Learning to Describe Scenes via Privacy-aware Optical Lens

Paula Arguello[1], Jhon Lopez[1], Karen Sanchez[2], Carlos Hinojosa[2], Hoover Rueda-Chacón[1], Henry Arguello[1]
[1]Universidad Industrial de Santander, [2]King Abdullah University of Science and Technology

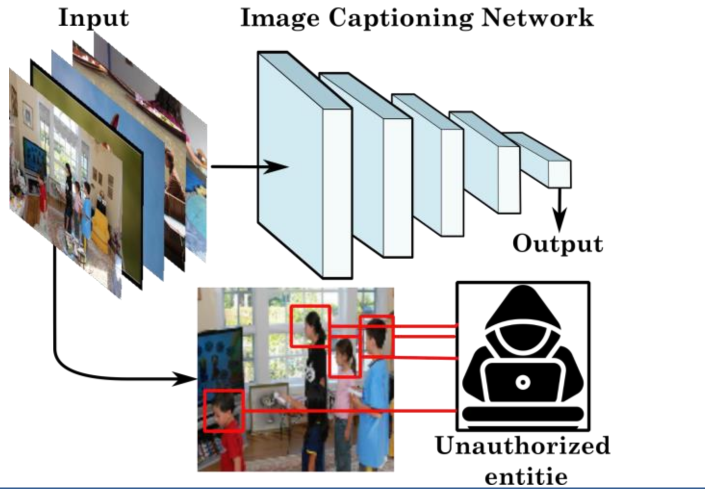
High Dimensional Signal Processing Research Group
Universidad Industrial de Santander
KAUST
LXAI
CVPR JUNE 17-21, 2024 SEATTLE, WA

Introduction

Image captioning: Create short informative texts for images, using natural language, that relates the visual content and context of an image.

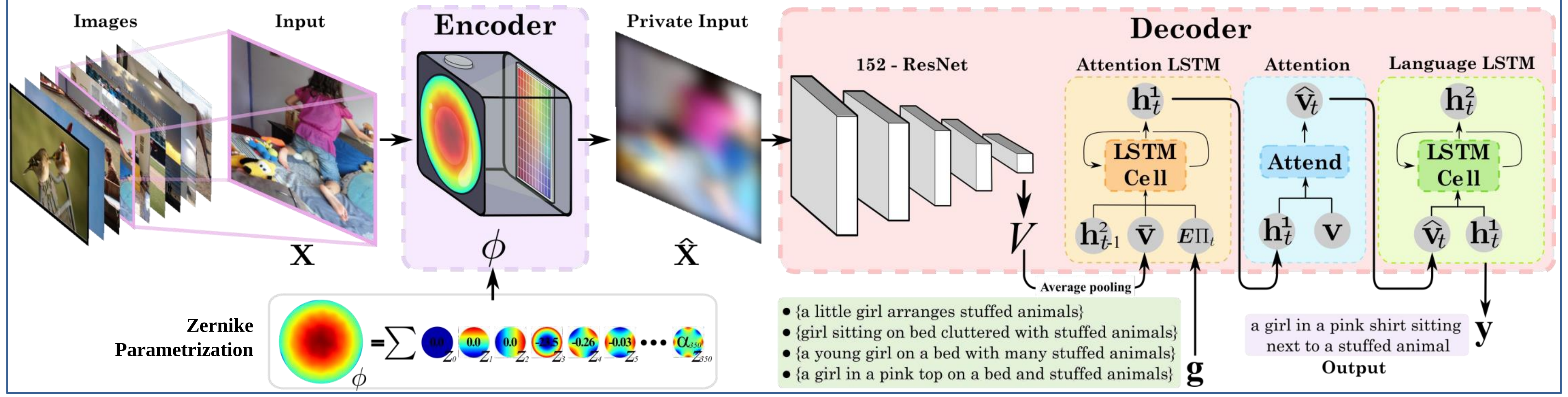
a small girl sitting on a chair holding a white bear — a man helps a disabled baseball player on the mound

However, the acquired images may contain **privacy-sensitive** data



Proposed Method

We propose a Encoder-Decoder end-to-end arquitecture [1] to learn optics by backpropagating the gradients from the captioning network (**decoder**) to the optics layer (**encoder**).



Optical Encoder

Assuming spatially incoherent light, we formulate the wave-based image formation model following Fourier optics and define the point spread function (**PSF**) [2]:

$$H_\lambda(x',y') = |\mathcal{F}^{-1}\{\mathcal{F}\{A(x,y)t_\phi(x,y)t_l(x,y)U_\lambda(x,y)\}T_{d_2}(f_x,f_y)\}|^2$$

and the phase modulation represented by:

$$t_\phi(x,y) = e^{j\frac{2\pi}{\lambda}\phi(x,y)}$$

obtained from the lens surface profile:

$$\phi = \sum_{j=1}^{q} \alpha_j Z_j$$

where each **Zernike polynomial** represents a specific wavefront aberration, creating a linear combination. Combining these **aberrations** forms the resulting optical lens surface profile.

Finally, the acquired images for the RGB channels can be modeled as:

$$\hat{\mathbf{X}}_\ell = \mathcal{S}_\ell(\mathbf{H}_\lambda * \mathbf{X}_\ell) + \mathbf{N}_\ell$$

Loss Function

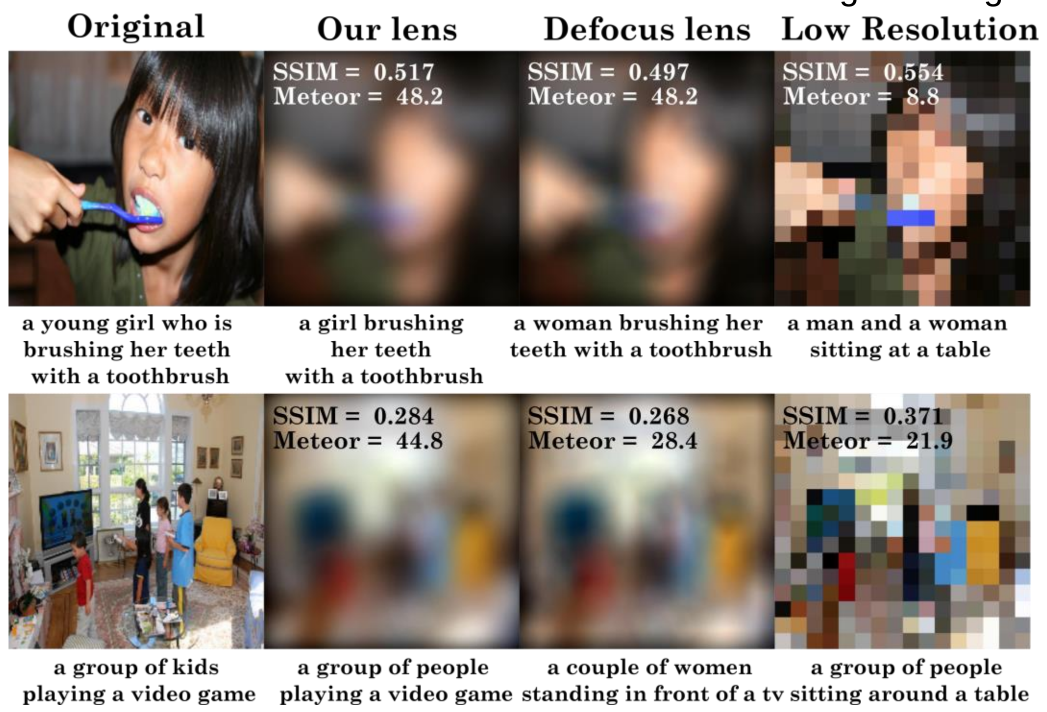
Our loss function combines multiple terms to increase **optical distortion** and preserve performance in **word generation**:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_{ce} + \mathcal{L}_d + \mathcal{L}_{\mathbf{H}}.$$

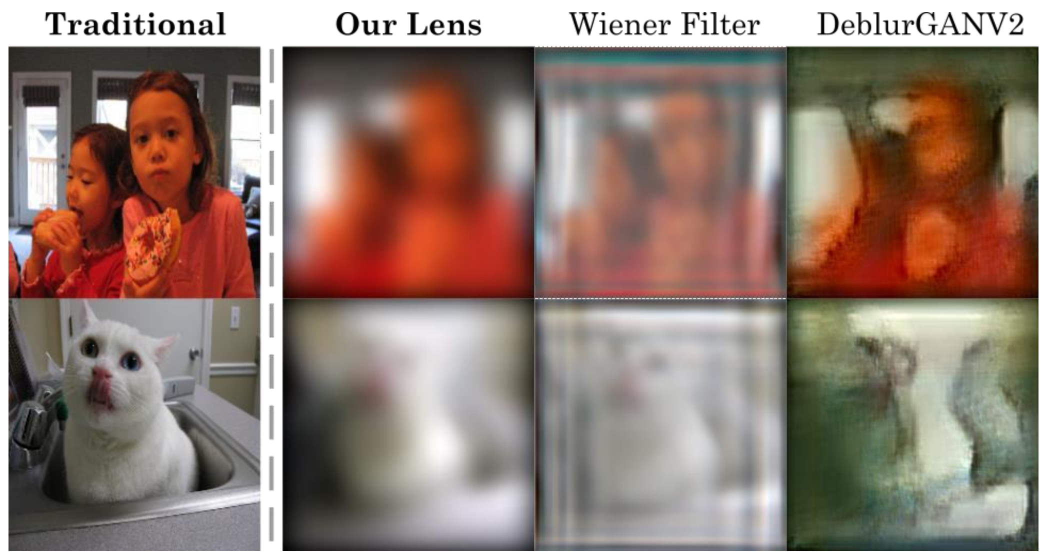
1. Promote distortion by maximizing the difference between the images: $\mathcal{L}_p = 1 - \|\hat{\mathbf{X}} - \mathbf{X}\|_2^2$
2. Multi-class cross-entropy, to guide the learning of the correct sequence of words for IC.
$$\mathcal{L}_{ce} = \sum_{c=1}^{C} \log\left(\frac{\exp(\mathbf{y}_c)}{\exp(\sum_{i=1}^{C}\mathbf{y}_i)}\right)\mathbf{g}_c.$$
3. Double regularization to attend every part of the distorted image
$$\mathcal{L}_d = -\log(p(\mathbf{y} \mid \mathbf{a})) + \lambda \sum_i^L \left(1 - \sum_t^C \boldsymbol{\theta}_{ti}\right)^2.$$
4. Regularization on the **PSF** promoting a centering on camera sensor
$$\mathcal{L}_H = \|(\mathbf{H}_\lambda * \mathbf{M}) - \mathbf{H}_\lambda\|_F$$
$$\mathbf{M}_{ij} = \begin{cases} 1, & \text{if } (i-p)^2 + (j-p)^2 \le r^2 \\ 0, & \text{otherwise.} \end{cases}$$

Qualitative Results

Qualitative results on two test set samples. Insets display the **SSIM** and **Meteor** between the distorted and original images.



Evaluation of the robustness of our lens-protected images against **deconvolution attacks**. Qualitative results show that the identities of individuals cannot be recovered after applying non-blind (**Wiener**) and blind (**DeblurGANv2**) deconvolution.



References & Contact

[1] P. Arguello, J. Lopez, C. Hinojosa, and H. Arguello, "Optics lens design for privacy-preserving scene captioning," *in ICIP Conf., 2022.*
[2] V. Sitzmann, S. Diamond, Y. Peng, X. Dun, S. Boyd, W. Heidrich, F. Heide, and G. Wetzstein, "End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging," *ACM, no. 4, 2018.*

Project Webpage
henarfu@uis.edu.co
http://hdspgroup.com/