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
{paula2191444@correo, jhon2208456@correo, karen.sanchez2@saber}uis.edu.co,
carlos.hinojosamontero@kaust.edu.sa, {hfarueda@, henarfu@}uis.edu.co

Abstract

Image caption generation textually summarizes the visual content of an image. This task has gained popularity at the turning point of computer vision (CV) and natural language processing (NLP). Images used to train image captioning models may contain sensitive data that should be confidential, such as faces, personal characteristics, documents, children, etc. This work focuses on protecting privacy in the image captioning task, directly from the image acquisition stage. For this, a refractive lens was designed to ensure privacy using an end-to-end deep learning-based optimization approach. The designed lens blurs sensitive visual attributes in the acquired image while extracting essential features to generate captions even from highly distorted images. The image caption network implements two long short-term memory networks (LSTMs) with an attention module in between to ensure high-quality captions. This method was tested and validated through simulations in the COCO dataset. The results showed a better balance between privacy and usability compared to traditional methods that do not consider privacy.

1. Introduction

Image captioning is the process of creating short informative texts for images, using natural language, that relates the visual content and context of an image. This process facilitates image search and simplifies content summarizing, it empowers virtual assistants and artificial intelligence systems by enriching educational materials [6] and improving communication in social networks [14, 22]. The image captioning task is considered significantly more complex than image classification or object recognition [26], since it requires describing not only the objects present, but also their relationship, attributes, and associated actions or activities.

Image captioning has been addressed using various techniques including convolutional neural networks (CNNs) to extract features from images, jointly with Long Short-Term Memory (LSTM) Networks capable of processing entire sequences to generate word-by-word captions [1]. This approach has substantially improved the contextual relevance and consistency of the generated captions [25, 28]. In addition, attention mechanisms have been integrated into these models, allowing them to focus on specific regions of the image [9]. The latter improves the quality of the captions by making them more closely linked to the visual content.

In most computer vision (CV) tasks, images are used to train deep neural networks (DNNs). However, available images may contain sensitive information, raising concerns about the associated risks of improper exploitation [15]. This has motivated privacy protection methods in different fields, such as healthcare [24], multimedia [29], social networking [13], and video surveillance [4]. Advances in optics and algorithms [23] have led to the development of privacy-preserving end-to-end systems for applications such as human pose estimation [7] and action recognition [8].

To generate privacy-preserving image captions, state-of-the-art has reported an encryption framework[16]. Although this method is effective for privacy, the accuracy of captioning is not as good as non-privacy-preserving approaches. An alternative approach [20], focusing on dietary intake images, trains a DNN using images in which people's faces have been previously masked to avoid direct use by nutritionists. These approaches operate on *already* acquired images, and therefore may still present vulnerabilities between the acquisition and processing stages. We propose to address this problem through image acquisition.

There exist different strategies for privacy protection during image acquisition, including defocus techniques that provide privacy restricted by the size of the camera sensor [18, 19]. Similarly, [7, 8] developed a hardware-based privacy solution that involved the design of a refractive lens that selectively blurs sensitive information while preserving the functionality needed for gesture recognition and human pose estimation. Inspired by these works, we propose to design a specialized optical lens that integrates privacy

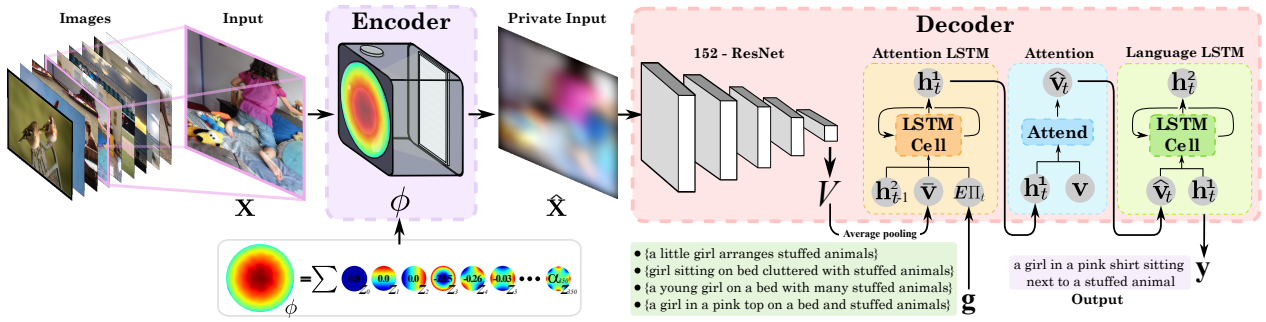

Figure 1. Proposed end-to-end model. The optical encoder incorporates a camera with a refractive lens, which is parametrized by a linear combination of Zernike polynomials. The decoder is formed by a convolutional feature extraction and an LSTM Network with attention, which produces a caption from the private image.

preservation into the generation of image captions. In particular, we rely on LSTMs for reliable image captioning and model the light propagation using Fourier optics, in an end-to-end framework. This method is expected to enable more effective privacy protection compared to existing methods, as will be demonstrated through rigorous testing and comparison with other techniques.

2. Proposed Method

Our proposed model comprises two components. First, an optical encoder consisting of a camera equipped with a refractive lens trained for privacy protection, where a linear combination of Zernike polynomials [17] is learned, as illustrated in Fig. 1, promoting distorted images as the response. This will obscure sensitive attributes such as personal objects, documents, and people's faces within the scene. The second component is a decoder module, which learns to generate image captions from distorted images. The decoder leverages a CNN to extract salient features, which are then processed by two LSTM networks, with an attention module in between. The two components work end-to-end, allowing the learning of optical parameters by back-propagating from the decoder to the optical layer.

2.1. Optical Encoder

Our encoder module is responsible for the image acquisition process, as shown in Fig. 1. As described earlier, our strategy to promote privacy modifies the camera lens by a learned refractive optical element. To facilitate the optimization of the camera lens, we employ a similar strategy from prior studies [7, 8, 23], which develop a differentiable module that takes into account the wave propagation and phase modulation processes within the camera.

Assuming spatially incoherent light, we formulate the wave-based image formation model following Fourier optics and define the point spread function (PSF) in terms of the lens surface profile parameters, which are learned. Harnessing the Fresnel approximation[5], we write the PSF as

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

where $T_\lambda(\cdot)$ denotes the transfer function involving the spatial frequencies (f_x, f_y) and the wavelength λ, the spatial coordinates on the camera plane are denoted by (x', y'), and on the lens by (x, y). $\mathcal{F}\{\cdot\}$ denotes the two-dimensional Fourier transform. W_λ is defined as $W_\lambda(x, y) = t_\phi(x, y)U_\lambda(x, y)$, with the wave field immediately preceding the lens as $U_\lambda(x, y)$ [23], and the phase modulation $t_\phi(\cdot)$ represented by $t_\phi(x, y) = e^{\mathrm{j}\frac{2\pi}{\lambda}\phi(x,y)}$, obtained from the lens surface profile $\phi = \sum_{j=1}^{q} \alpha_j Z_j$, [7, 8] where Z_j represents the j-th Zernike polynomial in Noll notation [17], and α_j is the corresponding coefficient. Each Zernike polynomial represents a specific wavefront aberration, and q denotes the total number of polynomials in the linear combination. Combining these aberrations forms the resulting surface profile, as shown at the bottom in Fig. 1.

Assuming the image formation is a shift-invariant convolution between the original image and the PSF, the acquired images for the RGB channels can be modeled as

$$\hat{\mathbf{X}}_\ell = \mathcal{S}_\ell(\mathbf{H}_\lambda * \mathbf{X}_\ell) + \mathbf{N}_\ell, \tag{2}$$

where the subscript $\ell \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$ indicates that operations are performed independently for each channel. The variable $\mathbf{X}_\ell \in \mathbb{R}+^{w\times h}$ represents the underlying scene with $w \times h$ pixels, and the matrix $\mathbf{H}_\lambda$ represents the discretized version of the PSF in Eq. (1). The term $\mathbf{N}_\ell \in \mathbb{R}^{w\times h}$ corresponds to the Gaussian noise present in the sensor. The function $\mathcal{S}_\ell(\cdot) : \mathbb{R}^{w\times h} \to \mathbb{R}^{w\times h}$ denotes the camera response function, and $*$ represents the 2D convolution. Note that in our optimization we want to learn the set of coefficients α_j.

2.2. Decoder

2.2.1 Feature extraction. To preserve privacy, we train our decoder to learn features directly from the distorted (private) images acquired with our camera. Specifically, we

employ the ResNet101 CNN to extract a set of L feature vectors $\mathcal{V}$, denoted as $\mathcal{V} = \{\boldsymbol{v}_i\}_{i=1}^{L}$, with $\boldsymbol{v}_i \in \mathbb{R}^D$.

2.2.2 Image Captioning Network. For the caption generation, we used a method [2] that combines an *attention* LSTM and a *language* LSTM. The attention LSTM uses visual features to focus on relevant areas of the image, and then, the language LSTM generates words sequentially.

Attention LSTM: The first LSTM network processes a single feature vector $\overline{\boldsymbol{v}} = \frac{1}{L}\sum_i \boldsymbol{v}_i$ representing the entire image. The input to the attention LSTM consists of the previous output of the language LSTM, $\boldsymbol{h}_{t-1}^2$, concatenated with $\overline{\boldsymbol{v}}$, and Π_t, a *one-hot* encoding of the input word at time step t. These inputs provide the maximum context regarding the state of the language LSTM, the overall content of the image, and the captions generated so far. The attention LSTM can be written as

$$\boldsymbol{h}_t^1 = \text{LSTM}([\boldsymbol{h}_{t-1}^2, \bar{\boldsymbol{v}}, \boldsymbol{E}\Pi_t], \boldsymbol{h}_{t-1}^1), \tag{3}$$

where superscript 1 denotes the first LSTM of the model. The matrix $\boldsymbol{E} \in \mathbb{R}^{m\times K}$ is a word embedding matrix for a vocabulary of size K.

Attention module: An attention module was used in-between the two LSTMs, to focus on the most relevant information within the images. The attention module receives $\boldsymbol{h}_t^1$ and $\boldsymbol{v}_i$, and performs $\alpha_t = \text{softmax}\,(a_t)$ with

$$a_{t,i} = \boldsymbol{w}_a^T \tanh\left(\boldsymbol{W}_{va}\boldsymbol{v}_i + \boldsymbol{W}_{ha}\boldsymbol{h}_t^1\right). \tag{4}$$

The attended image feature is used as input for the language LSTM, which is calculated as $\hat{\boldsymbol{v}}_t = \sum_{i=1}^{L} \alpha_{t,i}\boldsymbol{v}_i$.

Language LSTM: The language LSTM receives as input the attended image $\hat{\boldsymbol{v}}_t$ along with the output of the attention LSTM $\boldsymbol{h}_t^1$ as shown in Fig. 1, and operates as

$$\boldsymbol{h}_t^2 = \text{LSTM}([\hat{\boldsymbol{v}}, \boldsymbol{h}_t^1], \boldsymbol{h}_{t-1}^2), \tag{5}$$

where $\boldsymbol{h}_t^2$ corresponds to the language LSTM network denoted by the superscript 2. Finally, for each time step t, a conditional probability distribution is computed, applying the *softmax* function, for the current word given the history of previous words, as $p\left(y_t \mid y_{1:t-1}\right) = \text{softmax}\left(\boldsymbol{W}_p\boldsymbol{h}_t^2 + \boldsymbol{b}_p\right)$, where, $(y_1, ..., y_T)$ is a sequence of words, $\boldsymbol{W}_p \in \mathbb{R}^{K\times m}$ are the learned weights, and $\boldsymbol{b}_p \in \mathbb{R}^K$ the biases. The total probability of a complete sequence of words is calculated as the product of the conditional distributions, $p\left(y_{1:T}\right) = \prod_{t=1}^{T} p\left(y_t \mid y_{1:t-1}\right)$.

2.2.3 Loss Function. Our loss function combines four terms, $\mathcal{L}_p$, $\mathcal{L}_{ce}$, $\mathcal{L}_d$, and $\mathcal{L}_H$, chosen to increase optical distortion in the image acquisition, and to preserve the performance of word generation in image captioning. The term $\mathcal{L}_p$ represents the MSE between the original (undistorted) image $\mathbf{X}_\ell$ and the captured (distorted) sensor image $\hat{\mathbf{X}}_\ell$. This term aims to promote distortion by maximizing the difference between the two images via, $\mathcal{L}_p = 1 - \|\hat{\mathbf{X}} - \mathbf{X}\|_2^2$. $\mathcal{L}_{ce}$ represents the multi-class cross-entropy loss, used to guide the learning of the correct sequence of words for image captioning. It compares the predicted probabilities $\mathbf{y}$ to the ground truth caption $\mathbf{g}$ at each word c in the sequence of length C, following

$$\mathcal{L}_{ce} = \sum_{c=1}^{C} \log\left(\frac{\exp(\mathbf{y}_c)}{\exp(\sum_{i=1}^{C}\mathbf{y}_i)}\right)\mathbf{g}_c. \tag{6}$$

$\mathcal{L}_d$ incorporates a double regularization to encourage the model to attend every part of the distorted image, as in [28], given by

$$\mathcal{L}_d = -\log(p(\mathbf{y} \mid \mathbf{a})) + \lambda\sum_{i}^{L}\left(1 - \sum_{t}^{C}\boldsymbol{\theta}_{ti}\right)^2. \tag{7}$$

$\mathcal{L}_H$ performs a regularization on the PSF $\mathbf{H}_\lambda$, promoting the circular shape of the PSF by minimizing the values outside a circular mask $\mathbf{M}$. $\mathcal{L}_H$ is given by,

$$\mathcal{L}_H = \|(\mathbf{H}_\lambda * \mathbf{M}) - \mathbf{H}_\lambda\|_F, \tag{8}$$

where $\mathbf{M}$ is a binary matrix defined as,

$$\mathbf{M}_{ij} = \begin{cases} 1, & \text{if } (i-p)^2 + (j-p)^2 \leq r^2, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

with the elements inside the circle set to 1 and the elements outside set to 0. The purpose of $\mathbf{M}$ is to ensure a well-centered PSF on the camera sensor, where p represents the image center, while r denotes the expected PSF radius.

3. Experimental Results

We used the Common Objects in Context (COCO) 2014 dataset [12] for training, validation, and testing. To facilitate a meaningful comparison with our method, we incorporate concepts from existing privacy hardware-based approaches that involve the use of low-resolution cameras [21] and cameras equipped with defocusing lenses [18] to enhance visual privacy protection. Further, we conduct studies to demonstrate the robustness to deconvolution of the distorted images by attempting to recover the original images using two different models [11, 27].

3.1. Qualitative Results

We performed a comparative analysis with two different camera configurations. The first, referred to as the "defocus lens", shares the same optical architecture as our proposed method, but only optimizes the 4th Zernike polynomial, which induces defocus [17]. The second, so-called "low-resolution" camera, is a conventional camera equipped with a small sensor of 16×16 pixel dimensions. The results

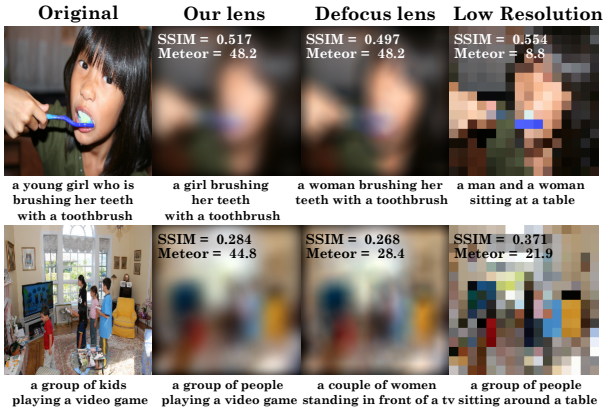

Figure 2. Qualitative results on two test set samples. Insets display the SSIM and Meteor between the distorted and original images.

Priv	Model	B-1	B-2	B-3	B-4	M	C
✗	BRNN [10]	64.2	45.1	30.3	20.1	19.5	66.6
	NIC [25]	66.6	46.1	32.9	24.6	23.7	-
	Hard Attn [28]	71.8	50.4	35.7	25.0	23.0	-
	2PSC-w [3]	72.1	54.8	40.4	29.6	**29.2**	89.2
	Proposed-w	**73.2**	**56.7**	**43.4**	**33.3**	29.0	**101.2**
✔	2PSC [3]	**68.9**	51.3	37.3	27.0	**28.1**	88.5
	Proposed	**68.9**	**51.7**	**38.5**	**29.0**	26.8	**89.0**
	Defocus	67.3	49.5	35.5	25.5	26.8	81.1
	Low-Res	61.6	42.7	29.1	19.9	23.3	58.8

Table 1. Bold results symbolize the best (highest), and underlined results symbolize the second-best, per dataset.

of these alternative approaches are presented in Figure 2, showing the original images along with their corresponding ground truth captions, as well as the privacy-preserving images and their respective captions. To quantitatively assess the level of degradation introduced by the different cameras compared to the original image, we calculated the structural similarity index measure (SSIM). The COCO test set reveals an average SSIM of 0.48 ± 0.12. It is crucial to note that, in all the approaches shown, the content of the images remains difficult to discern. Nonetheless, the proposed method is the one that achieves the most accurate caption of the scene, as expected, with the highest METEOR.

3.2. Quantitative Results

Table 1 presents the results in terms of the metrics *Bleu-1 (B-1), (B-2), (B-3), (B-4)*, *Meteor (M)* and *Cider (C)* respectively, which were used to evaluate the quality of captions in COCO dataset, the highest values being the best. The employed study includes a comparison between the implemented image captions model using original RGB images (denoted as "Proposed") and other image captions approaches using focused RGB images [3, 10, 25, 28]. This table also includes a quantitative evaluation of the privacy methods discussed above: defocusing and low-resolution cameras. In Table 1, the best results are indicated with boldface, while the second best results are underlined. As can

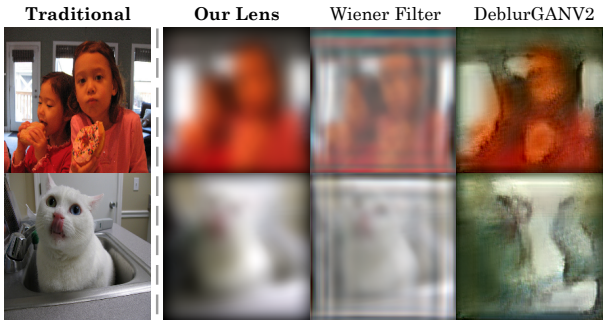

Figure 3. Evaluation of the robustness of our lens-protected images against deconvolution attacks. Qualitative results show that the identities of individuals cannot be recovered after applying non-blind (Wiener) and blind (DeblurGANv2) deconvolution.

be seen from the table, the proposed approach with privacy achieves the best balance between image distortion and caption accuracy, effectively balancing high efficiency with robust privacy protection.

Robustness to Deconvolution. To further validate our approach, we investigate how robust our lens is against deconvolution attacks. We consider both scenarios (blind and non-blind), where the attacker either knows the camera's PSF, thus able to use Wiener deconvolution, or an extensive collection of our private images, thus able to train a blind deconvolution network, e.g. DeblurGANv2 [11]. For the latter, we use 3,214 blurred images, partitioning into 2,103 for training and 1,111 for testing. Each distorted image was associated with its non-distorted image. We followed the same training strategy as in [8, 11].

We show some visual results from the testing set in Fig. 3. The reconstructed images exhibit average SSIM scores of 0.38 ± 0.10 and 0.40 ± 0.11 for Wiener and DeblurGAN deconvolutions, respectively. As observed for both scenarios (blind and non-blind deconvolution), our optimized lens distorts enough to safeguard facial details, preserving people's anonymity.

4. Conclusion

We presented an end-to-end image caption generation model based on LSTMs and an attention module, which integrates a refractive optical element in the image acquisition, to enhance privacy preservation by introducing a controlled distortion. The approach does not only effectively preserves the privacy of individuals, objects, and places depicted in the images, but also achieves optimal caption performance in the COCO dataset, as demonstrated by high scores on the metrics *Bleu* and *Meteor*. To validate the robustness of the method, deconvolution attacks were attempted to evaluate privacy preservation. The deconvolution results reinforce the idea that the method effectively preserves privacy by preventing the recovery of sensitive

information from distorted images.

References

[1] Jafar A Alzubi, Suresh Satapathy, Soham Taneja, Paras Gupta. Deep image captioning using an ensemble of cnn and lstm based deep neural networks. *Journal of Intelligent & Fuzzy Systems*, 40(4):5761–5769, 2021. 1

[2] Peter Anderson, Mark Johnson, Stephen Gould, Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 3

[3] Paula Arguello, Jhon Lopez, Carlos Hinojosa, Henry Arguello. Optics lens design for privacy-preserving scene captioning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3551–3555. IEEE, 2022. 4

[4] Ling Du, Huazhu Fu, Wenqi Ren, Xinpeng Zhang. An efficient privacy protection scheme for data security in video surveillance. *Journal of visual communication and image representation*, 59:347–362, 2019. 1

[5] Joseph W Goodman. *Introduction to Fourier optics*. Macmillan Learning, 4 edition, 2017. 2

[6] Mohammad Nehal Hasnine, Hiroaki Ogata, Kousuke Mouri, Noriko Uosaki. Vocabulary learning support system based on automatic image captioning technology. In *Distributed, Ambient and Pervasive Interactions: 7th International Conference, DAPI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*, pages 346–358. Springer, 2019. 1

[7] Carlos Hinojosa, Juan Carlos Niebles, Henry Arguello. Learning privacy-preserving optics for human pose estimation. In *IEEE/CVF International Conference on Computer Vision*, pages 2573–2582, 2021. 1, 2

[8] Carlos Hinojosa, Ehsan Adeli, Li Fei-Fei, Juan Carlos Niebles. Privhar: Recognizing human actions from privacy-preserving lens. *Preprint arXiv:2206.03891*, 2022. 1, 2, 4

[9] Lun Huang, Wenmin Wang, Jie Chen, Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019. 1

[10] Andrej Karpathy, Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 4

[11] Orest Kupyn, Tetiana Martyniuk, Junru Wu, Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8878–8887, 2019. 3, 4

[12] Tsung-Yi Lin, Deva Ramanan, Piotr Dollár, C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 3

[13] Chi Liu, Tianqing Zhu, Jun Zhang, Wanlei Zhou. Privacy intelligence: A survey on image privacy in online social networks. *ACM Computing Surveys*, 55(8):1–35, 2022. 1

[14] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, Edward Cutrell. Understanding blind people's experiences with computer-generated captions of social media images. In *proceedings of the 2017 CHI conference on human factors in computing systems*, pages 5988–5999, 2017. 1

[15] Karl Manheim, Lyric Kaplan. Artificial intelligence: Risks to privacy and democracy. *Yale JL & Tech.*, 21:106, 2019. 1

[16] Antoinette Deborah Martin, Ezat Ahmadzadeh, Inkyu Moon. Privacy-preserving image captioning with deep learning and double random phase encoding. *Mathematics*, 10(16):2859, 2022. 1

[17] Kuo Niu, Chao Tian. Zernike polynomials and their applications. *Journal of Optics*, 24(12):123001, 2022. 2, 3

[18] Francesco Pittaluga, Sanjeev J Koppal. Privacy preserving optics for miniature vision sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 314–324, 2015. 1, 3

[19] Francesco Pittaluga, Sanjeev Jagannatha Koppal. Pre-capture privacy for small vision sensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2215–2226, 2016. 1

[20] Jianing Qiu, Megan A McCrory, Edward Sazonov, others. Egocentric image captioning for privacy-preserved passive dietary intake monitoring. *arXiv preprint arXiv:2107.00372*, 2021. 1

[21] Michael S Ryoo, Brandon Rothrock, Charles Fleming, Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 3

[22] Kurt Shuster, Hexiang Hu, Antoine Bordes, Jason Weston. Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[23] Vincent Sitzmann, Wolfgang Heidrich, Felix Heide, Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM*, 37(4):1–13, 2018. 1, 2

[24] Asokan Sivaprakash, Samuel NE Rajan, Sundaramoorthy Selvaperumal. Privacy protection of patient medical images using digital watermarking technique for e-healthcare system. *Current Medical Imaging*, 15(8):802–809, 2019. 1

[25] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pages 3156–3164, 2015. 1, 4

[26] Qingzhong Wang, Jia Wan, Antoni B Chan. On diversity in image captioning: Metrics and methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):1035–1049, 2020. 1

[27] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT press, 1949. 3

[28] Kelvin Xu, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057. PMLR, 2015. 1, 3, 4

[29] Jinao Yu, Yu Wang, Shibing Zhu, Ming Ding. Gan-based differential private image privacy protection framework for the internet of multimedia things. *Sensors*, 21(1):58, 2020. 1