Lensless Imaging with Focusing Sparse URA Masks in Long-Wave Infrared and its Application for Human Detection

Ilya Reshetouski, Hideki Oyaizu, Kenichiro Nakamura, Ryuta Satoh, Suguru Ushiki, Ryuichi Tadano, Atsushi Ito, and Jun Murayama

Sony Corporation, Tokyo, JAPAN
{Ilya.Reshetouski,Hideki.Oyaizu,Kenichiro.Nakamura,Ryuta.Satoh,
Suguru.Ushiki,Ryuichi.Tadano,Atsushi.C.Ito,Jun.Murayama}@sony.com

Abstract. We introduce a lensless imaging framework for contemporary computer vision applications in long-wavelength infrared (LWIR). The framework consists of two parts: a novel lensless imaging method that utilizes the idea of local directional focusing for optimal binary sparse coding, and lensless imaging simulator based on Fresnel-Kirchhoff diffraction approximation. Our lensless imaging approach, besides being computationally efficient, is calibration-free and allows for wide FOV imaging. We employ our lensless imaging simulation software for optimizing reconstruction parameters and for synthetic image generation for CNN training. We demonstrate the advantages of our framework on a dual-camera system (RGB-LWIR lensless), where we perform CNN-based human detection using the fused RGB-LWIR data.

Keywords: Lensless imaging, long-wave infrared (LWIR) imaging, diffractive optics, image reconstruction, diffraction simulation, pedestrian detection, human detection, visible-infrared image fusion, Faster R-CNNs

1 Introduction

Lensless cameras have been extensively used in X-ray and γ-ray imaging where focusing is not feasible [11], [22]. These cameras are based on the principle that the signal from the scene is encoded by specially designed aperture such that it can be successfully decoded from the sensor image. The reconstruction quality of the coded aperture imaging system is crucially dependent on the underlying aperture mask design, and the apertures based on Uniformly Redundant Arrays (URA) have theoretically the best imaging properties [14].

The growing demand for smaller and cheaper visible light or infrared cameras brings attention to lensless imaging in this area. Unfortunately, in these wavebands the sizes of sensor pixels are comparable with wavelengths which significantly increases the diffraction contribution. Therefore, it is not feasible to directly use URA or similar types of blocking masks in longer wavelengths [12].

In this paper, we present a lensless imaging framework which adopts the efficient URA coding aperture technology for LWIR. Our framework consists

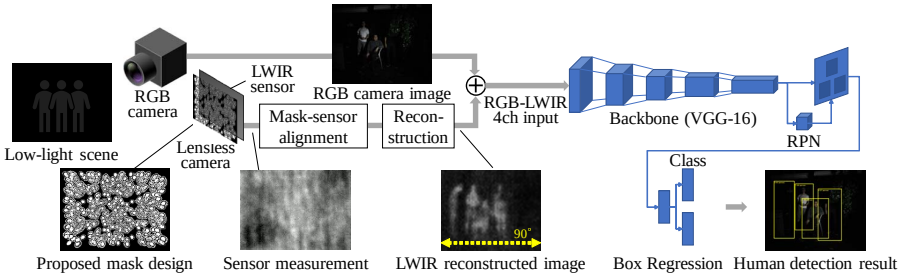


Fig. 1: The architecture of our RGB-LWIR lensless camera system for human detection: The lensless camera is equipped with our focusing sparse URA mask and has horizontal FOV $\approx 90°$. Four channels (RGB and LWIR) are fed into the CNN, and a robust human detection is performed even in low-light conditions.

of novel focusing sparse URA mask design, diffraction-based lensless imaging simulation software, and fast reconstruction algorithm. Our masks features high SNR, wide Field of View (FOV) and low-cost manufacturing. We show that our lensless system produces good quality reconstructions results without hardware-based calibration, but only with lensless imaging simulation. Finally, we demonstrate an application of our lensless framework for CNN-based human detection using fused RGB and LWIR lensless input data, see Fig. 1.

2 Related Work

LWIR imaging: LWIR cameras provide the ability to 'see' the temperature in a scene as an image [17]. Some early works showed that LWIR images can give us better performance on pedestrian detection task, which has been intensively studied in the context of autonomous driving [5], [13]. In [31], LWIR images were superior to visible light images (RGB images) in terms of detection performance. In combination with RGB images, [28] achieved even more improvement than the LWIR images only case. Both owe to the informative expression of human body in LWIR images even in cluttered backgrounds or bad low intensity environments at night in which ordinal RGB cameras will suffer. In response to this trend, several benchmark datasets, which are composed of RGB images synchronized with images of invisible light such as LWIR, were released [3], [21], [23]. They are inviting sophisticated approaches with Convolutional Neural Network (CNN) right now [27], [29], [39], [42].

Despite the benefit described above, there are some constrains when utilising LWIR cameras in practical products. Optical elements for LWIR are generally made with expensive materials like germanium and its compound such as chalcogenide, because of their high transmittance percentages. The challenge for exploring new designs using components made of cheap materials such as silicon or polyethylene has started now. Though these components can be massively replicated by photolithography or molding process, transmissions in the LWIR

spectral band are apparently lower than germanium. Therefore, the allowable thickness of a silicon lens is thinner than a germanium lens for keeping the camera sensitivity. It means that the entire LWIR camera system size is limited by the diameter of silicon lens. It also means that detectors having small pixel number are available. In recent work [20], nearly flat Fresnel lenses on a thin substrate are proposed for improving transmittance percentages. In this paper, to overcome these limitations in LWIR optics, we will take advantage of lensless imaging modality in which compact and cost efficient manufacturing is feasible.

Lensless imaging: Recently, multiple lensless architectures for visible and infrared light diapason have been proposed. These lensless cameras have demonstrated passive incoherent imaging using amplitude masks [2], [24], diffractive masks [18], diffuser masks [1], random reflective surfaces [16], [35], and modified microlens arrays [37].

Many of these solutions have multiple drawbacks including a) narrow FOV and/or large sensor-to-mask distance, usually due to the requirement to have constant PSF [1], [12]; b) reconstruction complexity is too high [25]; c) calibration is difficult and time consumptive [2], [24, 25]; d) calibration required for each unit [1], [25]. These drawbacks limit the application areas of existing lensless systems and make them less attractive for mass-production.

Additionally, the methods above target mostly visible light or short infrared. However, in the case of LWIR, the diffraction blur is stronger (especially for large incident light angles). In [18], even though it was aimed for thermal sensing, such side effect was not treated. We take care of this by incorporating Fresnel Zone Plates (FZP) into our mask design, which will be detailed in Section 3.1. FZP, which consists of concentric transparent and opaque rings (or zones), can be used to focus light and form an image using diffraction [4], [19]. Light, hitting a zone plate, diffracts around the opaque regions and interferes constructively at the focal point. Zone plates can be used in place of pinholes or lenses to form an image. One advantage of zone plates over pinholes is their large transparent area, which provides better light efficiency. In contrast with lenses, zone plates can be used for imaging wavelengths where lenses are either expensive or difficult to manufacture [10], [26].

3 Proposed Method

3.1 Mask Design

Ultimately, we want to bring exceptional reconstruction properties of URA coded apertures to lensless imaging in LWIR. Additionally, since focusing in LWIR is much simpler than in X-ray diapason, instead of simple blocking light at the opaque mask features, we can manipulate it more wisely.

Focusing Sparse URA Mask: Our basic idea is: a) To substitute binary blocking URA mask with an optical element that acts as a lens with multiple sub-lenses with optical centers distributed in the same pattern as the original binary mask and with coincident focal and sensor planes; b) To use sparse URA

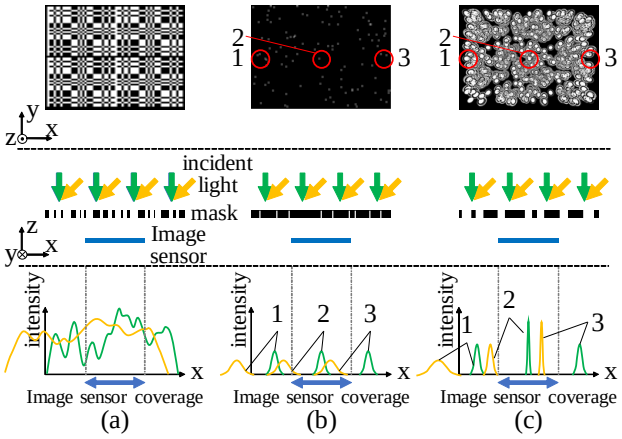


Fig. 2: Projection patterns for different mask designs: (a) MURA, (b) Sparse URA as pinhole array and (c) Sparse URA as FZP array (proposed)
The mask pattern (c) resolves the drawbacks in (a) and (b) by increasing image sharpness while maintaining a large amount of light.

masks in order to reduce the diffraction crosstalk between transparent features of the mask and to increase average focusing area for each feature. In other words, sparsity allows each sub-lens to have a much larger numerical aperture than the transparent feature of the original binary mask.

The advantages of a) and b) are illustrated on the Fig. 2. Here, three different mask patterns are shown in the top row. The middle row depicts a side view of each mask together with the image sensor and two different directions of the incident light (green and yellow arrows). One dimensional charts in the bottom row indicate the intensity of projected light for the two incident light directions respectively. Note that for simplicity, projection from only three selected mask locations is described in (b) and (c). (a) Modified Uniformly Redundant Array (MURA) mask or similar kind of pseudo-random mask, which is commonly used for lensless imaging. In this case, the shadow of the mask is significantly blurred due to the diffraction. For the same reason, for incident light with large-angle (yellow line), the shadow of the mask is blurred even stronger. (b) Pinholes arrayed in a sparse URA pattern make a sharper projection on the sensor surface due to the smaller diffraction crosstalk between holes. Nevertheless, the pinhole array has two major drawbacks: less amount of light and still relatively large blur caused by diffraction. (c) is our proposed method: Focused with FZP sparse URA pattern. This mask pattern resolves the drawbacks in (a) and (b) by increasing image sharpness while maintaining a large amount of light.

Optical Axes Tuning: We aim to create a system with wide FOV. Substituting the standard binary mask with lenses with multiple optical centers helps to increase Signal-to-Noise-Ratio (SNR), but, to maintain the quality over entire FOV, each sub-lens should have a sharp focus for all meaningful light directions.

Note that for a typical cyclic aperture system, the variation of scene directions passing through a transparent element depends on its position. In particular, the closer the element to the edges of the mask, the smaller is the variation, see Fig. 3. The figure shows the difference between meaningful incident light directions variation angles depending on the optical element position on the mask. These light directions are restricted by the directions (the red dotted lines) corresponding to the maximum meaningful angle defined by the relationship between the edge of the mask and that of the Focal Plane Array (FPA). The meaningful incident light variation angle from FZP B to the FPA is θ_B, and that from FZP C is θ_C. We use this observation to optimize the average sharpness of each sub-lens by aligning its optical axis with its average incoming light direction. Schematic result of the axes adjustment is illustrated in Fig. 2: for non-optimized mask (b), in the sensor area, the yellow signal is more blurred and less strong than the green signal, while for optimized mask (c) the yellow and green signals are equally sharp. Note that the shapes of FZPs in the center and near the edge of our prototype mask, as shown Fig. 4, are different for this purpose.

3.2 Scene Reconstruction

Our lensless imaging method is designed to project an $M \times N$ scene on the $M \times N$ sensor. We assume, that the imaging is linear and can be described as:

$$y = Fx + n \tag{1}$$

where F - an $(MN) \times (MN)$ real-valued matrix, $x \in R^{MN}$ - 1D column vector representing the scene intensities, $y \in R^{MN}$ - 1D column vector of the sensor measurements, $n \in R^{MN}$ - 1D column vector of the noise.

Ideally, if each scene direction can be precisely focused onto a set of uniformly bright sharp spots arranged as sparse URA pattern, we can use very efficient reconstruction algorithms [15] with computational complexity $O((MN)\log(MN))$. In this case $F = U$ - imaging matrix for ideal URA coded aperture system. However, since our implementation is based on blocking-type masks, sharp uniform focusing of light from a relatively broad spectrum over a large range of incoming light directions is problematic. Therefore, the direct application of reconstruction methods designed for URA masks is prone to artifacts.

One way to improve the reconstruction quality is to use more general reconstruction techniques, but they often require significantly larger computational and/or memory expenses. For example, scene reconstruction can be done with regularized matrix inversion approach

$$\hat{x} = F^* y \tag{2}$$

where F^* - an $(MN) \times (MN)$ real-valued matrix, calculated with, for example, Tikhonov regularization method [38]. But this method requires $O((MN)^3)$ operations to multiply F^* with y and sufficient memory to store F^*.

To balance the quality, speed and memory requirements we propose to approximate F^* with U^*:

$$U^* = \alpha U^{-1} \beta + u^T v \tag{3}$$

$$\|U^* - F^*\| \xrightarrow[\alpha,\beta,u,v]{} \min \tag{4}$$

where α, and β - $(MN) \times (MN)$ diagonal matrices and u^T, $v \in R^{MN}$ - 1D row and column vectors correspondingly, norm here is Frobenius matrix norm. Note, that the computational complexity of the calculation of the product $U^* y$ using (3) is the same as $U^{-1} y$ (i.e. URA reconstruction), can be performed with fast Fourier transform (FFT) and doesn't require explicit calculation or storage of U^* or U^{-1}. For more details regarding are computation of $U^* y$ please refer to the Supplementary Materials.

3.3 Diffraction Simulation for Imaging Matrix Estimation

As was mentioned in Section 3.2, due to drawbacks of binary FZP-type focusing, the imaging matrix F is not exactly URA, and therefore it has to be evaluated.

One way to estimate F is to perform calibration step, for example by capturing point light source or set of patterns. However, this method is not practical. Instead, we perform an estimation of F using diffraction simulation based on known mask design and mask to sensor distance.

Since our mask consists of multitude of FZP's with diameters not much smaller than the gap between the mask and the sensor, we decided to employ the Fresnel-Kirchhoff diffraction model [6]:

$$U(P) = -\frac{ia}{2\lambda} \int_M \frac{e^{ik(r+s)}}{rs} \left[cos(\alpha_{n,r}) - cos(\alpha_{n,s})\right] dM \tag{5}$$

where $U(P)$ - complex amplitude at the point P on the sensor, a - magnitude of the scene point, λ - wavelength, M - transparent area of the mask, k - wavenumber, r - distance from the scene point to the mask element, s - distance from the mask element to the sensor point, $(\alpha_{n,r})$ and $(\alpha_{n,s})$ - angle between the normal to the mask and the direction from the scene (sensor) point to the mask element.

To estimate imaging matrix F, we performed simulation of the sensor image from each scene point, see details in the implementation Section 4.4.

4 Implementation and Results

4.1 System Overview

Our RGB-LWIR lensless camera is equipped with a LWIR sensor (PICO384 Gen2$^{\text{TM}}$, spectral range: 8-14μm, resolution: 384×288, pixel pitch: 17μm), and with RGB camera (resolution: 1920×1080, horizontal FOV: 84°), see Fig. 5 (a)-(b). These two cameras were aligned to have similar scene view.

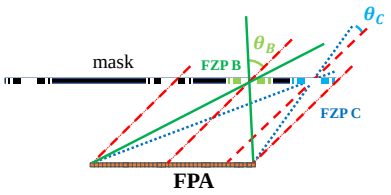


Fig. 3: θ_B is a meaningful incident light variation angle from FZP B, and θ_C is that from FZP C.

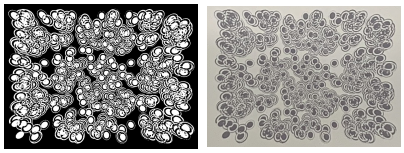


Fig. 4: Our mask design (*left*) and a photo of a real mask on Si wafer (*right*)

4.2 Mask Prototyping

Because irradiance received by the FPA falls off with an incoming angle, to record a reasonable range of sensor values from the entire FOV, we chose 90° as FOV. Since the final URA mask is twice bigger than the sensor, to have FOV 90°, the distance between the mask and the sensor should be 3.2 mm. Correspondingly, given the wide FOV and relatively large spectral range of the bolometer, to maintain uniform focusing, we choose to use FZPs with maximum 3 zones. This limits the output resolution to approximately 60μm.

Therefore, we choose sparse Singer URA mask with parameters $q = 107$ and $w = 3$, see [7]. The selected mask has basic resolution 127×91, only 0.9% of transparent features and matches the resolution of the bolometer after 3×3 pixel averaging (effective pixel size $17 \times 3 = 51\mu$m). Final sparse URA mask was generated as 2×2 mosaic of the basic URA pattern and had resolution 253×181.

For each transparent feature of the final URA mask we calculated average incoming light direction, see Section 3.1, and generated corresponding FZP using interference model. In case if one or more FZPs partially overlap, we resolved situation with greedy approach maximizing total focused light. Finally, to reduce variation of intensities of the focused mask, we performed iterative reduction (0.1 of the zone thickness) of the FZP diameter for the most bright spots.

Fig. 4 shows our mask design and the photo of the prototype. The mask was produced by chrome etching on an optical grade silicon wafer.

4.3 Implementation of the Scene Reconstruction

Our reconstruction software consists of a capturing part and a scene reconstruction part. The capturing part acquires RAW LWIR images and applies a two-points Non-Uniformity Correction (NUC) to the LWIR data in a similar manner as described in [32]. To measure NUC parameters we used an extended area blackbody radiation source. Before reconstructing lensless images, we estimated the parameters of the model, described in equation (3)-(4). In order to do so, we performed the following steps:

1. Calculated the imaging matrix F $[(127 * 91) \times (127 * 91)]$, see Section 3.2, for our mask design with our diffraction simulation software, see Section 4.4.

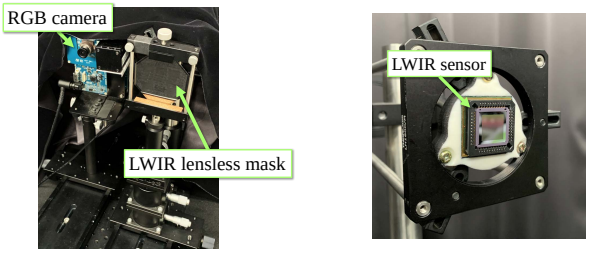


Fig. 5: Our prototype system: An RGB camera and an LWIR lensless camera are placed in parallel (*left*). The bare LWIR sensor is placed behind a mask (*right*).

2. Calculated the regularized inverse matrix F^* using Tikhonov regularization.
3. Calculated the theoretical inverse matrix U^{-1} for our URA mask.
4. Optimized the equation (4) with BFGS minimization algorithm and received approximation of the parameters of the model $\hat{\alpha}$, $\hat{\beta}$, $\hat{u}$, $\hat{v}$.

Finally, to recover a scene x from lensless LWIR image y, we calculated U^*y using FFT after expanding the equation (3) with the parameters $\hat{\alpha}$, $\hat{\beta}$, $\hat{u}$, $\hat{v}$.

Fig. 6 illustrates various examples of reconstructions. Images (a) - (d) show performance of our method for scenes with people. Note, that the results are sharp enough to distinguish poses and other details. (e) shows behaviour of our solution on the edge of the FOV. (f) shows the result with hot objects (half-empty cup with hot coffee and bottles with hot water).

We analyzed the performance of our lensless prototype by measuring its modulation transfer function (MTF) for different angles of incidence. For comparison, we performed MTF evaluation for a simple LWIR camera with a molded-in plastic Fresnel lens (FL = 0.0094m) and for the standard lensless camera with a 37x37 MURA mask (mask distance = 0.002m), see Fig. 7. By comparing the results in Fig. 7, we can conclude that our system has relatively uniform quality for wide FOV ($2 \times 42°$), it outperforms MURA lensless camera and better than Fresnel lens camera when incident angle is sufficiently large (more than 12°).

4.4 Implementation of the Imaging Matrix Simulation

We evaluated imaging F by simulating sensor image from each scene point for multiple wavelengths in the range of 8-14 μm. The final result was calculated as the weighted average of simulations for different wavelengths with weights corresponding to the sensor spectral response multiplied with expected scene spectral profile. In our implementation, we used sensor spectral response provided by the sensor manufacturer and the scene spectral profile was chosen to be the blackbody spectral profile at the average human body temperature $\approx 309K$.

The evaluation of the imaging matrix F using Fresnel-Kirchhoff integration is computationally expensive; however, we need to perform it only once. Additionally, images from each scene point and for each wavelength can be calculated

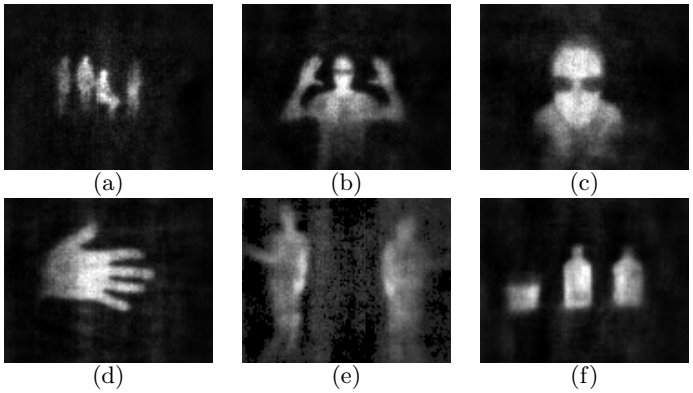


Fig. 6: Examples of reconstructed image

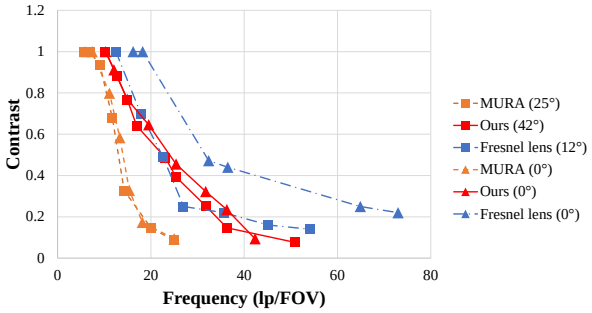


Fig. 7: Measured MTF's for different angles of incidence for Fresnel lens (*blue lines*), MURA mask (*orange lines*) and our prototype (*red lines*)

independently. We implemented our simulation software with C++/CUDA. It took approximately one week to estimate the F-matrix for our prototype on a machine with a single GPU (NVIDIA GeForce RTX 2080 Ti). Fig. 8 shows the comparison between real and simulated images of a point light source.

4.5 Robustness to the Alignment Error

As we explained in the previous section, position alignment between a sensor and a mask is important, especially for manufacturing real products. So we should know how large misalignment is acceptable for reconstructed image quality. We evaluated quantitative relationships between sensor-mask misalignment and reconstructed image quality. We measured a lace curtain before a blackbody as a thermally textured object from two different positions. Then we evaluated them by Complex Wavelet Structural Similarity (CW-SSIM) [40] index. Its insensibility to spatial translation and size difference is suitable for our evaluation. We measured parallel shift (ΔX), gap distance shift (ΔZ), and roll ($\Delta\theta$). Fig. 9 is

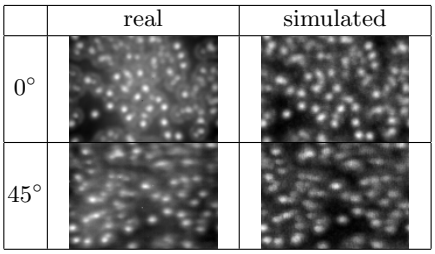


Fig. 8: Comparison between the real and the simulated sensor images of a point light source at 0° and 45° incidence angles: We could acquire assumed sensor image even in the corner of the image.

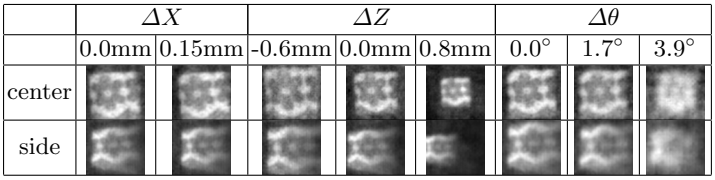
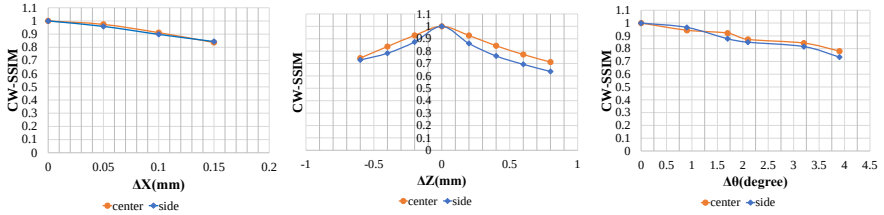


Fig. 9: Reconstructed samples in the robustness evaluation for misalignment

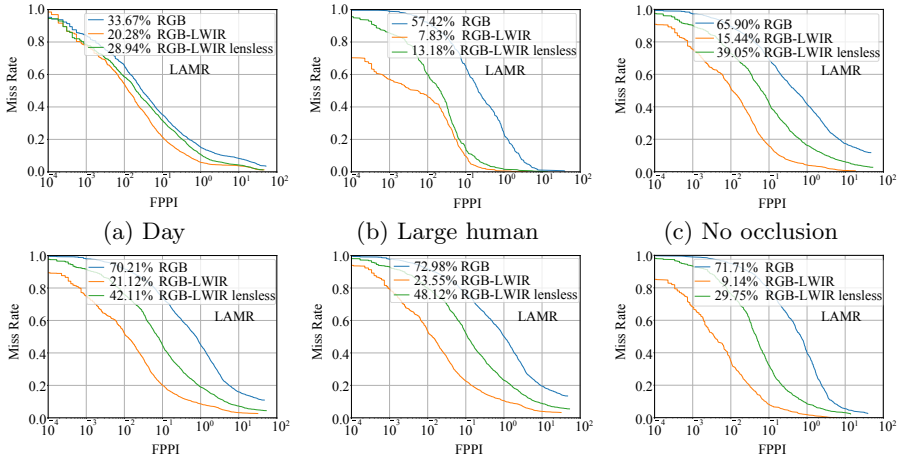


(a) ΔX (b) ΔZ (c) $\Delta\theta$

Fig. 10: Evaluation of misalignment error of our mask in CW-SSIM index: (a) parallel position misalignment, (b) gap misalignment and (c) roll misalignment

a sample of reconstructed images. We can see that the larger misalignment, the more degradation of a lace pattern. Fig. 10 is graphs of misalignment values and CW-SSIM values. In general, it is said that if the CW-SSIM value is larger than 0.9, the compared image quality can be treated as good. By these criteria, the allowable value of ΔX, ΔZ, and $\Delta\theta$ are $100\mu m$, $150\mu m$, and 1.5° respectively.

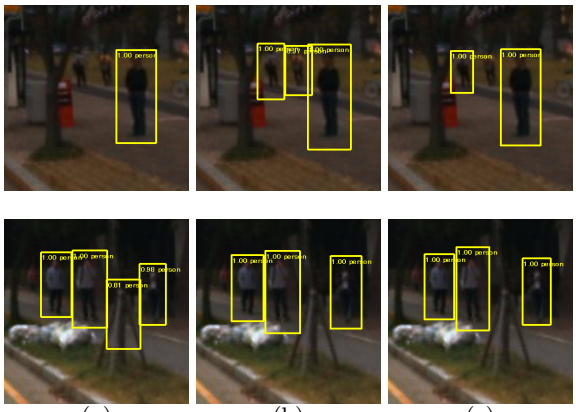
These values are much larger than the manufacturing accuracy, and therefore we can conclude that per-unit mask calibration will not be required in the production of our lensless camera.



(a) Day (b) Large human (c) No occlusion

(a') Night (b') Small human (c') Partial occlusion

Fig. 11: Comparison of human detection miss rate by LAMR: (a)(a') day and night scenes, (b)(b') human size in the scenes and (c)(c') different scenes about human occluded rates



(a) (b) (c)

Fig. 12: Examples of false-negatives and false-positives in synthetic evaluation: (a) detected result from RGB image, (b) RGB-LWIR and (c) RGB-LWIR lensless

5 Experiments in Inference Task

Human detection is still a current challenge in computer vision research as it is an essential task for the safety of the transportation system, including pedestrian detection in automobile applications.

5.1 Human Detection Algorithm Pipeline

For human body detection experiments, we chose a Faster R-CNN model [33] as a baseline because it is most commonly used in the image-based object detection field. The network architecture which we used is shown in Fig. 1. The Faster R-CNN has a two-stage structure with a backbone network (VGG16) followed by a Region Proposal Network (RPN) and a classifier. For the RGB-LWIR fusion, we modified the first convolution layer of the backbone CNN to accept 4ch input and fed additional LWIR plane along with RGB planes.

5.2 Performance Evaluation with Synthetic Dataset

Experimental Setup: To show the effectiveness of our LWIR lensless imaging system in an inference task, we used an open dataset i.e. KAIST Multispectral Pedestrian Detection Benchmark dataset [21] as a base for synthetic evaluation. It contains aligned RGB and LWIR images of day and night traffic taken from a vehicle. To retrieve people's images efficiently, we cropped two regions out from the lower half of each original 640x480 image and resized them to 127x91, which is the image size of our LWIR lensless camera. We took over most of the original annotations with minor modifications: We ignored "cyclist" labels and altered "person" labels to "occluded person" if the person is occluded more than 60% by other objects. Consequently, we prepared 26658 labels from 24719 frames for training, and 32981 labels from 27888 frames for testing. To acquire RGB-LWIR lensless pairs, we simulated a full lensless encoding-decoding cycle and applied it to the selected above 127x91 LWIR images. Our lensless encoding-decoding cycle was organized as follows: first, we produced synthetic lensless sensor images using simulated matrix F, see Section 3.3; second, we added Gaussian sensor noise with $\mu = 0$, $\sigma = 0.14\%$; finally, we reconstructed noisy synthetic sensor images.
Analysis: We created LWIR lensless training data from KAIST dataset by lensless simulation (Section 3.3) and utilized it as 'LWIR lensless' channel for our human detector. We trained the detector with following three configurations: 1) 3ch RGB input, 2) 4ch RGB-LWIR input, and 3) 4ch RGB-LWIR lensless input.

Evaluation results of False Positive Per Image (FPPI) to miss rate and Log Average Miss Rate (LAMR) are shown in Fig. 11. (a) and (a') show the comparison between daylight and night scenes. One can see the LWIR channel improves LAMR in night scenes (a') compared to that of daylight scenes (a). We also evaluated the miss rate between different human scales in (b) and (b'). LAMR naturally improves when detecting larger scale humans, and both LWIR and 'LWIR lensless' channels exhibit higher contribution for the near scale. (c) and (c') are the comparison of detection between no-occluded and partially-occluded person. Both LWIR and 'LWIR lensless' channels consistently maintain good LAMR improvements.

From the above, it is obvious that adding LWIR channel to RGB improves LAMR. Though the contribution of 'LWIR lensless' which was generated by lensless simulation is inferior to that of original LWIR, it is still effective to increase the ability of human detection.

Examples of false-negatives and false-positives are shown in Fig. 12. In the first row, RGB (a) missed some people in rather darker scenes while LWIR (b) and 'LWIR lensless' (c) didn't. On the other hand in the second row, (a) incorrectly detected an object as a person that (b) and (c) didn't.

5.3 Experiment with Real-World Dataset

Experimental Setup: For the data acquisition, we used the RGB-LWIR lensless camera system described in Section 4.1, which simultaneously captures RGB and LWIR lensless images. The next step was to apply a geometrical calibration between an LWIR lensless camera and an RGB camera. Since the resolution of reconstructed LWIR images is relatively low, the classical camera calibration approach [43] is not reliable. Therefore, we calibrated the intrinsic parameters of the RGB camera by applying [43]'s method and of the LWIR camera by using design parameters. Then we calibrated the extrinsic parameters by showing a high-temperature halogen light at two different known positions.
Analysis: We captured real image samples of humans by RGB-LWIR lensless cameras. These images were resized to fit the KAIST dictionary and were evaluated by the human detection algorithm described in Section 5.1. We show the evaluation result of human detection from 531 frame images in Fig. 13. As we wrote before, we can see that the result of the RGB-LWIR lensless camera is better than that of the RGB camera only. We show human detection samples in Fig. 14, which upper side is the results of an RGB-LWIR lensless camera, and the lower side is of the RGB camera. (a) and (a') show the same scene that four persons are standing in a low-light environment. While nobody is detected in (a'), all persons are recognized in (a). (b) and (b') show a backlight situation what is a difficult scene to detect humans for RGB camera. (c) and (c') show an occluded scene. While two persons are heavily occluded by a plate or leaves, and one person is standing in a dark area, all persons are detected in (c). RGB-LWIR lensless camera can detect such an occluded person in the dark by strengthening information about the human body with two different types of images.

6 Conclusion

We have presented a novel lensless imaging framework in the long-wave infrared. We proposed a wide FOV lensless camera system with novel modulation principle and an efficient reconstruction algorithm. With our experimental prototype, we evaluated the reconstruction performance qualitatively and quantitatively. We also have shown the robustness of the misalignment between a mask and a sensor, and the required accuracy is lower than manufacturing accuracy. As a background core technology, we constructed a precise diffraction simulator that can easily create imaging matrix F without calibration. We demonstrated, that our simulator can contribute to generating a massive number of training data for CNN-based inference algorithm and evaluated its performance in human detection tasks. Throughout evaluations with fusion data of LWIR lensless and

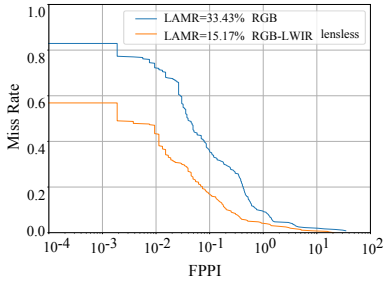


Fig. 13: Evaluation of human detection rate by our RGB-LWIR lensless camera: The result of the RGB-LWIR lensless camera is significantly better than that of the RGB camera only.

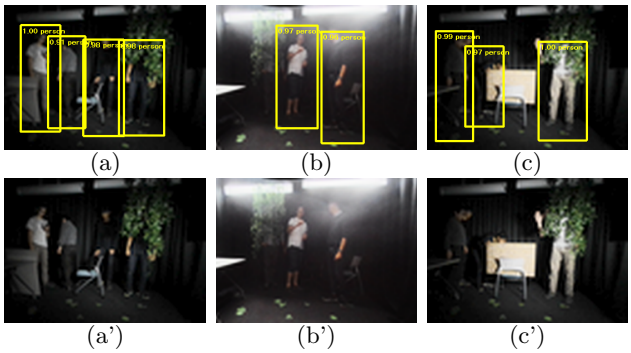


Fig. 14: Comparison of human recognition results between RGB-LWIR lensless images (*upper row*) and only RGB images (*lower row*)

RGB images, we show that performance with our dataset was superior to with just RGB dataset, and was comparable to RGB-LWIR with a lens.

Limitations and Future Work: One limitation is the quality degradation due to difficulty to achieve sharp uniform focusing of all open features of the URA mask with diffractive mask. The solution to this problem could be to substitute FZP-type mask with free-form lens. Another limitation is the mask distance. Reducing the mask distance to the sensor increases the FOV, however, the FOV can't be extended much more than 90° due to the angular intensity fall-off.

Although our mask design provides a good starting point for LWIR lensless imaging, we believe that data-driven approach which proved to be efficient for optics optimization (see [8, 9], [30], [34], [36], [41]) can be coupled with our lensless imaging simulation system for further mask design optimization.

A spectral extension is the most interesting future work. Recently, there are several proposals for improved detectors for not only long-wave infrared but middle-wave infrared or terahertz imaging. We consider that our diffraction simulation framework and focusing modulation can be applied in this direction.

References

1. Antipa, N., Kuo, G., Heckel, R., Mildenhall, B., Bostan, E., Ng, R., Waller, L.: DiffuserCam: lensless single-exposure 3D imaging. Optica **5**(1), 1 (jan 2018)
2. 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 (jul 2016)
3. Barrera Campo, F., Lumbreras Ruiz, F., Sappa, A.D.: Multimodal Stereo Vision System: 3D Data Extraction and Algorithm Evaluation. IEEE Journal of Selected Topics in Signal Processing **6**(5), 437–446 (sep 2012)
4. Barrett, H.H.: Fresnel zone plate imaging in nuclear medicine. Journal of Nuclear Medicine **13**(6), 382–385 (jun 1972)
5. Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 8926, pp. 613–627. Springer Verlag (2015)
6. Born, M., Wolf, E.: Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light. Cambridge University Press, 7th edn. (1999)
7. Busboom, A., Elders-Boll, H., Schotten, H.: Uniformly redundant arrays. Experimental Astronomy **8**, 97–123 (06 1998)
8. Chakrabarti, A.: Learning sensor multiplexing design through back-propagation. Advances in Neural Information Processing Systems pp. 3089–3097 (2016)
9. Chang, J., Wetzstein, G.: Deep optics for monocular depth estimation and 3D object detection. Proc. of Computer Vision and Pattern Recognition (CVPR) pp. 10192–10201 (2019)
10. Chu, Y.S., Yi, J.M., De Carlo, F., Shen, Q., Lee, W.K., Wu, H.J., Wang, C.L., Wang, J.Y., Liu, C.J., Wang, C.H., Wu, S.R., Chien, C.C., Hwu, Y., Tkachuk, A., Yun, W., Feser, M., Liang, K.S., Yang, C.S., Je, J.H., Margaritondo, G.: Hard-x-ray microscopy with Fresnel zone plates reaches 40 nm Rayleigh resolution. Applied Physics Letters **92**(10) (mar 2008)
11. Cieślak, M.J., Gamage, K.A., Glover, R.: Coded-aperture imaging systems: Past, present and future development – A review. Radiation Measurements **92**, 59–71 (sep 2016)
12. DeWeert, M.J., Farm, B.P.: Lensless coded-aperture imaging with separable Doubly-Toeplitz masks. Optical Engineering **54**(2), 023102 (feb 2015)
13. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(4), 743–761 (2012)
14. Fenimore, E.E., Cannon, T.M.: Coded aperture imaging with uniformly redundant arrays. Applied Optics **17**(3), 337 (feb 1978)
15. Fenimore, E.E., Cannon, T.M.: Uniformly redundant arrays: digital reconstruction methods. Applied Optics **20**(10), 1858 (may 1981)
16. Fergus, R., Torralba, A., Freeman, W.T.: Random Lens Imaging. MIT-CSAIL-TR-2006-058 (sep 2006)
17. Gade, R., Moeslund, T.B.: Thermal cameras and applications: A survey. Machine Vision and Applications **25**(1), 245–262 (jan 2014)
18. Gill, P.R., Tringali, J., Schneider, A., Kabir, S., Stork, D.G., Erickson, E., Kellam, M.: Thermal Escher sensors: Pixel-efficient lensless imagers based on tiled optics. In: Optics InfoBase Conference Papers. vol. Part F46-COSI 2017, p. CTu3B.3. OSA - The Optical Society (jun 2017)

19. Goodman, J.W.: introduction to Fourier Optics, 3rd edition. Roberts (2005)
20. Grulois, T., Druart, G., Guérineau, N., Crastes, A., Sauer, H., Chavel, P.: Extra-thin infrared camera for low-cost surveillance applications. Optics Letters **39**(11), 3169 (jun 2014)
21. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proc. of Computer Vision and Pattern Recognition (CVPR). vol. 07-12-June, pp. 1037–1045. IEEE Computer Society (oct 2015)
22. in't Zand, J.: A coded-mask imager as monitor of Galactic X-ray sources. Ph.D. thesis, Space Research Organization Netherlands, Sorbonnelaan 2, 3584 CA Utrecht, The Netherlands (1992)
23. Karasawa, T., Watanabe, K., Ha, Q., Tejero-De-Pablos, A., Ushiku, Y., Harada, T.: Multispectral object detection for autonomous vehicles. In: Proc. of Thematic Workshops of ACM Multimedia. pp. 35–43. Association for Computing Machinery, Inc, New York, New York, USA (oct 2017)
24. Khan, S.S., R, A.V., Boominathan, V., Tan, J., Veeraraghavan, A., Mitra, K.: Towards Photorealistic Reconstruction of Highly Multiplexed Lensless Images. In: Proc. of International Conference on Computer Vision (ICCV). pp. 7859–7868. IEEE (oct 2019)
25. Kim, G., Isaacson, K., Palmer, R., Menon, R.: Lensless photography with only an image sensor. Applied Optics **56**(23), 6450–6456 (aug 2017)
26. Kirz, J.: PHASE ZONE PLATES FOR X RAYS AND THE EXTREME UV. J Opt Soc Am **64**(3), 301–309 (mar 1974)
27. Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., Teutsch, M.: Fully Convolutional Region Proposal Networks for Multispectral Person Detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). vol. 2017-July, pp. 243–250. IEEE Computer Society (aug 2017)
28. Leykin, A., Ran, Y., Hammoud, R.: Thermal-visible video fusion for moving target tracking and pedestrian classification. In: Proc. of Computer Vision and Pattern Recognition (CVPR) (2007)
29. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. In: British Machine Vision Conference (BMVC). vol. 2016-Septe, pp. 73.1–73.13 (2016)
30. Metzler, C.A., Ikoma, H., Peng, Y., Wetzstein, G.: Deep Optics for Single-shot High-dynamic-range Imaging. Proc. of Computer Vision and Pattern Recognition (CVPR) (aug 2020)
31. Miezianko, R., Pokrajac, D.: People detection in low resolution infrared videos. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2008)
32. Mudau, A.E., Willers, C.J., Griffith, D., Le Roux, F.P.: Non-uniformity correction and bad pixel replacement on LWIR and MWIR images. In: Saudi International Electronics, Communications and Photonics Conference (SIECPC) (2011)
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (jun 2017)
34. 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 **37**(4) (2018)

35. Stylianou, A., Pless, R.: SparkleGeometry: Glitter Imaging for 3D Point Tracking. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 919–926. IEEE Computer Society (dec 2016)
36. Sun, Q., Zhang, J., Dun, X., Ghanem, B., Peng, Y., Heidrich, W.: End-to-end Learned, Optically Coded Super-resolution SPAD Camera. ACM Transactions on Graphics **39**(2), 1–14 (2020)
37. Tanida, J., Kumagai, T., Yamada, K., Miyatake, S., Ishida, K., Morimoto, T., Kondou, N., Miyazaki, D., Ichioka, Y.: Thin observation module by bound optics (TOMBO): concept and experimental verification. Applied Optics **40**(11), 1806 (apr 2001)
38. Tikhonov, A.N., Arsenin, V.Y.: Solutions of ill-posed problems. W.H.˜Winston (1977)
39. Wagner, J., Fischer, V., Herman, M., Behnke, S.: Multispectral pedestrian detection using deep fusion convolutional neural networks. In: Proc. of European Symposium on Artificial Neural Networks (ESANN) (2016)
40. Wang, Z., Simoncelli, E.P.: Translation insensitive image similarity in complex wavelet domain. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP). vol. II (2005)
41. Wu, Y., Boominathan, V., Chen, H., Sankaranarayanan, A., Veeraraghavan, A.: PhaseCam3D - Learning Phase Masks for Passive Single View Depth Estimation. IEEE International Conference on Computational Photography (ICCP) (2019)
42. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: Proc. of Computer Vision and Pattern Recognition (CVPR). vol. 2017-Janua, pp. 4236–4244. Institute of Electrical and Electronics Engineers Inc. (nov 2017)
43. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(11), 1330–1334 (nov 2000)