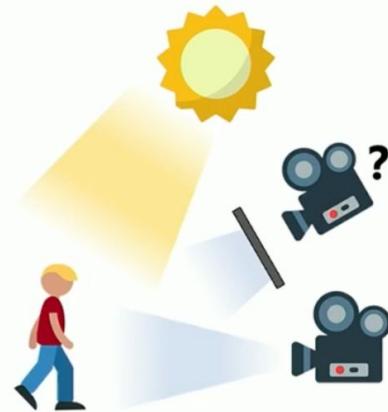


SLNet: A Spectrogram Learning Neural Network for Deep Wireless Sensing

Presentor: Kechen Liu

The limitations in vision systems



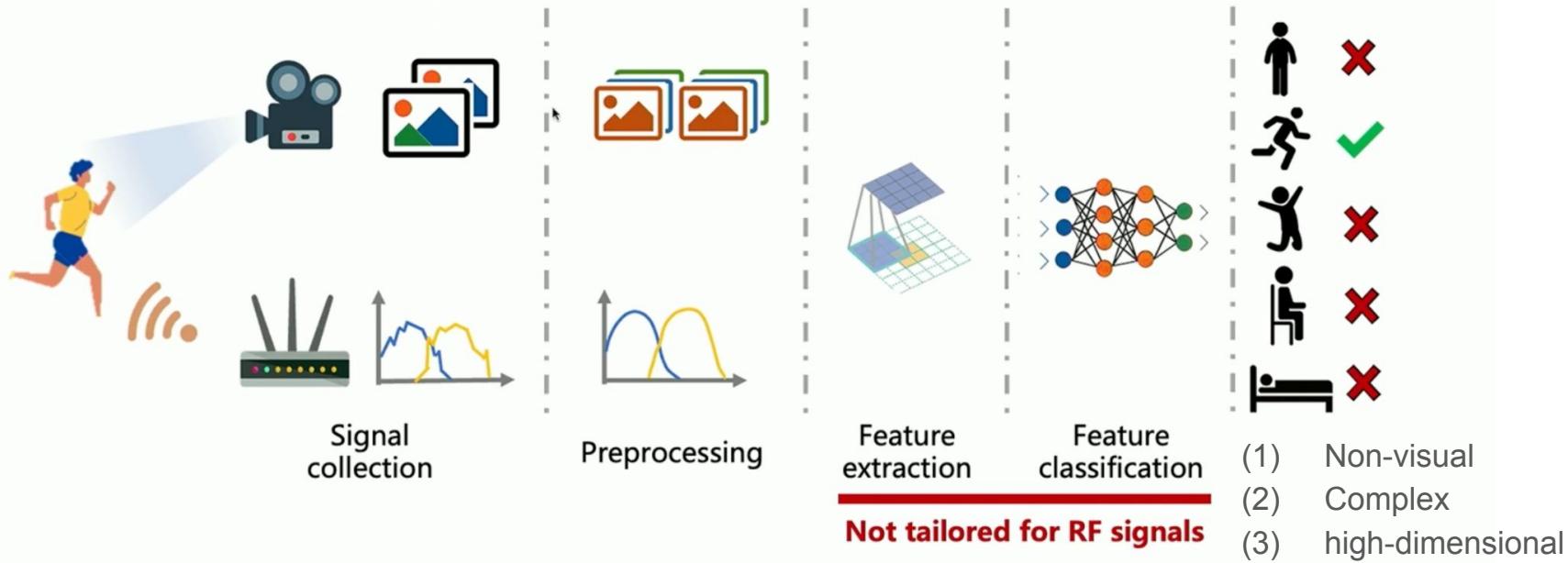
Occlusions



Privacy

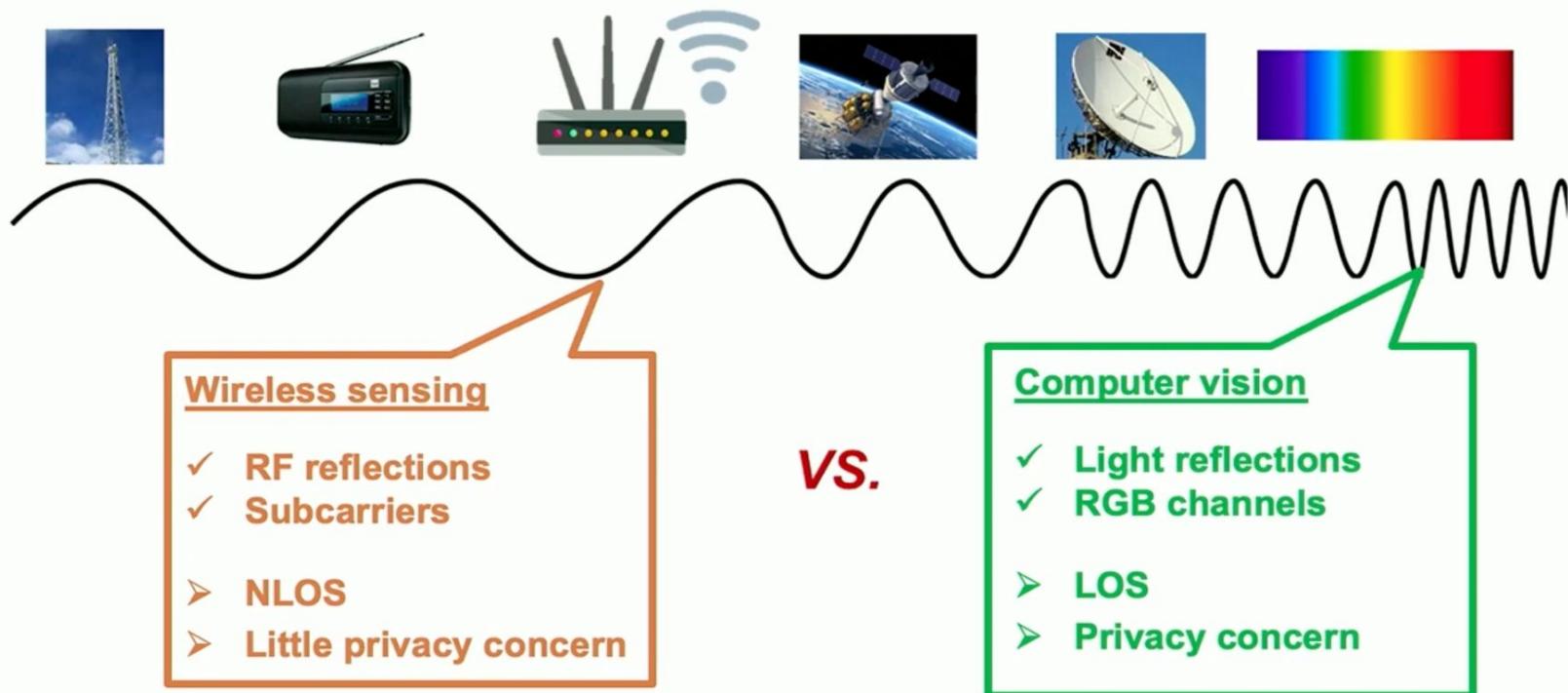
Despite being effective, CV technologies still has limitations regarding **occlusions** and **privacy** leakage.

Conventional DL-based wireless sensing



How to design specific deep neural networks for wireless signals?

The chance in wireless signals



Primer: Pre-processing of RF data

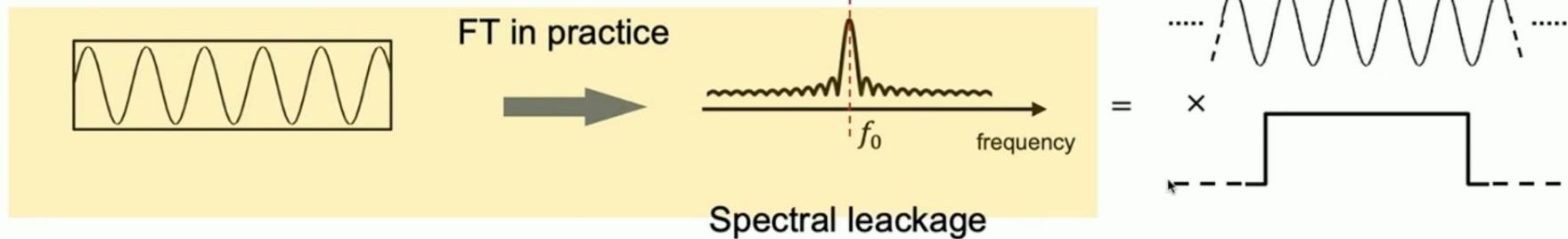
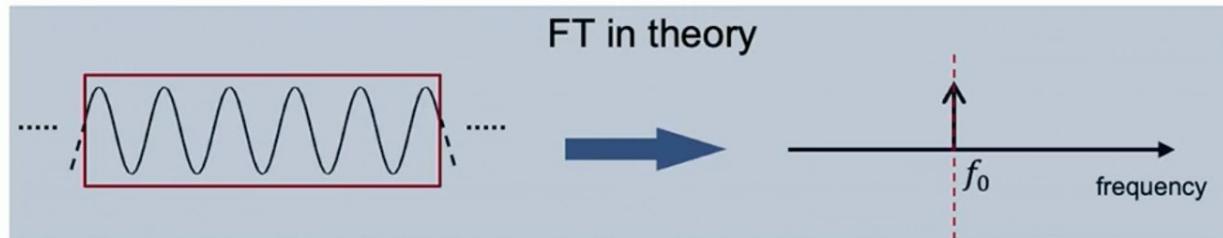
Measured CSI:

$$H(t) = H_s + \sum_{l=1}^L \alpha_l(t) e^{j2\pi \int_{-\infty}^t f_{D_l}(u) du} + n(t), \quad (1)$$

STFT processing:

$$S(f, t) = \text{STFT}[H(t)] = \text{FFT}[\omega] * \sum_{l=1}^L \alpha_l(t) \delta(f - f_{D_l}) + N(f, t), \quad (2)$$

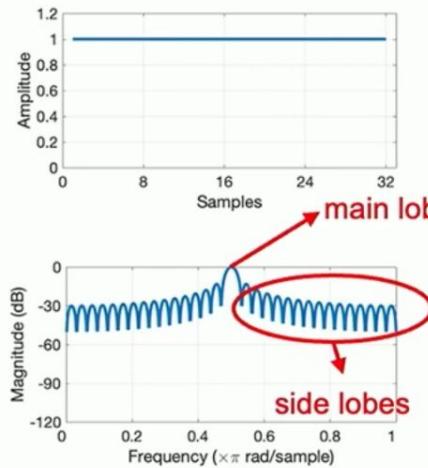
Challenges in wireless sensing



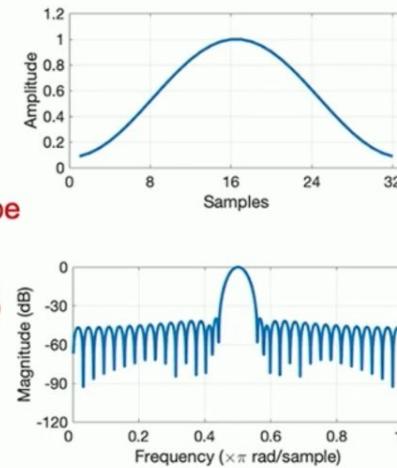
Practical Fourier transform gives an approximated while blurred version of the expected spectrograms.

Challenges in wireless sensing

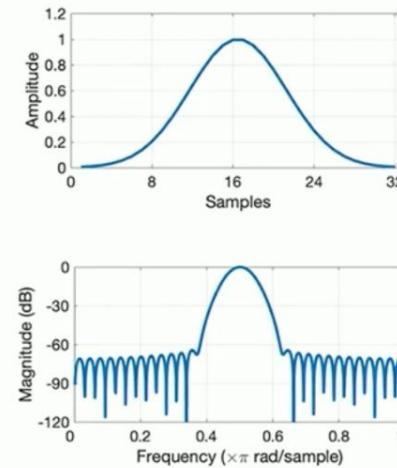
Spectral leakage can be mitigated by window functions.



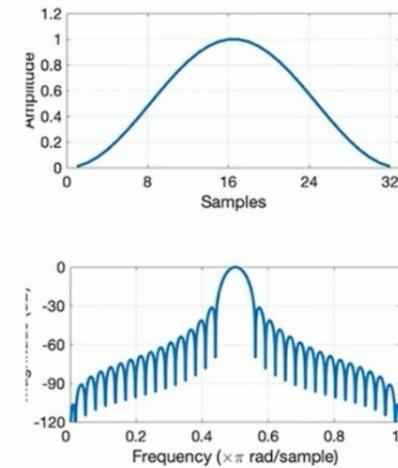
Rectangle



Hamming



Gaussian

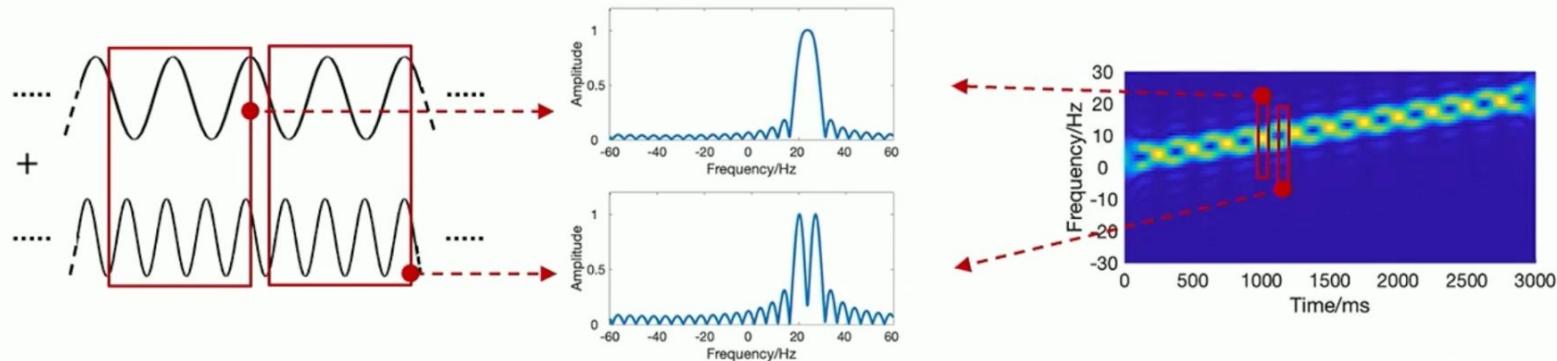


Hanning

Frequency resolution and amplitude resolution can hardly be balanced.

Challenges in wireless sensing

The interference caused by spectral leakage is **unstable**.

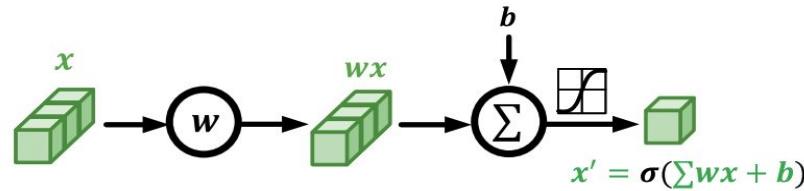


Different initial phase cause different interference patterns.

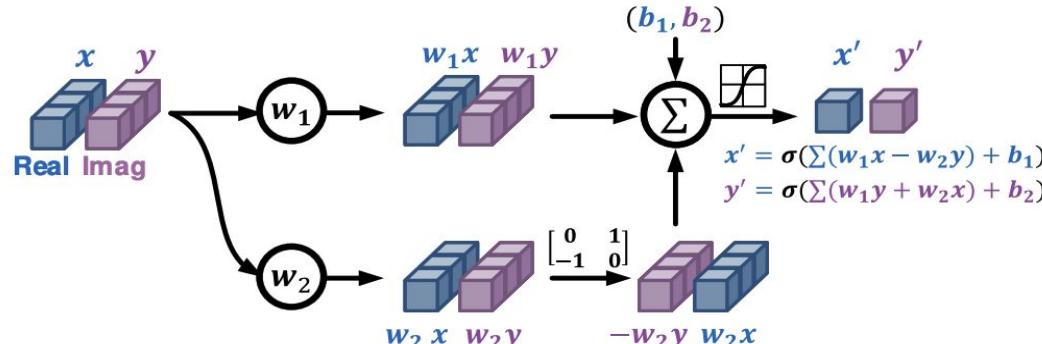


How to restore the ideal spectrums from the unpredictable interference?

Complex-valued neural network



(a)



(b)

SLnet architecture

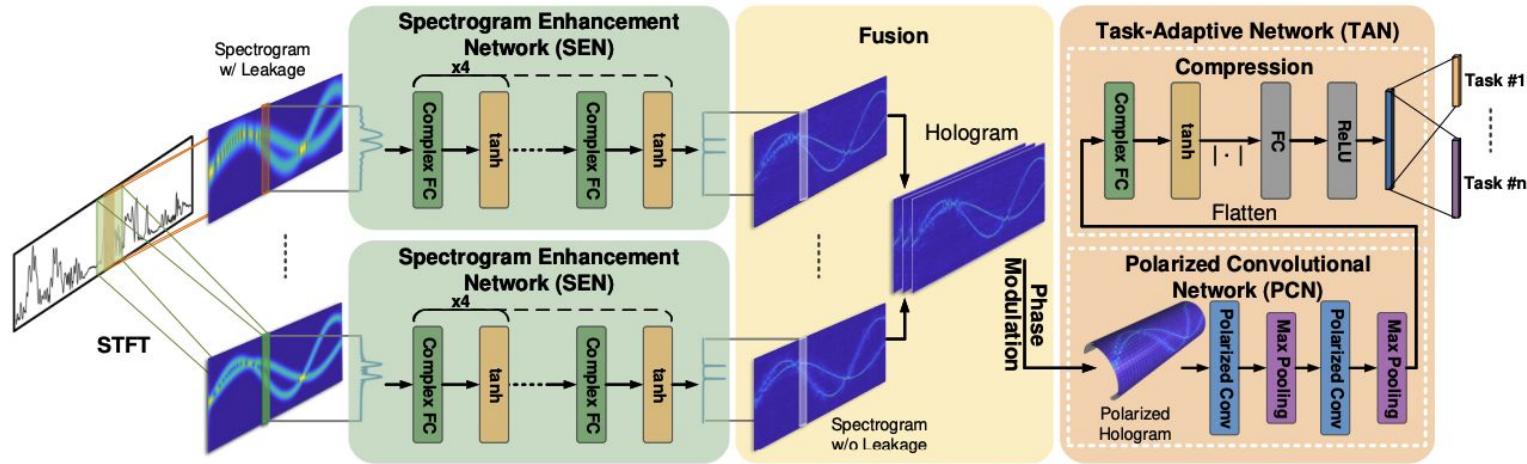
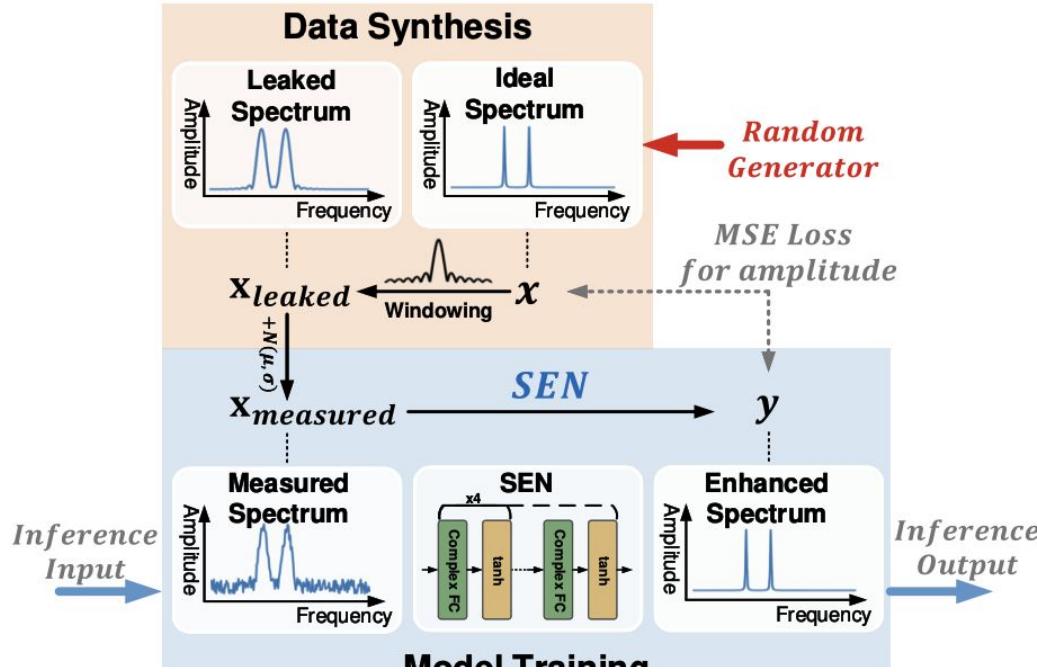


Figure 3: Overview of SLNET. The temporal CSI signal is transformed into spectrograms via a bank of STFT operators with different temporal and frequency resolutions. Each spectrogram is fed into the SEN to remove spectral leakage. Then, a hologram of spectrograms is generated by stacking all enhanced spectrograms and modulating them with linear phases. Next, the hologram is processed with the PCN to generate feature maps, and the compression networks to generate abstract features for specific learning tasks.

3.1 learning-assisted spectrogram enhancement



A compact complex-valued MLP with four fully connected layers and tanh activations.

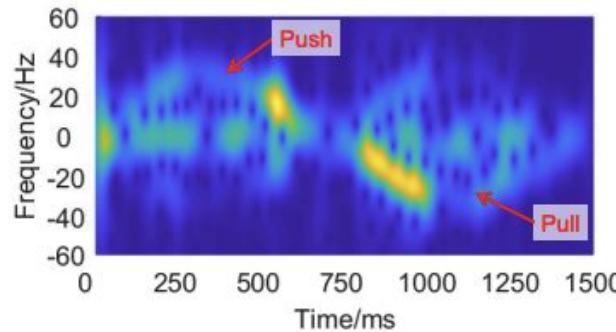
Data synthesis

Loss and normalization

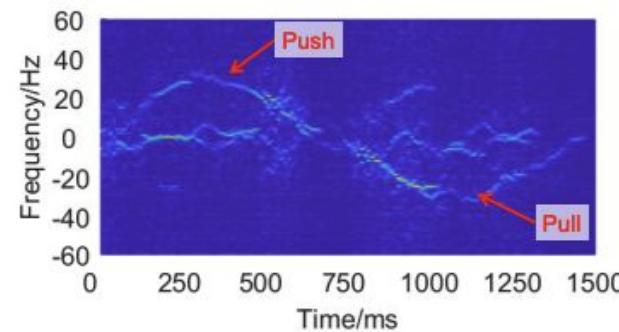
Per-window specialization

SEN results

Frequency components are more distinguishable after SEN enhancement.



(a)

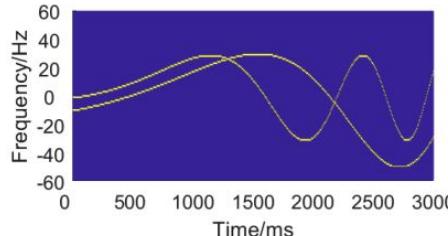


(b)

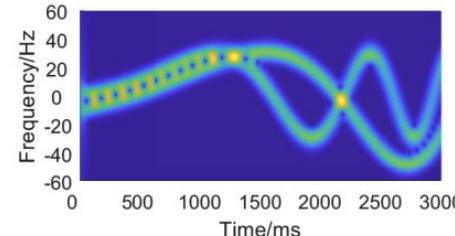
Figure 6: Illustration of the spectrogram of a pushing and pulling gesture. (a) The measured spectrogram and (b) the enhanced spectrogram from SEN.

3.2 multi-resolutions spectrogram fusions

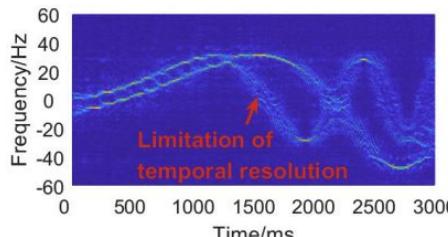
STFTs at multiple window sizes → SEN per window → Stack channels = “hologram” → (Pipeline) phase modulation → PCN



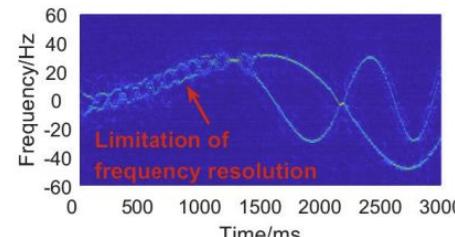
(a)



(b)

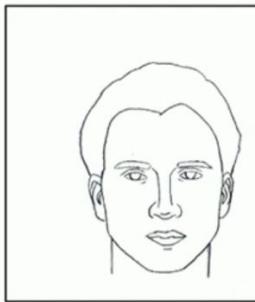
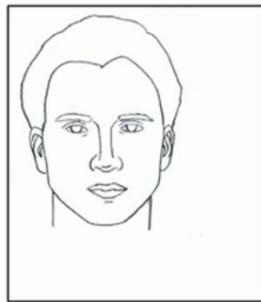


(c)

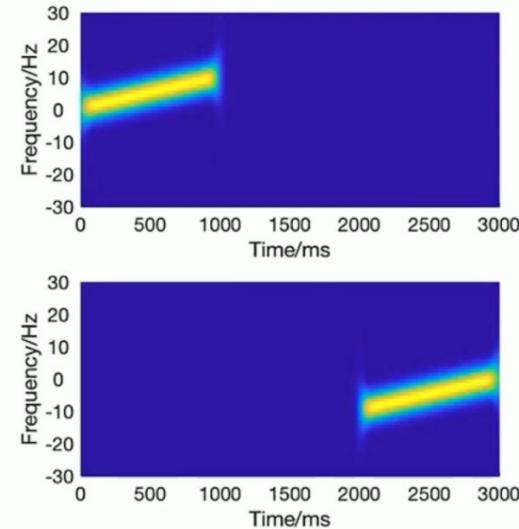


(d)

Challenges in deep learning



CNN is tailored for images since it is invariant to shifts.

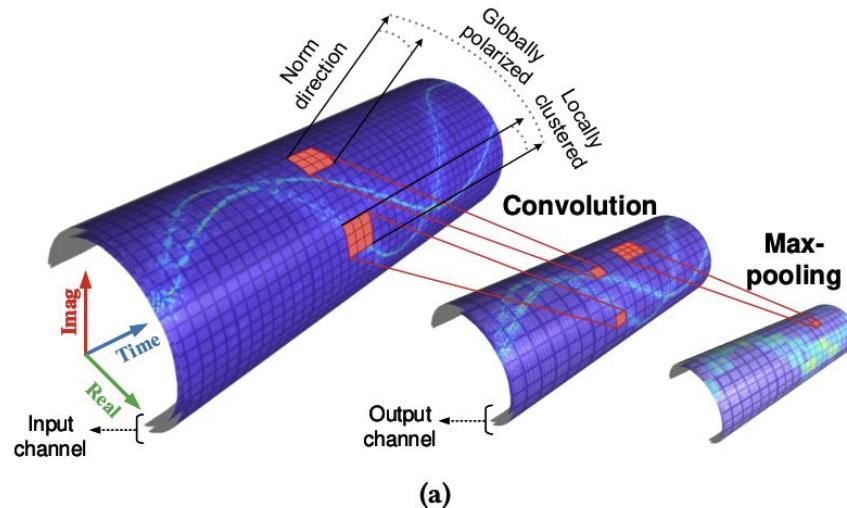


Wireless signals require global discriminations.

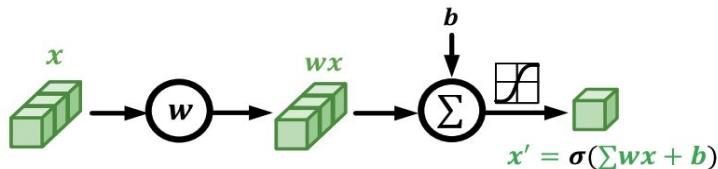
3.3 task-adaptive network

Phase polarized feature extraction

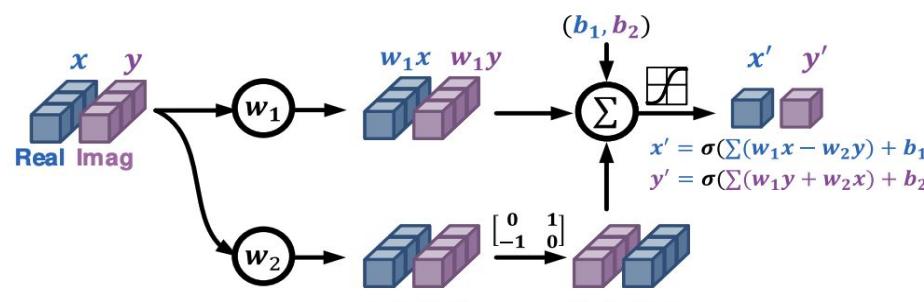
$$\phi_i = i \frac{\phi_h - \phi_l}{M} + \phi_l, \quad (6)$$



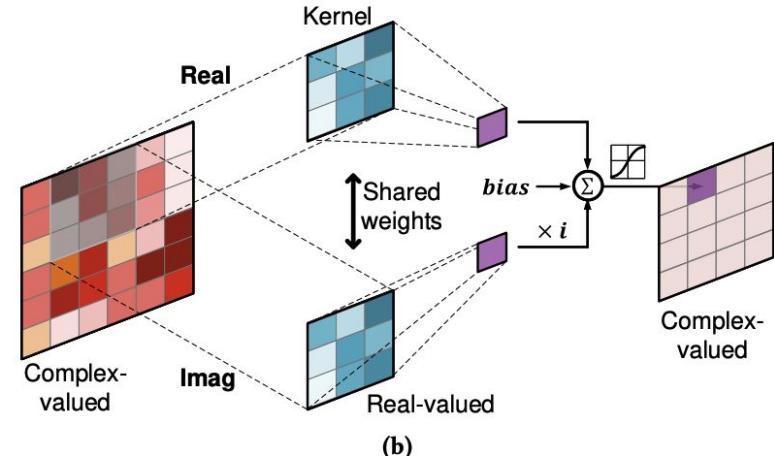
Complex-valued neural network



(a)



(b)



(b)

Review of the overall architecture

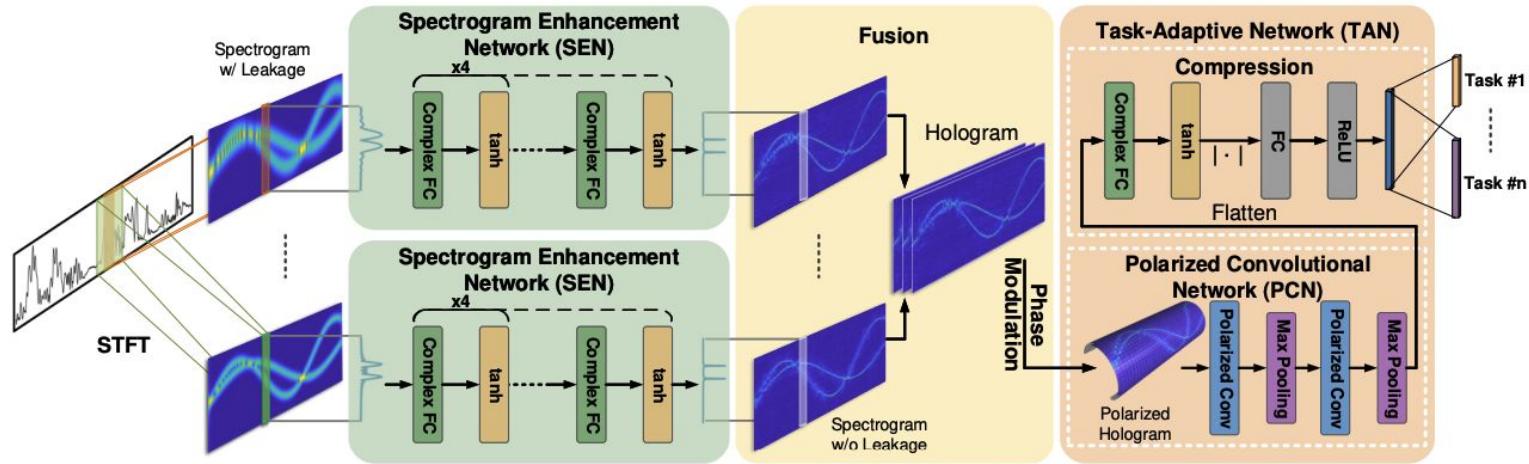


Figure 3: Overview of SLNET. The temporal CSI signal is transformed into spectrograms via a bank of STFT operators with different temporal and frequency resolutions. Each spectrogram is fed into the SEN to remove spectral leakage. Then, a hologram of spectrograms is generated by stacking all enhanced spectrograms and modulating them with linear phases. Next, the hologram is processed with the PCN to generate feature maps, and the compression networks to generate abstract features for specific learning tasks.

Feature compression

After PCN, a compact head reduces dimensionality and adapts to tasks:

1. **Complex FC + tanh**,
2. **Magnitude** (absolute value) bridge to real domain,
3. **Real FC + ReLU**, then optional task-specific layers (e.g., **softmax** for N-way gesture classes or **sigmoid** for fall probability)

Experiment setup

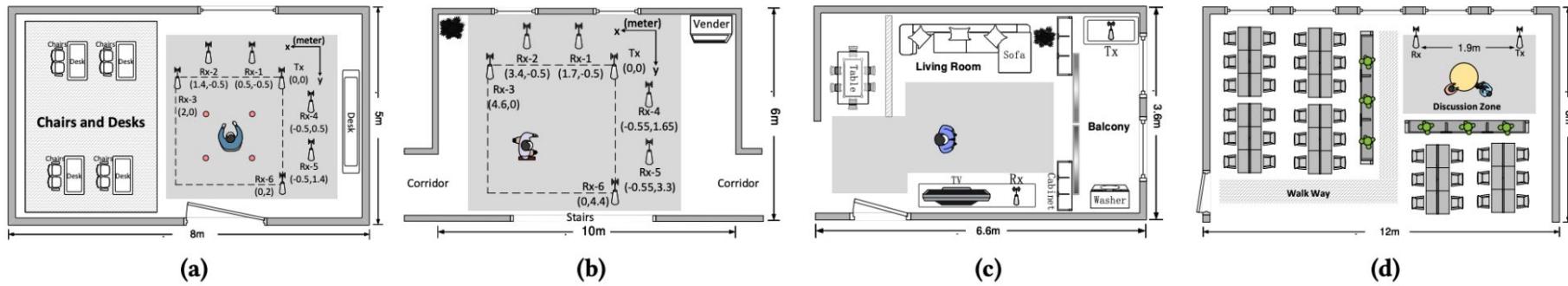
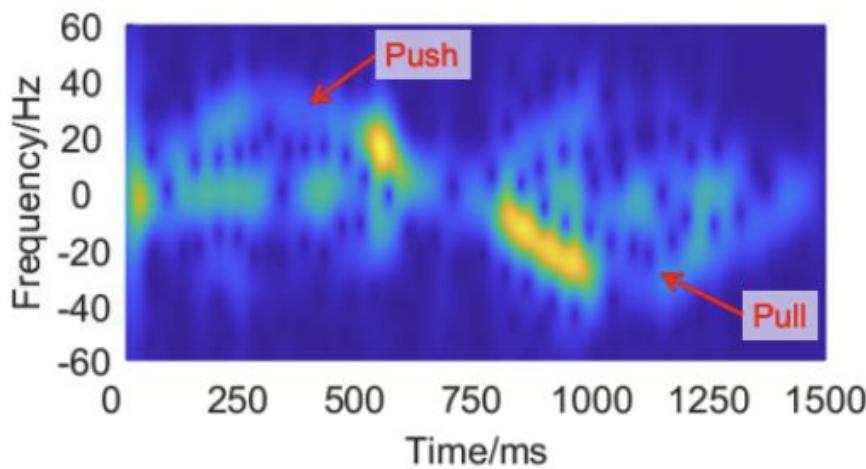
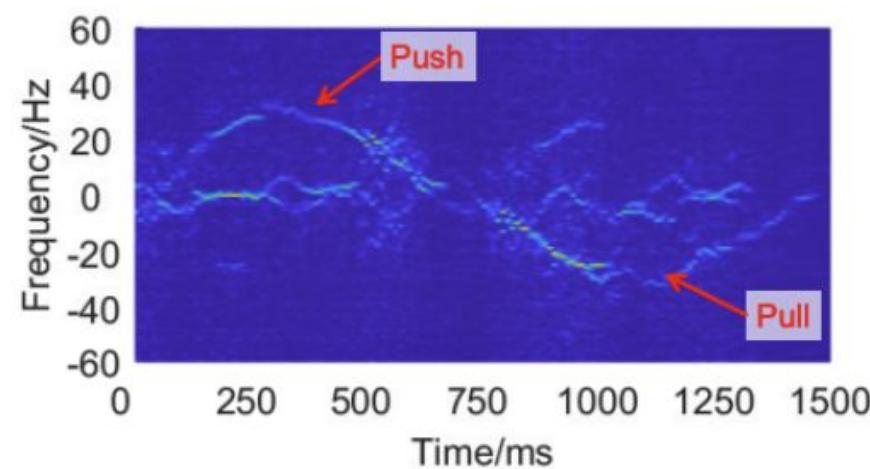


Figure 9: Experimental settings established in SLNET. (a) Classroom for gesture recognition. (b) Hall for gait identification. (c) Apartment for fall detection. (d) Office for breath estimation.

Performance—Human gestures

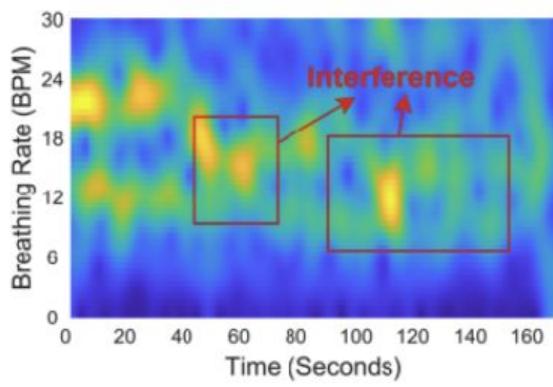


(a)

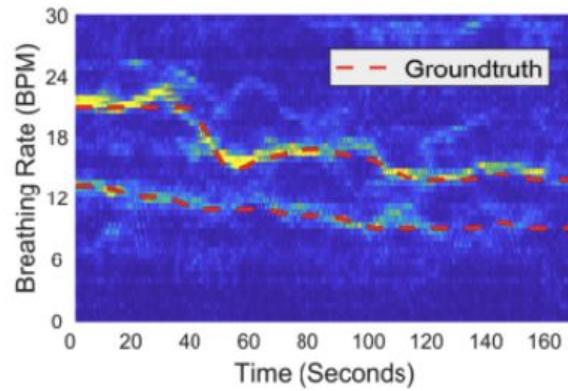


(b)

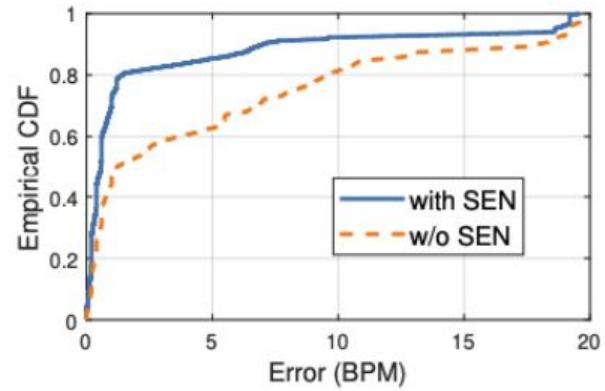
Performance—human breadth



(a)



(b)

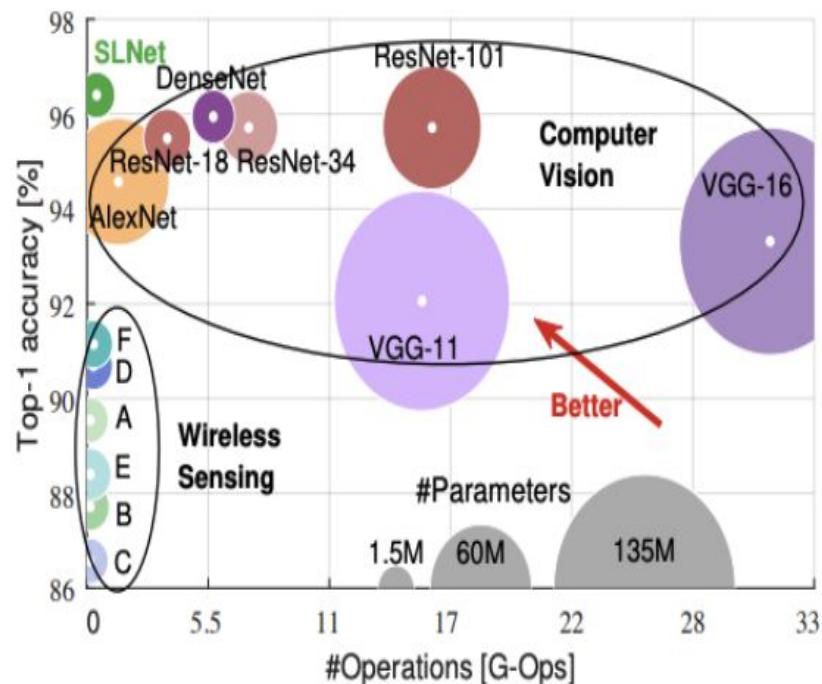


(c)

Performance-Recognition tasks

Modality	Ref.	Gesture	Gait	Fall ¹	Para ²
WiFi	[23, 90]	90.6%	95.1%	92.8%, 96.3%	1.07M
	[8, 22]	89.0%	96.6%	96.4%, 84.3%	2.72M
	[39, 79]	84.3%	83.3%	96.8%, 93.8%	5.77M
	[73] ³	78.9%	70.9%	95.5%, 96.8%	0.06M
FMCW	[87]	88.0%	95.4%	96.0%, 96.0%	1.06M
	[84, 86]	91.6%	96.4%	99.7%, 95.7%	2.76M
Acoustic	[30]	89.6%	95.4%	90.6%, 98.3%	6.08M
Vision	[40]	88.3%	90.1%	95.3%, 95.3%	128.8M
	[15]	91.9%	96.6%	97.0%, 95.6%	11.18M
	[20]	91.0%	97.7%	99.8%, 96.3%	6.96M
CVNN	[17, 32]	72.3%	96.0%	95.2%, 93.7%	115.6M
	[46]	92.0%	96.3%	98.4%, 93.8%	2.94M
WiFi	SLNET	96.6%	98.9%	99.8%, 97.2%	1.48M

Table 2: Comparison against 12 baseline models. ¹ The two metrics are precision and recall. ² Number of parameters in Million. ³ Trained with 10,000 epochs to converge.



Questions

Let's check Perusall comments!

My thoughts

Synthetic SEN training: Real hardware has quirks (CFO, I/Q imbalance). Will SEN still help if those don't match the simulator?

Generalization: Accuracy drops in new rooms/users—how robust is it across layouts and day-to-day variation?

Hyperparameters: The “phase polarization” strength seems important—how sensitive are results to that choice?