Vision Problems under Adverse Imaging Conditions

Zhu Li

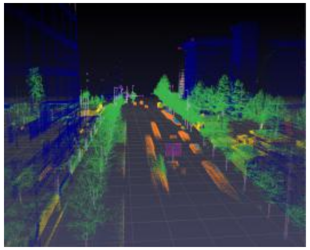
Director, UMKC NSF Center for Big Learning

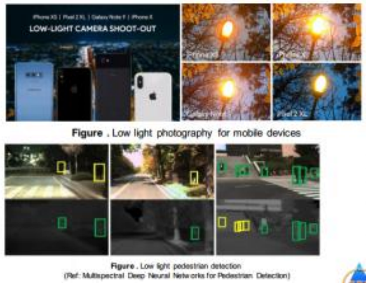
Dept of Computer Science & Electrical Engineering

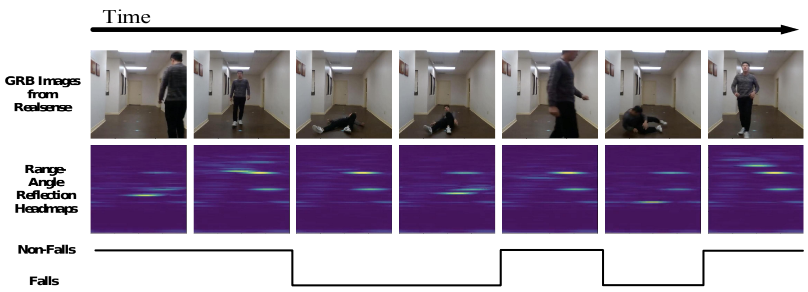
University of Missouri, Kansas City

Email: zhu.li@ieee.org, lizhu@umkc.edu

Web: http://l.web.umkc.edu/lizhu







Outline

- Overview of Research at MCC Lab
- Vision Problems under Adverse Imaging Conditions
  - Dark image enhancement from sensor field
  - Gradient image super resolution for key point repeatability
  - Human action recognition from RF signal domain
- Summary

Short Bio:

Research Interests:

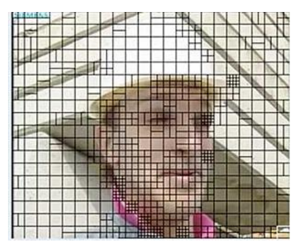
- Immersive Media Communication: light field, point cloud and 360 video capture, coding and low latency communication.
- Data & Image Compression: video, medical volumetric data, DNA sequence, and graph signal compression with deep learning
- Remote Sensing & Vision: vision problem under low resolution, blur, and dark conditions, hyperspectral imaging, sensor fusion
- Edge Computing & Federated Learning: gradient compression, light weight inference engine, retina features, fast edge cache for video CDN

NSF I/UCRC Center *for* Big Learning

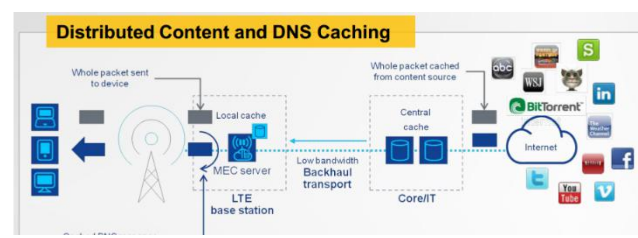
Creating Intelligence

signal processing and learning

image understanding



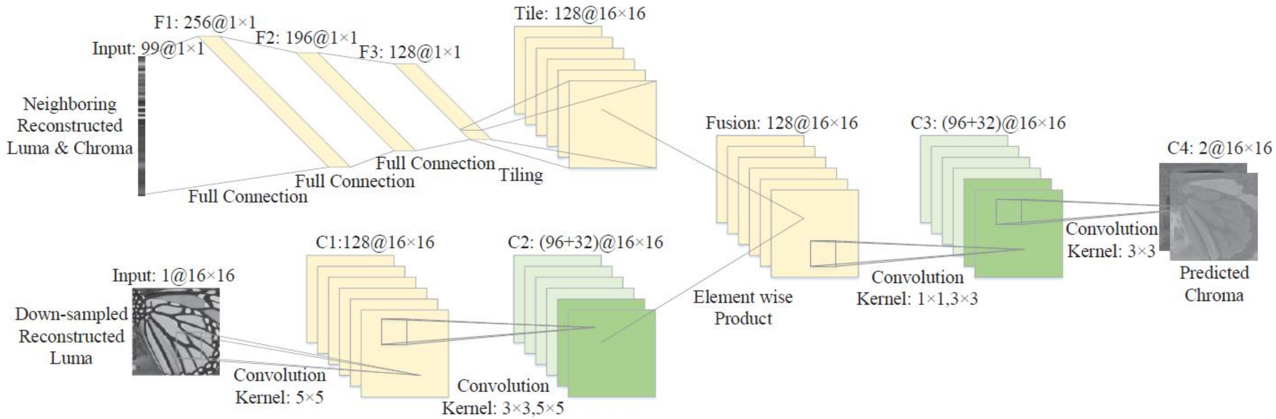
visual communication



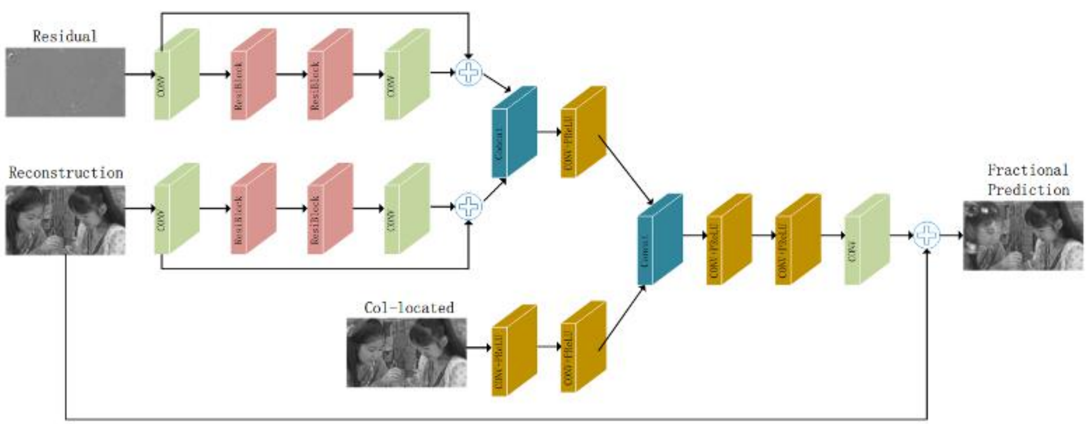
mobile edge computing & communication

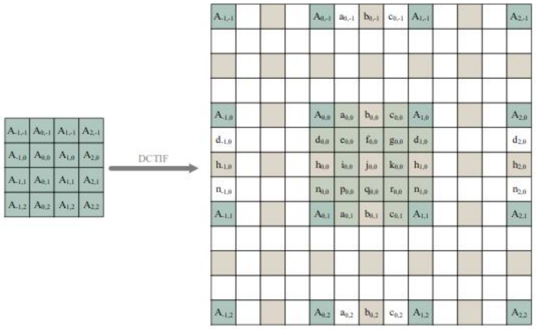
Data & Image Compression Highlights (NSF/IUCRC)

- “Neural Network Based Cross-Channel Intra Prediction”, *ACM Trans on Multimedia Computing Communication and Applications* (TOMM), 2021.



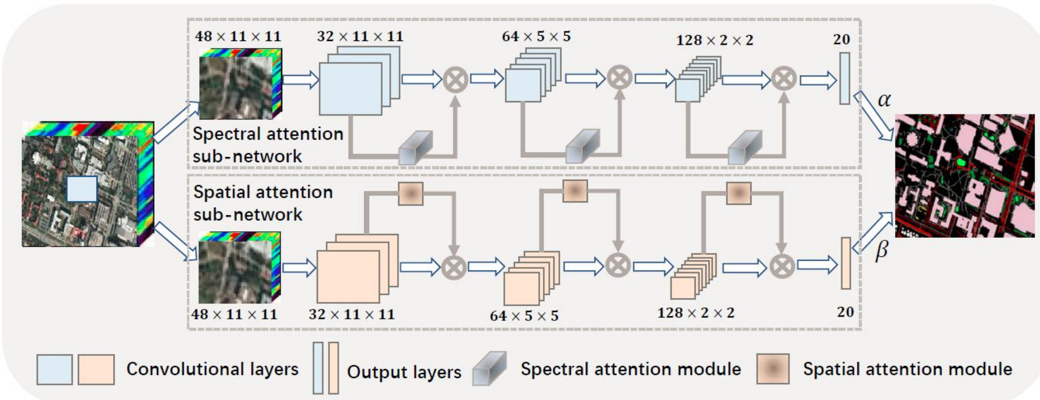
- “Compression Priors Assisted Convolutional Neural Network for Fractional
or Video Tech. (T-CSVT), 2020





Remote Sensing & Vision Highlights (AFOSR, ONR)

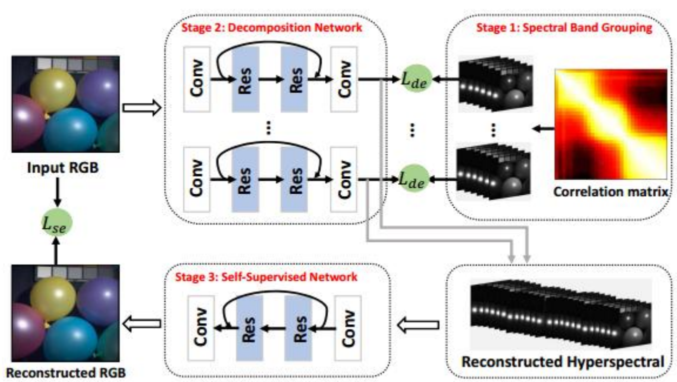
- **"Hyperspectral Image Classification with Attention Aided CNNs"**, *IEEE Trans. on Geoscience & Remote Sensing* (T-GRS), 2020.



Attention CNN for Hyperspectral Image Classification

- Introducing a dual stream network architecture with separate attention model for spatial and spectral feature maps
- Achieving the SOTA performance.

- **“PRINET: A ... uper-Resolution Network”**, *IEEE International Conf on* ... 020.



PRINET: Spectral Super Resolution

- Super-resolve hyper-spectral info from RGB inputs
- A dual loss network that learn a correlation decomposed HSI images
- Achieving the new SOTA performance.

Edge Media Computing & Federated Learning

- "Referenceless Rate-Distortion Modeling with Learning from Bitstream and Pixel Features", *ACM Multimedia* (MM), Seattle, 2020.

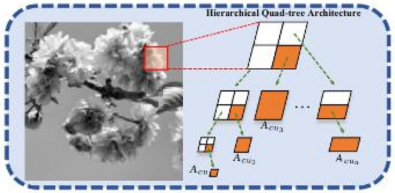


Fig 4. Segmentation Mappings

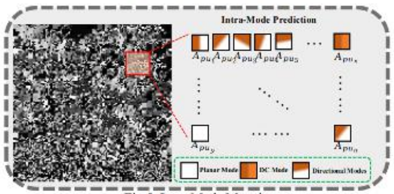
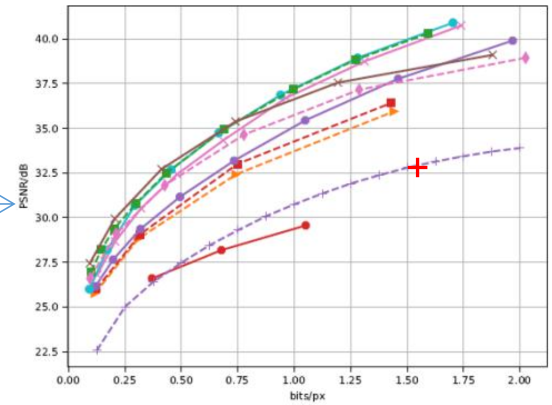


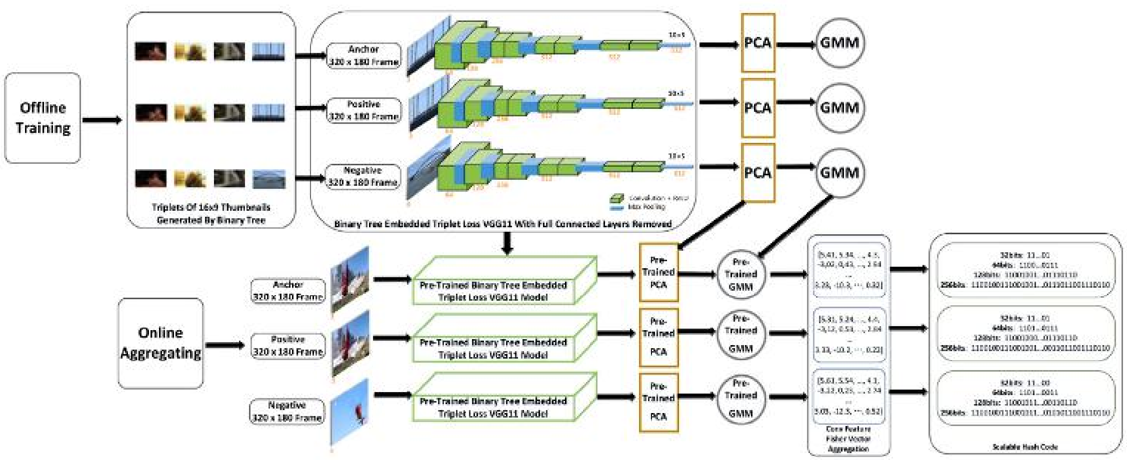
Fig 5. Intra-Mode Mappings

learn from
one encoding

ref-less R-D
modeling

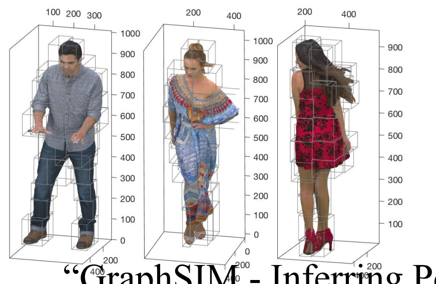


- “Scalable Hash From Triplet Loss Feature Aggregation for Video De-Duplication”, *Journal o*



0-FPR deduplication at
<10ms latency for very
large repository

Immersive Media Coding & Communication (NSF/IUCRC)




- ❑ "GraphSIM - Inferring Point Cloud Quality via Graph Similarity", *IEEE Trans on Pattern Analysis & Machine Intelligence* (T-PAMI), 2021.
- ❑ "Efficient Projected Frame Padding for Video-based Point Cloud Compression", *IEEE Trans on Multimedia*(T-MM), 2020.
- ❑ "Rate Control for Video-based Point Cloud Compression", *IEEE Transactions on Image Processing* (T-IP), 2020.
- ❑ "λ-domain Perceptual Rate Control for 360-degree Video Compression", *IEEE Journal of Selected Topics in Signal Processing* (JSTSP), 2020.
- ❑ "Advanced 3D Motion Prediction for Video Based Dynamci Point Cloud Compression", *IEEE Trans on Image Processing*(T-IP), 2019.
- ❑ "Quadtree-based Coding Framework for High Density Camera Array based Light Field Image", *IEEE Trans on Circuits and Systems for Video Tech*(T-CSVT), 2019.
- ❑ "Advanced Spherical Motion Model and Local Padding for 360 Video Compression", *IEEE Trans on Image Processing* (T-IP) vol. 28, no. 5, pp. 2342-2356, May 2019.
- ❑ "Scalable Point Cloud Geometry Coding with Binary Tree Embedded Quadtree", *IEEE Int'l Conf. on Multimedia & Expo* (ICME) ,San Diego, USA, 2018.
- ❑ "Pseudo sequence based 2-D hierarchical coding structure for light-field image compression", *IEEE Journal of Selected Topics in Signal Processing* (JSTSP), Special Issue on Light Field, 2017.

Motivation

- Low-light photography
  - Almost all smart phone camera have dedicated section for low-light imaging

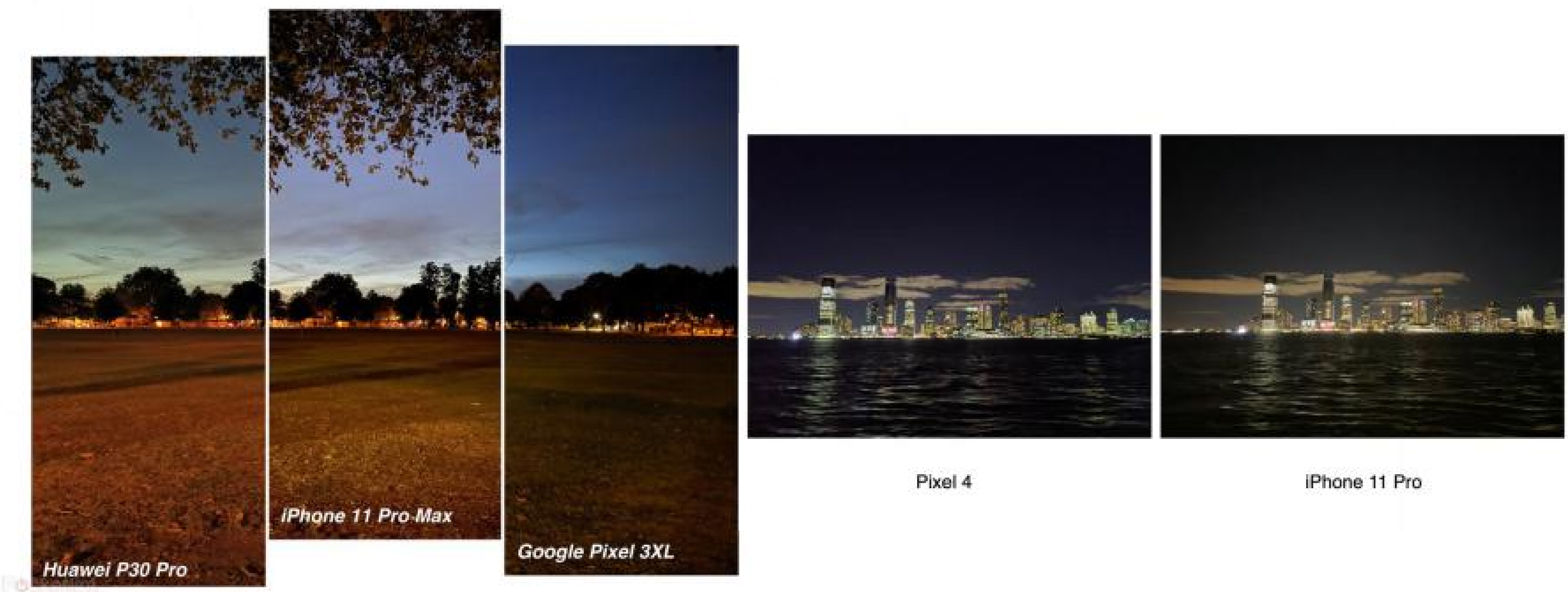


Figure 1. Low-light camera comparison for different smartphones

Motivation

- Low-light vision task
  - Object detection
  - Face recognition
  - Surveillance

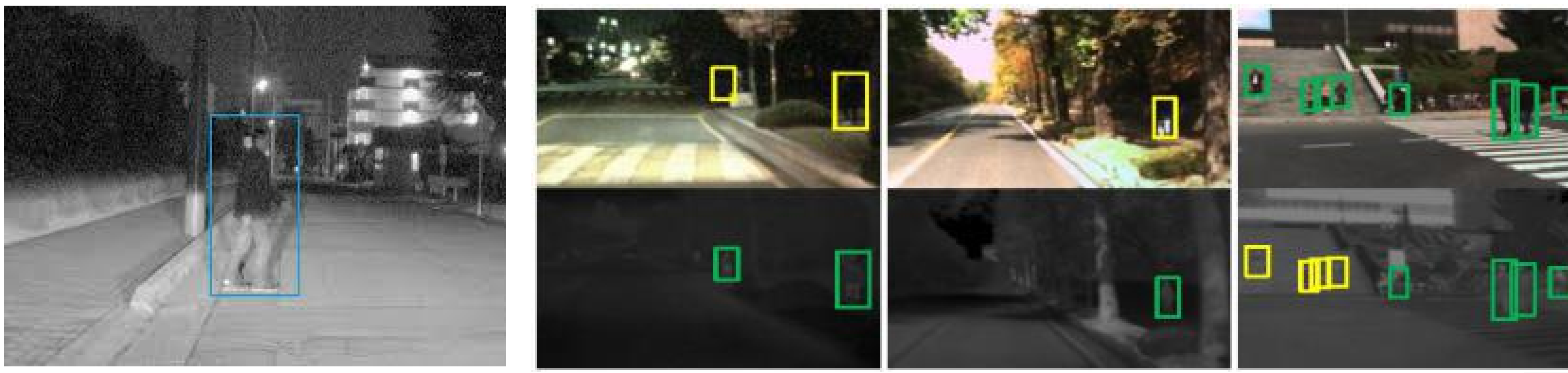


Figure 2. Low light pedestrian detection (Ref: Multispectral Deep Neural Networks for Pedestrian Detection)

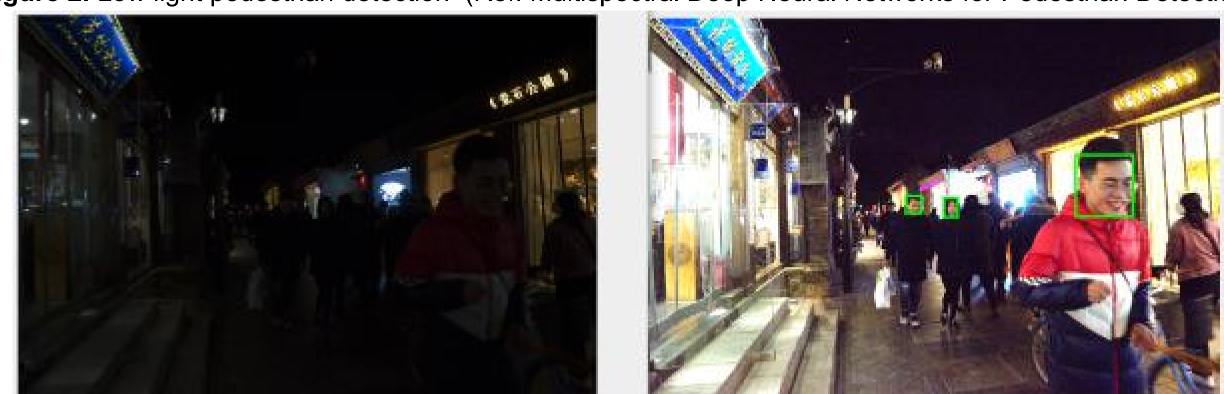


Figure 3. Low light pedestrian detection (Ref: Multispectral Deep Neural Networks for Pedestrian Detection)

Objective

- To design network to denoise the low-light image in Bayer domain
- To use wavelet decomposition to divide and conquer the problem by learning sensor field sub images using separate networks

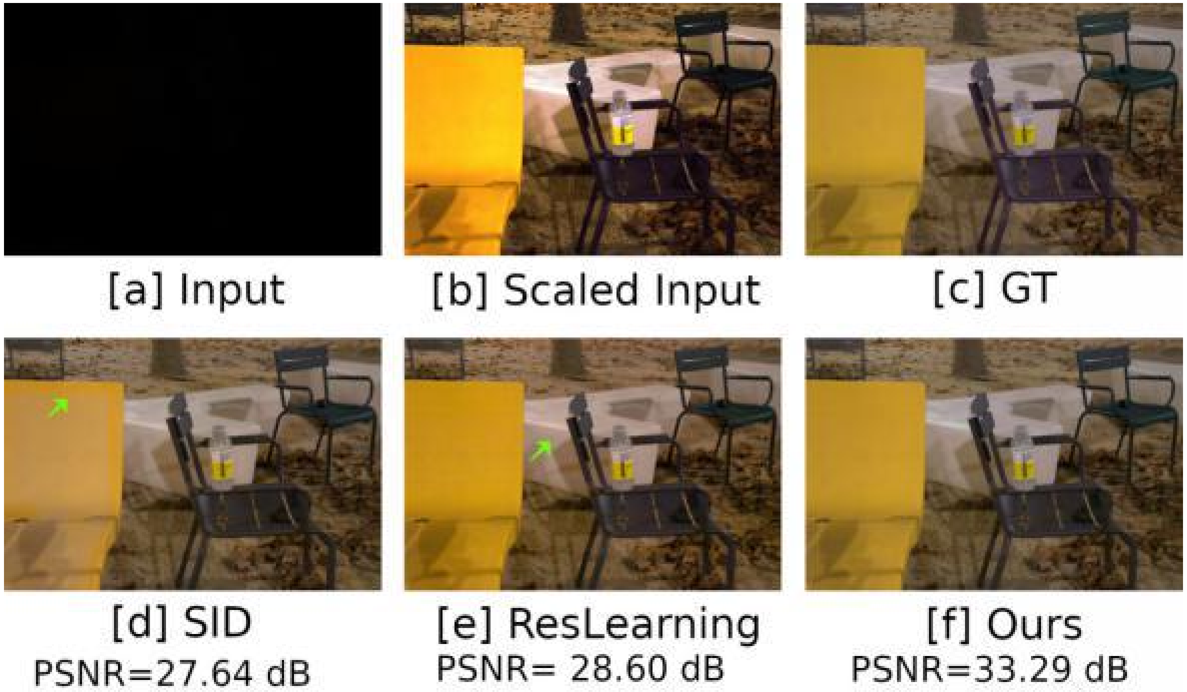


Figure 4: [a] Extreme low-light image from Sony a7S II exposed for 1/25 second . [b] 250x intensity scaling of image in [a]. [c] Ground truth image captured with 10 second exposure time. [d] Output from SID[]. SID introduced some artifacts around the edge of the chair as shown by green arrow. [e] Output from ResLearning[]. The white region as indicated by arrow in image is not properly reconstructed as white compared to that in ground truth image. [f] Our result.

Introduction

- Under low-light condition image sensor suffers from low signal-to-noise ratio
- Generates noisy image, as not enough photon reaches the camera sensors
- Enlarging aperture will reduce the depth of field –blurry image
- Extending the exposure time cause motion blur
- Increasing the ISO will also amplify the noise signals

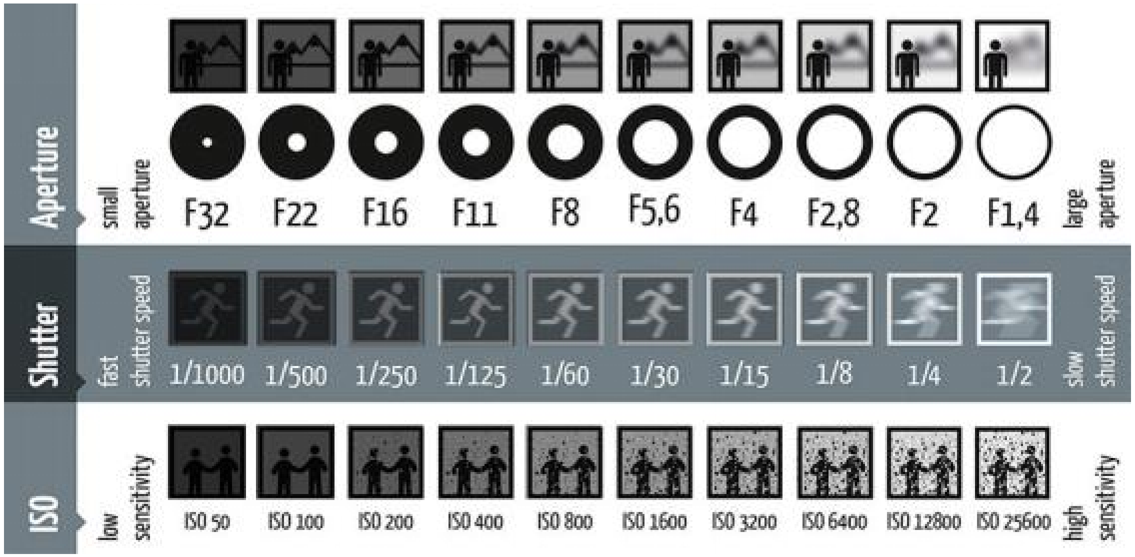


Figure 5. Effect of aperture, shutter speed and ISO in camera for low light imaging

Main Contribution

- Proposed a novel method of denoising before ISP (can be more useful for machine vision instead of human consumption)
- Decomposed the input raw image into low and high frequency subimages using wavelet transform
- A new loss function for learning high frequency components of our proposed wavelet decomposition network

Dataset

- See-In-Dark Dataset: Real world extreme low-light images with corresponding noise-free ground truth
- Illumination less than 0.5 lux
- Three different exposure of 1/10th 1/25th and 1/30th seconds and corresponding ground truth of 10 seconds
- The time difference between the shutter speed is taken as the amplification ratio

Figure 6. Sample of low-light image and its corresponding ground truth image

Wavelet Decomposition

- Used Haar wavelet as decomposition filter
- *g(n)* is low pass filter, *h(n)* is high pass filter
- The resulting output is downsampled by half in rows and columns
- LL is equivalent to low freq while LH, HL and HH equavalent to horizontal, vertical and diagonal component respectively

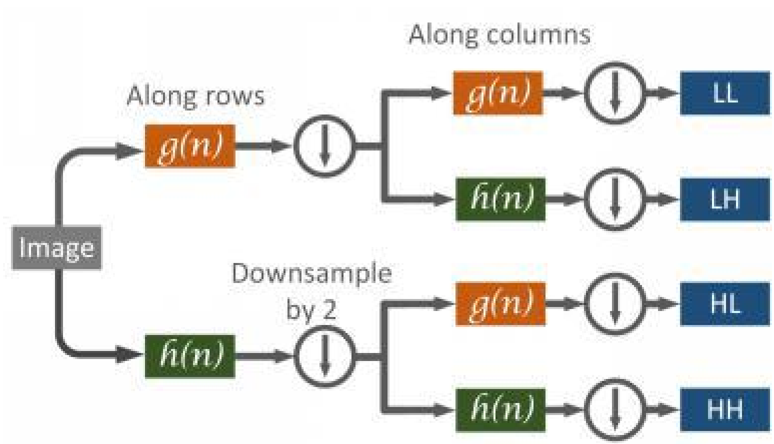


Figure 7. One layer decomposition using wavelet transform.

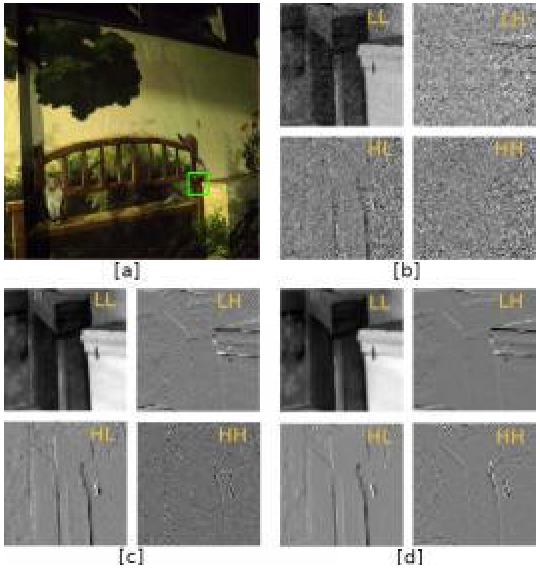


Figure 8. Decomposition of image using wavelet transform. [a] Noisy low-light image converted to sRGb by using Rawpy library [b] Wavelet decomposition of small patch of [a]. [c] Wavelet decomposition of corresponding ground truth image [d] Prediction from our network for LL, LH, HL and HH component with combination of L1 and SSIM
for high frequency component.

Methodology

- Two stages: first stage is the denoising network while the second stage is the off-the-shelf camera ISP
- Trained four different network for LL, LH, HL and HH component of wavelet
- Combined the information of LL to LH, HL and HH for better prediction of high frequency imformation

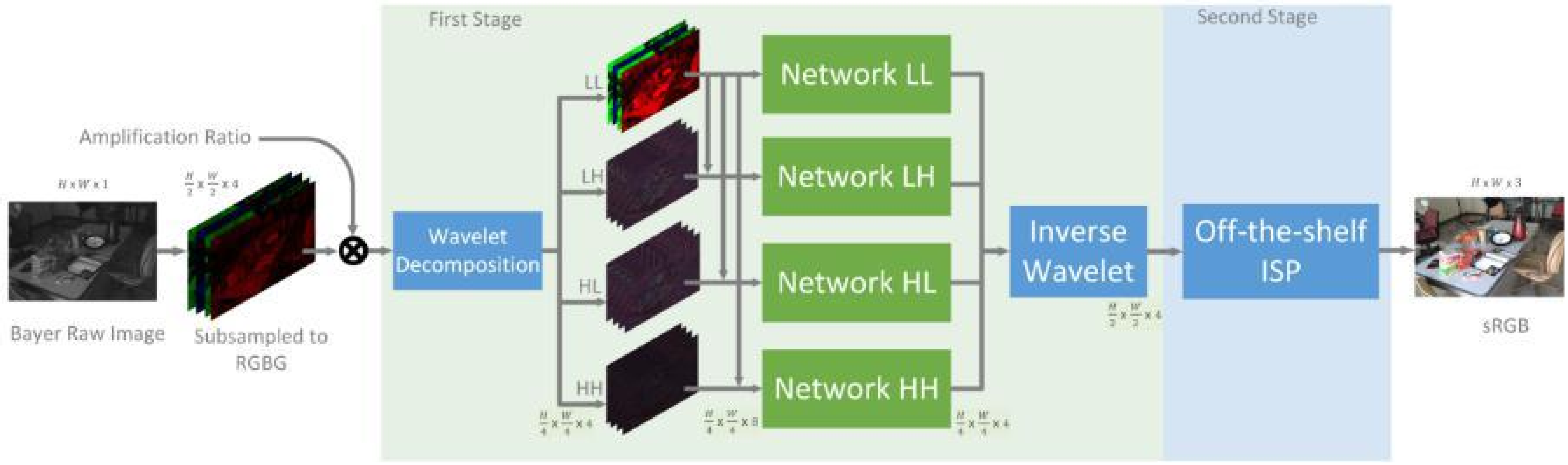


Figure 9: Overview of our wavelet decomposition based network. The first stage learns the decomposed image and used the inverse wavelet to reconstruct the denoised 4 channel image. The second stage uses the off-the-shelf ISP to enhances the image and converts into 3 channel sRGB image.

Network Architecture

- Network based on residual learning
- Consists of 32 residual blocks for LL while only 8 residual blocks were used for LH, HL and HH network
- LeakyReLU as activation function
- Residual block followed by Squeeze-and-Excitation block-converges the network faster and increases the performance
- While training, patch size is 256 x 256, learning rate of 0.0001, and 64 filters at each conv layer
- L1 as loss function, Adam as optimizer, and each network trained for 4000 epochs

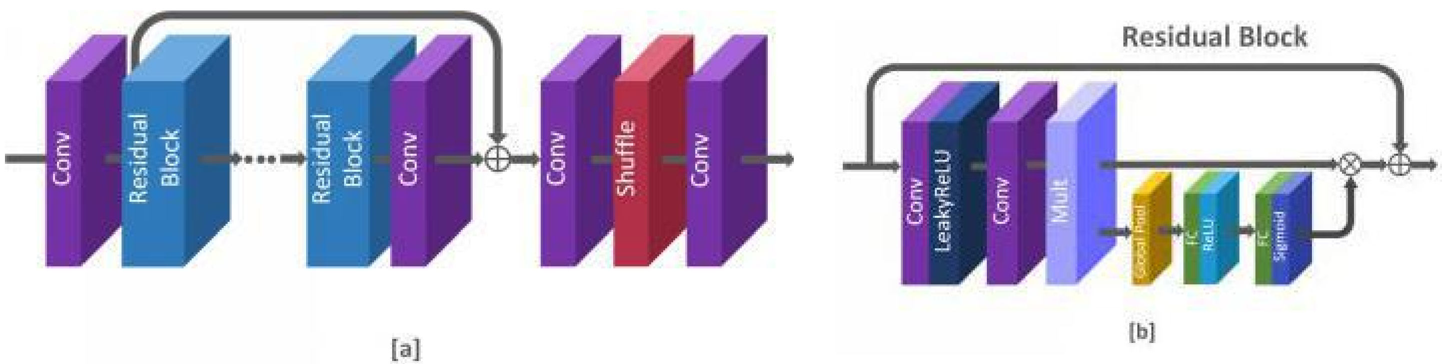


Figure 10. [a]Residual network [b] Residual block with LeakyReLU as activation function and squeeze-and-excitation block

Subband Image Adaptive Loss Function

- We use L1 loss for learning low frequency component (LL),

$$\mathcal{L}_1 = \|\hat{x} - x\| \quad (1)$$

- For high frequency component LH, HL and HH we used adaptive loss of L1 and SSIM loss

$$\mathcal{L}_{sturctural} = 1 - SSIM(\hat{x}, x) \quad (2)$$

$$\mathcal{L}_{HF} = \alpha * \mathcal{L}_1 + \mathcal{L}_{sturctural} \quad (3)$$

Quality Metrics

- Evaluation against the current SOTA

Table 1. Comparison of our proposed method of denoising before ISP with the existing method of joint denoising and demosaicing. **(Higher value of PSNR is better. Lower value of RMSE and NIQE is better.)**

Experiments	PSNR	RMSE	NIQE
SID[2]	28.97	0.03956	5.1904
ResLearning[1]	29.16	0.03926	5.8507
Ours	**30.02**	**0.03568**	**4.6166**

Table 2. Ablation study of our proposed method in terms of PSNR.

Experiments	100x	250x	300x	Total
UNet + ISP	31.95	29.49	27.64	29.52
ResLearning + ISP	32.25	29.70	27.70	29.70
Wavelet+L1+ISP	32.14	29.85	27.97	29.82
Wavelet+L1 & SSIM + w/o data separation + ISP	32.10	29.61	28.22	29.84
Ours	**32.34**	**29.97**	**28.22**	**30.02**

Experimental Results

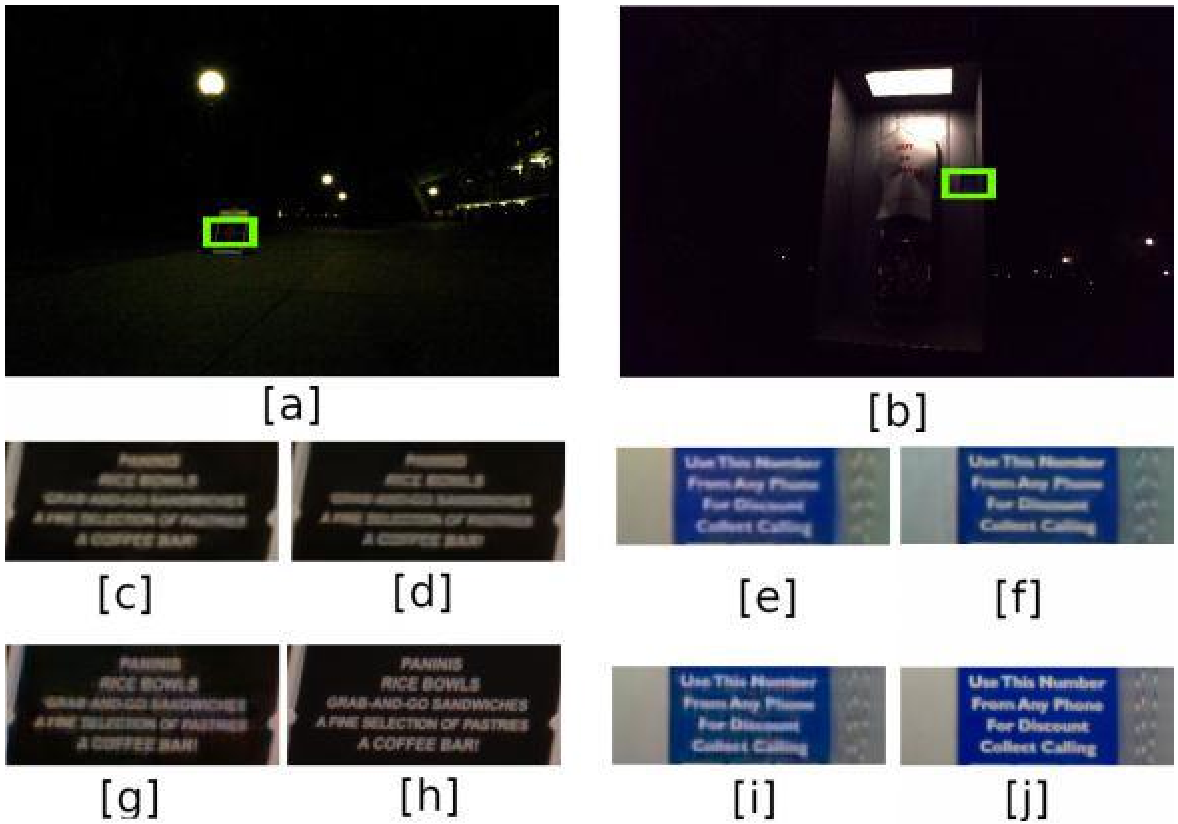


Figure 11. Results showing image details using our method in comparison to SID[] and ResLearning[]. [a, b] Dark input images [c, e] Outputs from SID []. The text are blurred and color is different from ground truth. [d, f] Output from ResLearning. Though the image has lots of details than [c,e], the text is still blurred. [g, i] Outputs from our network. The text are much cleaner and color is much closer to the ground truth. [h, j] Zoomed version of corresponding ground truth images.

Experimental Results

Figure 12. [a] Subjective results from our method in comparison with BM3D[],SID[] and ResLearning[] [a] Extreme low-light image captured by Sony a7S II. [b] Intensity scaled version of [a] converted to RGB by rawpy library [c] Denoised by BM3D and demosaic and enhanced by Rawpy library. We used different sigma values of 10,20,40, and 60 and selected the one with best PSNR. BM3D was not able to denoise properly as seen in the zoom image [d] Output from SID. We can see some artifacts indicated by arrow and bounding box [e] Output from ResLearning. The color reproduction in accurate. [f] Our result. Denoised in Bayer domain using wavelet decomposition and demosaic and enhanced by Rawpy library [g] Corresponding ground truth image.

More Results

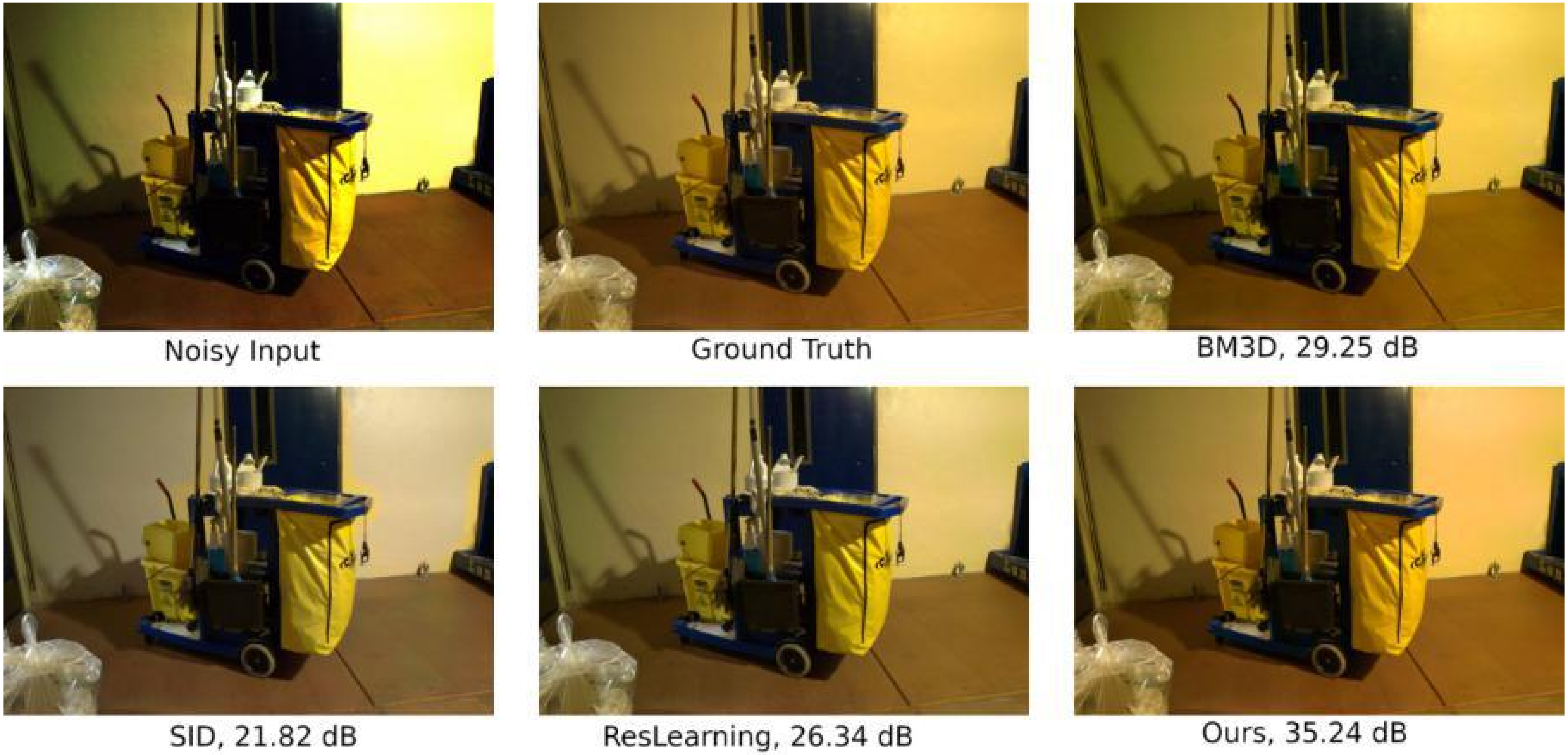


Figure 13. Comparison of our method with BM3D[2], SID[1] and ResLearning[3] in terms of PSNR for the indoor imageunder extreme low-light condition. The color in the wall and the floor is well reproduced and closer to the ground truth image.

More Results

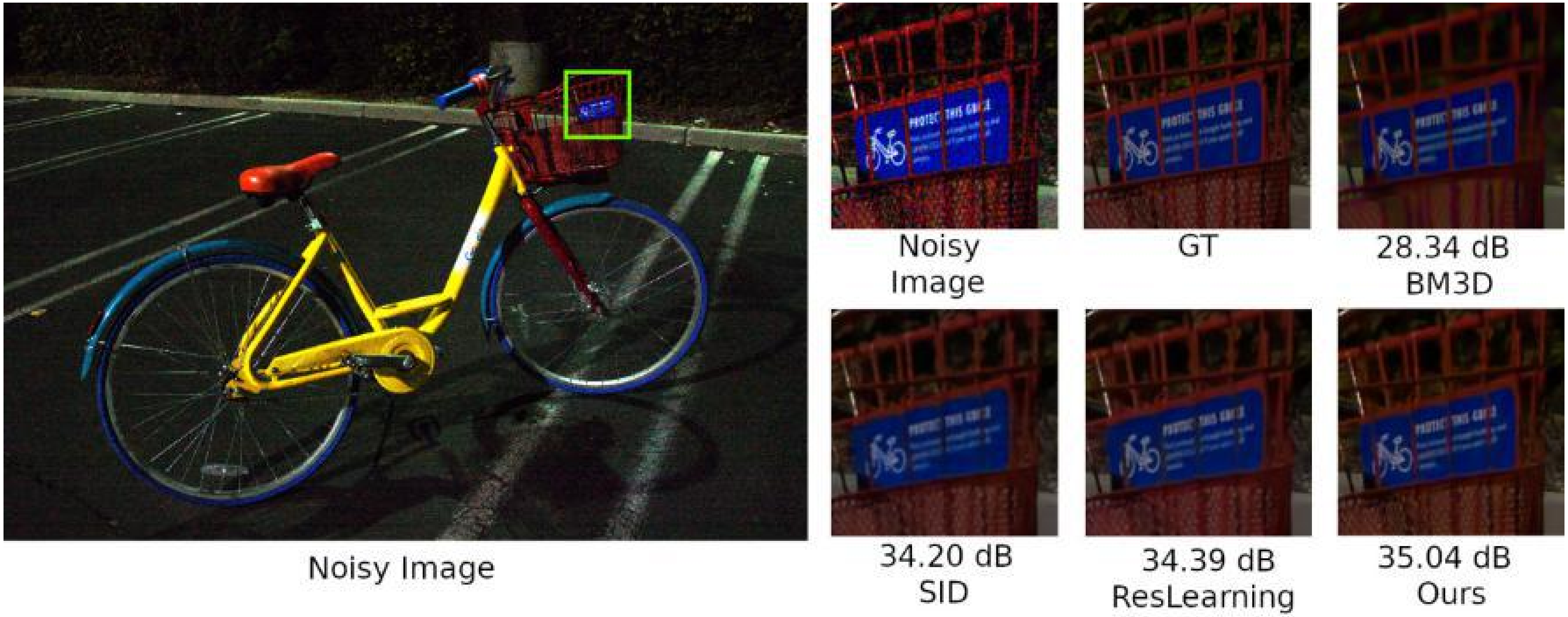


Figure 14. Comparison of our method with BM3D[2], SID[1] and ResLearning[3] in terms of PSNR for the outdoor imageunder extreme low-light condition. The detail in the image produced by our method is much closer to ground truth image.

More Results

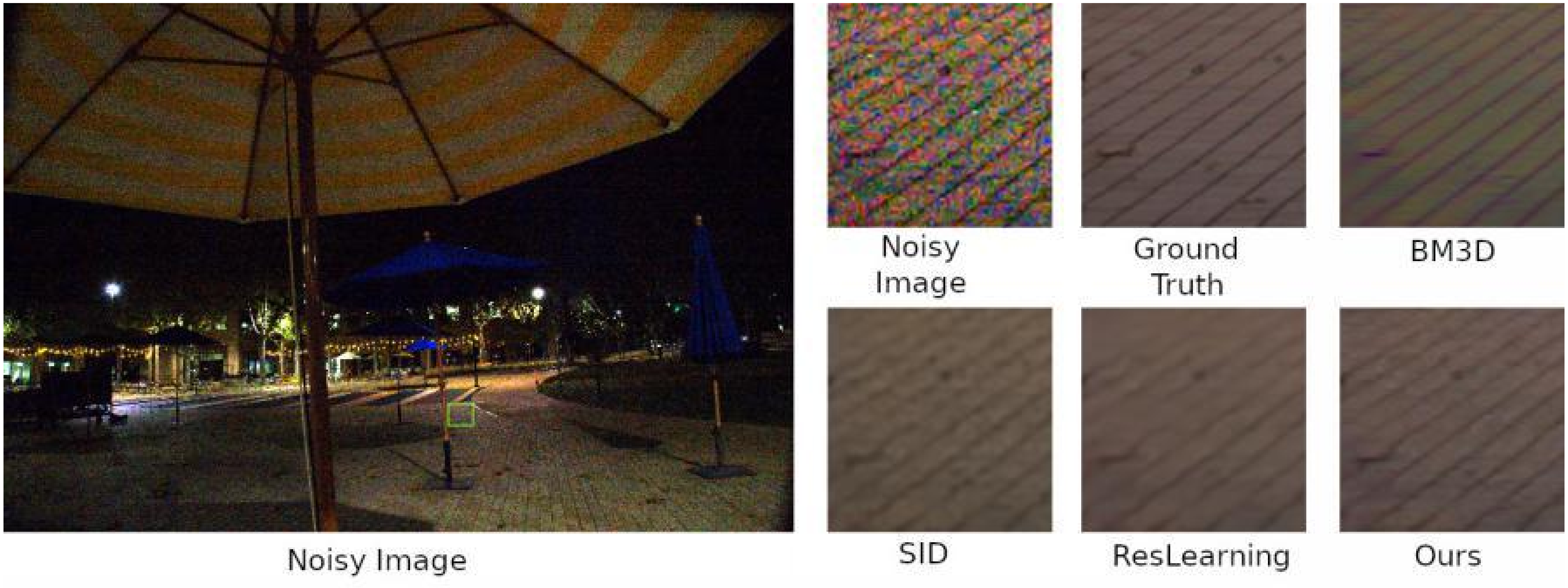


Figure 15. Another example showing both color and details from our proposed method which is closer to the ground truthimage. BM3D[2] uses the sigma value of 5. Though the texture is preserved, the color is different from output. SID[1] andResLearning[3] have missing details and are blurred.

Conclusion and Future Work

- We propose a nobel method of direct sensor field denoising solution by exploiting the strong prior obtained from wavelet decomposition
- We achieved significant gain in terms of PSNR via our decomposition network and loss function adaptation
- The time complexity for our network is less than typical implementation, as we are processing approx two-third less information than sRGB image.
- Inference time is 21x faster (11 ms per 4K frame) than prior state of the art.
- In future, we will explore different wavelet functions, develop prefiltering and design adaptive loss function for even more performance gain

Gradient Image and Multi-scale Representation

- **Gradient image** generally refers to a change in the direction of the intensity or color of an image. In a gradient image, in a certain direction, each pixel finds out the change in intensity of that same point in the original image

- **Harris Detector** is used to find out the edges and extract corners of the image as well as discovering the infer features of the image

- **Laplacian of Gaussian** is used for blob detection. It detects points that are continuously local maxima or minima with respect to both scale and space

- In **SIFT**, difference of Gaussian (**DoG**) is used for feature detection. From DoG images, maxima and minima are computed to find key points in SIFT detection

Harris Edge Detection

LoG Blob Detection

SIFT Feature Detection

Proposed Method Formulation

- Let , I(x,y) is the original image; G is the Gaussian Kernel,

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \quad (1)$$

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \quad (2)$$

L is the function which denotes the scale space of the input image I

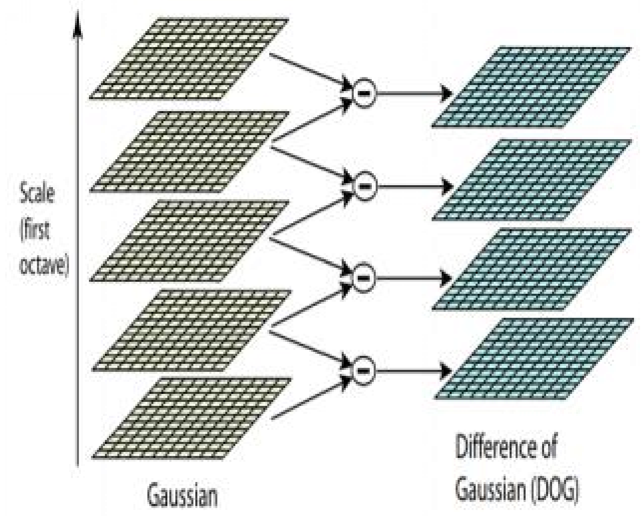


Figure : DoG in SIFT

- Therefore Difference of Gaussian will be:

$$D(x,y,\sigma_1,\sigma_2) = (G_1(x,y,\sigma_1) - G_2(x,y,\sigma_2)) * I(x,y)) \quad (3)$$

$$D(x,y,\sigma_1,\sigma_2) = L_1(x,y,\sigma_1) - L_2(x,y,\sigma_2) \quad (4)$$

- The standard deviation values , σ are 1.24 , 1.54 ,1.94 , 2.45, 3.09 for formulating 4 different DoGs

Proposed Method Formulation

- The loss function E is the MSE loss between the DoG of the super-resolved blurred generated image and the DoG from convolution with original image:

$$E(\hat{D}, D_{original}) = \sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{D}^{ij} - D^{ij}_{original})^2 \qquad (5)$$

Where $\widehat{D}$ is the predicted DoG image which is upscaled and $D_{original}$ is the DoG image computed from of the original one convolved with Gaussian filter.

- The gradient descent of the loss function will be:

$$\frac{\delta E}{\delta \hat{D}} = \frac{\delta(\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{D}^{ij} - D^{ij}_{original})^2)}{\delta \hat{D}} \qquad (6)$$

$$\begin{aligned}\frac{\delta E}{\delta \hat{D}} =& 2\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{D}^{ij} - (\frac{1}{2\pi{\sigma_1}^2}P - \frac{1}{2\pi{\sigma_2}^2}Q)) \\ & (1 - (\frac{1}{2\pi{\sigma_1}^2}\frac{\delta P}{\delta \hat{D}} - \frac{1}{2\pi{\sigma_2}^2}\frac{\delta Q}{\delta \hat{D}}))\end{aligned} \qquad (7)$$

$$P = e^{\frac{-(x_i^2+y_j^2)}{2{\sigma_1}^2}} * I(x_i, y_j), Q = e^{\frac{-(x_i^2+y_j^2)}{2{\sigma_2}^2}} * I(x_i, y_j) \qquad (8)$$

- The simplified loss function can be written as MSE between Gaussian blurred images and computing DoG images separately.

$$E(\hat{L}, L_{original}) = \sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{L}^{ij} - L^{ij}_{original})^2 \qquad (9)$$

Network Implementation

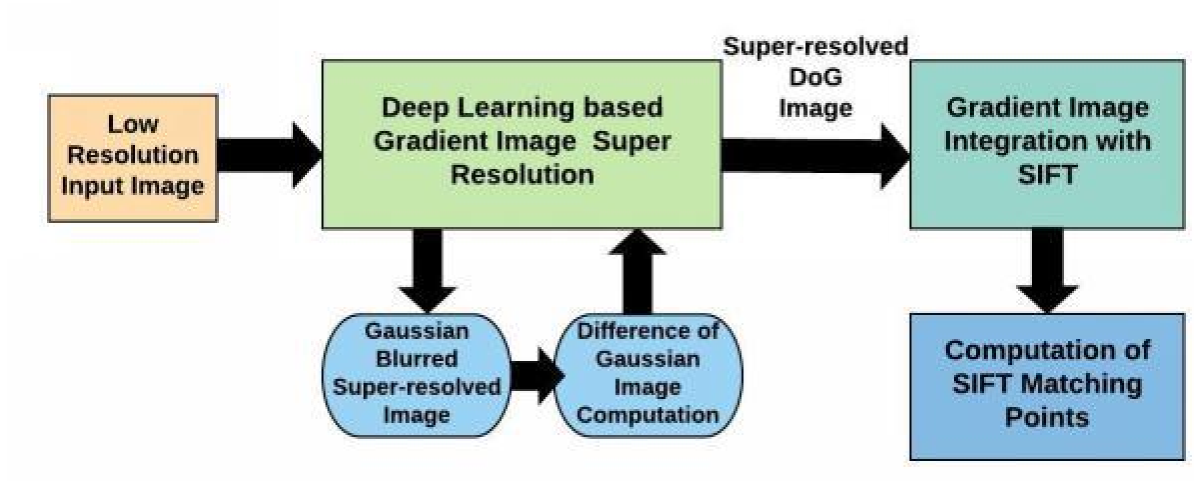


Figure: Proposed Network Architecture

- Low Resolution input images will be passed through a deep learning based gradient image super resolution stage. There are five SR networks for the purpose
- Each SR network produces a super-resolved Gaussian blurred image with different σ values [σ ={1.24 , 1.54 ,1.94 , 2.45, 3.09}]
- Four Gradient images (DoG image) are computed from five Gaussian Blurred images
- Four Gradient images are integrated to SIFT method for the computation of key matching points

Network Implementation

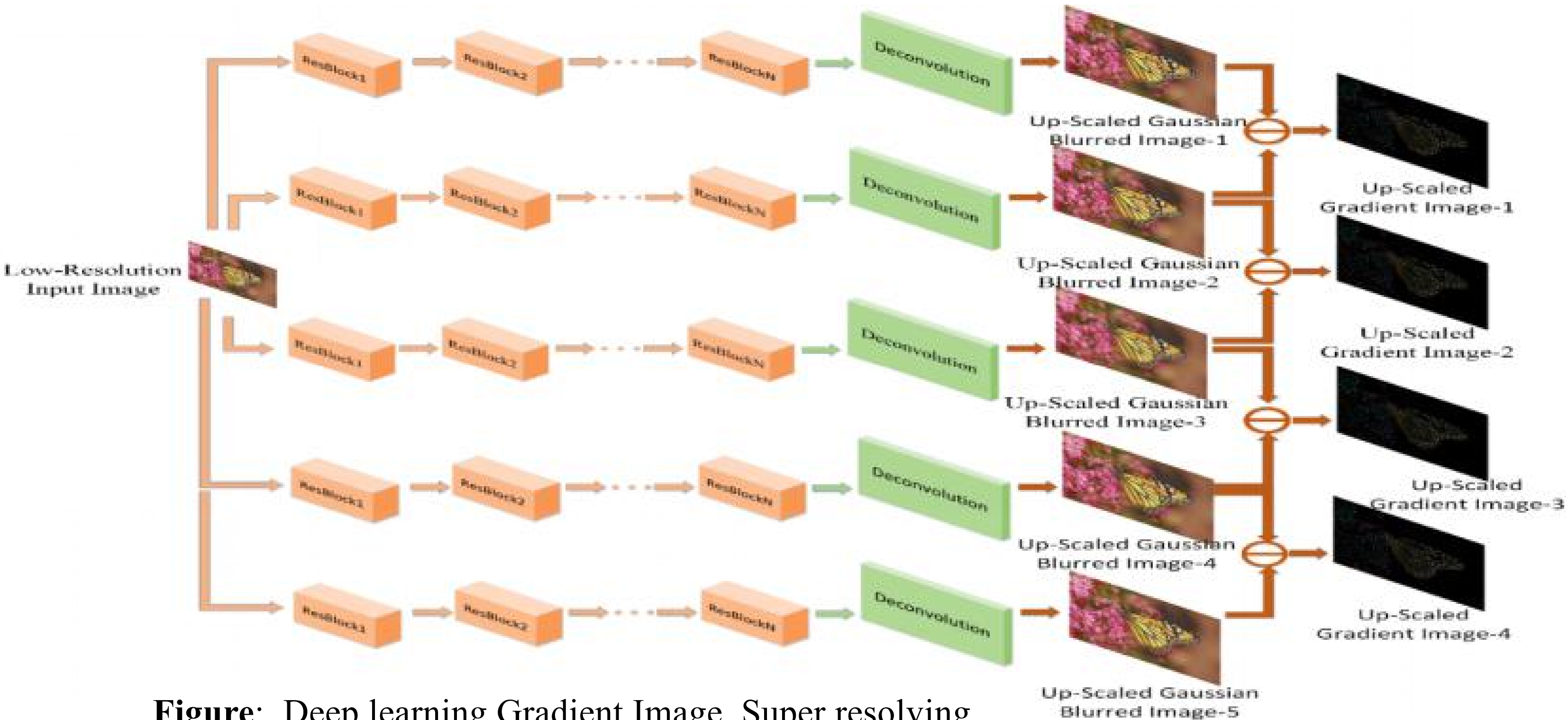


Figure: Deep learning Gradient Image Super resolving network to compute upscaled gradient image

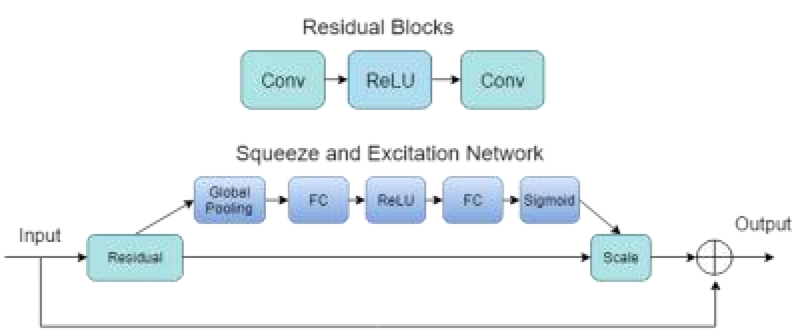


Figure: Residual Blocks

- Filter kernel size of 3X3 with 64 number of features
- Deconvolutional Layer is used to upscale.

Alternative Network Implementation

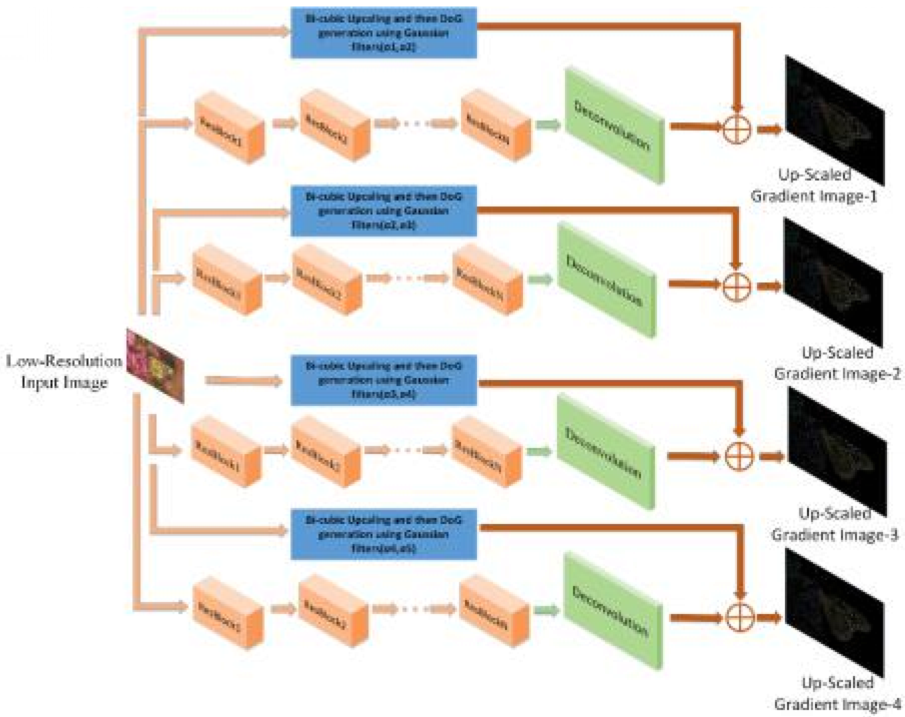


Figure: Deep learning Gradient Image Super resolving network to compute upscaled gradient image

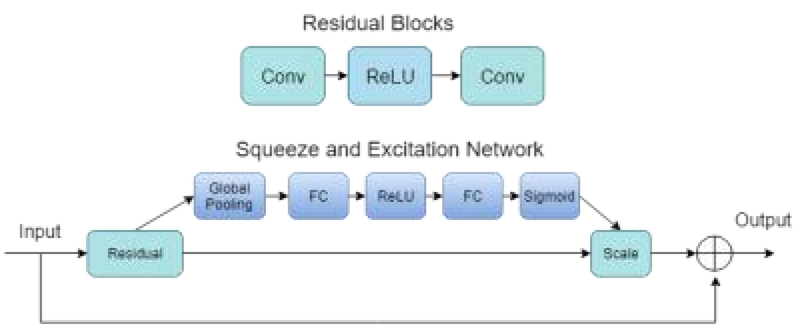


Figure: Residual Blocks

- Filter kernel size of 3X3 with 64 number of features
- Deconvolutional Layer is used to upscale.

Experimental Dataset

Training Dataset:

1. CVPR DIV_2k dataset with 800 images is used for training.
2. They are first downsampled by 2 /4 times
3. Cropped patch size:32X32.
4. Total input data 300k

Test Dataset:

1. MPEG CDVS Full dataset.
2. MPEG CDVS is a comprehensive collection of images of various objects which consists of 186k labeled images of CDs and book covers, paintings, video frames, buildings and common objects
3. Oxford building dataset
4. Paris building dataset
5. 200 matching pairs from each category were chosen
6. They are first downsampled by 2 /4 times

Results

MPEG CDVS Full dataset results:

Table 1: Average number of SIFT matching points for 200 matching image pairs from each category

Category	Upscaling Factor	Avg. no. of matching SIFT points for the original image	Avg. no. of matching SIFT points using proposed method-1	Avg. no. of matching SIFT points using proposed method-2	Avg. no. of matching SIFT points using EDSR	Avg. no. of matching SIFT points using SRCNN	Avg. no. of matching SIFT points using SRGAN	Avg. no. of matching SIFT points using bi-cubic interpolation
Building	2	125.8	124.5	130.4	116.3	114.5	115.8	112.4
Building	4	125.8	110.8	115.4	105.6	104.2	104.3	100.4
Graphics	2	101.6	99.8	102.8	94.5	93.8	94.2	92.8
Graphics	4	101.6	87.2	90.4	86.7	86.1	86.8	85.4
Objects	2	115.3	113.9	118.5	106.9	103.9	104.8	102.6
Objects	4	115.3	105.1	108.8	99.1	98.2	98.5	96.2
Painting	2	114.4	114.7	120.5	105.9	104.4	104.9	100.7
Painting	4	114.4	106.1	109.8	101.5	100.1	100.2	96.1
Video	2	94.3	90.3	94.4	87.2	86.2	85.8	85.2
Video	4	94.3	82.2	85.5	80.1	79.4	79.6	79.2

Results

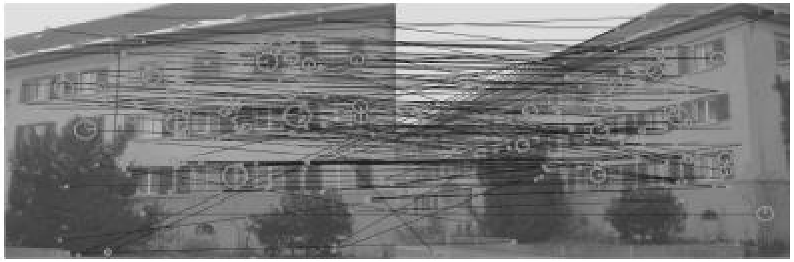
Oxford dataset results: Average number of SIFT matching points for 200 matching image pairs

Upscaling Factor	Avg. no. of matching SIFT points for the original image	Avg. no. of matching SIFT points using proposed method-1	Avg. no. of matching SIFT points using proposed method-2	Avg. no. of matching SIFT points using EDSR	Avg. no. of matching SIFT points using SRCNN	Avg. no. of matching SIFT points using SRGAN	Avg. no. of matching SIFT points using bi-cubic interpolation
2	105.4	101.4	107.3	97.1	96.2	96.4	94.2
4	105.4	93.2	97.8	91.1	90.4	90.3	89.9

Paris dataset results: Average number of SIFT matching points for 200 matching image pairs

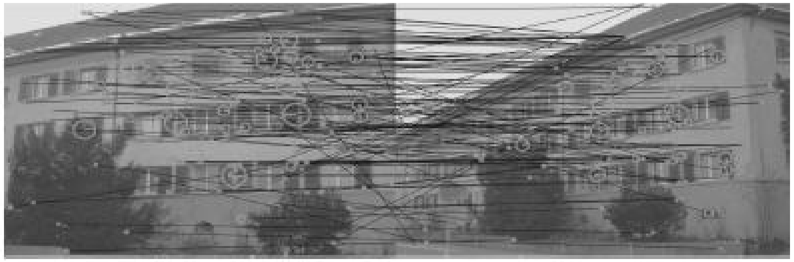
Upscaling Factor	Avg. no. of matching SIFT points for the original image	Avg. no. of matching SIFT points using proposed method-1	Avg. no. of matching SIFT points using proposed method-2	Avg. no. of matching SIFT points using EDSR	Avg. no. of matching SIFT points using SRCNN	Avg. no. of matching SIFT points using SRGAN	Avg. no. of matching SIFT points using bi-cubic interpolation
2	110.5	107.9	113.2	101.4	99.2	99.8	99.1
4	110.5	99.8	102.4	97.1	95.4	95.9	95.3

Comparative Results for SIFT matching points

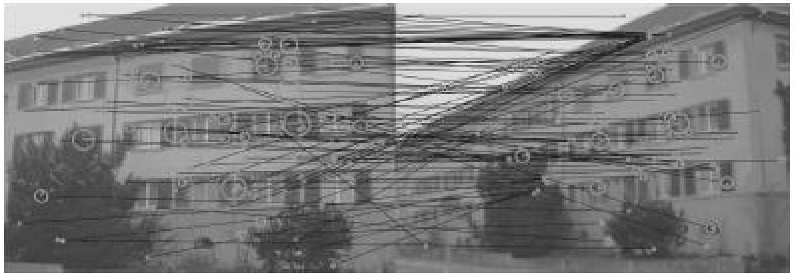


(a) SIFT matching points for original image (102 points)

(b) SIFT matching points using proposed method (112 points)



(c) SIFT matching points using EDSR (100 points)



(d) SIFT matching points using bi-cubic interpolation(96 points)

Figure: SIFT Matching Points Comparison for a sample matching image pair with 2x upscaling

Privacy-Preserving Fall Detection with Deep Learning on mmWave Radar Signal

Outline

- Introduction
- Framework
- Radar Signal Processing
- Experimental Devices
- Network
- Experimental Results

Introduction

- Fall injuries lead the accidental death and nearly $34 billion in direct medical costs annually for seniors.
- Conventional solutions:
  - Wearable portable alert devices, e.g. automatic bracelets.
    - Pros: Accuracy and low-latency
    - Cons: Skin discomfort and inconvenience
  - Nonwearable alert system, e.g. camera-based surveillance equipment.
    - Pros: Accuracy and low-latency
    - Cons: High power consumption, invasion of privacy and high sensitiveness at extreme environment
- Related works:
  - Doppler based radar detection [1]
  - The changes of different WiFi channel solution [2]
  - 3D-CNN radar frequency detection [3]

[1] L. Liu, M. Popescu, M. Skubic, M. Rantz, T. Yardibi, and P. Cuddihy, "Automatic fall detection based on doppler radar motion signature," in 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops. IEEE, 2011, pp. 222–225.
[2] S. Palipana, D. Rojas, P. Agrawal, and D. Pesch, "Falldefi: Ubiquitous fall detection using commodity wi-fi devices," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 1, no. 4, p. 155, 2018.
[3] Y. Tian, G.-H. Lee, H. He, C.-Y. Hsu, and D. Katabi, "Rf-based fall monitoring using convolutional neural networks," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 3, p. 137, 2018.

Framework

- Motivated by the 3D-CNN RF-based solution, we propose an LSTM-based fall detection method based on the mmWave radar signal.
  - Characterize the radar reflections based on distance from the human body along with the vertical and horizontal angles of arrays.
  - Capture locality and velocity components simultaneously.
  - Radar signal low-dimension embedding algorithm (RLDE) with LSTM reduces the complex and save chip memory.

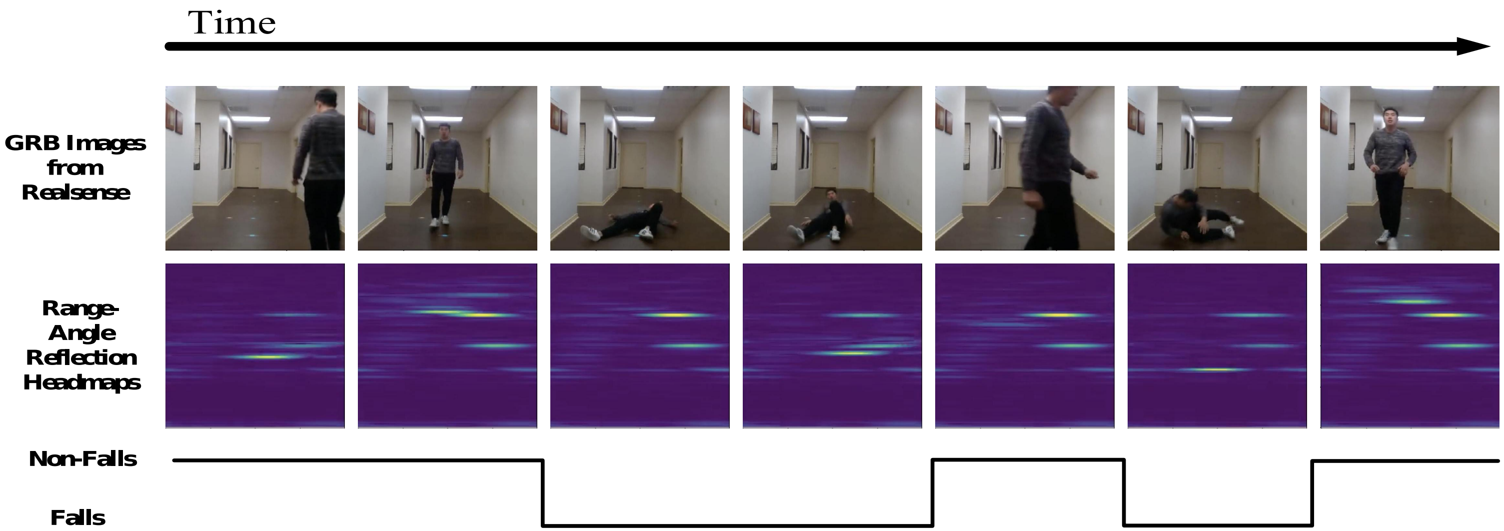


Figure 1. mmWave Radar based Fall Detector

Proposed radar signal-based fall detection

- Human activities are regarded as the changes in terms of range, angle, and speed, which can be caught by a pair of IWR1642 radar devices.
- The time interval and intensity of signal between the receiver (RX) and transmitter (TX) can be recorded and correlated to fundamental attributes by training.
- The proposed method comprises two subtasks:
  - Radar signal processing
  - Neural network processing

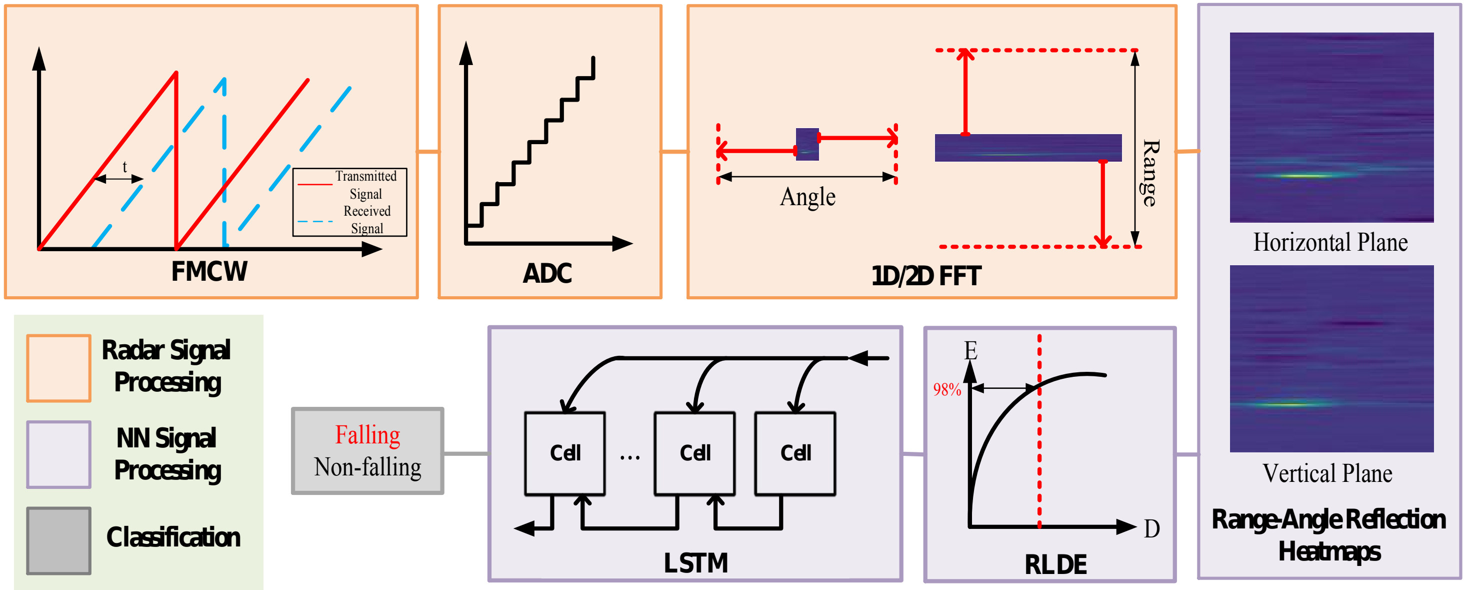


Figure 2. Framework of Proposed Detector

Radar Signal Processing

- This procedure performs the frequency modulated continuous wave (FMCW) signal conversion to analyzable digital form in the spatial domain (reflection heatmaps).
  - ADC (Analog-digital converter): modulate continuous form to discrete form.
  - Range-FFT (Range domain Fast Fourier Transform): convert the signal from the time domain to the spatial (range) domain.
  - Angle-FFT (Angle domain Fast Fourier Transform): catch phase difference between each RX antenna.

Table 1. Core Parameters of Radar Device

Parameters	Values	Parameters	Values
Max. Range	10 m	Wave Form	FMCW
Range Res.	4 cm	Frequency	77-81 GHz
Num of RX	8	Num of TX	4
Field of View	120°	Angular Res.	15°
Max. Velocity	6.5 m/s	ADC samples	256
Velocity Res.	0.2 m/s	Frame rate	25 f/s
Wavelength	3.9 mm	Max. Bandwidth	3,750 MHz

Neural network processing

- Human activities are continuous dynamic patterns that can be recognized in both spatial and temporal dependencies. We use successive radar reflection heatmaps as the representative of human activities.
  - PCA is adopted as RLDE algorithm to project reflection heatmaps $\{H_t, \ V_t\}$ to a low-dimension subspace P as the elimination of spatial redundancies,
  - The proposed RNN with LSTM units utilizes the changes of motion at the temporal domain. The softmax layer operates as a classifier. The cross-entropy function is adopted as the objective function.

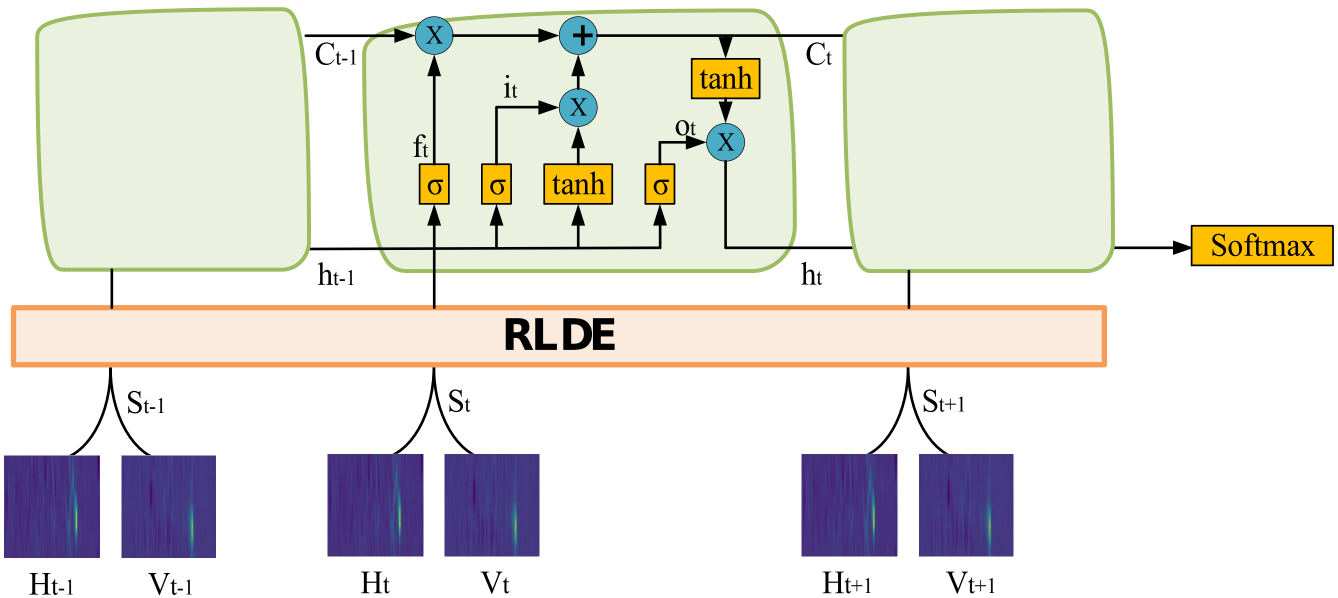


Figure 3. Architecture of RNN with LSTM units

Experimental Results

- 4,126 samples (2.56s for each sample) consist of 128 frames of reflection heatmaps, divided into two classes: fall and non-fall.

Table 1: the comparison on accuracy and processing time between 3DCNN and LSTM with or w/o RLDE implementation

	Model	Precision	Recall	F1-Score	Training time (s)
w/o RLDE	$3DCNN$	95.3%	96.6%	96.0%	181.21
	$LSTM$	100.0%	93.6%	96.7%	94.29
with RLDE	$LSTM^{64}$	100.0%	97.9%	98.9%	56.83
	$LSTM^{32}$	100.0%	95.8%	97.8%	37.22
	$LSTM^{16}$	100.0%	97.7%	98.9%	22.21
	$LSTM^{8}$	97.9%	100.0%	98.9%	20.33
	$LSTM^{4}$	100.0%	97.7%	98.9%	17.08
	$LSTM^{2}$	97.5%	88.6%	92.9%	15.12

Extensive experiment

- Multiple human activities detections: 7 categories of human activities are labeled: Boxing, Falling, Jogging, Jump, Pick up, Stand up & Walking.

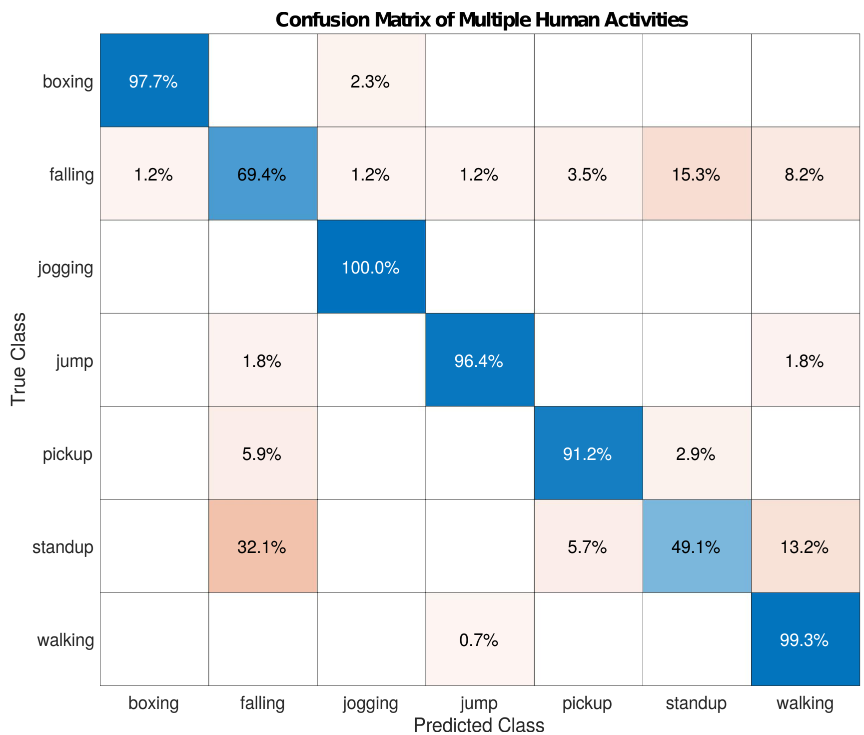


Figure 4. Accuracy of Multiple Human Activities Detecting

Average Inference Time Complexity:

RLDE + LSTM: 0.06042 sec

3DCNN: 7.336 sec

Conclusion & Future Work

- Summary
  - Radar signal domain contains enough info for a variety of vision tasks, while have the feature of privacy preserving
  - Introducing deep learning schemes with rich prior constraints of radar signal can potentially achieve better performances
  - This is an initial work that shows promising results
- Future Work
  - Larger data set with richer and fine granular labeling of human actions automatically and semiautomatically from cameras
  - Pontetial compressive sensing + deep learning to by-pass the radar signal processing pipeline after ADC