ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10887978

Facial Expression Recognition with DToF Sensing

Chengxiao Li, Xie Zhang, Chenshu Wu
Department of Computer Science, School of Computing and Data Science
The University of Hong Kong, HK SAR, China
{chengxiaoli, zhangxie}@connect.hku.hk, chenshu@cs.hku.hk

***Abstract*—Facial Expression Recognition (FER) is crucial for understanding human emotions, with applications spanning from mental health assessment to marketing recommendation systems. However, existing camera-based methods raise privacy concerns, while RF-based approaches suffer from limited environmental generalizability and high cost. In this work, we propose ToFace, a FER system leveraging a low-cost (4.8$) Direct Time-of-Flight (DToF) sensor that has been available on commodity smartphones. This sensor provides an extremely low-resolution 8×8 depth map and a clear Field of View (FoV), significantly mitigating privacy concerns while avoiding the impact of ambient objects. Despite the benefits, the low-resolution depth map introduces significant challenges for precise expression recognition due to limited facial structure information. We first develop a physical model to extract additional spatial information from the intermediate sensor output, *i.e.*, the transient histograms. We then propose a physics-integrated neural network to reconstruct a facial structure map comprising both depth and orientation for accurate expression recognition. We conduct real-world experiments with 12 users and compare our model with several baselines. The results demonstrate that ToFace achieves the highest recognition accuracy of 75%.**

***Index Terms*—facial expression recognition, direct time-of-flight sensor, sensing AI, internet of things.**

I. Introduction

Facial expression recognition (FER) plays a pivotal role in understanding emotional states [1]–[3]. By analyzing facial expressions, we can assess emotional flexibility and monitor changes in emotions, providing valuable feedback for quantifying customer interest [2] and depression symptoms [3]. FER has been extensively studied, particularly in the computer vision field, with methods such as de-expression learning with a single RGB image [4] and part-based hierarchical bidirectional RNNs for facial sequences analysis [5]. However, vision-based approaches experience performance degradation in low-light environments and raise privacy concerns [6]. Recently, mmFER [7] proposes a millimeter-wave radar-based FER system, which significantly mitigates privacy concerns and the impact of light conditions. However, mmFER [7] suffers from limited environmental generalizability due to multipath interference and high costs, which hinder its widespread deployment. In a nutshell, existing vision-based and RF-based methods face issues including privacy intrusion, environmental vulnerability, and high cost.

To address these issues, we propose ToFace, a privacy-preserving and environmentally robust FER system, with a low-cost (4.8$) Direct Time-of-Flight (DToF) sensor. A DToF sensor is an integrated gride of Single-Photon Avalanche Diodes (SPADs) that estimate the distance between the sensor and target by measuring the time of flight (ToF) of infrared (IR) photons, providing an 8×8 low-resolution depth map. The extremely low-resolution depth map mitigates privacy risks, making them less likely to identify individuals and easier to obtain consent for their use. Additionally, as an active sensor operating in the near-infrared spectrum, DToF can function in the dark, and its clear Field of View (FoV) ($45° \times 45°$) facilitates the mitigation of interference from ambient objects. Although DToF sensing is well-suited for privacy-preserving and environmentally robust FER, the limited facial structure information in the low-resolution depth map [8], [9] poses challenges for both face detection and expression recognition.

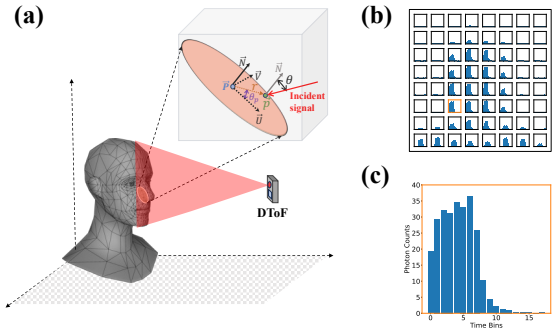


Fig. 1. DToF sensing diagram and data sample illustration. (a) shows the diffuse reflection of infrared light on the face. (b) is the transient histograms of 8×8 zones of VL53L8CH sensor. (c) is the transient histogram of one zone.

We propose two novel techniques for DToF sensing to overcome these challenges. First, based on an in-depth understanding of DToF sensing principles, we present a *physical structure estimation model* that can estimate fine-grained structural information, including multiple depths and orientations, from the intermediate output of the DToF sensor, i.e., the transient histogram. Then we devise ToFace, a *physics-integrated neural network* that combines the above physical model and deep learning modules to achieve accurate face detection and expression recognition. Specifically, for accurate face detection, we propose a simplified detection network to estimate the bounding box of the target user's face in the depth map by utilizing the property that the reflectance intensity of human skin is significantly lower than that of ambient objects. For expression recognition, we design a super-resolution module based on the physical structure estimation model to estimate higher-resolution depth and orientation maps using

the transient histogram as input. We then filter out regions corresponding to ambient objects in the depth and orientation maps based on the facial bounding box and propose a classifier module to perform accurate expression estimation.

To evaluate the effectiveness of ToFace, we conduct experiments with 12 users in two different environments: an office and a living room. Compared with other baseline approaches, ToFace achieves the highest recognition accuracy of 75.02%, which is comparable to state-of-the-art CV-based accuracy [10]. Additionally, for depth and orientation estimation, our model demonstrates the best performance, with errors of 23.55 mm and 0.042 rad, respectively.

II. DToF Sensing Principle

A DToF sensor estimates the distance between the sensor and targets by measuring the propagation time of individual IR photons using a grid of SPADs. This approach offers lower power consumption (100 mW [11]) and higher depth resolution (e.g., 1.5 mm [12]) compared to indirect ToF sensors, which rely on phase-shift measurements [13]. In this work, we build our prototype using the STMicroelectronics VL53L8CX sensor, which outputs an 8×8 depth map, along with intermediate measurements: an 8×8 reflectance map and transient histograms, as shown in Fig. 1.

Specifically, we introduce the sensing principle as follows. As depicted in Fig. 1(a), the IR light emitted by a dToF sensor is diffusely reflected off the surfaces of the human face according to the Lambertian reflection model, described by the following equation:

$$f\left(\theta, R\right) = \frac{\beta cos(\theta)}{4\pi^2 R^4}, \tag{1}$$

where $f(\cdot,\cdot)$ denotes the reflection factor, θ is the incident angle of IR light, R is the distance between the target and the sensor, and β is the skin's reflectivity.

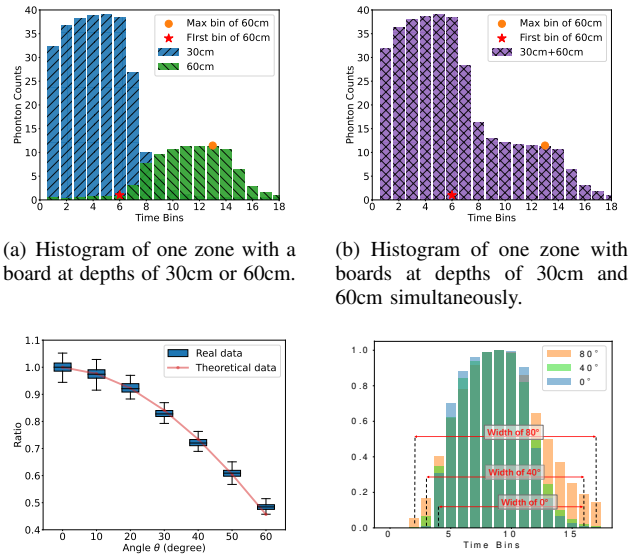
To extend this model to a patch with a central point $\vec{P}$ and surface normal $\vec{N}$, we establish a local polar coordinate system on that patch with $\vec{P}$ as the origin. Hence, the reflection factor for a point $\vec{p}$ on the patch is given by:

$$f(\vec{P}, \vec{N}, \theta_p, r) = \frac{\beta \vec{P} \cdot \vec{N}}{4\pi^2 |\vec{p}|^5}. \tag{2}$$

where r and θ_p represent the polar coordinates of the point. $\vec{p}$ can be represented as $\vec{p} = \vec{P} + r\cos(\theta_p)\vec{U} + r\sin(\theta_p)\vec{V}$, where $\vec{U}$ and $\vec{V}$ are two orthogonal vectors on the patch. Consequently, the received signal $\psi(t)$ for one zone in the dToF sensor is:

$$\psi(t) = \sum_{P}^{Z} \int_0^{r_0} \int_0^{2\pi} f(\vec{P}, \vec{N}, \theta_p, r) \cdot \gamma(t - \frac{|p|}{2c}) d\theta_p dr, \tag{3}$$

where γ represents the emitted infrared signal, and c denotes the speed of light and Z is the set of all patchs in the dToF zone. $\psi(t)$ is the transient histogram ultimately obtained, as shown in Fig. 1(c).



(a) Histogram of one zone with a board at depths of 30cm or 60cm.

(b) Histogram of one zone with boards at depths of 30cm and 60cm simultaneously.

(c) Ratio of the maximum value in transient histogram w.r.t. orientations.

(d) Normalized histogram of one zone with target across orientations.

Fig. 2. Verification experimental results.

III. Methodology

In this section, we first introduce the physical model that enables the estimation of multiple depths and orientations of the target, which serves as a generic model for DToF sensing. Next, we detail our physics-integrated neural network design for face detection and expression recognition.

A. Physical Structure Estimation Model

Through in-depth analysis of the collected real data samples and a thorough understanding of the sensing principle, we establish the relationship between the shape of the transient histogram and the target's fine-grained structure, which inspires us to develop the following multi-depth and orientation estimation method.

From Transient Histogram to Multi-Depth Estimation: To estimate a single depth, we place a panel at 30cm or 60cm respectively, and get the histograms as shown in Fig. 2(a). We identify the first bin in the transient histogram that surpasses the noise threshold [14]. This bin marks the time delay of IR signals, enabling depth calculation by multiplying the bin index by half the speed of light. For multi-depth estimation, we place two boards at 30cm and 60cm at the same time to get the corresponding transient histogram. As shown in Fig. 2(b), the transient histogram appears as a superposition of individual histograms. Although the overlap between objects may obscure the first bins, the maximum bin of the peak is more distinguishable and easier to detect than the first bin, making it a more reliable marker for estimating depth. The depth of the i-th target is then calculated as $r_i = \frac{c}{2}(T_i - \tau)$, where T_i is the time index of the i-th peak. τ represents the time between the start of the emitted pulse and the moment when the peak intensity is reached.

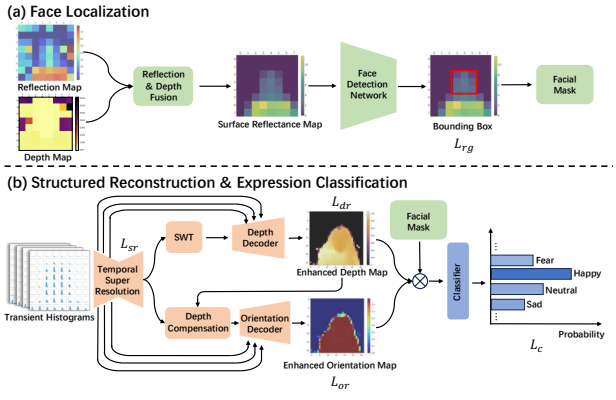


Fig. 3. Model overview of ToFace.

From Transient Histogram to Orientation Estimation: To determine the orientation of a patch, we discovered two properties of the transient histogram, *i.e.*, *orientation-maximum relation* and *orientation-width relation* as:

$$\theta = \text{arc cos}(\frac{\max(\psi_\theta)}{\max(\psi_0)}), \tag{4}$$

$$\theta = \text{arc sin}(\frac{c \cdot (T_h - T_l)}{4r}), \tag{5}$$

where ψ_θ is the histogram at orientation θ, ψ_0 is the histogram at $\theta = 0$, T_h is the histogram width, T_l is the duration of the IR signal, and r is the patch radius. Our experiments further validate these findings, as shown in Fig. 2(c) and Fig. 2(d).

B. Physics-Integrated Neural Network

The ToFace model comprises three main components: face detection, structured reconstruction, and expression recognition, as depicted in Fig 3. We first preprocess the reflection and depth maps to obtain the surface reflectance map, which is used to localize the face via a bounding box based on Eq.1. Then, as described in §III-A, the transient histogram is used to extract additional spatial information and reconstruct higher-resolution depth and orientation maps. These maps are then masked by the face bounding box and used for expression classification.

Face Detection: As shown in Fig. 3(a), the DToF sensor generates a reflection map $f(\cdot,\cdot)$ as described in Eq. 1. By *fusing* the depth map $\mathbf{d}^{8\times8}$ with the fourth power of the depth values, we can get the surface reflectance map $\mathbf{S}^{8\times8}$. The surface reflectance at position (i,j), $\mathbf{S}_{i,j}$, is given by $S_{i,j} = \frac{\beta cos(\theta)}{4\pi^2}$, which depends on the object's reflectivity β and orientation angle θ. Since the reflectivity and curvature of the face are consistent, background noise can be filtered by thresholding, yielding the surface reflectance map as shown in Fig. 3(a).

The surface reflectance map $\mathbf{S}^{8\times8}$ has lower spatial resolution than standard images, making typical computer vision models unnecessary and inefficient. To address this, we design a simplified CNN-based *face detection model* tailored for low-resolution inputs, using the same regression loss $\mathbf{L}_{rg}$ as SSD [15]. Despite its simplicity, this network achieves IOU performance comparable to standard detection models.

Structured Reconstruction: As discussed in §III-A, accurate depth extraction from a transient histogram requires identifying its peaks. Given the limited temporal resolution (e.g., 13ns [11]), improving this resolution is essential for precise depth estimation. While methods such as fitting Gaussian functions to histograms have been proposed [11], they are computationally expensive and struggle to handle multiple peaks efficiently. Instead, we use a *Temporal Super Resolution model* based on U-Net [16] improve the temporal resolution of the original transient histogram $\mathbf{T}^{8\times8\times n}$ to a higher-resolution histogram $\mathbf{T}_h^{8\times8\times(20\times n)}$. To reduce noise and ensure continuity in facial movements, multiple consecutive frames of the histogram are used as input. Due to the lack of ground truth high-resolution transient histograms, we applied a weakly supervised approach. The generated $\mathbf{T}_h$ is downsampled as $\mathbf{T}_{DS}$ through summation, and the temporal super-resolution loss $\mathbf{L}_{sr}$ is calculated using Mean Squared Error (MSE) as $\mathbf{L}_{sr} = \mathbf{MSE}(\mathbf{T}_{DS}, \mathbf{T})$.

To further enhance depth extraction accuracy, we incorporate wavelet-based preprocessing to optimize peak identification. In §III-A, we explored extracting multiple object depths by identifying peaks. To improve peak extraction, we applied the *Stationary Wavelet Transform (SWT)* to preprocess the input histogram. *SWT* preserves translation invariance and enables multi-scale analysis, ensuring that peak detection remains unaffected by peak position while identifying peaks across different frequency ranges. As shown in Fig. 4(a), the transient histogram and the corresponding Level 2 Detail Coefficients of the SWT are illustrated for two boards positioned at 30cm and 60cm. The SWT feature consistently aligns with the peaks of the transient histogram, with the feature's magnitude reflecting the intensity of each peak.

Following the preprocessing step, the next phase involves reconstructing depth map with higher space resolution by using a specialized neural network approach. We use a CNN-based *Depth Decoder* to recover depth values from the SWT features. To prevent information loss and enhance feature propagation, we adopted bypass connections inspired by U-Net. For the depth reconstruction loss, we chose the Mean Absolute Error (MAE) for the depth reconstruction loss L_{dr} because it penalizes all errors equally, making it more suitable for super-resolution tasks than MSE, which focuses more on large errors [17]. The loss is defined as $\mathbf{L}_{dr} = \mathbf{MAE}(\mathbf{D}, \mathbf{D_{gt}})$, where $\mathbf{D}$ is the reconstructed depth map and $\mathbf{D}_{gt}$ is the downsampled ground truth from the depth map obtained by Femto Mega.

As described in Eq. 4 and Eq. 5, the orientation angle θ is influenced by both the peak intensity and width. However, as noted in Eq. 2, peak intensity also depends on the distance between the DToF sensor and the target. To address this, we first calculate a *depth compensation* factor $\mathbf{C}$ using the depth map, and the compensated transient histogram $\mathbf{T_c}$ is derived as $\mathbf{T_c} = \mathbf{C} \times \mathbf{T_h}$ which is then passed through a CNN-based Orientation Decoder to generate the orientation map.

Similar to the Depth Decoder, the *Orientation Decoder*

employs bypass connections and Mean Absolute Error (MAE) as the orientation reconstruction loss: $\mathbf{L}_{or} = \mathbf{MAE}(\mathbf{O}, \mathbf{O_{gt}})$, where $\mathbf{O_{gt}}$ is the ground truth of orientation map which is calculated by the depth map of Femto Mega.

Expression Recognition: From the reconstructed depth $\mathbf{D}$ and orientation $\mathbf{O}$ maps, we extract the facial region using a mask derived from the bounding box, yielding $\mathbf{D}_{\text{mask}}$ and $\mathbf{O}_{\text{mask}}$. These masked maps are used for expression classification with cross-entropy loss $\mathbf{L}_c$.

IV. EXPERIMENT SETUP

A. Dataset

To evaluate ToFace, we collected 46,369 samples from video frames captured in two environments (office and living room) with IRB approval. Twelve volunteers (9 males, 3 females) performed 7 expressions: neutral, happiness, anger, sadness, fear, and disgust [18]. Participants moved freely during data collection, introducing variations in orientation, position, and depth to test the model's generalization across diverse scenarios. We split the dataset into training, validation, and testing sets with a ratio of 8:1:1, respectively.

B. Baseline Models

For face detection and expression classification, we compare our approach with Faster R-CNN [19] and SSD [15]. In terms of depth and orientation reconstruction, we benchmark our method against U-Net [16] and MobileNetV2 [20]. These comparisons allow us to comprehensively evaluate the performance of our model across different aspects.

C. Metrics

Accuracy: The accuracy of facial expression classification.

Intersection over Union (IOU): It indicates the accuracy of predicted bounding boxes in face detection by measuring the overlap between the predicted area and the ground truth area.

Mean Absolute Error(MAE): The performance of facial reconstruction is evaluated by calculating the Mean Absolute Error (MAE) between the masked depth map D_{mask} and orientation map O_{mask} and their corresponding ground truth maps, after applying the bounding box mask. The units for depth and orientation are millimeters and radians, respectively.

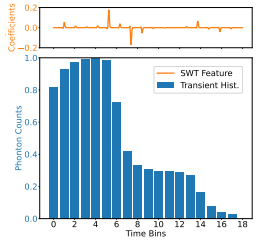
V. EVALUATION

Overall performance: As shown in Table. I, ToFace outperformes most aspects of baselines, demonstrating the superiority of our physics-inspired design. ToFace achieves the highest FER accuracy with 75%. Besides, ToFace yields an IOU similar to the baseline, but use a simpler detection network structure. Moreover, the ToFace model outperforms all baselines in terms of reconstruction performance.

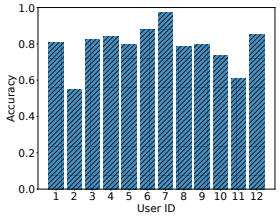
Cross user: Robust performance across users is essential for practical FER systems. We evaluate ToFace using data from 12 participants (U1–U12). As shown in Fig.4(b), while accuracy for U2 and U11 is slightly lower, it remains consistently high for most users, demonstrating the model's overall robustness.

TABLE I
OVERALL PERFORMANCES.

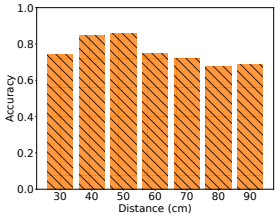
Model	Acc	IOU	Depth Err (mm)	Orient. Err (rad)
ToFace	**0.7502**	0.8814	**23.55**	**0.042**
Faster Rcnn	0.5719	0.8872	✗	✗
SSD	0.521	**0.8898**	✗	✗
Unet	✗	✗	24.7	0.047
MobileNet	✗	✗	52.28	0.1176



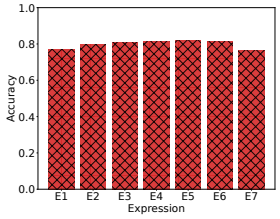
(a) Histogram of one zone with a board at depth of 30cm and 60cm.



(b) FER accuracy of different users.



(c) FER accuracy of users at different depths.



(d) FER accuracy of different expressions.

Fig. 4. Cross evaluation.

Impact of distance: We evaluate the impact of distance between the users and the DToF sensor and observe a slight performance drop beyond 50 cm, as shown in Fig. 4(c).

Impact of Expression: ToFace exhibits consistent performance across facial expressions, achieving comparable accuracy for all expressions, as shown in Fig. 4(d).

VI. CONCLUSION

We propose ToFace, a FER system utilizing a low-cost DToF sensor that ensures both privacy preservation and environmental robustness. Through an in-depth exploration of the DToF sensing principles, we introduce a physical model for fine-grained target structure estimation, serving as a general method for DToF sensing. By integrating this physical model with deep learning modules, ToFace achieves accurate face detection and expression recognition. We develop a prototype using a commodity DToF sensor and conduct extensive real-world experiments, demonstrating the remarkable performance of ToFace. Our future work will explore finer-grained facial representation or reconstruction and improve ToFace's generalizability for cross-domain scenarios, *e.g.*, new users.

VII. ACKNOWLEDGMENTS

This paper is partly supported by the NSFC under grant No. 62222216, Hong Kong RGC ECS under grant No. 27204522, and GRF under grant No. 17212224.

REFERENCES

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[2] G. Yolcu, I. Oztel, S. Kazan, C. Oz, and F. Bunyak, "Deep learning-based face analysis system for monitoring customer interest," *Journal of ambient intelligence and humanized computing*, vol. 11, pp. 237–248, 2020.

[3] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald, "Social risk and depression: Evidence from manual and automatic facial expression analysis," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.

[4] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2168–2177.

[5] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[6] GDPR-info.eu, "General Data Protection Regulation (GDPR) – Legal Text," https://gdpr-info.eu/, 2024, [Accessed 25-June-2024].

[7] X. Zhang, Y. Zhang, Z. Shi, and T. Gu, "mmfer: Millimetre-wave radar based facial expression recognition for multimedia iot applications," in *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, 2023, pp. 1–15.

[8] E. Charbon, "Single-photon imaging in complementary metal oxide semiconductor processes," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 372, no. 2012, p. 20130100, 2014.

[9] P. Padmanabhan, C. Zhang, and E. Charbon, "Modeling and analysis of a direct time-of-flight sensor architecture for lidar applications," *Sensors*, vol. 19, no. 24, p. 5464, 2019.

[10] S. Zhang, Y. Zhang, Y. Zhang, Y. Wang, and Z. Song, "A dual-direction attention mixed feature network for facial expression recognition," *Electronics*, vol. 12, no. 17, p. 3595, 2023.

[11] C. Callenberg, Z. Shi, F. Heide, and M. B. Hullin, "Low-cost spad sensing for non-line-of-sight tracking, material classification and depth imaging," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–12, 2021.

[12] L. Cester, A. Lyons, M. C. Braidotti, and D. Faccio, "Time-of-flight imaging at 10 ps resolution with an iccd camera," *Sensors*, vol. 19, no. 1, p. 180, 2019.

[13] seeed studio, "Deptheye s2," 2023. [Online]. Available: https://www.seeedstudio.com/DepthEye-S2-VGA-Resolution-ToF-Camera-p-5095.html

[14] S. Chan, A. Halimi, F. Zhu, I. Gyongy, R. K. Henderson, R. Bowman, S. McLaughlin, G. S. Buller, and J. Leach, "Long-range depth imaging using a single-photon detector array and non-local data fusion," *Scientific reports*, vol. 9, no. 1, p. 8075, 2019.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[16] O. Ronneberger, P. Fischer, and T. Brox, "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*. Springer, 2015, pp. 234–241.

[17] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[18] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.