Seeing Around Street Corners: Non-Line-of-Sight Detection and Tracking In-the-Wild Using Doppler Radar

Nicolas Scheiner[*1] Florian Kraus[*1] Fangyin Wei[*2] Buu Phan[3]
Fahim Mannan[3] Nils Appenrodt[1] Werner Ritter[1] Jürgen Dickmann[1]
Klaus Dietmayer[4] Bernhard Sick[5] Felix Heide[2,3]

[1]Mercedes-Benz AG [2]Princeton University [3]Algolux [4]Ulm University [5]University of Kassel

Abstract

Conventional sensor systems record information about directly visible objects, whereas occluded scene components are considered lost in the measurement process. Non-line-of-sight (NLOS) methods try to recover such hidden objects from their indirect reflections – faint signal components, traditionally treated as measurement noise. Existing NLOS approaches struggle to record these low-signal components outside the lab, and do not scale to large-scale outdoor scenes and high-speed motion, typical in automotive scenarios. In particular, optical NLOS capture is fundamentally limited by the quartic intensity falloff of diffuse indirect reflections. In this work, we depart from visible-wavelength approaches and demonstrate detection, classification, and tracking of hidden objects in large-scale dynamic environments using Doppler radars that can be manufactured at low-cost in series production. To untangle noisy indirect and direct reflections, we learn from temporal sequences of Doppler velocity and position measurements, which we fuse in a joint NLOS detection and tracking network over time. We validate the approach on in-the-wild automotive scenes, including sequences of parked cars or house facades as relay surfaces, and demonstrate low-cost, real-time NLOS in dynamic automotive environments.

1. Introduction

Conventional sensor systems capture objects in their direct line of sight, and, as such, existing computer vision methods are capable of detecting and tracking only the visible scene parts [13, 15, 38, 37, 12, 23, 53, 30], whereas occluded scene components are deemed lost in the measurement process. Non-line-of-sight (NLOS) methods aim at recovering information about these occluded objects from their indirect reflections or shadows on visible scene surfaces, which are again in the line of sight of the detector. While performing scene understanding of occluded objects may enable applications across domains, including remote sensing or medical imaging, especially autonomous driving applications may benefit from detecting approaching traffic participants that are occluded.

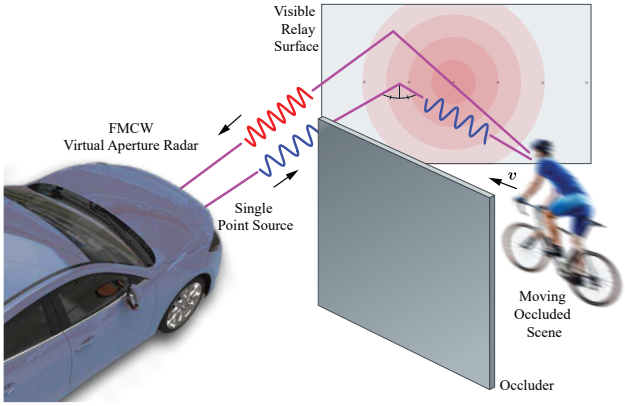


Figure 1: We demonstrate that it is possible to recover moving objects outside the direct line of sight in large automotive environments from Doppler radar measurements. Using static building facades or parked vehicles as relay walls, we jointly classify, reconstruct, and track occluded objects.

Existing NLOS imaging methods struggle outside controlled lab environments, and they struggle with large-scale outdoor scenes and high-speed motion, such as in typical automotive scenarios. The most successful NLOS imaging methods send out ultra-short pulses of light and measure their time-resolved returns [46, 34, 14, 8, 45, 5, 33, 29]. In contrast to a conventional camera, such measurements allow existing methods to unmix light paths based on their travel time [1, 21, 32, 34], effectively trading angular with temporal resolution. As a result, pulse widths and detection at a time scale of < 10 ps is required for room-sized scenes, mandating specialized equipment which suffers from low photon efficiency, high cost, and slow mechanical scanning. As intensity decreases quartically with the distance to the visible relay wall, current NLOS methods are limited to meter-sized scenes even when exceeding the eye-safety limits for a Class 1 laser (*e.g.* Velodyne HDL-64E) by a factor of 1000 [28]. Moreover, these methods are impractical for dynamic scenes as scanning and reconstruction takes up minutes [29, 5]. Unfortunately, alternative ap-

[*]Equal contribution.

proaches based on amplitude-modulated time-of-flight sensors [16, 18, 17] suffer from modulation bandwidth limitations and ambient illumination [25], and intensity-only methods [11, 42, 6] require highly reflective objects. Large outdoor scenes and highly dynamic environments remain an open challenge.

In this work, we demonstrate that it is possible to detect and track objects in large-scale dynamic scenes outside of the direct line-of-sight using automotive Doppler radar sensors, see Fig. 1. Departing from visible-wavelength NLOS approaches which rely on diffuse indirect reflections on the relay wall, we exploit the fact that specular reflections dominate on the relay wall for mm-wave radar signals, *i.e.*, when the structure size is an order of magnitude larger than the wavelength. As such, in contrast to optical NLOS techniques, phased array antenna radar measurements preserve the angular resolution and emitted radio frequency (RF) power in an indirect reflection, which enables us to achieve longer ranges. Conversely, separating direct and indirect reflections becomes a challenge. To this end, we recover indirectly visible objects relying on their Doppler signature, effectively suppressing static objects, and we propose a joint NLOS detection and tracking network, which fuses estimated and measured NLOS velocity over time. We train the network in an automated fashion, capturing training labels along with data with a separate positioning system, and validate the proposed method on a large set of automotive scenarios. By using facades and parked cars as reflectors, we show a first application of NLOS collision warning at urban intersections.

Specifically, we make the following contributions:

- We formulate an image formation model for Doppler radar NLOS measurements. Based on this model, we derive the position and velocity of an occluded object.
- We propose a joint NLOS detection and tracking network, which fuses estimated and measured NLOS velocity over time. For occluded object labeling, we acquire our data with an automated positioning system.
- We validate our system on in-the-wild automotive scenarios, and as a first application of this new imaging modality, demonstrate collision warning for vulnerable road users before seeing them in direct line of sight.
- The experimental training and validation data sets and models will be published[1].

2. Related Work

Optical Non-Line-of-Sight Imaging A growing body of work explores optical NLOS imaging techniques [34, 46, 14, 18, 33, 45, 5, 35, 50, 29]. Following Kirmani et al. [21], who first proposed the concept of recovering occluded objects from time-resolved light transport, these methods directly sample the temporal impulse response of a scene by sending out pulses of light and capturing their response using detectors with high temporal precision of $< 10\,\mathrm{ps}$, during which the pulses travel a distance of $3\,\mathrm{mm}$. While early work relies on costly and complicated streak camera setups [46, 47], a recent line of work uses single photon avalanche diodes (SPAD) [8, 33, 29]. Katz et al. [20, 19] demonstrate that correlations in the carrier wave itself can be used to realize fast single shot NLOS imaging, however, limited to scenes at microscopic scales [19].

Non-Line-of-Sight Tracking and Classification Several recent works use conventional intensity images for NLOS tracking and localization [22, 9, 10, 6, 11]. The ill-posedness of the underlying inverse problem limits these methods to localization with highly reflective targets [6, 11], sparse dark background, or only scenes with additional occluders present [42, 6]. Unfortunately, recent acoustic methods [27] are currently limited to meter-sized lab scenes and minutes of acquisition time. All of these existing methods have in common that they are impractical for large and dynamic outdoor environments.

Radio Frequency Non-Line-of-Sight Imaging A further line of work has explored imaging, tracking, and pose estimation through walls using RF signals [2, 3, 4, 39, 49, 52]. However, RF signals are not reflected when traveling through typical interior wall material, such as drywall, drastically simplifying through-the-wall vision tasks. As a result, only a few works have explored NLOS radar imaging and tracking [44, 36, 51]. These methods backpropagate raytraced high-order-bounce signals, which requires scenes with multiple known (although they are occluded) hidden relay walls. For the in-the-wild scenarios tackled in this work without prior scene knowledge, only third-bounce measurements, and with imperfect relay walls, *e.g.*, a parked sequence of vehicles, these methods are impractical. Moreover, traditional filtering and backprojection estimation suffers from large ambiguities at more than $10\,\mathrm{m}$ in the presence of realistic measurement noise [36]. In this work, we address this challenge with a data-driven joint detection and tracking method, allowing us to demonstrate practical NLOS detection in-the-wild using radar systems which have the potential for low-cost mass production in the near future.

3. Observation Model

When a radar signal gets reflected off a visible wall onto a hidden object, some of the signal is scattered and reflected back to the wall where it can be observed, see Fig. 2. Next, we derive a forward model including such observations.

[1] https://github.com/princeton-computational-imaging/doppler_nlos for code and models.

3.1. Non-Line-of-Sight FMCW Radar

Radar sensors emit electromagnetic (EM) waves, traveling at the speed of light c, which are reflected by the scene and received by the radar sensors. In this work, we use a frequency-modulated continuous-wave (FMCW) Doppler radar with multiple input multiple output (MIMO) array configuration, which can resolve targets in range r, azimuthal angle ϕ, and radial Doppler velocity v_r. Instead of a single sinusoidal EM wave, our FMCW radar sends out linear frequency sweeps [7] over a frequency band B starting from the carrier frequency f_c, that is

$$g(t) = \cos\left(2\pi f_c t + \pi \frac{B}{T_m} t^2\right), \tag{1}$$

with T_m being the sweep rise time. The instantaneous frequency of this signal is $1/2\pi \, d/dt \left(2\pi f_c t + \pi \, B/T_m t^2\right) = f_c + B/T_m t$, that is a linear sweep varying from f_c to $f_c + B$ with slope B/T_m, which is plotted in Fig. 3.

The emitted signal g propagates through the visible and occluded parts of the scene, that is, this signal is convolved with the scene's impulse response. For a given emitter position $\mathbf{l}$ and receiver position $\mathbf{c}$ the received signal becomes

$$s(t, \mathbf{c}, \mathbf{l}, \mathbf{w}) = \int_\Lambda \alpha(\mathbf{x}) \, \rho(\mathbf{x} - \mathbf{w}, \mathbf{w} - \mathbf{x}) \cdot \tag{2}$$
$$\frac{1}{(r_{\mathbf{lw}} + r_{\mathbf{xw}})^2} \frac{1}{(r_{\mathbf{xw}} + r_{\mathbf{wc}})^2} \, g\left(t - \frac{r_{\mathbf{lw}} + 2\, r_{\mathbf{xw}} + r_{\mathbf{wc}}}{c}\right) d\sigma(\mathbf{x}),$$

see Fig. 2, with $\mathbf{w}$ and $\mathbf{x}$ being the positions on the relay wall and the object surface Λ, the surface measure σ on Λ, α as the albedo, and ρ denoting the bi-directional reflectance distribution function (BRDF), which depends on the incident direction $\omega_\mathbf{i} = \mathbf{x} - \mathbf{w}$ and outgoing direction $\omega_\mathbf{o} = \mathbf{w} - \mathbf{x}$. The distance r describes here the distance between the subscript positions, and its squared inverse in Eq. (2) models the intensity falloff due to spherical travel, which we approximate as not damped by the specular wall, and diffuse backscatter from object back to the receiver $\mathbf{c}$.

Reflection Model The scattering behavior ρ depends on the surface properties. Surfaces that are flat, relative to the wavelength λ of $\approx 5\,\mathrm{mm}$ for typical 76 GHz-81 GHz automotive radars, will result in a specular response. As a result, the transport in Eq. (2) treats the relay wall as a mirror, see Fig. 2. We model the reflectance of the hidden and directly visible targets following [11] with a diffuse and specular term as

$$\rho(\omega_\mathbf{i}, \omega_\mathbf{o}) = \alpha_d \, \rho_d(\omega_\mathbf{i}, \omega_\mathbf{o}) \; + \underbrace{\alpha_s \, \rho_s(\omega_\mathbf{i}, \omega_\mathbf{o})}_{\approx 0}. \tag{3}$$

In contrast to recent work [11, 27], we cannot rely on the specular component ρ_s, as for large standoff distances, the relay walls are too small to capture the specular reflection. Indeed, completely specular facet surfaces are used as "stealth" technology to hide targets [31]. As retroreflective radar surfaces are extremely rare in nature [39], the diffuse part ρ_d dominates ρ. Note that $\alpha(\mathbf{x})\rho(\mathbf{x} - \mathbf{w}, \mathbf{w} - \mathbf{x})$ in Eq. (2) is known as the intrinsic radar albedo, describing backscatter properties, *i.e.*, the radar cross section [41].

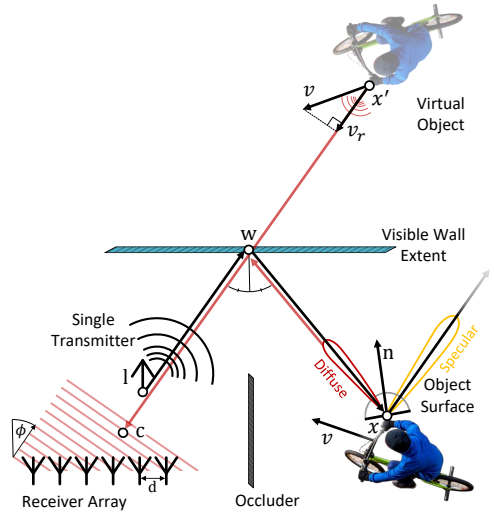


Figure 2: Radar NLOS observation. For mm-wavelengths, typical walls appear flat, and reflect radar waves specularly. We measure distance, angle and Doppler velocity of the indirect diffuse backscatter of an occluded object to recover its velocity, class, shape, and location.

Range Measurement Assuming an emitter and detector position $\mathbf{c} = \mathbf{l} = \mathbf{w}$ and a static single target ξ at distance $r = \|\mathbf{c} - \mathbf{x}\|$ with roundtime reflection $\tau_\xi = 2r/c$, Eq. (2) becomes a single sinusoid

$$s_\xi(t) = \alpha_\xi g(t - \tau_\xi), \tag{4}$$

where α_ξ describes here the accumulated attenuation along the reflected path. FMCW radars mix the received signal s_ξ with the emitted signal g, resulting in a signal p_ξ consisting of the sum and a difference of frequencies. The sum is omitted due to low-pass filtering in the mixing circuits, *i.e.*:

$$p_\xi(t) = s_\xi(t) \cdot g(t) \approx \frac{\alpha_\xi}{2} \cos\left(2\pi f_\text{beat} t + 4\pi \frac{f_c r}{c}\right). \tag{5}$$

The remaining difference due to the time difference between transmitted and received chirp, see Supplemental Material, results in a frequency shift with beat frequency

$$f_\text{beat} = \frac{B}{T_m} \frac{2r}{c}, \qquad \text{and} \qquad r = c \, \frac{f_\text{beat} T_m}{2B}. \tag{6}$$

The range can be estimated from this beat note according to Eq. (6). To this end, FMCW radar systems perform a Fourier analysis, where multiple targets with different path lengths (Eq. (2)) appear in different beat frequency bins.

Doppler Velocity Estimation For the case when the object is moving, radial movement v_r along the reflection path results in an additional Doppler frequency shift in the received signal

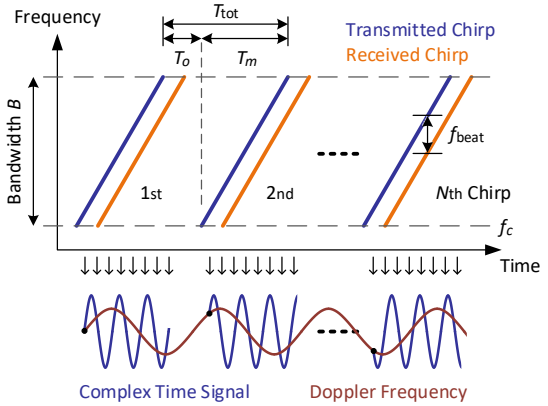


Figure 3: Chirp sequence modulation principle for a single receiver-transmitter antenna: N consecutive frequency ramps are sent and received with a frequency shift f_{beat} corresponding to the distance of the reflector. Each frequency ramp is sampled and the phase of the received signal is estimated at each chirp and range bin. The phase shift between consecutive chirps corresponds to the Doppler frequency.

$$f_{\text{Doppler}} = 2 \cdot \frac{v_r}{\lambda}. \tag{7}$$

To avoid ambiguity between a frequency shift due to round-trip travel opposed to relative movement, the ramp slope B/T_m is chosen high, so that Doppler shifts are negligible in Eq. (6). Instead, this information is recovered by observing the phase shift θ in the signals between two consecutive chirps with spacing T_{tot}, see Fig. 3, that is

$$v_r = \frac{\lambda \cdot \theta}{4\pi \cdot T_{\text{tot}}} = \mathbf{v} \cdot \frac{\mathbf{x}' - \mathbf{c}}{\|\mathbf{x}' - \mathbf{c}\|}. \tag{8}$$

This velocity estimate is the radial velocity, see Fig. 2. Akin to the range estimation, the phase shift θ (and velocity) is also estimated by a Fourier analysis, but applied on the phasors of N sequential chirps for each range bin separately.

Incident Angle Estimation To resolve incident radiation directionally, radars rely on an array of antennas. Under a far field assumption, *i.e.*, $r \gg \lambda$, for a single transmitter and target, the incident signal is a plane wave. The incident angle ϕ of this waveform causes a delay of arrival $d\sin(\phi)/c$ between the two consecutive antennas with distance d, see Fig. 2, resulting in a phase shift $\Omega = 2\pi d \sin(\phi)/\lambda$. Hence, we can estimate

$$\phi = \arcsin \frac{\Omega\lambda}{2\pi d}. \tag{9}$$

For this angle estimation, a single transmitter antenna illuminates and all receiver antennas listen. A frequency analysis on the sequence of phasors corresponding to peaks in the 2D range-velocity spectrum assigns angles, resulting in a 3D range-velocity-angle data cube.

3.2. Sensor Post-Processing

The resulting raw 3D measurement cube contains $1024 \times 512 \times 64$ bins for range, angle, and velocity, respectively. For low-reflectance scenes, typical noise, and clutter, tens of millions of non-zero reflection points can be measured. To tackle such measurement rates in real-time, we implement a constant false alarm rate filter to detect high RCS values $\tilde{\sigma}$ following [40]. We retrieve a radar point cloud $\tilde{\mathbf{U}}$ with less than 10^4 points, allowing for efficient inference:

$$\tilde{\mathbf{U}} = \left\{(\tilde{\phi}, \tilde{r}, \tilde{v_r}, \tilde{\sigma})_i \mid 1 \leq i \leq R\right\} \text{ with } R < 10^4. \tag{10}$$

See Supplemental Material for details on post-processing.

4. Joint NLOS Detection and Tracking

In this section, we propose a neural network for the detection and tracking of hidden objects from radar data.

4.1. Non-Line-of-Sight Detection

The detection task is to estimate oriented 2D boxes for pedestrians and cyclists, given a Bird's-eye-view (BEV) point cloud $\tilde{\mathbf{U}}$ as input. The overall detection pipeline consists of three main stages: (1) input parameterization that converts a BEV point cloud into a sparse pseudo-image; (2) high-level representation encoding from the pseudo-image using a 2D convolutional backbone; and (3) 2D bounding box regression and detection with a detection head.

Input Parameterization We denote $\mathbf{u}$ as a d-dimensional ($d = 4$) point in a raw radar point cloud $\tilde{\mathbf{U}}$ with coordinates x, y (derived from the polar coordinates $\tilde{\phi}, \tilde{r}$), velocity $\tilde{v_r}$, and amplitude $\tilde{\sigma}$. We use the logarithm of the amplitude to get an intensity measure $s = \log \tilde{\sigma}$. As a first step, the point cloud is discretized into an evenly spaced grid in the x-y plane, resulting in a pseudo-image of size $(d-2, H, W)$ where H and W indicate the height and width of the grid.

High-level Representation Encoding To efficiently encode high-level representations of the hidden detections, the backbone network contains two modules: a pyramid network and a zoom-in network. The pyramid network contains two consecutive stages to produce features at increasingly small spatial resolutions. Each stage downsamples its input feature map by a factor of two using three 2D convolutional layers. Next, a zoom-in network upsamples and concatenates the two feature maps from the pyramid network. This zoom-in network performs transposed 2D convolutions with different strides. As a result, both upsampled features have the same size and are then concatenated to form the final output. All (transposed) convolutional layers use kernels of size 3 and are interlaced with BatchNorm and ReLU, see Supplemental Material for details.

Detection Head The detection head follows the setup of Single Shot Detector (SSD) [26] for 2D object detection. Specifically, each anchor predicts a 3-dimensional vector for classification (background / cyclist / pedestrian) and a

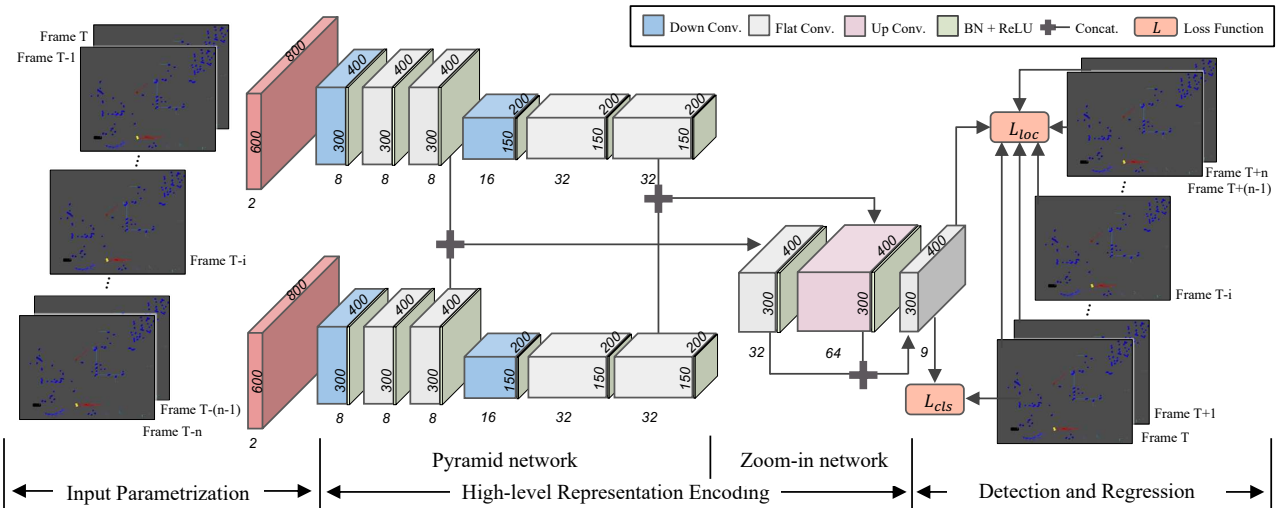


Figure 4: NLOS detection and tracking architecture. The network accepts the current frame T and the past n radar point cloud data as input, and outputs predictions for frame T and the following n frames. The features are downsampled twice in the pyramid network, and then upsampled and concatenated by the zoom-in network. We merge the features from different frames at both levels to encode high-level representation and fuse temporal information.

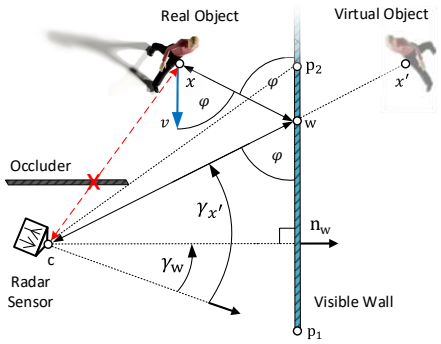


Figure 5: NLOS geometry and velocity estimation from indirect specular wall reflections. The hidden velocity v can be reconstructed from the radial velocity v_r by assuming that the road user moves parallel to the wall, *i.e.*, on a road.

6-dimensional vector for bounding box regression (center, dimension, orientation, and velocity of the box).

Relay Wall Estimation We use first-response pulsed lidar measurements of a separate front-facing lidar sensor to recover the geometry of the visible wall. Specifically, we found that detecting line segments in a binarized binned BEV is robust using [48], where each bin with size $0.01\,\mathrm{m}$ is binarized with a threshold of 1 detection per bin. We filter out segments with a length shorter than $1\,\mathrm{m}$, constraining the detected wall to smooth surfaces that our NLOS model holds for, see Supplemental Material. Each segment is represented by its endpoints $\mathbf{p}_1$ and $\mathbf{p}_2$, cf. Fig. 5.

Third-Bounce Geometry Estimation Next, we derive the real location $\mathbf{x}$ of a third-bounce or virtual detection $\mathbf{x}'$, for reference see Fig. 2 and Fig. 5. In order to decide whether a point is a virtual detection, we first derive its intersection $\mathbf{w}$ with the relay wall $\mathbf{p} = \mathbf{p}_2 - \mathbf{p}_1$, that is

$$\mathbf{w} = \mathbf{c} + \frac{(\mathbf{p}_1 - \mathbf{c}) \times \mathbf{p}}{(\mathbf{x}' - \mathbf{c}) \times \mathbf{p}} (\mathbf{x}' - \mathbf{c}), \tag{11}$$

where $\times$ is the 2D cross product $\mathbf{a} \times \mathbf{b} = a_1 b_2 - a_2 b_1$. For a detection $\mathbf{x}'$ to be a third-bounce detection, we have two criteria. First, $\mathbf{x}'$ and the receiver $\mathbf{c}$ must be on opposite sides of the relay wall. We define the normal of the relay wall $\mathbf{n}_\mathbf{w}$ as pointing away from the receiver $\mathbf{c}$. Second, the intersection $\mathbf{w}$ must be between $\mathbf{p}_1$ and $\mathbf{p}_2$, both expressed as

$$\begin{aligned} \mathbf{n}_\mathbf{w} \cdot (\mathbf{x}' - \mathbf{p}_1) \geq 0 \wedge \|\mathbf{w} - \mathbf{p}_1\| \leq \|\mathbf{p}\| \\ \wedge \|\mathbf{w} - \mathbf{p}_2\| \leq \|\mathbf{p}\| . \end{aligned} \tag{12}$$

The first term is the signed distance, indicating whether $\mathbf{x}'$ and $\mathbf{c}$ are on opposite sides of the wall and the other terms determine whether $\mathbf{w}$ lies between $\mathbf{p}_1$ and $\mathbf{p}_2$. If Eq. (12) is true, *i.e.*, $\mathbf{x}'$ is a third-bounce detection, we reconstruct the original point $\mathbf{x}$ as

$$\mathbf{x} = \frac{\left(\mathbf{w} - \mathbf{c} - 2\left(\mathbf{n}_\mathbf{w} \cdot (\mathbf{w} - \mathbf{c})\right) \mathbf{n}_\mathbf{w}\right) \|\mathbf{w} - \mathbf{x}'\|}{\|\mathbf{w} - \mathbf{c}\|}. \tag{13}$$

Third-Bounce Velocity Estimation After recovering $\mathbf{x}$, we estimate the real velocity vector $\mathbf{v}$ under the assumption that the real velocity is parallel to the relay wall, see Fig. 5. Specifically, it is

$$\mathbf{v} = \|\mathbf{v}\| \operatorname{sgn}(v_r) \cdot \operatorname{sgn}(\gamma_{\mathbf{x}'} - \gamma_\mathbf{w}) \frac{\mathbf{p}}{\|\mathbf{p}\|}. \tag{14}$$

Here, $\gamma_{\mathbf{x}'}$ and $\gamma_\mathbf{w}$ are the angles of $\mathbf{x}' - \mathbf{c}$ and $\mathbf{n}_\mathbf{w}$ relative to the sensor's coordinate system, respectively. In Eq. (14), the sign of v_r distinguishes approaching and departing hidden object targets, while $\operatorname{sgn}(\gamma_{\mathbf{x}'} - \gamma_\mathbf{w})$ determines the object's allocation to the left or right half-plane with respect to the normal $\mathbf{n}_\mathbf{w}$. By convention, we define that $\mathbf{p}$ is rotated $\frac{\pi}{2}$ counterclockwise relative to $\mathbf{n}_\mathbf{w}$. Using the measured radial velocity $v_r = \|\mathbf{v}\| \cdot |\cos\varphi|$, we get

$$\mathbf{v} = \operatorname{sgn}(v_r) \cdot \operatorname{sgn}(\gamma_{\mathbf{x}'} - \gamma_{\mathbf{w}}) \cdot \frac{|v_r|}{|\cos\varphi|} \cdot \frac{\mathbf{p}}{\|\mathbf{p}\|}, \quad (15)$$

with φ being the angle between $\mathbf{x}' - \mathbf{c}$ and $\mathbf{v}$, cf. Fig. 5. See the Supplemental Material for detailed derivations.

4.2. Non-Line-of-Sight Doppler Tracking

Our model jointly learns tracking with future frame prediction, inspired by Luo et al. [30]. At each timestamp, current and its n preceding frames form the input, and predictions are for the current plus the following n future frames.

One of the main challenges is to fuse temporal information. A straightforward solution is to add another dimension and perform 3D convolutions over space and time. However, this approach is not memory-efficient and computationally expensive given the sparsity of the data. Alternatives can be early or late fusion as discussed in [30]. Both fusion schemes first process each frame individually, and then start to fuse all frames together.

Instead of such one-time fusion, our approach leverages the multi-scale backbone and performs fusion at different levels. Specifically, we first perform separate input parameterization and high-level representation encoding for each frame as described in Sec. 4.1. After the two stages of the pyramid network, we concatenate the $n+1$ feature maps along the channel dimension for each stage. This results in two feature maps of sizes $\left((n+1)C_1, \frac{H}{2}, \frac{W}{2}\right)$ and $\left((n+1)C_2, \frac{H}{4}, \frac{W}{4}\right)$, which are then concatenated as inputs to the two upsampling modules of the zoom-in network, respectively. The rest of the model is the same as before. By aggregating temporal information across $n+1$ frames at different scales, the model is allowed to capture both low-level per-frame details and high-level motion features. We refer to Fig. 4 for an illustration of our architecture.

4.3. Loss Functions

Our overall objective function contains a localization term and a classification term

$$L = \alpha L_{loc} + \beta L_{cls}. \quad (16)$$

The localization loss is a sum of the localization loss for the current frame T and n frames into the future:

$$L_{loc} = \sum_{t=T}^{T+n} L_{loc_t} \quad \text{with} \quad L_{loc_t} = \sum_{u \in \{x,y,w,l,\theta,v\}} \alpha_u |\Delta u|, \quad (17)$$

where Δu is the localization regression residual between ground truth (gt) and anchors (a) defined by (x, y, w, l, θ, v):

$$\Delta x = x^{gt} - x^a, \quad \Delta y = y^{gt} - y^a, \quad \Delta v = v^{gt} - v^a,$$
$$\Delta w = \log \frac{w^{gt}}{w^a}, \quad \Delta l = \log \frac{l^{gt}}{l^a}, \quad \Delta\theta = \sin(\theta^{gt} - \theta^a). \quad (18)$$

We do not distinguish the front and back of the object, therefore all θ's are within the range $[-\frac{\pi}{2}, \frac{\pi}{2})$. For classification, we adopt the focal loss L_{cls} from [26].

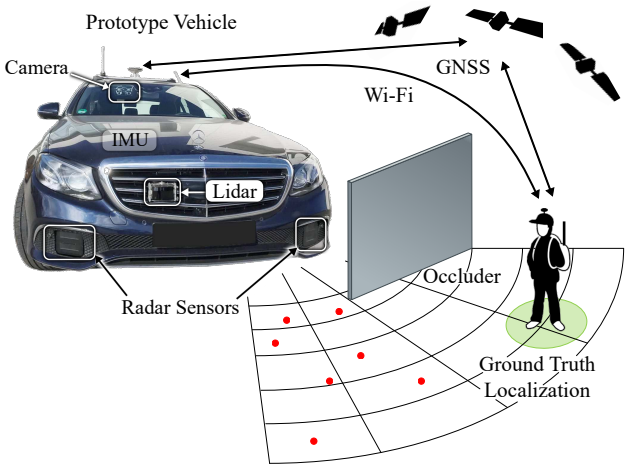
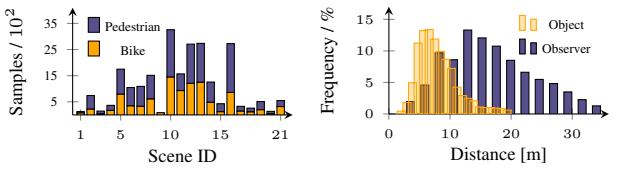


Figure 6: Prototype vehicle with measurement setup (top left) and automated ground-truth localization system (right). To acquire training data in an automated fashion, we use GNSS and IMU for a full pose estimation of ego-vehicle and the hidden vulnerable road users.

5. Data Acquisition and Training

Prototype Vehicle Setup The observation vehicle prototype is shown in Fig. 6. We use experimental FMCW radar prototypes, mounted in the front bumper, with frequency band 76 GHz to 77 GHz and chirp sequence modulation, see Sec. 3. We use a mid-range configuration with 153 m maximum range and FoV of 140°, *i.e.*, for urban scenarios or intersections. A single measurement takes 22.6 ms, with a resolution of 0.15 m, 1.8°, and 0.087 m s^{-1}. Similar sensors are available as development kits for a few hundred USD, *e.g.* Texas Instruments AWR1642BOOST; the mass-produced version costing a small fraction. The radar sensors are complemented by an experimental 4-layer scanning lidar with 0.25° and 0.8° resolution in azimuth and elevation. With a wide FoV of 145°, a single sensor installed in the radiator grill suffices for our experiments. We use a GeneSys ADMA-G PRO localization system consisting of a combined global navigation satellite system (GNSS) receiver and an inertial measurement unit (IMU) to track ego-pose using an internal Kalman filter. The system has an accuracy of up to 0.8 cm and 0.01 m s^{-1} for the position and velocity. For documentation purposes, we use a single AXIS F1015 camera with 97° FoV behind the test vehicle's windshield. See Supplemental Material for details on all sensors along with required coordinate system transforms.

Automated Ground-Truth Estimation Unfortunately, humans are not accustomed to annotating radar measurements, and NLOS annotations are even more challenging. We tackle this problem by adopting a variant of the tracking device from [43]. We equip vulnerable road users, *i.e.*, occluded pedestrians or bicyclists, with a hand-held GeneSys ADMA-Slim tracking module synced with the capture ve-



(a) NLOS detection sample distribution over scenarios.

(b) Distance of hidden object and observer to wall.

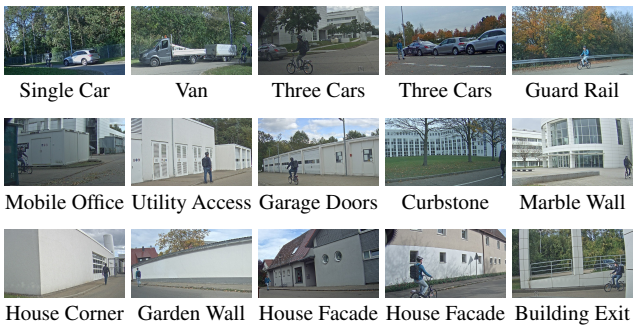


Figure 7: NLOS training and evaluation data set for large outdoor scenarios. Top: Data set statistics (a), and hidden object and observer distances (b) to the relay wall. Bottom: Camera images including the (later on) hidden object.

hicle via Wi-Fi, see Fig. 6. In contrast to [43] we do not purely rely on GNSS, but also use the IMU for pose estimation of the hidden object, see Supplemental Material.

Training and Validation Data Set We capture a total of 100 sequences in-the-wild automotive scenes with 21 different scenarios, *i.e.*, we repeat scenarios with different NLOS trajectories multiple times. The wide range of relay walls appearing in this dataset is shown in Fig. 7 and includes plastered walls of residential and industry buildings, marble garden walls, a guard rail, several parked cars, garages, a warehouse wall, and a concrete curbstone. The dataset is equally distributed among hidden pedestrians and cyclists, and adds up to over 32 million radar points, see Supplemental Material. We opt for these two kinds of challenging road users, as bigger, faster, and more electrically conductive objects such as cars are much easier to detect for automotive radar systems. We split the dataset into non-overlapping training and validation sets, where the validation set consists of four scenes with 20 sequences and 3063 frames.

6. Assessment

Evaluation Setting and Metrics For both, training and validation, the region of interest is a large area of $60\,\mathrm{m} \times 80\,\mathrm{m}$. We use resolution $0.1\,\mathrm{m}$ to discretize both x, z axes into a 600×800 grid. We assign each ground truth box to its highest overlapping predicted box for training. The hidden classification and localization performance are evaluated with Average Precision (AP) and average distance between the predicted and ground truth box centers, respectively.

Class	Cyclist			Pedestrian			Object		
AP	@0.5	@0.25	@0.1	@0.5	@0.25	@0.1	@0.5	@0.25	@0.1
Ours	**29.35**	**56.43**	**62.40**	**44.74**	**62.19**	**68.15**	**41.36**	**66.34**	**75.41**
SSD [26][2]	10.07	37.87	51.50	27.19	49.24	56.24	19.87	46.29	60.98
PointPillars [24][2]	2.02	15.02	28.00	7.83	22.16	26.76	9.61	45.69	58.68

Table 1: Detection classification (AP) comparison. We compare our model to an SSD detector and the PointPillars [24], details in Supplementary Material.

Localization (Box Center Distance)		
Model	MAE	MSE
Tracking (w. v)	0.12	0.03
Tracking (w/o. v)[3]	0.13	0.04

Model	Visibility	MOTA	MOTP
Tracking (w. v)	NLOS	0.58	0.93
	LOS	0.85	0.91
Tracking (w/o. v)[3]	NLOS	0.52	0.94
	LOS	0.81	0.90

Table 2: Localization and tracking performance on NLOS and LOS data, with MAE and MSE in meters. Velocity prediction (and supervision) indicated by v.

Qualitative Validation Fig. 8 shows results for realistic automotive scenarios with different wall types. Note that the size of ground truth bounding box varies due to the characteristics of radar data. The third row shows a scenario where no more than three detected points are measured for the hidden object, and the model has to rely on velocity and orientation of these sparse points to make a decision on box and class prediction. Despite such noise, we do observe that the model outputs stable predictions. As illustrated in the fourth row, predicted boxes are very consistent in size and orientation across frames despite the extreme radar detection sparsity. The first frame in the fourth row shows a detection where a hidden object became visible by lidar but not radar. Note that all other scenes have occluder geometries visible in the lidar measurements. For rare cases where the ground truth information is imperfect due to jitter of the ground truth acquisition system, we can reason about sequences of frames instead of a single one. While the predicted box seemingly does not match the ground truth well in this particular frame, it is, in fact, detected correctly, validating the proposed joint detection and tracking approach. Fig. 9 shows qualitative tracking trajectories for two different scenes. The model is able to track an object only with occasional incorrect ID switch.

Quantitative Detection Results We report AP at IoU thresholds $0.1, 0.25$ and 0.5 for cyclist/pedestrian detection in Tab. 1. We also list the mean AP of predicting object/non-object by merging cyclist/pedestrian labels. We compare our model to a simplified SSD [26] and the PointPillars [24] for BEV point cloud detection, see Supplemental Material. Since most bounding boxes in our collected data are challenging small boxes with sizes smaller than $0.5\,\mathrm{m} \times 0.5\,\mathrm{m}$, a very small offset may significantly affect the detection performance at a high IoU threshold. However, in prac-

[2]Trained with proposed third-bounce geometry and velocity estimates.
[3]Input is velocity-based pre-processed data, see Supplemental Material.

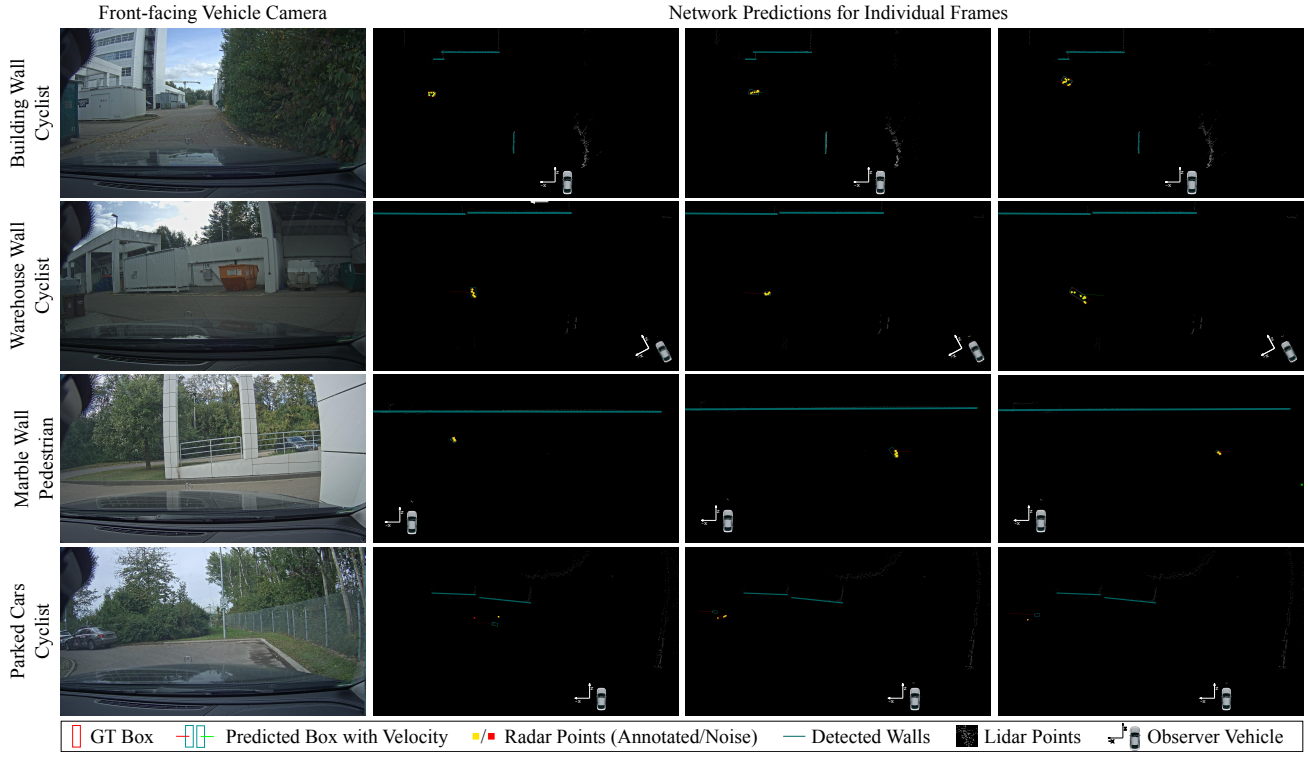
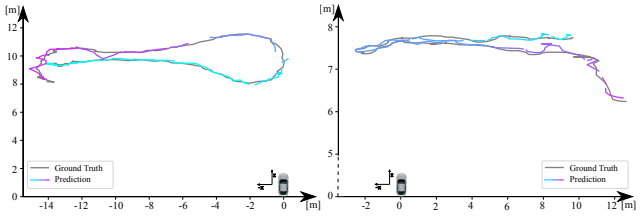


Figure 8: Joint detection and tracking results for automotive scenes with different relay wall type and object class in each row. The first column shows the observer vehicle front-facing camera view. The next three columns plot BEV radar and lidar point clouds together with bounding box ground truth and predictions. NLOS velocity is plotted as line segment from the predicted box center: red and green corresponds to moving towards and away from the vehicle.



(a) Cyclist (b) Pedestrian

Figure 9: Tracking trajectories for two NLOS scenes. The predictions consist of segments, with each corresponding to a different tracking ID visualized in different colors.

tice, a positive detection with an IoU as small as 0.1 is still a valid detection for collision warning applications. Combined with the high localization accuracy, see Tab. 2 (left), the proposed approach allows for long-range detection and tracking of hidden object in automotive scenarios, even for small road users as pedestrians and bicycles.

Quantitative Tracking Results Tab. 2 lists the localization and tracking performance of the proposed approach. Relying on multiple frames and measured Doppler velocity estimates, the proposed method achieves high localization accuracy of $0.1\,\mathrm{m}$ in MAE despite measurement clutter and small diffuse cross section of the hidden pedestrian and bicycle objects. We evaluate the tracking performance on NLOS and visible line-of-sight (LOS) frames separately in Tab. 2. For challenging NLOS data, while the number of unmatched objects (Multiple Object Tracking Accuracy – MOTA) increases, the model is still able to precisely locate most of the matched objects (Multiple Object Tracking Precision – MOTP). These results validate the proposed joint NLOS detector and tracker for collision avoidance applications. Tab. 2 also compares models with and without velocity supervision, showing that velocity supervision improves both localization and tracking accuracy.

7. Conclusion

In this work, we introduce a novel method for joint non-line-of-sight detection and tracking of occluded objects using automotive Doppler radar. Learning detection and tracking end-to-end from a realistic NLOS automotive radar data set, we validate that the proposed approach allows for collision warning for pedestrians and cyclists in real-world autonomous driving scenarios – before seeing them with existing direct line-of-sight sensors. In the future, detection from higher-order bounces, and joint optical and radar NLOS could be exciting next steps.

Acknowledgment

This research received funding from the European Union under the H2020 ECSEL program as part of the DENSE project, contract number 692449.

References

[1] N. Abramson. Light-in-flight recording by holography. *Optics Letters*, 3(4):121–123, 1978. 1

[2] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)*, 34(6):219, 2015. 2

[3] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller. 3d tracking via body radio reflections. In *11th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 14)*, pages 317–329, 2014. 2

[4] F. Adib and D. Katabi. See through walls with wifi! *ACM SIGCOMM Computer Communication Review*, 43(4), 2013. 2

[5] V. Arellano, D. Gutierrez, and A. Jarabo. Fast back-projection for non-line of sight reconstruction. *Optics Express*, 25(10):11574–11583, 2017. 1, 2

[6] K. L. Bouman, V. Ye, A. B. Yedidia, F. Durand, G. W. Wornell, A. Torralba, and W. T. Freeman. Turning corners into cameras: Principles and methods. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2289–2297, 2017. 2

[7] G. M. Brooker. Understanding millimetre wave fmcw radars. In *1st International Conference on Sensing Technology*, pages 152–157, 2005. 3

[8] M. Buttafava, J. Zeman, A. Tosi, K. Eliceiri, and A. Velten. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Optics express*, 23(16):20997–21011, 2015. 1, 2

[9] P. Caramazza, A. Boccolini, D. Buschek, M. Hullin, C. F. Higham, R. Henderson, R. Murray-Smith, and D. Faccio. Neural network identification of people hidden from view with a single-pixel, single-photon detector. *Scientific reports*, 8(1):11945, 2018. 2

[10] S. Chan, R. E. Warburton, G. Gariepy, J. Leach, and D. Faccio. Non-line-of-sight tracking of people at long range. *Optics express*, 25(9):10109–10117, 2017. 2

[11] W. Chen, S. Daneau, F. Mannan, and F. Heide. Steady-state non-line-of-sight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6799, 2019. 2, 3

[12] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 1

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1

[14] O. Gupta, T. Willwacher, A. Velten, A. Veeraraghavan, and R. Raskar. Reconstruction of hidden 3d shapes using diffuse reflections. *Opt. Express*, 20(17):19096–19108, Aug 2012. 1, 2

[15] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 1

[16] F. Heide, L. Xiao, W. Heidrich, and M. B. Hullin. Diffuse mirrors: 3d reconstruction from diffuse indirect illumination using inexpensive time-of-flight sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3222–3229, 2014. 2

[17] A. Kadambi, R. Whyte, A. Bhandari, L. Streeter, C. Barsi, A. Dorrington, and R. Raskar. Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles. *ACM Transactions on Graphics (ToG)*, 32(6):167, 2013. 2

[18] A. Kadambi, H. Zhao, B. Shi, and R. Raskar. Occluded imaging with time-of-flight sensors. *ACM Transactions on Graphics (ToG)*, 35(2):15, 2016. 2

[19] O. Katz, P. Heidmann, M. Fink, and S. Gigan. Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations. *Nature photonics*, 8(10):784, 2014. 2

[20] O. Katz, E. Small, and Y. Silberberg. Looking around corners and through thin turbid layers in real time with scattered incoherent light. *Nature photonics*, 6(8):549–553, 2012. 2

[21] A. Kirmani, T. Hutchison, J. Davis, and R. Raskar. Looking around the corner using transient imaging. In *IEEE International Conference on Computer Vision (ICCV)*, pages 159–166, 2009. 1, 2

[22] J. Klein, C. Peters, J. Martín, M. Laurenzis, and M. B. Hullin. Tracking objects outside the line of sight using 2d intensity images. *Scientific reports*, 6:32491, 2016. 2

[23] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018. 1

[24] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. 7

[25] R. Lange. *3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology*. PhD thesis, Universitt Siegen, 2000. 2

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4, 6, 7

[27] D. B. Lindell, G. Wetzstein, and V. Koltun. Acoustic non-line-of-sight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6780–6789, 2019. 2, 3

[28] D. B. Lindell, G. Wetzstein, and M. OToole. Wave-based non-line-of-sight imaging using fast f-k migration. *ACM Trans. Graph. (SIGGRAPH)*, 38(4):116, 2019. 1

[29] X. Liu, I. Guillén, M. La Manna, J. H. Nam, S. A. Reza, T. H. Le, A. Jarabo, D. Gutierrez, and A. Velten. Non-line-of-sight imaging using phasor-field virtual wave optics. *Nature*, pages 1–4, 2019. 1, 2

[30] W. Luo, B. Yang, and R. Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the*

IEEE conference on Computer Vision and Pattern Recognition, pages 3569–3577, 2018. 1, 6

[31] D. Lynch. *Introduction to RF stealth*. The SciTech radar and defense series. SciTech, 2004. 3

[32] N. Naik, S. Zhao, A. Velten, R. Raskar, and K. Bala. Single view reflectance capture using multiplexed scattering and time-of-flight imaging. *ACM Trans. Graph.*, 30(6):171, 2011. 1

[33] M. OToole, D. B. Lindell, and G. Wetzstein. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature*, 555(7696):338–341, 2018. 1, 2

[34] R. Pandharkar, A. Velten, A. Bardagjy, E. Lawson, M. Bawendi, and R. Raskar. Estimating motion and size of moving non-line-of-sight objects in cluttered environments. In *Proc. CVPR*, pages 265–272, 2011. 1, 2

[35] A. K. Pediredla, M. Buttafava, A. Tosi, O. Cossairt, and A. Veeraraghavan. Reconstructing rooms using photon echoes: A plane based model and reconstruction algorithm for looking around the corner. In *IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2017. 2

[36] O. Rabaste, J. Bosse, D. Poullin, I. Hinostroza, T. Letertre, T. Chonavel, et al. Around-the-corner radar: Detection and localization of a target in non-line of sight. In *2017 IEEE Radar Conference (RadarConf)*, pages 0842–0847. IEEE, 2017. 2

[37] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[38] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[39] M. A. Richards, J. Scheer, W. A. Holm, and W. L. Melvin. *Principles of modern radar*. Citeseer, 2010. 2, 3

[40] H. Rohling. Radar CFAR Thresholding in Clutter and Multiple Target Situations. *IEEE Transactions on Aerospace and Electronic Systems*, AES-19(4):608–621, 1983. 4

[41] K. Sarabandi, E. S. Li, and A. Nashashibi. Modeling and measurements of scattering from road surfaces at millimeter-wave frequencies. *IEEE Transactions on Antennas and Propagation*, 45(11):1679–1688, 1997. 3

[42] C. Saunders, J. Murray-Bruce, and V. K. Goyal. Computational periscopy with an ordinary digital camera. *Nature*, 565(7740):472, 2019. 2

[43] N. Scheiner, N. Appenrodt, J. Dickmann, and B. Sick. Automated Ground Truth Estimation of Vulnerable Road Users in Automotive Radar Data Using GNSS. In *IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, pages 5–9, Detroit, MI, USA, 2019. IEEE. 6, 7

[44] A. Sume, M. Gustafsson, M. Herberthson, A. Janis, S. Nilsson, J. Rahm, and A. Orbom. Radar detection of moving targets behind corners. *IEEE Transactions on Geoscience and Remote Sensing*, 49(6):2259–2267, 2011. 2

[45] C.-Y. Tsai, K. N. Kutulakos, S. G. Narasimhan, and A. C. Sankaranarayanan. The geometry of first-returning photons for non-line-of-sight imaging. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[46] A. Velten, T. Willwacher, O. Gupta, A. Veeraraghavan, M. Bawendi, and R. Raskar. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature Communications*, 3:745, 2012. 1, 2

[47] A. Velten, D. Wu, A. Jarabo, B. Masia, C. Barsi, C. Joshi, E. Lawson, M. Bawendi, D. Gutierrez, and R. Raskar. Femto-photography: Capturing and visualizing the propagation of light. *ACM Trans. Graph.*, 32, 2013. 2

[48] R. G. Von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. Lsd: a line segment detector. *Image Processing On Line*, 2:35–55, 2012. 5

[49] J. Wilson and N. Patwari. Through-wall tracking using variance-based radio tomography networks, 2009. 2

[50] F. Xu, G. Shulkind, C. Thrampoulidis, J. H. Shapiro, A. Torralba, F. N. C. Wong, and G. W. Wornell. Revealing hidden scenes by photon-efficient occlusion-based opportunistic active imaging. *OSA Opt. Express*, 26(8):9945–9962, 2018. 2

[51] R. Zetik, M. Eschrich, S. Jovanoska, and R. S. Thoma. Looking behind a corner using multipath-exploiting uwb radar. *IEEE Transactions on aerospace and electronic systems*, 51(3):1916–1926, 2015. 2

[52] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018. 2

[53] Y. Zhou and O. Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018. 1