An End-to-End Optimized Lensless System for Privacy-Preserving Face Verification

Xin Cai[1,2], Hailong Zhang[3], Chenchen Wang[1,3], Wentao Liu[1,4], Jinwei Gu[2], Tianfan Xue[2]

[1]Shanghai AI Laboratory, [2]The Chinese University of Hong Kong, [3]Tsinghua University, [4]SenseTime

{cx023, tfxue}@ie.cuhk.edu.hk, jwgu@cuhk.edu.hk, {zhanghl21,wcc20}@mails.tsinghua.edu.cn, liuwentao@sensetime.com

Abstract. Lensless cameras, innovatively replacing traditional lenses for ultra-thin, flat optics, encode light directly onto sensors, producing images that are not immediately recognizable. This compact, lightweight, and cost-effective imaging solution offers inherent privacy advantages, making it attractive for privacy-sensitive applications like face verification. Typical lensless face verification adopts a two-stage process of reconstruction followed by verification, incurring privacy risks from reconstructed faces and high computational costs. This paper presents an end-to-end optimization approach for privacy-preserving face verification directly on encoded lensless captures, ensuring that the entire software pipeline remains encoded with no visible faces as intermediate results. To achieve this, we propose several techniques to address unique challenges from the lensless setup which precludes traditional face detection and alignment. Specifically, we propose a face center alignment scheme, an augmentation curriculum to build robustness against variations, and a knowledge distillation method to smooth optimization and enhance performance. Evaluations under both simulation and real environment demonstrate our method outperforms two-stage lensless verification while enhancing privacy and efficiency. Project website: lenslessface.github.io.

1 Introduction

With recent advances in optical engineering and computational photography, lensless imaging systems have emerged as a new, ultra-compact, lightweight, and cost-effective imaging system. Unlike traditional cameras that use large, heavy-weight focusing lenses, lensless cameras capture light using a simple and flat optic, such as an amplitude [4, 5] or phase mask [3, 6, 25]. These optics transform incoming light into highly multiplexed patterns. Consequently, images captured by lensless cameras are not immediately recognizable and require detailed knowledge of the system's mask pattern to reconstruct the original captured scenes. The low cost, small size, lightweight, and potential for increased privacy make lensless imaging systems a promising option for use in wearable devices, augmented and virtual reality, and biometric identification.

Specifically, lensless cameras offer a promising hardware solution for privacy-preserving imaging. Taking face recognition as an example, traditional recognition systems capture a clear face image using a convergence lens camera and
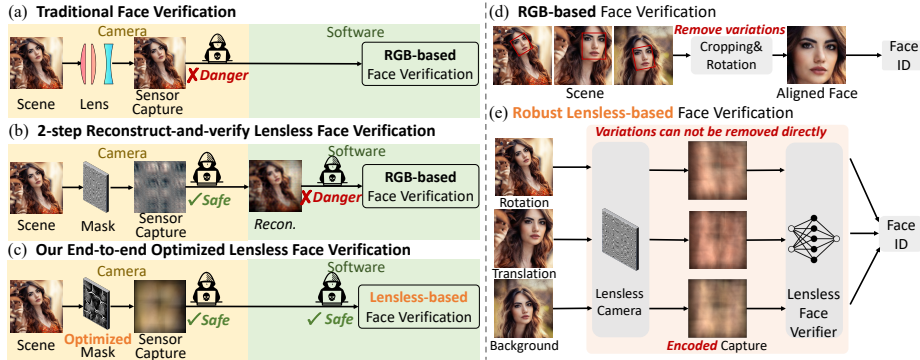
Fig. 1: Left: Comparison of different face verification approaches: lens-based (top) v.s. two-step lensless (middle) v.s. our optimized one-step lensless method. Our approach enhances privacy against software attacks as well as maintains robust face verification performance. **Right: Challenges for robust lensless-based face verification as compared to RGB-based methods.** Our robust lensless-based face verification is designed to accurately distinguish identities with encrypted captures, demonstrating resilience to variations in background, rotation, shift, and scales.

recognize facial identity through a verification algorithm, as shown in Fig. 1(a). Such a verification system may be vulnerable to the potential of a network attack, which may hijack camera data. On the other hand, lensless cameras only capture unrecognizable images and thus protect facial images from security attacks.

A typical lensless verification system uses a *reconstruct-and-verify* approach. This approach first restores the original image from the encoded one and then subsequently applies conventional verification algorithms. However, this method has its weaknesses. First, while lensless imaging can prevent hackers from hijacking camera data during transmission, it remains vulnerable to software attackers who can directly access reconstructed images, as illustrated in Figure 1(b). Second, the recognition accuracy is much lower than that of traditional RGB face verification systems, both because the reconstructed images are often blurry and the mask patterns designed for reconstruction are sub-optimal for face verification. Lastly, the additional reconstruction is not mandatory and may increase computational costs, making it a less desirable solution for edge computing.

Therefore, to improve verification accuracy while preserving privacy better, we propose a novel reconstruction-free method with lensless cameras for facial verification tasks. Unlike the reconstruct-and-verify process, our approach is a one-step solution that directly verifies faces on the encoded sensor images, without reconstructing facial images as intermediate results in the pipeline, as illustrated in Fig.1(c). This design has three advantages. First, the entire system is invulnerable to both camera data hijacking and software attacks, and it also reduces computational cost by skipping the reconstruction step. Second, the mask is also optimized to improve verification accuracy, capturing key facial features for verification. Therefore, it outperforms hand-crafted masks that are only designed for high-quality reconstruction. In summary, our approach is an end-

to-end optical-electronic neural network for face verification, offering enhanced security, cost efficiency, and improved verification accuracy.

Moreover, a simple end-to-end learning strategy for the proposed optical-electronic verification may result in poor accuracy (70% on LFW [20]). To deal with face rotation, translation, or background changes, traditional face verification pipelines need to first align input images to a canonical pose using simple transformations and cropping, as shown in   Fig. 1(d). However, in a lensless setup, it only captures an encoded image that does not keep the original spatial structure, making traditional alignment impossible, as shown in Fig. 1(e).

In response to these challenges, we propose a novel robust scheme to make end-to-end optimization with various face images practicable (from 70% to 95% on LFW). First, we design a new face-center alignment strategy, aligning the face center with the center of lensless images, enhancing system resilience against face translation. Second, to make our model robust to variations including rotation, scale, and background changes, we adopt an augmentation curriculum of variations. This curriculum gradually increases the complexities of variations, allowing the model to initially focus on learning face verification and then progressively build robustness against these variations. Third, to circumvent the local minimum during training, we proposed a novel cross-modality knowledge distillation method. This technique transfers verification capabilities from an RGB face verification model to our lensless verification model, smoothing the loss function landscape and aligning face verification performance to the RGB model. It is noted that the three strategies function as a holistic system, with each process being essential for achieving satisfactory optimization results.

In summary, our contributions are as follows: 1) We introduce an end-to-end optimization-based face verification system using lensless cameras. The proposed method brings improvements in privacy, accuracy, and speed when compared to conventional two-stage lensless face verification techniques. 2) We propose a novel training scheme to make the end-to-end verification network trainable. The training scheme includes a face center alignment tailored for lensless setup, a new cross-modality knowledge distillation to avoid local minima, and an augmentation curriculum that improves the robustness of our model under different variations. 3). We evaluate the efficacy of our method in both simulated and real environments. Experiments show that our method outperforms the two-step solutions, moving closer to the accuracy of traditional RGB-based approaches while inheriting hardware-level privacy protections from the lensless design.

2   Related Work

2.1   Lensless imaging

Lensless imaging [7] has become an emerging topic in visible light computational imaging due to its potential for miniaturizing camera systems. Unlike traditional cameras, lensless imaging systems use a mask element to modulate an incoming scene instead of a focusing lens. This technique can employ a variety of masks, including phase gratings [41], random diffusers [3, 25], hand-crafted designed phase masks [6, 10], amplitude masks [2, 4, 39], compressive samplers [18, 37],

and spatial light modulators [17, 49, 50]. Replacing the lens with masks results in multiplexed sensor captures that lack visual resemblance to the scene. This provides inherent privacy protection as the original scene is non-visible.

Transforming lensless images into identifiable images usually involves the use of computational reconstruction algorithms, which can be broadly categorized into optimization-based methods and data-driven methods. Optimization based approaches model the reconstruction as a regularized least squares problem, with prominent examples such as sparsity in the spatial domain [4, 6] and frequency domain [34]. Conversely, data-driven methods [23, 30, 35] use deep neural networks to learn a mapping from the sensor captures to the reconstructed images.

2.2   Perception with Lensless Camera

Due to the inherent hardware-level privacy protection feature, lensless cameras have been widely utilized to perform perception tasks [15, 21, 28, 29, 31, 45], especially those associated with humans and faces, like face attribute detection [5], face recognition [14, 38, 42, 44, 47], and action recognition [45]. Some [21, 44] adopt a two-stage reconstruction and perception pipeline, which could be suboptimal in terms of accuracy, privacy, and efficiency. Others [14, 29, 31] try to perform the perception tasks directly on encoded sensor images. However, due to the lack of mask optimization in the process, their performance remains inferior compared to traditional RGB-based methods. More recently, for enhanced privacy or performance, some work [5, 38] have performed end-to-end optimization of lensless masks and neural networks for downstream perception tasks.

Our work is similar to [5, 38] in performing joint optical-electric neural network optimization. However, Bezzam et al. [5] optimize for privacy by downsampling encoded images, making the identities hard to distinguish. Shi et al. [38] only deal with aligned faces, and their face verification performance lags behind RGB-based methods. In contrast to these approaches, we develop a method for lensless-based face verification to withstand various variances, making it more suitable for real-world applications.

3   End-to-end Lensless Optics Formation

Deep optics [40, 46], or end-to-end methods for optimizing optical components, seek to co-design the optics and corresponding computational algorithm to maximize performance for a specific application. Based on this concept, we present an end-to-end-differentiable pipeline that supports joint optimization of the lensless camera mask and post-capturing processing network.

Imaging formulation model and simulation pipeline. In our setup, we place an image sensor with a lensless mask at a known distance from the scene to be captured. Based on the scalar diffraction theory [12], we have such an assumption: for a given wavelength λ, image formation between two parallel planes acts as a linear shift-invariant system. For incoherent light of a distant scene, the theory enables modeling the image formation $S_\lambda(x, y)$ on the sensor as a convolution:

$$S_\lambda(x, y) = (l_\lambda * p_\lambda)(x, y), \tag{1}$$

where $*$ is the convolution operator, $l_\lambda(x, y)$ represents the scene's spatial intensity distribution. $p_\lambda(x, y)$ denotes the point spread function (PSF). In our

simulation, we use RGB images as the scene intensity distribution and simulate PSF based on the mask pattern. More details are discussed in the supplementary.

Based on this model, the lensless sensor capture S is formulated as a linear function $S = g(I, M)$ of the RGB image I and the mask pattern M. This linear shift-invariant assumption is based on the scene-to-mask distance obeying the paraxial criterion [40]. The majority of standard face verification scenarios meet this criterion, thereby validating the assumption for such applications.

Trainable mask to PSF. We formulate our lensless mask M with learnable parameters for transparency $w(x, y)$ at varying positions (x, y). The modulation for the amplitude $A_m(x, y)$ attributed to the incident wave is expressed as: $A_m(x, y) = w(x, y)$. We can optimize the mask parameters $\theta_M = \{w(x, y)\}$ to obtain the desired light modulations and PSF, hence enhancing the results of subsequent tasks. Note that in amplitude masks, the PSF $p_\lambda(x, y)$ remains consistent across any wavelength λ and we can drop the wavelength parameter.

End-to-end training objective. Leveraging a differentiable image formulation, we can simultaneously optimize the optical mask and electronic model for desired tasks. Unlike previous methods that only optimized the post-capture processing model, our end-to-end joint training also finds an optimal mask that extracts useful features for the target tasks. These features are stored in encoded images and are processed by the subsequent model, as illustrated in Fig. 2.

Given a training dataset $\{I_i, t_i\}_{i=1}^N$ with input scene image I_i and the target label t_i, the optimization objective is:

$$\min_{\theta, \theta_M} \sum_{i=1}^N \mathcal{L}(t_i, \mathcal{F}_\theta(g(I_i, M))) - \alpha \sum_{x,y} w(x, y), \qquad (2)$$

where the first term $\mathcal{L}$ is for the loss function of the target task. For face verification, $\mathcal{L}$ is the ArcFace loss function [11]. $\mathcal{F}_\theta$ is the downstream task model, like the face verification model, with parameter θ. $g(\cdot, \cdot)$ is the differentiable rendering process described above The second term penalizes small masks, to ensure the learned mask's transparent regions are large enough to yield unidentifiable images for privacy-preserving. The α is a factor to balance the losses and in our experiments, we set $\alpha = 0.01$.

4 Lensless Face Verification Pipeline

Our goal is to make the optical mask and face verification model optimal for face verification through an end-to-end training process. To increase efficiency and safeguard privacy, our pipeline skips the reconstruction process and directly executes face verification on encoded captures. Following the previous setups in [44], we assume that only one face appears in the captured scene.

One challenge of this pipeline is that the encoded captures it produces are not inherently suitable for face recognition. As shown in Fig. 1(d), in traditional RGB face verification, the input images are first cropped and aligned to a canonical space before being sent to the verification model, to reduce the impact of rotation, translation, and cluttered backgrounds. However, this process cannot be applied to a lensless camera, as the encoded captures encapsulate all scene
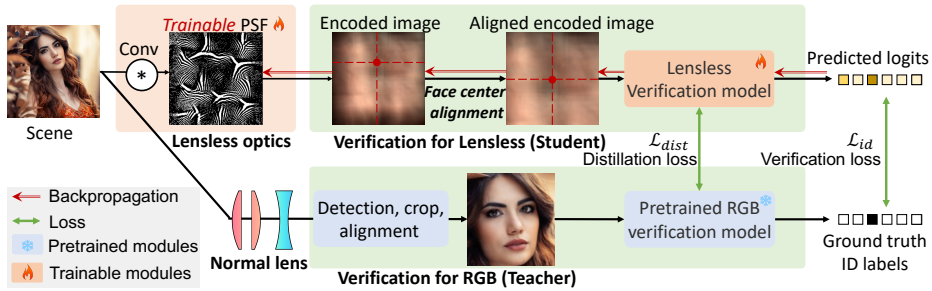
Fig. 2: Overview of our end-to-end optimization pipeline. The top pathway simulates lensless imaging using a learnable mask to encode a facial scene into a sensor capture. The capture after face-center alignment is fed into the lensless student model. In the bottom pathway, the same scene is aligned by preprocessing, and then processed by a trained RGB teacher model. The teacher's output features and ID labels supervise the training of both the lensless student model and the learnable mask.
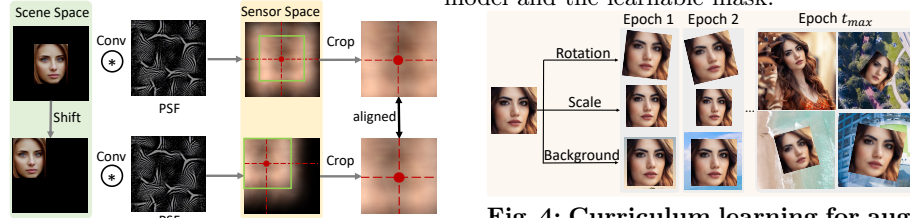

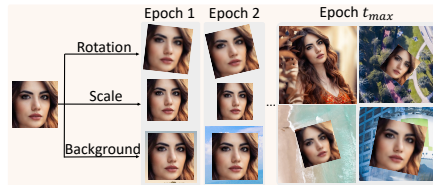**Fig. 3: Center alignment** for lensless capture of facial scenes.

Fig. 4: Curriculum learning for augmentation. The augmentation intensity gradually increases with training epochs.

information, making direct cropping or alignment infeasible. A simple way to ensure robustness to different variations is using a training set that includes these variations or augmenting existing ones with such variations. Nonetheless, lensless captures lack the spatial structures vital for face recognition, which complicates the use of simple end-to-end optimization methods. Consequently, these methods alone have not been successful in establishing robust lensless face verification.

To train a robust face verification that can handle translation, rotation, and background changes, we propose a holistic end-to-end optimization framework comprising three critical components: 1) A novel physics-based alignment algorithm tailored for lensless captures, designed to counteract facial translations before processing by the face verification model; 2) A new augmentation curriculum to increase the robustness of the face verification model; 3) A cross-modality distillation method to ensure facial feature alignment and effective optimization.

4.1 Alignment before Verification

First, we explore the face-center alignment method, a preprocessing step for lensless captures before they are processed by the verification model. To start with, we define two spaces: *scene space* and *sensor space*, as shown in Fig. 3. Images before capture (optical mask) are in the *scene space* and images captured by the sensor after light passes through the mask, are in the *sensor space*.

Unlike RGB images where the alignment is done by cropping faces from backgrounds, lensless images embed entire scenes across every pixel, making tradi-

tional alignment infeasible. To solve this issue, we apply face-center alignment in the *sensor space* to center and crop an initially off-centered encoded facial capture to a fixed size within the *sensor space*. This alignment translates a face into a canonical position, reducing the training and testing gap and enhancing the accuracy of off-center faces. In contrast to RGB-based face alignment which handles all types of similarity transformation, our lensless alignment only focuses on translation due to translation equivalence. Other similarity transformations, like rotation and scaling, are addressed in a later augmentation curriculum step.

The key idea of lensless face alignment is based on the principle of translation equivalence in convolution. According to this principle, the shift of the convolution output equals the convolution of the shifted input image, except on image boundaries. Recall that lensless capturing can be formulated as a convolution, as shown in Eq. (1). Therefore, if we can determine the face center position in the *scene space*, we can simply center the face in the *sensor space*, as the centered sensor image (the bottom row of Fig. 3) is the same as if we centered the face in the scene space and captured it (the top row of Fig. 3).

Based on this, we train a face center detection model $\mathcal{G}_\phi$ on sensor images with the following objective, similar to Eq. (2):

$$\min_{\phi, \theta_M} \sum_{i=1}^{N} \mathcal{L}(y_i, \mathcal{G}_\phi(g(I_i, M))), \tag{3}$$

where $\{I_i, y_i\}_{i=1}^{N}$ is a face-center dataset with unaligned images I_i and alignment y_i, and M is the mask to be optimized. Then, we use the output of this detection model to center the face in sensor space.

4.2 Verification Model Training

After centering the face, we train a face verification model that can accurately recognize the face, under variations in rotation, scaling, and background changes. However, training a model to directly recognize encrypted captures under these conditions is difficult, often resulting in a model with poor generalization ability, as already verified by previous works on RGB-based face verification [43]. To address this issue, we propose two strategies to make this training process tractable, the curriculum learning of augmentation and cross-modality distillation.

Curriculum learning of augmentation. To enhance the robustness under different variations, we first implement random augmentations, including rotation, scaling, and background changes, to the face images within the *scene space*, as shown in Fig. 4. This enhances the diversity of our face recognition dataset. We then execute an end-to-end training process with the augmented data.

However, directly applying all augmentations to face images from scratch makes the training hard to converge. Therefore, we adopt an augmentation curriculum that progressively increases in difficulty. This enables the model to initially focus on acquiring face recognition capabilities through easy examples, and gradually develop robustness against variations with more challenging examples. We control the degree of augmentations, such as rotation angles and background sizes, and incrementally increase these augmentations, as illustrated in Fig. 4. Specifically, we amplify the augmentation of variations following a cos-annealing

warm-up strategy. Specifically, to augment a particular variation eventually with a maximum magnitude $\eta_{\max}$, we initiate the training with a minimum magnitude $\eta_{\min}$, and gradually increase the magnitude at each epoch t as:

$$\eta_t = \eta_{\min} + \frac{1}{2}\left(\eta_{\max} - \eta_{\min}\right)\left(1 - \cos\left(\frac{t}{t_{max}}\pi\right)\right), \tag{4}$$

where t_{max} is the total number of epochs. Experiments in Fig. 10 show that this curriculum learning (CL) greatly improves the stability of the training process.

Cross-modality distillation. Directly training face verification models on encoded images is very challenging, even when the impact of different variations is minimized through previous alignment and augmentation. This is because the encoded image does not keep the original spatial structure, making the training process hard to converge. Consequently, most previous methods reconstruct original images first before proceeding with downstream recognition tasks.

We propose a novel model distillation scheme to solve this challenge. Model distillation is widely used in transferring knowledge from a well-trained *teacher* model to another to-be-learned *student* model [16]. In light of this, we propose a cross-modality distillation method to employ an already trained RGB-based face recognition model to distill knowledge into the hybrid optical-electric lensless model, enhancing its generalization capabilities and optimization efficiency. The RGB model extracts highly discriminative features for identity verification, and through distillation, we aim to empower our lensless model with similar discriminative capabilities. Considering the substantial difference in modality between lensless and RGB images, we apply relational knowledge distillation [32] which focuses on transferring the structural relations of features, rather than the features themselves, to mitigate the effects of the modality gap. By transferring the relations of features, the inputs of teacher and student need not be pixel-matched as long as they represent the same concept. As shown in Fig. 2, the RGB *teacher* model's output features of the aligned face are used to supervise the training of the hybrid optical-electric *student* model with unaligned samples.

Moreover, distillation also brings additional advantages. It enhances the alignment of features from lensless captures at both instance-level and identity-level. The distillation process ensures that features from unaligned lensless images share a similar feature with the aligned RGB images and thus improves the robustness of the model under variations. Furthermore, given that RGB-based models proficiently associate different images of the same person with similar features, the distillation also makes features of lensless captures from a singular identity closely align with that identity's feature.

Optimization objective. Based on our above discussion on training strategies, we define the optimization objective as the summation of the identification loss and the distillation loss. Specifically, given a face dataset $\{I_i, t_i\}_{i=1}^N$ with the identity labels t_i and a pretrained RGB face recognition model $\mathcal{F}_t$, our training finds the optimal masks M and the parameters θ for lensless face verification model, by minimizing the following loss:

$$\min_{\theta,\theta_M}\sum_{i=1}^{N}\mathcal{L}_{id}(y_i, \mathcal{F}_\theta(\mathcal{C}(g(\mathcal{A}(t_i), M)))) + \mathcal{L}_{dist}(\mathcal{F}_t(\hat{I}_i), \mathcal{F}_\theta(\mathcal{C}(g(\mathcal{A}(I_i), M)))). \tag{5}$$
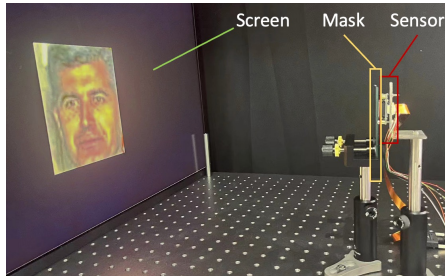
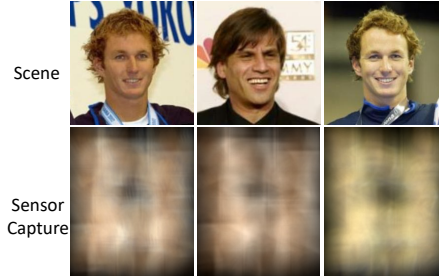Fig. 5: **Prototype** of our lensless camera for face verification.

Fig. 6: **Visualization** of scenes (top) and real captures (bottom).

In the loss function, a RGB face image has sequentially gone through data augmentation $\mathcal{A}(\cdot)$, lensless imaging simulation $g(\cdot)$, and face alignment $\mathcal{C}(\cdot)$, resulting in the aligned capture of the augmented image $\mathcal{C}(\mathcal{S}(\mathcal{A}(I_i), M))$. The first term is the ArcFace identification loss [11] that compares the recognition result on this image with the ground truth t_i. The second term is a distillation loss that minimizes the difference in recognition results between the RGB and lensless pipelines, following the relational knowledge distillation framework [32].

5   Experiments

In this section, we evaluate the performance and robustness of our end-to-end face verification pipeline, on both simulated and real-world data. Additionally, we analyze the contribution of each component, and also validate the robustness of the proposed approach under different lighting conditions.

5.1   Experimental Setup

Hardware configuration. The experimental setup consisted of three main components: scenes displayed on a screen, a jointly-optimized binary amplitude mask, and a CMOS image sensor, as shown in Fig. 5. The image sensor used was a Sony IMX250, located inside a FLIR BFS-U3-51S5C-BD2 camera. It is a 2056×2464 RGGB Bayer sensor, with each pixel being 3.45μm wide. The full sensor array measures approximately 7.1mm×8.5mm. The binary amplitude mask was fabricated on a chrome-coated glass substrate using photolithography. The feature size (width of each unit grid) of the mask was 0.0276mm, around 8 times the pixel width. The mask has 257×308 units, making its size approximately the same as the sensor. The mask-sensor distance is set to 2mm.

Datasets. We utilized extensive public datasets for training and evaluating our proposed methods. For face verification tasks, we used the Asian-Celeb dataset [1] containing 2.8M images from 94K identities for training and used both LFW [20] and FCFD [44] datasets for evaluation. For both datasets, we followed the "unrestricted with labeled outside data" protocol proposed in [19] and presented average accuracy rates on 10-fold cross-validation sets.

The optimization of the optical-electric model during the training phase was conducted entirely via simulation. However, our evaluation approach encompassed both simulated and real-world settings. In real-world evaluations, we capture the displayed facial scene on a screen with a lensless camera.

Evaluation protocols. We tried two evaluation protocols: aligned face evaluation and random face evaluation. In the aligned face evaluation, the algorithm takes the original aligned face in the dataset as input, retaining a constant face height of 27cm and a scene-to-camera distance of 50cm. For the random face evaluation, we further apply randomly perturbation to the facial image, including face center shifting of [-15cm, 15cm], face height scaling ranging from [22cm, 30cm], face rotating within angles [-30°, 30°], and scene-to-camera distance shifting between [40cm, 60cm]. Unlike the aligned face evaluation, the random face evaluation is factored into the scene background because the whole scene is larger than the face[1]. The background dataset BG-20K [26] is used to augment face background. To avoid data leakage, different background images are used for training and evaluation. Some augmentation examples are shown in Fig. 4.

5.2   Implementation Details

We adopted T2T-ViT [48] as the backbone of the face verification model. For efficient training, all lensless captures and optical convolution kernels were resized to 200×240. To simulate the spatial intensity of scenes with RGB images, the face images were rescaled to specific sizes based on the physical settings (face height, scene-to-camera distance, mask-to-sensor distance). with details provided in the supplementary. We trained face verification models using the Adam [24] optimizer with a learning rate 5×10^{-4} until 50 epochs. The batch size is 64×8.

For face-center alignment, we followed ArcFace [11] to normalize faces and defined the centers of normalized faces as the face centers. With the center, we center-cropped the lensless images to 60% of their original size in both width and height. During training, we used the ground truth of the face center for alignment. For evaluation, we trained a separate face-center detection model to output the face-center positions. This model was trained with the optimized mask from the face verification model and the same backbone and training settings.

The augmentation curriculum was implemented during the initial 20 epochs of the 50-epoch training period. The training began with augmentation parameters (shift, rotation, scale, background) identical to those in the aligned face evaluation setting. Over these 20 epochs, we progressively adjusted these parameters to align with those employed in the random face evaluation setting.

To implement cross-modality distillation, we first trained an RGB teacher model with aligned RGB images from the training dataset. Subsequently, we trained the optoelectronic (lensless) model on the same dataset, this time with lensless optics simulation, and incorporated distillation loss to enhance the learning process. It's noted that both the RGB teacher model and the lensless student model share the same model architecture, except for the simulated optics.

---

[1] The field of view of our camera is limited by the $2\theta_{CRA}$ (Chief Ray Angle) of the sensor [7], which is 45° in our experiment. For example, if the scene-to-camera distance $d_{sc} = 50$cm, the width of the whole scene is around $2d_{sc}\tan\theta_{CRA} \approx 42$cm.

Table 1: Decoded Image quality for the varying number of plaintext attacks.

#attacks	10	100	1,000	10,000
PSNR/SSIM	12.1/0.31	13.9/0.39	16.2/0.50	19.1/0.59

Scene   Sensor Capture   Random PSF   Decoded Image
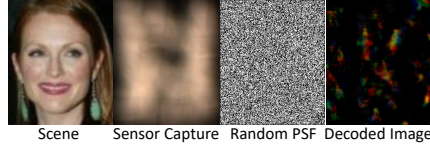
Fig. 7: Example of decoding lensless captures using a random PSF.

Scene   Sensor Capture   Decoded image
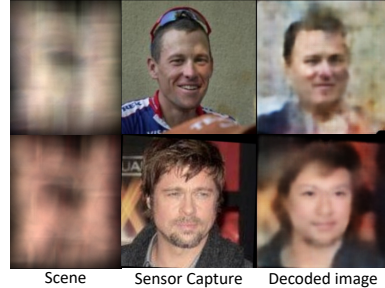
Fig. 8: Decoded examples by decoder trained with 10,000 attack pairs.

5.3 Privacy Protection with Learned Mask

To demonstrate the proposed does protect user privacy, we did the following experiments. First, we showed that the raw lensless captures are unrecognizable to both humans and normal RGB-based face verification models. Fig. 6 presents some visual examples from our lensless camera. The lensless captures of two faces with similar backgrounds appear highly similar to each other (left and middle ones), while captures of the same face under different backgrounds appear quite distinct (left and right ones). This observation indicates that the raw lensless data lack distinguishable cues for human perception of identity, as variations in background overwhelm subtle differences between faces. Quantitatively, a conventional RGB-based face verification model only achieves 51.2% accuracy (50% by random guessing) on lensless captures from the LFW dataset, further validating its potential for preserving privacy.

Moreover, we demonstrate that our lensless imaging system also preserves privacy against reconstruction attacks. We assumed attackers cannot access the camera's point spread function (PSF) [22], and evaluated two attack scenarios: i)**Attackers only access lensless captures only:** In this setup, we utilized 50 random mask patterns as surrogate PSFs and solved the inverse imaging problems with ADMM [8]. The reconstructed images yielded very low average quality, with only 11.7dB PSNR and 0.25 SSIM. Some reconstruction examples are shown in Fig. 7. These results convincingly indicate that random PSF guesses fail to reconstruct the origin scenes meaningfully. ii)**Attackers access to RGB-lensless face pairs**: In this setup, we utilized RGB-lensless pairs, created from the LFW dataset, to train a U-Net [36] as a decoder for simulating reconstruction attacks. The performance of the trained decoder with the varying numbers of pairs, as measured by image quality (PSNR/SSIM) on the left LFW dataset, is summarized in Tab. 1. The results indicate that with fewer than 1,000 pairs, the quality of the reconstructed images is very low, showing that attackers can hardly recover the original scene with a limited number of plaintext attacks. We also show some visual examples of reconstruction results by the decoder trained with 10,000 pairs in Fig. 8. The visual results show that humans can hardly recognize facial identity, further demonstrating that privacy protection ability.

Table 2: Comparison with lensless face verification methods. Input type refers to the input data type for face verification models.

	Methods	WebCam†	FlatCam	Ours	FlatCam	FlatVer	FlatDCT	Ours
	Input type	Captured RGB	Reconstructed RGB (real)*	Lensless (real)	Reconstructed RGB (sim.)	Lensless (sim.)	Lensless (sim.)	Lensless (sim.)
Datasets	LFW (A)	99.32	-	91.51	92.18	68.13	80.13	**96.78**
	LFW (R)	98.82	-	89.88	90.91	63.80	66.25	**94.98**
	FCFD (A)	98.31	82.15	90.23	87.33	57.35	77.81	**94.34**
	FCFD (R)	97.90	80.75	87.45	85.98	51.05	60.87	**92.81**

†Upper Bound.   * We use the published real-captured dataset [44] for evaluation.

5.4 Comparison with Face Verification Methods

To evaluate the face verification performance with the lensless camera under both constrained and unconstraint conditions, we compare our methods with other state-of-the-art face verification methods under two protocols: aligned face evaluation (A) and random face evaluation (R).

Lensless face verification methods can be divided into two classes: 2-step reconstruct-and-verify and 1-step verify from the lensless capture. (1) For the 2-step method, *WebCam* is a traditional RGB-based face verification method. *FlatCam* [44] reconstructs the face images from encoded images captured by flatcam [4], one type of lensless camera, and then verifies faces on reconstructed images. For 2-step random face evaluation, face detection [33] and alignment [9] are conducted before verification. (2) For the 1-step method, *FlatVer* is a network that directly verifies faces on raw flatcam data without reconstruction. *FlatDCT* [14] is a method that transforms the flatcam captures into a multi-resolution DCT subband representation and then verifies faces on this DCT representation. The training of 1-step networks uses a similar augmentation as in our random face evaluation. For a fair comparison, we reproduce all these methods using the same backbone as our methods and we use the same training dataset to train the RGB teacher in our distillation method, without using extra training data.

Results in Tab. 2 show that our method outperforms both the 2-step and 1-step methods by a significant margin, on both synthetic and real data. The performance gap between the simulated and the real experiment is caused by various factors, including inconsistencies in the image processing pipeline, noise in ambient lighting, inaccuracies in the physical simulation model, and errors in the hardware setup parameters. This performance gap is common in lensless experiments [23, 38], and our gap is in an acceptable region. Even with this performance gap, our model on real images outperforms *FlatVer* and *FlatDCT* on synthetic ones and achieves better performance with *FlatCam* on real images. All these results validate the effectiveness of our method.

5.5 Ablation Study

We investigate the effectiveness of each component in our method under two evaluation settings: aligned face evaluation and random face verification.

In the aligned face evaluation, we first show how mask optimization generates better masks compared with randomly initialized baselines, and then

Table 3: Ablation studies of (a) aligned and (b) random face evaluation.

(a) Aligned face evaluation.

Methods		Datasets	
Mask optim.	Cross-modal distillation	LFW (A)	FCFD (A)
×	×	68.77	63.15
×	✓	92.70	88.18
✓	×	71.33	64.82
✓	✓	**96.78**	**94.34**

(b) Random face evaluation.

Methods				Datasets	
Mask optim.	Face-center alignment	Augmentation curriculum	Cross-modal distillation	LFW (R)	FCFD (R)
×	✓	✓	✓	91.36	82.43
✓	×	✓	✓	84.77	79.95
✓	✓	×	✓	83.65	80.32
✓	✓	✓	×	73.45	59.97
✓	✓	✓	✓	**94.98**	**92.81**

Random Initial Mask   Optimized Mask
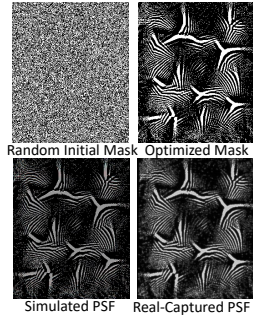
Simulated PSF   Real-Captured PSF

Fig. 9: Top: Mask shape before and post optimization. Bottom: Simulated PSF is close to the real-captured PSF.
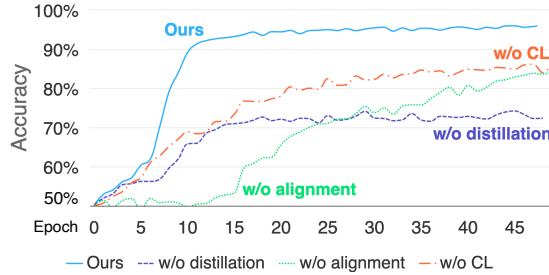

Fig. 10: Comparison of test accuracy of our method against ablated variations (CL refers to curriculum learning) on LFW (R) dataset at different training epochs.

demonstrate the effectiveness of distillation. Fig. 9 illustrates how the mask optimization changes its shape and corresponding PSF. Tab. 3(a) shows that mask optimization brings 3-4% performance boosting and the the proposed distillation significantly boosts the accuracy by about 25%.

In the random face evaluation, we ablate all four key components: 1) mask optimization, 2) face-center alignment, 3) augmentation curriculum, and 4) cross-modality distillation. Tab. 3(b) shows that all four components have positive contributions, with the distillation being the most important one. Interestingly, even with a random unlearned mask, the other components still provide reasonable face verification results. We also visualize the epoch-validation accuracy curves of our procedures in Fig. 10, when individual components are omitted.

This experiment demonstrated that naively training for our end-to-end pipeline may result in poor performance, as we claimed in the introduction. All four key components are important to the robustness of our system. Without any one part, the accuracy drops below 85%, making the whole system infeasible.

5.6   Discussion about Cross-modality Distillation

Our ablation study indicates that cross-modality distillation is a critical component for optimization. To demonstrate this, we visualize the facial features of ten randomly selected individuals from the FCFD (R) dataset using t-SNE [27] for dimensionality reduction. We compare these visualizations between models with and without cross-modality distillation. Fig. 11a shows that cross-modality
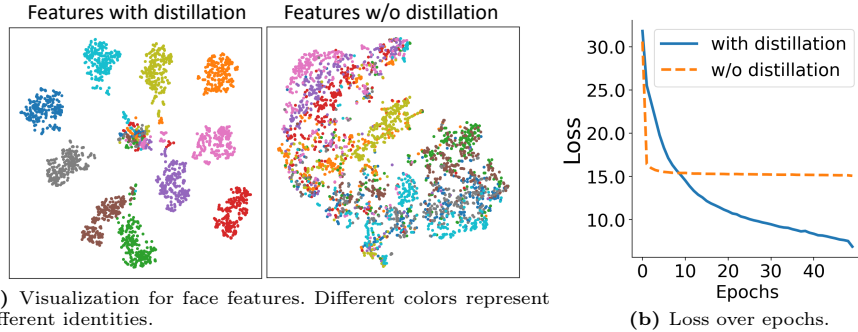
(a) Visualization for face features. Different colors represent different identities.

(b) Loss over epochs.

Fig. 11: Ablation study of distillation. (a) Features for different identities are better grouped after distillation. (b) Distillation improves the convergence.

distillation effectively separates features from different identities, enhancing the performance of face verification tasks. Furthermore, we plot the face verification loss curves for both models during training, as depicted in Fig. 11b. The comparison reveals that the model lacking distillation tends to converge to a local minimum with poor generalization capabilities. In contrast, incorporating distillation allows the model to avoid these local minima during optimization.

Our distillation, coupled with the novel augmentation and alignment, enables our model to learn transformation-invariant features, evidenced by consistent performance in both aligned and random evaluations in Tab. 2. In contrast, relying on distillation and normal augmentation alone, alternative approaches such as FlatVer and FlatDCT [14] show limited accuracy, achieving only 75.6% and 78.4% on LFW(R) and 66.0% and 72.6% on FCFD(R) respectively.

5.7   Robustness to Varying Lighting Conditions

Our lensless face verification system can handle diverse lighting conditions because the training simulation incorporates various noise levels. Specifically, we introduce noise with an average signal-to-noise ratio (SNR) of 30dB during training using the camera noise model [13]. On the LFW(A) dataset, the verification accuracy under SNR 20dB, 25dB, 30dB, 35dB, and 40dB are 94.20%, 96.03%, 96.78%, 96.62%, and 96.68%, respectively. This result shows the robustness of our model against different lighting conditions. Furthermore, optimizing the electronic part of our lensless model for low-light conditions (SNR=20dB) can further improve the accuracy (from 94.20% to 96.02%). This optimization further demonstrates the adaptability and efficiency of our approach in challenging lighting scenarios. More details can be found in the supplementary.

6   Conclusion and Limitation

We introduce an innovative end-to-end optimization approach for robust face verification that operates directly on raw lensless captures. This method enhances privacy, accuracy, and processing speed, representing a significant advancement over traditional two-stage lensless face verification. We validate our approach through simulations and further confirm its effectiveness with a real-world lensless camera prototype. When compared with traditional RGB-based face verifications, the lensless face verification system provides inherent hardware-level

privacy protections with a small, efficient, and low-cost device and presents an attractive performance tradeoff. One limitation is that there is still a performance gap between simulations and real-world environments, as one of the common problems of lensless networks trained on simulated data.

References

1. http://trillionpairs.deepglint.com/ 9
2. Adams, J.K., Boominathan, V., Avants, B.W., Vercosa, D.G., Ye, F., Baraniuk, R.G., Robinson, J.T., Veeraraghavan, A.: Single-frame 3d fluorescence microscopy with ultraminiature lensless flatscope. Science advances **3**(12), e1701548 (2017) 3
3. Antipa, N., Kuo, G., Heckel, R., Mildenhall, B., Bostan, E., Ng, R., Waller, L.: Diffusercam: lensless single-exposure 3d imaging. Optica **5**(1), 1–9 (2018) 1, 3
4. Asif, M.S., Ayremlou, A., Sankaranarayanan, A., Veeraraghavan, A., Baraniuk, R.G.: Flatcam: Thin, lensless cameras using coded aperture and computation. IEEE Transactions on Computational Imaging **3**(3), 384–397 (2016) 1, 3, 4, 12
5. Bezzam, E., Vetterli, M., Simeoni, M.: Privacy-enhancing optical embeddings for lensless classification. arXiv preprint arXiv:2211.12864 (2022) 1, 4
6. Boominathan, V., Adams, J.K., Robinson, J.T., Veeraraghavan, A.: Phlatcam: Designed phase-mask based thin lensless camera. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(7), 1618–1629 (2020) 1, 3, 4
7. Boominathan, V., Robinson, J.T., Waller, L., Veeraraghavan, A.: Recent advances in lensless imaging. Optica **9**(1), 1–16 (2022) 3, 10
8. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning **3**(1), 1–122 (2011) 11
9. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1021–1030 (2017) 12
10. Chi, W., George, N.: Optical imaging with phase-coded aperture. Optics Express **19**(5), 4294–4300 (2011) 3
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019) 5, 9, 10
12. Goodman, J.W.: Introduction to Fourier optics. Roberts and Company publishers (2005) 4
13. Hasinoff, S.W.: Photon, poisson noise. Computer Vision, A Reference Guide **4**(16), 1 (2014) 14
14. Henry, C., Asif, M.S., Li, Z.: Privacy preserving face recognition with lensless camera. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 4, 12, 14
15. Henry, C., Kathariya, B., Asif, M.S., Li, Z., York, G.: Aerial image classification through thin lensless camera. In: 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval. pp. 27–30. IEEE (2022) 4
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 8
17. Hua, Y., Nakamura, S., Asif, M.S., Sankaranarayanan, A.C.: Sweepcam—depth-aware lensless imaging using programmable masks. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(7), 1606–1617 (2020) 4

18. Huang, G., Jiang, H., Matthews, K., Wilford, P.: Lensless imaging by compressive sensing. In: IEEE International Conference on Image Processing. pp. 2101–2105. IEEE (2013) 3

19. Huang, G.B., Learned-Miller, E.: Labeled faces in the wild: Updates and new reporting procedures. Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep **14**(003) (2014) 9

20. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007) 3, 9

21. Ishii, Y., Sato, S., Yamashita, T.: Privacy-aware face recognition with lensless multi-pinhole camera. In: European Conference on Computer Vision Workshops. pp. 476–493. Springer (2020) 4

22. Ishii, Y., Sato, S., Yamashita, T.: Privacy-aware face recognition with lensless multi-pinhole camera. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020 Workshops. pp. 476–493. Springer International Publishing, Cham (2020) 11

23. Khan, S.S., Sundar, V., Boominathan, V., Veeraraghavan, A., Mitra, K.: Flatnet: Towards photorealistic scene reconstruction from lensless measurements. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(4), 1934–1948 (2020) 4, 12

24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10

25. Kuo, G., Liu, F.L., Grossrubatscher, I., Ng, R., Waller, L.: On-chip fluorescence microscopy with a random microlens diffuser. Optics Express **28**(6), 8384–8399 (2020) 1, 3

26. Li, J., Zhang, J., Maybank, S.J., Tao, D.: Bridging composite and real: towards end-to-end deep image matting. International Journal of Computer Vision **130**(2), 246–266 (2022) 10

27. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 13

28. Nguyen Canh, T., Nagahara, H.: Deep compressive sensing for visual privacy protection in flatcam imaging. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 4

29. Pan, X., Chen, X., Nakamura, T., Yamaguchi, M.: Incoherent reconstruction-free object recognition with mask-based lensless optics and the transformer. Optics Express **29**(23), 37962–37978 (2021) 4

30. Pan, X., Chen, X., Takeyama, S., Yamaguchi, M.: Image reconstruction with transformer for mask-based lensless imaging. Optics Letters **47**(7), 1843–1846 (2022) 4

31. Pan, X., Nakamura, T., Chen, X., Yamaguchi, M.: Lensless inference camera: incoherent object recognition through a thin mask with lbp map generation. Optics Express **29**(7), 9758–9771 (2021) 4

32. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019) 8, 9

33. Qi, D., Tan, W., Yao, Q., Liu, J.: Yolo5face: why reinventing a face detector. In: European Conference on Computer Vision. pp. 228–244. Springer (2022) 12

34. Reddy, D., Veeraraghavan, A., Chellappa, R.: P2c2: Programmable pixel compressive camera for high speed imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 329–336. IEEE (2011) 4

35. Rick Chang, J., Li, C.L., Poczos, B., Vijaya Kumar, B., Sankaranarayanan, A.C.: One network to solve them all–solving linear inverse problems using deep projection

models. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5888–5897 (2017) 4

36. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 11

37. Satat, G., Tancik, M., Raskar, R.: Lensless imaging with compressive ultrafast sensing. IEEE Transactions on Computational Imaging **3**(3), 398–407 (2017) 3

38. Shi, W., Huang, Z., Huang, H., Hu, C., Chen, M., Yang, S., Chen, H.: Loen: Lensless opto-electronic neural network empowered machine vision. Light: Science & Applications **11**(1), 121 (2022) 4, 12

39. Shimano, T., Nakamura, Y., Tajima, K., Sao, M., Hoshizawa, T.: Lensless light-field imaging with fresnel zone aperture: quasi-coherent coding. Applied optics **57**(11), 2841–2850 (2018) 3

40. Sitzmann, V., Diamond, S., Peng, Y., Dun, X., Boyd, S., Heidrich, W., Heide, F., Wetzstein, G.: End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. ACM Transactions on Graphics (TOG) **37**(4), 1–13 (2018) 4, 5

41. Stork, D.G., Gill, P.R.: Lensless ultra-miniature cmos computational imagers and sensors. Proc. Sensorcomm pp. 186–190 (2013) 3

42. Taheri, M., Mozaffari, S., Keshavarzi, P.: Cancelable face verification using optical encryption and authentication. JOSA A **32**(10), 1772–1779 (2015) 4

43. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1701–1708 (2014) 7

44. Tan, J., Niu, L., Adams, J.K., Boominathan, V., Robinson, J.T., Baraniuk, R.G., Veeraraghavan, A.: Face detection and verification using lensless cameras. IEEE Transactions on Computational Imaging **5**(2), 180–194 (2018) 4, 5, 9, 12

45. Wang, Z.W., Vineet, V., Pittaluga, F., Sinha, S.N., Cossairt, O., Bing Kang, S.: Privacy-preserving action recognition using coded aperture videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 4

46. Wetzstein, G., Ozcan, A., Gigan, S., Fan, S., Englund, D., Soljačić, M., Denz, C., Miller, D.A., Psaltis, D.: Inference in artificial intelligence with deep optics and photonics. Nature **588**(7836), 39–47 (2020) 4

47. Wu, M.H., Lee, Y.T.C., Tien, C.H.: Lensless facial recognition with encrypted optics and a neural network computation. Applied Optics **61**(26), 7595–7601 (2022) 4

48. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 558–567 (2021) 10

49. Yuan, X., Pu, Y.: Parallel lensless compressive imaging via deep convolutional neural networks. Optics Express **26**(2), 1962–1977 (2018) 4

50. Zomet, A., Nayar, S.K.: Lensless imaging with a controllable aperture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 339–346. IEEE (2006) 4