Optics Lens Design for Privacy-Preserving Scene Captioning




Paula Arguello Jhon Lopez Carlos Hinojosa Henry Arguello

Universidad Industrial de Santander

Universidad Industrial de Santander

High Dimensional Signal Processing Research Group

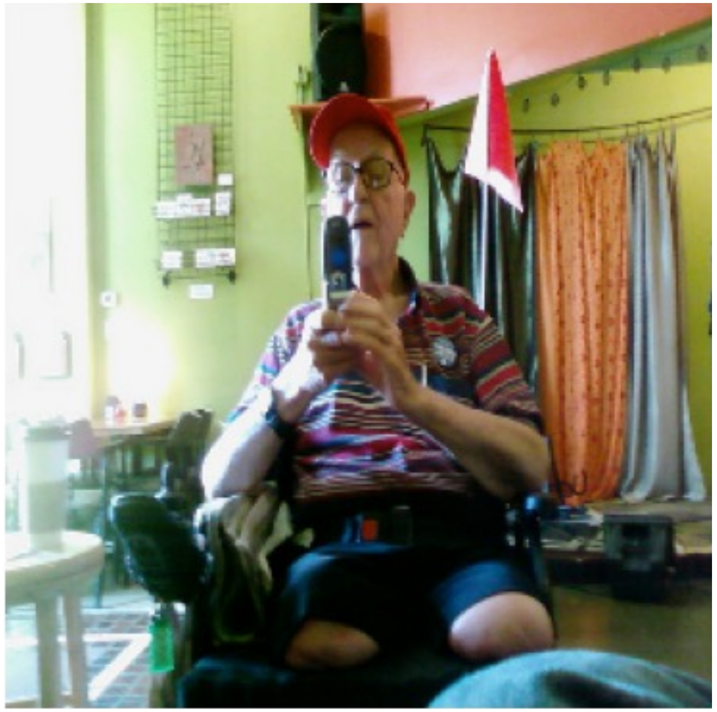
Image Captioning

a man that is next to a child with bread

a large giraffe standing next to a forest

people are playing volleyball on the sandy beach

Related Problem



Certain images may include content that should be private.
Sensitive content: Faces, Medical Eviroments, Elders, Toddlers.

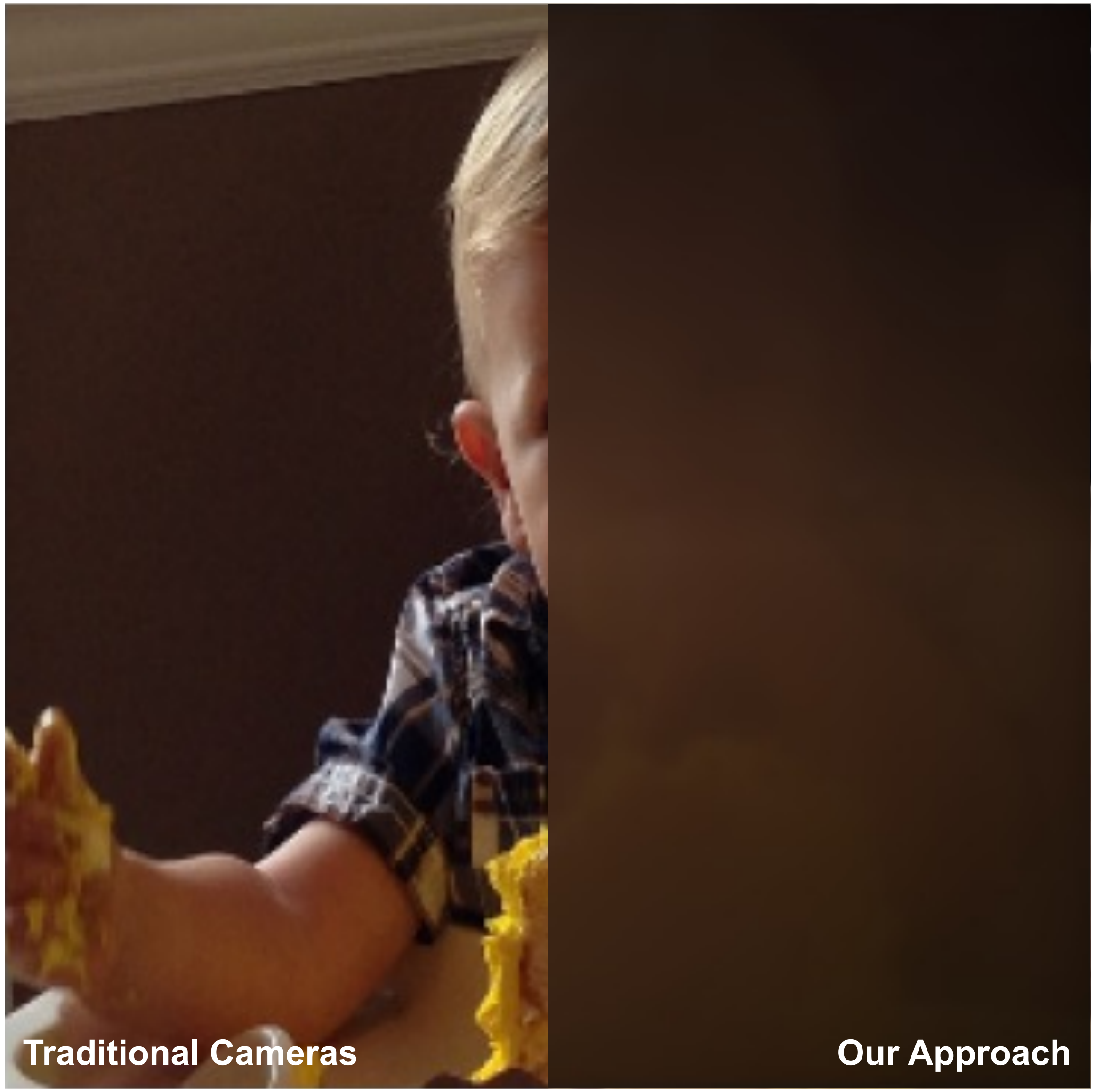
Not-private

a baby is eating a piece of cake

Private

a toddler is eating a cake

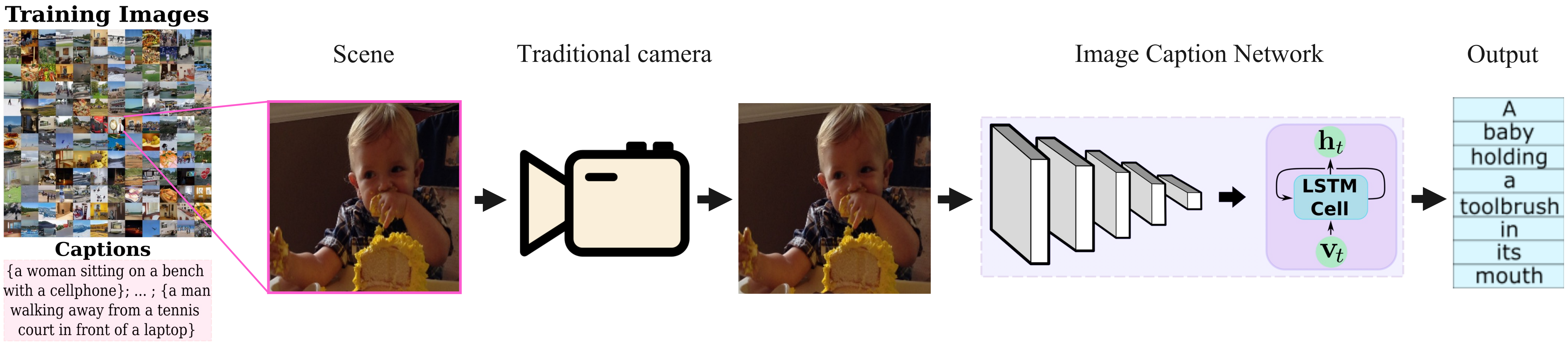
Not-private



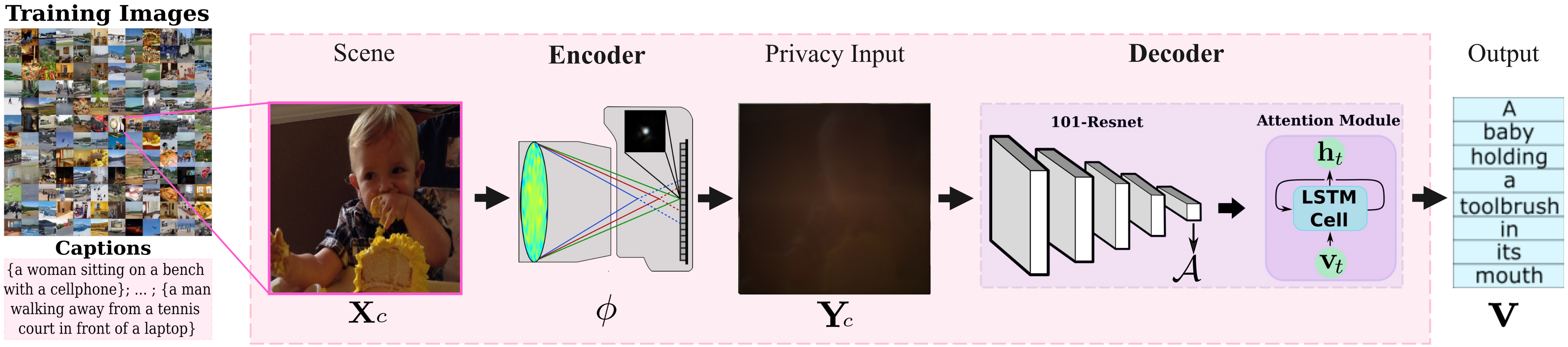
Private

Let's perform image captioning!

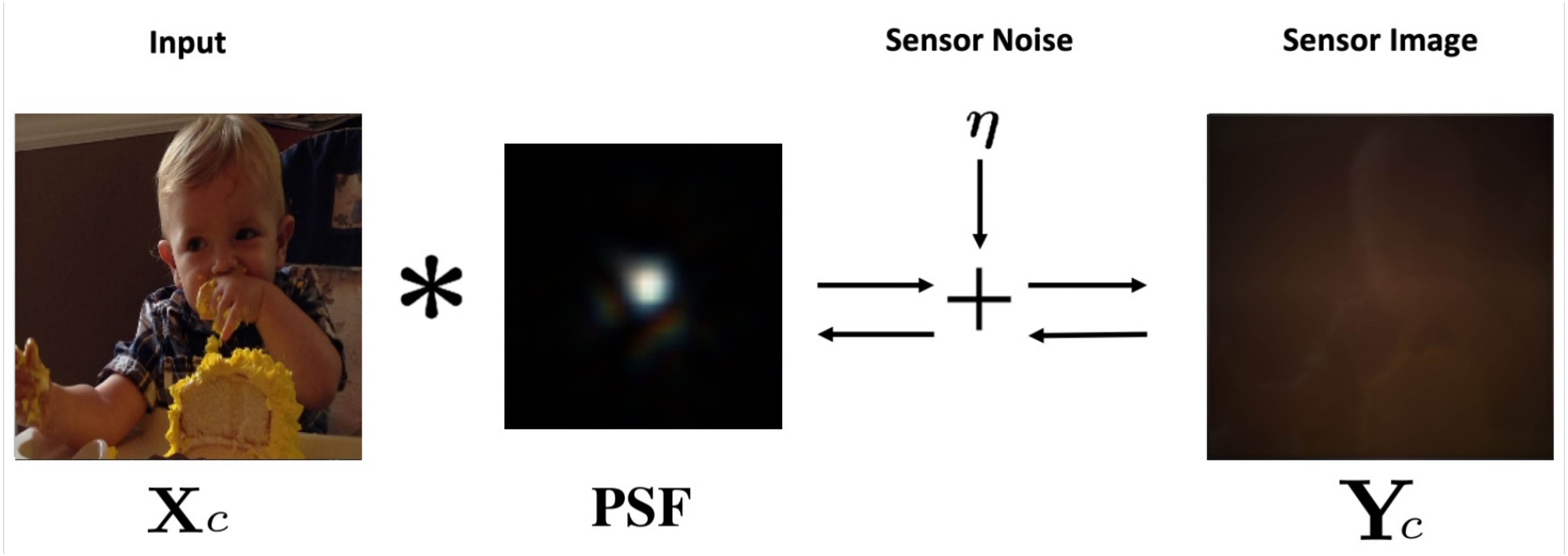
Traditional Approaches



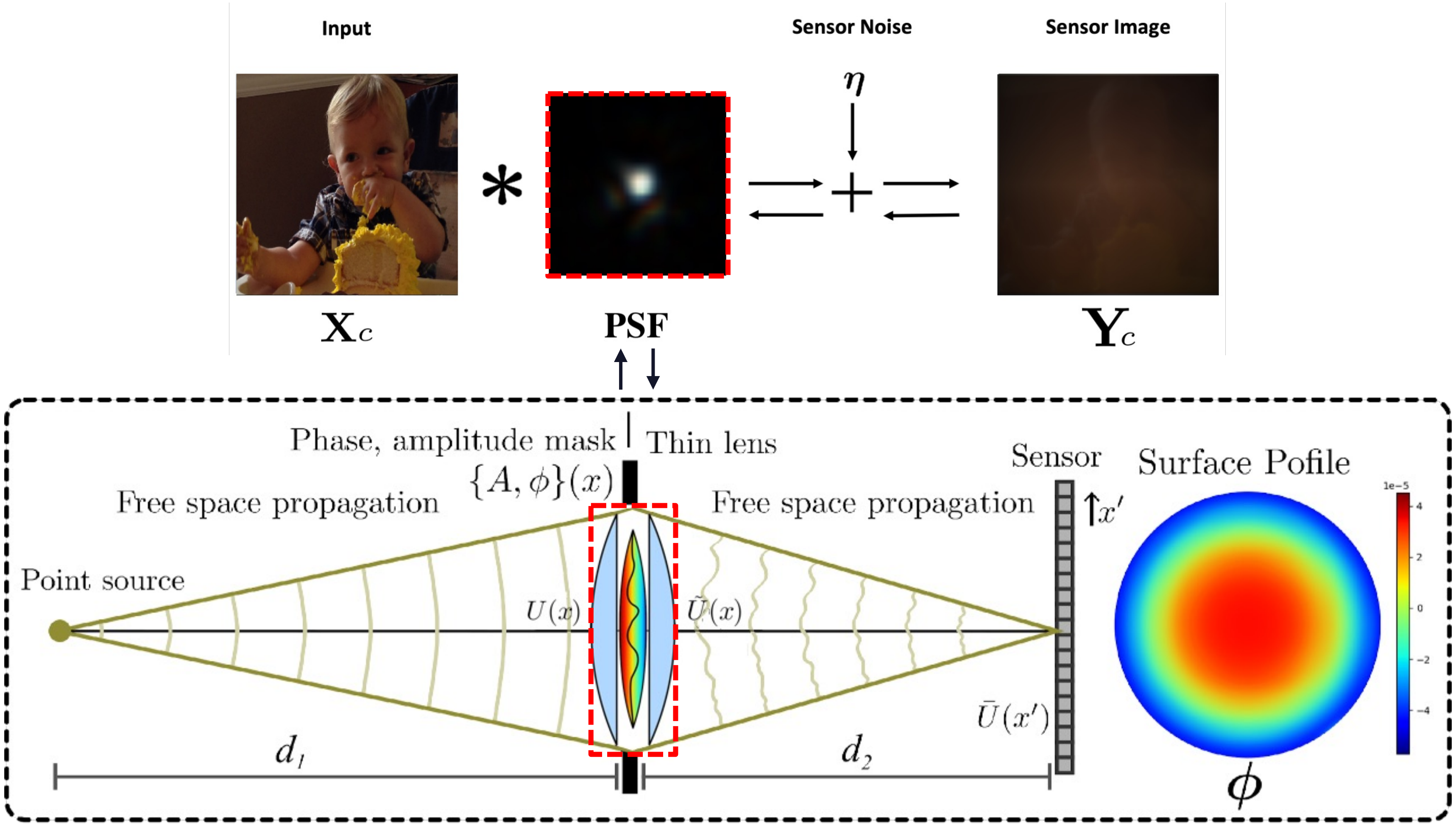
Proposed Method



Optical Encoder



Optical Encoder

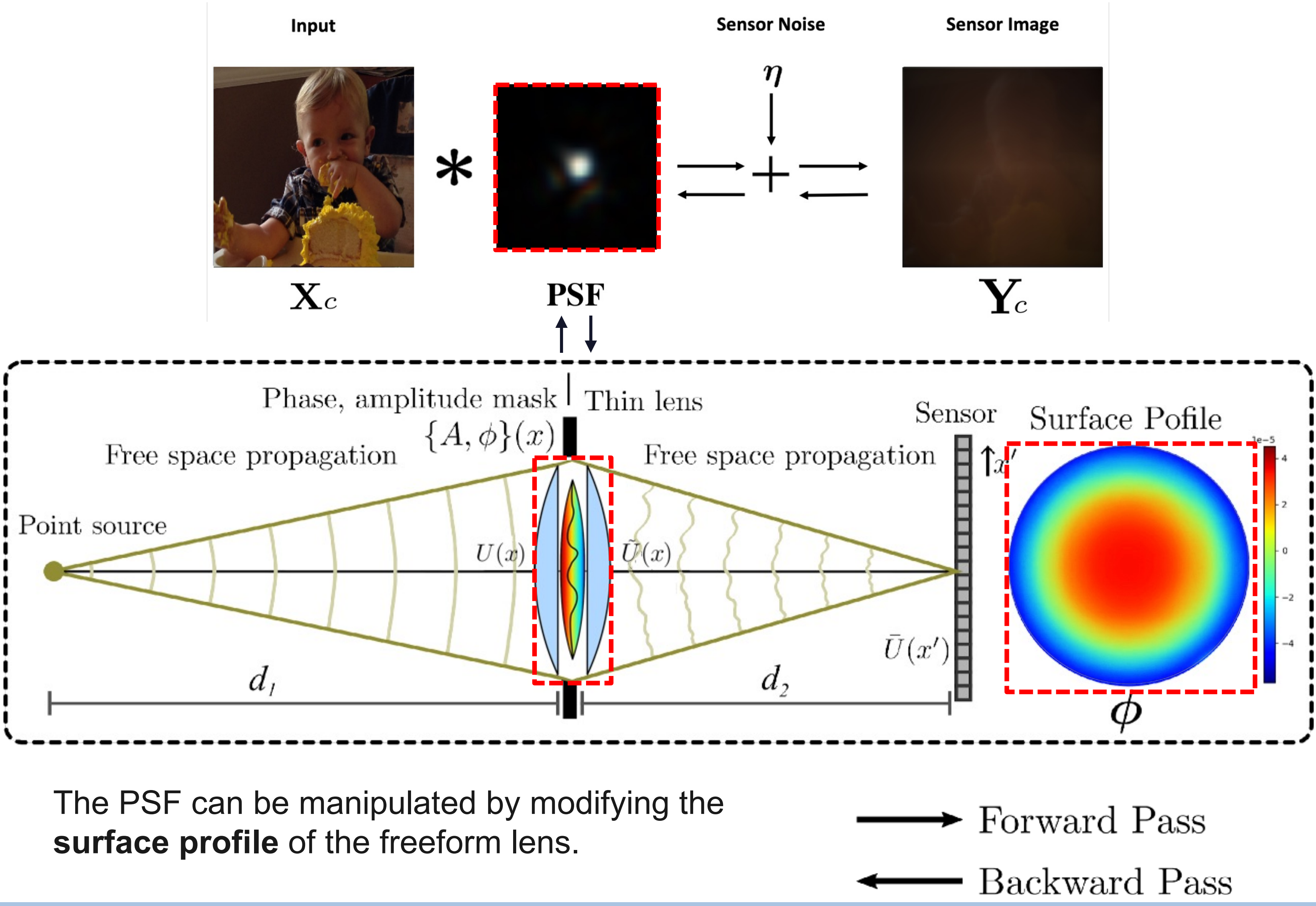


Our optical system consists of a convex thin lens and a refractive optical element (freeform lens) add-on.

→ Forward Pass

← Backward Pass

Optical Encoder

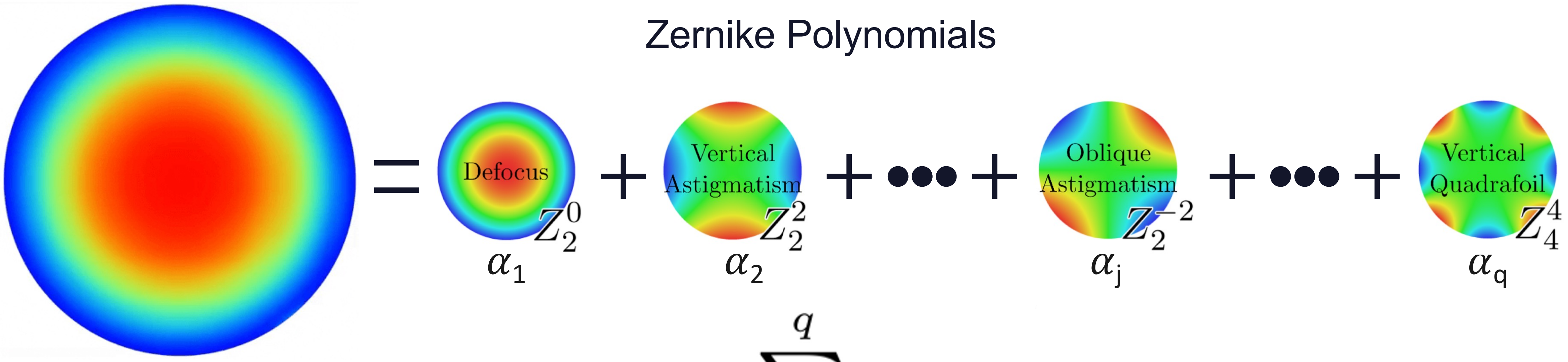


The PSF can be manipulated by modifying the **surface profile** of the freeform lens.

Optical Encoder

Surface Profile

Zernike Polynomials



ϕ

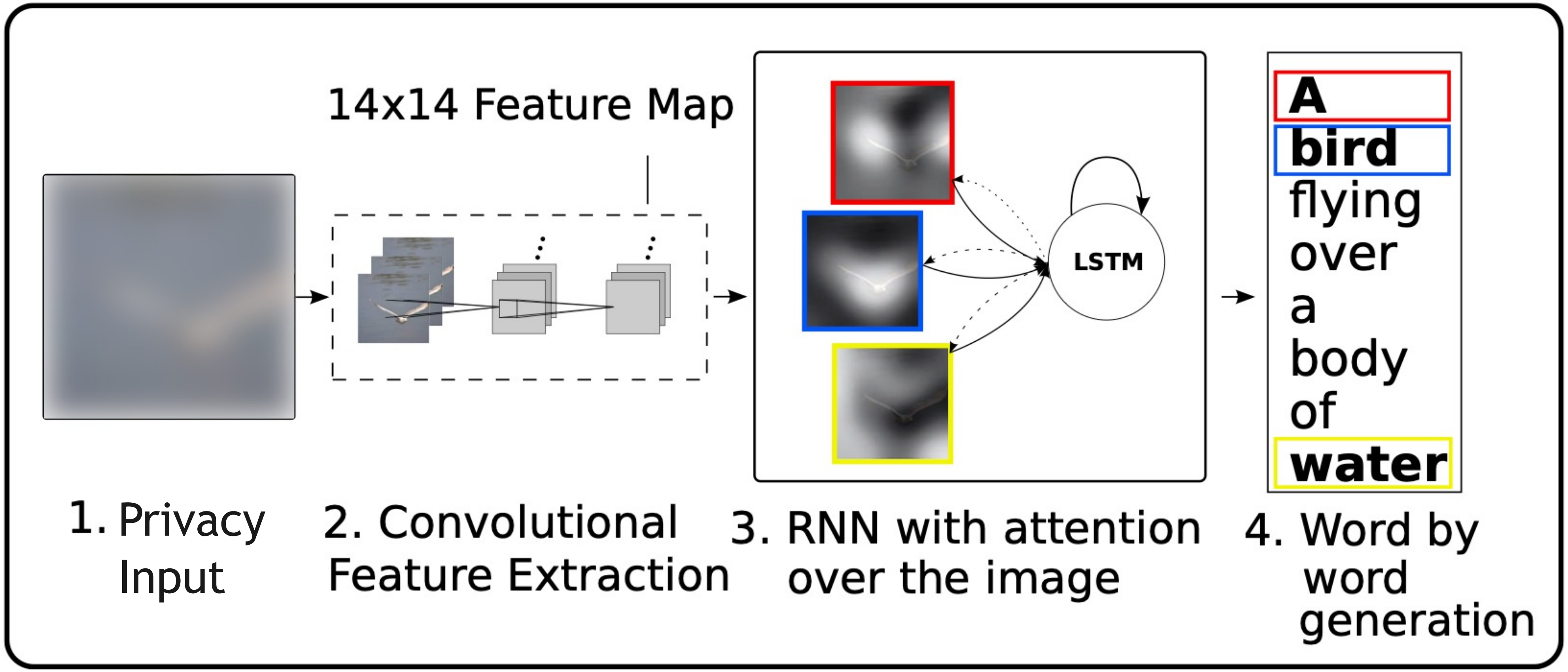
$$\phi = \sum_{j=1}^{q} \alpha_j \mathbf{Z}_j \, ,$$

* We learn α_j

We optimize the PSF by learning to add optical aberrations to the system.

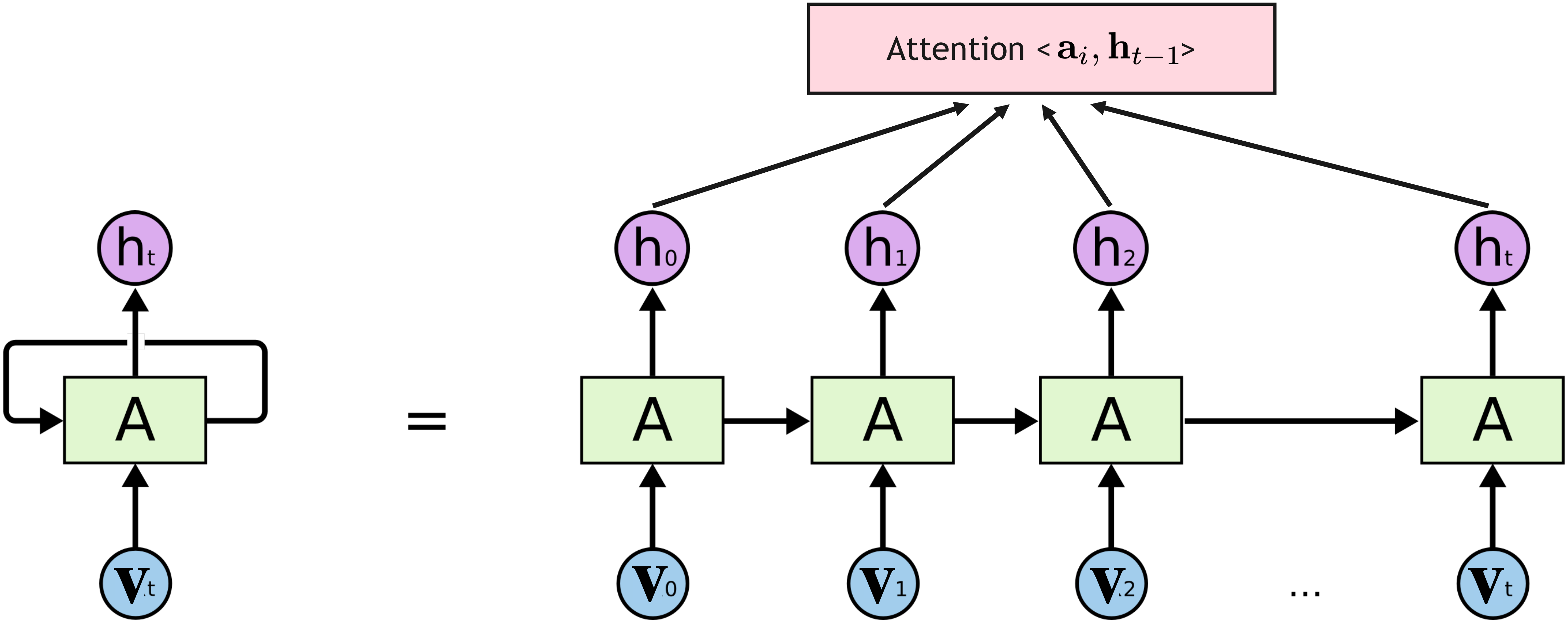
[1] Carlos Hinojosa, Juan Carlos Niebles, Henry Arguello; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2573-2582

Decoder

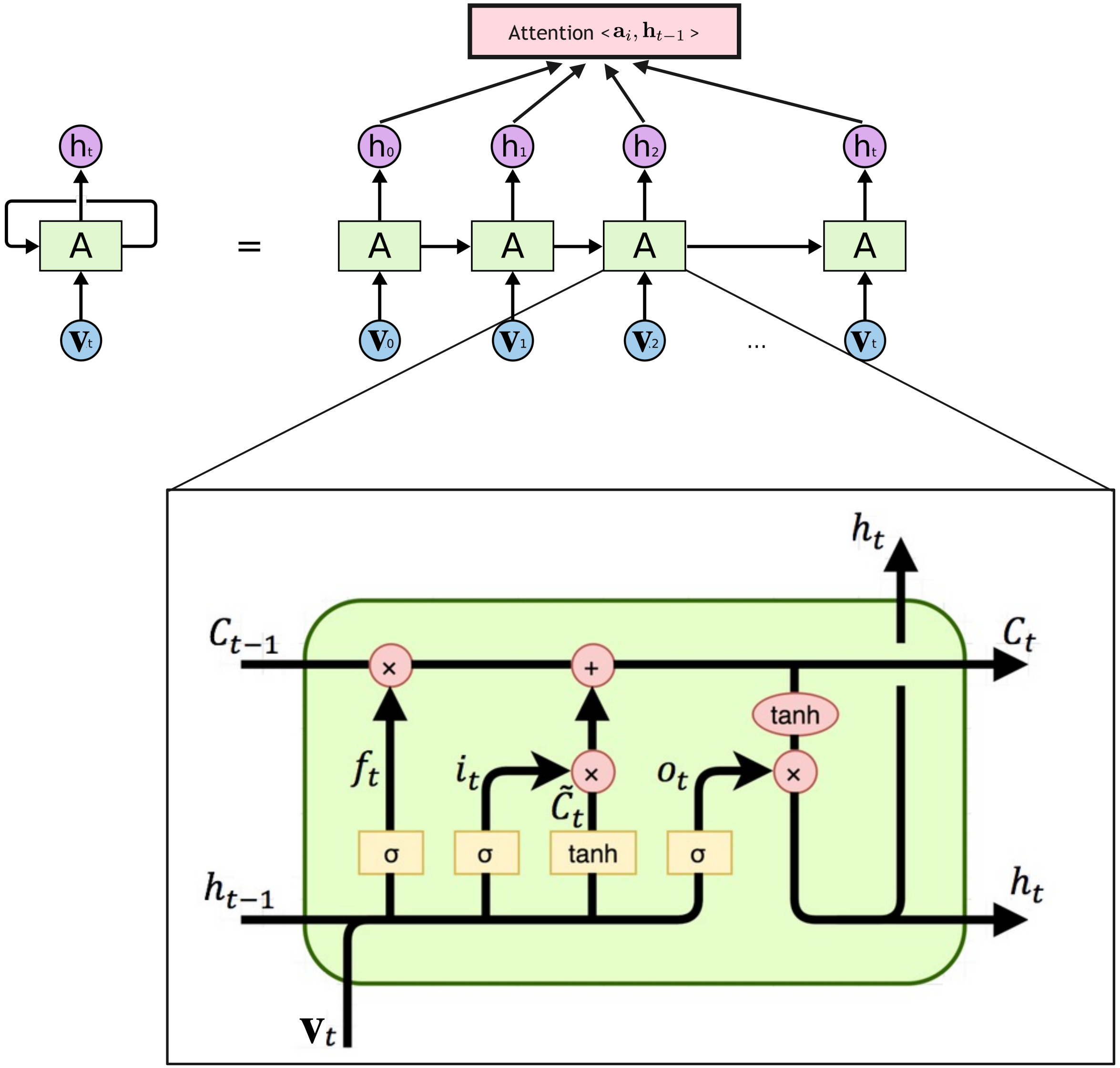


[2] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in ICML. PMLR, 2015, pp. 2048–2057.

Decoder: Recurrent Neural Network



Decoder: Recurrent Neural Network



$\mathbf{i}_t$ Input

$\mathbf{f}_t$ Forget

$\mathbf{c}_t$ Memory

$\mathbf{o}_t$ Output

$\mathbf{h}_t$ Hidden

Loss Function

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

Doubly stochastic regularization

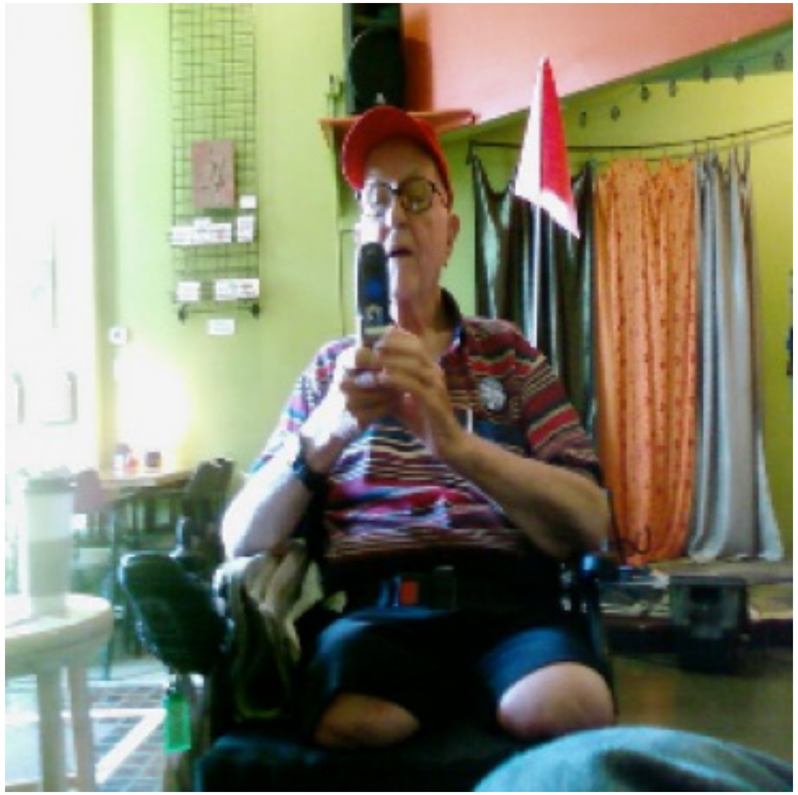
Multi-class cross-entropy loss

Mean squared error

Qualitative Results

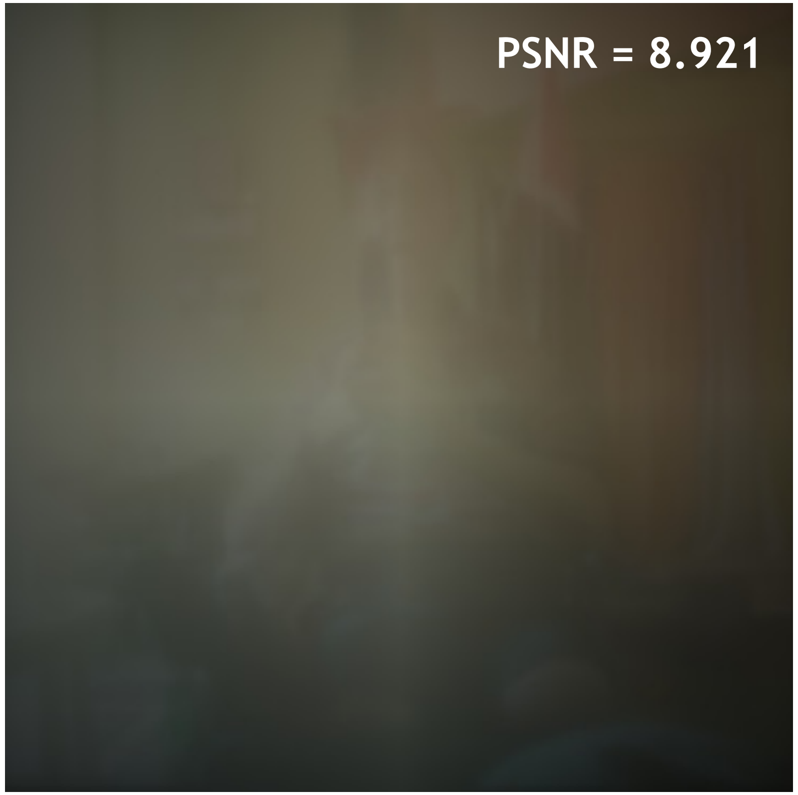
Original Image

Not-private



an elderly man looks at a cell phone

Sensor Image



Private

an old man looks at a cell phone screen

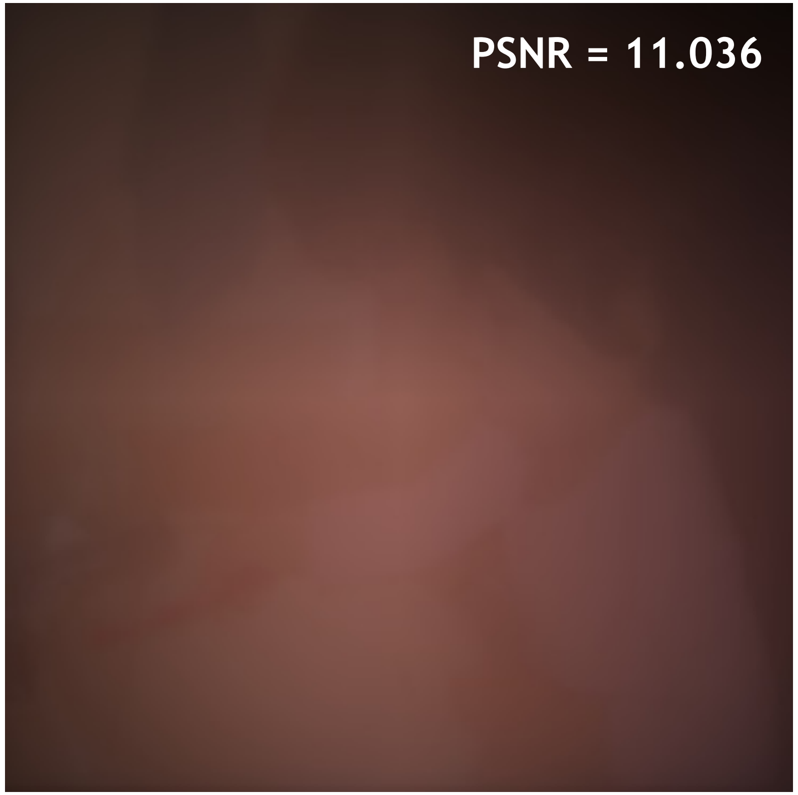
Qualitative Results

Original Image

Not-private

two children standing at the sink brushing their teeth

Sensor Image



Private

a little girl is brushing her teeth in a bathroom

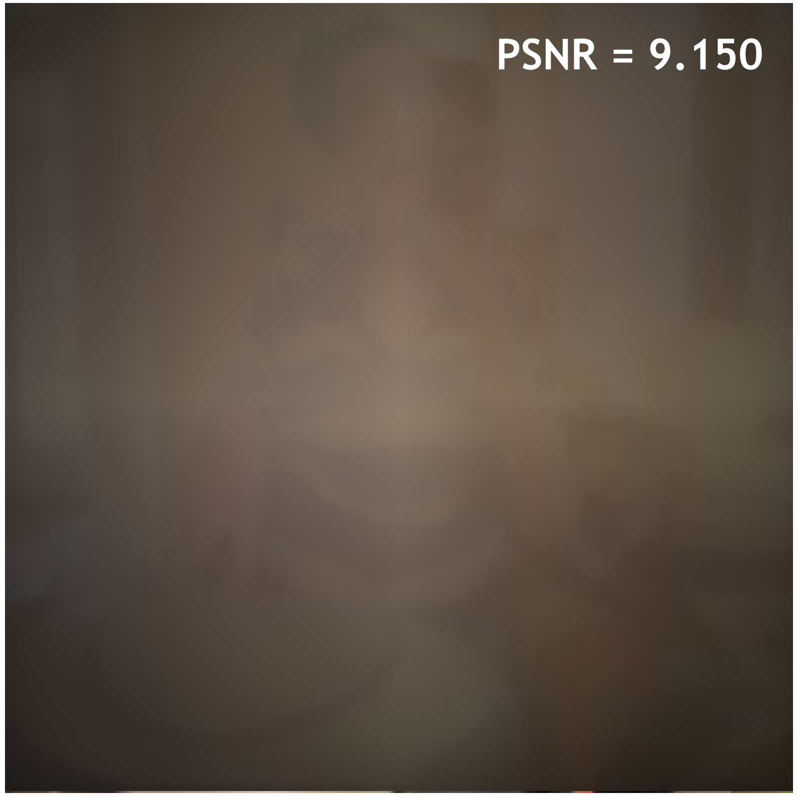
Qualitative Results

Original Image

Not-private

a man sitting at a table in a wheelchair while on a phone

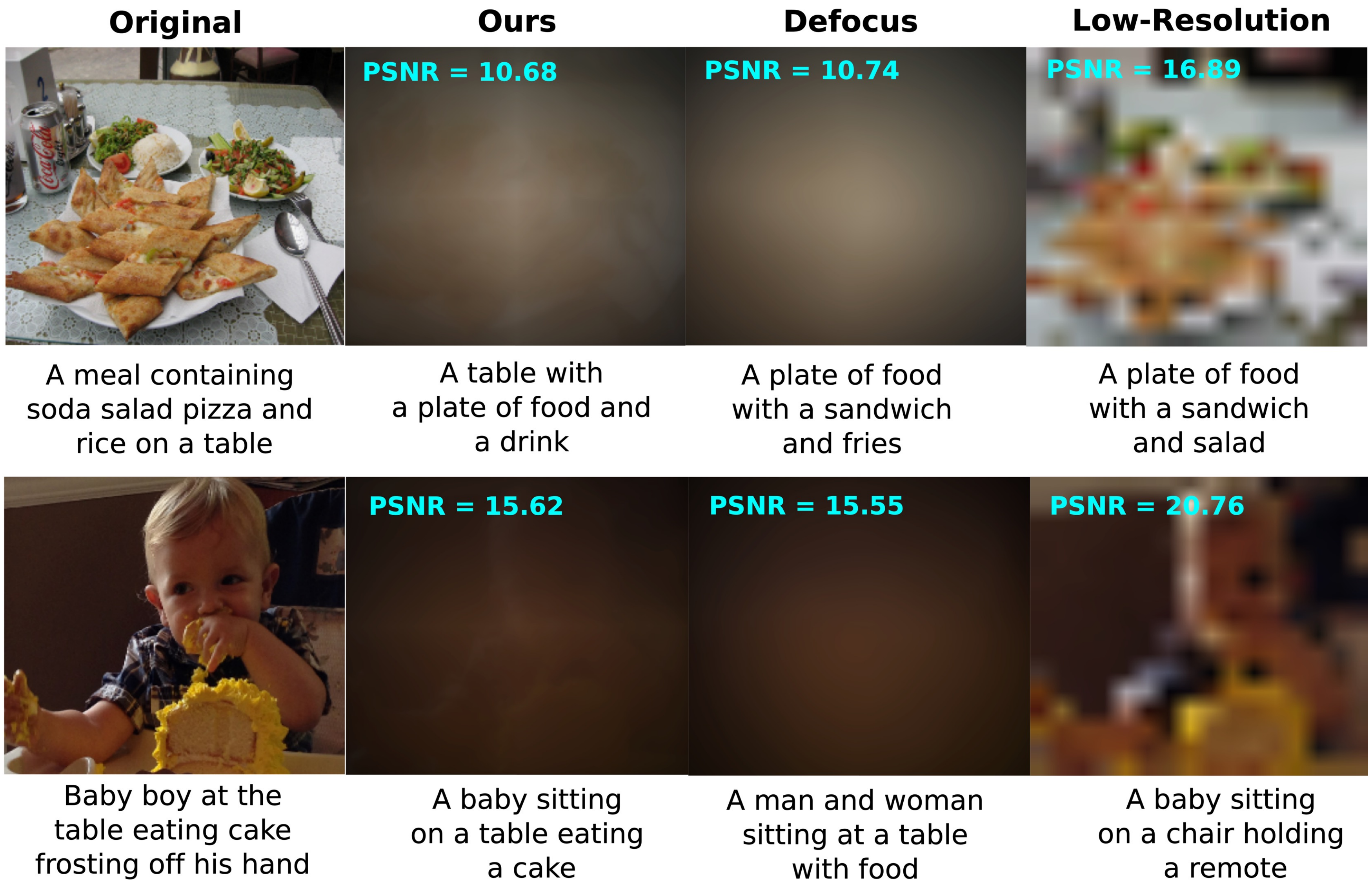
Sensor Image



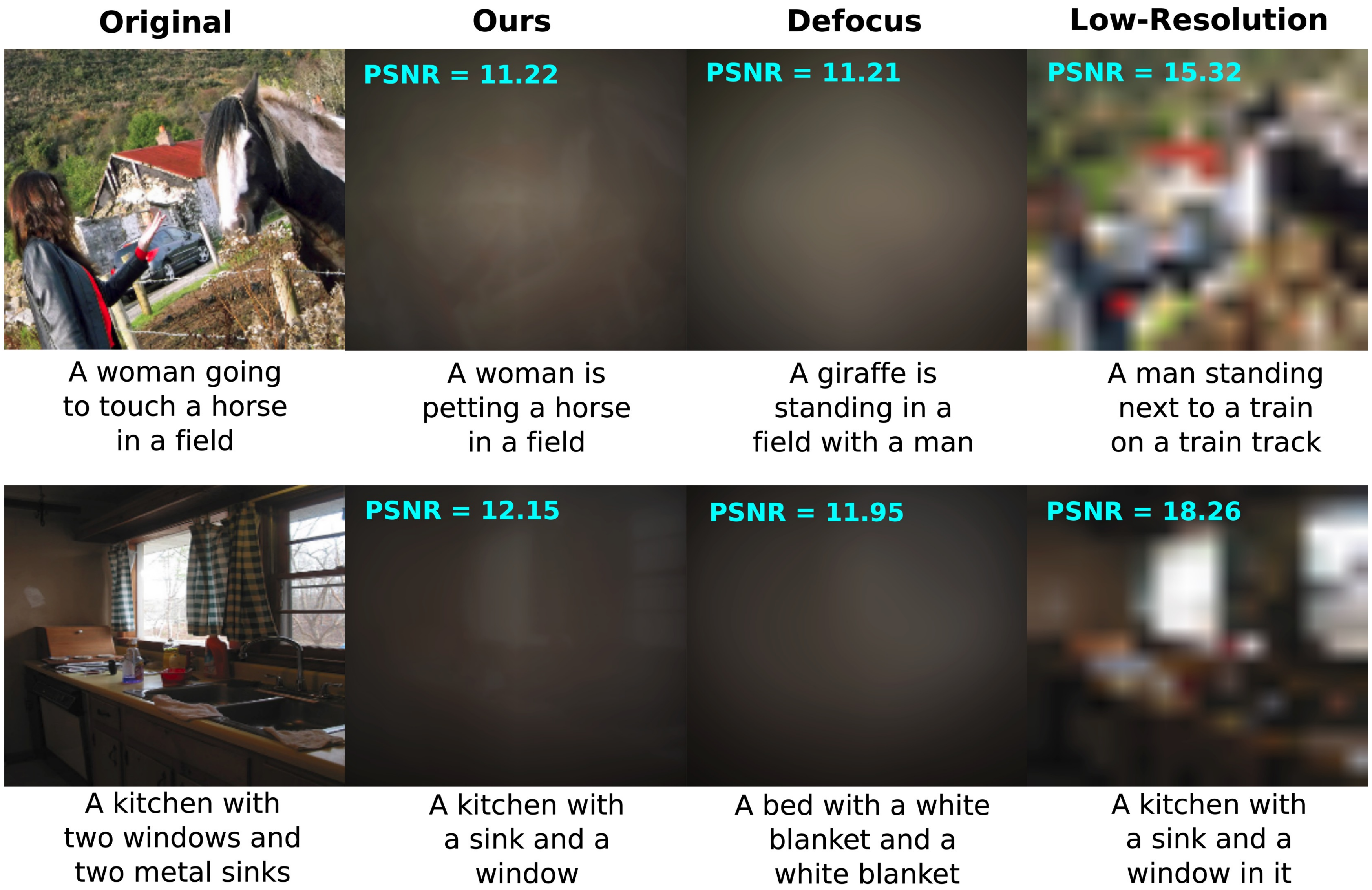
Private

a person in a wheelchair talking on a telephone

Ablation Studies



Ablation Studies

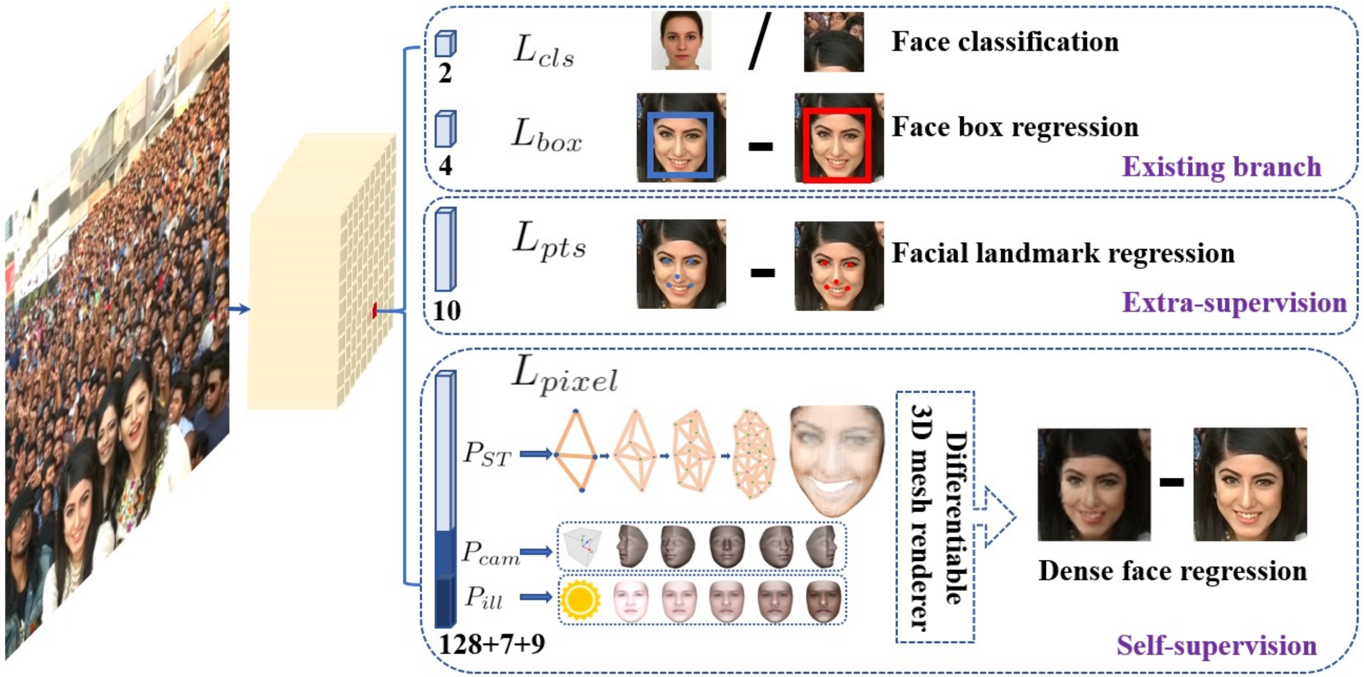


Quantitative Results

	Model	Bleu - 1	Bleu - 2	Bleu - 3	Bleu - 4	Meteor
Non - Privacy	BRNN [1]	64.2	45.1	30.3	20.1	19.5
	NIC [2]	66.6	46.1	32.9	24.6	23.7
	CutMix [3]	64.2	-	-	24.9	23.1
	AAIC [4]	71.0	-	-	27.7	23.8
	Hard Attn [5]	71.8	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	53.5	39.4	28.9	29.0
	Defocus	56.1	36.7	24.2	16.3	20.4
	Low-Resolution	57.3	37.8	25.2	17.4	20.9

Privacy Validation

pixel-wise face localisation on various scales of faces



[3] 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.

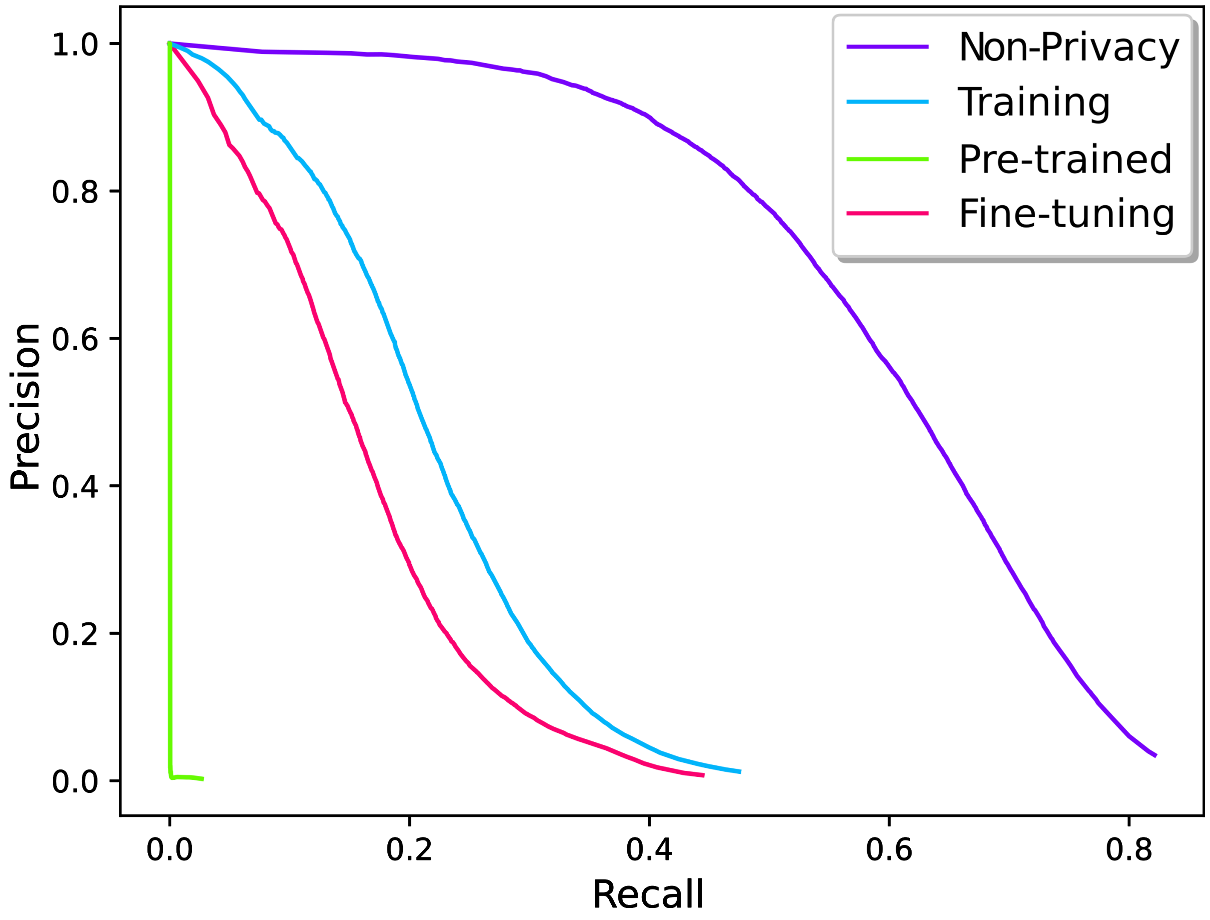
Privacy Validation

1. Non-privacy: We trained the face detection model from scratch with original images resized.

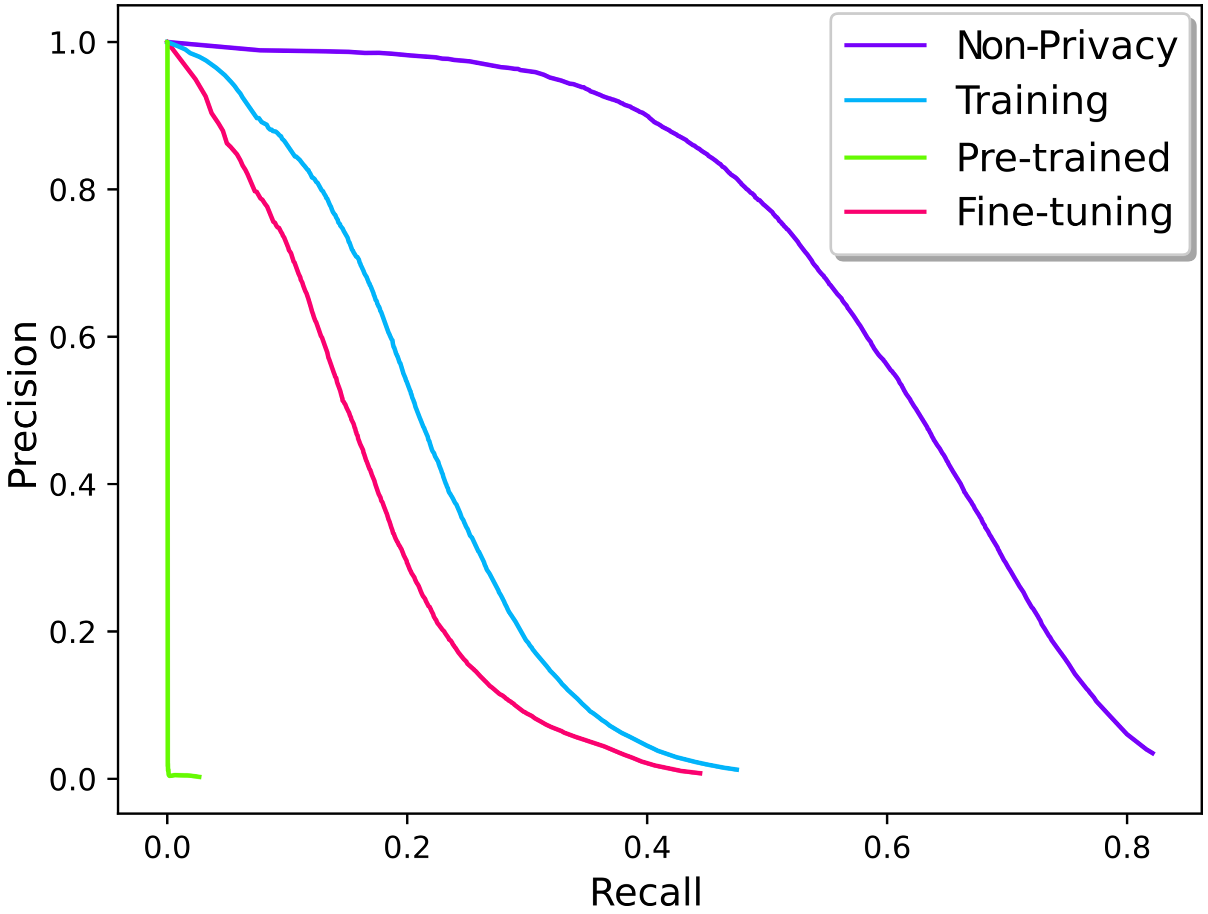
2. Training: We trained the face detection model from scratch using blurred images.

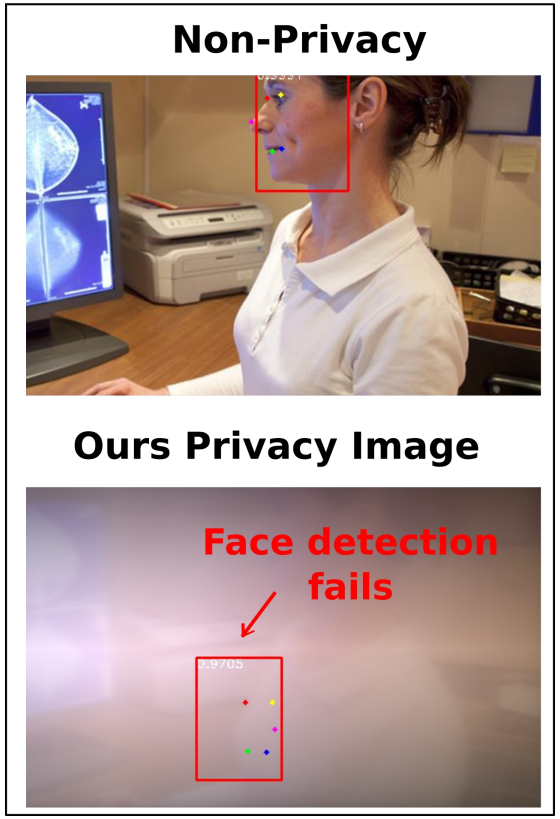
3. Pre-trained: We evaluated the previous experiment (Non-privacy) on distorted images.

4. Fine-tuning: We perform fine-tuning on the Non-privacy experiment using the blurred images.



Privacy Validation





Conclusions

- We propose an image captioning model based on attention, which promotes privacy of the input images, causing a blurred visual effect on them.
- The people, objects, and places involved in the input images can be reserved.
- We maintain high performance on the BLEU metric with the COCO dataset despite visual distortion.
- We trained a face detector on our private images to validate our method's effectiveness.

Thank you!
Any questions?

Universidad Industrial de Santander

High Dimensional Signal Processing Research Group