Supplementary Material: What You Can Learn by Staring at a Blank Wall

1. Sample Scenes in the Dataset

In Figure 1, we show a small subset of the observed scenes and a sequence of amplified signal observed in that sequence. Please refer to the dataset available at wallcamera.csail.mit.edu.

2. Neural Network Architectures

We present the network architectures and training procedure used for the two tasks.

2.1. Classification of the number of people

The input to the neural network is a horizontal space-time plot of dimension $64 \times 256 \times 3$, corresponding to a 256-frame RGB video downsampled to 64 pixels in width. We use a total of 5 convolutional blocks followed by a shallow 2-layer fully connected network.

In the first four convolutional blocks, each convolution uses kernel size of 5×5, stride of 1, and uses zero-padding. The feature channel count in every intermediate layer is 64. The convolution in each block is followed by a leaky-ReLU non-linearity (with negative slope of 0.1) [3], and a 2×2 max-pooling layer to reduce the spatial and temporal dimensions. We apply batch normalization to the output of each convolution [1].

The fifth convolutional block is similar to the preceding ones, but has a convolution kernel size of 4×4 with no padding. This reduces the spatial dimension to 1 and temporal dimension to 13 time instants. This is the temporal summary of the space-time plot.

These 13 dimensional feature vectors are collapsed using a max-pooling operation over the time dimension, resulting in a 64-dimensional feature vector. This vector is the input to the linear layer followed by a leaky-ReLU, followed by a final linear layer which outputs a vector with 3 values. We apply the softmax function to the three-dimensional output to get the probabilities for the three classes (0, 1, and 2 persons).

2.2. Activity Recognition

The input to the neural network are both horizontal and vertical space-time plots corresponding to the same video input of 256 frames. Each space-time plot is of dimension $64 \times 256 \times 3$. We have two initial branches, one to process each of the horizontal and vertical space-time plots. Each branch consists of the 5 convolutional blocks similar to the initial part of the network used for classifying the number of people, outputting a vector of spatial dimension 1 and temporal dimension 13. We then perform a max-pool operation over the time dimension, resulting in 64-dimensional feature vector for each branch. We further concatenate these two 64-dimensional feature vectors to get a 128-dimensional representation for both the horizontal and vertical space-time plot. This 128-dimensional vector is input to a linear layer, which outputs a 64-dimensional vector followed by leaky-ReLU. This is followed by another linear layer which outputs a 32-dimensional vector, followed by a leaky-ReLU layer. Finally, we use this 32-dimensional vector to output raw scores for the 5 classes using an output linear layer.

For both networks, we use standard cross-entropy loss and optimize using RMSProp [4] with a learning rate of `1e-3`.

3. Derivations of the SNR Formulas

We present the derivations for the formulas used in Section 6.3 of the main document.

3.1. Power of the Radiance Signal

As described in the main text, we consider an idealized scene, where we are observing a diffuse wall of albedo (color) α, surrounded by a static and constant-colored environment from where the incident radiance is L_s. Facing the wall is a person at distance d, of approximately circular shape, of radius r and area $A = \pi r^2$ (as projected on the wall). The radiance from the person is L_p.

The *rendering equation* [2] predicts that the radiance L_o observed from the wall towards the camera, at the point x perpendicular to the person, is

$$L_o(x, \omega_o) = \int_\Omega L_i(x, \omega_i) f(x, \omega_i, \omega_o) \cos \omega_i \, \mathrm{d}\omega_i. \quad (1)$$

Here, Ω is the hemisphere of directions ω_i surrounding the point x, and L_i is the hemispherical incident radiance towards that point. f is the reflectance function (BRDF) that indicates the transmission of light between any two direc-

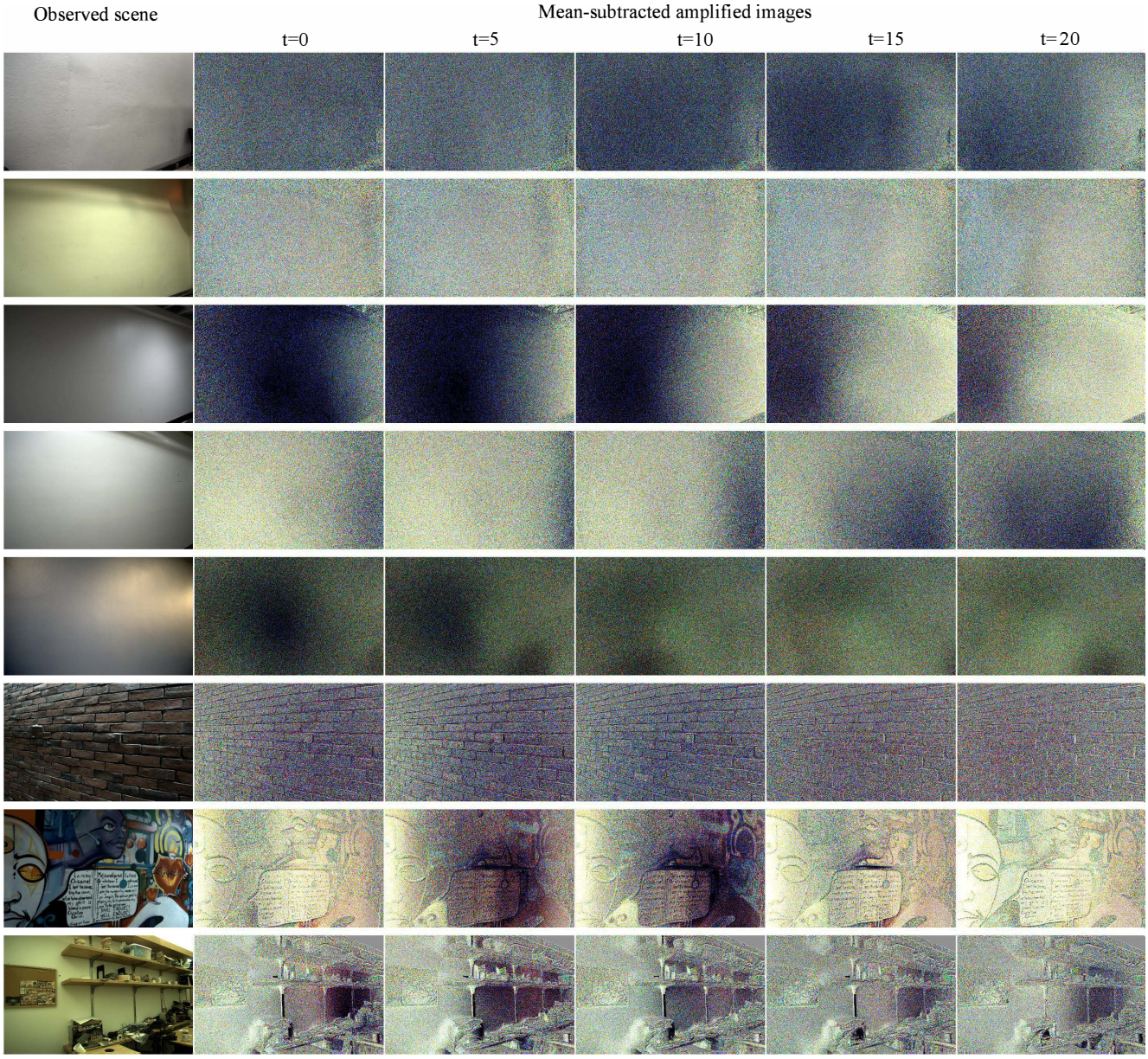

Figure 1: Observed scenes and a sequence of the amplified signal as one person walks in the hidden scene.

tion. For our diffuse (Lambertian) wall, the BRDF f is simply α/π. The incident radiance $L_i(\omega_i)$ is L_s in directions that see the environment, and L_p in the regions that are occluded by the person.

For the simple scene described above, the rendering equation becomes

$$L_o(x,\omega_o) = \frac{\alpha}{\pi}\int_\Omega [\mathbb{1}_p(\omega)L_p + (1-\mathbb{1}_p(\omega))L_s]\cos\omega_i \,\mathrm{d}\omega_i, \tag{2}$$

where $\mathbb{1}_p$ is a binary indicator function that corresponds to directions covered by the person, and $1-\mathbb{1}_p$ is its complement, i.e. the indicator for directions covered by the background. This rearranges to

$$L_o(x,\omega_o) = \alpha L_p \int_\Omega \mathbb{1}_p(\omega)\frac{\cos\omega_i}{\pi}\,\mathrm{d}\omega_i + \alpha L_s\left(1-\int_\Omega \mathbb{1}_p(\omega)\frac{\cos\omega_i}{\pi}\,\mathrm{d}\omega_i\right), \tag{3}$$

where the integral (note that it is the same in both two terms) represents the fraction that the person occupies of the projected unit hemisphere. It can be computed analytically as follows.

First, by elementary geometry we see that the person subtends a circular section of opening angle $\alpha = 2\ \mathrm{atan}\frac{r}{d}$ on the hemisphere Ω. Switching to projected solid angle coordinates, the cosine vanishes and the section is projected into a circle that occupies the fraction $\tilde{\theta}(r,d) := \frac{r^2}{r^2+d^2}$ of the full unit circle. This is the value of the integral.

For $d >> r$, this ratio can be approximated as $\theta(r,d) = \frac{r^2}{d^2} = \frac{A}{\pi d^2}$. Note that the fraction is low when the person is far away or small, and vice versa. Substituting this formula as the value of the integral in the previous equation, we find that the final observed radiance is a blend of the the person's radiance and the background radiance, weighted by this fraction, times the albedo of the wall:

$$L_o \approx \alpha \left[\theta(r,d)L_p + [1-\theta(r,d)]L_s\right], \tag{4}$$

Let us then make the approximation that the average observation of the pixel over the video segment is simply αL_s, as the person is moving and is not significantly covering the wall in most frames. Subtracting this mean from the above, the zero-mean signal is then $\tilde{L}_o \approx \frac{\alpha A}{\pi d^2}(L_p - L_s)$.

Finally, as the person is moving and only occasionally near any given point on the wall (a fraction $q \in [0,1]$ of time), the average power of the signal is $q\tilde{L}_o^2$. This coarse approximation is derived by treating the signal as a Bernoulli random variable where with probability $(1-q)$ we observe 0, and with probability q we observe L_o. The variance is then $(1-q)(q)L_o^2$, or approximately qL_o^2 when q is small. Substituting the formula for L_o, we arrive at Eq. 1 in the main text.

3.2. Power of Signal and Noise in Real Data

A video V (with zero mean over time) can be viewed as a sum of two unknown videos $V = S + N$, one containing the signal and the other the noise. In the following, we show how to estimate the power of each component. Let $\mathcal{T}$ be the (linear) operation that shifts a video in time by one frame. We can reasonably assume that our shadow signal varies slowly in time, so that any two consecutive frames are almost identical. Then, $S - \mathcal{T}(S) \approx 0$. In contrast, the noise between frames is independent, and a neighboring frame subtraction merely boosts the noise: $\mathrm{Var}(N - \mathcal{T}(N)) \approx 2\ \mathrm{Var}\ N$. Combining these two findings, we have $\mathrm{Var}\ (V - \mathcal{T}(V)) = \mathrm{Var}\ [S + N - \mathcal{T}(S + N)] = \mathrm{Var}\ [S - \mathcal{T}(S)] + \mathrm{Var}\ [N - \mathcal{T}(N)] = 0 + 2\ \mathrm{Var}\ N$, giving us a way to estimate the variance of N using only V. Furthermore, the power (which coincides with variance) of the noise *plus* signal is simply $\mathrm{Var}\ V$. An estimate of the signal power alone can be recovered by subtracting the earlier estimate, giving $\mathrm{Var}\ S \approx \mathrm{Var}\ V - \frac{1}{2}\mathrm{Var}\ (V - \mathcal{T}(V))$.

3.3. Experiment for SNR in Real Data

To study the effect of distance of the person to the wall and light intensity in the scene, we collected data in a controlled environment controlling these parameters. One person performed four activities, walking, crouching, waving hands, and jumping, at four different distances (1.75, 2.45, 3.2, 3.95 meters) and at three different lighting intensities (12, 48, 125 lux). The layout of the room is shown in Figure 2.

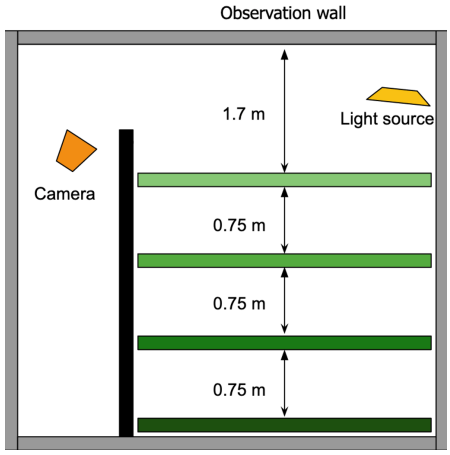

Figure 2: Setup diagram for collecting data with varying distances and lighting intensities.

Supporting the theory and the graphs shown in the main document, Figure 3 and 4 show the variation in the space-time plots for the activities for the variations in distance of the person to the wall and lighting intensity. As expected, the signal correlated to the motion becomes faint as the distance of the person to the wall increases. Similar downtrend in signal can be seen as the lighting intensity is reduced in the scene.

4. Synthetic Data Pipeline

The flatland setup consists of two parallel walls – a receiver wall that is observed, and an emitter plane that emulates the lighting environment, as shown in Figure 5a. Between these, one or two flat occluders move in distorted sinusoidal patterns. The light transport is simulated according to an approximate two-dimensional version of the rendering equation [2]. The lighting environment, diffuse albedo variation on receiver, occluder colors, motion patterns, ambient lighting, noise, and other aspects of the simulation are randomized. Figure 5b and 5c show a selection of one and two-person space-time plots generated in this manner. While not fully identical to the real data, the plots show similar qualitative effects – for example, the streaks resulting from cross-overs of the two person case can also be observed in these plots.

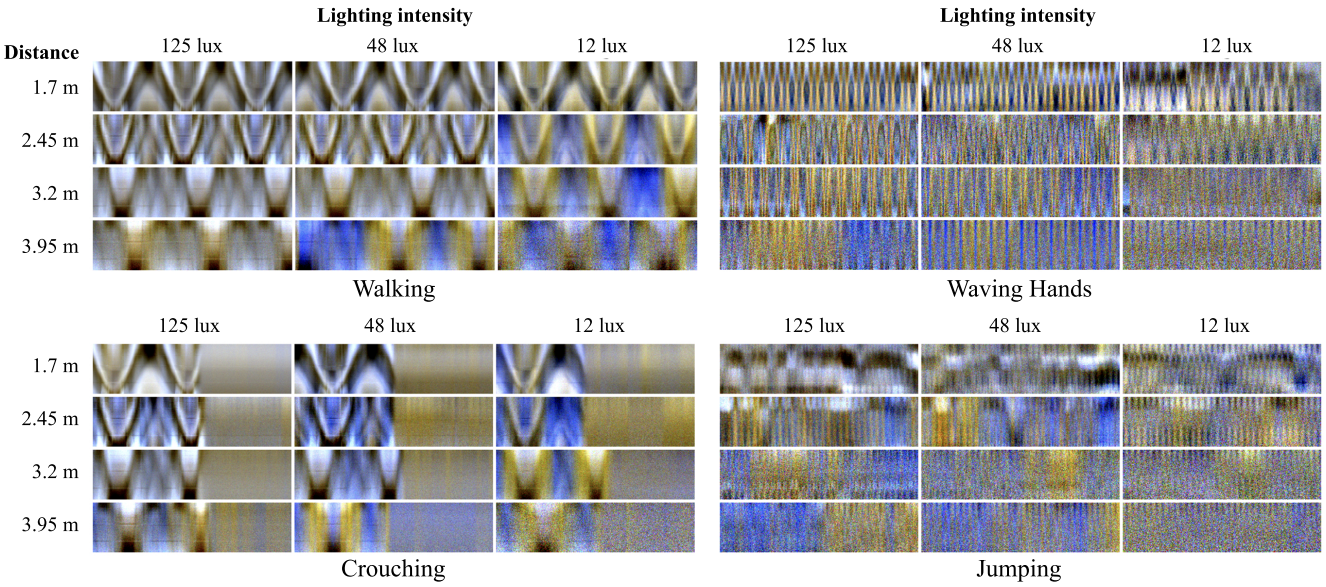

Figure 3: Horizontal space-time plots over different distance of the person to the observation wall and lighting intensities.

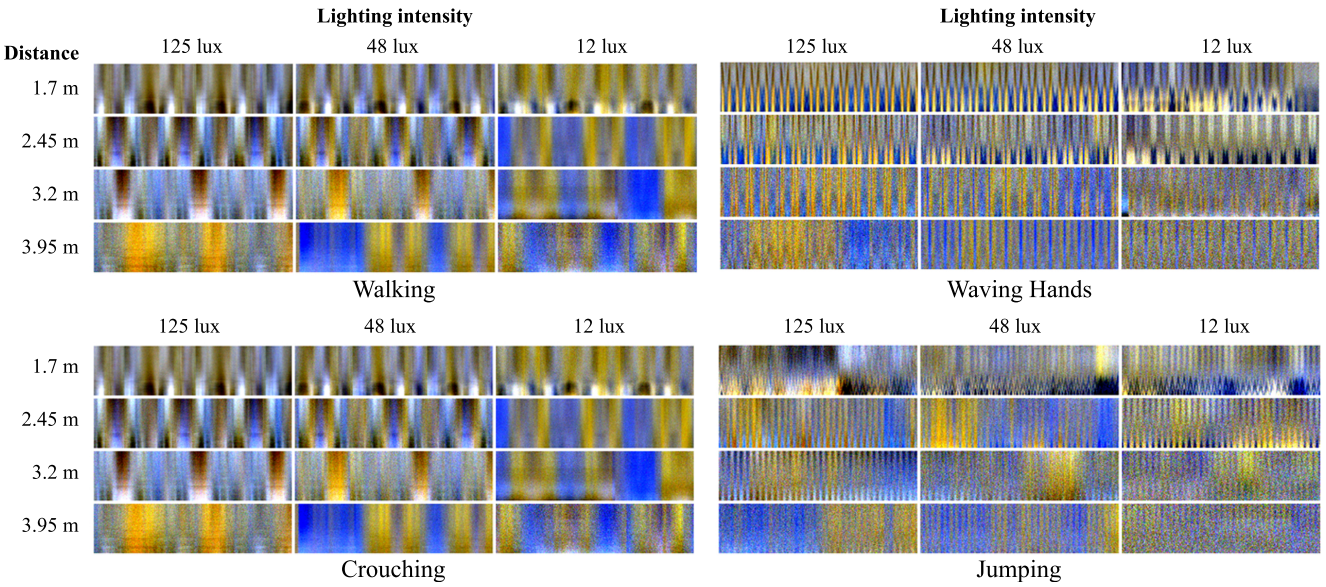

Figure 4: Vertical space-time plots over different distance of the person to the observation wall and lighting intensities.

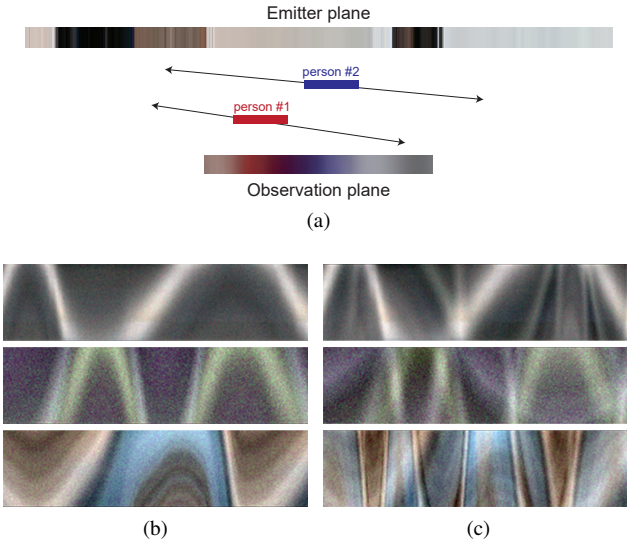

Figure 5: (a) Two-dimensional flatland setup for synthetic data generation. The two blockers representing persons move back and forth along random directions. A 1D image is rendered at the observation plane, taking into account the mutual visibility between the blockers and the back wall acting as an illuminant. (b) Samples of synthetic space-time plots for one person scenario. (c) Samples of synthetic space-time plots for two people, in the same flatland scenario as (b).

References

[1] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 1

[2] James T. Kajiya. The rendering equation. In *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '86, pages 143–150, New York, NY, USA, 1986. ACM. 1, 3

[3] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013. 1

[4] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012. 1