OPTICS LENS DESIGN FOR PRIVACY-PRESERVING SCENE CAPTIONING

Paula Arguello, Jhon Lopez, Carlos Hinojosa, Henry Arguello

Department of Computer Science, Universidad Industrial de Santander
Bucaramanga, 680002, Colombia

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897535

ABSTRACT

Image captioning is a challenging task that connects two major artificial intelligence fields: computer vision and natural language processing. Image captioning models use traditional images to generate a natural language description of the scene. However, the scene could contain private information that we want to hide but still generate the captions. Inspired by the trend of jointly designing optics and algorithms, this paper addresses the problem of privacy-preserving scene captioning. Our approach promotes privacy preservation, by hiding the faces in the images, during the acquisition process with a designed refractive camera lens while extracting useful features to perform image captioning. The refractive lens and an image captioning deep network architecture are optimized end-to-end to generate descriptions directly from the blurred images. Simulations show that our privacy-preserving approach degrades private visual attributes (e.g., face detection fails with our distorted images) while achieving comparable captioning performance with traditional non-private methods on the COCO dataset.

Index Terms— Image Captioning; Privacy-preserving Lens Design; Deep Optics; Computational Optics.

1. INTRODUCTION

The image captioning task consists of describing the content observed in an image. This challenging task has received significant attention in recent years as it lies at the intersection of two major artificial intelligence fields: computer vision and natural language processing. Image captioning is applicable in various scenarios, e.g., usage in virtual assistants, support of the disabled, etc. Previous works have addressed the image captioning problem from different approaches. Most of them use recurrent neural networks for processing long sequences on an element-by-element basis due to their ability to make predictions based on temporality [1, 2]. Another approach consists of using short-term memory neural networks [3] that can process whole sequences data (image descriptions). Moreover, other works have relied on the attention technique [4] in their deep network models [5] to enable deeper image understanding through fine-grained analysis. In a traditional non-privacy image captioning pipeline, cameras are used to acquire multiple high-fidelity images, and then the image captioning network is tuned to improve accuracy. However, if the acquired images contain privacy-sensitive data, they could be exposed in an attack. Visual privacy protection has become very popular because of its high demand in security, medical, social networking, and other areas where privacy is essential [6]. Inspired by the trend of jointly designing optics and algorithms [7], recent work in [8] proposes an end-to-end framework to preserve privacy in the human pose estimation.

For the particular problem of describing images while preserving their privacy, the authors in [9] propose to use radio signals to obtain the descriptions of a house. Specifically, by joining the data obtained from the house floor map with the radio signals, the floor map is encoded from the perspective of the person's location. Similarly, 3D human skeletons are extracted from the radio signals while the reference system changes at each time step. Therefore, authors in [9] avoid the use of videos by using only the radio signals and the floor map of the house, preserving the privacy of the household members. Although this approach preserves privacy, the use of radio signals requires a radio-frequency device, increasing the costs of the addressed task, and it does not directly preserve privacy in the images but avoids their use. Further, an attacker could easily inject malicious radio signals to attack the proposed system directly. Another privacy-preserving image captioning approach was proposed in [10], where authors obtain detailed descriptions from the images about dietary intake, hence preventing the direct use of the images by nutritionists and reducing the risk of privacy leakage. However, authors in [10] do not propose a method to preserve user privacy over original images but only train a deep neural network with images where faces of people are masked.

Paper Contribution. In this work, we are interested in the preserving-privacy image captioning task. Our main observation is that by adopting a similar strategy as in [7, 8, 11], we can learn a refractive lens that can distort the scene and allow extracting useful features to perform image captioning. Our main contribution consists of jointly optimizing a refractive optical element and a convolutional neuronal network (CNN) to extract the description of the scenes and obtain a simulated lens design. Accordingly, the input images for the CNN first pass through the designed camera that contains our learned refractive element, which produces a distortion effect in the images and protects user privacy while extracting relevant features to perform image captioning. As far as we know, this is the first work incorporating the lens design with an image captioning workflow to preserve user privacy end-to-end.

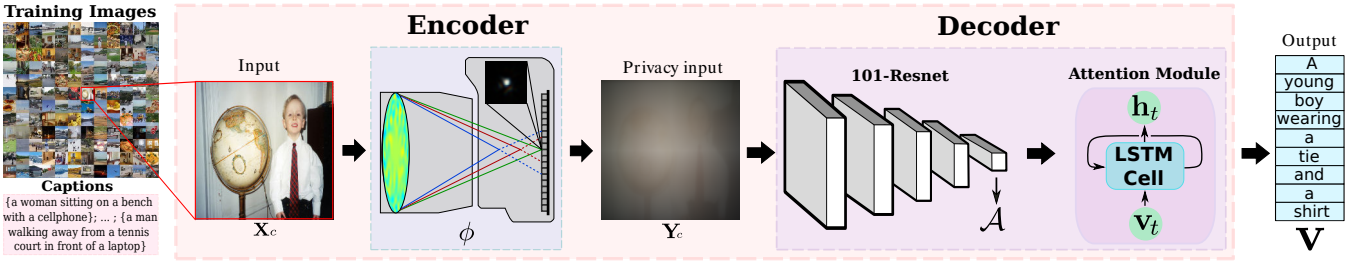

Fig. 1: Proposed end-to-end (2PSC) model. The optical encoder consists of a camera with a refractive lens. The decoder consists of convolutional feature extraction and an LSTM with attention, which generates a description from the privacy image.

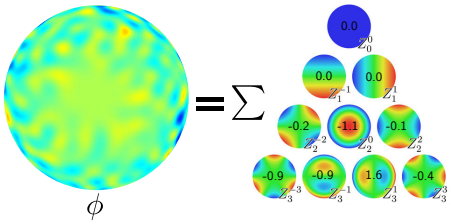

Fig. 2: Optimized refractive lens (left) and the corresponding first 9 Zernike Polynomials (right).

2. PROPOSED METHOD

In our proposed model, the optical encoder comprises a camera with a refractive optical element specifically learned for privacy preservation. A refractive optical element is designed by learning a linear combination of Zernike polynomials (Fig 2). Our end-to-end network learns optics by backpropagating the gradients from the image captioning network decoder to the optics layer. Then, the output of our optical encoder is distorted images such that objects, people, and places are anonymized. The second part of our proposed architecture is a decoder that learns and generates captions from the blurred images acquired from the designed lens. Specifically, the decoder first uses a CNN to extract features from the distorted images. Finally, it uses a long-short-term memory (LSTM) network that produces the captions of the images by generating a word at each step conditioned by a context vector.

2.1. Optical Encoder

The main goal of the optical encoder in our framework (Fig. 1) is to design a refractive optical element to encode the physical characteristics of an image and preserve essential features for the image captioning. Therefore, to achieve an end-to-end model in which we can optimize the camera lens for privacy-preserving image captioning, the camera lenses must be parameterized to perform backpropagation. In brief, we adopted a similar strategy as the authors in [8, 7, 11] to couple the modeling and design of two essential operators in the imaging system: the wave propagation and phase modulation. We model the image acquisition process using the point spread function (PSF) defined in terms of the lens surface profile to emulate the wavefront propagation and train the parameters of the refractive lens. Specifically, using the Fresnel approximation and the paraxial regime [12], the PSF is described as

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

where $T_\lambda(\cdot)$ represents the transfer function [12] with (f_x, f_y) as the spatial frequencies and λ the wavelength; $U_\lambda(x,y)$ denotes the complex-valued wave field immediately before the lens [7]; $t_\phi(\cdot)$ denotes the phase modulation introduced by the lens and the wave propagation; $\mathcal{F}\{\cdot\}$ denotes the 2D Fourier transform; (x', y') are the spatial coordinates on the camera plane and (x, y) are the coordinates on the lens plane. The phase modulation function $t_\phi(x,y) = e^{\mathrm{j}\frac{2\pi}{\lambda}\phi(x,y)}$ in Eq. (1) is produced by the lens surface profile ϕ [8], as

$$\phi = \sum_{j=1}^{q} \alpha_j Z_j, \quad (2)$$

where Z_j is the j-th Zernike polynomial in Noll notation, α_j is the corresponding coefficient, and q is the number of employed polynomials. Each Zernike polynomial represents a wavefront aberration; therefore, the linear combination of these aberrations will form the surface profile, see left on Fig. 2. Assuming that the image formation is a shift-invariant convolution of the image and the PSF, the acquired private images for the three RGB channels can be modeled as:

$$\mathbf{Y}_\ell = \mathcal{S}_\ell(\mathbf{H}_\lambda * \mathbf{X}_\ell) + \mathbf{N}_\ell, \quad (3)$$

where the sub-index ℓ denotes the color channel; $\mathbf{X}_\ell \in \mathbb{R}_+^{w\times h}$ represents the underlying scene with $w \times h$ pixels; $\mathbf{H}_\lambda$ denotes the discretized version of the PSF in Eq. (1); $\mathbf{N}_\ell \in \mathbb{R}^{w\times h}$ represents the Gaussian noise in the sensor; $\mathcal{S}_\ell(\cdot) : \mathbb{R}^{w\times h} \to \mathbb{R}^{w\times h}$ is the camera response function, which is modeled as linear operator; and $*$ denotes the 2D convolution operation. It is worth noting that the optimization of the camera lens aims to achieve the maximum visual distortion in the acquired images by learning the set of coefficients $\{\alpha_1, \cdots, \alpha_q\}$.

2.2. Decoder

2.2.1. *Feature extraction*

In contrast with traditional approaches that learn features from a full resolution image, the objective of the decoder in the proposed architecture is to learn from the encoded sensor images acquired with Eq. (3). Hence, we employ a CNN to extract a set of feature vectors $\mathcal{A} = \{\mathbf{a}_i\}_{i=1}^{L}$, where $\mathbf{a}_i \in \mathbb{R}^D$ corresponds to parts of the blurred image. Specifically, we use the 101-layer Residual Network (ResNet-101) trained on ImageNet, because of its high performance to extract features. In addition, we remove the last two classification layers of ResNet since our objective is not image classification.

2.2.2. LSTM network

Similar than the work proposed by Kelvin et.al. [1], we use LSTM network with attention to generate a caption by computing one word at each time step t conditioned on a context vector $\hat{\mathbf{z}}_t$, the previous hidden state $\mathbf{h}_{t-1}$, and the previously generated word $\mathbf{v}_{t-1}$. The context vector $\hat{\mathbf{z}}_t$ is a dynamic representation of the most important part of the image at time t and is obtained from the feature vectors $\mathbf{a}_i$, and a function ψ with parameters $\boldsymbol{\theta}_t$, i.e., $\hat{\mathbf{z}}_t = \psi(\mathbf{a}_1, \cdots, \mathbf{a}_L; \boldsymbol{\theta}_t)$.The LSTM network can be described as

$$\mathbf{i}_t = \sigma\left(\mathbf{W}_i \mathbf{E}\mathbf{v}_{t-1} + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{Q}_i \hat{\mathbf{z}}_t + \mathbf{b}_i\right),$$
$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f \mathbf{E}\mathbf{v}_{t-1} + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{Q}_f \hat{\mathbf{z}}_t + \mathbf{b}_f\right),$$
$$\mathbf{c}_t = \mathbf{f}_t \mathbf{c}_{t-1} + \mathbf{i}_t \tanh\left(\mathbf{W}_c \mathbf{E}\mathbf{v}_{t-1} + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{Q}_c \hat{\mathbf{z}}_t + \mathbf{b}_c\right),$$
$$\mathbf{o}_t = \sigma\left(\mathbf{W}_o \mathbf{E}\mathbf{v}_{t-1} + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{Q}_o \hat{\mathbf{z}}_t + \mathbf{b}_o\right),$$
$$\mathbf{h}_t = \mathbf{o}_t \tanh\left(\mathbf{c}_t\right),$$

where $\mathbf{i}_t, \mathbf{f}_t, \mathbf{c}_t, \mathbf{o}_t$, and $\mathbf{h}_t$ are the input, forget, memory, output and hidden state of the LSTM, respectively. Additionally, the $\mathbf{W}_s, \mathbf{U}_s, \mathbf{Q}_s \in \mathbb{R}^{n \times m}$ terms denote learned weight matrices and the $\mathbf{b}_s \in \mathbb{R}^n$ terms denote bias vectors. The sub-index $s = \{i, f, c, o\}$ indicates which variable is computed from the learned matrices and biases (e.g., $\mathbf{U}_i$ is used to compute the input). $\mathbf{E} \in \mathbb{R}^{m \times K}$ is the embedding matrix, m and n denote the embedding and LSTM dimensionality, respectively, and K is the vocabulary size. σ is the logistic sigmoid activation and $\tanh(\cdot)$ represents hyperbolic tangent activation function. For more details, we refer the interested reader to [1].

2.3. Loss Function

Two important aspects were considered to find an appropriate cost function for our approach. The first is maintaining the visual image distortion, and the second is the performance at word generation. Then, our loss function is defined as:

$$\mathcal{L} = -\log(p(\mathbf{v} \mid \mathcal{A})) + \lambda \sum_{i=1}^{L} \left(1 - \sum_{t=1}^{C} \boldsymbol{\theta}_{ti}\right)^2 - \sum_{c=1}^{C} \log \frac{\exp\left(\mathbf{v}_c\right)}{\exp\left(\sum_{i=1}^{C} \mathbf{v}_i\right)} \mathbf{g}_c + \left(1 - \frac{1}{J} \sum_{l=1}^{3} \|\mathbf{Y}_\ell - \mathbf{X}_\ell\|^2\right),$$

where C represents the length of the caption, $\mathbf{g} \in \mathbb{R}^C$ represents the ground truth caption, and J is the total image pixels for all channels. The first two terms are a doubly stochastic regularization to encourage the model to pay equal attention to every part of the blurred image [1]. The third term is the multi-class cross-entropy loss to generate a correct sequence of words. The last term uses the mean squared error to maximize the difference between original $\mathbf{X}_\ell$ and sensor $\mathbf{Y}_\ell$ images to achieve private images.

2.4. Training details

We used the Common Objects in Context (COCO) 2014 dataset [13] for training (83K images), validation (41K images), and testing (41K images). We trained our end-to-end model on an Nvidia Geforce RTX 3090 during 90 epochs and a batch size of 32. We used the Adam optimizer in all the models and learning rates of $1e-02$, $1e-04$, $5e-04$ for the optical encoder, the feature extraction layers of the decoder, and the LSTM network, respectively.

	Model	B-1↑	B-2↑	B-3↑	B-4↑	M↑
Non-Privacy	BRNN [16]	64.2	45.1	30.3	20.1	19.5
	NIC [2]	66.6	46.1	32.9	24.6	23.7
	CutMix [17]	64.2	-	-	24.9	23.1
	AAIC [18]	71.0	-	-	27.7	23.8
	Hard Attn [1]	<u>71.8</u>	50.4	35.7	25.0	23.0
	2PSC-w (ours)	**72.1**	**54.8**	**40.4**	**29.6**	**29.2**
Privacy	2PSC (ours)	70.7	<u>53.5</u>	<u>39.4</u>	<u>28.9</u>	<u>29.0</u>
	Defocus	56.1	36.7	24.2	16.3	20.4
	Low-Resolution	57.3	37.8	25.2	17.4	20.9

Table 1: Comparisons on the COCO test set. **B-#** denotes the Bleu metrics, and **M** is the Meteor metric. Results with (-) were not reported by authors in the corresponding papers.

3. RESULTS

This section shows the obtained results from simulations of our proposed privacy-preserving image captioning approach over the COCO test set. In general, visual privacy methods are not well explored in the literature. Therefore, to compare our method, we adapt the ideas of using low-resolution cameras [14] and cameras with a defocus lens [15] to provide visual privacy protection. Specifically, we change the encoder in our proposed end-to-end architecture (Fig. 1) with a low-resolution camera, which resizes the input image to 16×16 pixels, and a defocus lens, respectively.

Qualitative Results. Figure 3 shows a visual comparison of our proposed method using the optimized lens against from defocus lens and low-resolution cameras. The figure shows the image and caption outputs of each model. Additionally, we compute and show the PSNR between the original and distorted images by different approaches. As observed, in all approaches, the content of the images cannot be easily recognized; however, our method achieves the best description of the scene and, as expected, minimum PSNR.

To quantitatively evaluate our proposed approach, we use the standard BLEU [19] and Meteor metrics. BLEU-1,2,3,4 scores (values between 0 and 1) indicate how similar the candidate text is to the reference texts, with values closer to 1 representing more similar texts. The indexes {1,2,3,4} denote the evaluation of the precision for a contiguous sequence of $\tilde{n}$ items from a given sample ($\tilde{n}$-grams), where $\tilde{n} \in \{1, \cdots, 4\}$. Meteor metric [20] scores output captions model by aligning them to a set of references. Alignments are based on exact, synonym, and paraphrase matches among words and phrases.

Quantitative Results. Table 1 shows the quantitative results where we compare our model against the following non-privacy methods, i.e., these methods use cameras with standard lens: BRNN [16], Google NIC [2], Cutmix [17], AAIC [18], the LSTM network model Hard-Attention [1], 2PSC-w (our proposed model without the optimized lens). This table also provides a quantitative comparison with the privacy methods explained above: defocus and low-resolution cameras. The bold values in Table 1 represent the best re-

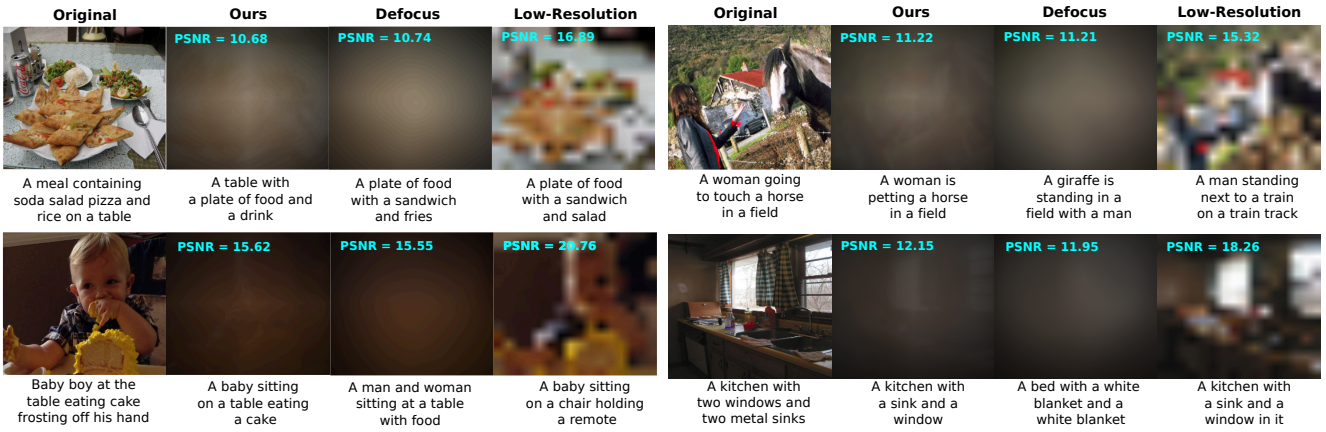

Fig. 3: Qualitative results on the COCO dataset test set. Under each privacy image, we show the caption obtained by each model, and under the original image, we show the ground truth caption. We compute the PSNR between the original and distorted images for each approach.

sults, and the underline values correspond to the second-best result. As shown in Table 1, our proposed architecture without privacy (2PSC-w) obtains the best results; because of our training improvements and the use of ResNet for feature extraction. Additionally, it can also be observed that our 2PSC approach achieves the best trade-off between image distortion and accuracy. Specifically, the performance of our proposed model does not decrease considerably compared with the non-privacy methods while providing privacy protection.

Privacy validation. To validate privacy, we evaluate the performance of a face detection network on the distorted images obtained by simulations of our designed lens and compare it with its performance when using normal (non-privacy) images. We used the face detection network proposed by Deng et al. [21] and performed the following experiments:

1) *Non-privacy:* We trained the face detection model from scratch by using original images resized to 256×256.

2) *Pre-trained:* We only evaluated the face detector performance trained in the previous experiment (Non-privacy) on distorted images acquired using our optimized lens.

3) *Training:* We trained the face detection model from scratch using blurred images from our optimized lens.

4) *Fine-tuning:* We first load the learned weights from the trained model in *Non-privacy* experiment and then perform fine-tuning using the blurred images acquired using our lens.

For each mentioned experiment, we trained the face detection network during 250 epochs on the WIDER FACE dataset [22], with default training parameters by authors [21]. Figure 4 (a) shows the precision-recall curves obtained by evaluating the face detectors obtained at each experiment on the validation set. As expected, the curves show that the best performance on face detection is achieved for the *Non-privacy* experiment since the faces are visible on the images. On the other hand, it is observed that the face detector performance decreases when using distorted images acquired by our lens since the model cannot correctly infer where the faces are located. Although the *Training* experiment achieves the best performance using our privacy-protected images, its overall performance is significantly worst compared with the traditional *Non-privacy* model. Additionally, Fig. 4 (b) shows the visual face detection results, where it can be seen the successful detection of the woman's face when using the model from the *Non-privacy* experiment and the failed detection when using the model from *Training* experiment on our distorted images. These results are expected due to our optimized lens.

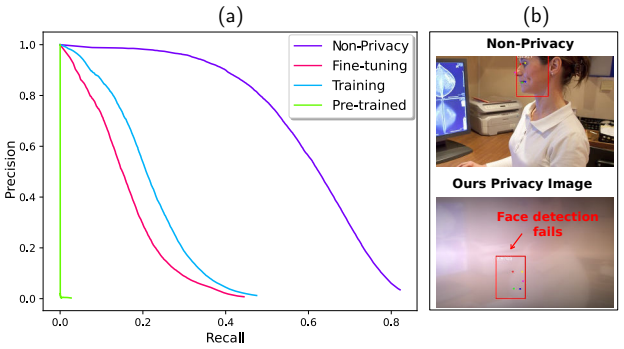

Fig. 4: Privacy validation via face detection task. (a). Precision-Recall curves on ablations experiments. (b). Comparison between the non-privacy model (top) and our proposed model (bottom).

4. CONCLUSION

In this work, we propose an image captioning model based on attention, which promotes privacy of the input images through a refractive optical element, causing a blurred visual effect on them. In this way, the people, objects, and places involved in the input images can be reserved. In addition, we maintain high performance on the BLEU metric with the COCO dataset despite visual distortion. Additionally, we trained a face detector on our private images to validate our method's effectiveness. As observed, the face detector fails; hence our learned distortion preserves privacy.

5. ACKNOWLEDGMENT

This research was sponsored by the Army Research Office/Laboratory under Grant Number W911NF-21-1-0099, and the VIE project entitled Dual blind deconvolution for joint radar-communications processing.

6. REFERENCES

[1] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*. PMLR, 2015, pp. 2048–2057.

[2] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *IEEE/CVF CVPR*, 2015, pp. 3156–3164.

[3] Ying Hua Tan and Chee Seng Chan, "Phrase-based image caption generator with hierarchical lstm network," *Neurocomputing*, vol. 333, pp. 86–100, 2019.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and 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*, 2018, pp. 6077–6086.

[6] Albert Haque, Arnold Milstein, and Li Fei-Fei, "Illuminating the dark spaces of healthcare with ambient intelligence," *Nature*, vol. 585, no. 7824, pp. 193–202, 2020.

[7] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein, "End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging," *ACM*, vol. 37, no. 4, pp. 1–13, 2018.

[8] Carlos Hinojosa, Juan Carlos Niebles, and Henry Arguello, "Learning privacy-preserving optics for human pose estimation," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2573–2582.

[9] Lijie Fan, Tianhong Li, Yuan Yuan, and Dina Katabi, "In-home daily-life captioning using radio signals," in *ECCV*. Springer, 2020, pp. 105–123.

[10] Jianing Qiu, Frank P-W Lo, Xiao Gu, Modou L Jobarteh, Wenyan Jia, Tom Baranowski, Matilda Steiner-Asiedu, Alex K Anderson, Megan A McCrory, Edward Sazonov, et al., "Egocentric image captioning for privacy-preserved passive dietary intake monitoring," *arXiv preprint arXiv:2107.00372*, 2021.

[11] Carlos Hinojosa, Miguel Marquez, Henry Arguello, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles, "Privhar: Recognizing human actions from privacy-preserving lens," *Preprint arXiv:2206.03891*, 2022.

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

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

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

[15] Francesco Pittaluga and Sanjeev J Koppal, "Privacy preserving optics for miniature vision sensors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 314–324.

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

[17] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *IEEE/CVF CVPR*, 2019, pp. 6023–6032.

[18] Chenggang Yan, Yiming Hao, Liang Li, Jian Yin, Anan Liu, Zhendong Mao, Zhenyu Chen, and Xingyu Gao, "Task-adaptive attention for image captioning," *IEEE Transactions on Circuits and Systems for Video technology*, vol. 32, no. 1, pp. 43–51, 2021.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Association for Computational Linguistics*, 2002, pp. 311–318.

[20] Satanjeev Banerjee and Alon Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Association for Computational Linguistics*, 2005, pp. 65–72.

[21] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *EEE/CVF CVPR*, 2020, pp. 5203–5212.

[22] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Wider face: A face detection benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.