la red *LSTM de atención*, respectivamente, con superíndice 1 dado que corresponde a la primera LSTM del modelo. La matriz $\boldsymbol{E} \in \mathbb{R}^{m\times K}$ es una matriz de incrustación de palabras para un vocabulario de tamaño K[79]. Las matrices $\boldsymbol{W}_s \in \mathbb{R}^{n\times m}$ y los vectores $\boldsymbol{b}_s \in \mathbb{R}^n$ representan los pesos aprendidos, y sesgos, respectivamente.

- **Módulo de atención**

  Dada la salida de la *LSTM de atención*, se utilizó un módulo de atención con el fin de que el modelo se enfoque en la información más relevante dentro de las imágenes. La entrada del módulo de atención esta dada por la salida $\boldsymbol{h}_t^{(1)}$ y las características de imágenes $\boldsymbol{v}_i$:

$$
\begin{aligned}
a_{t,i} &= \boldsymbol{w}_a^T \tanh\left(\boldsymbol{W}_{va}\boldsymbol{v}_i + \boldsymbol{W}_{ha}\boldsymbol{h}_t^{(1)}\right), \\
\alpha_{t,i} &= \operatorname{softmax}\left(a_{t,i}\right).
\end{aligned}
\tag{27}
$$

  La característica de imagen atendida se utiliza como entrada para la *LSTM de lenguaje*, la cual se calcula como una combinación de todas las características

[79] Anderson et al., ver n. 78.

de entrada:

$$\hat{\boldsymbol{v}}_t = \sum_{i=1}^{K} \alpha_{t,i} \boldsymbol{v}_i. \tag{28}$$

- **Red LSTM de lenguaje**

  La *LSTM de lenguaje* recibe como entrada la imagen atendida $\hat{\boldsymbol{v}}_t$ junto con la salida de la *LSTM de atención* $\boldsymbol{h}_t^{(1)}$, lo cual se puede expresar mediante las siguientes ecuaciones:

$$\begin{aligned}
\boldsymbol{i}_t^{(2)} &= \sigma\left(\boldsymbol{W}_{\mathbf{i}}^{(2)}\left[\boldsymbol{h}_{t-1}^{(2)}, \boldsymbol{h}_t^{(1)}, \hat{\boldsymbol{v}}_t\right] + \boldsymbol{b}_{\mathbf{i}}^{(2)}\right), \\
\boldsymbol{f}_t^{(2)} &= \sigma\left(\boldsymbol{W}_{\mathbf{f}}^{(2)}\left[\boldsymbol{h}_{t-1}^{(2)}, \boldsymbol{h}_t^{(1)}, \hat{\boldsymbol{v}}_t\right] + \boldsymbol{b}_{\mathbf{f}}^{(2)}\right), \\
\bar{\boldsymbol{c}}_t^{(2)} &= \tanh\left(\boldsymbol{W}_{\mathbf{c}}^{(2)}\left[\boldsymbol{h}_{t-1}^{(2)}, \boldsymbol{h}_t^{(1)}, \hat{\boldsymbol{v}}_t\right] + \boldsymbol{b}_{\mathbf{c}}^{(2)}\right), \\
\boldsymbol{c}_t^{(2)} &= \boldsymbol{f}_t^{(2)} \boldsymbol{c}_{t-1}^{(2)} + \boldsymbol{i}_t^{(2)} \bar{\boldsymbol{c}}_t^{(2)}, \\
\boldsymbol{o}_t^{(2)} &= \sigma\left(\boldsymbol{W}_{\mathbf{o}}^{(2)}\left[\boldsymbol{h}_{t-1}^{(2)}, \boldsymbol{h}_t^{(1)}, \hat{\boldsymbol{v}}_t\right] + \boldsymbol{b}_{\mathbf{o}}^{(2)}\right), \\
\boldsymbol{h}_t^{(2)} &= \boldsymbol{o}_t^{(2)} \tanh\left(\boldsymbol{c}_t^{(2)}\right).
\end{aligned} \tag{29}$$

  Las variables $\boldsymbol{i}_t^{(2)}$, $\boldsymbol{f}_t^{(2)}$, $\boldsymbol{c}_t^{(2)}$, $\boldsymbol{o}_t^{(2)}$ y $\boldsymbol{h}_t^{(2)} \in \mathbb{R}^n$ corresponden a los estados de la red *LSTM de lenguaje* denotadas por su superíndice 2.

  Finalmente, para cada paso de tiempo t, se calcula una distribución de probabilidad condicional, aplicando la función *softmax*, para la palabra actual dado el historial de palabras previas, usando

$$p\left(y_t \mid y_{1:t-1}\right) = \operatorname{softmax}\left(\boldsymbol{W}_p \boldsymbol{h}_t^{(2)} + \boldsymbol{b}_p\right), \tag{30}$$

  donde, $\boldsymbol{W}_p \in \mathbb{R}^{K \times m}$ son los pesos aprendidos, y $\boldsymbol{b}_p \in \mathbb{R}^K$ los sesgos. Por último, la probabilidad total de una secuencia completa de palabras $(y_1, ..., y_T)$,

se calcula como el producto de las distribuciones condicionales de todas las palabras generadas en la secuencia:

$$p\left(y_{1:T}\right) = \prod_{t=1}^{T} p\left(y_t \mid y_{1:t-1}\right). \tag{31}$$

3.3. Función de costo para la descripción de imágenes con privacidad

La función de costo combina múltiples términos para lograr dos objetivos importantes: aumentar la distorsión óptica en la adquisición (con el fin de maximizar la privacidad) y preservar el rendimiento de la generación de palabras en la descripción de imágenes. Para lograr estos objetivos se diseñó la función de pérdida incorporando cuatro términos: $\mathcal{L}_p$, $\mathcal{L}_{ce}$, y $\mathcal{L}_{\mathbf{H}}$.

El término $\mathcal{L}_p$ representa el error cuadrático medio entre la imagen nítida $\hat{x}_\lambda$ que ha sido discretizada como $\hat{\boldsymbol{X}}_\lambda$ y la imagen que ha sido optimizada y parametrizada con polinomios de Zernike x_λ que ha sido discretizada como $\boldsymbol{X}_\lambda$. Este término pretende promover la distorsión de las imágenes capturadas maximizando la diferencia entre las dos imágenes en cada canal de color independientemente. Esta función se define como,

$$\mathcal{L}_p = 1 - \|\hat{\boldsymbol{X}}_\lambda - \boldsymbol{X}_\lambda\|_2^2. \tag{32}$$

$\mathcal{L}_{ce}$ representa la pérdida de entropía cruzada multiclase (más conocida como *categorical cross-entropy* en inglés), que se utiliza para guiar el aprendizaje de la secuencia correcta de palabras para la descripción de imágenes[80]. Este término, compara las probabilidades previstas $\boldsymbol{y}$ con la descripción de referencia $\boldsymbol{g}$ en cada

[80] Anqi Mao, Mehryar Mohri y Yutao Zhong. «Cross-entropy loss functions: Theoretical analysis and applications». En: *International conference on Machine learning*. PMLR. 2023, págs. 23803-23828.

palabra c y promueve la precisión en las descripciones. El término C corresponde a la longitud de la descripción. La ecuación para $\mathcal{L}_{ce}$ es

$$\mathcal{L}_{ce} = \sum_{c=1}^{C} \log \frac{\exp(\boldsymbol{y}_c)}{\exp(\sum_{i=1}^{C} \boldsymbol{y}_i)} \boldsymbol{g}_c. \tag{33}$$

Por último, el término $\mathcal{L}_H$ realiza una regularización en la PSF h_λ que ha sido discretizada como $\boldsymbol{H}_\lambda$, promoviendo la forma circular de la PSF minimizando los valores fuera de una máscara circular $\boldsymbol{M}$. Se calcula como la norma Frobenius de la diferencia entre la PSF convolucionada con la máscara y la PSF original. Formalmente, la ecuación se puede escribir como:

$$\mathcal{L}_{\mathbf{H}} = \|(\boldsymbol{H}_\lambda * \boldsymbol{M}) - \boldsymbol{H}_\lambda\|_F, \tag{34}$$

donde $\boldsymbol{M}$ es una matriz binaria definida por la máscara circular, con los elementos dentro del círculo puestos a 1 y los elementos fuera puestos a 0. Entonces, $\boldsymbol{M}$ se define como:

$$\boldsymbol{M}_{ij} = \begin{cases} 1, & \text{sí } (i-p)^2 + (j-p)^2 \leq r^2, \\ 0, & \text{en caso contrario.} \end{cases} \tag{35}$$

El propósito de $\boldsymbol{M}$ es garantizar una PSF bien centrada en el sensor de la cámara, donde p representa el centro de la imagen, mientras que r denota el radio esperado de la PSF, que permanece fijo desde el principio. Al emplear esta máscara en el enfoque de aprendizaje de extremo a extremo, se mejora la convergencia de la PSF y se evita la pérdida de energía en su centro.

4. RESULTADOS

4.1. Bases de datos

Para los resultados del método propuesto se utilizaron dos conjuntos de datos: el conjunto de datos COCO[81] y Flickr8k[82]. El conjunto de datos COCO consta de 82.783 imágenes para el entrenamiento, 41.000 para la validación y 41.000 para las pruebas. Flickr8k contiene 8.000 imágenes, que posteriormente se particionaron en 6.800 imágenes para el entrenamiento y 1.200 para validación. Cabe señalar que, mientras que el conjunto de datos Flickr8k contiene cinco frases de referencia para cada imagen, el conjunto de datos COCO puede tener más de cinco referencias para algunas imágenes. Sin embargo, para mantener la uniformidad entre los conjuntos de datos, se descartó cualquier referencia adicional de forma aleatoria, garantizando así, que los dos conjuntos de datos tuvieran exactamente 5 referencias por imagen. Los experimentos se realizaron en dos pasos. En primer lugar, se entrenaron los modelos sin el componente del codificador óptico, es decir con imágenes nítidas, para garantizar la precisión de las descripciones. Después, se realizó un entrenamiento de extremo a extremo con el método propuesto, con el objetivo de maximizar la precisión de las descripciones mientras se distorsionan las imágenes. Para ello, se utilizaron los dos conjuntos de datos mencionados anteriormente.

---

81 Lin et al., ver n. 27.

82 Vicente Ordonez, Girish Kulkarni y Tamara Berg. «Im2Text: Describing Images Using 1 Million Captioned Photographs». En: *Advances in Neural Information Processing Systems*. Ed. por J. Shawe-Taylor et al. Curran Associates, Inc., 2011.

Figura 11. Muestras de las bases de datos descritas: COCO 2014 (izquierda) y Flickr8k (derecha), cada imagen contiene 5 descripciones asociadas a su contenido.

4.2. Métricas de evaluación

- **BLEU:**[83] Los puntajes BLEU-1, BLEU-2, BLEU-3 y BLEU-4, todos dentro del rango $[0, 1]$, sirven como medidas de la similitud entre la descripción candidata y las descripciones de referencia. Valores más altos, más cercanos a 1, indican una mayor semejanza entre descripciones. Estos índices BLEU, denotados como $n = \{1, 2, 3, 4\}$, evalúan la precisión de secuencias contiguas de n elementos (conocidos como n-gramas) dentro de una muestra dada, donde n toma valores de 1 a 4. La formula general de la métrica BLEU es:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log\left(p_n\right)\right), \tag{36}$$

donde N es el número máximo de n, w_n son los pesos asignados a cada precisión (generalmente iguales, sumando 1 en total) y p_n es la precisión de los n-gramas para cada n, el cual se calcula como la proporción del número de n-gramas coincidentes en la descripción candidata frente al total en la descripción de referencia. Por último, BP es el factor de penalización a las descripciones demasiado cortas, que se calcula de la siguiente manera

$$BP = \begin{cases} 1 & \text{si } c > r, \\ e^{(1-r/c)} & \text{si } c \leq r, \end{cases} \tag{37}$$

donde c es la longitud de la descripción candidata y r es la longitud de las referencias.

---

[83] Kishore Papineni et al. «Bleu: a method for automatic evaluation of machine translation». En: *Association for Computational Linguistics*. 2002.

- **METEOR:**[84] Esta métrica evalúa la calidad de las descripciones generadas por un modelo al compararlas con descripciones de referencia, teniendo en cuenta coincidencias exactas, sinónimos y paráfrasis, proporcionando así una evaluación completa de su calidad. La fórmula se compone de varios componentes:

  - La proporción del número de unigramas coincidentes en la descripción generada y el número total de unigramas en la descripción: $P = \frac{m}{w_{\mathbf{y}}}$, donde m es el número de unigramas coincidentes, y $w_{\mathbf{y}}$ es el número total de unigramas en la predicción.
  - La proporción del número de unigramas coincidentes y el número total de unigramas en la descripción de referencia: $R = \frac{m}{w_{\mathbf{g}}}$, donde $w_{\mathbf{g}}$ es el número total de unigramas en la referencia.
  - El promedio armónico de los terminos anteriores: $F_{\text{mean}} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$.
  - Penalización de las coincidencias desordenadas para tener en cuenta el orden correcto: $p = 0.5 \cdot \left(\frac{c}{m}\right)^3$, donde c es el número de fragmentos contiguos de unigramas coincidentes.

  Dado los anteriores componentes, la ecuación final de la métrica se define como:

  $$M = F_{\text{mean}} \cdot (1 - p). \tag{38}$$

  Entre más alto sea el valor de esta métrica, las descripciones generadas son más precisas.

- **CIDEr:**[85] Mide el consenso entre las descripciones generadas y las referen-

---

[84] Satanjeev Banerjee y Alon Lavie. «METEOR: An automatic metric for MT evaluation with improved correlation with human judgments». En: *Association for Computational Linguistics*. 2005.

[85] Ramakrishna Vedantam, C Lawrence Zitnick y Devi Parikh. «Cider: Consensus-based image description evaluation». En: *Proceedings of the IEEE Conference on CVPR*. 2015.

cias calculando la similitud coseno entre sus respectivas características de n-gramas. Las descripciones se representan como vectores de frecuencia de términos ponderados (TF-IDF)[86]. Sea y_i una descripción generada y g_j una de las descripciones de referencia, se utiliza la similitud de coseno entre sus vectores TF-IDF:

$$\text{sim}(y, g) = \frac{\sum_w \text{tf-idf}(w, y) \cdot \text{tf-idf}(w, g)}{\sqrt{\sum_w (\text{tf-idf}(w, y))^2} \cdot \sqrt{\sum_w (\text{tf-idf}(w, g))^2}}. \tag{39}$$

Seguidamente, se promedia la similitud de coseno entre la descripción generada y todas las descripciones de referencia, de la siguiente manera

$$\text{CIDEr}(y) = \frac{1}{5} \sum_{j=1}^{5} \text{sim}(y, g_j), \tag{40}$$

donde hay 5 descripciones de referencia. Por último, se toma el promedio sobre todas las imágenes del conjunto de datos

$$\text{CIDEr} = \frac{1}{N} \sum_{i=1}^{N} \text{CIDEr}(y_i). \tag{41}$$

Un valor alto en la metrica CIDEr simboliza una buena generación de descripción de imágenes.

- **ROUGE:**[87] Calcula la superposición entre la descripción generada y una o más descripciones de referencia, evaluando la calidad de las descripciones en términos de n-gramas coincidentes, mediante la ecuación

---

[86] Hans Christian y Mikhael Pramodana Agus. «Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)». En: *ComTech* (2016).

[87] Chin-Yew Lin. «Rouge: A package for automatic evaluation of summaries». En: *Text summarization branches out*. 2004.

$$\text{ROUGE-N} = \frac{\sum_{s \in g} \sum_{n \in s} \text{mín}(\text{Cont}(n, y), \text{Cont}(n, g))}{\sum_{s \in g} \sum_{n \in s} \text{Cont}(n, g)}, \tag{42}$$

donde, Cont(n, y) es la frecuencia del n-grama n en la descripción generada, y, Cont(n, g) es la frecuencia del n-grama n en la descripción de referencia. Entre más alto sea su valor, más precisión tendrán las descripciones obtenidas.

- **PSNR:** Esta métrica se utiliza para evaluar la distorsión entre la imagen original $\mathbf{X}$ y la imagen generada distorsionada $\hat{\mathbf{X}}$. Para calcular el PSNR se hace uso de la siguiente ecuación,

$$\text{PSNR}(\mathbf{X}, \hat{\mathbf{X}}) = 10 \cdot \log_{10} \left(\frac{255^2}{\text{MSE}(\mathbf{X}, \hat{\mathbf{X}})} \right), \tag{43}$$

donde el error cuadrático medio (MSE) se obtiene con la siguiente ecuación

$$\text{MSE}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{X} - \hat{\mathbf{X}}\|_2^2. \tag{44}$$

Menor valor de PSNR indica imágenes diferentes, y mayor distorsión por parte del lente refractivo.

- **SSIM:**[88] La métrica SSIM igualmente se usa para evaluar la distorsión de las imágenes obtenidas $\hat{\mathbf{X}}$, con respecto a las imágenes originales $\mathbf{X}$. Se calcula con la siguiente ecuación

$$\text{SSIM}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{(2\mu_{\mathbf{X}}\mu_{\hat{\mathbf{X}}} + c_1)\,(2\sigma_{\mathbf{X}\hat{\mathbf{X}}} + c_2)}{\left(\mu_{\mathbf{X}}^2 + \mu_{\hat{\mathbf{X}}}^2 + c_1\right)\left(\sigma_{\mathbf{X}}^2 + \sigma_{\hat{\mathbf{X}}}^2 + c_2\right)}, \tag{45}$$

---

[88] Mohammed Hassan y Chakravarthy Bhagvati. «Structural similarity measure for color images». En: *International Journal of Computer Applications* (2012).

donde $\mu_{\mathbf{X}}$ denota la media de $\mathbf{X}$, $\sigma^2_{\mathbf{X}}$ representa la varianza de $\mathbf{X}$, $\sigma_{\mathbf{X}\hat{\mathbf{X}}}$ simboliza la covarianza entre $\mathbf{X}$ y $\hat{\mathbf{X}}$. Por otro lado, $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ son dos variables que estabilizan la división, $L = 1$ es el rango de valores de los píxeles, $k = 0.01$ y $k_2 = 0.03$ son dos parámetros preestablecidos. Entre más bajo es el valor de la métrica SSIM, mayor distorsión hay en la imagen obtenida.

4.3. Simulaciones

4.3.1. Resultados cualitativos

Para evaluar el rendimiento del método propuesto a la hora de mantener la privacidad frente a otras alternativas, se realizó un análisis comparativo entrenando dos configuraciones de cámara diferentes. La primera cámara, denominada “lente de desenfoque”, comparte la misma arquitectura óptica que el método propuesto, pero sólo optimiza el 4º polinomio de Zernike, que induce el desenfoque. La segunda, denominada “cámara de baja resolución”, es una cámara convencional equipada con un pequeño sensor de 16×16 píxeles. Los resultados de estos enfoques alternativos se presentan en la Figura 12, que muestra las imágenes originales junto con sus correspondientes descripciones de referencia, así como las imágenes que preservan la privacidad y sus respectivas descripciones. La métrica *Meteor* se calculó para cada descripción obtenida comparándola con su referencia, donde un valor alto de la métrica simboliza mayor precisión en la descripción generada. Además, para evaluar cuantitativamente el nivel de degradación introducido por las distintas cámaras en comparación con la imagen original, se calculó la métrica SSIM. Es crucial señalar que, en los enfoques mostrados, el contenido de las imágenes sigue siendo difícil de discernir. Sin embargo, el método que se propone es el que consigue la descripción más precisa de la imagen.

Figura 12. Resultados cualitativos en 4 muestras aleatorias del conjunto de datos COCO. En la esquina superior izquierda de cada imagen se muestra la métrica SSIM calculada entre las imágenes distorsionadas y las originales, junto con la métrica Meteor de las descripciónes generadas y las descripciones de referencia.

4.3.2. Resultados cuantitativos

En el Cuadro 2 se presentan los resultados las métricas *BLEU-1, BLEU-2, BLEU-3, BLEU-4*, *Meteor*, *Rouge* y *Cider* respectivamente, las cuales se utilizaron para evaluar la calidad de las descripciones en los conjuntos de datos COCO y Flickr8k, los valores más altos simbolizan mayor precisión en las descripciones. El estudio empleado incluye una comparación entre el modelo de descripciones de imágenes implementado sin el codificador óptico (denominado **Tesis**), y otros enfoques de descripciones de imágenes que emplean imágenes RGB sin distorsiones, como lo son los enfoques BRNN[89]. Adicionalmente se realiza una evaluación cuantitativa del método propuesto que preserva la privacidad con otros enfoques similares, incluyendo los comentados anteriormente: desenfoque y cámaras de baja resolución. En el Cuadro 2, los mejores resultados se indican con valores en negrita, mientras que los segundos mejores resultados aparecen subrayados; esta valoración se realiza por separado para cada conjunto de datos y cada enfoque (Sin privacidad y con privacidad). Como se puede ver en la cuadro, el enfoque **Tesis** obtiene resultados satisfactorios en la precisión de las descripciones en ambos conjuntos de datos. Adicionalmente se puede observar que el método **Tesis-Priv** con privacidad logra el mejor equilibrio entre distorsión de la imagen y precisión en las descripciones, equilibrando eficazmente una alta eficiencia con una sólida protección de la privacidad.

4.3.3. Resultados de técnicas de ataques de privacidad

Se planearon dos tipos de técnicas de deconvolución para validar el enfoque de privacidad propuesto. Se desea demostrar que la información sensible, como los

---

[89] Andrej Karpathy y Li Fei-Fei. «Deep visual-semantic alignments for generating image descriptions». En: *Proceedings of the IEEE Conference on CVPR*. 2015.

Database	Privacy	Model	B-1	B-2	B-3	B-4	M	R	C
COCO	✗	BRNN[90]	64.2	45.1	30.3	20.1	19.5	–	66.6
		NIC[91]	66.6	45.1	32.9	24.6	23.7	–	–
		Hard Attn[92]	71.8	50.4	35.7	25.0	23.0	–	–
		2PSC-w[93]	72.1	54.8	40.4	29.6	**29.2**	39.0	89.2
		Thesis	**73.2**	**56.7**	**43.4**	**33.3**	29.0	**39.4**	**101.2**
	✓	2PSC	**68.9**	51.3	37.3	27.0	**28.1**	**38.1**	88.5
		Thesis-Priv	**68.9**	**51.7**	**38.5**	**29.0**	26.7	37.6	**89.0**
		Defocus	65.1	47.3	34.4	25.3	24.4	35.7	74.0
		Low-Res	50.4	30.7	19.6	13.4	16.9	29.2	30.2
Flickr8k	✗	BRNN	57.9	38.3	24.5	16.0	–	–	–
		NIC	63.0	41.0	27.0	–	–	–	–
		Hard Attn	**67.0**	45.7	31.4	21.3	20.3	–	–
		DRPE[94]	–	–	–	22.5	21.9	**49.6**	**78.2**
		2PSC-w	65.7	47.6	33.9	23.8	25.5	35.6	57.9
		Thesis	65.5	**48.9**	**35.5**	**25.0**	**26.3**	35.9	65.4
	✓	DRPE	48.3	28.9	17.2	10.1	13.6	36.1	22.5
		2PSC	63.5	45.2	31.4	21.5	24.7	34.7	51.8
		Thesis-Priv	**64.6**	**45.8**	**32.0**	**21.6**	**26.2**	**35.5**	**59.3**

Cuadro 1. Los resultados en negrita simbolizan los mejores (los más altos), y los subrayados, los segundos mejores, por conjunto de datos.

Conjunto de Datos	Privacidad	Modelo	B-1	B-2	B-3	B-4	M	R	C
COCO	✗	BRNN[95]	64.2	45.1	30.3	20.1	19.5	–	66.6
		NIC[96]	66.6	45.1	32.9	24.6	23.7	–	–
		Hard Attn[97]	71.8	50.4	35.7	25.0	23.0	–	–
		2PSC-w[98]	72.1	54.8	40.4	29.6	**29.2**	39.0	89.2
		Tesis	**73.2**	**56.7**	**43.4**	**33.3**	29.0	**39.4**	**101.2**
	✔	2PSC	**68.9**	51.3	37.3	27.0	**28.1**	**38.1**	88.5
		Tesis-Priv	**68.9**	**51.7**	**38.5**	**29.0**	26.7	37.6	**89.0**
		Defocus	65.1	47.3	34.4	25.3	24.4	35.7	74.0
		Low-Res	50.4	30.7	19.6	13.4	16.9	29.2	30.2
Flickr8k	✗	BRNN	57.9	38.3	24.5	16.0	–	–	–
		NIC	63.0	41.0	27.0	–	–	–	–
		Hard Attn	**67.0**	45.7	31.4	21.3	20.3	–	–
		DRPE[99]	–	–	–	22.5	21.9	**49.6**	**78.2**
		2PSC-w	65.7	47.6	33.9	23.8	25.5	35.6	57.9
		Tesis	65.5	**48.9**	**35.5**	**25.0**	**26.3**	35.9	65.4
	✔	DRPE	48.3	28.9	17.2	10.1	13.6	36.1	22.5
		2PSC	63.5	45.2	31.4	21.5	24.7	34.7	51.8
		Tesis-Priv	**64.6**	**45.8**	**32.0**	**21.6**	**26.2**	**35.5**	**59.3**

Cuadro 2. Los resultados en negrita simbolizan los mejores (los más altos), y los subrayados, los segundos mejores, por conjunto de datos.

rostros, seguirá siendo irreconocible (protegida) incluso si se invierte el enfoque de distorsión de imágenes propuesto. En concreto, se considera la posibilidad de que un atacante tenga acceso a la PSF implementada o a una amplia cantidad de imágenes con sus correspondientes imágenes sin privacidad. En tal caso, el atacante podría utilizar técnicas de deconvolución (Ciega y No Ciega) sobre las imágenes privadas para recuperar la identidad de las personas. Aquí, se considera que ambos escenarios son posibles, es decir, si el atacante tiene acceso a la cámara, entonces la PSF se puede conocer mediante la imagen de un punto de luz de origen; por lo tanto, se podría utilizar un método de deconvolución no ciega como el *Filtro Wiener*. Del mismo modo, si el atacante no tiene acceso a la cámara pero sí a una gran cantidad de imágenes adquiridas, podría entrenar una red de deconvolución ciega, por ejemplo una red generativa adversarial para la tarea de restauración de imágenes[100]. Los resultados visuales de este experimento se pueden ver en la Figura 13.

- **Deconvolución ciega:** Se entrenó la red *DeblurGANv2* en 3,214 pares de imágenes nítidas y distorsionadas adquiridas por la cámara optimizada propuesta. El conjunto de imágenes se particionó en 2,103 para el entrenamiento y 1,111 para validación. Se implementó la misma configuración de entrenamiento que en[101] y se entrenó el modelo por 40 épocas.
- **Deconvolución no ciega**: Para realizar esta técnica de deconvolución se implemento el Filtro Wiener para generar una estimación del estado original de las imágenes distorsionadas por la cámara.

---

[100] Kupyn et al., ver n. 75.

[101] Kupyn et al., ver n. 75; Hinojosa et al., ver n. 18.

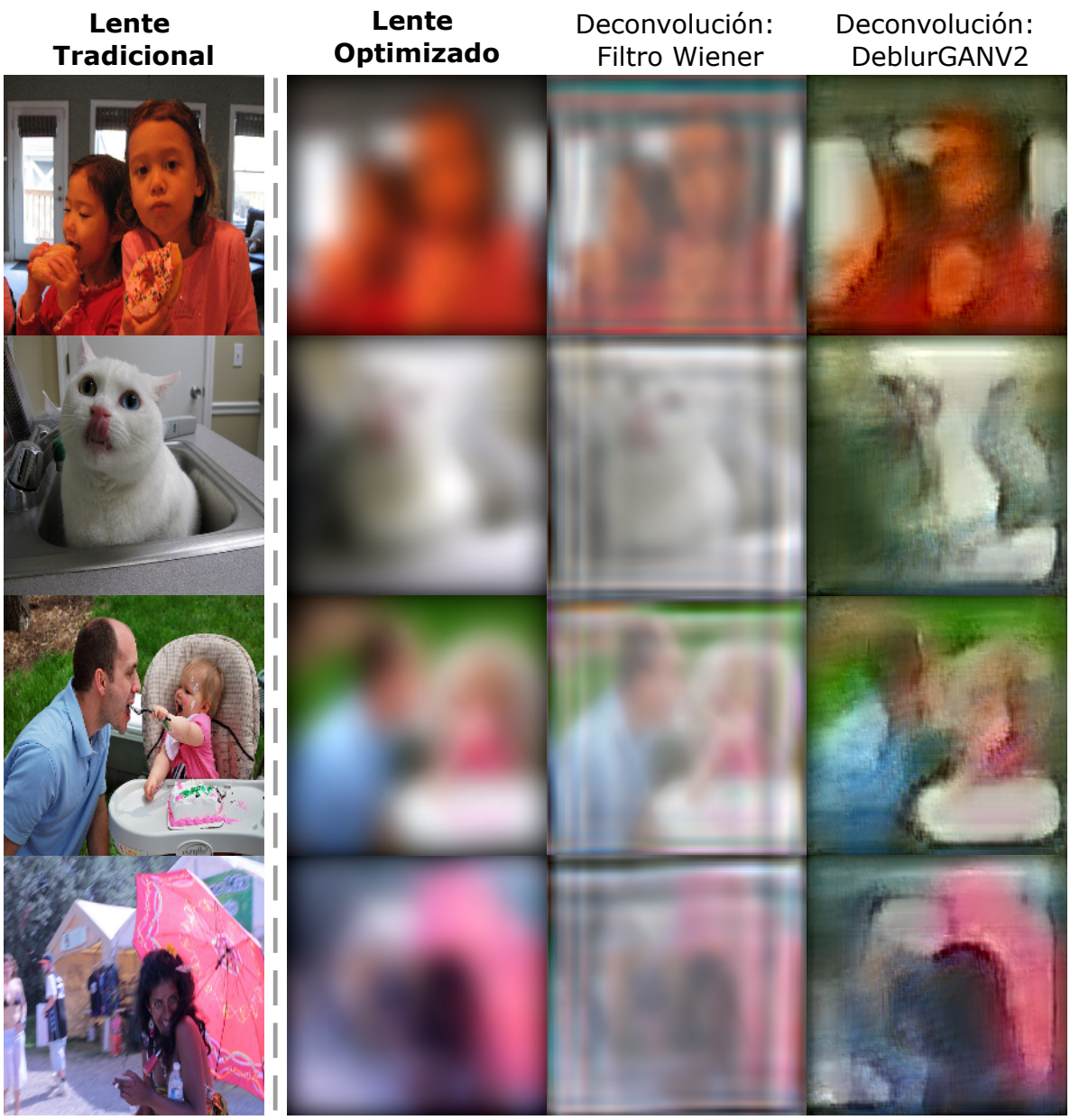

Figura 13. Resultados cualitativos que demuestran la robustez del lente optimizado propuesto frente a ataques de deconvolución.

En la Figura 13 se muestran algunos resultados visuales del conjunto de *Testing* del dataset COCO. Como se observa para ambos escenarios (deconvolución ciega y no ciega), la distorsión conseguida por la lente optimizada es suficiente para no recuperar los detalles de la cara y, por tanto, la identidad de las personas queda protegida. En el Cuadro 4 las filas (sombreadas) que corresponden a *Imágenes privadas* corresponden a las medias de las métricas SSIM y PSNR entre imágenes privadas y las correspondientes nítidas en los conjuntos de *Testing*, tanto para la lente propuesta optimizada como para la lente desenfocada por el polinomio 4 de Zernike. Las filas restantes corresponden a la medida de distorsión (SSIM y PSNR) tras el ataque a la privacidad con los métodos de deconvolución *DeblurGANv2* y *Filtro Wiener*.

		SSIM	PSNR (dB)
Proposed Lens	**Private Images**	$\mathit{0.48 \pm 0.12}$	$\mathit{15.19 \pm 1.60}$
	DeblurGANv2	0.38 ± 0.10	$\mathbf{14.87 \pm 1.51}$
	Wiener Filter	$\mathbf{0.40 \pm 0.11}$	12.65 ± 1.66
Defocus Lens	**Private Images**	$\mathit{0.48 \pm 0.12}$	$\mathit{15.50 \pm 1.59}$
	DeblurGANv2	0.38 ± 0.10	$\mathbf{15.06 \pm 1.34}$
	Wiener Filter	$\mathbf{0.44 \pm 0.11}$	14.68 ± 1.81

Cuadro 3. Medida de distorsión de las imágenes adquiridas con la lente optimizada y lente desenfocada y su correspondiente medida de distorsión tras la deconvolución con DeblurGAN y Fitro Wiener.

		SSIM	PSNR (dB)
Lente Propuesto	**Imágenes privadas**	$\mathit{0.48 \pm 0.12}$	$\mathit{15.19 \pm 1.60}$
	DeblurGANv2	0.38 ± 0.10	$\mathbf{14.87 \pm 1.51}$
	Filtro Wiener	$\mathbf{0.40 \pm 0.11}$	12.65 ± 1.66
Lente Desenfocado	**Imágenes privadas**	$\mathit{0.48 \pm 0.12}$	$\mathit{15.50 \pm 1.59}$
	DeblurGANv2	0.38 ± 0.10	$\mathbf{15.06 \pm 1.34}$
	Filtro Wiener	$\mathbf{0.44 \pm 0.11}$	14.68 ± 1.81

Cuadro 4. Medida de distorsión de las imágenes adquiridas con la lente optimizada y lente desenfocada y su correspondiente medida de distorsión tras la deconvolución con DeblurGAN y Fitro Wiener.

Los resultados en negrita indican los valores más altos de SSIM y PSNR, que simbolizan que las reconstrucciones obtenidas con los métodos de deconvolución mencionados son más similares a las imágenes originales. Como se puede observar, los valores de SSIM y PSNR para el *Lente Desenfocado* (polinomio 4 de Zernike), correspondientes a las deconvoluciones con la red DeblurGANv2 y el Filtro Wiener, son más altos en comparación con las métricas del *Lente Propuesto*. Esto significa que las imágenes con el *Lente Desenfocado* son más fácilmente reconstruibles en comparación con los resultados obtenidos con el *Lente Propuesto*. Esta característica muestra la robustez del método implementado con el *Lente Propuesto* frente a los métodos de deconvolución para la reconstrucción de imágenes privadas. Asímismo, en los Anexos 7.1 y 7.2 se presentan resultados adicionales sobre la validación de la privacidad.

4.4. Resultados experimentales en el laboratorio

Para evaluar experimentalmente la eficacia del enfoque propuesto de descripción de imágenes preservando la privacidad, se diseñó un prototipo de sistema óptico para realizar una prueba de concepto que se muestra en la Figura 14.

En el laboratorio de óptica del grupo de investigación HDSP y con apoyo de los laboratoristas, se construyó un banco de pruebas que sigue el flujo de trabajo descrito en la Figura 10. Para el prototipo, se utilizó la cámara CANON EOS M50 con una dimensión espacial de 6.000×4.000 píxeles colocados en el plano de la imagen de la configuración óptica. El plano medio de la imagen está formado por una lente objetivo (lente CANON de 28-80 mm F/3,5-5,6), que es retransmitida por un par de lentes transformadoras de Fourier de 100mm (Thorlabs AC254-075-A-ML). En un lado de un divisor de haz (BS, Thorlabs CCM1-BS013), se colocó un espejo deformable (DM, Thorlabs DMP40-P01) en el plano de la pupila a una distancia de $2f = 200\ mm$ del plano intermedio de la imagen.

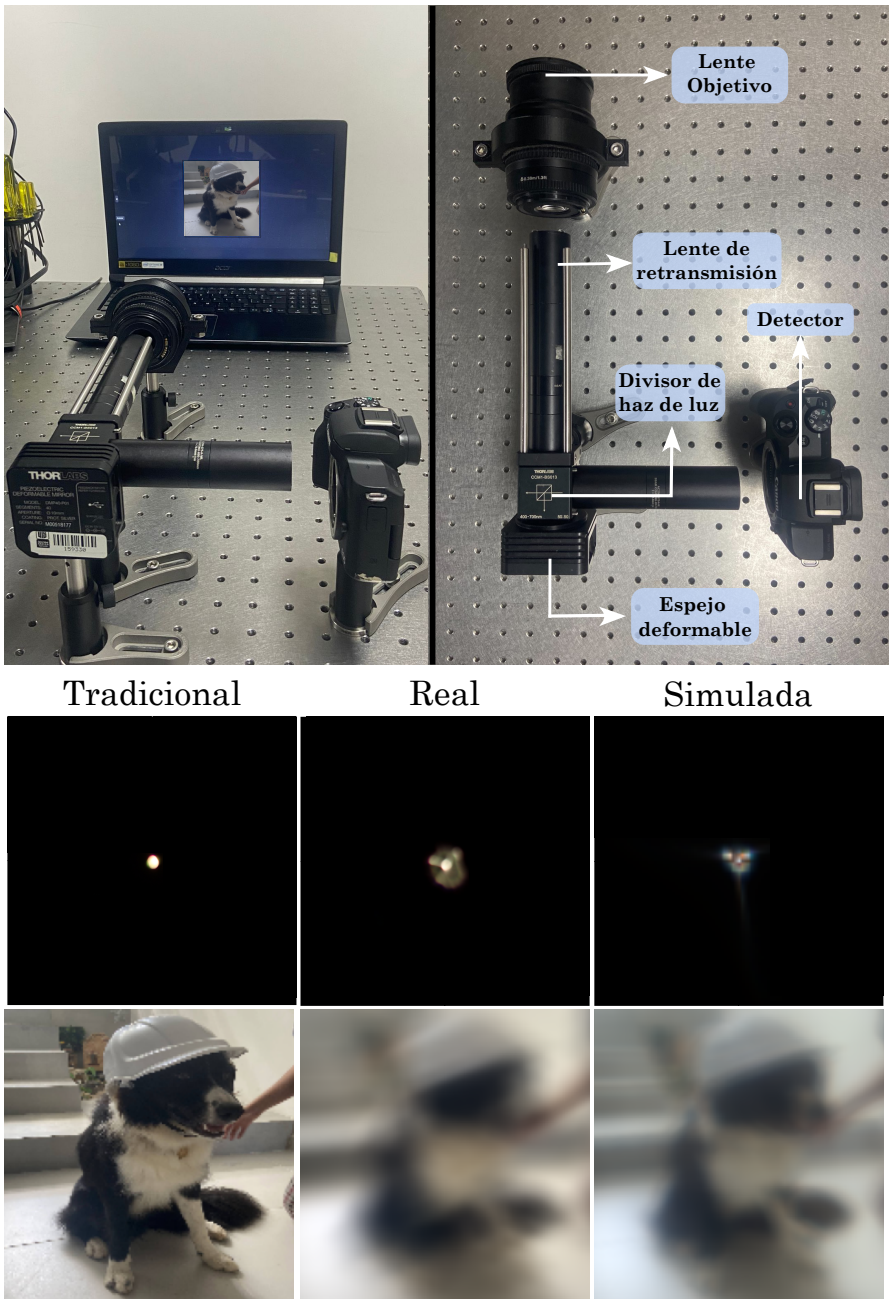

Figura 14. (Arriba) Configuración de prototipo hardware experimental para el método propuesto de descripciones de imágenes con preservación de la privacidad. (Abajo) PSFs y resultados cualitativos en un ejemplo de imagen adquirida con una cámara convencional (izquierda), la cámara para la prueba de concepto (centro) y una cámara simulada (derecha).

Una vez calibrado el sistema enfocado para obtener imágenes RGB naturales sin privacidad, se capturó una imagen de una fuente puntual de luz blanca utilizando una cámara *pinhole*[102] de $20\ \mu m$. A continuación, se calibró el sistema para simular los polinomios de Zernike optimizados por el algoritmo propuesto y aplicarlos al espejo deformable, que modifica el frente de onda de la luz incidente. A continuación, se adquirió la PSF con el punto fuente de luz blanca, utilizando el espejo deformable con los polinomios de Zernike optimizados. En la fila central de la Fig. 14 se muestra la lente tradicional, la PSF real, y la PSF simulada por software, respectivamente. Es importante señalar que el espejo deformable impone una limitación importante al diseño experimental, ya que sólo puede utilizar 15 polinomios de Zernike, lo que restringe el grado de distorsión de la escena que se puede conseguir Vasudevan Lakshminarayanan y Andre Fleck. «Zernike polynomials: a guide». En: *Journal of Modern Optics* 58.7 (2011), págs. 545-561. No obstante, los resultados obtenidos con el prototipo muestran la elevada protección de la identidad personal que proporciona el método propuesto. Estos resultados se obtuvieron a partir del conjunto de pruebas COCO. Para ello, se capturaron imágenes con el prototipo desde una pantalla de ordenador portátil colocada frente a la óptica, que mostraba imágenes del conjunto de pruebas COCO.

Para realizar un análisis comparativo de las imágenes adquiridas por el laboratorio y las obtenidas mediante simulación, se probaron ambos modelos, como se ilustra en la Figura. 15. Como se observa, las distorsiones en las imágenes presentan notables disparidades debido a las distorsiones inherentes que pueden existir a priori en el entorno de laboratorio. No obstante, las descripciones de las imágenes arrojan resultados coherentes y mantienen una gran similitud con las descripciones de las imágenes originales (sin distorsiones).

---

[102] MH Eggar. «Pinhole cameras, perspective, and projective geometry». En: *The American mathematical monthly* 105.7 (1998), págs. 618-630.

Para evaluar cuantitativamente el prototipo el sistema propuesto, primero se seleccionaron 618 imágenes del conjunto de datos COCO y se mostraron en un monitor; a continuación, se utilizó el prototipo de cámara para adquirir las imágenes, véase la Figura 14. De las imágenes adquiridas, se seleccionaron 500 imágenes para ajustar la red decodificadora del enfoque propuesto, como se ilustra en la Figura. 10. Se utilizaron las 118 imágenes restantes para las pruebas y se evaluó la precisión del método propuesto utilizando métricas para las descripciones. Los resultados cuantitativos se muestran en el cuadro experimental, donde podemos observar que el método propuesto alcanzó una notable puntuación *BLEU-4* del $25.3\,\%$.

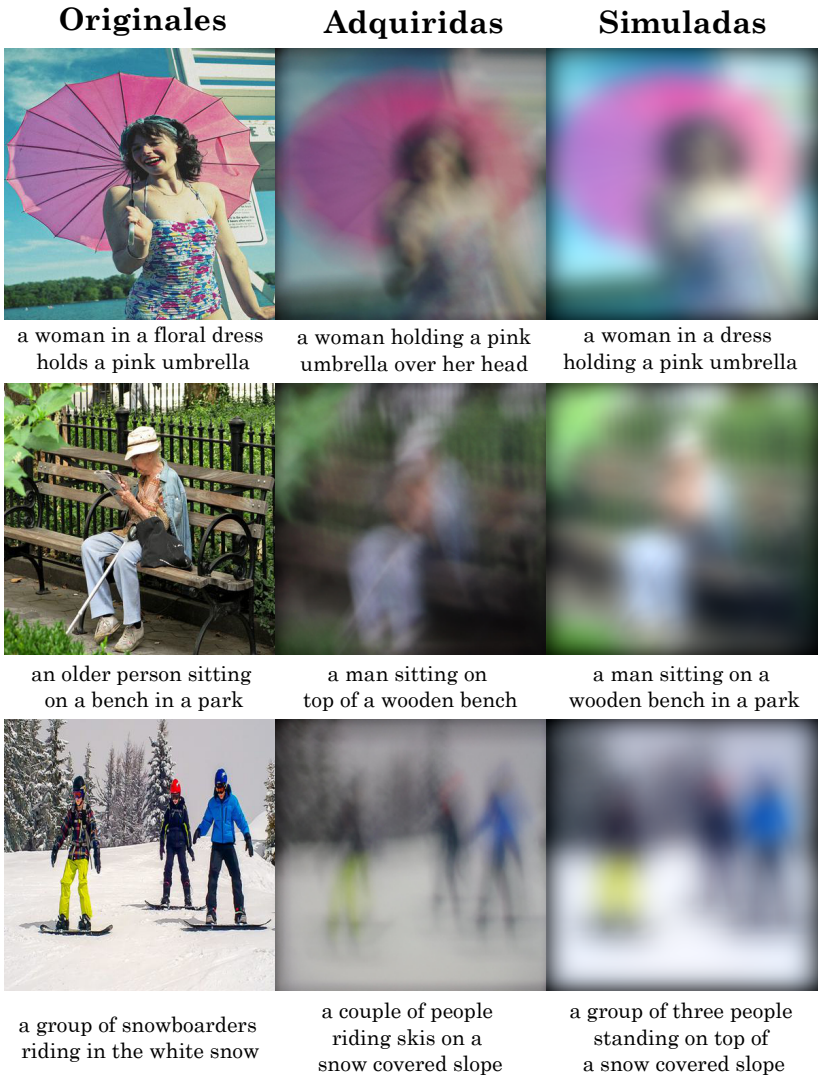

Figura 15. Resultados cualitativos de imágenes originales sin privacidad (izquierda), imágenes adquiridas de la implementación de hardware (centro) y las imágenes simuladas (derecha). Cada imagen con su correspondiente descripción.

Imágenes Totales	Imágenes de entrenamiento	Imágenes de prueba	B-1	B-2	B-3	B-4	M	C
618	500	118	63.0	45.9	33.85	25.3	24.4	74.8

Cuadro 5. Evaluación cuantitativa de las imágenes adquiridas en el laboratorio. **B-1** y **B-4** denotan las métricas BLEU-1 y BLEU-4, respectivamente, **M** representa la métrica *Meteor*, y **C** se refiere a la métrica CIDEr.

5. CONCLUSIONES

En este trabajo se presentó un modelo de generación de descripciones de imágenes basado en LSTMs, en conjunto con un módulo de atención, que integra un elemento óptico refractivo para mejorar la privacidad de las imágenes en la etapa de adquisición, mediante la introducción de distorsión. El enfoque no sólo preserva ópticamente de forma eficaz la privacidad de los individuos, objetos y lugares representados en las imágenes, sino que también consigue un rendimiento óptimo en las descripciones en varios conjuntos de datos, como COCO, Flickr8k, como demuestran las altas puntuaciones en las métricas *BLEU* y *Meteor*. Por otra parte, al realizar la implementación óptica, utilizando datos del mundo real en el hardware, también se generan descripciones ideales. Para validar la robustez del método a ataques, se implementaron técnicas de deconvolución para evaluar la solidez de la preservación de la privacidad. Los resultados de la deconvolución validan la hipótesis de que el método preserva eficazmente la privacidad al impedir la recuperación de información sensible a partir de imágenes distorsionadas. Uno de los puntos fuertes del enfoque es la preservación de la privacidad de extremo a extremo.

6. TRABAJO FUTURO

El trabajo futuro incluye la implementación de técnicas adicionales para la validación de la privacidad. Como se vió en este trabajo de grado, las imágenes adquiridas por la cámara codificadora preservan la privacidad de las personas involucradas en los escenarios en los que se encuentran, y con las pruebas realizadas de robustez a la deconvolución se pudo observar que las imágenes son resistentes a ataques de terceros. Sin embargo, para aumentar aún más la robustez del método frente a otros ataques a la privacidad que intentan recuperar las imágenes originales nítidas, más allá de los dos métodos implementados en este trabajo de grado, se podría emplear un método iterativo de recuperación basado en algoritmos. Estos métodos suelen requerir pocos datos y a menudo alcanzan soluciones óptimas, superando en algunos casos a las redes neuronales. Por lo tanto, su uso puede demostrar que las imágenes capturadas por la cámara están protegidas contra cualquier tipo de ataque.

7. Anexos

7.1. Resultados de validación de privacidad: Reconocimiento facial

Para evaluar las implicaciones para la privacidad, se llevaron a cabo experimentos adicionales para evaluar la resistencia del método propuesto de privacidad frente a una red de reconocimiento facial. Para esta prueba, se hizo uso de la red AdaFace[103] y se analizó el rendimiento en dos conjuntos de datos: el *Cross-Pose Labeled Faces in the Wild* (CPLFW)[104], y el *Labeled Faces in the Wild (LFW)*[105]. Estos conjuntos de datos contienen imágenes de caras de personas reales etiquetadas con sus respectivos nombres.

Los gráficos de la Fig. 16 representan curvas ROC (*Receiver Operating Characteristic*) utilizada comúnmente para mostrar el desempeño de un clasificador binario donde los ejes x y y estan dadas por las siguientes ecuaciones

$$\text{(Eje } y\text{): Tasa de Verdaderos Positivos} = \frac{VP}{VP + FN}, \tag{46}$$

donde VP simbolizan los *Verdaderos Positivos*, que representan los casos en los que el modelo ha predicho correctamente el reconocimiento de un rostro, y FN los *Falsos Negativos*, que se dan cuando el modelo erra al no reconocer un rostro que debería haber identificado.

---

[103] Minchul Kim, Anil K Jain y Xiaoming Liu. «Adaface: Quality adaptive margin for face recognition». En: *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

[104] Tianyue Zheng y Weihong Deng. «Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments». En: *Beijing University of Posts and Telecommunications, Tech. Rep* (2018).

[105] Xiangxin Zhu y Deva Ramanan. «Face detection, pose estimation, and landmark localization in the wild». En: *2012 IEEE Conference on CVPR*. IEEE. 2012.

$$\text{(Eje } x\text{): Tasa de Falsos Positivos} = \frac{FP}{FP + VN}, \tag{47}$$

donde FP simboliza los *Falsos Positivos*, que ocurren cuando el modelo reconoce incorrectamente un rostro que no debería haber reconocido y VN los *Verdaderos Negativos*, que se dan cuando el modelo correctamente no reconoce un rostro que no está presente. Estas tasas ayudan a evaluar cuántos casos positivos y negativos son correctamente identificados por el clasificador. Por lo tanto la curva correspondiente a un buen clasificador debe ser cercana a la esquina superior izquierda.
Los gráficos muestran que las imágenes faciales distorsionadas con nuestro sistema propuesto dio lugar a una disminución de la precisión media (AP) para el reconocimiento facial de aproximadamente $20\,\%$ en los conjuntos de datos evaluados, en comparación a los resultados con las imágenes faciales originales. Esto significa que nuestro sistema es efectivo para proteger la privacidad de las imágenes faciales, al hacer más difícil para los algoritmos de reconocimiento facial identificar correctamente a las personas.

7.2. Resultados de validación de privacidad: Reconocimiento de atributos privados

Para validar que las imágenes adquiridas protegen información sensible, se realizó una clasificación multietiqueta en el conjunto de datos *VISPR*[106], que comprende 68 clases distintas asociadas a atributos de privacidad potencialmente violados en imágenes como «desnudez», «género», «color de piel», «cultura» y «rostro». Inicialmente, utilizamos una *ResNet101* y un clasificador basado en máquinas de soporte vectorial en imágenes nítidas para establecer una línea de base para la comparación

[106] Tribhuvanesh Orekondy, Bernt Schiele y Mario Fritz. «Towards a visual privacy advisor: Understanding and predicting privacy risks in images». En: *Proc. ICCV Conf.* 2017.

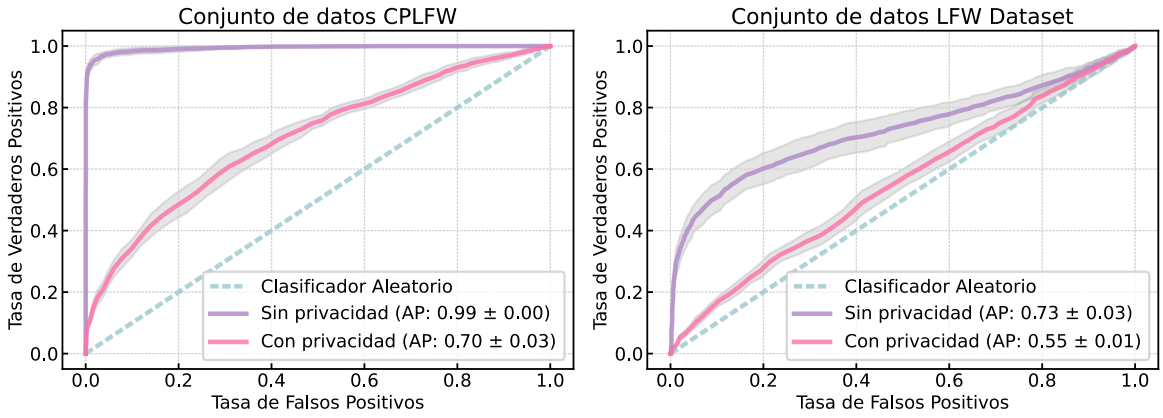

Figura 16. Curvas ROC de un modelo de reconocimiento facial en dos conjuntos de datos: CPLFW y LFW. *Sin privacidad* representa el rendimiento utilizando imágenes RGB estándar, mientras que *Con privacidad* representa los resultados utilizando imágenes privadas distorsionadas por la lente optimizada. Además, se muestra el rendimiento de un *clasificador aleatorio* con fines comparativos.

del rendimiento. En la Figura 17 se han incluido las curvas medias ROC, destacando los cinco atributos de privacidad más pertinentes: «Desnudez», «Color de Piel», «Orientación Sexual», «Género» y «Rostro» (parcial y completo).

En particular, aunque las imágenes procesadas obtienen en general peores resultados en la mayoría de las clases críticas, es esencial subrayar que el atributo del color de la piel sigue siendo especialmente vulnerable debido a que la metodología no altera el espectro de color.

Además, en la Fig. 18, se muestran resultados de cuatro atributos específicos diferentes de este experimento: cultura, edad, peso, y ocupación. Los resultados de este experimento ilustran la eficacia de nuestro enfoque para salvaguardar estos atributos frente a posibles violaciones de la privacidad.

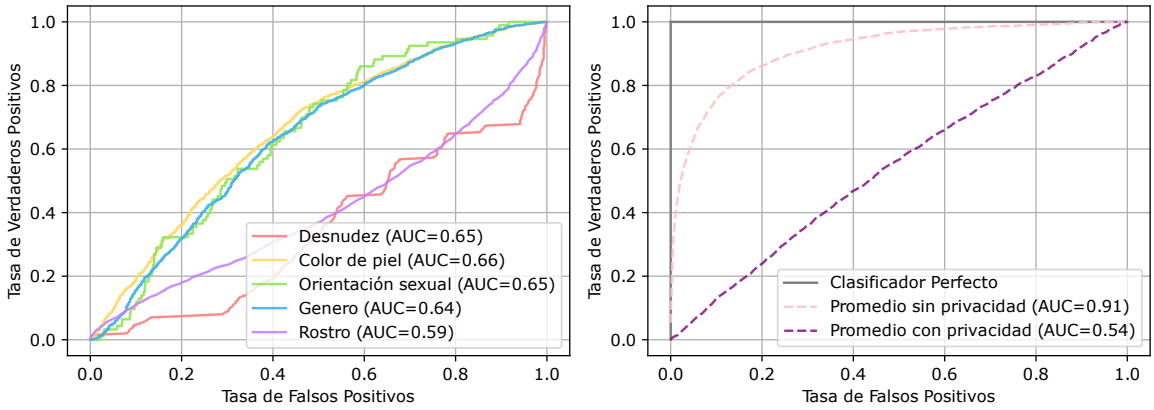

Figura 17. Evaluación de atributos de privacidad del método propuesto. Curvas ROC de cinco clases del conjunto de datos VISPR (Izquierda), junto con el reconocimiento medio de las 68 clases VISPR en las imágenes nítidas y privadas (Derecha).

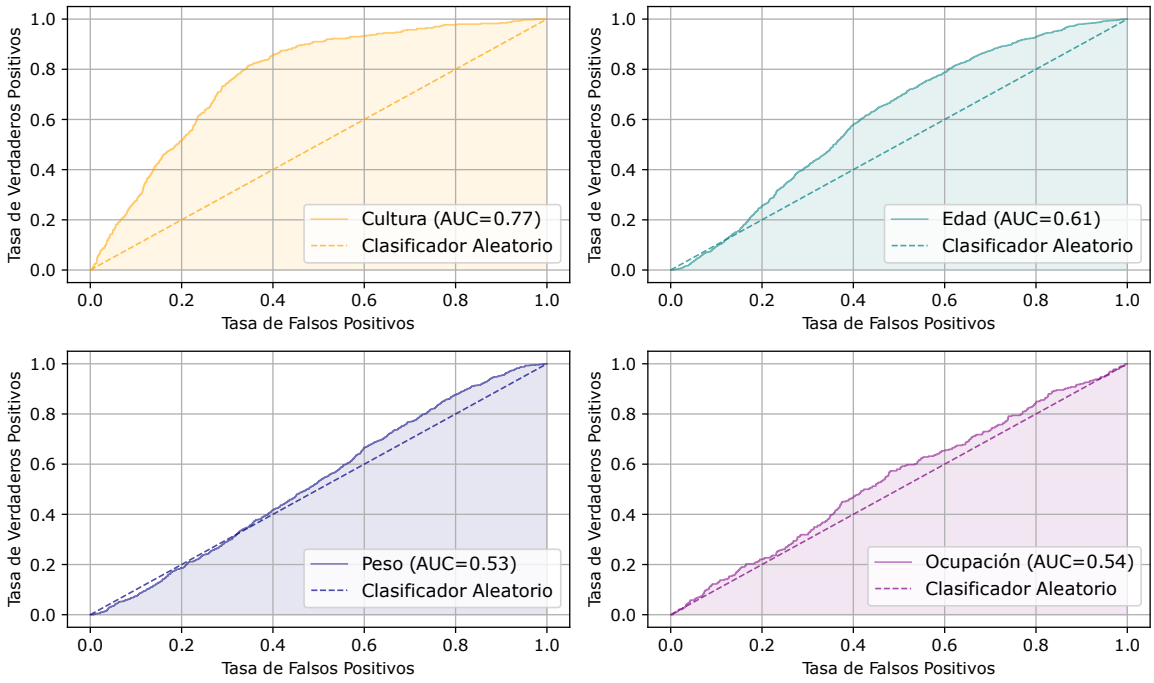

Figura 18. Evaluación de atributos de privacidad: Curvas ROC del reconocimiento de atributos de cultura, edad, peso, y ocupación en el conjunto de datos VISPR privados.

BIBLIOGRAFÍA

Agrawal, Prachi y PJ Narayanan. «Person de-identification in videos». En: *IEEE TCSVT* (2011) (vid. pág. 28).

Anderson, Peter et al. «Bottom-up and top-down attention for image captioning and visual question answering». En: *Proceedings of the IEEE Conference on CVPR*. 2018 (vid. págs. 45, 46).

Arguello, Paula et al. «Optics Lens Design for Privacy-Preserving Scene Captioning». En: *2022 IEEE ICIP*. IEEE. 2022 (vid. pág. 34).

Banerjee, Satanjeev y Alon Lavie. «METEOR: An automatic metric for MT evaluation with improved correlation with human judgments». En: *Association for Computational Linguistics*. 2005 (vid. pág. 53).

Bhalekar, Madhuri y Mangesh Bedekar. «D-CNN: A new model for generating image captions with text extraction using deep learning for visually challenged individuals». En: *Engineering, Technology & Applied Science Research* (2022) (vid. págs. 13, 19).

Christian, Hans y Mikhael Pramodana Agus. «Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)». En: *ComTech* (2016) (vid. pág. 54).

Ciregan, Dan, Ueli Meier y Jürgen Schmidhuber. «Multi-column deep neural networks for image classification». En: *2012 IEEE Conference on CVPR*. IEEE. 2012 (vid. pág. 20).

Colton, David. *Surveys on solution methods for inverse problems*. Springer Science & Business Media, 2000 (vid. pág. 37).

Du, Ling et al. «An efficient privacy protection scheme for data security in video surveillance». En: *Journal of VCIR* (2019) (vid. pág. 15).

Dubey, Shiv Ram, Satish Kumar Singh y Bidyut Baran Chaudhuri. «Activation functions in deep learning: A comprehensive survey and benchmark». En: *Neurocomputing* (2022) (vid. pág. 25).

Dun, Xiong et al. «Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging». En: *Optica* (2020) (vid. pág. 33).

Eggar, MH. «Pinhole cameras, perspective, and projective geometry». En: *The American mathematical monthly* 105.7 (1998), págs. 618-630 (vid. pág. 66).

Fan, Lijie et al. «In-Home daily-life captioning using radio signals». En: *ECCV*. Springer. 2020 (vid. pág. 36).

Goodfellow, Ian, Yoshua Bengio y Aaron Courville. *Deep learning*. MIT press, 2016 (vid. pág. 20).

Goodman, Joseph W. *Introduction to Fourier optics*. Macmillan Learning, 4 edition, 2017 (vid. pág. 31).

Hasnine, Mohammad Nehal et al. «Vocabulary learning support system based on automatic image captioning technology». En: *Distributed, Ambient and Pervasive Interactions: 7th International Conference, DAPI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*. Springer. 2019 (vid. pág. 13).

Hassan, Mohammed y Chakravarthy Bhagvati. «Structural similarity measure for color images». En: *International Journal of Computer Applications* (2012) (vid. pág. 55).

He, Kaiming et al. «Deep residual learning for image recognition». En: *Proceedings of the IEEE Conference on CVPR*. 2016 (vid. págs. 23, 44).

Hinojosa, Carlos, Juan Carlos Niebles y Henry Arguello. «Learning Privacy-preserving Optics for Human Pose Estimation». En: *IEEE/CVF ICCV*. 2021 (vid. págs. 15-17, 30, 35, 41).

Hinojosa, Carlos et al. «Privhar: Recognizing human actions from privacy-preserving lens». En: *ECCV*. Springer. 2022 (vid. págs. 15-17, 30, 35, 41, 61).

Hochreiter, Sepp et al. *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. 2001 (vid. pág. 23).

Huang, Lun et al. «Attention on attention for image captioning». En: *Proceedings of the IEEE/CVF ICCV*. 2019 (vid. págs. 14, 19).

Karpathy, Andrej y Li Fei-Fei. «Deep visual-semantic alignments for generating image descriptions». En: *Proceedings of the IEEE Conference on CVPR*. 2015 (vid. pág. 58).

Khan, Salman et al. *A guide to convolutional neural networks for computer vision*. Springer, 2018 (vid. pág. 22).

Khan, Salman et al. «Transformers in vision: A survey». En: *ACM Computing Surveys (CSUR)* (2022) (vid. pág. 26).

Kim, Minchul, Anil K Jain y Xiaoming Liu. «Adaface: Quality adaptive margin for face recognition». En: *Proceedings of the IEEE/CVF Conference on CVPR*. 2022 (vid. pág. 71).

Kupyn, Orest et al. «Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better». En: *Proceedings of the IEEE/CVF ICCV*. 2019 (vid. págs. 39, 61).

Lakshminarayanan, Vasudevan y Andre Fleck. «Zernike polynomials: a guide». En: *Journal of Modern Optics* 58.7 (2011), págs. 545-561 (vid. págs. 36, 66).

Leotta, Maurizio, Fabrizio Mori y Marina Ribaudo. «Evaluating the effectiveness of automatic image captioning for web accessibility». En: *Universal access in the information society* (2023) (vid. págs. 13, 19).

Li, Zewen et al. «A survey of convolutional neural networks: analysis, applications, and prospects». En: *IEEE TNNLS* (2021) (vid. pág. 22).

Lin, Chin-Yew. «Rouge: A package for automatic evaluation of summaries». En: *Text summarization branches out*. 2004 (vid. pág. 54).

Lin, Tsung-Yi et al. «Microsoft coco: Common objects in context». En: *ECCV*. Springer. 2014 (vid. págs. 19, 50).

Liu, Chi et al. «Privacy intelligence: A survey on image privacy in online social networks». En: *ACM Computing Surveys* (2022) (vid. pág. 15).

Liu, Xiaoxiao, Qingyang Xu y Ning Wang. «A survey on deep neural network-based image captioning». En: *The Visual Computer* (2019) (vid. pág. 21).

MacLeod, Haley et al. «Understanding blind people's experiences with computer-generated captions of social media images». En: *proceedings of the 2017 CHI CAFCS*. 2017 (vid. págs. 13, 19).

Manheim, Karl y Lyric Kaplan. «Artificial intelligence: Risks to privacy and democracy». En: *Yale JL & Tech.* (2019) (vid. págs. 14, 28).

Mao, Anqi, Mehryar Mohri y Yutao Zhong. «Cross-entropy loss functions: Theoretical analysis and applications». En: *International conference on Machine learning*. PMLR. 2023, págs. 23803-23828 (vid. pág. 48).

Martin, Antoinette Deborah, Ezat Ahmadzadeh e Inkyu Moon. «Privacy-Preserving Image Captioning with Deep Learning and Double Random Phase Encoding». En: *Mathematics* (2022) (vid. págs. 15, 36).

Medsker, Larry R y LC Jain. «Recurrent neural networks». En: *Design and Applications* (2001) (vid. pág. 24).

Nehme, Elias et al. «DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning». En: *Nature Methods* (2020) (vid. pág. 33).

Ordonez, Vicente, Girish Kulkarni y Tamara Berg. «Im2Text: Describing Images Using 1 Million Captioned Photographs». En: *Advances in Neural Information Processing Systems*. Ed. por J. Shawe-Taylor et al. Curran Associates, Inc., 2011 (vid. pág. 50).

Orekondy, Tribhuvanesh, Bernt Schiele y Mario Fritz. «Towards a visual privacy advisor: Understanding and predicting privacy risks in images». En: *Proc. ICCV Conf.* 2017 (vid. pág. 72).

Padilla-López, José Ramón, Alexandros Andre Chaaraoui y Francisco Flórez-Revuelta. «Visual privacy protection methods: A survey». En: *Expert Systems with Applications* (2015) (vid. pág. 28).

Papineni, Kishore et al. «Bleu: a method for automatic evaluation of machine translation». En: *Association for Computational Linguistics*. 2002 (vid. pág. 52).

Pittaluga, Francesco, Sanjeev Koppal y Ayan Chakrabarti. «Learning privacy preserving encodings through adversarial training». En: *2019 IEEE WACV*. IEEE. 2019 (vid. pág. 28).

Pittaluga, Francesco y Sanjeev J Koppal. «Privacy preserving optics for miniature vision sensors». En: *Proceedings of the IEEE Conference on CVPR*. 2015 (vid. págs. 16, 29).

Pittaluga, Francesco y Sanjeev Jagannatha Koppal. «Pre-capture privacy for small vision sensors». En: *IEEE TPAMI* (2016) (vid. págs. 16, 29).

Poor, H. «On robust Wiener filtering». En: *IEEE TAC* (1980) (vid. pág. 37).

Qiu, Jianing et al. «Egocentric image captioning for privacy-preserved passive dietary intake monitoring». En: *IEEE Transactions on Cybernetics* (2023) (vid. págs. 16, 36).

Ravi, Siddharth, Pau Climent-Pérez y Francisco Florez-Revuelta. «A review on visual privacy preservation techniques for active and assisted living». En: *Multimedia Tools and Applications* (2023) (vid. pág. 15).

Ryoo, Michael, Kiyoon Kim y Hyun Yang. «Extreme Low Resolution Activity Recognition With Multi-Siamese Embedding Learning». En: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018 (vid. pág. 29).

Ryoo, Michael S et al. «Privacy-preserving human activity recognition from extreme low resolution». En: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017 (vid. pág. 29).

Satish, Pooja, Mallikarjunaswamy Srikantaswamy y Nataraj Kanathur Ramaswamy. «A Comprehensive Review of Blind Deconvolution Techniques for Image Deblurring.» En: *Traitement du Signal* (2020) (vid. pág. 37).

Shechtman, Yoav et al. «Optimal point spread function design for 3D imaging». En: *Physical Review Letters* (2014) (vid. pág. 34).

Shuster, Kurt et al. «Engaging Image Captioning via Personality». En: *Proceedings of the IEEE/CVF Conference on CVPR*. Jun. de 2019 (vid. pág. 13).

Sitzmann, Vincent et al. «End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging». En: *ACM Transactions on Graphics (TOG)* (2018) (vid. págs. 15, 16, 30, 35, 41).

Sivaprakash, Asokan, Samuel NE Rajan y Sundaramoorthy Selvaperumal. «Privacy protection of patient medical images using digital watermarking technique for E-healthcare system». En: *Current Medical Imaging* (2019) (vid. pág. 15).

Staniūtė, Raimonda y Dmitrij Šešok. «A Systematic Literature Review on Image Captioning». En: *Applied Sciences* (2019) (vid. págs. 19, 22).

Szegedy, Christian, Alexander Toshev y Dumitru Erhan. «Deep neural networks for object detection». En: *Advances in neural information processing systems* (2013) (vid. pág. 21).

Vaswani, Ashish et al. «Attention is all you need». En: *Advances in neural information processing systems* (2017) (vid. pág. 26).

Vedantam, Ramakrishna, C Lawrence Zitnick y Devi Parikh. «Cider: Consensus-based image description evaluation». En: *Proceedings of the IEEE Conference on CVPR*. 2015 (vid. pág. 53).

Vinyals, Oriol et al. «Show and tell: A neural image caption generator». En: *IEEE/CVF Conference on CVPR*. 2015 (vid. págs. 14, 20).

Wang, Qingzhong, Jia Wan y Antoni B Chan. «On diversity in image captioning: Metrics and methods». En: *IEEE TPAMI* (2020) (vid. pág. 13).

Wang, Zihao W et al. «Privacy-preserving action recognition using coded aperture videos». En: *Proceedings of the IEEE Conference on CVPR Workshops*. 2019 (vid. pág. 29).

Wiatowski, Thomas y Helmut Bölcskei. «A mathematical theory of deep convolutional neural networks for feature extraction». En: *IEEE TIT* (2017) (vid. pág. 22).

Wikipedia. *Wiener-Filter*. `https://de.wikipedia.org/wiki/Wiener-Filter`. Accedido 17-02-2024. 2023 (vid. pág. 38).

Wu, Qi et al. «Image captioning and visual question answering based on attributes and external knowledge». En: *IEEE TPAMI* (2017) (vid. pág. 19).

Wu, Zhenyu et al. «Privacy-Preserving Deep Action Recognition: An Adversarial Learning Framework and A New Dataset». En: *IEEE TPAMI* (2020) (vid. pág. 28).

Xu, Kelvin et al. «Show, attend and tell: Neural image caption generation with visual attention». En: *ICML*. PMLR. 2015 (vid. pág. 14).

Yao, Lirong y Yazhuo Guan. «An Improved LSTM Structure for Natural Language Processing». En: *2018 IEEE IICSPI*. 2018 (vid. pág. 14).

Yu, Jinao et al. «Gan-based differential private image privacy protection framework for the internet of multimedia things». En: *Sensors* (2020) (vid. pág. 15).

Zheng, Tianyue y Weihong Deng. «Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments». En: *Beijing University of Posts and Telecommunications, Tech. Rep* (2018) (vid. pág. 71).

Zhu, Xiangxin y Deva Ramanan. «Face detection, pose estimation, and landmark localization in the wild». En: *2012 IEEE Conference on CVPR*. IEEE. 2012 (vid. pág. 71).