Object Classification through Scattering Media with Deep Learning on Time Resolved Measurement

Guy Satat Matthew Tancik, Otkrist Gupta,
Barmak Heshmat, and Ramesh Raskar

guysatat@mit.edu

Abstract

We demonstrate an imaging technique that allows identification and classification of objects hidden behind scattering media and is invariant to changes in calibration parameters within a training range. Traditional techniques to image through scattering solve an inverse problem and are limited by the need to tune a forward model with multiple calibration parameters (like camera field of view, illumination position etc.). Instead of tuning a forward model and directly inverting the optical scattering, we use a data driven approach and leverage convolutional neural networks (CNN) to learn a model that is invariant to calibration parameters variations within the training range and nearly invariant beyond that. This effectively allows robust imaging through scattering conditions that is not sensitive to calibration. The CNN is trained with a large synthetic dataset generated with a Monte Carlo (MC) model that contains random realizations of major calibration parameters. The method is evaluated with a time-resolved camera and multiple experimental results are provided including pose estimation of a mannequin hidden behind a paper sheet with 23 correct classifications out of 30 tests in three poses (76.6% accuracy on real-world measurements). This approach paves the way towards real-time practical non line of sight (NLOS) imaging applications.

1 Introduction

Sensing through optical scattering is a challenging and important problem which enables applications such as image dehazing [1], cloud tomography [2], underwater imaging [3] and recovery of material scattering properties [4]. When considering strongly scattering media and non line of sight (NLOS) imaging, such as seeing around corners, traditional imaging with ambient illumination may prove to be insufficient. Many active methods have been demonstrated for imaging through scattering such as optical coherence tomography [5], wavefront

shaping [6, 7, 8, 9], speckle correlations [10, 11], acousto-optic [12] and photo-acoustic [13, 14]. Active time-of-flight (ToF) methods are used to overcome scattering with the advantage of wide-field imaging, and minimal (or no) raster scan [15]. ToF systems are either impulse [16, 17, 18] or phase [19, 20, 21, 22] based. Different aspects of imaging have been demonstrated with ToF systems like full scene reconstruction [23], pose estimation [24], tracking [25], character recognition [26], and medical applications such as seeing through tissue [27]. NLOS imaging has been demonstrated with impulse based systems like streak camera [17] and single photon avalanche photodiode (SPAD) [28, 29]. Recently a system with an active illumination and a regular camera demonstrated tracking around a corner [30]. All computational methods that rely on scattering inversion are sensitive to model calibration.

The requirement for calibration when performing imaging through scattering is directly related to the need of a physical model that explains the measurements. Such models simulate light transport and depend on the geometry and other physical parameters of the system. Since inverting scattering is an ill-posed problem, any mismatch between the physical model and the actual measurement will degrade performance. As a result, accurate calibration of imaging system parameters like illumination position, camera orientation etc. is needed (the full list of calibration parameters considered here is provided in table 1). This limits many inversion based techniques to scale to real-world applications. The approach presented here allows robust classification of objects hidden behind scattering media or beyond the line of sight of the camera that is not sensitive to calibration.

To tackle the problem of calibration-invariant imaging we use convolutional neural networks (CNN). Recently, CNN has become the main workhorse in many computer vision tasks. CNN are especially appealing for our application due to their ability to capture invariants [31, 32], reduce dimensionality from noisy data [33], and classify objects [34]. Initial uses of data driven approaches for imaging problems have been suggested in microscopy [35], compressive imaging [36], synthetic aperture radar [37], remote sensing [38, 39], dehazing [40], phase imaging [41], medical imaging [42], and classification with coherent light [44, 43]. In our case the CNN is trained with synthesized data that includes variations in calibration parameters. By training the CNN with synthetic and diverse data the network learns a model that is not only invariant under traditional transformations like translation, but also invariant to changes in calibration parameters within the training range and nearly invariant beyond that range. Thus, we eliminate the need to precisely calibrate a computational model for NLOS object classification.

The imaging procedure is shown in Fig. 1 and is partitioned into two halves. First (offline process) a Monte Carlo (MC) model is used to synthesize a large training dataset of potential measurements drawn from the distribution of all target variations and calibration parameters. The synthesized dataset is used to train a CNN. The resulting CNN is invariant to changes in calibration parameters within the training range, effectively allowing calibration-invariant object classification through scattering. This enables the second (online phase)

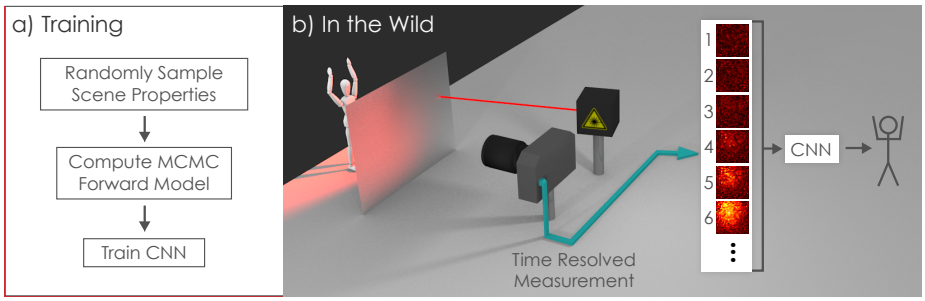

Figure 1: Calibration-invariant object classification through scattering. a) Training phase is an offline process in which the user defines random distributions of physical model parameters (based on approximate measurements or prior knowledge). The distributions are used to generate synthetic measurements with an MC forward model. The synthetic data is used to train a CNN for classification. b) Once the CNN is trained the user can simply place the camera (demonstrated here with a time sensitive SPAD array) and an illumination source in the scene, capture measurements (six examples of time resolved frames are shown), and classify with the CNN without having to precisely calibrate the system.

in which the user can simply place the camera and illumination without calibration and classify, in real time, hidden objects behind scattering media. We experimentally demonstrate this method with human pose estimation behind a scattering layer. Pose estimation behind scattering has many important applications in areas such as: privacy preserving human computer interaction systems (e.g. human pose estimation without access to face or body images) as well as search and rescue missions.

The key contributions of our approach are: 1. A measurement-independent training method that uses only synthetic data (based on a Monte Carlo renderer) to train a CNN; before acquiring any actual measurements. 2. A technique for imaging through scattering that is invariant to variations in system calibration parameters within the training range. 3. The technique allows real-time classification through scattering medium and beyond line of sight.

2 Imaging Procedure

2.1 Measurement System

The optical setup is shown in Fig. 1. A pulsed source (NKT photonics SuperK) with a repetition rate of $80MHz$ and pulse duration of $5ps$ is spectrally filtered to a band of $580\pm10nm$. The camera is a single photon avalanche diode (SPAD) array (Photon Force PF32) with 32×32 pixels, and a time resolution of $56ps$. The laser is incident on the diffuser at $\sim 45°$. The camera is focused on the diffuser (regular paper sheet which presents non-uniform scattering properties).

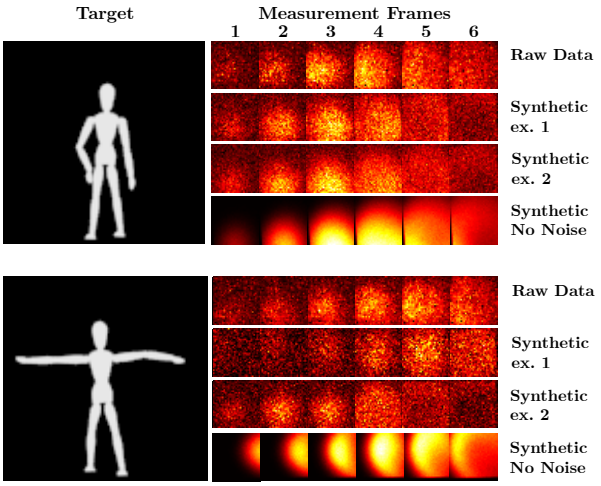

Figure 2: Comparison of SPAD measurement and forward model. The targets are two poses of a mannequin placed behind a paper sheet (diffuser). The data shows six frames (each frame is 32×32 pixels) of raw SPAD measurements, examples of two synthetic results generated by the MC forward model with similar measurement quality, and a synthetic result with high photon count and no additive noise. Note that differences between synthetic ex. 1, 2 and the raw measurement are due to the fact that the forward model was never calibrated to this specific setup. The synthetic images represent different instances chosen randomly from the dataset. The synthetic example with high photon count helps to distinguish between measurement (or simulated) noise and the actual signal as well as to observe the full signal wavefront.

Table 1: Distributions for calibration and target parameters used in mannequin dataset.

Calibration parameters	
Laser	
- Incident position	$L_P \sim U(-4, 4)cm$
Diffuser	
- Scattering profile	$D_D \sim N(0, \sigma)$, $\sigma \sim U(0.8, 1.2)rad$
Camera	
- Position	$C_P \sim U(-1.5, 1.5)cm$
- Time resolution	$C_{TR} \sim N(0, \sigma)$, $\sigma \sim 56 + U(-5, 5)ps$
- Time jitter	$C_{TS} \sim U(0, 3 * 56)ps$
- Field of view	$C_{FV} \sim U(0.1, 0.2)rad$
- Homography	Normal distributions
Noise	
- Dark count	$N_{DC} \sim U(3000, 9000)$ photons
Target parameters	
- Position	$T_{P_{x,y,z}} \sim U(-4, 4)cm$
- Scale	$T_S \sim U(18, 30)cm$

A flexible mannequin is placed behind the diffuser ($20cm$ head to toe). A black screen separates the camera from the incident position of the laser on the diffuser (to prevent direct reflection from the diffuser to the camera). The optical setup demonstrates a reflection mode geometry. The first 64 time bins of the SPAD measurement are used, such that the data structure is of size $32 \times 32 \times 64$ (the large number of frames guarantees consistency and flexibility of the data structure). Several examples of the measurement frames are provided in Fig. 2.

2.2 Forward Model - Synthetic Data Generation

The proposed method is based on an MC model that renders the SPAD measurements. Since SPAD captures single photon events it fits well to an MC model that traces individual photons. MC is a very generic forward modeling technique that can be easily modified to simulate various system geometries like looking around corners and seeing through a scattering medium (which is the focus of this work). This is accomplished by modeling a wide range of physical parameters which can be broadly divided into target and calibration related.

The MC forward model is used to generate a dataset with random realizations of potential measurements for different target instances and calibration parameters. For each new simulated data point, the above parameters are randomly sampled from their given distributions to define a specific system, geometry and target (Table 1 provides the list of parameters and distributions used here). A ray tracer is used to simulate propagation of individual photons

from the illumination source, through the diffuser, onto the target, back to the diffuser and finally into the camera (see Algorithm 1). This process takes into account the propagation time. SPAD array measurements are only based on thousands of detected photons. Since there is no need to render the full time-dependent scene irradiance, the computational burden of ray tracing with MC is low (we simulate 10^6 photons for each data point, which takes $\sim$1 second on a regular desktop computer). Figure 2 compares raw measurements taken with the SPAD camera and instances of the forward model (chosen randomly from the synthetic dataset).

We note that while paper sheet is a strongly scattering media with multiple scattering events, it can be modeled as a single scatter event due to: 1) The propagation time through the paper ($\sim 10ps$) [45] is much smaller compared to the time resolution of the SPAD camera. 2) The scene size (target feature size and scene length scales) are much larger compared to the scatterer thickness, so we can approximate the photon exit coordinate to be equal to the entrance coordinate. In cases where these assumptions don't hold, it is easy to add a random walk simulation to the Monte Carlo renderer that would simulate the scattering process in the material (including the time and location dependency).

Each data point in the dataset corresponds to a specific example of a target measured by a system that is defined by a set of random target and calibration parameters (see table 1):

- The target is defined by a label and an instance (for example, a specific appearance of a digit in a handwritten digits dataset), these are simply selected from the dataset. The dataset may or may not include variations in parameters such as scale and orientation. For improved robustness it is preferred to add variability in all parameters, this is achieved by scaling the target with parameters that are drawn from distributions of plausible target size. Finally, the target is placed at a random 3D location behind the diffuser, the location is sampled from a uniform distribution which defines the NLOS volume of interest.

- The imaging system is defined by a realization of various calibration parameters that are sampled from random distributions. User input is involved only in determining the random distributions, which are defined based on approximate measurements, for example observation of the system geometry by the naked eye. If a parameter is easy to evaluate (for example the laser position on the diffuser) it can be modeled with a Gaussian distribution with the known mean and small variance. Or if it is hard to evaluate, it can be modeled with a uniform distribution.

Varying calibration parameters in the training data allows the CNN to be invariant to changes in those parameters within the training range (see section 2.3).

2.3 Learning Calibration Invariant Sensing with CNN

The synthetic random dataset generated with the MC forward model is used to train a CNN for classification of hidden objects behind a diffuser. CNNs

Algorithm 1 MC Forward Model

```
Initialize scene by randomly sampling:
    Target: label, instance, position, size
    Laser: incident position
    Diffuser: scattering profile
    Camera: position, time resolution, time jitter, field of view, homography parameters
for All photons do
    Calculate initial intersection point with diffuser
    Randomly sample diffuser local scattering profile
    Randomly sample photon's angle after diffuser
    Calculate photon's intersection point with target
    if does not hit target then
        continue to next photon
    end if
    Randomly sample angle after reflection from target
    Calculate photon's intersection point with diffuser
    if does not hit diffuser then
        continue to next photon
    end if
    Randomly sample diffuser local scattering profile
    Randomly sample photon's angle after diffuser
    Map photon to camera sensor using homography
    Randomly sample photon's arrival time jitter
    Store photon's arrival time (with jitter) and location
end for
Bin recorded photons into discrete time frames.
Add dark count noise to measurement
```

are a natural fit for this task since: 1) they have been shown to perform well in classification tasks, 2) they are designed to be invariant to translations, and 3) learn to be invariant to other data transformations like scaling, rotation and, as demonstrated here, variations in the system calibration parameters.

Several neural network architectures were considered. The data structure in our case is composed of several frames, which is similar to the case of action recognition and gesture classification from short videos. Works such as [46, 47] indicated that convolutional architectures produce robust classification in that task. Thus, multiple convolutional architectures were evaluated including VGG [48], ResNet [49], and several custom shallower networks with various combinations of layers. All architectures performed similarly on the classification task with marginally better performance for VGG. The VGG topology was selected and modified by extension of convolution filters into time domain (3D space-time filters). Filters were resized to $3 \times 3 \times 10$ where the last index denotes the time dimension (see further details in section 5). The training time

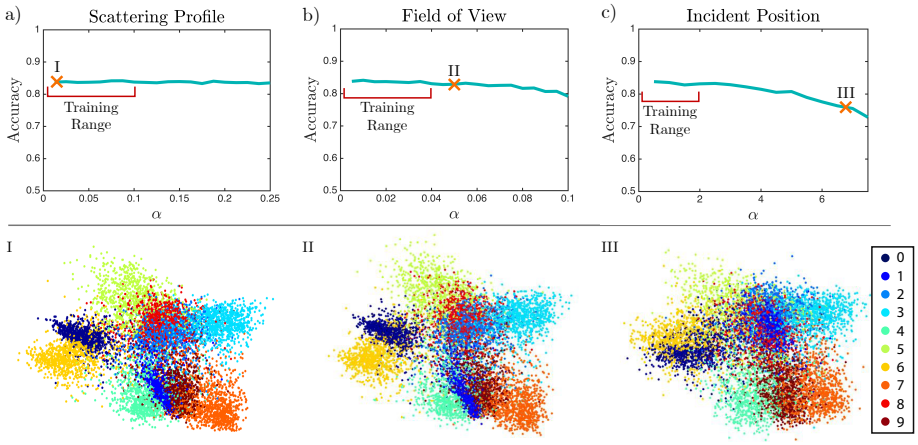

Figure 3: CNN learns to be invariant to model parameters. The CNN is trained with the complete random training set (based on the MNIST dataset), and evaluated with test sets in which all model parameters are fixed except for one that is randomly sampled from distributions with growing variance. Three parameters are demonstrated (other parameters show similar behavior). a) Diffuser scattering profile variance $D_D \sim N(0, \sigma)$, $\sigma \sim U(1-\alpha, 1+\alpha)$ radians, b) Camera field of view $C_{FV} \sim U(0.15-\alpha, 0.15+\alpha)$ radians, and c) Illumination source position $L_P \sim U(-\alpha, \alpha)$ cm. The top plots show the classification accuracy as a function of the parameter distribution variance in the test set. Red lines show the ranges used for training. The 'X' marks point to specific locations sampled for PCA projections in the bottom part of the figure. PCA projections show a color map where each digit has different color. Performance is maintained beyond the training range and starts to slowly degrade further from it, as can be observed in PCA projection III where more mixing is apparent at a test range $\times 2.5$ larger compared to the training set.

on $60,000$ data points is approximately two hours on an Nvidia Titan XP GPU.

To evaluate our approach, we used the well-known MNIST dataset of handwritten digits. The goal is to evaluate the CNN ability to classify hidden objects while being invariant to changes in calibration parameters. To that end, $60,000$ training samples and $10,000$ test samples are synthesized with the MC forward model. Each data point is a realization of a different set of target and calibration parameters. The result on the test set is an overall classification accuracy of 74% (compared to 10% random guess accuracy). These simulations demonstrate the ability to classify objects hidden behind a scattering layer without calibration. As a proof of concept lab experiment, we cut two targets from cardboard shaped like zero and one digits, placed them behind a paper sheet, and measured the response with the SPAD camera. The two time resolved measurements were correctly classified as zero and one using the above network. The

training dataset generation and network training were performed prior to this data acquisition. This demonstrates that our method is robust to variations in calibration parameters on raw data. Section 3 provides more challenging experimental results.

In order to evaluate the extent of the network's ability to handle changes in calibration parameters a set of controlled synthetic experiments were performed. We used the trained network with the MNIST dataset, and tested it with multiple test sets that were generated for the purpose of this evaluation. In each test set, all calibration parameters are held fixed (on the mean), except for one parameter that is randomly sampled from distributions with different variances. Thus, the CNN's sensitivity to variations in different parameters is probed independently. Specifically, for each calibration parameter to be investigated, multiple test sets are generated, each one with a different distribution variance. The variance is scanned starting from zero (i.e. just the mean) throughout the range that was used for training and then continues to grow beyond the training range up to at least $\times 2.5$ of the training range. Figure 3 demonstrates results for three calibration parameters (other parameters demonstrate similar behavior). As can be seen from the test accuracies, performance is maintained within the variance range used for training, and extended well beyond that range. This demonstrates the network ability to learn an invariant model to changes in the calibration parameters within the training range and nearly invariant beyond that range. For example, in Fig. 3(c) the network was trained with data that had the illumination position distributed uniformly within $5cm$ from the mean. Yet, the test performance starts to slightly drop only after the illumination position may be found within $10cm$ of the mean. Qualitative evaluation of these results are also presented in the bottom part of Fig. 3 with PCA projections of the activations from the penultimate layer of the CNN, these demonstrate sustained performance well beyond the training range.

This analysis shows that the network performance is maintained when the calibration parameters deviate from the mean within the training range. Furthermore, even if the network was trained under an assumption of certain ranges for system parameters, the performance degrades slowly if the actual calibration parameters are outside the training range.

3 Experimental Results - Hidden Human Pose Estimation

In order to demonstrate human pose estimation behind scattering medium, a flexible mannequin (length from head to toe $\sim 20cm$) is placed behind a regular paper sheet (Fig. 1). We define three different poses for the mannequin using various positions of hands and legs (Fig. 4).

CNN training is accomplished by synthesizing $24,000$ samples for training and $6,000$ samples for validation. Translations and perturbations to the mannequin's head and limbs are applied to create multiple instances of each pose.

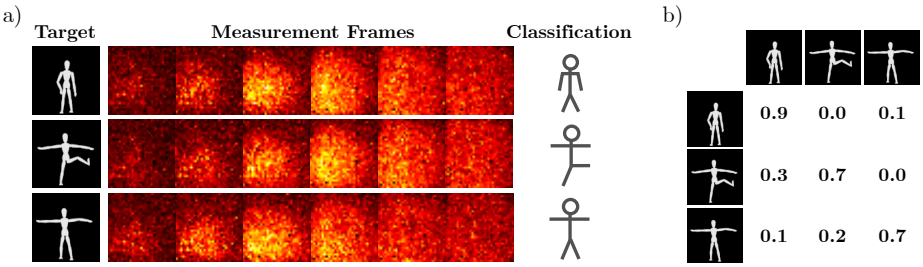

Figure 4: Lab experiments show successful estimation of hidden human pose. a) Three examples (rows) demonstrate target pose, raw SPAD measurement (first six frames), and the successful classification. b) Confusion matrix for classification of raw test set (10 samples per pose).

The test set is composed of 30 raw SPAD measurements, 10 per pose. For each measurement, the mannequin is moved around and the position of the hands, legs and head are adjusted. The CNN classifies correctly 23 out of the 30 tests (76.6% overall accuracy, compared to 33.3% random guess accuracy). Figure 4(a) shows examples of mannequin pose, SPAD measurements, and classification. Figure 4(b) shows the confusion matrix of this raw test set.

Here training is performed on one dataset (synthetic) and tested on another dataset (gathered by lab experiments). In general, it is challenging to train and test on different datasets and it is common to note performance degradation in such cases. The degradation in performance can potentially be mitigated with domain adaptation methods (e.g. [50]), we leave this to a future study.

To further explore the sensitivity to the number of poses we expanded the training set to include seven different poses (Fig. 5 bottom shows illustrations of the poses). The poses include a diverse combination of limb positions. For each label $8,000$ training examples and $2,000$ test examples were generated (total training set of $56,000$ examples and $14,000$ test set examples). Figure 5(a) shows a two dimensional student's t-distributed stochastic neighbor embedding (t-SNE) [51] visualization of activations from the CNN penultimate layer generated on the test set. This visualization demonstrates that the network correctly separates the classes. Figure 5(b) shows the confusion matrix for this synthetic test set. The network is able to classify the seven classes with 91.86% accuracy. The synthetic test accuracy for the network trained only on the three poses (Fig. 4) achieved 96.7%. This indicates the ability to experimentally classify among more poses without significant decrease in accuracy.

4 Evaluation

To evaluate our approach we compare its classification performance to several other classification techniques. The classification task is based on the three man-

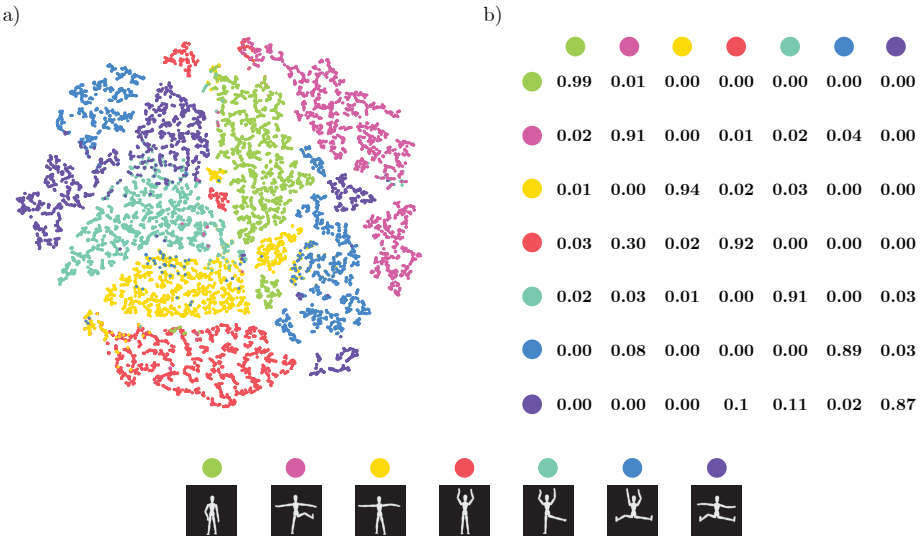

Figure 5: Classification among seven poses on synthetic test dataset. a) t-SNE visualization demonstrates the CNN ability to classify among the seven poses. b) Confusion matrix for classification on the synthetic test dataset.

nequin poses. We create two datasets for evaluation, each one consists of $24,000$ training examples and $6,000$ test examples. The clean dataset demonstrates the algorithms' sensitivity just to variation in calibration parameters (decoupling the sensitivity to measurement quality). The realistic dataset probes the algorithms' ability to classify on the actual lab experiments.

The results are summarized in table 2. While some of the traditional algorithms perform reasonably on the clean dataset, they fail on the realistic dataset. Our approach significantly outperforms the traditional methods on the clean dataset, and as demonstrated previously, it performs well on the lab measurement.

The details of the datasets are:

1. *Clean dataset*: This dataset aims to probe the ability to classify in extreme variation in calibration parameters in a noiseless measurement case. It is based on synthetic measurements with calibration parameters varying in ranges that are twice as large compared to the realistic dataset, and with 10^8 photons without any additive noise (Fig. 2 shows two noiseless examples from this dataset). In this case both training and testing datasets are synthetic.

2. *Realistic dataset*: this is the dataset used for training the network described in section 3. It is based on renderings with 10^6 photons with additive noise to approximate our SPAD measurements (see Fig. 2 syn-

Table 2: Comparison of different approaches on classification of the clean and realistic datasets. The CNN outperforms all methods in the clean dataset, and is the only method that achieves results that are better than random accuracy on the realistic dataset.

Training set	**Clean dataset**	**Realistic dataset**
Mean Example	33.3	33.3
KNN	53.0	30.0
SVM	57.1	20.0
Random forest	68.8	30.0
Single layer network	68.2	23.8
Our CNN	84.0	76.6

thetic examples 1 and 2). In this case the training is performed on the synthetic data and testing is based on the 30 lab measurements.

The different classification approaches that were used for comparison are:

1. *Mean example*: For each label we take the mean of the training data, such that we have one representative sample per label. Classification is performed based on nearest neighbor (closest sample in the dictionary to the measurement). This approach fails on both datasets.

2. *K-nearest neighbors*: Since this method may be sensitive to dictionary size, it is first evaluated on the clean dataset. We randomly choose varying number of samples from the training set to form different dictionary sizes. We consider two approaches here: a) Nearest neighbor — for each test point the chosen label is the label of the closest dictionary element. b) K-nearest neighbors (KNN) — for each test point the chosen label is the label of the majority of the K-nearest neighbors. K is chosen for each dictionary size with a validation set (taken from the training set). These results are presented in Fig. 6. The nearest neighbor approach shows decreased performance with increase in data size due to the increased ambiguity between dictionary elements. The K-nearest approach is able to overcome this limitation and provides classification accuracy in the range of 50% on the clean dataset, however it fails on the realistic dataset.

3. *Support vector machine (SVM)*: The SVM is evaluated with different kernels, and got the best performance with the linear kernel. After hyper parameters optimization we were able to achieve 57.1% classification accuracy on the clean dataset, and fail on the realistic dataset.

4. *Random forest*: A random forest is trained with 100 trees. The random forest achieves 68.2% accuracy on the clean dataset, and fails on the realistic dataset.

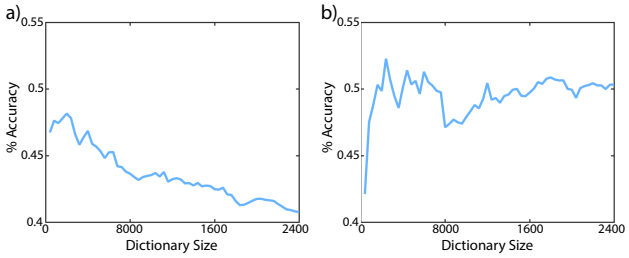

Figure 6: Performance of the K-nearest neighbor approach on the clean dataset. Classification accuracy with varying dictionary size for a) nearest neighbor classifier, and b) K-nearest neighbors classifier.

5. *Single layer network*: A neural network composed of one hidden layer. This network achieves 68.2% accuracy on the clean dataset, and like the previous methods, it fails on the realistic dataset.

This analysis presents the key difficulty and the requirement for both calibration invariance and robustness to noise. While some of the traditional approaches perform reasonably well on the clean data, they fail on the realistic dataset. Our approach is the only one that achieves results that are better than random accuracy.

5 Discussion

While our approach is invariant to variations of calibration parameters within the training range, it still requires some approximate measurements or knowledge of system parameters and geometry. This limitation is somewhat mitigated by the fact that the network can operate well beyond its training regime (see Fig. 3 for examples). Another limitation is the need to synthesize a dataset and train the CNN on different types of geometries, which might slow down the process when arriving to a completely new setting. Faster hardware for data generation and CNN training can potentially address this in the future. Active acquisition systems like the ones used here, may suffer from interference with ambient illumination. This can be more challenging with single photon counting sensors. One possible solution is the use of narrow-band spectral filters to pass only the source's wavelength. These filters are already used in systems such as LIDARs.

The measurement system suggested here uses time-resolved measurements with few spatial pixels (32×32). The importance of temporal resolution for classification when imaging through scattering media is evaluated with the suggested imaging pipeline. The MC model is used to create training and test sets (based on the MNIST dataset) with different time resolutions. The result is plotted in Fig. 7(a), where we note that the performance degrades slowly until the

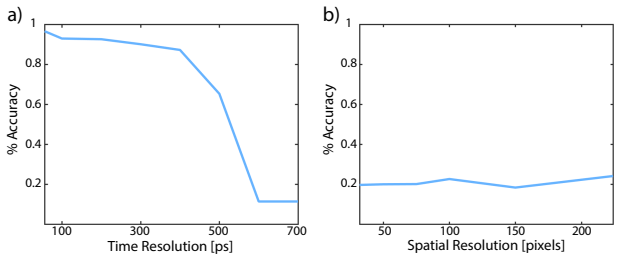

Figure 7: Time resolution is more important than number of pixels for imaging through scattering. a) Classification accuracy vs. time resolution (for 32x32 pixels). b) Classification accuracy vs. number of pixels (for non time-resolved system).

time resolution nears $400ps$ and then degrades rapidly. In the scenes discussed and analyzed here, the time between the first and last signal photons spans roughly $500ps$, so any time resolution better than that provides at least two frames with signal which allows the network to learn temporal filters. As seen from the measurements provided in Fig. 2, the spatial features have very little high frequency content, and therefore, unsurprisingly, low pixel count is sufficient for classification. To quantitatively evaluate this, we use the same pipeline to simulate no time dependency, while varying the pixel count. Figure 7(b) demonstrates that simply adding more pixels doesn't improve the classification accuracy. This analysis is limited to the particular scene considered here and evaluates two extremes: low pixel count with varying time resolution and, no time resolution with varying spatial resolution. This demonstrates theoretical performance of commercially available hardware variants. We leave further analysis of potential hardware with e.g. high pixel count and significant time resolution to a future study.

The importance of time-resolved data for classification with CNN can be observed from the filters the network learns (Fig. 8). Inspection of these indicate that the network performs derivatives in the time domain. Similar spatio-temporal features have been demonstrated when using CNNs for action recognition in videos [46]. The temporal features learned by our network combined with the strong dependency of classification accuracy on the SPAD's time resolution shows that network inference is computed using information in both space and time.

Several aspects can be taken into account when considering the potential of this approach to scale into real-world applications:

- **Hardware:** Our hardware is a SPAD camera. Since SPAD cameras are manufactured with scalable semiconductor processes, they can be commoditized. Other approaches like phase based ToF systems are also a possibility (probably with significantly lower time resolution, which would

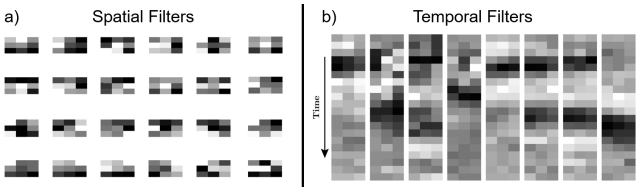

Figure 8: Examples of spatio-temporal filters learned by the CNN. The network generates both a) spatial and b) temporal filters for inference.

impact its ability to classify). We note that SPAD cameras are especially useful for imaging through scattering due to several reasons:

  - They are single-photon sensitive which is extremely useful in NLOS geometries where the optical signal is very weak.
  - The time resolution of $\sim 50ps$ corresponds to $1.5cm$ of spatial resolution, which is reasonable for room-sized scenes.
  - The low spatial resolution is not necessarily a drawback (see above analysis).

- **Real-time operation:** Since classification requires only a forward pass through the trained neural network, it can be performed in real time using specialized hardware (such as GPUs). The only caveat is the case of a completely new scene that requires rendering new synthetic dataset and training a CNN. This requires anticipatory preparation before the real-time operation.

- **Flexibility:** The suggested forward model is based on an MC ray tracer. The MC model is very flexible and can render a wide range of optical geometries and materials.

In this work we demonstrated the task of classification. As mentioned earlier, current machine learning models are very successful in classification tasks. Future work will explore more challenging imaging tasks such as full scene reconstruction. One promising direction is based on the recent research in generative models (e.g. [52, 53]). In this case the network can be composed of two parts. The first, an encoder that gathers the underlying information described by the scene from the captured measurement (this part is similar to the classification task demonstrated here). The second part is a decoder that would generate the full scene based on the recovered information (similar to a graphics rendering program). While each part has been demonstrated separately, the combination is challenging.

6 Conclusions

We have demonstrated a method for object classification through scattering that is effectively invariant to variations in calibration parameters. The approach leverages the ability of neural networks to learn invariants from synthetic random data. We show that the network is invariant to changes in the forward model parameters within the training range (and nearly invariant beyond that range) for the purposes of classification. An important cornerstone of our approach is its ability to generate synthetic data based on a generic forward model that is used to train and evaluate the neural network. This data driven approach can alleviate lengthy experimental calibrations that were needed before.

References

[1] D. Berman, T. treibitz, and S. Avidan, "Non-local image dehazing," The IEEE Conference on Computer Vision and Pattern Recognition (2016).

[2] V. Holodovsky, Y. Y. Schechner, A. Levin, A. Levis, and A. Aides, "In-situ multi-view multi-scattering stochastic tomography," The IEEE International Conference on Computational Photography (2016).

[3] M. Sheinin and Y. Y. Schechner, "The next best underwater view," The IEEE Conference on Computer Vision and Pattern Recognition (2016).

[4] I. Gkioulekas, S. Zhao, K. Bala, T. Zickler, and A. Levin, "Inverse volume rendering with material dictionaries," ACM Transactions on Graphics **32**, 162:1–162:13 (2013).

[5] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, and J. G. Fujimoto, "Optical coherence tomography," Science **254**, 1178–1181 (1991).

[6] Z. Yaqoob, D. Psaltis, M. S. Feld, and C. Yang, "Optical phase conjugation for turbidity suppression in biological samples," Nature Photonics **2**, 110–115 (2008).

[7] I. M. Vellekoop, A. Lagendijk, and A. P. Mosk, "Exploiting disorder for perfect focusing," Nature Photonics **4**, 320–322 (2010).

[8] O. Katz, E. Small, Y. Guan, and Y. Silberberg, "Noninvasive nonlinear focusing and imaging through strongly scattering turbid layers," Optica **1**, 170–174 (2014).

[9] R. Horstmeyer, H. Ruan, and C. Yang, "Guidestar-assisted wavefront-shaping methods for focusing light into biological tissue," Nature Photonics **9**, 563–571 (2015).

[10] J. Bertolotti, E. G. van Putten, C. Blum, A. Lagendijk, W. L. Vos, and A. P. Mosk, "Non-invasive imaging through opaque scattering layers." Nature **491**, 232–234 (2012).

[11] 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**, 784–790 (2014).

[12] X. Xu, H. Liu, and L. V. Wang, "Time-reversed ultrasonically encoded optical focusing into scattering media," Nature Photonics **5**, 154–157 (2011).

[13] X. Wang, Y. Pang, G. Ku, X. Xie, G. Stoica, and L. V. Wang, "Noninvasive laser-induced photoacoustic tomography for structural and functional in vivo imaging of the brain," Nature Biotechnology **21**, 803–806 (2003).

[14] L. V. Wang and S. Hu, "Photoacoustic tomography: in vivo imaging from organelles to organs," Science **335**, 1458–1462 (2012).

[15] G. Satat, B. Heshmat, N. Naik, A. Redo-Sanchez, and R. Raskar, "Advances in ultrafast optics and imaging applications," SPIE Defense+Security pp. 98350Q–98350Q–13 (2016).

[16] O. Gupta, T. Willwacher, A. Velten, A. Veeraraghavan, and R. Raskar, "Reconstruction of hidden 3d shapes using diffuse reflections," Optics Express **20**, 19096–19108 (2012).

[17] A. Velten, T. Willwacher, O. Gupta, A. Veeraraghavan, M. G. Bawendi, and R. Raskar, "Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging," Nature Communications **3** (2012).

[18] G. Satat, B. Heshmat, C. Barsi, D. Raviv, O. Chen, M. G. Bawendi, and R. Raskar, "Locating and classifying fluorescent tags behind turbid layers using time-resolved inversion," Nature Communications **6** (2015).

[19] F. Heide, L. Xiao, W. Heidrich, and M. B. Hullin, "Diffuse mirrors: 3d reconstruction from diffuse indirect illumination using inexpensive time-of-flight sensors," The IEEE Conference on Computer Vision and Pattern Recognition (2014).

[20] F. Heide, L. Xiao, A. Kolb, M. B. Hullin, and W. Heidrich, "Imaging in scattering media using correlation image sensors and sparse convolutional coding," Optics Express **22**, 26338–26350 (2014).

[21] A. Kadambi, H. Zhao, B. Shi, and R. Raskar, "Occluded imaging with time-of-flight sensors," ACM Transactions on Graphics **35**, 15:1–15:12 (2016).

[22] A. Bhandari, C. Barsi, and R. Raskar, "Blind and reference-free fluorescence lifetime estimation via consumer time-of-flight sensors," Optica **2**, 965–973 (2015).

[23] C. Jin, Z. Song, S. Zhang, J. Zhai, and Y. Zhao, "Recovering three-dimensional shape through a small hole using three laser scatterings," Optics Letters **40**, 52–55 (2015).

[24] D. Raviv, C. Barsi, N. Naik, M. Feigin, and R. Raskar, "Pose estimation using time-resolved inversion of diffuse light," Optics Express **22**, 20164–20176 (2014).

[25] G. Gariepy, F. Tonolini, R. Henderson, J. Leach, and D. Faccio, "Detection and tracking of moving objects hidden from view," Nature Photonics **10**, 23–26 (2015).

[26] A. Redo-Sanchez, B. Heshmat, A. Aghasi, S. Naqvi, M. Zhang, J. Romberg, and R. Raskar, "Terahertz time-gated spectral imaging for content extraction through layered structures," Nature Communications **7** (2016).

[27] G. Satat, B. Heshmat, D. Raviv, and R. Raskar, "All photons imaging through volumetric scattering," Scientific Reports **6** (2016).

[28] 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**, 20997–21011 (2015).

[29] M. Laurenzis, F. Christnacher, J. Klein, M. B. Hullin, and A. Velten, "Study of single photon counting for non-line-of-sight vision," Proc. SPIE **9492**, 94920K–94920K–8 (2015).

[30] 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** (2016).

[31] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence **35**, 1798–1828 (2013).

[32] S. Mallat, "Understanding deep convolutional networks," Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **374** (2016).

[33] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," Journal of Machine Learning Research **11**, 3371–3408 (2010).

[34] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," The IEEE Conference on Computer Vision and Pattern Recognition (2014).

[35] L. Waller and L. Tian, "Computational imaging: Machine learning for 3D microscopy," Nature **523**, 416–417 (2015).

[36] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "Reconnet: Non-iterative reconstruction of images from compressively sensed measurements," The IEEE Conference on Computer Vision and Pattern Recognition (2016).

[37] A. Profeta, A. Rodriguez, and H. S. Clouse, "Convolutional neural networks for synthetic aperture radar classification," Proc. SPIE **9843**, 98430M–98430M–10 (2016).

[38] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," Proceedings of the IEEE **104**, 2207–2219 (2016).

[39] R. Horisaki, R. Takagi, and J. Tanida, "Learning-based imaging through scattering media," Optics Express **24**, 13738–13743 (2016).

[40] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," IEEE Transactions on Image Processing **25**, 5187–5198 (2016).

[41] A. Sinha, J. Lee, S. Li, and G. Barbastathis, "Lensless computational imaging through deep learning," arXiv preprint arXiv:1702.08516 (2017).

[42] A. Abdolmanafi, L. Duong, N. Dahdah, and F. Cheriet, "Deep feature learning for automatic tissue classification of coronary artery using optical coherence tomography," Biomedical Optics Express **8**, 1203–1220 (2017).

[43] T. Ando, R. Horisaki, and J. Tanida, "Speckle-learning-based object recognition through scattering media," Optics express **23**, 33902–33910 (2015).

[44] A. Saade, F. Caltagirone, I. Carron, L. Daudet, A. Drémeau, S. Gigan, and F. Krzakala, "Random projections through multiple optical scattering: Approximating kernels at the speed of light," IEEE International Conference on Acoustics, Speech and Signal Processing pp. 6215–6219 (2016).

[45] J. Carlsson, P. Hellentin, L. Malmqvist, A. Persson, W. Persson, and C. G. Wahlström, "Time-resolved studies of light propagation in paper," Applied Optics **34**, 1528–1535 (1995).

[46] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," The IEEE Conference on Computer Vision and Pattern Recognition (2014).

[47] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," Advances in neural information processing systems pp. 568–576 (2014).

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2014).

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," The IEEE Conference on Computer Vision and Pattern Recognition (2016).

[50] S. Kumar and A. Savakis, "Robust domain adaptation on the l1-grassmannian manifold," The IEEE Conference on Computer Vision and Pattern Recognition Workshops (2016).

[51] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research **9**, 2579–2605 (2008).

[52] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," Advances in Neural Information Processing Systems pp. 2539–2547 (2015).

[53] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems pp. 2672–2680 (2014).