DISEÑO DE UN ALGORITMO COMPUTACIONAL BASADO EN APRENDIZAJE PROFUNDO PARA LA OPTIMIZACIÓN DE UN SISTEMA ÓPTICO DE ADQUISICIÓN DE IMÁGENES QUE PRESERVAN LA PRIVACIDAD PARA LA ESTIMACIÓN DE ACCIONES EN ENTORNOS CLÍNICOS

DAVID SANTIAGO MORALES NORATO
Ingeniero de Sistemas

Tesis presentada como requisito parcial para optar al título de Magíster en Matemática Aplicada.

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE FÍSICA
BUCARAMANGA

2025

DISEÑO DE UN ALGORITMO COMPUTACIONAL BASADO EN APRENDIZAJE PROFUNDO PARA LA OPTIMIZACIÓN DE UN SISTEMA ÓPTICO DE ADQUISICIÓN DE IMÁGENES QUE PRESERVAN LA PRIVACIDAD PARA LA ESTIMACIÓN DE ACCIONES EN ENTORNOS CLÍNICOS

DAVID SANTIAGO MORALES NORATO

Tesis de Maestría para optar al título de
Magíster en Matemática Aplicada

Director:
Henry Arguello Fuentes
Ph.D. Electrical and Computer Engineering

Codirector:
Hoover Fabián Rueda Chacón
Ph.D. Electrical and Computer Engineering

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE FÍSICA
MAESTRÍA EN MATEMÁTICA APLICADA
BUCARAMANGA
2025

DEDICATORIA

Para mi mamá Mariana, mi papá Hanssel,
mis hermanas María Silvana, Katherin y mi hermano Enrrique.
ellos son el círculo de personas que más amo.
Este trabajo de grado es el resultado de todo el esfuerzo y apoyo que mi familia me ha brindado, todo lo que soy y lo que seré, se lo debo a ellos.

AGRADECIMIENTOS

Agradezco al grupo de investigación HDSP por fomentar un entorno de excelencia, investigación y ambición. Este equipo me ha brindado un apoyo invaluable, tanto en lo académico como en lo personal, desde mi incorporación al grupo.

A la Universidad Pública y en especial a la Universidad Industrial de Santander, la institución que me proporcionó las herramientas necesarias para crecer, formarme profesionalmente y formalizar mi curiosidad científica.

Mi formación de maestría fue respaldada económicamente gracias a la convocatoria "Formación de Capital Humano de Alto Nivel para el Departamento de Santander" BPIN 2020000100536, impulsada por la Universidad Industrial de Santander y el Ministerio de Ciencia, Tecnología e Innovación de Colombia.

CONTENIDO

pág.

LISTA DE FIGURAS

pág.

LISTA DE CUADROS

pág.

RESUMEN

TÍTULO: DISEÑO DE UN ALGORITMO COMPUTACIONAL BASADO EN APRENDIZAJE PROFUNDO PARA LA OPTIMIZACIÓN DE UN SISTEMA ÓPTICO DE ADQUISICIÓN DE IMÁGENES QUE PRESERVAN LA PRIVACIDAD PARA LA ESTIMACIÓN DE ACCIONES EN ENTORNOS CLÍNICOS [*]

AUTOR: DAVID SANTIAGO MORALES NORATO [**]

PALABRAS CLAVE: Elementos ópticos difractivos, privacidad, reconocimiento de acciones, preservación de la privacidad visual, aprendizaje multimodal.

DESCRIPCIÓN: La preservación de la privacidad en sistemas de visión por computadora es esencial, especialmente en aplicaciones médicas donde se manejan datos sensibles de los pacientes. Las soluciones actuales, como sistemas óptico-computacionales que distorsionan las imágenes desde su adquisición, presentan limitaciones en el balance entre preservación efectiva de la privacidad y el desempeño en estimación de acciones. Este trabajo propone un algoritmo para optimizar un sistema óptico mediante la combinación de técnicas multimodales y parametrizaciones avanzadas de elementos ópticos difractivos (EODs), regularizados por métricas como el número de condición y la función de transferencia de modulación (MTF). Las funciones de regularización junto a las parametrizaciones de EODs propuestas promueven distorsiones del campo óptico más fuertes que las superficies tradicionales basadas en polinomios de Zernike, incrementando la privacidad visual. Además, la integración de datos visuales y textuales mediante un esquema multimodal mejora el rendimiento en la estimación de acciones y permite realizar estimaciones en escenarios con datos limitados mediante la técnica *Zero-shot*. El método fue evaluado utilizando conjuntos de datos de estimación de acciones en entornos clínicos, analizando tanto su rendimiento en estimación de acciones como sus propiedades ópticas mediante métricas como la resolución, la MTF y el número de condición, demostrando su eficacia en aplicaciones clínicas.

---

[*] Tesis de Maestría

[**] Facultad de Ciencias. Escuela de Física. Director: Henry Arguello Fuentes. Codirector: Hoover Fabián Rueda Chacón.

ABSTRACT

TITLE: End-to-End Design of Diffractive Optical Elements for Privacy-Preserving Action Recognition Systems in Clinical Environments [*]

AUTHOR: DAVID SANTIAGO MORALES NORATO [**]

KEYWORDS: Diffractive optical elements, Privacy-Preserving Action recognition, Multimodal learning.

DESCRIPTION: Privacy-preserving computer vision systems are essential, especially in medical applications where sensitive patient data is processed. Current methods involve Zernike-based optical-computational systems that distort the optical field before image acquisition; however, they face challenges in effectively preserving privacy while ensuring action recognition performance. This work proposes an algorithm to optimize an optical system through multimodal techniques and advanced parametrizations of diffractive optical elements (DOE), regularized by the condition number of the convolution matrix generated by the point spread function (PSF) and the modulation transfer function (MTF). The proposed regularization functions ensure stronger distortions than traditional Zernike-based continuous surfaces, enhancing visual privacy. Additionally, the multimodal scheme integrates visual and text information, improving action recognition performance and enabling estimations in data-limited scenarios through zero-shot learning. The proposed method was evaluated using multiple action recognition datasets, including those from clinical environments, to assess performance and optical properties. The evaluation included metrics such as resolution, MTF, and condition number, demonstrating that the optimized optical system successfully balances privacy and recognition performance. This approach offers a reliable framework for developing secure and efficient computer vision systems in clinical applications, ensuring that sensitive data remains protected without compromising functionality.

---

[*] Master Thesis

[**] Faculty of Sciences. Department of Physics. Advisor: Henry Arguello Fuentes. Co-advisor: Hoover Fabián Rueda Chacón

INTRODUCCIÓN

Los sistemas de visión por computadora han impulsado el desarrollo de productos capaces de extraer información valiosa de grandes conjuntos de datos, permitiendo la creación de sistemas avanzados de clasificación, detección y reconocimiento de acciones humanas, entre otros. Estas aplicaciones han tenido un impacto significativo en diversas industrias, como la agricultura[1], el transporte[2] y el sector salud[3]. En el ámbito de la salud, las tecnologías de visión por computadora facilitan la creación de herramientas para asistir, complementar y mejorar las actividades de cuidado[4,5], prevención[6] y diagnóstico[7,8] de enfermedades que afectan el bienestar de las personas. Estudios recientes han demostrado que es posible diagnosticar enfermedades con alta precisión mediante algoritmos de reconocimiento de acciones aplicados a

---

1 Andreas Kamilaris y Francesc X Prenafeta-Boldú. «Deep learning in agriculture: A survey». En: *Computers and Electronics in Agriculture* 147 (2018), págs. 70-90.

2 Sorin Grigorescu et al. «A survey of deep learning techniques for autonomous driving». En: *Journal of Field Robotics* 37.3 (2020), págs. 362-386.

3 Francesco Piccialli et al. «A survey on deep learning in medicine: Why, how and when?» En: *Information Fusion* 66 (2021), págs. 111-137.

4 Yiwen Xu et al. «Deep learning predicts lung cancer treatment response from serial medical imaging». En: *Clinical Cancer Research* 25.11 (2019), págs. 3266-3275.

5 Theodore Sakellaropoulos et al. «A deep learning framework for predicting response to therapy in cancer». En: *Cell Reports* 29.11 (2019), págs. 3367-3373.

6 Sung Hyun Kim et al. «Animal Infectious Diseases Prevention through Big Data and Deep Learning». En: *Journal of Intelligence and Information Systems* 24.4 (2018), págs. 137-154.

7 Ninon Burgos et al. «Deep learning for brain disorders: from data processing to disease treatment». En: *Briefings in Bioinformatics* 22.2 (2021), págs. 1560-1576.

8 Jeffrey De Fauw et al. «Clinically applicable deep learning for diagnosis and referral in retinal disease». En: *Nature Medicine* 24.9 (2018), págs. 1342-1350.

imágenes o videos[9]. Un caso destacado es la detección de la enfermedad de Parkinson, que puede lograrse a través del análisis de videos que estimaan las acciones y patrones de marcha de un paciente[10].

Tradicionalmente, los sistemas de estimación de acciones han utilizado modelos unimodales[11], es decir, modelos que se entrenan con un solo tipo de datos, como los videos. Estos métodos buscan aprovechar la información temporal presente en los videos mediante diversas técnicas[11,12,13,14], incluyendo arquitecturas como los *Transformers*[15]. Sin embargo, recientemente, con el auge de los grandes modelos de lenguaje y los llamados modelos fundacionales[16], ha crecido el interés por desarrollar modelos multimodales capaces de combinar información de diversas fuentes, incluyendo imágenes, texto y audio, para mejorar el reconocimiento de acciones, ya

---

[9] Fabio Martínez, Francisco Gómez y Eduardo Romero. «Análisis de vídeo para estimación del movimiento humano: una revisión». En: *Revista Med* 17.1 (2009), págs. 95-106.

[10] Luis C Guayacán, Brayan Valenzuela y Fabio Martinez. «Parkinsonian gait characterization from regional kinematic trajectories». En: *14th International Symposium on Medical Information Processing and Analysis*. Vol. 10975. International Society for Optics y Photonics. 2018, pág. 1097502.

[11] Christoph Feichtenhofer, Axel Pinz y Richard P. Wildes. «Spatiotemporal residual networks for video action recognition». En: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Curran Associates Inc., 2016, 3476–3484.

[12] Du Tran et al. «Learning spatiotemporal features with 3d convolutional networks». En: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, págs. 4489-4497.

[13] Linxi Fan et al. «Rubiksnet: Learnable 3d-shift for efficient video action recognition». En: *European Conference on Computer Vision*. Springer. 2020, págs. 505-521.

[14] Wenbo Li et al. «Adaptive RNN Tree for Large-Scale Human Action Recognition». En: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[15] Zhiwu Qing et al. «Mar: Masked autoencoders for efficient action recognition». En: *IEEE Transactions on Multimedia* (2023).

[16] Rishi Bommasani et al. «On the opportunities and risks of foundation models». En: *arXiv preprint arXiv:2108.07258* (2021).

que aprovechan la información semántica de frases que describen cada acción.
A medida que estas tecnologías avanzan y se integran en aplicaciones sensibles como la medicina, surge una preocupación fundamental: la privacidad y seguridad de los datos involucrados. Por lo tanto, en los últimos años, ha aumentado exponencialmente el interés por preservar y regular la privacidad de los datos en los sistemas de visión por computadora. A nivel mundial, y en particular en Europa[17], se han establecido regulaciones con el objetivo de garantizar que los datos de los usuarios sean adquiridos y procesados de manera anónima[18,19]. En Colombia, la Ley Estatutaria 1581 de 2012[20] para la protección de datos personales contiene preceptos que regulan el tratamiento de datos sensibles, entre los cuales se incluyen los datos biométricos. Dicha ley establece que la adquisición, almacenamiento y tratamiento de información sensible está prohibida, salvo en casos donde exista un acuerdo previo de consentimiento y confidencialidad entre ambas partes. El cumplimiento de estas regulaciones es de vital importancia, ya que se ha demostrado que si los datos utilizados en la etapa de entrenamiento de un algoritmo de aprendizaje profundo no consideran la privacidad, el sistema está sujeto a brechas de seguridad[21]. Estas

---

17 European Parliament Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*. Mayo de 2016.

18 Christopher G Schwarz et al. «Identification of anonymous MRI research participants with face-recognition software». En: *New England Journal of Medicine* 381.17 (2019), págs. 1684-1686.

19 Arvind Narayanan y Vitaly Shmatikov. «Robust de-anonymization of large sparse datasets». En: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, págs. 111-125.

20 EL CONGRESO DE COLOMBIA. *LEY ESTATUTARIA 1581 DE 2012*. Oct. de 2012.

21 Fanyu Bu et al. «Privacy preserving back-propagation based on BGV on cloud». En: *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*. IEEE. 2015, págs. 1791-1795.

brechas pueden ocurrir a través de los parámetros del modelo o de las predicciones, violando así la privacidad de los usuarios. En el área de visión por computadora aplicada a la medicina, ha surgido un creciente interés por proteger la privacidad de ciertos atributos sensibles de los pacientes, tales como el género, el color de piel, la desnudez y la identidad, entre otros[22].

En el contexto de entornos clínicos, las imágenes de pacientes pueden clasificarse en dos categorias: internas y externas, según la tecnología utilizada para su adquisición. En escenarios donde la información corresponde a órganos internos, se emplean imágenes de tomografías computarizadas, resonancias magnéticas, gammagrafías o ultrasonidos. En estos casos, el anonimato de los datos es más fácil de garantizar, ya que las imágenes no contienen información que permita identificar elementos como el color de piel, la desnudez o rasgos faciales. Sin embargo, en situaciones donde se utilizan sistemas de cámaras convencionales, como cámaras RGB, y se capturan imágenes del cuerpo externo del paciente, por ejemplo, en tratamientos y consultas de fisioterapia, monitoreo de cuidados intensivos o vigilancia de salas de espera, la identidad y privacidad del paciente se ven directamente comprometidas. Por lo tanto, es de gran interés desarrollar técnicas que permitan preservar la privacidad de los pacientes en los datos adquiridos, los cuales sirven como entrada para algoritmos de visión por computadora.

El estado del arte ha propuesto tanto estrategias basadas en *hardware* como en *software* para mantener la privacidad en sistemas de visión por computadora. Respecto a las soluciones de *software*, destacan aquellas que aplican técnicas de posprocesamiento, como el desenfoque de imágenes[23] o el uso de imágenes de menor

---

[22] Georgios Kaissis et al. «End-to-end privacy preserving deep learning on multi-institutional medical imaging». En: *Nature Machine Intelligence* 3.6 (2021), págs. 473-484.

[23] Francesco Pittaluga y Sanjeev Jagannatha Koppal. «Pre-capture privacy for small vision sensors». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.11 (2016),

resolución[24]. Sin embargo, estos enfoques se aplican únicamente después de la adquisición de los datos, lo que crea una brecha de seguridad desde la captura en el sensor hasta que los datos son procesados por estas técnicas.

Por otro lado, las soluciones basadas en *hardware* han desarrollado sistemas ópticocomputacionales que permiten preservar la privacidad antes de adquirir la información del paciente[25,26,2728]. Generalmente, estos trabajos proponen incluir un elemento óptico diseñable cuya superficie se modela utilizando la base de polinomios de Zernike[29,30]. Esto permite diseñar el sistema óptico mediante el aprendizaje de los coeficientes que acompañan a la base, utilizando aprendizaje profundo, con una técnica llamada entrenamiento de extremo a extremo[25,28,26,31]. Aunque estos méto-

---

págs. 2215-2226.

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

[25] Carlos Hinojosa, Juan Carlos Niebles y Henry Arguello. «Learning Privacy-preserving Optics for Human Pose Estimation». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 2573-2582.

[26] Carlos Hinojosa et al. «PrivHAR: Recognizing Human Actions from Privacy-Preserving Lens». En: *Computer Vision – ECCV 2022*. Ed. por Shai Avidan et al. Cham: Springer Nature Switzerland, 2022, págs. 314-332.

[27] Paula Arguello et al. «Optics Lens Design for Privacy-Preserving Scene Captioning». En: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, págs. 3551-3555.

[28] Paula Arguello et al. «Learning to Describe Scenes via Privacy-Aware Designed Optical Lens». En: *IEEE Transactions on Computational Imaging* 10 (2024), págs. 1069-1079. DOI: 10.1109/TCI.2024.3426975.

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

[30] Robert J. Noll. «Zernike polynomials and atmospheric turbulence∗». En: *J. Opt. Soc. Am.* 66.3 (1976), págs. 207-211.

[31] Henry Arguello et al. «Deep Optical Coding Design in Computational Imaging: A data-driven framework». En: *IEEE Signal Processing Magazine* 40.2 (2023), págs. 75-88.

dos logran superficies continuas y suaves, similares a las de los lentes refractivos, que provocan una distorsión en las imágenes que dificulta invertir y recuperar la información sensible de las escenas[26], el estado del arte no ha explorado el uso de parametrizaciones que generan superficies discontinuas, también conocidas como elementos ópticos difractivos[32]. Estas superficies discontinuas tienen el potencial de generar distorsiones más fuertes en las imágenes adquiridas y, por lo tanto, lograr una mayor privacidad visual.

Por otro lado, en el ámbito de la preservación de la privacidad y la estimación de acciones, el estado del arte ha implementado únicamente métodos basados en el paradigma unimodal, por lo que no ha contemplado el esquema multimodal, el cual combina múltiples fuentes de información para mejorar el desempeño en tareas de visión por computadora. La integración de enfoques multimodales puede potenciar la capacidad de los sistemas para reconocer acciones mientras se mantiene la privacidad de los datos. Por consiguiente, existe una oportunidad de investigar y desarrollar estas parametrizaciones de elementos ópticos difractivos (EODs) y aplicar esquemas multimodales para mejorar la preservación de la privacidad en sistemas de visión por computadora aplicados a entornos clínicos.

Para enfocar el diseño en una tarea específica de visión por computadora, estos trabajos minimizan una función de costo asociada a la tarea mientras maximizan la privacidad utilizando normas ℓ^p[25,28] o el índice de similitud estructural (SSIM)[26]. Sin embargo, optimizar estas métricas no es adecuado para cuantificar la privacidad, ya que existen transformaciones invertibles donde las métricas, muestran un alto valor de ℓ^p o SSIM, pero no se preserva la privacidad. Por lo tanto, es necesario plantear nuevas funciones de costo encargadas de maximizar la privacidad visual, mientras que permiten realizar las tareas de visión por computadora.

---

[32] Xin Liu et al. «Investigating deep optics model representation in affecting resolved all-in-focus image quality and depth estimation fidelity». En: *Opt. Express* 30.20 (2022), págs. 36973-36984.

En este trabajo, proponemos el uso de esquemas multimodales en la estimación de acciones, combinando diferentes modalidades de datos para mejorar el rendimiento en la tarea específica. Asimismo, investigamos diferentes parametrizaciones del elemento óptico a diseñar, con el objetivo de lograr superficies discontinuas o EODs que distorsionen las imágenes adquiridas de manera más efectiva. De esta forma, buscamos maximizar la privacidad visual utilizando métricas de invertibilidad y calidad, como el número de condición y la MTF, abordando así las limitaciones presentes en el estado del arte y aportando soluciones innovadoras en entornos clínicos.

NOTACIÓN Y NOMENCLATURA

Nomenclatura

EOD	Elemento óptico difractivo
PSF	Función de dispersión de punto
MTF	Función de transferencia de modulación
SSIM	Índice de similaridad estructural
ActionCLIP+PrivHAR	Método de linea base usado para las comparaciones

Notación general

$\mathbb{R}$	Conjunto de números reales
a	Escalar
$\mathbf{a}$	Vector
$\mathbf{A}$	Matriz
$\mathbf{A}^{-1}$	Matriz inversa
$\mathcal{A}$	Tensor
$\|\cdot\|_p$	Norma ℓ^p
$\mathcal{K}(\mathbf{A})$	Número de condición de la matriz $\mathbf{A}$
i	Constante imaginaria $\sqrt{-1}$
$\mathrm{mod}(\cdot)$	Función módulo
$\mathrm{rect}(\cdot)$	Función rectángulo
$\mathrm{circ}(\cdot)$	Función círculo
$\mathrm{supp}(\cdot)$	Conjunto de soporte de una función
$\mathbf{Z}_j$	j-ésimo término de la base de Zernike
$\mathcal{F}\{\cdot\}$	Operador transformada directa de Fourier
$\mathcal{F}^{-1}\{\cdot\}$	Operador transformada inversa de Fourier
δ	Delta de Dirac
$\mathbf{\Lambda}$	Vector de valores propios

Notación de imágenes computacionales

$\mathbf{x}$	Vector que representa la escena discreta
$\mathbf{y}$	Vector que representa la medida discreta
(x, y)	Coordenadas espaciales
λ	Longitud de onda
k_0	Número de onda
$n(\lambda)$	Índice de refracción
ϕ	Mapa de alturas de un elemento óptico
Φ	Parámetros de un elemento óptico difractivo
Ψ	Modulación en fase de un elemento óptico
h_Φ	función que representa la PSF continua de un sistema óptico
$\mathcal{H}_\Phi$	Tensor que representa la PSF discreta
$\mathbf{H}_{\Phi_k}$	Filtro 2D correspondiente al k-ésimo canal de color de $\mathcal{H}_\Phi$
$\mathbf{C}_{\Phi_k}$	Matriz de convolución lineal para el filtro $\mathbf{H}_{\Phi_k}$
$\mathring{\mathbf{C}}_{\Phi_k}$	Matriz de convolución circular para el filtro $\mathbf{H}_{\Phi_k}$
$\mathbf{C}_\Phi$	Matriz de sensado, construida a partir de cada $\mathbf{C}_{\Phi_k}$
$\mathring{\mathbf{C}}_\Phi$	Matriz de sensado, construida a partir de cada matriz de convolución circular $\mathring{\mathbf{C}}_{\Phi_k}$
$\mathcal{R}_{\mathsf{SC}}(\Phi)$	Función de regularización de soporte compacto
$\mathcal{R}_{\mathrm{cond}}(\Phi)$	Función de regularización del número de condición
$\mathcal{R}_{\mathrm{MTF}}(\Phi)$	Función de regularización de la MTF

Notación estimación de acciones

$\mathcal{L}_{tarea}(\cdot, \cdot)$	Función de costo para la tarea de estimación de acciones
$\mathcal{L}_{privacidad}(\cdot, \cdot)$	Función de costo que promueve la privacidad
$\mathcal{V}_\Theta(\cdot)$	Red neuronal para la extracción de características visuales
Θ	Parámetros entrenables de la red neuronal $\mathcal{V}_\Theta$
$\mathcal{T}_\Xi(\cdot)$	Red neuronal para la extracción de características textuales

Ξ	Parámetros entrenables de la red neuronal $\mathcal{T}_\Xi$
$\mathbf{d}$	Vector que representa la etiqueta de estimación de acciones
$\mathcal{D}$	Conjunto de datos compuesto por parejas de escenas y acciones
$\mathbf{f}^{\text{visual}}$	Vector de características visuales de una escena
$\mathbf{F}^{\text{texto}}$	Matriz de características textuales de un conjunto de etiquetas de acciones
$\mathbb{E}_{\mathbf{x},\mathbf{d}\sim\mathcal{D}}\,\mathcal{L}(\cdot,\cdot)$	Valor esperado de una función de costo $\mathcal{L}$ evaluada sobre el dataset $\mathcal{D}$
$S(\cdot,\cdot)$	Función de similaridad
$\mathcal{R}(\cdot)$	Función de regularización sobre algunos pesos entrenables

OBJETIVOS

Objetivo general: Diseñar un algoritmo de aprendizaje profundo orientado a la optimización de un sistema óptico difractivo para la adquisición de imágenes que preservan la privacidad para la estimación de acciones en entornos cliínicos.

Objetivos específicos

- Modelar matemáticamente el proceso de adquisición de imágenes desde un sistema óptico difractivo diferenciable que preserve la privacidad de los sujetos presentes en la escena.
- Optimizar los parámetros de un sistema óptico difractivo mediante un algoritmo de aprendizaje profundo para la adquisición de imágenes que preservan la privacidad y permiten la estimación de acciones en entornos clínicos.
- Evaluar el desempeño del algoritmo computacional en términos de tiempo de ejecución, preservación de la privacidad y reconocimiento de acciones.

1. MARCO TEÓRICO Y ESTADO DEL ARTE

1.1. PRIVACIDAD VISUAL EN SISTEMAS DE VISIÓN POR COMPUTADORA

Debido al avance significativo de los sistemas de reconocimiento basados en inteligencia artificial en industrias como los deportes[33], la seguridad[34] y, particularmente, los entornos clínicos[3], la protección de la privacidad visual de las personas se ha convertido en un área de estudio de creciente importancia dentro de la comunidad científica[35]. En la literatura, los métodos para proteger la privacidad visual se dividen en enfoques basados en *software* y *hardware*. En la Figura 1, se presenta una comparación visual de cómo una cámara de baja resolución espacial[24], un lente convencional desenfocado y el lente optimizado propuesto en[25,26,27] logran preservar la privacidad en la escena.

1.1.1. Métodos basados en *Software*. Los métodos de privacidad basados en software son aquellos que actúan después de adquirir una imagen, específicamente, buscan realizar degradaciones sobre esta para proteger la identidad de las perso-

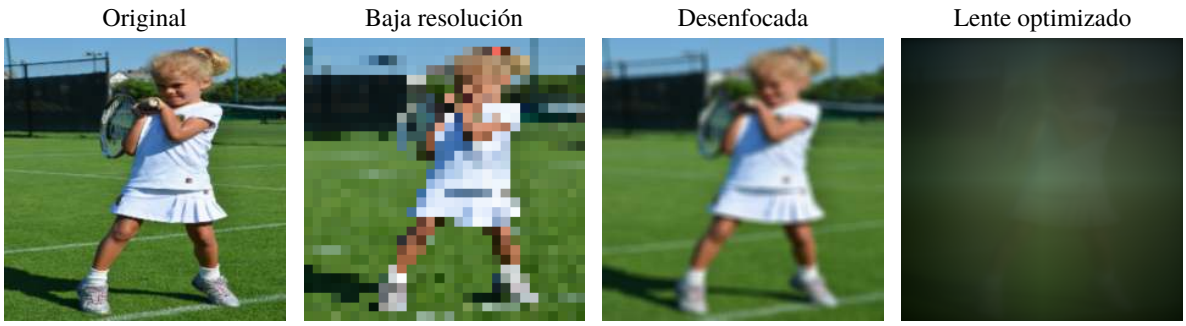
---

[33] Anthony Cioppa et al. «SoccerNet 2023 challenges results». En: *Sports Engineering* 27.2 (2024), pág. 24.

[34] GSDMA Sreenu y Saleem Durai. «Intelligent video surveillance: a review through deep learning techniques for crowd analysis». En: *Journal of Big Data* 6.1 (2019), págs. 1-27.

[35] José Ramón Padilla-López, Alexandros Andre Chaaraoui y Francisco Flórez-Revuelta. «Visual privacy protection methods: A survey». En: *Expert Systems with Applications* 42.9 (2015), págs. 4177-4195.

Figura 1. Comparación entre diferentes enfoques de privacidad visual, (a) imagen original, (b) imagen de baja resolución, (c) imagen desenfocada y (d) imagen obtenida con lente optimizado[26].



nas[36]. Este tipo de procesamiento consiste en distorsiones visuales[37], encriptación[38] o de eliminación, donde se borra la información sensible y se repinta la imagen o el video[39]. Sin embargo, los sistemas de protección basados en software sólo preservan la privacidad en la última etapa de la aplicación, es decir, las imágenes que se adquieren directamente en el dispositivo no cuentan con ninguna forma de protección.

1.1.2. Métodos basados en *Hardware*. En contraste, los métodos basados en hardware se caracterizan por agregar una capa de seguridad previa a la adquisición, modificando el sistema óptico con el objetivo de eliminar datos sensibles. En este

---

[36] Jong Wook Kim, Beakcheol Jang y Hoon Yoo. «Privacy-preserving aggregation of personal health data streams». En: *PloS one* 13.11 (2018), e0207639.

[37] Andrea Frome et al. «Large-scale privacy protection in Google Street View». En: *2009 IEEE 12th International Conference on Computer Vision*. 2009, págs. 2373-2380.

[38] Wenjun Zeng y S. Lei. «Efficient frequency domain selective scrambling of digital video». En: *IEEE Transactions on Multimedia* 5.1 (2003), págs. 118-129.

[39] Jasper Tan et al. «CANOPIC: pre-digital privacy-enhancing encodings for computer vision». En: *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2020, págs. 1-6.

grupo se pueden destacar métodos que usan sensores con restricciones físicas como lo son cámaras en el rango infrarrojo[40], cámaras de muy baja resolución[24], o lentes convencionales fuera de foco[41]. Aunque estos métodos aprovechan sistemas ópticos existentes, no optimizan el sistema en sí, por lo que no logran mantener un balance adecuado entre el desempeño en tareas de visión por computadora y la preservación de la privacidad visual. Adicionalmente, existen numerosos trabajos que buscan abordar las restricciones físicas inherentes a estos enfoques, intentando resolver los problemas inversos asociados con los sistemas convencionales, como lo son eliminación de ruido[42,43], superresolución[44,45], corrección de desenfoque[46,47]. Para evitar que sea posible reconstruir las imágenes adquiridas y promover un balance entre la tarea de visión por computadora y la preservación de la privacidad visual, investigaciones recientes han propuesto optimizar los elementos ópticos del

---

[40] Ralph Gross et al. «Integrating utility into face de-identification». En: *International Workshop on Privacy Enhancing Technologies*. Springer. 2005, págs. 227-242.

[41] Francesco Pittaluga y Sanjeev Jagannatha Koppal. «Pre-Capture Privacy for Small Vision Sensors». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.11 (2017), págs. 2215-2226.

[42] Zunlin Fan et al. «Dim infrared image enhancement based on convolutional neural network». En: *Neurocomputing* 272 (2018), págs. 396-404.

[43] Yu Binbin. «An improved infrared image processing method based on adaptive threshold denoising». En: *EURASIP Journal on Image and Video Processing* 2019.1 (2019), pág. 5.

[44] Nhat Nguyen, Peyman Milanfar y Gene Golub. «A computationally efficient superresolution image reconstruction algorithm». En: *IEEE Transactions on Image Processing* 10.4 (2001), págs. 573-583.

[45] Saeed Anwar, Salman Khan y Nick Barnes. «A Deep Journey into Super-resolution: A Survey». En: *ACM Comput. Surv.* 53.3 (2020).

[46] Kaihao Zhang et al. «Deep image deblurring: A survey». En: *International Journal of Computer Vision* 130.9 (2022), págs. 2103-2130.

[47] Lu Yuan et al. «Image deblurring with blurred/noisy image pairs». En: *ACM SIGGRAPH 2007 Papers*. SIGGRAPH '07. Association for Computing Machinery, 2007, 1–es.

sistema usando aprendizaje profundo[25,26,27]. Este método será detallado a profundidad en el Capítulo 1.4 de este trabajo de investigación.

1.2. OPTIMIZACIÓN DE SISTEMAS ÓPTICOS DIFRACTIVOS

1.2.1. Elementos ópticos difractivos. En la óptica refractiva, el trayecto que sigue la luz en un medio es una línea recta. Para calcular la nueva dirección de un rayo al ser refractado por un material se usa la Ley de Snell, donde la nueva dirección del rayo está definida por la geometría de la superficie y propiedades físicas del medio, por ejemplo, el índice de refracción[48]. La Ley de Snell explica el comportamiento de la luz al incidir sobre los lentes convencionales o prismas.

En el caso de la óptica difractiva, la propagación del frente de onda de la luz se explica mediante el principio de Huygens[48,49], el cual establece que cada punto del frente de onda que incide sobre una superficie actúa como una fuente secundaria que contribuye al frente de onda propagado total. Comúnmente, en la literatura, los elementos ópticos difractivos (EOD) se clasifican según si presentan una superficie discontinua o si la estructura de su superficie o apertura tienen dimensiones comparables a la longitud de onda de la luz utilizada[50].

Para ilustrar los EODs, en la Figura 2 se muestra una comparación de un lente refractivo tradicional 2(a) contra un lente difractivo comúnmente conocido como lente de Fresnel 2(b). La superficie del lente refractivo es continua y suave, por el contrario, el lente difractivo puede presentar múltiples discontinuidades. Sin embargo, para

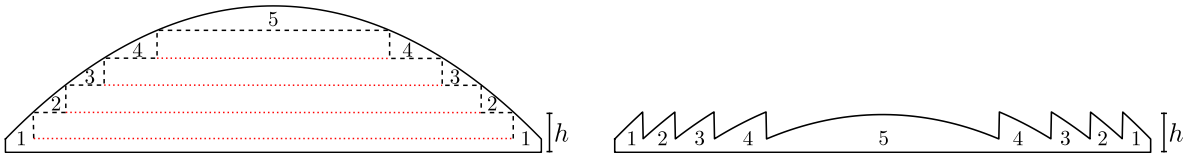
---

[48] Max Born y Emil Wolf. *Principles of optics: Electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.

[49] Joseph W Goodman. «Introduction to Fourier Optics, Roberts & Co». En: *Publishers, Englewood, Colorado* (2005).

[50] Anand Vijayakumar y Shanti Bhattacharya. *Design and fabrication of diffractive optical elements with MATLAB*. SPIE, 2017.

Figura 2. Comparación del mapa de alturas de (a) un lente refractivo y (b) un lente difractivo.



Lente refractivo (a) Lente de Fresnel (b)

este ejemplo, el efecto de (a) o (b) sobre un frente de onda incidente es el mismo[51], es decir

$$e^{ik_0\Delta n(\lambda)\phi(x,y)} = e^{ik_0\Delta n(\lambda)\hat{\phi}(x,y)}, \tag{1}$$

donde $k_0 = \frac{2\pi}{\lambda}$ es el número de onda, $\Delta n(\lambda)$ es el cambio de índice de refracción entre medios para una longitud de onda λ, $\phi(x, y)$ es la altura del lente refractivo y $\hat{\phi}(x, y)$ la altura del lente difractivo. Esto se produce debido a la periodicidad de 2π de la exponencial compleja $(e^{i\beta+ik2\pi} = e^{i\beta}, \quad \forall\beta \in \mathbb{R})$, por lo que es posible construir un elemento difractivo $\hat{\phi}(x, y)$[51] siguiendo

$$\hat{\phi}(x, y) = mod(\phi(x, y), h), \tag{2}$$

donde $h = \frac{\lambda}{\Delta n(\lambda)}$ corresponde a la altura que genera un aporte de 2π en la fase, de tal manera que, se cumple (1). Bajo este paradigma, es posible obtener elementos ópticos de menor peso y grosor que generen el mismo efecto en el frente de onda saliente. Los EODs, cuando se diseñan correctamente, no solo permiten reproducir el efecto de superficies refractivas, sino que también pueden inducir distorsiones complejas y diversas en el frente de onda. Esto ha sido explorado en campos como

---

[51] Warren J Smith. *Modern optical engineering: the design of optical systems*. SPIE Press, 2008.

imágenes espectrales[52,53], estimación de la profundidad[54] y aumento del rango de color[55].

1.2.2. Modelo de formación de imagen. En general, el campo óptico sobre el sensor $f_s(x, y; \lambda)$, está dado por

$$f_s(x, y; \lambda) = \iint h_\Phi(x - u, y - v; \lambda) f_0(u, v; \lambda) du dv, \tag{3}$$

donde $f_0(u, v; \lambda)$ es el campo incoherente que incide sobre el sistema óptico. La función h_Φ, conocida como la función de respuesta al impulso del sistema en su versión continua (PSF, por sus siglas en inglés *point spread function*), depende de la posición de los elementos ópticos, así como de la superficie del EOD, determinada por por el conjunto de parámetros Φ. La Figura 3 muestra dos configuraciones ópticas ampliamente usadas, en las que se incluye un EOD con el objetivo de distorsionar el campo óptico. En este trabajo de investigación exploramos ambas configuraciones para el contexto de preservación de la privacidad visual.

- **Configuración #1:** Ilustrada en la Figura 3(a), ha sido ampliamente utilizada debido a su capacidad para modular de manera versátil el frente de onda inci-

---

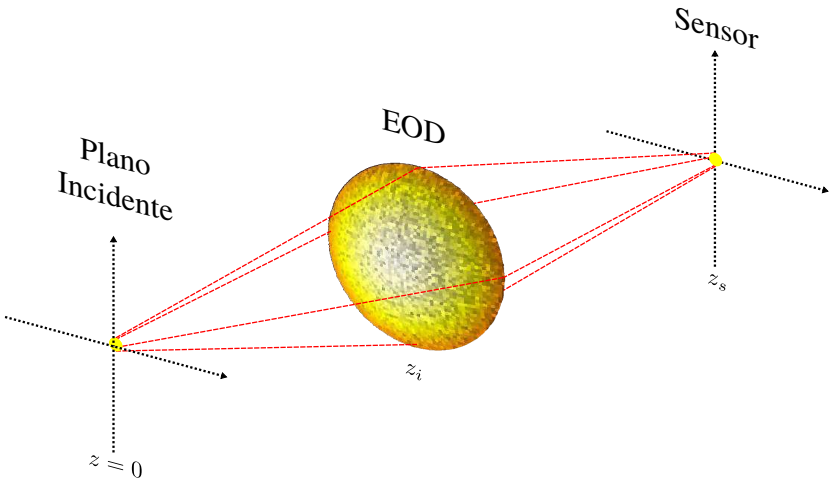
[52] Henry Arguello et al. «Shift-variant color-coded diffractive spectral imaging system». En: *Optica* 8.11 (2021), págs. 1424-1434.

[53] Daniel S. Jeon et al. «Compact Snapshot Hyperspectral Imaging with Diffracted Rotation». En: *ACM Trans. Graph.* 38.4 (2019).

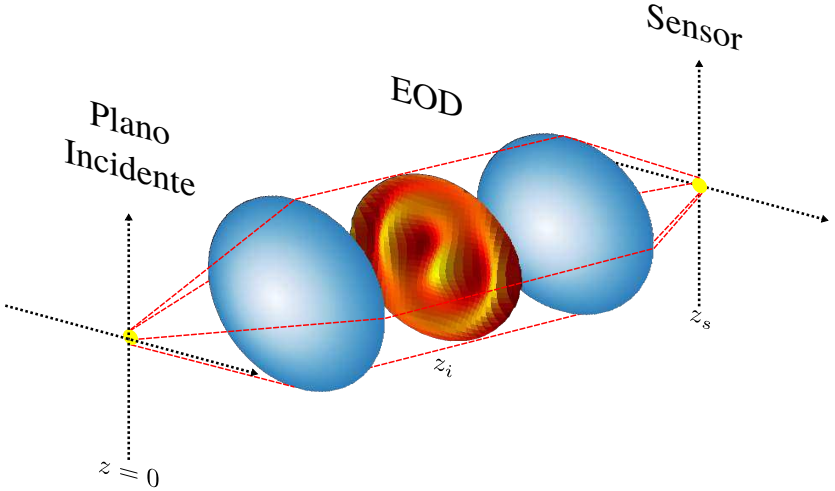
[54] Seung-Hwan Baek et al. «Single-shot hyperspectral-depth imaging with learned diffractive optics». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 2651-2660.

[55] Xiong Dun et al. «Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging». En: *Optica* 7.8 (2020), págs. 913-922.

Figura 3. Esquema de configuraciones ópticas que incluyen EODs.



(a) Configuración #1. Un EOD modula el frente de onda, logrando un sistema óptico compacto que distorsiona el campo óptico para preservar la privacidad visual.



(b) Configuración #2. Un arreglo $4f$, junto a un EOD en el plano pupila permite modular la transformada de Fourier del campo incidente mediante un EOD.

dente, permitiendo la implementación de un sistema óptico compacto[53,54,56,57]. La PSF de esta configuración está determinada por

$$h_\Phi(x, y; \lambda) = \left| \mathcal{F}^{-1}\{T_{z_s}(k_x, k_y; \lambda) \cdot \mathcal{F}\{U_{z_i}(x', y'; \lambda) \cdot \Psi_\Phi(x', y'; \lambda)\} \right|^2 \quad (4)$$

donde $U_{z_i}(x, y; \lambda)$ es el frente de onda esférico incidente desde una fuente puntual a una distancia z_i del EOD, $T_{z_s}(k_x, k_y; \lambda) = \exp\left(-iz_s\sqrt{(\frac{2\pi}{\lambda})^2 - k_x^2 - k_y^2}\right)$ es la función de transferencia del campo óptico propagado una distancia z_s hasta el sensor, Ψ_Φ corresponde al efecto de modulación de fase del elemento difractivo como $\Psi_\Phi = \exp(ik_0\Delta n(\lambda)\phi_\Phi(x', y'))$, donde $A(x', y')$ es la función de apertura del elemento óptico, $\phi_\Phi(x', y')$ representa la superficie dada por unos parámetros Φ, $\mathcal{F}$ y $\mathcal{F}^{-1}$ son los operadores de transformada de Fourier directa e inversa y $|\cdot|^2$ representa la intensidad de la luz captada por el sensor,.

- **Configuración #2:** Ilustrada en la Figura 3(b), es ampliamente utilizada por su capacidad de modular la transformada de Fourier del campo incidente, permitiendo modular las frecuencias espaciales de manera independiente para así

---

[56] 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)* 37.4 (2018), págs. 1-13.

[57] Xiong Dun et al. «Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging». En: *Optica* 7.8 (2020), págs. 913-922.

obtener PSFs diversas[58,59,60]. Su PSF se calcula como

$$h_\Phi(u, v; \lambda) = \left| \mathcal{F}\{\Psi_\Phi(k_x, k_y) \cdot \mathcal{F}\{U_{z_i}(x, y; \lambda)\}\right|^2, \tag{5}$$

donde $U_{z_i}(x, y; \lambda)$ es el frente de onda incidente a una distancia z_i del sistema óptico, Ψ_Φ corresponde al efecto de modulación de fase del elemento difractivo y el arreglo de lentes $4f$ dado por

$$\Psi_\Phi(k_x, k_y) = A(k_x, k_y) \exp\left(j\phi(k_x, k_y) + j\phi_{4f}(k_x, k_y)\right),$$

donde $A(x', y')$ es la función de apertura del EOD con mapa de alturas, $\phi_\Phi(x', y')$ determinado por los parámetros Φ, $\mathcal{F}$ es el operador de transformada de Fourier directa y $|\cdot|^2$ representa la intensidad de la luz captada por el sensor.

La versión discreta de h_Φ, denotada como $\mathcal{H}_\Phi \in \mathbb{R}^{N_y \times N_x \times N_\lambda}$ está dada por

$$(\mathcal{H}_\Phi)_{i,j,k} = \iiint h_\Phi(x, y; \lambda) \cdot \mathrm{rect}\Big(\frac{x}{\Delta_p} - j, \frac{y}{\Delta_p} - i\Big) \cdot \delta(\lambda - \lambda_k) dx dy d\lambda, \tag{6}$$

con $i \in \{1, \ldots, N_y\}$, $j \in \{1, \ldots, N_x\}$ indexando las filas y columnas del sensor, $k \in \{1, \ldots, N_\lambda\}$ indexando cada longitud de onda capturada y Δ_p es el tamaño del píxel espacial. Note que, que la PSF está compuesta por un filtro de convolución $\mathbf{H}_{\Phi k} = \mathcal{H}_{\Phi k} \in \mathbb{R}^{N_y \times N_x}$ para cada canal de color correspondiente a la longitud de

---

[58] Elias Nehme et al. «DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning». En: *Nature Methods* 17.7 (2020), págs. 734-740.

[59] Yoav Shechtman et al. «Precise Three-Dimensional Scan-Free Multiple-Particle Tracking over Large Axial Ranges with Tetrapod Point Spread Functions». En: *Nano Letters* 15.6 (2015). PMID: 25939423, págs. 4194-4199. DOI: 10.1021/acs.nanolett.5b01396. eprint: https://doi.org/10.1021/acs.nanolett.5b01396.

[60] Adam Greengard, Yoav Y. Schechner y Rafael Piestun. «Depth from diffracted rotation». En: *Opt. Lett.* 31.2 (2006), págs. 181-183.

onda λ_k, donde, cada filtro se puede representar como la matriz de convolución 2D, $\mathbf{C}_{\Phi k} \in \mathbb{R}^{N_y N_x \times N_y N_x}$. Para este trabajo de investigación asumimos un sensor RGB, por lo que $N_\lambda = 3$. El modelo discreto de la Ecuación (3) se puede representar matricialmente como,

$$\mathbf{y} = \mathbf{C}_\Phi \mathbf{x}, \tag{7}$$

donde $\mathbf{x} \in \mathbb{R}^{N_y N_x N_\lambda}$ es la escena, $\mathbf{y} \in \mathbb{R}^{N_y N_x N_\lambda}$ es la medida adquirida y $\mathbf{C}_\Phi \in \mathbb{R}^{N_y N_x N_\lambda \times N_y N_x N_\lambda}$ es la matriz de convolución, construida a partir de cada matriz de convolución $\mathbf{C}_{\Phi k}$, por bloques diagonales como

$$\mathbf{C}_\Phi = \begin{bmatrix} \mathbf{C}_{\Phi 1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\Phi 2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_{\Phi 3} \end{bmatrix}. \tag{8}$$

Es importante resaltar que para preservar la privacidad visual, se requiere generar $\mathbf{C}_\Phi$ de tal manera que $\mathbf{x} \neq \mathbf{C}_\Phi^{-1}\mathbf{y}$. Para lograr esto, hay que diseñar correctamente el sistema óptico y el conjunto de parámetros Φ que provocarán este comportamiento sobre h_Φ.

1.2.3. Optimización de sistemas ópticos usando aprendizaje profundo. En la literatura, diversos estudios han explorado el diseño del conjunto de parámetros Φ para la adquisición de imágenes. Estos métodos se pueden clasificar en dos categorías principales: por un lado, aquellos basados en propiedades matemáticas de $\mathbf{C}_\Phi$, tales como la independencia lineal[61], la coherencia mutua[62] o la concentración

---

[61] Yuri Mejia y Henry Arguello. «Binary Codification Design for Compressive Imaging by Uniform Sensing». En: *IEEE Transactions on Image Processing* 27.12 (2018), págs. 5775-5786. DOI: 10.1109/TIP.2018.2857445.

[62] Michael Elad. «Optimized Projections for Compressed Sensing». En: *IEEE Transactions on Signal Processing* 55.12 (2007), págs. 5695-5702. DOI: 10.1109/TSP.2007.900760.

de la medida[63]; por otro lado, los métodos basados en datos, que aprovechan el aprendizaje profundo[31] como una estrategia para optimizar Φ considerando la tarea específica para la cual será utilizado el sistema óptico.

Estos enfoques basados en datos permiten, por ejemplo, extender el campo de visión (*field of view*) de las cámaras[56], mejorar la calidad de la reconstrucción espectral[52,53], o aumentar la precisión en la estimación de profundidad[64]. En este contexto, el presente trabajo de investigación se enfoca en aprovechar el diseño basado en datos para promover la privacidad visual en las adquisiciones. El problema de optimización que permite diseñar los sistemas ópticos usando aprendizaje profundo puede escribirse matemáticamente como

$$\arg\min_{\Theta,\Phi} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{d}\sim\mathcal{D}}[\mathcal{L}_{tarea}(\mathbf{d}, \mathcal{V}_\Theta(\mathbf{C}_\Phi\mathbf{x})) + \mathcal{R}(\Phi)], \tag{9}$$

donde se busca encontrar los parámetros Φ y Θ, de tal manera que, se minimice una función de regularización $\mathcal{R}(\cdot)$ encargada de promover restricciones físicas de diseño sobre los parámetros Φ[31], a la vez que una red neuronal $\mathcal{V}_\Theta\colon \mathbf{C}_\Phi\mathbf{x} \mapsto \hat{\mathbf{d}}$ disminuya la función de costo $\mathcal{L}_{tarea}(\cdot,\cdot)$ en valor esperado $\mathbb{E}$, sobre un conjunto de datos $\mathcal{D} = \{\mathbf{x}^{(j)}, \mathbf{d}^{(j)}\}_{j=1}^{N_d}$, de una tarea deseada como lo puede ser clasificación, reconstrucción de imágenes, estimación de profundidad, entre otras. Para resolver (9) se usan métodos iterativos basados en la minimización del gradiente[65] implementados

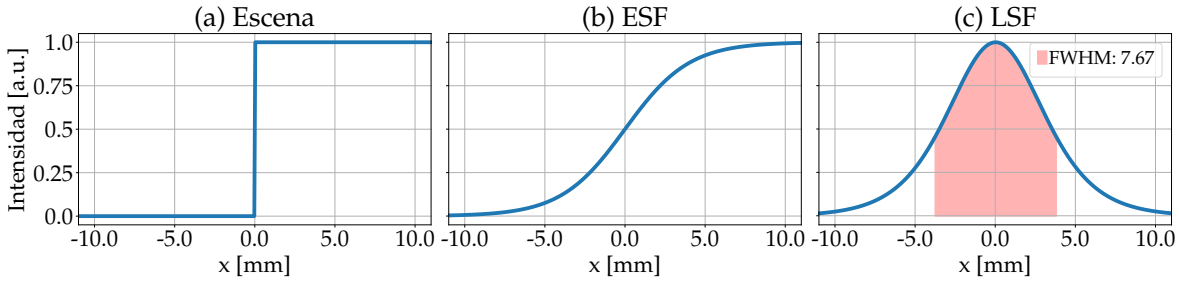
---

[63] Gonzalo R. Arce et al. «Compressive Coded Aperture Spectral Imaging: An Introduction». En: *IEEE Signal Processing Magazine* 31.1 (2014), págs. 105-115. DOI: 10.1109/MSP.2013.2278763.

[64] Hayato Ikoma et al. «Depth from defocus with learned optics for imaging and occlusion-aware depth estimation». En: *2021 IEEE International Conference on Computational Photography (ICCP)*. IEEE. 2021, págs. 1-12.

[65] Diederik P Kingma. «Adam: A method for stochastic optimization». En: *arXiv preprint arXiv:1412.6980* (2014).

Figura 4. Ilustración del método para calcular la resolución de un sistema óptico. Primero, una escena con un cambio espacial fuerte (a) es adquirida por el sistema óptico, generando la ESF (b), luego se calcula la LSF (c) derivando la ESF y se calcula el FWHM de esta curva.



usando herramientas de diferenciación automática como *PyTorch*[66].

1.2.4. Métricas de calidad. Tradicionalmente en el campo de la óptica, se han usado diversas métricas para cuantificar la capacidad de un sistema para adquirir información. A continuación se resumen las más relevantes:

- **Resolución:** Según el criterio de Rayleigh, la resolución espacial de un sistema óptico se define como la distancia mínima en la que se puede observar una separación de al menos el 16 % entre los picos de intensidad de dos fuentes puntuales de luz[48]. Aunque esta metodología es ampliamente conocida en la literatura óptica[51], presenta dificultades prácticas para su medición experimental.

  Por otra parte, tanto la ciencia como la industria fotográfica han adoptado es-

66 Adam Paszke et al. «PyTorch: An Imperative Style, High-Performance Deep Learning Library». En: *Advances in Neural Information Processing Systems*. Ed. por H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

tándares[67,68,69], ilustrados en la Figura 4, los cuales usan una escena con un cambio espacial fuerte, que es capturada usando el sistema óptico a analizar. Esta captura se denomina ESF (por sus siglas en inglés *edge spread function*), a partir de la cual se deriva la LSF (por sus siglas en inglés *line spread function*) y donde se determina la resolución mediante el FWHM (por sus siglas en inglés *full width at half maximum*) de esta curva.

Dado que la privacidad en este trabajo se fundamenta en la degradación de la calidad visual para prevenir la identificación de información sensible, el FWHM se adopta como métrica para cuantificar el grado de preservación de la privacidad. Valores elevados de FWHM indican que la LSF se ha expandido, resultando en una imagen más difusa y menos interpretable, lo que refuerza la protección de la identidad y el contenido visual.

- **Función de Transferencia de Modulación:** La función de transferencia de modulación (MTF por sus siglas en inglés *modulation transfer function*) cuantifica la capacidad del sistema óptico para determinar la información presente en la captura para cada frecuencia espacial de manera independiente. Usualmente se define la MTF como la amplitud de la función de transferencia óptica (OTF por sus siglas en inglés *optical transfer function*),

$$\mathrm{MTF}(\mathbf{H}_{\Phi k}) = |\mathrm{OTF}(\mathbf{H}_{\Phi k})| = |\mathcal{F}\{\mathbf{H}_{\Phi k}\}|. \tag{10}$$

---

[67] Ken Parulski et al. «Creation and evolution of ISO 12233, the international standard for measuring digital camera resolution». En: *Electronic Imaging* 34 (2022), págs. 1-7.

[68] Jingdan Liu et al. «Swept coded aperture real-time femtophotography». En: *Nature Communications* 15.1 (2024), pág. 1589.

[69] Hyundeok Hwang et al. «MTF assessment of high resolution satellite images using ISO 12233 slanted-edge method». En: *Image and Signal Processing for Remote Sensing XIV*. Vol. 7109. SPIE. 2008, págs. 34-42.

Cuando la MTF presenta valores cercanos a cero, indica que el sistema óptico atenúa significativamente ciertas frecuencias de la señal adquirida, indicando que la PSF impide la reconstrucción de la señal utilizando técnicas como el filtro *Wiener*[70].

1.2.5. Número de condición como métrica de invertibilidad. En numerosas aplicaciones de imágenes computacionales, el proceso de adquisición se modela usando la Ecuación (7), donde comúnmente $\mathbf{C}_\Phi$ es una matriz mal condicionada, es decir, no existe una matriz inversa[71], tal que $\mathbf{x} = \mathbf{C}_\Phi^{-1}\mathbf{y}$. Por lo tanto, para estimar $\mathbf{x}$ a partir de $\mathbf{y}$, se emplean algoritmos basados en optimización numérica para resolver el problema inverso asociado[72], como por ejemplo,

$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{C}_\Phi \mathbf{x}\|_2^2 + \mathcal{R}(\mathbf{x}), \tag{11}$$

donde $\|\cdot\|_2^2$ es la norma ℓ^2, que mide la consistencia de la estimación a las medidas, y $\mathcal{R}(\mathbf{x})$ es una función de regularización sobre $\mathbf{x}$ para imponer propiedades como suavidad o escasez, entre otras. Resolver (11) implica realizar numerosas operaciones involucrando $\mathbf{C}_\Phi$ y $\mathbf{C}_\Phi^T\mathbf{C}_\Phi$, por lo que se requiere que $\mathbf{C}_\Phi$ tenga propiedades que garanticen estabilidad numérica[73]. Particularmente, el número de condición es

---

[70] Amit Agrawal y Yi Xu. «Coded exposure deblurring: Optimized codes for PSF estimation and invertibility». En: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, págs. 2066-2073.

[71] K. Forbes y V. V. Anh. «Condition of system matrices in image restoration». En: *J. Opt. Soc. Am. A* 11.6 (1994), págs. 1727-1735.

[72] Henry Arguello y Miguel Marquez. «Convex Optimization for Image Reconstruction». En: *Coded Optical Imaging*. Springer, 2024, págs. 37-53.

[73] BISWA NATH DATTA. «CHAPTER 3 - SOME FUNDAMENTAL TOOLS AND CONCEPTS FROM NUMERICAL LINEAR ALGEBRA». En: *Numerical Methods for Linear Control Systems*. Ed. por BISWA NATH DATTA. San Diego: Academic Press, 2004, págs. 33-78.

una medida matemática que cuantifica la estabilidad numérica de una matriz[73], por ello, es una métrica comúnmente usada para cuantificar la facilidad de encontrar un mínimo en (11)[71,72,74,75]. El número de condición se define como

$$\mathcal{K}(\mathbf{C}_\Phi) = \frac{\sigma_{max}}{\sigma_{min}} \geq 1, \tag{12}$$

donde σ_{max} y σ_{min} son los valores singulares máximos y mínimos de $\mathbf{C}_\Phi$. Cuando $\mathcal{K}(\mathbf{C}_\Phi) = 1$, la inversa ${\mathbf{C}_\Phi}^{-1}$ existe y resolver (11) es trivial, mientras que para un valor alto del número de condición, $\mathcal{K}(\mathbf{C}_\Phi) \to \infty$, se considera que no se puede resolver (11).

Definición 1. *(Sistema óptico que promueve la privacidad en imágenes). En el contexto de este trabajo de investigación, un sistema óptico se considera que promueve la privacidad si obtiene un valores altos de FWHM y número de condición, mientras que obtiene valores cercanos a cero para la norma de la MTF.*

1.3. RECONOCIMIENTO DE ACCIONES EN ENTORNOS CLÍNICOS

La estimación de acciones humanas es un área de amplio estudio en la literatura. Esta tarea consiste en reconocer las acciones realizadas por un usuario mediante el análisis de aspectos, tales como, la movilidad, la pose y los gestos[76]. Para rea-

---

[74] J Carlos Santamarina y Dante Fratta. *Discrete signals and inverse problems: an introduction for engineers and scientists*. John Wiley & Sons, 2005.

[75] Yanan Zhao et al. «Deep, Convergent, Unrolled Half-Quadratic Splitting for Image Deconvolution». En: *IEEE Transactions on Computational Imaging* 10 (2024), págs. 574-588. DOI: 10.1109/TCI.2024.3377132.

[76] Hong-Bo Zhang et al. «A comprehensive survey of vision-based human action recognition methods». En: *Sensors* 19.5 (2019), pág. 1005.

lizar la estimación de acciones se suelen usar imágenes adquiridas con diferentes tipos de sistemas ópticos. Particularmente, se han usado imágenes RGB adquiridas utilizando cámaras tradicionales[76] o sistemas más complejos de imágenes RGB y profundidad basados en el sistema Kinect[77]. El reconocimiento de acciones tiene múltiples aplicaciones en el área de la seguridad privada y vigilancia. Recientemente, el reconocimiento de acciones ha sido utilizado en escenarios clínicos, donde el objetivo principal es crear sistemas que permitan asistir en tareas de cuidado[4,5], prevención[6] y diagnóstico[7,8] de enfermedades para así salvaguardar la salud de los pacientes[78,79]. Tradicionalmente, los trabajos de reconocimiento de acciones basados en técnicas de aprendizaje profundo se componen de modelos unimodales[11], es decir que se entrenan únicamente con un tipo de dato, como por ejemplo, los videos. Dentro de estos métodos tradicionales se busca el diseño de arquitecturas que exploten la información temporal de los videos usando capas convolucionales 3D[12], aprendiendo permutaciones y desplazamientos de los canales temporales[11,13], o extrayendo características globales aplicando capas recurrentes[14] o la arquitectura *Transformer*[15].

1.3.1. Entrenamiento multimodal para la estimación de acciones. Recientemente, con el auge de los grandes modelos del lenguaje y modelos fundacionales[16], potenciados en el campo del procesamiento de lenguaje natural, ha crecido el interés por generar modelos multimodales capaces de extraer información de diversas

---

77 Ali Seydi Keceli y Ahmet Burak Can. «Recognition of basic human actions using depth information». En: *International Journal of Pattern Recognition and Artificial Intelligence* 28.02 (2014).

78 ASRM Ahouandjinou, C Motamed y EC Ezin. «A temporal belief-based hidden markov model for human action recognition in medical videos». En: *Pattern Recognition and Image Analysis* 25.3 (2015), págs. 389-401.

79 Edward Chou et al. «Privacy-preserving action recognition for smart hospitals using low-resolution depth images». En: *arXiv preprint arXiv:1811.09950* (2018).

fuentes de datos incluyendo, imágenes, texto, audio, entre otros, y mejorar así el reconocimiento de acciones; por lo tanto, han surgido nuevos sistemas de reconocimiento de acciones basados en el paradigma multimodal[80]. Para el contexto de este trabajo exploramos la estimación de acciones multimodal propuesta en[81,82].

La Figura 5 muestra la comparación de la estimación de acciones entre el paradigma unimodal y el multimodal, donde para ambos casos, una red neuronal $\mathcal{V}_\Theta$, con pesos entrenables Θ, está encargada de extraer las características visuales $\mathbf{f}^{\text{visual}} = \mathcal{V}_\Theta(\mathbf{x}) \in \mathbb{R}^{N_c}$. En el caso unimodal, $\mathcal{V}_\Theta$ se entrena para predecir las etiquetas $\mathbf{d} \in \mathbb{R}^{N_c}$ usando la codificación tipo *one-hot*, en donde a cada clase $c \in \{1, \ldots, N_c\}$ se le asigna una etiqueta $\mathbf{d} = \vec{\mathbf{e}}_c = \delta_{i,c}$ correspondiente a un vector de la base canónica de $\mathbb{R}^{N_c} = \vec{\mathbf{e}}_1 \times \vec{\mathbf{e}}_2 \times \cdots \times \vec{\mathbf{e}}_{N_c}$. Para encontrar Θ se resuelve el siguiente problema de optimización

$$\arg\min_{\Theta} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{d}\sim\mathcal{D}} [\mathcal{L}_{tarea}(\mathbf{d}, \mathcal{V}_\Theta(\mathbf{x}))]. \tag{13}$$

Por el contrario, en el caso multimodal, se resuelve el siguiente problema de optimización,

$$\arg\min_{\Theta} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{d},\mathbf{F}^{\text{texto}}\sim\mathcal{D}} [\mathcal{L}_{tarea}(\mathbf{d}, S(\mathbf{F}^{\text{texto}}, \mathcal{V}_\Theta(\mathbf{x})))], \tag{14}$$

donde $\mathcal{L}_{tarea}$ corresponde a la entropía cruzada, $\mathbf{F}^{\text{texto}} \in \mathbb{R}^{N_c \times N_f}$ es una matriz, con las filas siendo características textuales, $\mathbf{F}^{\text{texto}}_{c,:} = \mathcal{T}_\Xi(\mathbf{t}_c)^T$, extraídas usando una red neuronal de procesamiento de lenguaje natural $\mathcal{T}_\Xi(\cdot)$, con parámetros fijos Ξ, donde $\mathbf{t}_c$ es una cadena de texto que describe cada acción, por ejemplo $\mathbf{t}_c =$ "Una

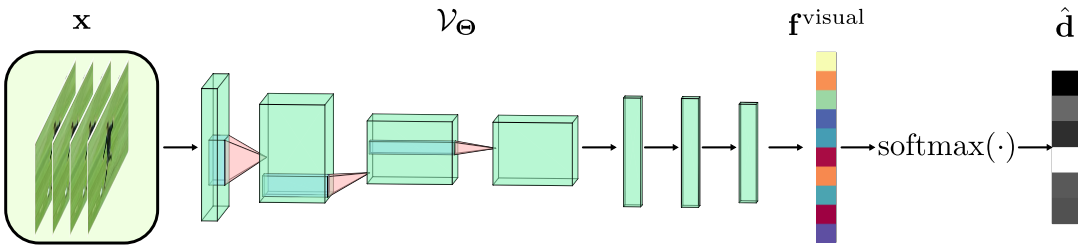
---

[80] Jiquan Ngiam et al. «Multimodal deep learning». En: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, págs. 689-696.

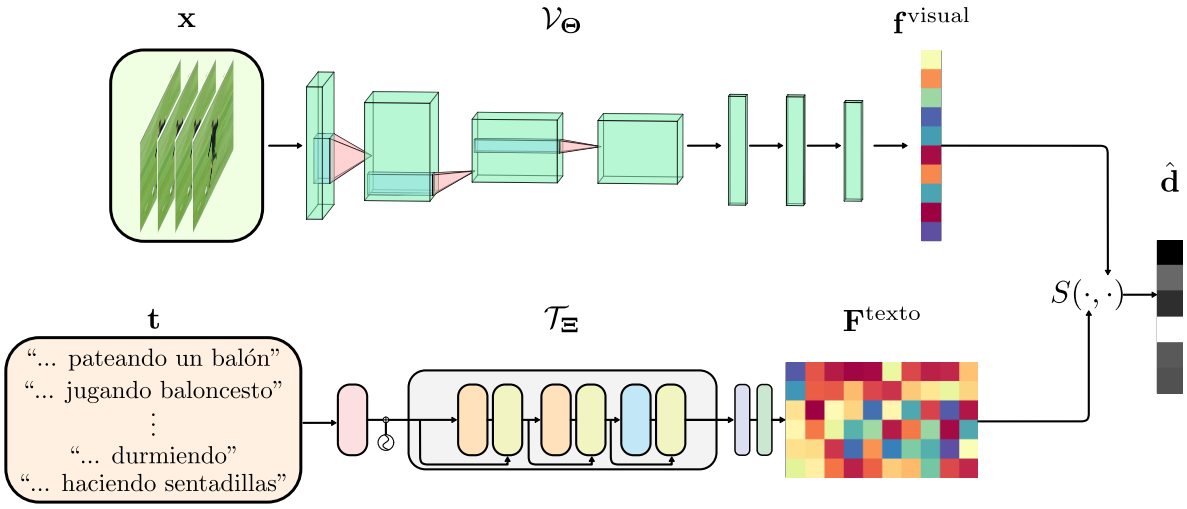
[81] Mengmeng Wang et al. «Actionclip: Adapting language-image pretrained models for video action recognition». En: *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[82] Wenhao Wu, Zhun Sun y Wanli Ouyang. «Revisiting Classifier: Transferring Vision-Language Models for Video Recognition». En: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.3 (2023), págs. 2847-2855.

Figura 5. Comparación de los paradigmas para la estimación de acciones. En el paradigma (a) unimodal, una red extrae características visuales de la escena y asigna una clase utilizando una función de activación $\mathrm{softmax}(\cdot)$. En el paradigma (b) multimodal, se extraen características visuales que se comparan con características textuales de posibles acciones mediante una función de similaridad $S(\cdot,\cdot)$, la clase asignada corresponde a la clase donde existe la máxima similaridad entre las características visuales y textuales.

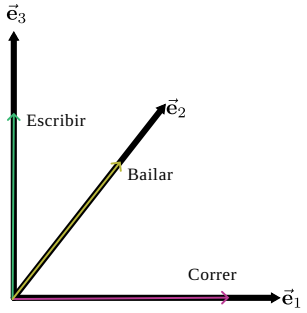


(a) Paradigma unimodal

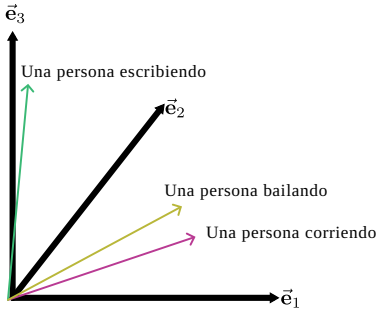


(b) Paradigma multimodal

Figura 6. Ilustración de las características visuales extraídas a partir de los paradigmas de estimación de acciones, en el caso (a) unimodal, la red neuronal se entrena para extraer características, de tal manera que sean vectores de la base canónica. En el caso (b) multimodal, la red neuronal se entrena para extraer características que sean similares a características extraídas a partir de texto, de esta manera se conservan propiedades semánticas de las acciones.



(a) características paradigma unimodal



(b) características paradigma multimodal

persona corriendo". $S(\cdot,\cdot)$ corresponde a una función que calcula la similaridad $S(\mathbf{F}^{\text{texto}}, \mathbf{f}^{\text{visual}}) = \text{softmax}\left(\frac{\mathbf{F}^{\text{texto}}\mathbf{f}^{\text{visual}}}{\|\mathbf{F}^{\text{texto}}\|\cdot\|\mathbf{f}^{\text{visual}}\|}\right) \in \mathbb{R}^{N_c}$, entre las características textuales $\mathbf{F}^{\text{texto}}$ y las características visuales $\mathbf{f}^{\text{visual}} = \mathcal{V}_\Theta(\mathbf{x}) \in \mathbb{R}^{N_f}$.

Con este paradigma es posible entrenar Θ de tal manera que se aprovecha la información semántica de las acciones. Como lo ilustra la Figura 6, las características extraídas de acciones similares como "correr" y "bailar" están más cerca en el espacio de características que otras clases no relacionadas.

1.3.2. Aprendizaje sin ejemplos. En la investigación sobre inteligencia artificial, uno de los mayores retos es desarrollar modelos capaces de realizar tareas como la clasificación con una menor cantidad de datos de entrenamiento. En particular, el área conocida como aprendizaje sin ejemplos (*zero-shot learning*) busca que los modelos puedan reconocer clases (en este caso, acciones) que nunca han sido

explícitamente vistas durante el proceso de entrenamiento[83,84]. Este enfoque es especialmente relevante en aplicaciones con conjuntos de datos limitados o en los que existe un desequilibrio significativo entre clases, como ocurre en el ámbito clínico, donde obtener grandes volúmenes de datos es costoso y requiere mucho tiempo[85]. En contraste a los enfoques unimodales tradicionales, los modelos multimodales para la estimación de acciones permiten una generalización más amplia, lo que posibilita la estimación de acciones aún cuando no existen ejemplos previos de pares de datos etiquetados (como video y acción). Esto es posible gracias al entrenamiento del modelo para relacionar el espacio de características visuales con el espacio de características textuales. De esta manera, durante la fase de prueba, se pueden introducir nuevas descripciones textuales que correspondan a acciones no vistas previamente, y si la similitud entre las características visuales y textuales es lo suficientemente alta, el modelo puede clasificar correctamente el video dentro de esta nueva clase.

1.4. LENTES OPTIMIZADOS PARA PRESERVAR LA PRIVACIDAD

El estado del arte ya ha estudiado el como optimizar un lente refractivo usando aprendizaje profundo, maximizando la privacidad visual en las imágenes adquiridas mientras que se minimiza el error de un algoritmo computacional en tareas tales

---

[83] Yongqin Xian, Bernt Schiele y Zeynep Akata. «Zero-shot learning-the good, the bad and the ugly». En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, págs. 4582-4591.

[84] Bernardino Romera-Paredes y Philip Torr. «An embarrassingly simple approach to zero-shot learning». En: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. por Francis Bach y David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, págs. 2152-2161.

[85] Dwarikanath Mahapatra, Behzad Bozorgtabar y Zongyuan Ge. «Medical Image Classification Using Generalized Zero Shot Learning». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2021, págs. 3344-3353.

como, estimación de la pose humana[25], descripción de escenas[27,28], estimación de la profundidad[86], segmentación[87], reconocimiento de depresión[88] y reconocimiento de acciones[26]. En general estos trabajos han descrito la superficie del elemento óptico usando la base de Zernike en la configuración #1[29,30]

$$\phi(x, y) = \sum_{i=1}^{N_z} c_i \cdot \mathbf{Z}_i(x, y), \tag{15}$$

donde N_z es la cantidad de términos, $\mathbf{Z}_i \in \mathbb{R}^{N_x \times N_y}$ y $c_i \in \mathbb{R}$ son el i-ésimo término y coeficiente de la base de Zernike usando la notación de Noll[30]. La parametrización de Zernike ha sido ampliamente utilizada en la física debido a que los parámetros diseñables $\Phi = \{c\}_{i=1}^{N_z}$ controlan aberraciones ópticas, como desenfoque, coma y astigmatismo, ofreciendo flexibilidad para ajustar la superficie del lente. Sin embargo, los elementos descritos utilizando la base de Zernike están limitados a modelar superficies suaves y continuas, lo que los hace adecuados para superficies refractivas, pero no permiten describir las discontinuidades características de los EOD.
La optimización de Φ y Θ se logra resolviendo

$$\arg\min_{\Theta,\Phi} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{d}\sim\mathcal{D}} \Big[\alpha_{\mathsf{t}}\, \mathcal{L}_{tarea}(\mathbf{d}, \mathcal{V}_\Theta(\mathbf{C}_\Phi \mathbf{x}) + \alpha_{\mathsf{p}}\, \mathcal{L}_{privacidad}(\mathbf{x}, \mathbf{C}_\Phi \mathbf{x})\Big], \tag{16}$$

donde $\mathcal{V}_\Theta$ es la red neuronal entrenada bajo el paradigma unimodal, $\mathcal{L}_{tarea}$ es la

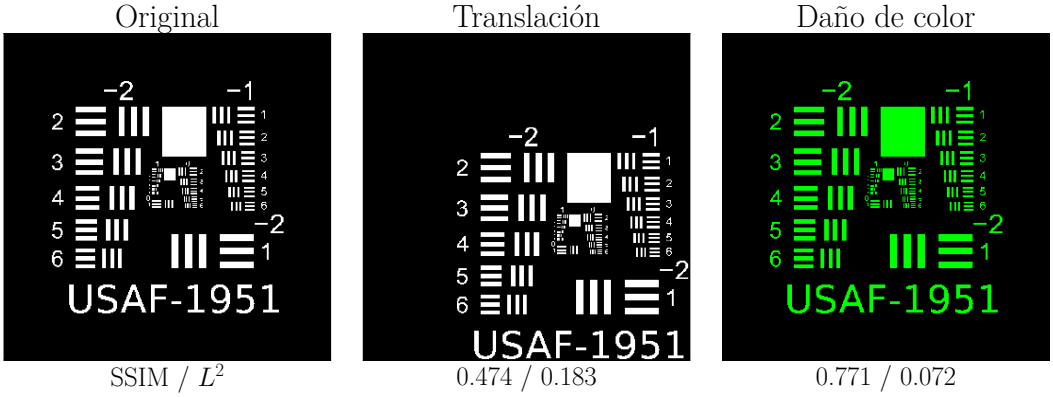
---

86 Zaid Tasneem et al. «Learning phase mask for privacy-preserving passive depth estimation». En: *European Conference on Computer Vision*. Springer. 2022, págs. 504-521.

87 Marius Dufraisse et al. «Physics Based Camera Privacy: Lens and Network Co-Design to the Rescue». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 1410-1419.

88 Yuchen Pan et al. «OpticalDR: A Deep Optical Imaging Model for Privacy-Protective Depression Recognition». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 1303-1312.

Figura 7. Ilustración de las desventajas de las métricas que cuantifican el error para preservar la privacidad. Aunque las transformaciones aplicadas realmente no preserven la privacidad, el SSIM y la norma ℓ^2 obtienen un error alto.



función de costo que depende de cada tarea y $\mathcal{L}_{privacidad}$ es una función de costo que busca maximizar el error entre la medida $\mathbf{C}_\Phi\mathbf{x}$ y la escena $\mathbf{x}$. Con respecto a $\mathcal{L}_{privacidad}$, tradicionalmente se han usado las normas ℓ^p, el índice de similitud estructural o esquemas de entrenamiento adversarial para promover privacidad. A continuación se detalla cada una con sus desventajas:

- **Normas ℓ^p:**[25,27,28,87,88] Debido a que las normas ℓ^p cuantifican el error en un espacio vectorial, se ha propuesto una estrategia que maximiza el error mediante la minimización de la norma negativa,

$$\mathcal{L}_{privacidad}(\mathbf{x}, \mathbf{C}_\Phi\mathbf{x}) = -\|\mathbf{x} - \mathbf{C}_\Phi\mathbf{x}\|_p, \tag{17}$$

  con $p \in \{1, 2\}$. Sin embargo, el negativo de la norma ℓ^p no está acotado por debajo, lo que significa que no tiene un valor mínimo al que converger, generando problemas de inestabilidad en el entrenamiento. Adicionalmente, optimizar esta métrica no es adecuada para cuantificar la privacidad debido a que existen transformaciones invertibles que no preservan la privacidad donde el valor de la norma norma ℓ^p es alto, como lo ilustra la Figura 7.

- **Índice de similitud estructural (SSIM):**[26] El SSIM es una métrica ampliamente utilizada para evaluar la calidad de una imagen en relación con una referencia. Para promover la privacidad, algunos trabajos han propuesto minimizar el SSIM de la siguiente manera:

$$\mathcal{L}_{privacidad}(\mathbf{x}, \mathbf{C}_\Phi \mathbf{x}) = \mathrm{SSIM}(\mathbf{x}, \mathbf{C}_\Phi \mathbf{x}). \tag{18}$$

  Como el SSIM está acotado entre $[0, 1]$, resuelve uno de los problemas de maximizar la norma ℓ^p, como la falta de cota inferior. Sin embargo, como lo ilustra la Figura 7, esta métrica no es adecuada para promover la privacidad debido a que existen transformaciones que no la preservan.

- **Entrenamiento adversarial:** [26,87] Para promover la privacidad visual, también se ha intentado incluir esquemas de entrenamiento adversarial maximizando el error de una red adversaria que intenta reconstruir imágenes[87] o extraer información privada[26]. Aunque efectivo para mejorar la privacidad, presenta problemas de inestabilidad, ya que requiere un equilibrio entre ambas redes. Además, incrementa la complejidad computacional al entrenar dos redes simultáneamente, lo que aumenta los tiempos de entrenamiento y los recursos necesarios.

2. MÉTODO PROPUESTO

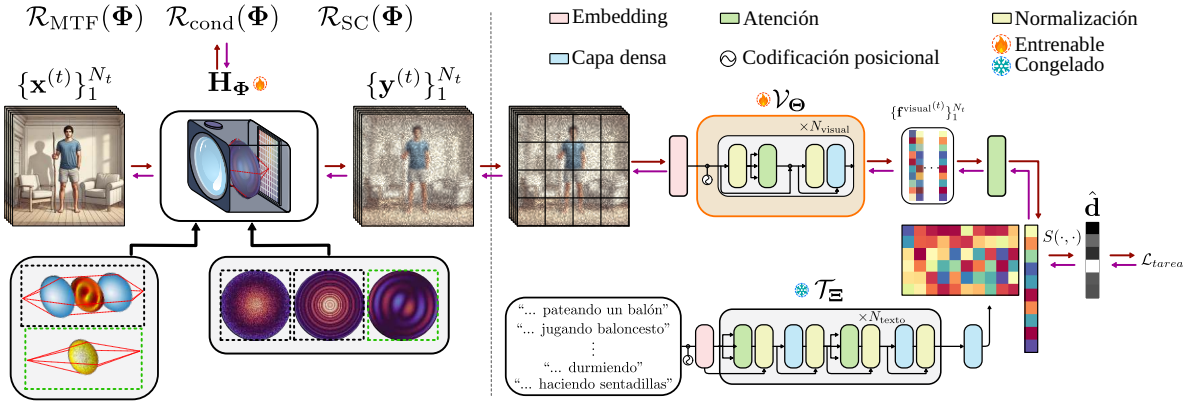
Este capítulo presenta un resumen del conjunto de elementos que componen el esquema propuesto de entrenamiento de extremo a extremo para el diseño de EODs, para la tarea de estimación de acciones en entornos clínicos usando el paradigma multimodal. Las contribuciones de este trabajo de investigación, mostradas en la Figura 8, son las siguientes:

- El estudio de un conjunto de parametrizaciones para los EODs en dos configuraciones de sistema ópticos diferentes para promover la preservación de la privacidad visual.
- El desarrollo de dos funciones de regularización sobre los parámetros del EOD (Φ) que promueven la privacidad de manera estable y con bajo costo computacional durante el entrenamiento, desarrolladas a partir de la teoría matemática y física de los modelos de adquisición de imágenes.
- Aprovechar el entrenamiento multimodal para la estimación de acciones en el contexto de preservación de la privacidad, permitiendo obtener un sistema óptico que promueva la privacidad sin sacrificar en gran medida el desempeño de estimación de acciones.
- Evaluar el desempeño del conjunto de propuestas sobre diversos conjuntos de datos y particularmente en un conjunto de datos de estimación de acciones en entornos clínicos.

2.1. PARAMETRIZACIÓN DE EODS

La superficie del elemento difractivo dentro del sistema óptico tiene la capacidad de provocar diferentes efectos sobre h_Φ, que resultarán en propiedades que pueden

Figura 8. Esquema de entrenamiento de extremo a extremo para el diseño de EODs en la tarea de estimación de acciones utilizando el paradigma multimodal. La escena de un paciente en un entorno clínico es adquirida por el sistema óptico cuya configuración es optimizada a través de la regularización, sobre los parámetros entrenables del EOD, obteniendo la medida privada. La arquitectura de red neuronal infiere la acción que se está realizando en la escena usando una red basada en el paradigma multimodal de estimación de acciones a partir de las medidas privadas.



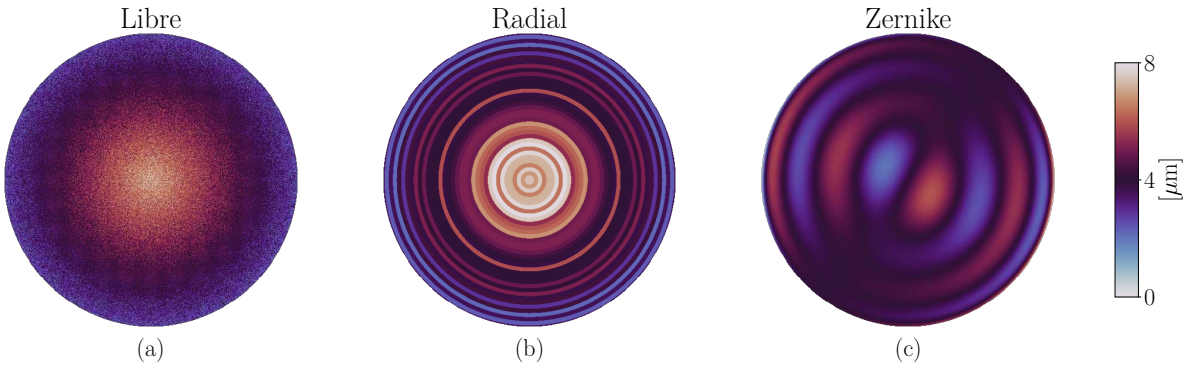
mejorar la distorsión del campo óptico. Para promover la privacidad visual, se busca evitar la reconstrucción directa de la imagen incidente, por lo tanto, la selección adecuada de la parametrización del elemento óptico difractivo es fundamental. La Figura 9 muestra las parametrizaciones que serán analizadas en el presente estudio[32].

2.1.1. Parametrización Libre. En este caso, se discretiza el mapa de alturas $\phi(x, y)$ del EOD como un arreglo bidimensional[58] siguiendo

$$\phi(x,y) = \sum_{i=1}^{N_y} \sum_{j=1}^{N_x} \Gamma_{ij} \cdot \mathrm{rect}\left(\frac{x}{d_p} - j, \frac{y}{d_p} - i\right), \tag{19}$$

donde $\mathbf{\Gamma} \in \mathbb{R}^{N_x \times N_y}$ es la matriz de alturas del EOD, $\mathrm{rect}(\cdot)$ es la función rectángulo que determina la discretización con un tamaño de píxel d_p, N_x y N_y son la cantidad de píxeles en el eje horizontal y vertical, respectivamente, y los parámetros a diseñar

Figura 9. Ilustración del mapa de alturas de un EOD usando diferentes tipos de parametrizaciones. (a) Libre, (b) Radial, (c) Zernike.



del modelo son las alturas de cada pixel y están denotados por $\Phi = \{\mathbf{\Gamma}_{ij}\}_{i=1,j=1}^{N_y,N_x}$.

2.1.2. Parametrización Radial. Debido a las propiedades que tienen los elementos ópticos que tienen invarianza rotacional[55], la Ecuación (20) describe una parametrización del EOD con esta propiedad mediante la suma de anillos concéntricos,

$$\phi(x,y) = \sum_{r=1}^{N_r} \gamma_r \cdot [\mathrm{circ}_{r+1}(x,y) - \mathrm{circ}_r(x,y)], \tag{20}$$

con $\mathrm{circ}_r(x,y)$ representando una máscara binaria circular definida como,

$$\mathrm{circ}_r(x,y) = \begin{cases} 1, & \text{si } x^2 + y^2 \leq R_r^2 \\ 0, & \text{en otro caso} \end{cases}, \tag{21}$$

donde, $r = 1, \ldots, N_r$ indexa la sección radial de la máscara con N_r el número total de radios. Los parámetros a diseñar en este modelo son $\Phi = \{\gamma_r\}_{r=1}^{N_r}$, donde γ_r corresponde a la altura de los radios.

2.2. REGULARIZACIONES QUE PROMUEVEN LA PRIVACIDAD VISUAL

Para promover la privacidad, los métodos del estado del arte han propuesto maximizar el error usando las funciones descritas en el Capítulo 1.4. Sin embargo, esta funciones pueden presentar inestabilidad durante el entrenamiento o alto costo computacional. En este trabajo proponemos retomar los conceptos de óptica y matemática relacionados con la calidad y la invertibilidad del problema de optimización en la formación de imagen descritos en las secciones 1.2.4 y 1.2.5, proponiendo dos funciones de regularización que promueven la privacidad a la vez que se calculan de manera diferenciable y eficiente en tiempo de ejecución.

2.2.1. Minimización de la MTF. En este trabajo aprovechamos la capacidad de la MTF, descrita en la sección 1.2.4, para medir la calidad de un sistema óptico. Por lo tanto, proponemos minimizar la energía de la MTF, promoviendo así la privacidad visual de las imágenes adquiridas. Esto se logra usando la siguiente regularización

$$\mathcal{R}_{\mathrm{MTF}}(\Phi) = \sum_{k=1}^{N_\lambda} \frac{1}{N_\lambda} \|\mathrm{MTF}(\mathbf{H}_{\Phi k})\|_2^2, \tag{22}$$

encargada de promediar la energía de la MTF por cada canal de color $k = \{1, \ldots, N_\lambda\}$, de la PSF $\mathcal{H}_\Phi$. Debido a que (22) involucra únicamente el calculo de una transformada de Fourier 2D de cada filtro $\mathbf{H}_{\Phi k}$, esta regularización es diferenciable y de costo computacional $O(n \log(n))$, con $n = N_x N_y$.

2.2.2. Maximización del número de condición. Un número de condición elevado $\mathcal{K}(\mathbf{C}_\Phi)$ indica la inestabilidad numérica[73] de calcular $\mathbf{C}_\Phi^{-1}\mathbf{x}$, lo que favorece la privacidad visual al hacer la reconstrucción numéricamente complicada. Sin embargo, esto no ha sido explorado con profundidad en el estado del arte debido al alto costo computacional de calcular $\mathcal{K}(\mathbf{C}_\Phi)$. Por lo tanto, en este trabajo analizamos las

matrices de convolución para optimizar el cálculo de $\mathcal{K}(\mathbf{C}_\Phi)$, de manera diferenciable y de bajo costo computacional.

Hasta este momento $\mathbf{C}_\Phi$ se ha presentado como una matriz de convolución lineal, en la que calcular los valores singulares es computacionalmente costoso. Aprovechando la teoría de tratamiento de señales es posible construir la matriz de convolución circular $\mathring{\mathbf{C}}_\Phi$, de tal manera que, realizando un rellenado con ceros adecuado en la escena[89], representado como $\mathbf{x}^{\text{pad}}$, se logra que $\mathbf{C}_\Phi\mathbf{x} = \mathring{\mathbf{C}}_\Phi\mathbf{x}^{\text{pad}}$. Por lo tanto, se puede maximizar $\mathcal{K}(\mathbf{C}_\Phi)$, maximizando $\mathcal{K}(\mathring{\mathbf{C}}_\Phi)$. Dado que $\mathring{\mathbf{C}}_\Phi$ es una matriz circulante, su espectro de valores propios $\mathbf{\Lambda} \in \mathbb{C}^n$ puede calcularse de manera eficiente[90,91,92] utilizando:

$$\mathbf{\Lambda} = \mathcal{F}^{1D}(\mathbf{h}_\Phi), \tag{23}$$

donde $\mathbf{h}_\Phi = [\text{vec}(\mathcal{H}_{\Phi_1}), \text{vec}(\mathcal{H}_{\Phi_2}), \text{vec}(\mathcal{H}_{\Phi_3})]^T$, $\text{vec}(\cdot)$ vectoriza cada PSF y $\mathcal{F}^{1D}$ es la transformada de Fourier unidimensional. Debido a que $\mathring{\mathbf{C}}_\Phi$ es una matriz normal, es decir ($\mathring{\mathbf{C}}_\Phi^T\mathring{\mathbf{C}}_\Phi = \mathring{\mathbf{C}}_\Phi\mathring{\mathbf{C}}_\Phi^T$), el número de condición se puede calcular como[90,91,92]

$$\mathcal{K}(\mathring{\mathbf{C}}_\Phi) = \frac{\max(|\mathbf{\Lambda}|)}{\min(|\mathbf{\Lambda}|)}. \tag{24}$$

Note que, para maximizar $\mathcal{K}(\mathring{\mathbf{C}}_\Phi)$ se debe minimizar el denominador de (24). Por

---

89 Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

90 Robert M Gray et al. «Toeplitz and circulant matrices: A review». En: *Foundations and Trends® in Communications and Information Theory* 2.3 (2006), págs. 155-239.

91 Bassam Bamieh. «Discovering transforms: A tutorial on circulant matrices, circular convolution, and the discrete fourier transform». En: *arXiv preprint arXiv:1805.05533* (2018).

92 Teodoro Lara. «Matrices Circulantes». En: *Divulgaciones Matemáticas* 9.1 (2001), págs. 85-102.

consiguiente, se propone la siguiente función de regularización

$$\mathcal{R}_{\text{cond}}(\Phi) = \min(|\mathbf{\Lambda}|), \tag{25}$$

Particularmente, la derivada de (25) puede calcularse de manera eficiente dentro del marco de optimización de aprendizaje profundo. En este trabajo, se empleó *PyTorch*, una librería que permite la autodiferenciación automática de funciones, lo que hace viable la optimización de (25) mediante técnicas de *backpropagation* [66]. Esta estrategia permite maximizar el número de condición de $\mathring{\mathbf{C}}_{\Phi}$ con costo computacional de $O(n\log(n))$, con $n = N_x N_y$ ya que involucra únicamente el calculo de una transformada de Fourier 1D de la PSF vectorizada $\text{vec}(\mathbf{h}_{\Phi})$, asegurando una implementación eficiente dentro del proceso de entrenamiento del sistema óptico.

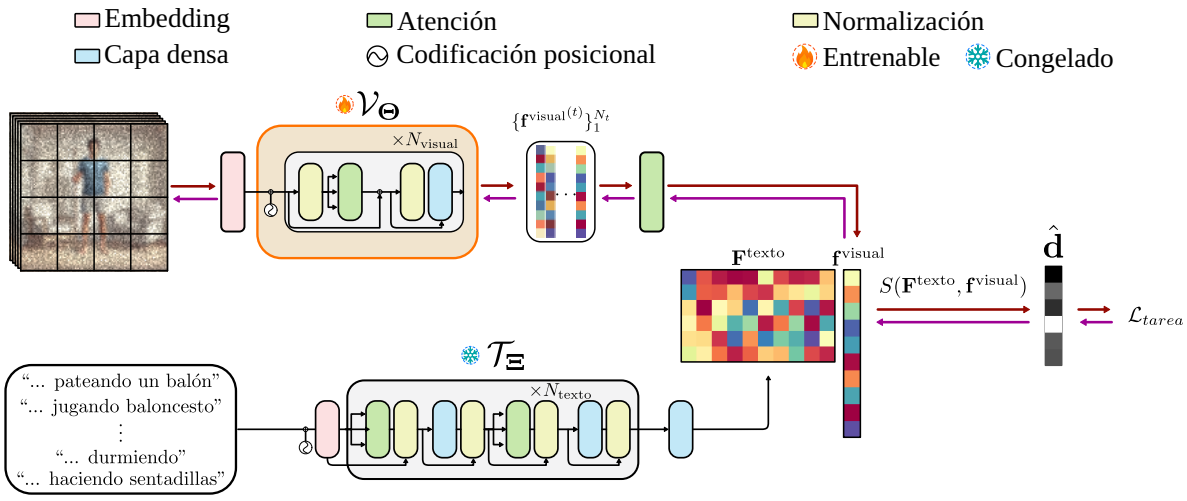
2.2.3. PSF de soporte compacto. Las regularizaciones previamente mencionadas promueven la privacidad. Sin embargo, para asegurar la factibilidad de implementación del sistema óptico, incluimos una función de regularización que garantiza que la energía de la PSF se concentre en una región específica del sensor[27,28]. Matemáticamente, esto se representa imponiendo que el soporte de la PSF sea compacto:

$$\text{supp}(\mathcal{H}_{\Phi}) = \{(i,j) \mid i^2 + j^2 \leq r^2\}, \tag{26}$$

donde (i,j) son los píxeles espaciales y $r \in \mathbb{Z}^+$ es una distancia definida en píxeles. Esta condición se puede fomentar durante el entrenamiento mediante la siguiente función de regularización:

$$\mathcal{R}_{\text{SC}}(\Phi) = \frac{1}{N_\lambda} \sum_{k=1}^{N_\lambda} \|\mathbf{M} \odot \mathcal{H}_{\Phi k}\|_{\text{F}}, \tag{27}$$

Figura 10. Arquitectura de red neuronal para la estimación de acciones basado en el paradigma multimodal a partir de medidas privadas.



donde $\|\cdot\|_F$ es la norma Frobenius y $\mathbf{M}$ es una máscara binaria que selecciona la sección del sensor donde la PSF no debería concentrarse, definida como

$$\mathbf{M}_{i,j} = \begin{cases} 1, & \text{si } i^2 + j^2 \geq r^2, \\ 0, & \text{en otro caso.} \end{cases} \tag{28}$$

2.3. OPTIMIZACIÓN DE EXTREMO A EXTREMO

2.3.1. Red neuronal para la estimación de acciones. En este trabajo de investigación aprovechamos las propuestas del estado del arte para la estimación de acciones bajo el paradigma multimodal[81,82]. Concretamente, exploramos el uso de la arquitectura CLIP[93] (del inglés, *Contrastive Language–Image Pre-training*), entrenada para relacionar las características visuales y textuales extraídas por los modelos $\mathcal{V}_\Theta$ y $\mathcal{T}_\Xi$ en la Figura 10.

---

93 Alec Radford et al. «Learning transferable visual models from natural language supervision». En: *International Conference on Machine Learning*. PMLR. 2021, págs. 8748-8763.

$\mathcal{V}_\Theta$ corresponde a una arquitectura de *Transformer* de visión[94], la cual toma cada fotograma del video de manera independiente y genera parches espaciales de la escena para calcular un espacio de representación, llamado *embedding*, para cada parche. Luego, utiliza $N_{\text{visual}} = 6$ bloques de capas de atención, normalización y capas densas para calcular las características visuales de los fotogramas privados. Posteriormente, emplea un *Transformer* temporal para relacionar estas características y generar un vector de características visuales $\mathbf{f}^{\text{visual}}$ de la escena.

En paralelo, $\mathcal{T}_\Xi$ corresponde a una arquitectura de *Transformer*[95] de texto, que toma los *embeddings* del conjunto de acciones y genera la matriz de características $\mathbf{F}^{\text{texto}}$ utilizando $N_{\text{texto}} = 12$ bloques de módulos de atención, normalización y capas densas.

Por último, la acción realizada en la escena se estima como aquella que tiene la máxima similitud entre los vectores de características textuales y visuales $S(\mathbf{F}^{\text{texto}}, \mathbf{f}^{\text{visual}})$.

Como punto de partida para la tarea de estimación de acciones realizada en este trabajo, se inicializan los pesos Ξ a partir de un preentrenamiento sobre 400 millones de pares de imágenes y textos extraídos de internet[93], y los pesos de Θ con el preentrenamiento para estimación de acciones en video realizado en[82].

2.3.2. Problema de optimización. Para obtener un algoritmo de estimación de acciones que preserva la privacidad visual, se propone optimizar Θ y Φ, usando el paradigma de estimación de acciones multimodal y las parametrizaciones propuestas, dentro de las configuraciones ópticas #1 y #2, mediante el siguiente problema

---

[94] Alexey Dosovitskiy et al. «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale». En: *International Conference on Learning Representations*. 2021.

[95] A Vaswani. «Attention is all you need». En: *Advances in Neural Information Processing Systems* (2017).

de optimización

$$\arg\min_{\Theta,\Phi} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{d}\sim\mathcal{D}} \alpha_{\mathsf{t}}[\mathcal{L}_{tarea}(\mathbf{d}, S(\mathbf{F}^{\text{texto}}, \mathcal{V}_\Theta(\mathbf{C}_\Phi \mathbf{x})))] + \alpha_{\mathsf{p}}[\mathcal{R}_{\text{cond}}(\Phi) + \mathcal{R}_{\text{MTF}}(\Phi) + \mathcal{R}_{\text{SC}}(\Phi)]. \tag{29}$$

El problema de optimización busca minimizar el valor esperado $\mathbb{E}$ de la función de costo $\mathcal{L}_{tarea}$ asociada al problema el reconocimiento de acciones, en este caso la entropía cruzada, y las funciones de regularización $\mathcal{R}_{\text{cond}}(\Phi)$ y $\mathcal{R}_{\text{MTF}}(\Phi)$, sobre los parámetros ópticos Φ que maximizan el número de condición y minimizan la MTF, respectivamente. Para tener un balance entre la tarea y la preservación de la privacidad visual, los coeficientes $\alpha_{\mathsf{t}} \in \mathbb{R}$ y $\alpha_{\mathsf{p}} \in \mathbb{R}$ controlan la ponderación de la función de costo de la tarea $\mathcal{L}_{tarea}$ y las regularizaciones $\mathcal{R}_{\text{cond}}(\Phi), \mathcal{R}_{\text{MTF}}(\Phi)$, respectivamente.

2.3.3. Algoritmo de optimización. El Algoritmo 1 resume el proceso de entrenamiento de extremo a extremo propuesto para optimizar los parámetros Φ del EOD y los parámetros de la red neuronal Θ. En la Línea 2, se inicializan los parámetros Φ en cero y los parámetros Θ con valores preentrenados sobre el problema de estimación de acciones de manera no privada[81,82]. El entrenamiento se realiza a lo largo de $\mathcal{E} \in \mathbb{N}$ épocas (ver Línea 3), y en cada una de ellas se itera sobre $N_d \in \mathbb{N}$ ejemplos del conjunto de datos $\mathcal{D}$ (ver Línea 4). En la Línea 5 se genera la simulación de la adquisición óptica de la escena a través del sistema $\mathbf{y}^{(j)} = \mathbf{C}_\Phi \mathbf{x}^{(j)}$ para cada ejemplo j. Posteriormente en la Línea 6, se extraen las características visuales $\mathbf{f}^{\text{visual}} = \mathcal{V}_\Theta(\mathbf{y}^{(j)})$ y se comparan con las características textuales $\mathbf{F}^{\text{texto}}$ para calcular la similaridad $\hat{\mathbf{d}} = S(\mathbf{F}^{\text{texto}}, \mathcal{V}_\Theta(\mathbf{y}^{(j)}))$ en la Línea 7. El valor de las funciones de costo para la tarea y la privacidad se obtienen en las Líneas 8 y 9 y se calcula la suma ponderada en la Línea 10. Por último, se usa el algoritmo Adam[65] en las Líneas 11 y 12 con una taza de aprendizaje β_{tarea} y $\beta_{\text{privacidad}}$ para optimizar los parámetros ópticos Φ y los parámetros de la red neuronal Θ, respectivamente. Al finalizar el entrenamiento, los parámetros optimizados Φ y Θ se retornan como salida del

algoritmo.

Algoritmo 1 Entrenamiento de extremo a extremo para resolver (29)

Entrada: Conjunto de entrenamiento $\mathcal{D} = \{\mathbf{x}^{(j)}, \mathbf{d}^{(j)}\}_{j=1}^{N_d}$, Matriz de características textuales $\mathbf{F}^{\text{texto}}$.
Inicialización parámetros: $\Phi = \mathbf{0}$, $\Theta = \Theta^{preentrenados}$
para $e = 1 : \mathcal{E}$ **hacer** ▷ $\mathcal{E}$ épocas
    para $j = 1 : N_d$ **hacer** ▷ N_d ejemplos
        $\mathbf{y}^{(j)} = \mathbf{C}_{\Phi}\mathbf{x}^{(j)}$ ▷ Adquisición óptica
        $\mathbf{f}^{\text{visual}} \leftarrow \mathcal{V}_{\Theta}(\mathbf{y}^{(j)})$ ▷ Extracción características visuales
        $\hat{\mathbf{d}} = S(\mathbf{F}^{\text{texto}}, \mathbf{f}^{\text{visual}})$ ▷ Cálculo similaridad
        $\ell_{\text{tarea}} = \mathcal{L}_{tarea}(\mathbf{d}, \hat{\mathbf{d}})$ ▷ Costo reconocimiento de acciones
        $\ell_{\text{privacidad}} = \mathcal{R}_{\text{cond}}(\Phi) + \mathcal{R}_{\text{MTF}}(\Phi) + \mathcal{R}_{\text{SC}}(\Phi)$ ▷ Regularización privacidad
        $\ell_{\text{total}} = \alpha_{\text{t}}\ell_{\text{tarea}} + \alpha_{\text{p}}\ell_{\text{privacidad}}$ ▷ costo toal
        $\Phi \leftarrow \text{ADAM}(\beta_{\text{tarea}}, \Phi, \nabla_{\Phi}\ell_{\text{total}})$ ▷ Optimización sobre Φ
        $\Theta \leftarrow \text{ADAM}(\beta_{\text{privacidad}}, \Theta, \nabla_{\Theta}\ell_{\text{total}})$ ▷ Optimización sobre Θ
    fin para
fin para
Salida: Parámetros óptimos: Θ, Φ.

3. SIMULACIONES Y RESULTADOS

3.1. CONFIGURACIÓN DE LAS SIMULACIONES

A continuación, se describen los detalles de las simulaciones con el objetivo de replicar los resultados.

3.1.1. Método de línea base: ActionCLIP + PrivHAR Con el fin de comparar las parametrizaciones propuestas con la parametrización de Zernike previamente explorada en el estado del arte[25,26,27,28], se implementó la arquitectura óptica del esquema PrivHAR[26]. Este esquema consiste en utilizar la función (17) para promover la privacidad en las escenas adquiridas, empleando un sistema óptico compuesto por un lente en la configuración #1, parametrizado usando la base de Zernike. Para evaluar el efecto de la parametrización óptica, se adaptó PrivHAR para utilizar el paradigma multimodal de estimación de acciones. Este método se utilizará como línea base para las comparaciones y se denominará "ActionCLIP+PrivHAR" en los resultados.

3.1.2. Infraestructura usada. Los experimentos reportados en este trabajo fueron ejecutados en un computador con CPU Intel Core i7-10700K a 3.8 GHz con 32 GB de memoria RAM, equipado con una GPU NVIDIA GeForce RTX 3080 Ti de 12 GB VRAM. Las simulaciones de las arquitecturas ópticas, la red neuronal y el algoritmo de extremo a extremo fueron implementadas usando la librería de Pytorch[66].

3.1.3. Algoritmo de optimización. En el Algoritmo 1 se inicializaron las tasas de aprendizaje $\beta_{\text{tarea}} = 5 \times 10^{-5}, \beta_{\text{privacidad}} = 1 \times 10^{-2}$ y se aplicó el esquema de actualiza-

ción de coseno[96] para variar gradualmente la tasa de aprendizaje. El entrenamiento se realizó a lo largo de 30 épocas, con un tamaño de lote de 3, resultando en un tiempo de entrenamiento aproximado de 8 horas.

3.1.4. Inicialización de parámetros entrenables. Los métodos de visión por computadora que consideran la privacidad dependen fuertemente del punto inicial de los pesos Φ y Θ. Para todos los experimentos realizados en este trabajo, los pesos iniciales de la red neuronal Θ se establecieron utilizando los pesos preentrenados proporcionados por[82]. Dado que esta inicialización no considera la privacidad visual, los parámetros ópticos Φ se inicializaron de manera que las configuraciones ópticas adquieran imágenes sin distorsiones. En la Configuración $\#1$, esto se logró fijando la altura del mapa de fase como la suma de un lente de Fresnel que forma imagen y un EOD entrenable, es decir, $\phi = \phi_{\text{Fresnel}} + \phi_{\text{EOD entrenable}}$. Para la Configuración 2, se estableció $\phi = \phi_{\text{EOD entrenable}}$, ya que el uso del arreglo óptico $4f$ genera el efecto deseado de formación de imagen sin necesidad de un lente adicional.
Para ambas configuraciones, los parámetros iniciales del elemento óptico entrenable se fijaron a cero. Específicamente, se estableció $c_i = 0$, $\gamma_r = 0$ y $\mathbf{\Gamma}_{ij} = 0$ para las parametrizaciones Zernike, Radial y Libre, respectivamente. Esto permite que, al inicio del entrenamiento el sistema óptico comience desde una parametrización que forma imágenes convencionales, pero al entrenarse ϕ se promueve la privacidad visual en las imágenes.

3.1.5. Detalles de la óptica. En esta sección, se evalúan las parametrizaciones propuestas utilizando las configuraciones ópticas $\#1$ y $\#2$, ilustradas en la Figura 3. En el lente usado como inicialización en configuración $\#1$, junto a los lentes del

---

[96] Ilya Loshchilov y Frank Hutter. «SGDR: Stochastic Gradient Descent with Warm Restarts». En: *International Conference on Learning Representations*. 2022.

Figura 11. Ejemplo de imágenes presentes en los conjuntos de datos HMDB51, HPTE, DDPD.

arreglo óptico $4f$, se empleó un mapa de alturas de lentes de Fresnel.

Para la simulación de los lentes de EOD, se utilizó *polidimetilsiloxano* (PDMS), un material ampliamente empleado en la literatura para la fabricación de EODs[52].

En ambas configuraciones, la distancia entre la escena y el sistema óptico fue fijada en $z_i = 3$m. En la configuración $\#1$, la distancia entre el EOD y el sensor fue de aproximadamente 50 mm. En la configuración $\#2$, se estableció una distancia entre lentes de 200 mm en el arreglo $4f$, con el EOD posicionado a la mitad de esta distancia, es decir, a 100 mm del primer lente. Así, la distancia entre el último lente del arreglo y el sensor fue de aproximadamente 75 mm.

3.2. CONJUNTOS DE DATOS

Para entrenar y evaluar la arquitectura de aprendizaje profundo para la estimación de acciones, en conjunto con los parámetros del EOD (ver Fig. 8), usando el Algoritmo 1 propuesto, se hicieron uso de los siguientes conjuntos de datos, mostrados en la Figura 11.

- ***Human Motion Database* (HMDB51)[97]:** Este conjunto de datos cuenta con $6,849$ videos recopilados principalmente de películas, y bases de datos públicas como el archivo Prelinger, YouTube y Google. El conjunto de datos contiene etiquetas de 51 categorías de acciones, cada una con un mínimo de 101 videos. La división de los videos usada para entrenamiento y validación corresponde a la sugerida por los autores con una proporción de $78\,\%$ y $22\,\%$, respectivamente. Este conjunto fue usado principalmente para realizar estudios de ablación y así encontrar la combinación óptima de los coeficientes $\alpha_{\mathsf{t}}, \alpha_{\mathsf{p}}$.

- ***Home-Based Physiotherapy Exercises* (HPTE) [98]:** Este conjunto de datos cuenta con 240 videos de 5 pacientes realizando 8 ejercicios de fisioterapia en 6 repeticiones. Para obtener resultados más robustos, se utilizó la estrategia de validación cruzada *K-fold* con $K = 5$. En cada iteración, el conjunto de entrenamiento corresponde a los datos de cuatro pacientes y se evalúa en el paciente restante, resultando en una división de $80\,\%$ para entrenamiento y $20\,\%$ para validación, respectivamente. Este conjunto de datos fue usado para evaluar el aprendizaje de los EODs sobre un conjunto en ambiente clínico.

- ***Dual-Pixel Defocus Deblurring* (DDPD)[99]:** Este conjunto de datos cuenta con 2000 imágenes de 1680×1120 píxeles, capturadas en 500 escenas distintas. Para reducir la complejidad del entrenamiento y mejorar el desempeño del modelo, se aplicó aumento de datos mediante recortes aleatorios de tamaño 256×256. La partición del conjunto de datos se realizó con una proporción

---

97 H. Kuehne et al. «HMDB: a large video database for human motion recognition». En: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2011.

98 Ilktan Ar y Yusuf Sinan Akgul. «A Computerized Recognition System for the Home-Based Physiotherapy Exercises Using an RGBD Camera». En: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22.6 (2014), págs. 1160-1171.

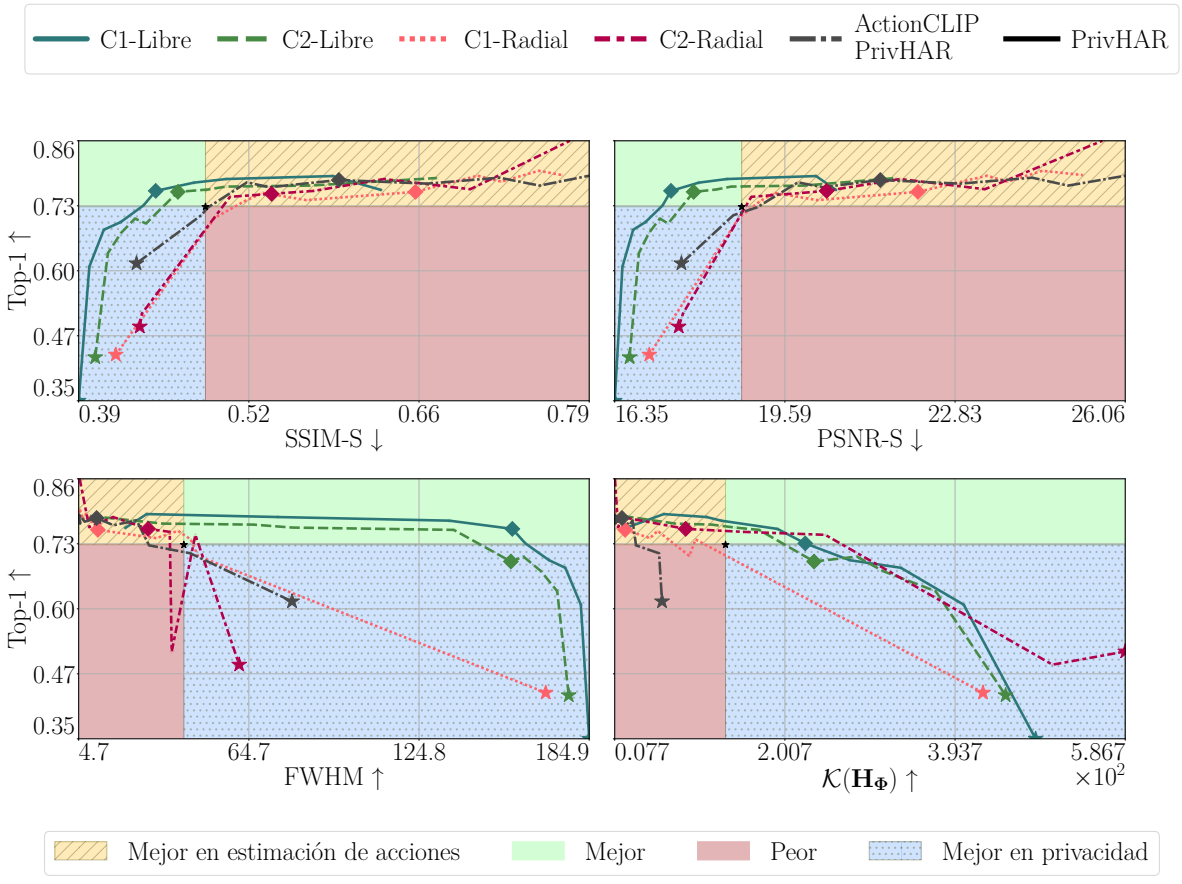
de 80 % para entrenamiento y 20 % para evaluación. Este conjunto de datos fue usado para realizar el entrenamiento de una red neuronal encargada de realizar la reconstrucción de las medidas privadas simuladas en este trabajo de investigación.

3.3. BALANCE ENTRE RECONOCIMIENTO DE ACCIONES Y PRIVACIDAD

En el campo de los métodos de visión por computadora que consideran la privacidad existe un balance entre el desempeño de la tarea y la privacidad visual, controlado por los coeficientes α_{t} y α_{p}. Para estudiar este equilibrio, se fijó $\alpha_{\mathsf{t}} = 1 - \alpha_{\mathsf{p}}$ y se realizó un estudio de ablación utilizando 10 valores equidistantes de $\alpha_{\mathsf{p}} \in [0, 1]$. La Figura 12 presenta los resultados obtenidos para las configuraciones $\#1$ y $\#2$, comparando las parametrizaciones Zernike, Libre, Radial con el método de referencia PrivHAR[26]. Para evaluar el equilibrio entre privacidad y desempeño, se calculó la exactitud de clasificación en el conjunto de datos HMDB51 y se comparó con diversas métricas de preservación de la privacidad. Estas métricas incluyen el FWHM, el número de condición $\mathcal{K}(\mathbf{C}_\Phi)$, y las métricas SSIM y PSNR. En particular, una mayor privacidad visual se asocia al incremento en FWHM y el número de condición $\mathcal{K}(\mathbf{C}_\Phi)$, y una disminución de SSIM y PSNR. Los resultados del método PrivHAR[26] se incluyeron como línea base, definiéndose cuatro regiones: resultados superiores tanto en privacidad como en reconocimiento de acciones, resultados inferiores en ambas métricas, y aquellos que superan al estado del arte solo en términos de estimación de acciones (marcados con /) o de privacidad (marcados con $\cdot$). En la Figura 12, ♦ y ★ simbolizan el punto más balanceado entre privacidad y estimación de acciones y el punto que maximiza la privacidad para cada parametrización, respectivamente. Los resultados numéricos de estos puntos se comparan en el Cuadro 1.

En resumen, la parametrización Libre en la configuración $\#1$ alcanza el mejor equilibrio entre privacidad y estimación de acciones, debido a que el codo de la curva tie-

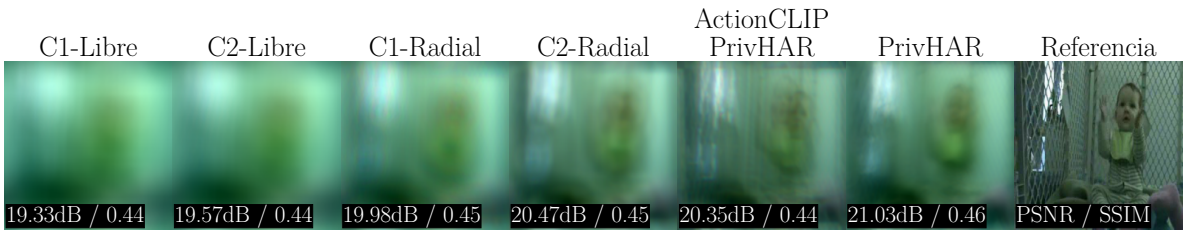
Figura 12. Balance entre estimación de acciones (exactitud) y preservación de la privacidad visual (SSIM, PSNR, FWHM y $\mathcal{K}(\mathbf{C}_\Phi)$).



Cuadro 1. Resultados de evaluación sobre el conjunto HMDB en las métricas de estimación de acciones, como la exactitud Top-1 y Top-5, así como las métricas de privacidad: el número de condición $\mathcal{K}(\mathbf{C}_\Phi)$, el FWHM y el SSIM de la medida privada. Las flechas $^\uparrow$ indican que métricas más altas son mejores, mientras que $^\downarrow$ indican que métricas más bajas son mejores.

Óptica	**Top-1** $\uparrow$	**Top-5** $\uparrow$	**SSIM-S** $\downarrow$	$\mathcal{K}(\mathbf{H}_\Phi)$ $\uparrow$	**FWHM** $\uparrow$
Mejor balance entre privacidad y estimación de acciones (♦ en Fig. 12)					
C1-Libre	0.73	0.92	**0.44**	223.83	**162.38**
C2-Libre	0.70	0.93	0.44	**233.79**	157.27
C1-Radial	0.76	**0.95**	0.65	19.54	11.45
C2-Radial	0.76	0.94	0.57	88.21	29.17
ActionCLIP+PrivHAR	**0.78**	0.95	0.59	16.03	10.99
Máxima privacidad (★ en Fig. 12)					
C1-Libre	0.35	0.66	**0.39**	485.06	**184.92**
C2-Libre	0.43	0.75	0.40	451.04	177.69
C1-Radial	0.44	0.77	0.42	425.38	169.56
C2-Radial	0.52	0.81	0.44	**586.68**	37.51
ActionCLIP+PrivHAR	**0.62**	**0.88**	0.43	61.40	79.99

Figura 13. Ejemplo de una adquisición del sensor utilizando cada una de las parametrizaciones, para el caso de máxima privacidad (★ en Fig. 12), usando una imagen del conjunto de datos HMDB51.



ne el mejor resultado de clasificación para altos niveles de privacidad visual, destacándose como la opción más relevante para los usuarios interesados en un sistema balanceado. En el caso que el mayor interés sea la tarea, la parametrización Radial en la configuración $\#2$ obtiene el mejor desempeño en estimación de acciones. En términos de privacidad, la parametrización Libre en la configuración $\#2$ ofrece la mayor privacidad, obteniendo un número de condición $\times 14$ veces mayor que el de la parametrización de Zernike. Adicionalmente, la Figura 13 muestra un ejemplo de una adquisición del sensor para cada una de las parametrizaciones, donde se puede observar que las parametrizaciones propuestas ofrecen mayor privacidad visual, en comparación con el estado del arte. Estos resultados demuestran que las parametrizaciones propuestas superan el estado del arte al ofrecer alternativas tanto en privacidad visual como en desempeño en la estimación de acciones.

Es importante destacar que los tiempos de entrenamiento e inferencia del método propuesto están influenciados por el tipo de parametrización empleada. El Cuadro 2 presenta una comparación de los tiempos de entrenamiento e inferencia, los *Frames per second* (FPS) y la cantidad de parámetros optimizables para cada tipo de parametrización.

Aunque las parametrizaciones de Zernike y Radial requieren la menor cantidad de parámetros entrenables, estas requieren el cálculo del mapa de alturas resultante de la suma de cada coeficiente c_i y γ_r. Este enfoque resulta eficiente en términos

Cuadro 2. Resumen costo computacional del método propuesto representado por el tiempo de entrenamiento sobre el conjunto de datos HMDB, el tiempo de inferencia de una imagen, *Frames per second* (FPS) alcanzados y el número de parámetros entrenables del EOD.

Parametrización	**Entrenamiento** $[h]$	**FPS**	**#** Φ
C1-Libre	$\mathbf{6{,}37 \pm 0{,}17}$	$\mathbf{92{,}00 \pm 0{,}27}$	10^6
C2-Libre	$6{,}59 \pm 0{,}56$	$90{,}38 \pm 0{,}81$	10^6
C1-Radial	$7{,}23 \pm 0{,}15$	$82{,}45 \pm 0{,}28$	$\mathbf{150}$
C2-Radial	$7{,}28 \pm 0{,}18$	$81{,}63 \pm 0{,}23$	$\mathbf{150}$
Zernike	$7{,}45 \pm 0{,}16$	$82{,}08 \pm 0{,}19$	$\mathbf{150}$

de uso de memoria, pero conlleva un alto costo computacional, incrementando los tiempos de entrenamiento e inferencia.

Por otro lado, la parametrización Libre describe la superficie del EOD Γ_{ij} para cada píxel, lo que simplifica el cálculo del mapa de alturas. Como consecuencia, esta parametrización no solo disminuye los tiempos de entrenamiento, sino que también incrementa la cantidad de FPS.

3.4. ESTIMACIÓN DE ACCIONES EN ENTORNOS CLÍNICOS

En esta sección se presentan los resultados de la evaluación del método propuesto para la estimación de acciones en entornos clínicos, utilizando el conjunto de datos HPTE[98]. Se reportan las métricas de estimación de acciones, como la exactitud Top-1 y Top-5, así como las métricas de privacidad: el número de condición $\mathcal{K}(\mathbf{C}_\Phi)$, el FWHM y el valor de SSIM de la medida privada. Las flechas $^\uparrow$ indican que métricas más altas son mejores, mientras que $^\downarrow$ indican que métricas más bajas son preferibles. Para evaluar el desempeño del método propuesto, se diseñaron dos casos experimentales comunes en sistemas de reconocimiento en entornos clínicos[8,85].

3.4.1. Entrenamiento con Ejemplos. El Cuadro 3 presenta los resultados del experimento en el que se dispone de un conjunto de datos etiquetado de videos

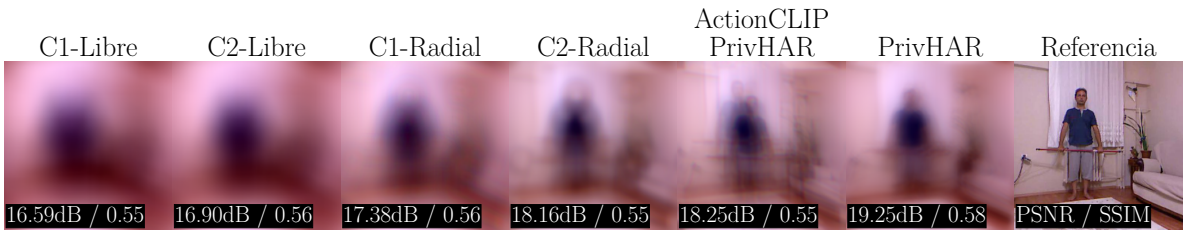
Cuadro 3. Resultados del experimento de entrenamiento con ejemplos, reportando los resultados de evaluación sobre el conjunto HTPE en las métricas de estimación de acciones, como la exactitud Top-1 y Top-5, así como las métricas de privacidad: el número de condición $\mathcal{K}(\mathbf{C}_\Phi)$, el FWHM y el SSIM de la medida privada. Las flechas $^\uparrow$ indican que métricas más altas son mejores, mientras que $^\downarrow$ indican que métricas más bajas son mejores.

Óptica	**Top-1** $\uparrow$	**Top-5** $\uparrow$	**SSIM-S** $\downarrow$	$\mathcal{K}(\mathbf{H}_\Phi)$ $\uparrow$	**FWHM** $\uparrow$
C1-Libre	0.69	0.90	**0.42**	465.60	**184.81**
C2-Libre	0.64	0.90	0.44	438.87	177.61
C1-Radial	0.76	0.92	0.50	348.95	80.83
C2-Radial	0.82	0.93	0.50	**525.58**	45.66
ActionCLIP+PrivHAR	**0.96**	**1.00**	0.50	86.30	35.55

y acciones en entornos clínicos. El método de estimación de acciones se entrenó utilizando el conjunto de datos HPTE de acciones de fisioterapia. Se reporta el promedio de validación de cinco experimentos, empleando la estrategia de validación cruzada *K-fold* con $K = 5$. En cada iteración, el modelo se entrenó con los datos de cuatro pacientes y se evaluó en el paciente restante, variando en cada caso el paciente utilizado para la evaluación.

Aunque la parametrización de Zernike alcanzó el mejor resultado en términos de exactitud de estimación de acciones, las métricas de privacidad como el SSIM, el número de condición $\mathcal{K}(\mathbf{C}_\Phi)$ y el FWHM indican que no proporciona el nivel óptimo de preservación de la privacidad. Esto se puede confirmar visualmente en la Figura 14, donde se observa un ejemplo de una adquisición para cada parametrización en el caso de máxima privacidad ($\bigstar$ en Fig. 12) . En contraste, las parametrizaciones Libre y Radial logran valores comparables de exactitud mientras mantienen una mejor privacidad de la escena. En particular, la parametrización Libre ofrece la máxima privacidad visual, mientras que la parametrización Radial se destaca como la más equilibrada en términos de estimación de acciones y preservación de la privacidad visual.

Figura 14. Ejemplo de una adquisición del sensor utilizando cada una de las parametrizaciones, resultantes de el experimento reportado en el Cuadro 3, usando una imagen del conjunto de datos HTPE.



3.4.2. Evaluación *zero-shot*. En el segundo experimento, se considera la situación en la que no es posible obtener un conjunto de datos de entrenamiento etiquetado. En este escenario, se aprovechó la capacidad del método propuesto para estimar las acciones realizadas en entornos clínicos sin haber sido entrenado explícitamente con estas clases, utilizando la técnica de aprendizaje *zero-shot* descrita en la Sección 1.3.2. Para este experimento se emplearon los pesos optimizados sobre el conjunto de datos HMDB, cuyos resultados se presentaron en el Cuadro 1, y se evaluó el modelo sobre todos los pacientes del conjunto de datos HPTE.

Para este experimento, la parametrización Radial se establece como el mejor resultado de estimación de acciones realizando *zero-shot*, mientras que logra la mejor privacidad visual representada por la $\mathcal{K}(\mathbf{C}_\Phi)$. Estos resultados sugieren que la parametrización Radial es particularmente adecuada para escenarios sin datos etiquetados, ofreciendo un equilibrio óptimo entre precisión en la estimación de acciones y preservación de la privacidad visual.

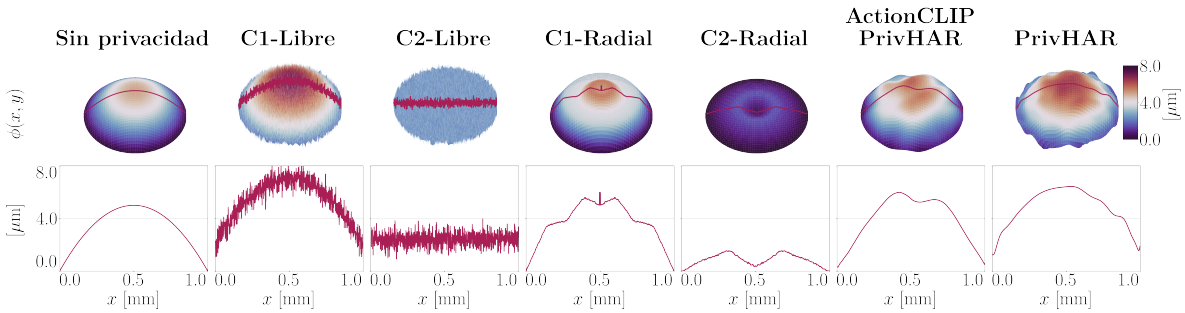
3.5. ANÁLISIS DE LOS EOD OPTIMIZADOS

La Figura 15 muestra los elementos ópticos difractivos que proporcionan mayor privacidad visual medida a través de un mayor número de condición, optimizados para la estimación de acciones. En particular, la parametrización Libre muestra un comportamiento pseudoaleatorio en su mapa de alturas tanto para la configuración $\#1$

Cuadro 4. Resultados del experimento *zero-shot*, reportando los resultados de evaluación sobre el conjunto HTPE en las métricas de estimación de acciones, como la exactitud Top-1 y Top-5, así como las métricas de privacidad: el número de condición $\mathcal{K}(\mathbf{C}_\Phi)$, el FWHM y el SSIM de la medida privada. Las flechas $^\uparrow$ indican que métricas más altas son mejores, mientras que $^\downarrow$ indican que métricas más bajas son mejores.

Óptica	**Top-1** $\uparrow$	**Top-5** $\uparrow$	**SSIM-S** $\downarrow$	$\mathcal{K}(\mathbf{H}_\Phi)$ $\uparrow$	**FWHM** $\uparrow$
Mejor balance entre privacidad y estimación de acciones (♦ en Fig. 12)					
C1-Libre	0.35	0.81	**0.47**	223.83	**162.38**
C2-Libre	0.31	0.91	0.47	**233.79**	157.27
C1-Radial	**0.50**	**0.91**	0.68	19.54	11.45
C2-Radial	0.29	0.87	0.61	88.21	29.17
ActionCLIP+PrivHAR	0.48	0.83	0.63	16.03	10.99
Máxima privacidad (★ en Fig. 12)					
C1-Libre	0.27	0.68	**0.42**	485.06	**184.92**
C2-Libre	0.22	0.75	0.44	451.04	177.69
C1-Radial	0.32	0.74	0.46	425.38	169.56
C2-Radial	**0.35**	0.79	0.49	**586.68**	37.51
ActionCLIP+PrivHAR	0.35	**0.84**	0.48	61.40	79.99

Figura 15. Elementos ópticos difractivos que proporcionan mayor privacidad visual, medida a través de un mayor número de condición, optimizados para la estimación de acciones.



y $\#2$; por otro lado, la parametrización Radial muestra un mapa de alturas con cambios más suaves, pero conservando pequeñas variaciones entre anillos radiales. Note que la parametrización de Zernike genera un mapa de alturas continuo, con cambios suaves en su apertura.

La Figura 16 muestra la PSF y la amplitud y fase de la OTF, para cada mapa de alturas obtenido. La parametrización Libre es la que obtiene el mayor número de condición, por lo tanto, las PSFs resultantes tienen un comportamiento pseudoaleatorio en un área amplia sobre el sensor. En comparación, la parametrización Radial genera PSFs simétricas angularmente, con patrones de intensidad radiales. Por último, las PSFs generadas por la parametrización de Zernike resultan en patrones cáusticos, donde la amplitud de la OTF muestra que esta PSF no logra atenuar correctamente las bajas frecuencias de la señal. Por otro lado, la magnitud de la OTF de las parametrizaciones propuestas logra atenuar mejor las bajas frecuencias, proporcionando una mayor privacidad visual.

Para profundizar en el análisis de la distorsión generada por cada parametrización, la Figura 17 presenta el promedio radial y en longitud de onda de la MTF, junto con el valor de $\mathcal{R}_{\mathrm{MTF}}(\Phi)$ correspondiente a cada curva. Adicionalmente, se muestra el promedio por longitud de onda de la distribución de la magnitud de los valores propios de la matriz de convolución, acompañado del número de condición $\mathcal{K}(\mathbf{C}_{\Phi})$.

Figura 16. PSF, amplitud y fase de la OTF de los mapas de alturas (ϕ) optimizados para las parametrizaciones Libre y Radial en las configuraciones $\#1$ y $\#2$, comparadas con la parametrización de Zernike.

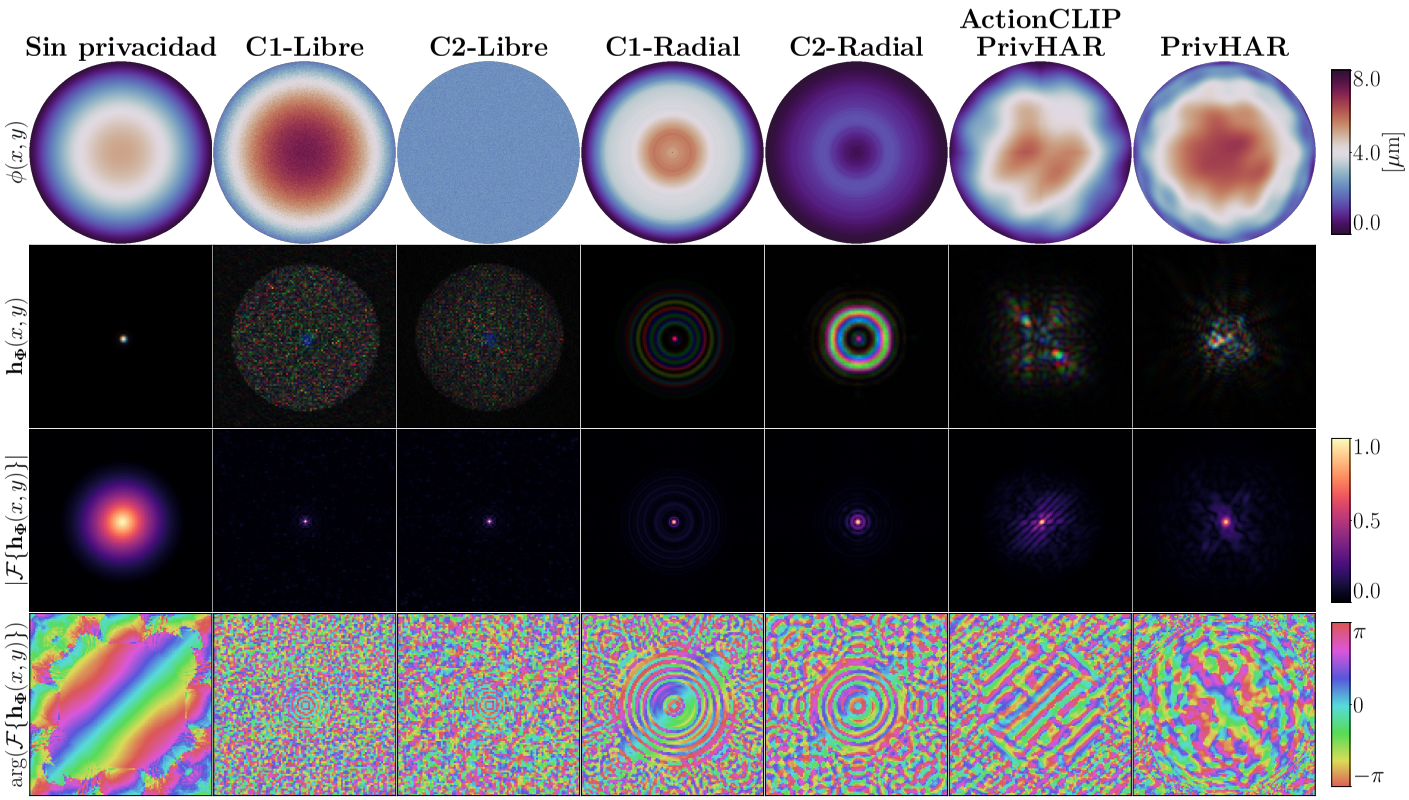
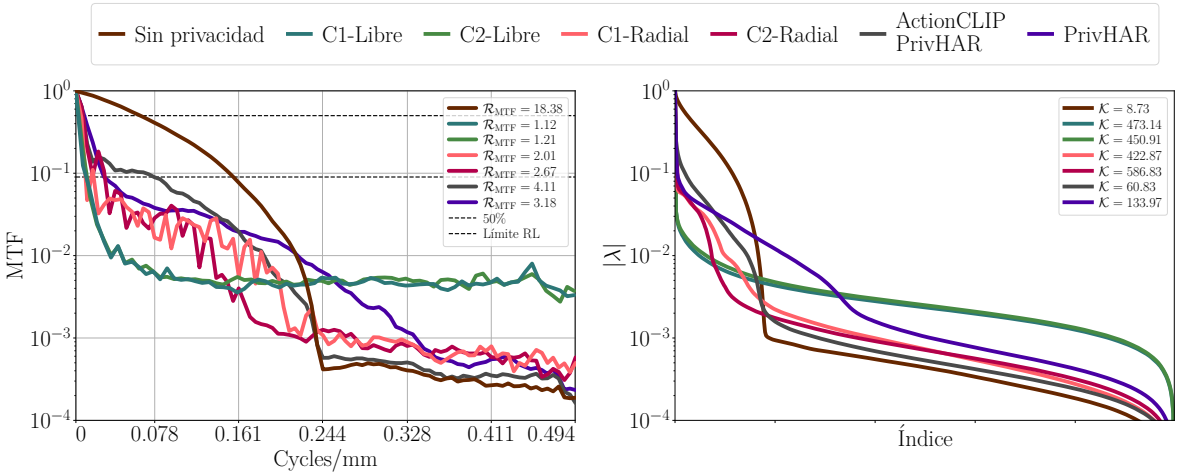


Figura 17. (Izquierda) Promedio radial y en longitud de onda de la MTF para cada parametrización, junto con el valor $\mathcal{R}_{\mathrm{MTF}}(\Phi)$. (Derecha) Promedio por longitud de onda de la distribución de la magnitud de los valores propios de la matriz de convolución, acompañado del número de condición $\mathcal{K}(\mathbf{C}_\Phi)$ para cada configuración.



Las parametrizaciones propuestas, tanto Libre como Radial, logran una atenuación más efectiva de las frecuencias, como lo refleja el menor valor de $\mathcal{R}_{\mathrm{MTF}}(\Phi)$. En particular, estas parametrizaciones tienden a mitigar las frecuencias de manera equilibrada en todas las direcciones. En contraste, la parametrización de Zernike preserva las bajas frecuencias, comportándose de forma similar a un lente sin características de privacidad. Además, la distribución de los valores propios revela que el número de condición para las parametrizaciones propuestas es significativamente mayor que el de la parametrización de Zernike, lo que confirma su superior capacidad para preservar la privacidad visual.

3.6. ATAQUES DE INVERTIBILIDAD

En los enfoques que preservan la privacidad, es fundamental evaluar mediante ataques adversarios cómo se desempeñan los sistemas ópticos ante la posibilidad de que un atacante, interesado en obtener información sensible de los pacientes en una escena, intente resolver el problema inverso descrito en la ecuación (7). Para

esto se plantearon dos enfoques que intentan resolverlo. En el primer enfoque, se asume que el atacante tiene acceso físico al sistema óptico. Esto le permite obtener la PSF del sistema y utilizar el algoritmo de ***Wiener***[99] para invertirla y reconstruir la imagen original. El segundo enfoque considera que el atacante no dispone de acceso directo al sistema óptico, pero sí cuenta con un conjunto de datos etiquetados de imágenes privadas y no privadas, además de los recursos computacionales necesarios para entrenar un algoritmo basado en aprendizaje profundo. En este segundo escenario, implementamos el algoritmo ***Restormer***[100], el cual está compuesto por bloques jerárquicos de la arquitectura *Transformer* que operan a múltiples escalas de la imagen, permitiendo así una reconstrucción de alta calidad. El entrenamiento de la arquitectura Restormer se realiza de manera independiente para cada una de las parametrizaciones, utilizando el conjunto de datos DDPD, lo que resulta en tiempos de entrenamiento de 24 horas por cada parametrización.

La Figura 18 muestra la distribución de puntos de las métricas de error SSIM y PSNR para cada algoritmo. El eje horizontal representa el valor de la métrica sobre la medida privada adquirida en el sensor (PSNR-S y SSIM-S), mientras que el eje vertical muestra el valor de las métricas después de aplicar los algoritmos de deconvolución (PSNR-D y SSIM-D).

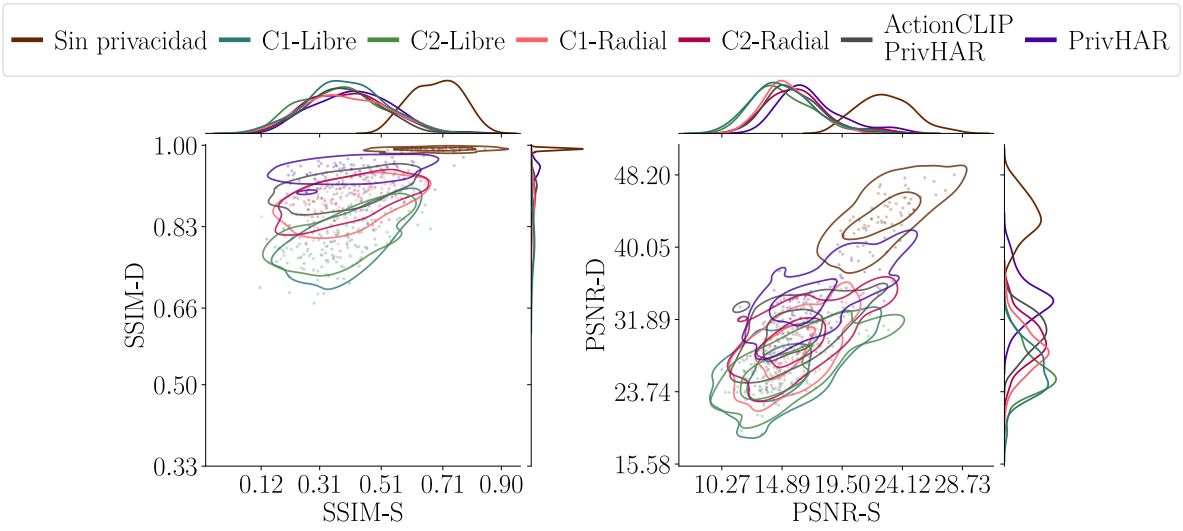
Es importante notar que los valores de las métricas sobre el sensor (SSIM-S y PSNR-S) presentan una distribución similar entre las diferentes parametrizaciones. Sin embargo, después de aplicar los algoritmos de deconvolución, se observa una reducción en las métricas (SSIM-D y PSNR-D) debido a la capacidad de las parame-

---

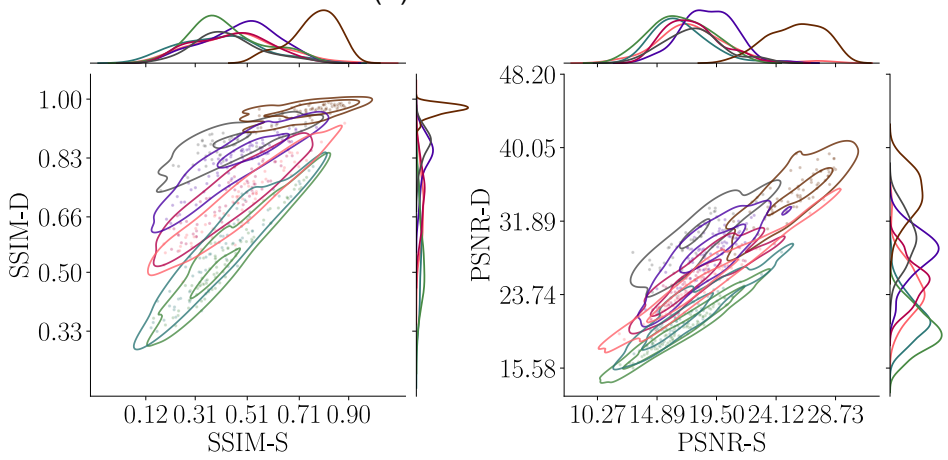
[99] François Orieux, Jean-François Giovannelli y Thomas Rodet. «Bayesian estimation of regularization and point spread function parameters for Wiener–Hunt deconvolution». En: *J. Opt. Soc. Am. A* 27.7 (2010), págs. 1593-1607. DOI: 10.1364/JOSAA.27.001593.

[100] Syed Waqas Zamir et al. «Restormer: Efficient Transformer for High-Resolution Image Restoration». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, págs. 5728-5739.

Figura 18. Distribución de puntos de las métricas SSIM y PSNR para las imágenes adquiridas en el sensor y sus reconstrucciones mediante algoritmos de deconvolución. El eje horizontal muestra los valores de la imagen adquirida en el sensor (S) y el eje vertical los de la reconstrucción (D). La mejor parametrización es la más cercana a la esquina inferior izquierda, indicando mayor preservación de la privacidad.



(a) Restormer



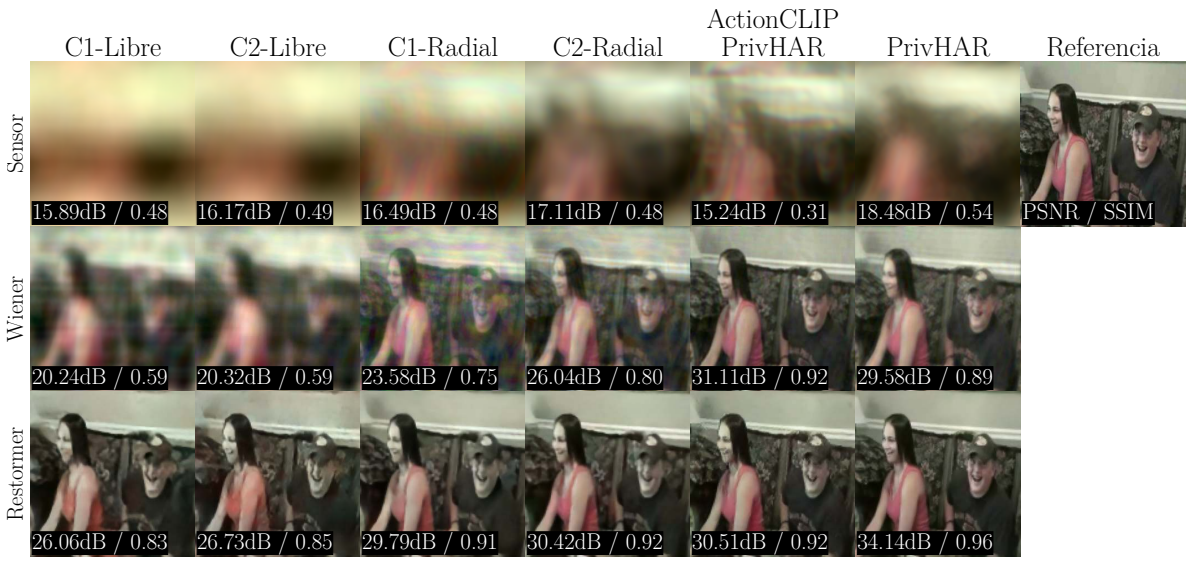
(b) Wiener

trizaciones propuestas para preservar la privacidad visual de mejor manera. Esto se relaciona con la Figura 7, donde se ilustra que las métricas de error sobre el sensor no necesariamente indican una mejor preservación de la privacidad visual.

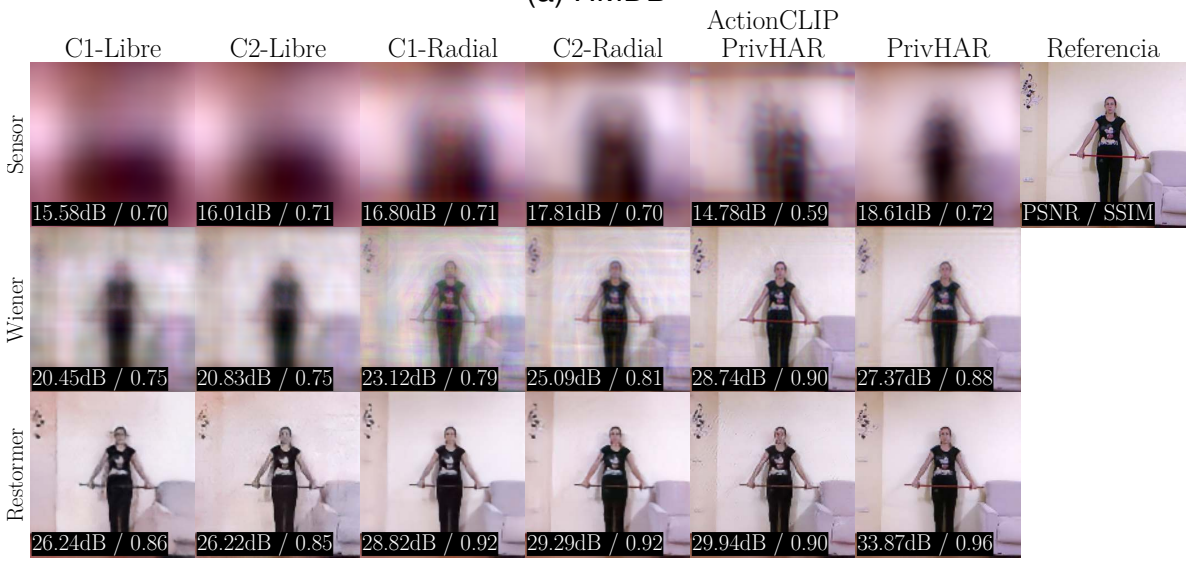
La mejor parametrización óptica para preservar la privacidad es aquella cuya distribución de puntos se acerca más a la esquina inferior izquierda del gráfico, donde se minimizan ambas métricas, indicando una reconstrucción de menor calidad. En este sentido, la parametrización Libre es la más eficaz para impedir la invertibilidad de las adquisiciones, ya que logra valores más bajos de SSIM-D y PSNR-D. Esto significa que, incluso tras aplicar técnicas avanzadas de reconstrucción, la calidad de la imagen recuperada es insuficiente para extraer información sensible. Por el contrario, la parametrización de *Zernike* ofrece una menor resistencia a la reconstrucción, evidenciando que es más susceptible a ser invertida por un atacante.

La Figura 19 presenta los resultados visuales de los ataques de invertibilidad aplicados a dos imágenes seleccionadas de los conjuntos de datos HMDB y HPTE. Las filas muestran las imágenes adquiridas en el sensor y las reconstrucciones obtenidas mediante los algoritmos Wiener y Restormer. Las parametrizaciones Libres generan mayor degradación de la imagen, reflejando valores más bajos de PSNR y SSIM, lo que evidencia su efectividad para preservar la privacidad al dificultar la reconstrucción, incluso utilizando algoritmos avanzados. Sin embargo, la parametrización de Zernike no logra preservar adecuadamente la privacidad en ninguno de los métodos. En contraste, la parametrización Radial es efectiva para evitar la reconstrucción con el algoritmo de Wiener, pero se vuelve vulnerable frente a Restormer, obteniendo resultados comparables a los de Zernike. En conjunto, los resultados confirman que las parametrizaciones Libres son las más eficaces para preservar la privacidad, ya que evitan reconstrucciones fieles incluso ante ataques sofisticados como los realizados con Restormer.

Figura 19. Resultados al aplicar los algoritmos de deconvolución de Wiener y Restormer sobre imágenes privadas capturadas con cada una de las parametrizaciones ópticas propuestas, utilizando dos conjuntos de datos diferentes.



(a) HMDB



(b) HPTE

4. CONCLUSIONES Y TRABAJO FUTURO

En este trabajo se presentó un enfoque para la optimización de sistemas ópticos difractivos mediante algoritmos de aprendizaje profundo, orientado a preservar la privacidad en la adquisición de imágenes y facilitar la estimación de acciones en entornos clínicos. Se estudiaron diferentes parametrizaciones de elementos ópticos difractivos y configuraciones de sistemas ópticos, con el propósito de inducir distorsiones ópticas que dificultan la reconstrucción de la información sensible, mientras mantienen el desempeño en la tarea de estimación de acciones. Adicionalmente, se propusieron dos funciones de regularización, $\mathcal{R}_{\mathrm{MTF}}(\Phi)$ y $\mathcal{R}_{\mathrm{cond}}(\Phi)$ para promover la privacidad durante el proceso de optimización, minimizando la MTF y maximizando el número de condición del sistema. Estas regularizaciones permiten la estabilidad numérica durante el entrenamiento así como un incremento en la preservación de la privacidad del método.

Adicionalmente, se propuso el empleo del paradigma multimodal para la estimación de acciones, lo que facilita la generalización del modelo en escenarios con datos limitados, como los evaluados mediante la técnica de aprendizaje *zero-shot*. Las evaluaciones realizadas con métricas como el número de condición, FWHM y SSIM indicaron que la parametrización Radial logró un equilibrio adecuado entre la preservación de la privacidad y la precisión en la estimación de acciones.

Los resultados obtenidos muestran que el diseño y optimización conjunta del sistema óptico y los modelos de aprendizaje profundo permiten mejorar tanto la privacidad visual como el rendimiento en la tarea de estimación de acciones. Este enfoque proporciona un marco de trabajo que puede ser extendido para el desarrollo de sistemas seguros y eficientes en aplicaciones clínicas, sin comprometer la funcionalidad ni la privacidad de los usuarios.

BIBLIOGRAFÍA

Agrawal, Amit y Yi Xu. «Coded exposure deblurring: Optimized codes for PSF estimation and invertibility». En: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, págs. 2066-2073 (vid. pág. 38).

Ahouandjinou, ASRM, C Motamed y EC Ezin. «A temporal belief-based hidden markov model for human action recognition in medical videos». En: *Pattern Recognition and Image Analysis* 25.3 (2015), págs. 389-401 (vid. pág. 40).

Anwar, Saeed, Salman Khan y Nick Barnes. «A Deep Journey into Super-resolution: A Survey». En: *ACM Comput. Surv.* 53.3 (2020) (vid. pág. 27).

Ar, Ilktan y Yusuf Sinan Akgul. «A Computerized Recognition System for the Home-Based Physiotherapy Exercises Using an RGBD Camera». En: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22.6 (2014), págs. 1160-1171 (vid. págs. 61, 66).

Arce, Gonzalo R. et al. «Compressive Coded Aperture Spectral Imaging: An Introduction». En: *IEEE Signal Processing Magazine* 31.1 (2014), págs. 105-115. DOI: 10.1109/MSP.2013.2278763 (vid. pág. 35).

Arguello, Henry y Miguel Marquez. «Convex Optimization for Image Reconstruction». En: *Coded Optical Imaging*. Springer, 2024, págs. 37-53 (vid. págs. 38, 39).

Arguello, Henry et al. «Deep Optical Coding Design in Computational Imaging: A data-driven framework». En: *IEEE Signal Processing Magazine* 40.2 (2023), págs. 75-88 (vid. págs. 18, 35).

Arguello, Henry et al. «Shift-variant color-coded diffractive spectral imaging system». En: *Optica* 8.11 (2021), págs. 1424-1434 (vid. págs. 30, 35, 60).

Arguello, Paula et al. «Learning to Describe Scenes via Privacy-Aware Designed Optical Lens». En: *IEEE Transactions on Computational Imaging* 10 (2024), págs. 1069-1079. DOI: 10.1109/TCI.2024.3426975 (vid. págs. 18, 19, 45, 46, 53, 58).

Arguello, Paula et al. «Optics Lens Design for Privacy-Preserving Scene Captioning». En: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, págs. 3551-3555 (vid. págs. 18, 25, 28, 45, 46, 53, 58).

Baek, Seung-Hwan et al. «Single-shot hyperspectral-depth imaging with learned diffractive optics». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 2651-2660 (vid. págs. 30, 32).

Bamieh, Bassam. «Discovering transforms: A tutorial on circulant matrices, circular convolution, and the discrete fourier transform». En: *arXiv preprint arXiv:1805.05533* (2018) (vid. pág. 52).

Binbin, Yu. «An improved infrared image processing method based on adaptive threshold denoising». En: *EURASIP Journal on Image and Video Processing* 2019.1 (2019), pág. 5 (vid. pág. 27).

Bommasani, Rishi et al. «On the opportunities and risks of foundation models». En: *arXiv preprint arXiv:2108.07258* (2021) (vid. págs. 15, 40).

Born, Max y Emil Wolf. *Principles of optics: Electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013 (vid. págs. 28, 36).

Bu, Fanyu et al. «Privacy preserving back-propagation based on BGV on cloud». En: *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*. IEEE. 2015, págs. 1791-1795 (vid. pág. 16).

Burgos, Ninon et al. «Deep learning for brain disorders: from data processing to disease treatment». En: *Briefings in Bioinformatics* 22.2 (2021), págs. 1560-1576 (vid. págs. 14, 40).

Chou, Edward et al. «Privacy-preserving action recognition for smart hospitals using low-resolution depth images». En: *arXiv preprint arXiv:1811.09950* (2018) (vid. pág. 40).

Cioppa, Anthony et al. «SoccerNet 2023 challenges results». En: *Sports Engineering* 27.2 (2024), pág. 24 (vid. pág. 25).

COLOMBIA, EL CONGRESO DE. *LEY ESTATUTARIA 1581 DE 2012*. Oct. de 2012 (vid. pág. 16).

DATTA, BISWA NATH. «CHAPTER 3 - SOME FUNDAMENTAL TOOLS AND CONCEPTS FROM NUMERICAL LINEAR ALGEBRA». En: *Numerical Methods for Linear Control Systems*. Ed. por BISWA NATH DATTA. San Diego: Academic Press, 2004, págs. 33-78 (vid. págs. 38, 39, 51).

De Fauw, Jeffrey et al. «Clinically applicable deep learning for diagnosis and referral in retinal disease». En: *Nature Medicine* 24.9 (2018), págs. 1342-1350 (vid. págs. 14, 40, 66).

Dosovitskiy, Alexey et al. «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale». En: *International Conference on Learning Representations*. 2021 (vid. pág. 55).

Dufraisse, Marius et al. «Physics Based Camera Privacy: Lens and Network Co-Design to the Rescue». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 1410-1419 (vid. págs. 45-47).

Dun, Xiong et al. «Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging». En: *Optica* 7.8 (2020), págs. 913-922 (vid. págs. 30, 50).

— «Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging». En: *Optica* 7.8 (2020), págs. 913-922 (vid. pág. 32).

Elad, Michael. «Optimized Projections for Compressed Sensing». En: *IEEE Transactions on Signal Processing* 55.12 (2007), págs. 5695-5702. DOI: 10.1109/TSP.2007.900760 (vid. pág. 34).

European Union, European Parliament Council of the. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*. Mayo de 2016 (vid. pág. 16).

Fan, Linxi et al. «Rubiksnet: Learnable 3d-shift for efficient video action recognition». En: *European Conference on Computer Vision*. Springer. 2020, págs. 505-521 (vid. págs. 15, 40).

Fan, Zunlin et al. «Dim infrared image enhancement based on convolutional neural network». En: *Neurocomputing* 272 (2018), págs. 396-404 (vid. pág. 27).

Feichtenhofer, Christoph, Axel Pinz y Richard P. Wildes. «Spatiotemporal residual networks for video action recognition». En: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Curran Associates Inc., 2016, 3476–3484 (vid. págs. 15, 40).

Forbes, K. y V. V. Anh. «Condition of system matrices in image restoration». En: *J. Opt. Soc. Am. A* 11.6 (1994), págs. 1727-1735 (vid. págs. 38, 39).

Frome, Andrea et al. «Large-scale privacy protection in Google Street View». En: *2009 IEEE 12th International Conference on Computer Vision*. 2009, págs. 2373-2380 (vid. pág. 26).

Goodman, Joseph W. «Introduction to Fourier Optics, Roberts & Co». En: *Publishers, Englewood, Colorado* (2005) (vid. pág. 28).

Gray, Robert M et al. «Toeplitz and circulant matrices: A review». En: *Foundations and Trends® in Communications and Information Theory* 2.3 (2006), págs. 155-239 (vid. pág. 52).

Greengard, Adam, Yoav Y. Schechner y Rafael Piestun. «Depth from diffracted rotation». En: *Opt. Lett.* 31.2 (2006), págs. 181-183 (vid. pág. 33).

Grigorescu, Sorin et al. «A survey of deep learning techniques for autonomous driving». En: *Journal of Field Robotics* 37.3 (2020), págs. 362-386 (vid. pág. 14).

Gross, Ralph et al. «Integrating utility into face de-identification». En: *International Workshop on Privacy Enhancing Technologies*. Springer. 2005, págs. 227-242 (vid. pág. 27).

Guayacán, Luis C, Brayan Valenzuela y Fabio Martinez. «Parkinsonian gait characterization from regional kinematic trajectories». En: *14th International Symposium on Medical Information Processing and Analysis*. Vol. 10975. International Society for Optics y Photonics. 2018, pág. 1097502 (vid. pág. 15).

Hinojosa, Carlos, Juan Carlos Niebles y Henry Arguello. «Learning Privacy-preserving Optics for Human Pose Estimation». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, págs. 2573-2582 (vid. págs. 18, 19, 25, 28, 45, 46, 58).

Hinojosa, Carlos et al. «PrivHAR: Recognizing Human Actions from Privacy-Preserving Lens». En: *Computer Vision – ECCV 2022*. Ed. por Shai Avidan et al. Cham: Springer Nature Switzerland, 2022, págs. 314-332 (vid. págs. 18, 19, 25, 28, 45, 47, 58, 62).

Hwang, Hyundeok et al. «MTF assessment of high resolution satellite images using ISO 12233 slanted-edge method». En: *Image and Signal Processing for Remote Sensing XIV*. Vol. 7109. SPIE. 2008, págs. 34-42 (vid. pág. 37).

Ikoma, Hayato et al. «Depth from defocus with learned optics for imaging and occlusion-aware depth estimation». En: *2021 IEEE International Conference on Computational Photography (ICCP)*. IEEE. 2021, págs. 1-12 (vid. pág. 35).

Jeon, Daniel S. et al. «Compact Snapshot Hyperspectral Imaging with Diffracted Rotation». En: *ACM Trans. Graph.* 38.4 (2019) (vid. págs. 30, 32, 35).

Kaissis, Georgios et al. «End-to-end privacy preserving deep learning on multi-institutional medical imaging». En: *Nature Machine Intelligence* 3.6 (2021), págs. 473-484 (vid. pág. 17).

Kamilaris, Andreas y Francesc X Prenafeta-Boldú. «Deep learning in agriculture: A survey». En: *Computers and Electronics in Agriculture* 147 (2018), págs. 70-90 (vid. pág. 14).

Keceli, Ali Seydi y Ahmet Burak Can. «Recognition of basic human actions using depth information». En: *International Journal of Pattern Recognition and Artificial Intelligence* 28.02 (2014) (vid. pág. 40).

Kim, Jong Wook, Beakcheol Jang y Hoon Yoo. «Privacy-preserving aggregation of personal health data streams». En: *PloS one* 13.11 (2018), e0207639 (vid. pág. 26).

Kim, Sung Hyun et al. «Animal Infectious Diseases Prevention through Big Data and Deep Learning». En: *Journal of Intelligence and Information Systems* 24.4 (2018), págs. 137-154 (vid. págs. 14, 40).

Kingma, Diederik P. «Adam: A method for stochastic optimization». En: *arXiv preprint arXiv:1412.6980* (2014) (vid. págs. 35, 56).

Kuehne, H. et al. «HMDB: a large video database for human motion recognition». En: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2011 (vid. pág. 61).

Lakshminarayanan, Vasudevan y Andre Fleck. «Zernike polynomials: a guide». En: *Journal of Modern Optics* 58.7 (2011), págs. 545-561 (vid. págs. 18, 45).

Lara, Teodoro. «Matrices Circulantes». En: *Divulgaciones Matemáticas* 9.1 (2001), págs. 85-102 (vid. pág. 52).

Li, Wenbo et al. «Adaptive RNN Tree for Large-Scale Human Action Recognition». En: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017 (vid. págs. 15, 40).

Liu, Jingdan et al. «Swept coded aperture real-time femtophotography». En: *Nature Communications* 15.1 (2024), pág. 1589 (vid. pág. 37).

Liu, Xin et al. «Investigating deep optics model representation in affecting resolved all-in-focus image quality and depth estimation fidelity». En: *Opt. Express* 30.20 (2022), págs. 36973-36984 (vid. págs. 19, 49).

Loshchilov, Ilya y Frank Hutter. «SGDR: Stochastic Gradient Descent with Warm Restarts». En: *International Conference on Learning Representations*. 2022 (vid. pág. 59).

Mahapatra, Dwarikanath, Behzad Bozorgtabar y Zongyuan Ge. «Medical Image Classification Using Generalized Zero Shot Learning». En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2021, págs. 3344-3353 (vid. págs. 44, 66).

Martínez, Fabio, Francisco Gómez y Eduardo Romero. «Análisis de vídeo para estimación del movimiento humano: una revisión». En: *Revista Med* 17.1 (2009), págs. 95-106 (vid. pág. 15).

Mejia, Yuri y Henry Arguello. «Binary Codification Design for Compressive Imaging by Uniform Sensing». En: *IEEE Transactions on Image Processing* 27.12 (2018), págs. 5775-5786. DOI: 10.1109/TIP.2018.2857445 (vid. pág. 34).

Narayanan, Arvind y Vitaly Shmatikov. «Robust de-anonymization of large sparse datasets». En: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, págs. 111-125 (vid. pág. 16).

Nehme, Elias et al. «DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning». En: *Nature Methods* 17.7 (2020), págs. 734-740 (vid. págs. 33, 49).

Ngiam, Jiquan et al. «Multimodal deep learning». En: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, págs. 689-696 (vid. pág. 41).

Nguyen, Nhat, Peyman Milanfar y Gene Golub. «A computationally efficient superresolution image reconstruction algorithm». En: *IEEE Transactions on Image Processing* 10.4 (2001), págs. 573-583 (vid. pág. 27).

Noll, Robert J. «Zernike polynomials and atmospheric turbulence∗». En: *J. Opt. Soc. Am.* 66.3 (1976), págs. 207-211 (vid. págs. 18, 45).

Oppenheim, Alan V. *Discrete-time signal processing*. Pearson Education India, 1999 (vid. pág. 52).

Orieux, François, Jean-François Giovannelli y Thomas Rodet. «Bayesian estimation of regularization and point spread function parameters for Wiener–Hunt deconvolution». En: *J. Opt. Soc. Am. A* 27.7 (2010), págs. 1593-1607. DOI: 10.1364/JOSAA.27.001593 (vid. pág. 73).

Padilla-López, José Ramón, Alexandros Andre Chaaraoui y Francisco Flórez-Revuelta. «Visual privacy protection methods: A survey». En: *Expert Systems with Applications* 42.9 (2015), págs. 4177-4195 (vid. pág. 25).

Pan, Yuchen et al. «OpticalDR: A Deep Optical Imaging Model for Privacy-Protective Depression Recognition». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, págs. 1303-1312 (vid. págs. 45, 46).

Parulski, Ken et al. «Creation and evolution of ISO 12233, the international standard for measuring digital camera resolution». En: *Electronic Imaging* 34 (2022), págs. 1-7 (vid. pág. 37).

Paszke, Adam et al. «PyTorch: An Imperative Style, High-Performance Deep Learning Library». En: *Advances in Neural Information Processing Systems*. Ed. por H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019 (vid. págs. 36, 53, 58).

Piccialli, Francesco et al. «A survey on deep learning in medicine: Why, how and when?» En: *Information Fusion* 66 (2021), págs. 111-137 (vid. págs. 14, 25).

Pittaluga, Francesco y Sanjeev Jagannatha Koppal. «Pre-capture privacy for small vision sensors». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.11 (2016), págs. 2215-2226 (vid. pág. 17).

— «Pre-Capture Privacy for Small Vision Sensors». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.11 (2017), págs. 2215-2226 (vid. pág. 27).

Qing, Zhiwu et al. «Mar: Masked autoencoders for efficient action recognition». En: *IEEE Transactions on Multimedia* (2023) (vid. págs. 15, 40).

Radford, Alec et al. «Learning transferable visual models from natural language supervision». En: *International Conference on Machine Learning*. PMLR. 2021, págs. 8748-8763 (vid. págs. 54, 55).

Romera-Paredes, Bernardino y Philip Torr. «An embarrassingly simple approach to zero-shot learning». En: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. por Francis Bach y David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, págs. 2152-2161 (vid. pág. 44).

Ryoo, Michael S et al. «Privacy-preserving human activity recognition from extreme low resolution». En: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017 (vid. págs. 18, 25, 27).

Sakellaropoulos, Theodore et al. «A deep learning framework for predicting response to therapy in cancer». En: *Cell Reports* 29.11 (2019), págs. 3367-3373 (vid. págs. 14, 40).

Santamarina, J Carlos y Dante Fratta. *Discrete signals and inverse problems: an introduction for engineers and scientists*. John Wiley & Sons, 2005 (vid. pág. 39).

Schwarz, Christopher G et al. «Identification of anonymous MRI research participants with face-recognition software». En: *New England Journal of Medicine* 381.17 (2019), págs. 1684-1686 (vid. pág. 16).

Shechtman, Yoav et al. «Precise Three-Dimensional Scan-Free Multiple-Particle Tracking over Large Axial Ranges with Tetrapod Point Spread Functions». En: *Nano Letters* 15.6 (2015). PMID: 25939423, págs. 4194-4199. DOI: 10.1021/

acs.nanolett.5b01396. eprint: https://doi.org/10.1021/acs.nanolett.5b01396 (vid. pág. 33).

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)* 37.4 (2018), págs. 1-13 (vid. págs. 32, 35).

Smith, Warren J. *Modern optical engineering: the design of optical systems*. SPIE Press, 2008 (vid. págs. 29, 36).

Sreenu, GSDMA y Saleem Durai. «Intelligent video surveillance: a review through deep learning techniques for crowd analysis». En: *Journal of Big Data* 6.1 (2019), págs. 1-27 (vid. pág. 25).

Tan, Jasper et al. «CANOPIC: pre-digital privacy-enhancing encodings for computer vision». En: *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2020, págs. 1-6 (vid. pág. 26).

Tasneem, Zaid et al. «Learning phase mask for privacy-preserving passive depth estimation». En: *European Conference on Computer Vision*. Springer. 2022, págs. 504-521 (vid. pág. 45).

Tran, Du et al. «Learning spatiotemporal features with 3d convolutional networks». En: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, págs. 4489-4497 (vid. págs. 15, 40).

Vaswani, A. «Attention is all you need». En: *Advances in Neural Information Processing Systems* (2017) (vid. pág. 55).

Vijayakumar, Anand y Shanti Bhattacharya. *Design and fabrication of diffractive optical elements with MATLAB*. SPIE, 2017 (vid. pág. 28).

Wang, Mengmeng et al. «Actionclip: Adapting language-image pretrained models for video action recognition». En: *IEEE Transactions on Neural Networks and Learning Systems* (2023) (vid. págs. 41, 54, 56).

Wu, Wenhao, Zhun Sun y Wanli Ouyang. «Revisiting Classifier: Transferring Vision-Language Models for Video Recognition». En: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.3 (2023), págs. 2847-2855 (vid. págs. 41, 54-56, 59).

Xian, Yongqin, Bernt Schiele y Zeynep Akata. «Zero-shot learning-the good, the bad and the ugly». En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, págs. 4582-4591 (vid. pág. 44).

Xu, Yiwen et al. «Deep learning predicts lung cancer treatment response from serial medical imaging». En: *Clinical Cancer Research* 25.11 (2019), págs. 3266-3275 (vid. págs. 14, 40).

Yuan, Lu et al. «Image deblurring with blurred/noisy image pairs». En: *ACM SIGGRAPH 2007 Papers*. SIGGRAPH '07. Association for Computing Machinery, 2007, 1–es (vid. pág. 27).

Zamir, Syed Waqas et al. «Restormer: Efficient Transformer for High-Resolution Image Restoration». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, págs. 5728-5739 (vid. pág. 73).

Zeng, Wenjun y S. Lei. «Efficient frequency domain selective scrambling of digital video». En: *IEEE Transactions on Multimedia* 5.1 (2003), págs. 118-129 (vid. pág. 26).

Zhang, Hong-Bo et al. «A comprehensive survey of vision-based human action recognition methods». En: *Sensors* 19.5 (2019), pág. 1005 (vid. págs. 39, 40).

Zhang, Kaihao et al. «Deep image deblurring: A survey». En: *International Journal of Computer Vision* 130.9 (2022), págs. 2103-2130 (vid. pág. 27).

Zhao, Yanan et al. «Deep, Convergent, Unrolled Half-Quadratic Splitting for Image Deconvolution». En: *IEEE Transactions on Computational Imaging* 10 (2024), págs. 574-588. DOI: 10.1109/TCI.2024.3377132 (vid. pág. 39).