Applied Numerical Methods in Characterizing
Optical Nanomaterials and Nanostructures

A Dissertation in
Electrical Engineering

by

Joshua Alton Noble

© 2020 Joshua Alton Noble

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2020
The dissertation of Josh Noble was reviewed and approved by the following:

Zhiwen Liu
Professor of Electrical Engineering
Dissertation Advisor
Chair of Committee

Tim Kane
Professor of Electrical Engineering

Xingjie Ni
Charles H. Fetter Assistant Professor of Electrical Engineering

Raj Kothapalli
Assistant Professor of Biomedical Engineering

Hans Hallen
Special Member
Professor of Physics
North Carolina State University

Kultegin Aydin
Professor of Electrical Engineering
Head of the Department of Electrical Engineering
Abstract

I am a tool builder

This dissertation represents the distillation of thousands of man hours and tens of thousands of lines of code for creating a novel set of tools for characterizing 2D materials, nanostructures, and metamaterials by combining established techniques in ultrafast, nonlinear, and holographic microscopy with applied numerical methods such as phase retrieval, SNR boosting, and machine learning. The goal is to create a toolbox for researchers to quickly and efficiently characterize new nanostructures. As such, we will focus more on the development and application of these tools and less so on the specific material properties.

Chapter 2 gives a broad introduction to the field of nanomaterials, nanostructures, and their applications, specifically in the field of optics. It addresses some of the current characterization techniques and their specific strengths and weakness. It also introduces certain applied numerical methods, including machine learning, and their capabilities in extracting additional information.

Chapter 3 discusses the use holographic microscopy for quantifying the refractive index and layer thicknesses of two-dimensional nanomaterials. Holography on its own can capture the complex field reflected off a substrate. We expand on this by first numerically modelling a 2D material on a substrate using known parameters. By fitting the simulation model to the experimental results, we retrieve the unknown parameters for the refractive index and layer thickness. Our results are in good agreement with the current gold standard techniques such as ellipsometry and atomic force microscopy (AFM).
Chapter 4 focuses on ultrafast optical pulse reconstruction. Frequency resolved optical gating (FROG) via an iterative reconstruction algorithm has been a well-established method since the 1990s. However, traditional FROG is typically implemented using an off-axis geometry. Experimentally, it is impossible to realize this geometry in the near-field. Instead, collinear FROG (cFROG) measurement can be performed and the off-axis FROG term is numerically obtained by digital filtering. However, this filter choice is subjective and it is difficult to make a quantitative error comparison back to the original cFROG spectrogram. To counter this, we develop a machine learning based reconstruction method which retrieves the complex field and incorporates any requisite filtering into the model, eliminating the subjectivity.

In Chapter 5, near-field scanning optical microscopy (NSOM) is used to characterize optical nanostructures. NSOM is a super-resolved optical technique because it collects the near-field signal generated from a nanostructure and is therefore, not subject to the diffraction limit. By combining NSOM, ultrafast optics, and pulse retrieval, we develop a tool that can be used for characterizing the full spatio-temporal field on the nanometer-femtosecond scale. Additionally, we develop a numerical routine for boosting the signal to noise ratio of low count optical measurements to extract meaningful data from low SNR scans. Furthermore, we are the first to retrieve the temporal near field and show that evanescently collecting the near-field using an NSOM probe does not distort its temporal profile via both the iterative pulse retrieval algorithm and our custom built convolutional neural network.

Finally, Chapter 6 summarizes how these techniques can be applied and extended to other materials and applications. Merging the techniques in this dissertation provides researchers with an extensive experimental and numerical toolbelt for investigating the basic physics of optical nanostructures.
# Table of Contents

List of Figures ................................................................. vi
List of Tables ..................................................................... xii
Acknowledgements ............................................................ xiii
1: Introduction ................................................................... 1
2: Review of 2D Materials and Nanostructures ......................... 5
   2.1 – Optical Properties of 2D Materials ............................... 5
   2.2 – Engineered Nanostructures and Metamaterials .................. 9
   2.3 – Nanostructure Characterization Techniques ..................... 12
   2.4 – Applied Numerical Methods in Nanomaterial Characterization 14
3: Quantitative Holographic Imaging of 2D Materials ................. 16
   3.1 – Current Techniques for Characterizing 2D Materials ........ 16
   3.2 – Formalism of Digital Holography ................................. 19
   3.3 – Numerical Modeling of TMD and Substrate ..................... 24
   3.3 – Holographic Imaging of TMDs ...................................... 28
   3.4 – Complex Refractive Index Prediction of TMDs ................... 33
   3.5 – Layer Deposition Prediction of TMDs ............................. 37
4: Ultrafast Optical Pulse Characterization ................................ 42
   4.1 – Frequency Resolved Optical Gating (FROG) ..................... 42
   4.2 – Iterative FROG Pulse Retrieval ..................................... 44
   4.3 – Collinear vs. Noncollinear FROG ................................. 47
   4.4 – cFROG Filtering ......................................................... 50
   4.5 – Experimental Retrieval of Ultrafast Optical Pulses .......... 54
   4.6 – Neural Network Reconstruction of Ultrafast Optical Pulses 61
   4.7 – Neural Network Replication of cFROG Noise .................. 74
5: Near-Field Microscopy of Nanomaterials ............................. 79
   5.1 – Near-Field Scanning Optical Microscopy (NSOM) ............ 79
   5.2 – Algorithm Design of SNR Enhancement .......................... 88
   5.3 – Super-Resolved Nonlinear Microscopy of 2D Materials ....... 99
   5.4 – Near-Field Optical Pulse Reconstruction ....................... 103
6: Conclusions and Future Work ........................................ 113
References ......................................................................... 118
List of Figures

Fig. 2.1: Atomic structure of a TMD containing transition metal atoms (blue) and chalcogen atoms (silver). A) Side view of TMD monolayer. B) Top down view of TMD monolayer. C) Side view of TMD bilayer with symmetry point shown. Figure courtesy of Dr. William Murray. .................................................. 7


Fig. 2.6: Second harmonic image of MoS\textsubscript{2} monolayers exhibiting optical edge effects. A) SHG excited at 1280nm. B) SHG excited at 1300nm. C) Line scans across both SHG images [59]. From X. Yin et al., “Edge nonlinear optics on a MoS\textsubscript{2} atomic monolayer,” Science (80-. ), vol. 344, no. 6183, pp. 488–490, May 2014, doi: 10.1126/science.1250564. Reprinted with permission from AAAS. ........................................................................................................ 13

Fig. 3.1: Holographic CCD capture of a sample containing multiple TMD flakes........................................ 20

Fig. 3.2: Fourier transform of the CCD capture showing the DC term, two sidebands, and filter................ 21

Fig. 3.3: Momentum space diagram for determining the holographic grating pitch from the sample and reference beams........................................................................................................ 22


Fig. 3.5: Expected amplitude (A) and phase response (B) for a MoS\textsubscript{2}/Si/SiO\textsubscript{2} stack vs. 2D material layer thickness at various wavelengths [90]. A. Ghosh, J. Noble, A. Sebastian, S. Das, and Z. Liu, “Digital holography for non-invasive quantitative imaging of two-dimensional materials,” J. Appl. Phys., vol. 127, no. 8, p. 084901, Feb. 2020, doi: 10.1063/1.5128135. Reproduced with the permission of AIP Publishing........................................................................................................ 27


Fig. 3.8: A) Phase reconstruction of a TMD flake showing phase wrapping from wavefront curvature. B) Line scan taken across (A). C) Amplitude reconstruction of a TMD flake showing a Gaussian envelope. D) Line scan taken across (C). ........................................................................................................ 32

Fig. 3.9: Mean squared error plots between the predicted and measured values. A) Amplitude error. B) Phase error. C) Cumulative error. ........................................................................................................ 34

Fig. 3.10: Reported refractive indices and extinction coefficients for monolayer WS$_2$ and MoS$_2$ [8], [83], [95], [98]. Figure courtesy of Dr. William Murray. ........................................................................................................ 35

Fig. 3.11: Predicted amplitude and phase response for monolayer WS$_2$ at 672nm as a function of refractive index and extinction coefficient. The red dots indicate the best fit within one standard deviation of the measured values with the mean value occurring at the center of the cluster [90]. A. Ghosh, J. Noble, A. Sebastian, S. Das, and Z. Liu, “Digital holography for non-invasive quantitative imaging of two-dimensional materials,” J. Appl. Phys., vol. 127, no. 8, p. 084901, Feb. 2020, doi: 10.1063/1.5128135. Reproduced with the permission of AIP Publishing. ........................................ 36


Fig. 4.1: Initial guess for the pulse shape and the corresponding simulated signal, $E_{\text{sig}}(t, \tau)$. ........................................................................................................ 44

Fig. 4.2: Schematic of the principal components generalized projections FROG pulse retrieval algorithm (PCGPA) [103]. ........................................................................................................ 45

Fig. 4.3: Original (simulated) and reconstructed FROG spectrograms with a weighted MSE of 0.13% after processing for 1000 iterations. ........................................................................................................ 46

Fig. 4.4: Noncollinear FROG acquisition geometry. The FROG term is spatially separated due to its distinct propagation direction. ........................................................................................................ 48

Fig. 4.5: Collinear FROG acquisition geometry. ........................................................................................................ 49
Fig. 4.6: Filtering a simulated cFROG spectrogram. A) Original simulated cFROG spectrogram. B) Fourier transform of cFROG along the delay axis. C) Filtered spectrogram after filtering to remove higher order terms. D) FROG spectrogram after filtering and removal of the background.................51

Fig. 4.7: Weighted MSE between the noncollinear and filtered FROG spectrograms of a simulated pulse as a function of filter size........................................................................................................52

Fig. 4.8: Optical set up used to collect cFROG data. A Ti-Sapphire laser beam is split in a Michelson interferometer with a delay stage in one arm. The two beams are focused and generate SHG off a WS$_2$ monolayer. This signal is epicollected and separated by a dichroic mirror for detection.55

Fig. 4.9: Raw experimental capture of a cFROG spectrogram generated off a WS$_2$ monolayer..............55

Fig. 4.10: Filtering an experimentally collected cFROG spectrogram. A) Fourier transform along the delay axis. B) Super-Gaussian filter. C) Original cFROG spectrogram. D) Resulting FROG spectrogram with a slight residue due to smearing of the higher order terms.......................................................57

Fig. 4.11: Reconstruction of the experimental Ti-Sapphire laser pulse. A) FROG spectrogram after preprocessing and filtering. B) Reconstructed FROG spectrogram from the PCGPA retrieval. C) Reconstructed field (time domain). D) Reconstructed field (frequency domain).....................58

Fig. 4.12: Raw and filtered autocorrelation corresponding to the previous cFROG spectrogram. The autocorrelation predicts a temporal FWHM of 69fs compared to 74fs from the full iterative pulse retrieval.................................................................59

Fig. 4.13: Weighted mean squared error between the original experimentally collected cFROG and the cFROG created after reconstruction. A) Weighted MSE as a function of spectral detuning. B) Weighted MSE as a function of time shift..........................................................60


Fig. 4.15: Reconstruction of 4 random pulses unknown to the training set. True pulses in blue and reconstructions in orange. The absolute value reconstructions are near exact while the phase reconstructions show more error.............................................................................63

Fig. 4.16: Overview of the DeepcFROG neural network structure.............................................................................................................64

Fig. 4.17: DeepcFROG convolutional neural network consisting of several hidden convolutional layers along with the custom filtering and FROG layers. .........................................................................................65

Fig. 4.18: True cFROG and reconstructed cFROG spectrograms after propagating through the neural net. ........................................................................................................................................66

Fig. 4.19: Time domain representation of the original pulse and the one reconstructed by the DeepcFROG neural network. Absolute value shown in the top plot and phase in the bottom [111].............67

Fig. 4.20: A) Experimental cFROG spectrogram after Fourier transforming along the delay axis into $I(\omega,k)$. B) $I(\omega,k)$ after cropping to a square region to isolate the baseband and background terms.....70

Fig. 4.21: Top) Comparison of the pulse intensity profiles between the CNN and iterative reconstruction. The FWHM is only different by 2fs, but the CNN reconstruction exhibits broadening. Bottom) Comparison of the phase between the CNN and iterative reconstructions..........................71

Fig. 4.22: WMSEs calculated between the experimental and reconstructed $I(\omega,k)$ terms for several shifts along both axes. The shifts strongly affect the error, but have minimal correlation to the pulse shape........................................................................................72
Fig. 4.23: A) Experimental $I_{\omega,k}$ generated by Fourier transforming along the $\tau$ axis and cropping the baseband and DC terms. B) Reconstructed $I(\omega,k)$ generated from the CNN. C) Reconstructed $I(\omega,k)$ generated by first iteratively retrieving the pulse via the PCGPA, constructing the full cFROG trace, Fourier transforming along the delay axis, then cropping............................................. 73

Fig. 4.24: A) Experimental $I_{\omega,k}$ term cropped to only the DC and baseband terms. B) Same term plotted logarithmically to better visualize the sources of error. Error due to background speckle circled in black. Error due to background speckle and the spectrometer noise floor is circled in white. Error due to imperfect stage translation and sampling is circled in red.................................................. 75

Fig. 4.25: A) Simulated noise free $I(\omega,k)$ spectrogram plotted logarithmically. B) Simulated noisy $I(\omega,k)$ spectrogram with the 3 noise sources tuned to match the experimental parameters.............................. 76

Fig. 4.26: Top) Comparison of the pulse intensities profiles between the noise trained CNN and the iterative reconstruction. Bottom) Comparison of the phases between the CNN and iterative reconstruction................................................................. 77

Fig. 5.1: A) Photoluminescent image of a WS$_2$ monolayer exhibiting edge enhancement. B) Zoomed in image showing that far-field imaging is not sufficient to resolve the scale on which edge enhancement occurs. Figure courtesy of Dr. Ethan Kahn................................................................. 80

Fig. 5.2: Diagram of near-field and far-field propagation. The near-field decays exponentially away from the sample surface........................................................................................................ 81

Fig. 5.3: Various geometries that can be used for near-field scanning optical microscopy. A) Apertureless NSOM (sSNOM) using an oscillating tip and collecting the scattered light. B) Transmission mode NSOM exciting through the tip and collecting in the far-field. C) Transmission mode NSOM exciting in the far-field and collecting the near-field through the tip. D) Reflection mode NSOM exciting in the far-field and collecting the near-field through the tip. E) Reflection mode NSOM exciting through the tip and collecting in the far-field. F) Excitation and collection both through the tip. Geometry (C) is used in our experiments.............................................................................. 84

Fig. 5.4: A) SEM image of an NSOM tip that produces super-resolved SHG images. B) Mechanical setup at the core of the NSOM. The tip oscillates at a fixed position while the sample (e.g., a 2D material) is placed on a transparent substrate which is kept in feedback by a piezo tube. Far-field illumination is provided via the other side in transmission mode. C) Resonance curve for an NSOM tip not yet in feedback. Peak resonance is slightly below 61kHz. D) Monitoring the feedback in signal over 500ms. Roughly 20mV of noise on the feedback in signal corresponds to a spatial instability of 19nm. (A) SEM image courtesy of Yao Duan. (B,D) Figures courtesy of Dr. William Murray................................................................. 86

Fig. 5.5: A) Homebuilt microscope designed by Prof. Hans Hallen of NC State and Dr. William Murray of Penn State. The entire NSOM system is enclosed in a Styrofoam and metal enclosure to minimize thermal fluctuations and air currents. The inside of the enclosure is wrapped in DynaMat and sandbags are placed on the table to dampen acoustic and mechanical vibrations. 1) Transparent substrate on which the sample is mounted. 2) Extender tube for the piezo scanner. 3) Clamp for attaching the piezo tube. 4) Aluminum base plate for attaching the 5-axis piezo driven stage to the main piezo tube. 5) 5-axis New Focus stage with picomotor piezo drivers. 6) Aluminum baseplate. 7) Block for mounting the tip assembly. 8) Pre-amp for the tip assembly. 9) NSOM probe tip. 10) Newport 562 3-axis translation stage. 11) Picomotors for transrating the Newport 562 3-axis stage. 12) Main base plate. 13) PCB to amplify and send signal from tip pre-amp to the lock-in amplifier.................................................................................................................. 87

Fig. 5.6: A) Far-field image of a WS$_2$ monolayer in the NSOM (green) and the tip nearby in reflection mode (white) before laser illumination. B) Far-field image of the sample after burning............................................. 89

Fig. 5.7: A) 1000ms exposure raster scan of a WS$_2$ monolayer at 2mW of incident power. B) 300ms exposure raster scan of a WS$_2$ monolayer at 250$\mu W$ of incident power................................................................. 90
Fig. 5.8: A) SHG spectrum at the center of the raster scan shown previously in Fig. 5.7A. B) SHG spectrum at the center of the raster scan shown previously in Fig. 5.7B. ................................................................. 91

Fig. 5.9: A) 250μW incident power, 300ms integration time SHG raster scan of a WS$_2$ monolayer scan after thresholding to remove hot pixels and summing along the spectral domain in the window of the SHG signal. B) Raw spectrum showing the bounds of where the spectral summation is performed. ........................................................................................................................................ 92

Fig. 5.10: A) High signal spectrum of an SHG region. B) High signal spectrum of a background region. C) Low signal spectrum of an SHG region. D) Low signal spectrum of a background region. ........................................ 92

Fig. 5.11: Reconstructions of the background and SHG regions for the SNR boosting regression-based CNN are virtually identical. ................................................................................................................................. 93

Fig. 5.12: Validation loss for the regression-based CNN plateaus very early after only 10 iterations. The model is unable to differentiate SHG vs. background regions............................................................ 94

Fig. 5.13: Spectral preprocessing sequence. A) Raw low signal experimental SHG spectrum. B) SHG spectrum after removing hot pixels and spectral filtering. C) SHG spectrum with Gaussian fit applied. ......................................................................................................................................... 95

Fig. 5.14: Example of a false fit to noise in a background region containing no SHG signal. ......................... 96

Fig. 5.15: ANN for classifying preprocessed spectrums into either SHG or background categories.............. 97

Fig. 5.16: A) Original low signal (250μW incident power and 300ms integration time) raster scan of a WS$_2$ monolayer. B) Low signal raster scan after classifying SHG vs background spectrums from the ANN. C) Low signal WS$_2$ monolayer raster scan after denoising and spectral integration. D) Low signal raster scan after total denoising algorithm and ANN classification with peak second harmonic signal 3-4 counts above the background. ................................................................................................. 98

Fig. 5.17: Far-field second harmonic raster scan of a WS$_2$ monolayer deposited on a quartz substrate. A) Raw SHG raster scan. B) Raster scan after applying the denoising and fitting algorithm. C) Raw line scan taken at x = 600nm. D) Line scan taken at x = 600nm from the fitted raster scan.... 100

Fig. 5.18: Near-field second harmonic raster scan of the same WS$_2$ monolayer deposited on a quartz substrate. A) Raw SHG raster scan. B) Raster scan after denoising and fitting algorithm. C) Raw line scan taken at x = 300nm. D) Line scan taken at x = 300nm from the fitted 2D raster scan. ......................................................................................................................................... 101

Fig. 5.19: Polar plot (in degrees) as a function of incident polarization for the coupled fundamental and second harmonic near-field signals................................................................. 102

Fig. 5.20: Experimental setup detailing the optics around the primary NSOM enclosure. A Ti:Sapphire femtosecond laser is first precompensated by a prism pair and then passed into an interferometer with an attached delay stage. The cFROG is generated from a TMD mounted on the end of the piezo tube in the NSOM. Far-field SHG is epi-collected and sent to the spectrometer while the near-field SHG is collected evanescently by the NSOM probe tip. Figure courtesy of Dr. William Murray [113] ................................................................. 104

Fig. 5.21: Reconstruction of the complex field from the cFROG spectrogram collected off a WS$_2$ monolayer and captured in the far-field. Top) FROG spectrogram after filtering and the reconstructed FROG spectrogram after the PCGPA. Middle) Reconstructed time domain field separated into absolute value and phase. Bottom) Reconstructed spectral domain field separated into absolute value and phase......................................................................................................................................... 105

Fig. 5.22: Reconstruction of the complex near-field from the cFROG spectrogram collected off a WS$_2$ monolayer and captured evanescently through an NSOM probe in feedback. Top) FROG spectrogram after filtering and the reconstructed FROG spectrogram after the PCGPA. Middle) Reconstructed time domain field separated into absolute value and phase. Bottom) Reconstructed spectral domain field separated into absolute value and phase................................. 106
Fig. 5.23: Comparison between the far-field and near-field pulse reconstructions via the iterative PCGPA after manually correcting for FROG ambiguities showing the overlaid absolute values (top) and phases (bottom)........................................................................................................................................107

Fig. 5.24: Comparison between the far-field and near-field CNN pulse reconstructions showing the overlaid absolute values (top) and phases (bottom).................................................................................................................................108

Fig. 5.25: Comparison between the PCGPA iteratively reconstructed near-fields at incident polarizations of 25, 38, 51, 65, 76, and 85 degrees showing the absolute values (top) and phases (bottom).........................................................109

Fig. 5.26: Comparison between the CNN reconstructed near-fields at incident polarizations of 25, 38, 51, 65, 76, and 85 degrees showing the absolute values (top) and phases (bottom).................................110

Fig. 5.27: A) Near-field polarization plot with each green dot indicating a location where the field is reconstructed. B) FWHM near-field pulse widths computed from the CNN retrieval, PCGPA retrieval, and autocorrelation. C) Weighted MSEs between the experimentally collected near-field cFROGs and reconstructed cFROGs for the CNN and PCGPA.................................................111

Fig. 6.1: Diagram of measuring a nanostructure in the NSOM. Moving the nonlinear interaction from the substrate to the tip permits temporal characterization of structures that do not intrinsically possess a second order nonlinear susceptibility.................................................................................................................................116
List of Tables

Table 6.1: Comparison between the lateral, axial, and temporal resolution of select microscopy techniques [15], [16], [76], [77], [101], [108], [121]. ............................................................................................................. 116
**Acknowledgements**

This research has been a massively collaborative effort and for that I consider myself very fortunate. No one truly learns in a vacuum and I would not be a tenth of the way without their help and support. I must first thank my old friend and fellow colleague, Dr. William Murray. We began this journey over a decade ago in undergrad and have shared the highs, lows, and sighs of relief though classes, research, and far too many side projects to count. He is one of the most patient, hardworking, and downright tenacious people I have ever met and there is no way I would have hung in for the long haul without him. Within the scope is this dissertation, he is responsible for most of the construction of the NSOM and providing much of the experimental data for raster scans, the NSOM, and FROG.

In a very close second, I must thank my advisor, Prof. Zhiwen Liu. His patience, support, and constructive criticism have been critical in leading me to where I am today. He has looked at me as both a student and a peer and in retrospect, I cannot think of a better person to be mentored by. I must also extend a special thanks to Prof. Hans Hallen of NC State for his collaboration on the NSOM. I doubt there is another living person who has attained such knowledge and mastery in so many areas of engineering, physics, software development, and more. Thanking him and his former student, Dr. Brandon Long, for their patience and assistance is an understatement. Thanks also to my other committee members, Prof. Tim Kane, Prof. Xingjie Ni, and Prof. Raj Kothapalli.

I would also like to specially thank Chen Zhou for his collaboration in the machine learning optical pulse reconstruction and Dr. Atriya Ghosh for his collaboration in the holography project and performing the AFM verification. I must
also thank all of our collaborators for providing us the materials to test and validate our methods including Dr. Ethan Kahn, Tianyi Zhang, Da Zhou and others.

I also thank many of the outstanding people along the way for their input and assistance including Cheng-Yu Wang, Landon Hernandez, Dr. Corey Janisch, Dr. Alex Cocking, and Dr. Chenji Zhang from Penn State and several of my old group members, Prof. Bob McLeod, Dr. David Glugla, Dr. Darren Forman, Dr. Martha Bodine, Dr. Ben Kowalski, and Dr. David Miller from my time at the University of Colorado at Boulder. Their guidance and input in my early years of graduate school was critical in my success.

I first decided to go to graduate school in the spring of my sophomore year after working with Prof. Gary Weisel and Prof. Darin Zimmerman at the Penn State Altoona Physics department. They were extraordinary mentors to me at the time and sparked my desire to conduct research and development. I must also thank the Penn State Electrical Engineering Department for both their financial support and mentoring over the past decade, rest in peace Prof. John Mitchell. Special thanks also to SherryDawn Jackson and Megan Bowersox for administrative assistance over the years. Finally, I must thank my fiancé Cat, and my friends and family for their support over the years. They’ve reminded me of what I’m working towards and have always kept me pointed true north.

Lastly, I would like to acknowledge the financial support from NSF awards ECCS 1711099 and EFRI 1433311. Our findings do not necessarily reflect the views of the funding agency.
1: Introduction

Nanomaterials, nanostructures, and metamaterials have become the fundamental building blocks for creating novel optical elements. A multitude of papers each year are published demonstrating new and novel properties. Many more show efficacy in creating new optical elements with wide ranging applications in photovoltaics, opto-electronics, wavefront shaping, and many more with far reaching applications to not only the research sector but also with direct commercial applications [1]–[3]. Characterizing these materials presents a unique challenge though due to spatial scale on which such physics occur. In Chapter 2, we present a broad introduction to such nanomaterials and artificially engineered nanostructures along with current characterization methods.

When examining 2D materials, the difference of a single atomic layer among few layered depositions will result in widely varying opto-electronic properties due to the difference in the band structure [4], [5]. Techniques such as atomic force microscopy (AFM) are proven to reliably resolve the difference between single and few-layered depositions, however, such scanning techniques are prohibitive for spatially large and time-sensitive substrates due to the requisite measurement time [6]. Similarly, contact mode AFM can be potentially destructive, especially to sensitive single layered substrates. In Chapter 3, we apply digital holography to resolve single and few layer atomic depositions. Because holography is single shot instead of a scanning technique, the measurement time does not scale with the sample size. Furthermore, we numerically model the multilayered sample/substrate system to predict the thickness of multi-layered depositions with an accuracy comparable to AFM. Our method also predicts the complex refractive index of TMD
monolayers with resultant values consistent with those reported in the literature [7], [8].

A similar set of challenges exist for characterizing artificially engineered optical nanostructures, commonly known as metasurfaces or metamaterials. The key attribute of engineered nanostructures for optical applications is that they modify the phase of an electric field on the subwavelength scale [9], [10]. Therefore, true characterization of optical nanostructures requires subwavelength, super-resolved characterization. However, the vast majority of optical characterization techniques rely on far-field signal collection, limiting the spatial scale at which such effects can be quantified. Some metasurfaces also exhibit a coupling between the spatial and temporal field evolutions on the nanometer-femtosecond scale [11]. Due to the lack of optical characterization tools, the corresponding near-field optical characteristics of these structures are often simulated but rarely measured [11]–[13].

Tools such as stimulated emission depletion microscopy (STED) have been demonstrated to resolve structures spatially down to below 30nm, well into super-resolved territory [14]. Furthermore, techniques such as fluorescent lifetime imaging (FLIM) can probe the time response of a sample on the nanosecond scale and can be fused together with STED to enable temporal gating [15]. However, with the exception of very specific resonant samples, such techniques require fluorescent protein markers [16]. Therefore, these techniques are unsuitable for characterizing materials without fluorescent labeling.

As many nanostructures have both a spatial and temporal response, we must address both challenges. In Chapter 5, we address the temporal characterization problem which includes both the intensity and phase profiles at the femtosecond scale. Well-developed approaches for retrieving the complex field include frequency resolved optical gating (FROG) and similar approaches such as GRENOUILLE,
SPIDER, and ptychographic reconstruction [17]–[19]. Specifically, FROG involves gating a pulse with itself and recovering the complex field via an iterative reconstruction algorithm. Recently, pulse reconstruction via machine learning (DeepFROG) has been demonstrated to outperform these iterative methods, especially at low signal to noise ratios [20]. All such methods though operate only on the typical off-axis FROG spectrogram. Separating this FROG term from the higher and background terms present in collinear FROG (cFROG) data is realized though either the noncollinear experimental geometry or by collecting the full collinear spectrogram and then numerically filtering. However, the noncollinear geometry is impossible to implement in the near-field, and the filtering step in cFROG is subjective, resulting in an unavoidable loss of information. To attack this problem, we develop a machine learning model to reconstruct the complex field from unfiltered cFROG data (DeepcFROG). Our method incorporates both the filtering and the reconstruction into the neural network and makes an error comparison directly back to the original cFROG spectrograms.

To address the spatial resolution problem, we demonstrate how apertured near-field scanning optical microscopy (NSOM) can be used to characterize such structures. Super-resolved techniques such as scattering scanning near-field optical microscopy (sSNOM) have previously been utilized to probe the near-field of nanostructures [21], [22]. However, apertureless sSNOM is highly susceptible to scattering interference, is challenging to perform spectrographic characterization, and enhances only the scattered z-polarized field. In comparison, apertured NSOM can alleviate these limitations.

Much like sSNOM, NSOM uses a resonant tip held in close proximity to a sample to probe the near-field. In our NSOM, the sample is excited in the far-field and the signal is collected through the tip in the near-field and sent back to a
detector. Using a subwavelength tip aperture, this ensures that only the evanescent field is coupled back [23]. Since this subwavelength resolution requires collection through the tip, cFROG is required to retrieve the temporal field. Our system consists of a homebuilt NSOM designed and constructed together with Prof. Hans Hallen of NC State and Dr. William Murray of Penn State. This system was used in conjunction with an ultrafast Ti-Sapphire laser to probe the nonlinear response of a monolayer transition metal dichalcogenide (TMD) 2D material with a spatial resolution approaching 100nm. Nonlinear signal collection is weak to begin with and this effect is compounded when probing the near-field. One solution is to increase the incident power, but this is limited by the burn threshold of the sample. Another is to increase the collection integration time, but this risks system instability such as laser drift or the tip falling out of feedback. To counter this, we create a numerical fitting algorithm to boost the signal to noise of NSOM raster scans. We also demonstrate that temporal characterization is possible with NSOM by collecting near-field cFROG spectrograms through the tip and reconstructing the corresponding complex field. This method can be extended to create scanning ultrafast near-field second harmonic optical microscopy (SUNSHOM). The synthesis of these techniques demonstrates at the minimum, a proof of concept tool for characterizing optically responsive nanomaterials and nanostructures at the nanometer-femtosecond scale.
2: Review of 2D Materials and Nanostructures

2.1 – Optical Properties of 2D Materials

Two-dimensional (2D) materials show extraordinary promise for their novel properties following the discovery and characterization of graphene, which was subsequently awarded the Nobel Prize [24]. The key characteristic of 2D materials is that their properties vary significantly as they are thinned down from bulk depositions to single layers [25]. For example, compare bulk amorphous carbon to graphene. Bulk sp² bonded graphite is weakly conductive and has limited applications in electronics and as optical components [26]. Graphene, however, exhibits strikingly different properties including a high thermal conductivity, the presence of edge states, and many others due to its ordered atomic structure which only becomes observable with examining single or few layer depositions [26]–[28]. Furthermore, it has been shown that the specific electrical, optical, and mechanical properties of 2D graphene can be broadly tuned with the addition of specific dopants and lead to novel properties such as a broadly tunable direct bandgap, high quantum efficiency, and many more [29], [30]. The properties of graphene can be even further modified by shaping monolayer sheets into structures such as carbon nanotubes and carbon cages (fullerene) via the insertion of multi-member carbon rings [31]–[35]. This same process, the thinning down of bulk materials into single and few layer depositions, has been applied to other materials. For example, bulk boron nitride has been thinned out into single layer hexagonal sheets with possible applications in high-power optical components due to its high thermal stability and behavior as a near prefect insulator [36], [37]. Similarly, 2D black phosphorus sheets display a tunable direct bandgap which can be controlled by modifying number of atomic layers [38].
In the field of optics, transition metal dichalcogenides (TMDs) are particularly interesting. TMD compounds are composed of a transition metal (tungsten, molybdenum, etc.) strongly covalently bonded with a chalcogenide (sulfur, selenium, tellurium, etc.) in a hexagonal structure of the form MX$_2$ [39]. Along the layered axis, these sheets are only weakly bonded together via Van der Walls forces which allows them to be easily separated into single and multilayered sheets grown through either or mechanical exfoliation or synthesized by chemical vapor deposition (CVD) [6], [39], [40]. In their bulk forms, these materials possess an indirect bandgap and have limited opto-electronic applications. For example, bulk MoS$_2$ is used as an additive in high wear lubricants and greases [41]. Like graphene though, these materials exhibit radically different properties when thinned down to single and few layer depositions, including a presence of a direct bandgap, and can be mechanically shaped into novel atomic scale structures such as sheets, rods, and cages [25], [42]. This transition to a direct band gap makes TMDs a prime candidate for applications requiring photoluminescent generation [25]. Furthermore, mechanical effects such as strain can influence the optical properties [42].
Fig. 2.1: Atomic structure of a TMD containing transition metal atoms (blue) and chalcogen atoms (silver). A) Side view of TMD monolayer. B) Top down view of TMD monolayer. C) Side view of TMD bilayer with symmetry point shown. Figure courtesy of Dr. William Murray.

A particularly interesting feature of TMDs is their high nonlinear susceptibility for single and odd layered depositions, exceeding traditional nonlinear mediums such as BBO and LiNO₃ by at least an order of magnitude due to the particular crystalline structure of the atomic deposition [5], [43]. Such a high nonlinear response makes TMDs a hot area of interest with potential applications in crafting ultrathin nonlinear optical components [44]. The high $\chi^{(2)}$ component in TMDs can be explored through density functional theory (DFT) of the band structure. In work described by Janisch et. al., the band structure of monolayer WS₂ features three regions with multiple conduction bands parallel to the valence band, leading to an increased joint density of states (JDOS) [5]. This increased JDOS produces an enhanced $\chi^{(2)}$ response.

2.2 - Engineered Nanostructures and Metamaterials

Beyond the field of 2D materials lies the field of metamaterials. Two-dimensional materials are fundamentally naturally occurring compounds, although selective engineering processes such as doping and mechanical rearrangement into structures such as nanorods significantly alter their electrical, mechanical, and optical characteristics [32], [36], [42]. Metamaterials, however, are a class of artificially constructed materials which exhibit properties and functionalities which cannot be realized in natural materials [45]. The power of metamaterials and similar optically responsive nanostructures is they exert control over the electromagnetic field such as the phase of a wavefront [9], [46], [47]. This is due to their spatial structure which consists of an engineered patterned structure usually consisting of resonant metallic or dielectric nanostructures [2], [46], [48]. Such structures can be fabricated to create negative index materials, ultrathin lenses and polarizers, broadband waveguides, and many more [10], [47], [49], [50].

Control over the electro phase requires engineered structures on the subwavelength scale. As a result, these structures are usually fabricated lithographically with resonant nanoantennae or nanorods, often gold, aluminum, or silver [46]–[48]. Other structures such as coupled split-ring resonators can be fabricated using a similar process [51]. Using deep ultraviolet lithography, it is possible to reliably fabricate deep subwavelength structures with specific dimensions below 50nm [51]. Furthermore, it has been shown that metamaterials can be formed by exploiting self-assembly of silver or gold functionalized nanoparticles with specific dimensions on the order of 5nm [52]. The process for designing, fabricating, and characterizing a new metasurface is usually:
1. Design structure based on desired optical parameters, element size, element shape, periodicity, etc.

2. Simulate optical response using finite difference time domain (FDTD), finite element methods (FEM), Greene's functions, or others [11], [53], [54]

3. Fabricate

4. Verify structure with AFM or electron microscopy [47]

5. Optically characterize

In the above sequence, the near-field optical response is simulated and the entire structure is often optically characterized in the far-field. Ideally, one would want to measure the near-field optical response as well as the far-field. While super-resolved methods such as scattering scanning near-field optical microscopy (sSNOM) have been employed in characterizing nanoscale systems below the diffraction limit, they are vulnerable to field distortion and enhance only the z-polarization state [21], [55].

In addition to their novel wavefront shaping capabilities, some metamaterials also exhibit a distinct spatio-temporal response. In Stockman et. al., the temporal phase modulation of an ultrafast excitation source was used to coherently control the spatial field distribution [11]. The conclusion of this is that the spatio-temporal response can actually be engineered to create novel pulse profiles at the nanometer-femtosecond scale. This includes not only wavefront shaping but also temporal compression and amplification [56]. It is already a challenge to measure the spatial near-field of such structures, let alone the corresponding temporal evolution at this scale. We will see later how temporal gating can be combined with NSOM to measure the electric field at the spatial and temporal scales required to fully understand such structures.

2.3 – Nanostructure Characterization Techniques

We have briefly described some of the novel properties of optically active nanomaterials, including 2D materials, TMDs, and metamaterials/metasurfaces. These structures exhibit novel optical properties at the nanoscale, both along the axial and transverse dimensions. Additionally, some nanoscale systems exhibit a novel temporal response along with spatio-temporal coupling. It should be evident that fully understanding the optical properties of these structures from a material researcher’s perspective requires an extensive experimental toolbox.

Along the axial dimension, the difference of a single atomic layer will result in dramatically different opto-electronic properties [5]. The current gold standard method for determining this is through atomic force microscopy (AFM). While AFM is an extremely reliable technique for determining the surface topography, it also has substantial limitations. AFM is a scanning technique which means the measurement time scales with the area to be measured. This makes AFM challenging for measuring large substrates. AFM also contains no optical information and can be potentially destructive as contact with the tip can damage the materials, especially when measuring fragile single atomic layered structures [57].

Measuring the optical characteristics of such materials are even more challenging. When characterizing these types of structures, it is rare to have a homogenous sample. Most samples contain inhomogeneous regions that are a mix of background and 2D material depositions of varying thicknesses. This makes it difficult to quickly measure samples using methods that do not simultaneously image. Refractive index determination is most commonly performed via ellipsometry which analyzes the polarization reflected from a substrate [8]. While this method is very accurate for determining the refractive index of a material deposition, much like
AFM, ellipsometry is not a single capture method. Imaging ellipsometry addresses this problem somewhat by reducing the scan dimensionality from 2D to 1D via imaging, but still requires the scan to complete before the complete final image is formed [58]. Because ellipsometry illuminates and collects light off normal incident, it is also not possible to use high NA objectives which further limits the system resolution.

Along the transverse dimension, there exists a different set of challenges. We have previously discussed the challenges that go along with measuring optically responsive nanomaterials, nanostructures, and engineered metasurfaces. For metasurfaces, it is obvious that subwavelength characterization is needed to fully probe the optical near-field. For nanomaterials, specifically TMDs, this is also the case. Monolayer TMDs can exhibit novel optical properties at the nanoscale due to defects, crystalline domain orientations, or the presence of edge states [59]–[61]. While methods such as electron microscopy yield information on the atomic structure along there lacks a method to directly optically characterize these structures on the necessary spatial scales. It should be evident by now that there is a critical need optically characterize these structures on the nanoscale as some of these material effects are still not well understood.

Fig. 2. 6: Second harmonic image of MoS₂ monolayers exhibiting optical edge effects. A) SHG excited at 1280nm. B) SHG excited at 1300nm. C) Line scans across both SHG images [59]. From X. Yin et al., “Edge nonlinear optics on a MoS₂ atomic monolayer,” Science (80-. ), vol. 344, no. 6183, pp. 488–490, May 2014, doi: 10.1126/science.1250564. Reprinted with permission from AAAS.
2.4 – Applied Numerical Methods in Nanomaterial Characterization

Even without microscopy, applied numerical methods are critical in understanding material behavior at the nanoscale. Such techniques can be applied prior to fabrication via simulations, during characterization, or after data acquisition. Simulation techniques such as finite difference time domain (FDTD), Fourier beam propagation, finite element methods (FEM), and others are critical in modeling nanostructures before fabrication [12], [13]. Because there are no experimental limitations at this step, such as working against the diffraction limit or loss of phase information when capturing with a CCD, it is possible to very accurately model the optical response of these structures on the deep subwavelength scale. Of course, a simulation model is only as good as the assumptions contained within it and experimental verification is always desirable. Post fabrication, applied numerical microscopy techniques such as wavefront shaping and compressive sensing can be used to extract additional information [62]. Following acquisition, techniques such as holography and frequency resolved optical gating (FROG) can be used via an algorithmic approach to recover either the spatial or temporal complex field [17], [63], [64]. In this work, we will mainly focus on techniques applied after data acquisition.

Most numerical methods follow an algorithmic approach: take data \((x,y)\), apply a specific analytical mapping or transfer function, and extract data \((x',y')\) at the end. In this approach, the transfer function must be carefully defined to achieve the desired outcome. For more complicated systems in which the transfer function is poorly defined or is nonanalytical, inversion techniques such as Monte Carlo methods can be used to recover information [65]–[67].

A specific subset of numerical methods utilizes machine learning to recover data. In deep learning, a large set of known inputs and outputs is used to train a
model while in recursive learning, the model is adapted continuously. This model is analogous to the transfer function which transforms a given input to an output. After training with a sufficient amount of data, the model can take in a never before seen input and determine the corresponding output. Machine learning can broadly be separated into two categories: classification and regression. Classification involves training a model to separate inputs into a certain number of classes. This can either be 1/0 in binary classification, or into several pre-defined groups. As a result, classification based machine learning has direct applications in pattern recognition methods such as speech recognition, early cancer detection, antibiotic susceptibility testing, and many others [68]–[71]. Regression based machine learning is an even more powerful tool by enabling full inversion analogous to constructing a nonanalytical transfer function given a set of inputs and outputs. When applied to optical characterization, such a method is especially relevant for situations where the transfer function of a system is either highly nonlinear or poorly defined. This approach has been used for optical ranging, analyzing the response of waveguides, error mitigation in communication systems, and dehazing of noisy images. [72]–[75]. Theoretically, machine learning can be used to replace any such technique, given a sufficient number of known inputs and outputs. However, the challenge in machine learning is that the model is only as good as the data used for training it. As a result, it is often challenging to construct a generalized model.
3: Quantitative Holographic Imaging of 2D Materials

3.1 – Current Techniques for Characterizing 2D Materials

Of particular interest are a subset of 2D materials known as transition metal dichalcogenides (TMDs) which, as previously discussed, display interesting optical properties making them prime candidates for opto-electronic applications including field effect transistors, photovoltaic cells, second harmonic generation, and many more [1], [7], [25], [29]. Monolayers in particular have been shown to possess a large $\chi^{(2)}$ [5].

During synthesis of these materials, it is rare to see a substrate consisting of only monolayers. Typically, a sample is an agglomeration of monolayers, bilayers, trilayers, and thicker atomic layer depositions. There is a need for scientists in the materials field to quickly differentiate between these different atomic layer depositions. The two most commonly used methods for this task are atomic force microscopy (AFM) and ellipsometry. While both of these techniques have their strengths, they also have several limitations.

Atomic force microscopy is a well-documented technique and is the leading gold standard method for quantifying the surface topography of a sample as it is capable of reliably determining surface topographic features down to a few nm [76]. In AFM, a resonating probe is scanned over a substrate. Specifically, a cantilevered tip is oscillated above with a slight bumping of the sample, close enough that forces such as Van Der Walls force, electrostatics, physical contact, or others cause deflection of the tip according to Hooke’s law [77], [78]. The exact method of deflection is dependent on the specific type of AFM used and the substrate being probed. While
AFM is considered the gold standard for surface profile measurement, it has its limitations.

One obvious limitation of AFM is that it is not an optical measurement technique, though for many samples of interest this is a non-issue. But because 2D materials exhibit such novel optical properties, it is desirable to have a single technique simultaneously characterize both the layer thickness and the optical properties. Also, AFM is a scanning technique as opposed to single capture. As such, the measurement time scales with area of the sample being measured. AFM can also be a relatively invasive technique. Even in non-contact oscillation modes, the interaction force between the tip and the substrate can be enough to damage the sample, especially when examining materials as fragile as single atomic layer depositions [57]. While AFM is the tool of choice for many samples in different fields, it leaves much to be desired when characterizing 2D materials.

Other techniques such as photoluminescence and Raman spectroscopy can be used to determine the layer thickness deposition [60], [79], [80]. For example, photoluminescence can be used to distinguish few layer TMD depositions across a broad substrate [25], [81]. Similarly, the Raman spectrum of a single point can be matched to predict the layer thickness [42]. Both of these methods though require thorough calibration of the system and fail for thicker layer depositions as their optical properties approach those of the bulk counterparts.

A technique commonly used to characterize the optical properties of 2D materials is ellipsometry [82], [83]. In ellipsometry, light of an engineered polarization state is directed onto a substrate of interest at a particular incident angle and the reflected light is collected. The collected light is analyzed by comparing the relative s and p components of the polarization ellipse. Ellipsometry has the advantage that it can precisely determine the refractive index of a substrate, but is still an imperfect
technique for charactering 2D materials. Because ellipsometry works by detecting relative chances in the polarization state of the reflected light, it is very sensitive to any scattering events or surface contaminants. A relatively uneven or dirty substrate will yield unreliable results. Another refractive index characterization method is Kramers-Kronig analysis which uses information from the reflection or transmission off a substrate, but requires measurement over a broad spectrum, which is not always possible [84].

The limitations of such techniques are particularly evident when characterizing optical nanomaterials such as TMDs. An ideal tool for probing these particular materials is one that would retain the topographic measurement capabilities of AFM and the refractive index characterization of ellipsometry while also being non-invasive and suitable for measuring large substrates. A practical tool would also permit simultaneous imaging as these samples are rarely homogenous.
3.2 - Formalism of Digital Holography

Digital holography has found numerous applications in fields ranging from biological science, nondestructive testing, advanced metrology, and many more [63], [85]–[87]. The strength of holography over simple optical microscopy is that it enables the retrieval of the full complex field. In simple microscopy, only one incident beam is used to measure a sample, thereby restricting measurement to only intensity. However, in digital holography, a second beam is interfered with the primary beam, for example, in a Michelson interferometer geometry. The interference with this second beam encodes the phase information.

One can imagine a set of two optical beams, the reference beam $R(x, y)$ with a known field profile, and the signal beam $S(x, y)$ with an unknown profile. Both beams contain an amplitude and phase profile along the transverse dimensions and can be defined in (1) and (2) where $\vec{k}_R$ and $\vec{k}_S$ are the wave propagation vectors:

$$R(x, y) = A_R(x, y)e^{i(\vec{k}_R \cdot \vec{r} - \omega t)}$$

$$S(x, y) = A_S(x, y)e^{i(\vec{k}_S \cdot \vec{r} - \omega t)}$$

If the beams are overlapped at a fixed plane in space $z = z_0$, and at a fixed time $t = 0$, the resulting field intensity profile becomes:

$$I_H(x, y, z = z_0) = |R(x, y, z = z_0) + S(x, y, z = z_0)|$$

$$I_H(x, y, z = z_0) = |R|^2 + |S|^2 + R^*S + RS^*$$

In (4), the intensity distribution is the sum of 4 distinct terms: the intensity of the reference arm $|R|^2$, the intensity of the signal arm $|S|^2$, and the interference of the signal arm with the reference arm $R^*S + RS^*$. It is this last term that encodes the
spatial phase distribution of the signal field to be measured. Expanding the interference term:

\[
R^*S + RS^* = 2A_R(x,y,z_0)A_S(x,y,z_0) \cos \left( (\vec{k}_S - \vec{k}_R) \cdot \vec{r} + (\phi_S(x,y,z_0) - \phi_R(x,y,z_0)) \right)
\] (5)

In (5), the interference pattern between the sample and reference arm is now encoded by the carrier term. This manifests itself as a periodic interference pattern when captured on a digital camera or similar medium.

Fig. 3.1: Holographic CCD capture of a sample containing multiple TMD flakes

To reconstruct the signal beam off the sample, the field of the reference arm must be known. In practice, this is often designed to be a planar wave with a constant amplitude profile and a linear phase profile. However, different reference fields exist including spherical, cylindrical, or custom wavefronts. While these geometries have special niche applications in optics, they are beyond the scope of this dissertation. For the sake of mathematical simplicity, we will assume that the
reference wavefront is planar, the amplitude is unity, and \( z_0 = 0 \) which reduces our expression for the intensity distribution to:

\[
I_H(x, y) = 1 + |S(x, y, 0)|^2 + A_S(x, y, 0) \cos \left( (\vec{k}_S - \vec{k}_R) \cdot \vec{r} + \phi_S(x, y, 0) \right)
\]  

(6)

Retrieving the term of interest from the above equation is now a simple matter of Fourier transforming the above equation and numerically filtering out one of the side bands. The CCD capture from Fig. 3.1 is Fourier transformed in Fig. 3.2, clearly showing the DC terms in the center and the two sidebands, each one containing the full complex field or its complex conjugate.

![Fourier transform of the CCD capture showing the DC term, two sidebands, and filter.](image)
In Fig. 3.2, it is important to understand the separation between the DC terms and the sidebands and the relation to filter size. Too small of a filter will result in information will be lost. Too large of a filter will introduce bleed through from the DC. For this reason, it is critical that the DC terms be sufficiently separate from the sidebands. Experimentally, this is performed by creating a slight angle between the two beams at the intersection point on the sample. This angular offset determines the spacing of the interferometric fringes which directly impacts the filtering and the quality of the resultant reconstruction. The holographic grating pitch can be determined as a function of the angular separation between the two incident beams and the excitation wavelength.

![Momentum space diagram](image)

Fig. 3.3: Momentum space diagram for determining the holographic grating pitch from the sample and reference beams

In Fig. 3.3, the wave vectors for the signal and reference arms are defined as 

\[ |\vec{k}_R| = |\vec{k}_S| = \frac{2\pi}{\lambda_0} \]

under the assumption of free space in both the signal and reference arms. The grating vector is denoted as 

\[ |\vec{G}| = \frac{2\pi}{|\Lambda|} \]

where \(|\Lambda|\) is the spatial width of the
grating pitch and $\theta$ denotes the angular spatial separation between the signal and reference beams. Using geometry to solve for the grating pitch, we arrive at:

$$|\Lambda| = \frac{\lambda_0}{2 \sin \left(\frac{\theta}{2}\right)}$$  \hspace{1cm} (7)
3.3 - Numerical Modeling of TMD and Substrate

By measuring the complex field, holographic microscopy is a powerful tool for quantifying the surface topography of a sample and has numerous applications in metrology [85], [88], [89]. By using numerical modeling and applying known parameters for the sample and substrate, we further extend the capabilities of holographic imaging to characterize 2D materials.

The TMDs we focus on characterizing are molybdenum disulfide (MoS$_2$) and tungsten disulfide (WS$_2$). When these materials are fabricated, they are deposited onto a commercially available substrate consisting of a thick (>1mm) silicon substrate with a thin 300nm SiO$_2$ layer. The differences in refractive indices between the Si, SiO$_2$, and TMD deposition create a reflective multi-layer structure. Therefore, when measuring the phase and amplitude shift of the reflected light off of a TMD deposition, one has to not only consider the thickness and index of the TMD, but must model the entire stack. Later, we will show how this effect can be exploited to ascertain additional information about material. The multi-layer model is shown in Fig. 3.4 (obviously not to scale.)
A₁ – A₄ denote the forward propagating fields while B₁ – B₄ denote the backwards propagating fields. The silicon substrate is very thick compared to the 2D/SiO₂ layer and has a non-negligible extinction coefficient over the wavelengths of interest to us. For this reason, we treat the Si layer as semi-infinite and assume no backward propagating wave. Using propagation matrices and the corresponding transmission and reflection coefficients for each interface in the stack, we arrive at an expression for the forward and backward propagating fields:

\[
\begin{pmatrix}
A_1 \\
B_1
\end{pmatrix} = \frac{1}{t_{12}} \begin{pmatrix}
1 & -r_{21} \\
r_{12} & 1
\end{pmatrix} \begin{pmatrix}
e^{-ik_2d_2} & 0 \\
0 & e^{ik_2d_2}
\end{pmatrix} \frac{1}{t_{23}} \begin{pmatrix}
1 & -r_{32} \\
r_{23} & 1
\end{pmatrix} \begin{pmatrix}
e^{-ik_3d_3} & 0 \\
0 & e^{ik_3d_3}
\end{pmatrix} \frac{1}{t_{34}} \begin{pmatrix}
1 & -r_{43} \\
r_{34} & 1
\end{pmatrix} \begin{pmatrix}
A_4 \\
0
\end{pmatrix}
\]  

(8)

We then define the complex reflection coefficient for the entire cavity as \( \bar{r} = \frac{B_1}{A_1} \) and can reduce (8) to (9):

\[
\begin{pmatrix}
A_1 \\
B_1
\end{pmatrix} = \frac{1}{t_{12}} \begin{pmatrix}
1 & -r_{21} \\
r_{12} & 1
\end{pmatrix} \begin{pmatrix}
e^{-ik_2d_2} & 0 \\
0 & e^{ik_2d_2}
\end{pmatrix} \frac{1}{t_{23}} \begin{pmatrix}
1 & -r_{32} \\
r_{23} & 1
\end{pmatrix} \begin{pmatrix}
e^{-ik_3d_3} & 0 \\
0 & e^{ik_3d_3}
\end{pmatrix} \frac{1}{t_{34}} \begin{pmatrix}
1 & -r_{43} \\
r_{34} & 1
\end{pmatrix} \begin{pmatrix}
A_4 \\
0
\end{pmatrix}
\]  

(9)
\[ \tilde{r} = \frac{r_{12}(1 - r_{32}r_{34}e^{2ik_2d_3}) + e^{2ik_2d_2}(r_{23} + r_{34}e^{2ik_3d_3})}{1 - r_{32}r_{34}e^{2ik_3d_3} - r_{21}e^{2ik_2d_2}(r_{23} + r_{34}e^{2ik_3d_3})} \]  

(9)

In (9), \( k_2 = \frac{2\pi n_2 d}{\lambda_0} \) represents the wave vector containing the 2D material layer.

For a bare substrate, \( k_2 = k_0 \) while \( d_2 \) remains constant to account for the air layer.

We then model the reflection coefficient of the entire stack by taking the ratio of the reflection coefficients where the 2D material is present against the portion where the 2D material is not present. We assume normal incidence and monochromatic illumination from the laser source and a uniform layer of SiO\(_2\). For verification, subsequent ellipsometry showed a relatively constant SiO\(_2\) layer of 299nm ± 2nm.

The expected phase and amplitude response are then defined as:

\[ A = \left| \frac{\tilde{r}_{2d}}{\tilde{r}_{\text{back}}} \right| \]  

(10)

\[ \phi = \arctan \left( \frac{\tilde{r}_{2d}}{\tilde{r}_{\text{back}}} \right) \]  

(11)

We then numerically model the sample as a function of both wavelength and TMD layer thickness. For relatively thick TMD depositions of MoS\(_2\), we set the refractive index to the bulk value which is well documented in the literature [84].

When simulating the behavior of the sample, we observe that the illumination wavelength is critical in determining the phase and amplitude response of the structure. This is due to two factors. One is that the effective cavity length scales with the wavelength. The other is that the refractive indices for the substrate layers also have a rather strong dependence on the wavelength used [82], [83], [91]. Both of these factors significantly alter the resonant behavior of the cavity. The conclusion is that by choosing an optimal wavelength for the illumination source, it is possible to exploit multi-layer interference within the cavity and maximize the respective
phase and amplitude shifts for a given range of TMD (or other 2D material) thicknesses.

Fig. 3.5: Expected amplitude (A) and phase response (B) for a MoS$_2$/Si/SiO$_2$ stack vs. 2D material layer thickness at various wavelengths [90]. A. Ghosh, J. Noble, A. Sebastian, S. Das, and Z. Liu, “Digital holography for non-invasive quantitative imaging of two-dimensional materials,” J. Appl. Phys., vol. 127, no. 8, p. 084901, Feb. 2020, doi: 10.1063/1.5128135. Reproduced with the permission of AIP Publishing.
3.3 - Holographic Imaging of TMDs

We then construct a digital holographic microscope in the Michelson interferometer configuration shown in Fig. 3.6 to perform the 2D material characterization. The laser used for the illumination source is centered at 672nm with a coherence length of approximately 0.2mm. We chose this specific laser for two distinct reasons. One is that at the chosen wavelength, we expect a strong resonant behavior from the cavity for detecting TMD depositions consisting only of a few layers. The other is the coherence length of the laser. When performing holography, it is desirable to have a coherence length much longer than the structure size being studied, but shorter than the separation between the optical elements of the system. This is shorter than both the distance between separate optical elements and the distance between optical element surfaces [92]. For common elements such as lenses and beam splitters, this corresponds to a coherence length shorter that approximately 2mm. Using an illumination source with a coherence length longer than these distances will create interference between the optical elements and will manifest itself as coherent artifacts on the CCD capture [92]. After the laser, a single mode fiber is used to spatially filter the beam before entering the Michelson interferometer.

In addition to laser choice, two-inch optical elements (lenses, mirrors, and beam splitters) are utilized to minimize wavefront curvature and aberrations within the two interferometric arms. To create the angular separation between the signal and reference beams, we employ a 4f imaging in the reference arm. This approach enables controlled wavefront tilting in the reference arm with minimum spatial deviation of the reference beam on the CCD.
A strength of holography is its ability to image a very weak source off the signal arm as the noise scales with the ratio of $|\frac{R}{S}|$. The measurement sensitivity of a digital holographic microscope is dependent on the time stability of the system between the signal and reference arms. When capturing the hologram, we use the shortest possible integration time permitted by the CCD and cover the entire optical set up with a thick Styrofoam enclosure to limit effects such as air currents and thermal drift. More extensive interferometric stability measures include placing the interferometer in a vacuum, spring or magnetic vibration damping, and post-capture numerical compensation [93], [94]. While these techniques are necessary for very high precision interferometry, such as detecting gravitational waves, they are unnecessary and cost-prohibitive for our application. We found that enclosing the system to eliminate air currents and reduce thermal drift was sufficient to obtain the phase resolution necessary for detecting monolayers.
The main objective used for is a 0.55 NA long working distance objective. Given our illumination wavelength of 672nm, this yields a theoretical maximum lateral resolution of \( r = \frac{\lambda_0}{2N_A} = 611\text{nm} \). To record the digital hologram at the plane of intersection we use an Apogee AP32ME charged coupled device (CCD) interfaced with a computer. The detector of this CCD has a resolution of 2184x1472 with a pixel size of 6.8μm. After collection, the raw CCD capture is processed and separated into the amplitude and phase images.

![CCD capture](image1)


It is evident in Fig. 3.7 that our microscope is able to detect TMD monolayers, and, by extension, thicker atomic layer depositions. However, the amplitude and phase reconstructions must both be flattened first to obtain an average value for the
phase and amplitude across the sample. Even using two-inch optics, the wavefront is still curved and needs to be corrected. The full expression for the signal and reference beams with wavefront curvature and ignoring higher order aberrations can be expressed as:

\[ R(x, y) = A_R(x, y) e^{-\frac{(x-x_r)^2+(y-y_r)^2}{2\sigma_r^2}} e^{i\phi_R(x, y)} e^{i\alpha(x-x_r)^2+(y-y_r)^2} \]  
\[ S(x, y) = A_S(x, y) e^{-\frac{(x-x_s)^2+(y-y_s)^2}{2\sigma_s^2}} e^{i\phi_S(x, y)} e^{i\beta((x-x_s)^2+(y-y_s)^2)} \]

It is possible to numerically correct for this curvature by fitting and subtracting a radial Gaussian function from the amplitude and a quadratic function from the phase. However, we achieved better results by translating the sample to an area containing only substrate and no 2D material, capturing a reference hologram, and subsequently using its amplitude and phase reconstruction to flatten the holographic reconstructions of interest. Fig. 3.8 shows the holographic amplitude and phase reconstructions of a monolayer. Taking a line scan across these images reveals the Gaussian and spherical background profiles.
Fig. 3.8: A) Phase reconstruction of a TMD flake showing phase wrapping from wavefront curvature. B) Line scan taken across (A). C) Amplitude reconstruction of a TMD flake showing a Gaussian envelope. D) Line scan taken across (C).
3.4 - Complex Refractive Index Prediction of TMDs

We demonstrated that our holographic microscope resolves single layer atomic depositions of TMDs. Using our prior numerical model for the sample, we further extend the capabilities of this tool to predict the complex refractive index of TMDs. In the world of 2D materials, this is not a trivial task. The complex index for bulk depositions of TMDs such as WS\(_2\) and MoS\(_2\) are well defined, however this is not the case for monolayers [8], [83], [95]. The optical properties of TMDs change significantly as they are thinned down from bulk depositions to trilayers, bilayers, and monolayers. A common method for refractive index measurement is ellipsometry. For single layer depositions though, this is a challenge which yields varying results. In Liu, et. al., they measured the optical properties of MoS\(_2\) using ellipsometry and recorded significant variations in the measured values, even from sample to sample [8]. Using our holographic microscope, we create a new method for quantifying the refractive index of 2D materials.

To quantify the complex refractive index \(\tilde{n} = n + ik\), we find the minimum of the mean squared error (MSE) function between the averaged measured values for the amplitude and phase responses, and the values predicted by our numerical model. As mentioned previously, there are three unknowns in our model, one from the TMD layer thickness and two from the refractive index as it is complex valued. For monolayers however, the thickness is well documented [39], [96], [97]. This constrains the number of unknowns to only two, and can now be solved as the holographic reconstruction yields two unique pieces of information contained separately in the amplitude and phase.

After reconstructing a hologram containing a 2D material, we obtain a value for the measured amplitude and phase \((A_m \text{ and } \phi_m)\) by flattening the reconstructions
and averaging over a smooth region of the sample. We then generate a map of the 
predicted amplitude and phase values \( A_p(n, \kappa) \) and \( \phi_p(n, \kappa) \) over a range of values for 
\( n \) and \( \kappa \). In Fig. 3.9, we plot the MSE function between the predicted and measured 
values.

\[
\delta_A(n, \kappa) = (A_p(n, \kappa) - A_m)^2
\]

(14)

\[
\delta_\phi(n, \kappa) = (\phi_p(n, \kappa) - \phi_m)^2
\]

(15)

Fig. 3.9: Mean squared error plots between the predicted and measured values. A) Amplitude error. B) Phase error. C) 
Cumulative error.

We notice the error plots in Fig. 3.9 are smooth with only a global minimum 
and no local minima. It follows then that the best fit for the refractive index is located 
at the global minimum of this error function. In (16) \( \alpha \) and \( \beta \) are normalized weighing 
constants between the amplitude and phase components defined as 
\( \alpha = \frac{1}{A_m^2} \) and \( \beta = \frac{1}{\phi_m^2} \). In practice, applying normalized weighting constants places equal importance on 
the both the phase and the amplitude measurements in the fitting algorithm. 
However, this may be modified in certain situations near resonance where either the 
phase or amplitude response can be expected to have a much larger relative shift and 
can be weighted stronger.
\[(n_{\text{best}}, \kappa_{\text{best}}) = \text{Argmin}_{n, \kappa} \left[ \alpha (A_p(n, \kappa) - A_m)^2 + \beta (\phi_p(n, \kappa) - \phi_m)^2 \right] \] (16)

Using this approach, we fit the refractive index and extinction coefficient to a WS\textsubscript{2} monolayer deposited on the Si/SiO\textsubscript{2} substrate. The triangular shape of the material deposition is the key factor that identifies this structure as a monolayer along with verification via photoluminescence. From the samples tested, we measure a mean amplitude response of 0.72 ± 0.08 and a mean phase shift of 0.21 ± 0.03 rad. Using these values, we predict a complex refractive index of \(n = (5.21 \pm 0.31) + (1.12i \pm 0.30i)\) for monolayer WS\textsubscript{2} at a wavelength of 672nm. These results are in agreement with values reported in the literature obtained with different techniques [7], [8], [83], [95].

Fig. 3.10: Reported refractive indices and extinction coefficients for monolayer WS\textsubscript{2} and MoS\textsubscript{2} [8], [83], [95], [98]. Figure courtesy of Dr. William Murray.
To visually examine the refractive index fitting, we plot the best fit for the refractive index over one standard deviation from the mean measurement value across several measured samples. As shown in Fig. 3.11, the red dots cluster closely together on both the amplitude and phase error plots. This supports the robustness and reliability of our technique. This technique can be extended to other TMDs which have well established layer thicknesses.

3.5 - Layer Deposition Prediction of TMDs

For monolayer depositions of TMDs, the refractive index is difficult to quantify while the thickness is well known. The inverse is true when examining thicker multilayer depositions. As the TMD layer thickness is increased, their optical properties quickly take on the characteristics of their bulk counterparts. In this regime, the refractive indices and optical properties are well explored in the literature, however, quantifying the exact thickness of the deposition becomes challenging. For example, differentiating between a 48 and 50-layer stack is very challenging as their optical responses will be almost identical. The tried and true method for this is AFM. While AFM is very reliable and accurate, it is time consuming when attempting to measure large samples. Using holography to measure these structures is faster, non-invasive and simultaneously images. These attributes make holographic imaging particularly relevant when characterizing large, non-uniform, and sensitive substrates. Through our approach, holographic microscopy coupled with numerical modeling can resolve TMD flakes with an accuracy comparable to AFM.

To demonstrate this technique, we analyze a substrate containing several relatively thick MoS$_2$ depositions with expected thicknesses of several tens of layers. Using our holographic microscope at 672nm incident light, we capture the hologram of several thick TMD depositions and subsequently retrieve the amplitude and phase responses.
We then apply our numerical fitting model in a similar fashion used to calculate the refractive index. In this situation, we fix the refractive index at the bulk value of $5.43 + 1.18i$ while letting the thickness vary in our numerical model. To predict the number of layers, we calculate the minimum of the MSE function between the measured amplitude and phase response and the predicted amplitude and phase response over a range of thicknesses. Using this technique, we predict an average thickness of $51 \pm 2$nm. Assuming each atomic layer has a uniform thickness of 0.7nm, and that the thickness scales linearly with the number of layers, we estimate this particular flake contains $78 \pm 3$ distinct atomic layers.

To verify our results, we compare against an AFM measurement of the same sample. The AFM field of view is much smaller than the holographic microscope as AFM is a scanning technique and would require a long time to scan the entire field. After locating the region of interest, we measure the surface topography of the larger flake. A line scan across the center of this flake is averaged to arrive at an estimate for the height of the flake. We find this flake has an average thickness of $56.2 \pm 0.6$nm compared to $51 \pm 2$nm obtained from the holographic microscope.
To further explore this technique, we repeat the procedure using a different sample containing thinner depositions of WS$_2$ instead of MoS$_2$. For WS$_2$, the refractive index is fixed at the bulk value of $4.1 + 0.05i$ and we determine this flake is $14 \pm 1$nm or $20 \pm 1$ layers thick. This compares well to the AFM characterization yielding an average thickness of $15.0 \pm 0.2$nm.


To thoroughly quantify the thickness comparison between holographic microscopy and AFM, we perform a Blant-Altman analysis using several samples consisting of thick WS$_2$ and MoS$_2$ flakes. In this analysis, we plot the predicted flake thickness from our holographic microscope results fitted with our numerical model against the values measured with AFM. In a hypothetical comparison where all points lie on the equivalency line, both techniques would yield identical results. In our situation, the holographic approach predicts the accuracy of any given flake within 10% of the values obtained from AFM. Over a set of measurements for flakes in the range of 15 – 50nm, the holographic microscope predicts the thickness of the flake to be 0.12 ± 2.42nm less than the AFM measurement. When examining the 95% confidence interval, we conclude with 95% certainty, that the holographic microscope will measure any given TMD flake to within about 8nm of values obtained through AFM. In this experiment though, all values obtained through holography are within 4nm of those obtained through AFM.

We have demonstrated a technique for the efficient, non-invasive, and accurate characterization of 2D materials by combining holographic microscopy with numerical modelling. Holography can be used to predict the complex refractive index of monolayer atomic depositions, similar to ellipsometry, while simultaneously imaging the substrate. For thicker flakes, holography can quickly predict the layer thickness with an accuracy comparable to the gold-standard AFM. Furthermore, this approach is suitable for delicate samples and its characterization time is fast due to single-shot holographic capture. While there will always be conditions where AFM and/or ellipsometry will be preferred, we have demonstrated that holographic microscopy in conjunction with numerical modeling can be a powerful asset in the researcher’s toolbox for studying 2D materials.
4: Ultrafast Optical Pulse Characterization

4.1 – Frequency Resolved Optical Gating (FROG)

Many nanomaterials and metamaterials also exhibit a temporal response occurring at the femtosecond scale [11]. Recovering this temporal information is accomplished through techniques such as autocorrelation, FROG, ptychography, and others [18], [99]–[101]. Simple intensity autocorrelation measurements quantify the pulse intensity envelope while others, such as frequency resolved optical gating (FROG), can recover the full complex field.

FROG was initially developed to recover the temporal structure of ultrafast lasers [17], [101]. The underlying challenge is that to measure a short duration event, you would often need an instrument that responds even faster. This is overcome by performing temporal pulse interference wherein we gate the pulse in question with itself, essentially performing “holography” in the time domain. This procedure encodes not only the intensity profile of the field, but the phase information as well. Characterizing the complex field requires two degrees of information, either the absolute value and phase or the real and imaginary components. Experimentally then, it is necessary to also provide two degrees of information. All such methods utilize spectrographic information for one degree of the information. The second degree of information is obtained via time gating wherein FROG incorporates a delay stage while approaches such as GRENOUILLE use a prism to introduce a relative delay between intersecting beams [18].

First let’s consider an ultrashort pulse whose electric field is described in the time domain as $\mathcal{E}(t)$ as the product of a slowly varying envelope $E(t)$ with the associated carrier $e^{i\omega t}$. This original pulse is separated into two pulses in a Michelson
interferometer, one of which is delayed relative to the other. After being delayed, this new pulse is represented as \( \mathcal{E}(t - \tau) = E(t - \tau)e^{i\omega(t - \tau)} \) where \( \tau \) is a delay in time resulting from the extra path length relative to the other pulse. These two pulses are then recombined and focused onto a nonlinear crystal with a second order nonlinear susceptibility for second harmonic generation, generating the term in (17). The spectrum of the SHG signal as a function of delay is then recorded on a spectrometer (18). In the off-axis FROG implementation, the two beams do not propagate collinearly and are overlapped only at the location of the nonlinear medium [102]. In both geometries, the particular term of interest is given by:

\[
\mathcal{E}_{\text{sig}}(t, \tau) \propto E(t)E(t - \tau) \tag{17}
\]

\[
I_{\text{sig}}(\omega, \tau) = \left| \mathcal{F}_t\{\mathcal{E}_{\text{sig}}(t, \tau)\} \right|^2 \tag{18}
\]

By translating the delay arm, we construct a temporal spectrogram pattern as a function of pulse delay. This spectrogram is used later to recover the original pulse profile. FROG can be realized with either with \( \chi^{(2)} \) or \( \chi^{(3)} \), although SHG is by far the most common [101]. Within the scope of this work, we will only consider SHG FROG [101]. It has been shown that the FROG spectrogram is a unique mapping to a complex field, minus the so called trivial ambiguities including time shift, time-reversal, and constant phase offset [102].
4.2 - Iterative FROG Pulse Retrieval

There are several approaches for ultrafast pulse reconstruction such as FROG, GRENOUILLE, and ptychography [18], [19], [64]. The difference in these approaches lies in the specific geometries used for data acquisition and the subsequent retrieval algorithms. We focus on the principal components generalized projections iterative phase retrieval algorithm (PCGPA) for its robustness and prevalence in the field of ultrafast pulse retrieval [64], [103]. In our reconstructions, both for simulated pulses and experimental acquisitions, we base our numerical model off the FROG package by Adam Wyatt, available via the MathWorks central file exchange [104].

In the implementation of the algorithm with experimental data, the specific sampling details can be strict and require careful attention when processing experimental data. The algorithm starts with an initial guess for the complex envelope of the pulse. This is typically chosen as a Gaussian or hyperbolic secant amplitude profile with a flat phase response. Using this initial guess, we construct the associated signal $\mathcal{E}_{\text{sig}}(t, \tau)$.

![Initial guess for the pulse shape and the corresponding simulated signal, $\mathcal{E}_{\text{sig}}(t, \tau)$.

Fig. 4.1: Initial guess for the pulse shape and the corresponding simulated signal, $\mathcal{E}_{\text{sig}}(t, \tau)$.}
We then Fourier transform this signal and apply the amplitude constraint of the retrieval algorithm. The amplitude constraint states that the amplitude of the guess must be equal to the amplitude of the experimental capture. This updated spectrogram now contains the amplitude of the experimentally measured FROG spectrogram with the phase resulting from the initial pulse guess. We then inverse Fourier transform the spectrogram back into the time domain and shift it to recover \( E(t)E(\tau) \). Singular value decomposition (SVD) is then used to recover the field. Only the first singular value is used since all information comprising \( E(t)E(\tau) \) is contained in one vector, \( E(t) \). This updated pulse is taken as the new guess and the algorithm is iterated several times until the error between the guessed FROG spectrogram and the experimentally measured FROG spectrogram converges to a minimum.

Fig. 4.2: Schematic of the principal components generalized projections FROG pulse retrieval algorithm (PCGPA) [103].

To quantify the reconstruction, we calculate the error between the reconstructed and ground truth FROG spectrograms. The weighted mean squared error (WMSE) is a good metric for these comparisons as it is independent of the array size and also works for sparse arrays.
\[
\sum_{i,j} \left( l_{\text{true}}(\omega_i, \tau_j) - l_{\text{pred}}(\omega_i, \tau_j) \right)^2

\]

(19)

In this situation, we reconstruct the FROG of a simulated pulse to get a baseline error for the reconstruction. In these first simulations, we assume no noise, both on the pulse and on the FROG spectrogram. Our simulation results show the WMSE between the reconstructed FROG spectrogram and the simulated ground truth drops to 0.13% after 1000 iterations. With more iterations, the WMSE asymptotically approaches zero and reconstructs the field perfectly for a noise free pulse with exception of the trivial ambiguities including time-reversal, time shift, and constant phase shift which cannot be mathematically resolved [17], [101]. This is true for asymmetric and more complex pulse profiles as well.

![Fig. 4.3](image)

Fig. 4.3: Original (simulated) and reconstructed FROG spectrograms with a weighted MSE of 0.13% after processing for 1000 iterations.

The PCGPA for reconstructing ultrafast optical pulses is a powerful tool and is considered the gold standard for characterizing the complex field for femtosecond optical sources. Other papers show it is able to reconstruct in real time and with high noise tolerance given sufficient signal and computing resources [103], [105]. All of these results however, are predicated on one critical assumption, the experimentally collected spectrogram passed into the reconstruction algorithm contains only the nonlinear mixing term \( E(t)E(t - \tau) \).
4.3 - Collinear vs. Noncollinear FROG

There are two experimental geometries that can be used to collect FROG data for ultrafast pulse reconstruction. These are typically referred to as collinear or noncollinear. Both perform the same task, but have certain advantages and disadvantages. Both methods involve splitting an ultrashort pulse into two copies with a variable time delay between them. The two pulses are then focused onto a nonlinear medium such as BBO or LiNbO₃ in the case of 𝜒⁽²⁾ FROG. This nonlinear mixing generates the FROG term, i.e. \( E(t)E(t - \tau) \) in the case of SHG FROG, which is used in the iterative reconstruction. For collinear FROG (cFROG) though, this nonlinear mixing creates additional terms. For second harmonic time gating, this can be expressed as:

\[
\mathcal{E}_{slg}(t, \tau) \propto \mathcal{E}(t)^2 + \mathcal{E}(t - \tau)^2 + 2\mathcal{E}(t)\mathcal{E}(t - \tau)
\]  

To recover the off-axis, or noncollinear FROG term \( E(t)E(t - \tau) \), it must be separated from the other terms. This can be performed by either spatially separating the beams or in numerical post processing. In noncollinear FROG, the two beams are spatially oriented such that they intersect only at the nonlinear medium and the phase matching condition dictates the off-axis FROG term can be spatially separated. In this noncollinear geometry, only the FROG term is recorded by the spectrometer. However, this approach can be difficult or even impossible to experimentally implement. This is because there must be sufficient space to spatially separate the FROG term. More importantly, such spatial separation becomes impossible in near-field measurements involving the use of a nanoscale tip. For these reasons, the noncollinear geometry typically only used to characterize strong pulses directly from their sources, such as ultrafast lasers.
In the collinear FROG (cFROG) acquisition geometry, the two beams propagate collinear to each other and are not spatially separated. After second harmonic generation, an optical component such as a filter in combination with a dichroic mirror separates the fundamental beams from the SHG signal. When the two interferometer arms are perfectly time matched, the SHG signal power will be 8x stronger than when the two pulses are completely separated in time. This approach is similar to an interferometric autocorrelation measurement, except uses a spectrometer to resolve the spectrum instead of a photo-diode or power detector. Because both beams propagate over each other, this geometry can be implemented in many more experimental situations. It is also easy to simultaneously image the nonlinear substrate during acquisition. We will show later why this geometry is necessary for characterizing nanostructures.
Fig. 4.5: Collinear FROG acquisition geometry.
4.4 – cFROG Filtering

After collecting collinear FROG (cFROG) data, it is necessary to numerically filter the spectrogram to extract the baseband FROG term as the nonlinear component contains many higher order terms along with the DC residue [106]. The expression for the cFROG spectrogram can be further expanded out.

\[ I_{\text{sig}}(\omega, \tau) = |\mathcal{F}_t\{E(t)^2 + E(t - \tau)^2 + 2E(t)E(t - \tau)\}|^2 \]  \hspace{1cm} (21)

In the case of cFROG, the terms \(E(t)^2\) and \(E(t - \tau)^2\) cannot be spatially removed separated from the FROG signal due to the collinear geometry. It is clear that fully expanding (21) will produce many other terms in addition to the FROG term of interest. We can define the complex field as the product of the field envelope with an attached carrier \(E(t) = E(t)e^{i\omega t}\) and the time delayed field as \(E(t - \tau) = E(t - \tau)e^{i\omega t}e^{-i\omega \tau}\). Using this replacement in (21) yields:

\[ I_{\text{sig}}(\omega, \tau) = |\mathcal{F}_t\{E(t)^2e^{2i\omega t} + E(t - \tau)^2e^{2i\omega t}e^{-2i\omega \tau} + 2E(t)E(t - \tau)e^{2i\omega t}e^{-i\omega \tau}\}|^2 \]  \hspace{1cm} (22)

While this expression may appear complicated, it is now evident that we have several distinct terms, each with a unique frequency shift. The FROG term is recovered by first Fourier transforming over the delay axis from \(\tau\) to \(k\) (not to be confused with the wave propagation vector \(\vec{k}\)), applying a filter \(H(\omega, k)\), then inverse Fourier transforming back to \(\tau\).

\[ I'_{\text{sig}}(\omega, \tau) = \mathcal{F}_t^{-1}\{\mathcal{F}_t\{I_{\text{sig}}(\omega, \tau)\}H(\omega, k)\} \]  \hspace{1cm} (23)

There are different possible filter shapes for pulling extracting \(E(t)E(t - \tau)\) from the cFROG spectrogram. The simplest is a square hi-low filter that only selects this term. However, the frequency domain representation of this interferogram does
not perfectly decay to zero between the FROG term and the higher order terms and using a filter with a hard cut off results in artifacts after filtering. A more suitable filter shape is a 4th order super-Gaussian of the form $H(\omega, k) = e^{-k^4/\sigma_k^4}$ with infinite width along the $\omega$ axis and a width along the $k$ axis chosen specifically to filter out the FROG term. After recovering $I'_{\text{sig}}(\omega, \tau)$, we still need to remove the background term $\mathcal{F}_t\{E(t)^2 + E(t-\tau)^2\}$. This is accomplished by subtracting the average of the spectrums at the two ends of scan in $\tau$ which contain only the background and not the FROG term. In Fig. 4.6, we show the original cFROG trace for a simulated pulse along with the result post filtering.

Fig. 4.6: Filtering a simulated cFROG spectrogram. A) Original simulated cFROG spectrogram. B) Fourier transform of cFROG along the delay axis. C) Filtered spectrogram after filtering to remove higher order terms. D) FROG spectrogram after filtering and removal of the background.
In Fig. 4.6, we demonstrate that the filtering algorithm is effective at separating out the FROG term of interest from the cFROG spectrogram. However, this process is not completely lossless, even with the appropriately sized filter applied. To explore this effect, we simulated both the collinear and noncollinear spectrograms. The collinear spectrogram is filtered, the background is removed, and the filtered cFROG is compared to the FROG term. We compute the weighted mean squared error between these two spectrograms as a function of filter size.

![Weighted MSE between the noncollinear and filtered FROG spectrograms of a simulated pulse as a function of filter size.](image)

We observe in Fig. 4.7 that using a too narrow filter results in a loss of information while a too broad filter allows bleed through from higher order terms. In the middle region, the error curve is relatively flat. For this particular trace, the best filter is 127 pixels wide and yields a weighted MSE of 0.086%. This minimum error may appear small, but the implication is that even when choosing the perfectly sized filter, information is still lost in the process. The effect is cascaded when applying
the iterative pulse reconstruction. Previously we reconstructed the pulse for a simulated FROG and calculated a weighted MSE of 0.13% after 1000 iterations with the error continuing to drop. We repeat this procedure for the same simulated pulse, this time generating the cFROG spectrogram and filtering it before reconstruction. This yields a weighted MSE of 0.46% after 1000 iterations, over 3x higher than when reconstructing the noncollinear trace.
4.5 - Experimental Retrieval of Ultrafast Optical Pulses

This effect of non-perfect filtering is magnified when reconstructing from real experimental data. As discussed earlier, it can be experimentally difficult or impossible to implement the noncollinear FROG measurement geometry. It is also impossible to implement the noncollinear geometry if attempting to temporally characterize the near-field as there is no way to separate the FROG term from the additional terms.

To reconstruct optical pulses originating from a Ti-Sapphire mode-locked laser, we construct a collinear FROG acquisition optical setup with a Michelson interferometer. Instead of using a standard nonlinear crystal such as BBO, we use a WS$_2$ monolayer to generate the SHG signal which is epicollected. A dichroic mirror separates the fundamental from the SHG signal in the reflected beam. The Ti-Sapphire femtosecond laser is centered near 800nm and has an expected temporal full width at half maximum (FWHM) bandwidth of around 50 – 80fs. To eliminate any aliasing effects, we record the spectrums at delay intervals of 0.5fs. The Nyquist criterion dictates we must sample with a temporal resolution finer than 0.68fs to fully sample each interference fringe along the delay axis. While we are only interested in reconstructing the pulse envelope and not the carrier, there are no downsides to sampling at a finer interval other than increased acquisition time.
Fig. 4.8: Optical set up used to collect cFROG data. A Ti-Sapphire laser beam is split in a Michelson interferometer with a delay stage in one arm. The two beams are focused and generate SHG off a WS$_2$ monolayer. This signal is epicollected and separated by a dichroic mirror for detection.

Fig. 4.9: Raw experimental capture of a cFROG spectrogram generated off a WS$_2$ monolayer
After capturing the spectrogram, the raw capture must be preprocessed. First, the spectrogram is interpolated from wavelength into frequency using the relationships \( c = \lambda f \) and \( \frac{\Delta \lambda}{\lambda_0} = -\frac{\Delta f}{f_0} \). The iterative MATLAB FROG reconstruction package also requires the spectrogram be interpolated onto a square grid such that the delay axis and frequency axes are Fourier transform pairs, \( f_{\text{span}} = \frac{1}{dt} \) and \( df = \frac{1}{t_{\text{span}}} \). These requirements make the interpolation and preprocessing of the experimental cFROG a nontrivial task which requires careful attention to detail. After interpolation, the spectrogram is further processed to remove the spectrometer background, eliminate any noise spikes, and normalized.

This cFROG spectrogram is then numerically filtered to remove the higher order terms, the background, and extract the FROG term used for the iterative reconstruction. After filtering, we observe a slight residue left over in the filtered FROG spectrogram in Fig. 4.10D that is not evident in the simulations. We mentioned before that the filtering step is an imperfect process, even when processing simulated, noise-free pulses. This effect is magnified when reconstructing experimental pulses due to several reasons. Collecting a FROG or cFROG spectrogram can take from several minutes to several hours depending on factors such as the resolution of the delay stage and the necessary spectrometer integration time. Factors on this time scale, such as thermal stability and laser drift, now play a significant role. This is clearly seen in Fig. 4.10A where we Fourier transform the cFROG spectrogram along the delay axis. Notice how the 5 terms are not clearly separated like they were when processing a simulated pulse. Instead, the information is smeared out and any choice of filter will result in either information loss or incomplete filtering.
After preprocessing and filtering the spectrogram, we reconstruct the complex field using the iterative principal components generalized projections algorithm. After running for 100 iterations, the weighted MSE plateaus around 2.5%. Qualitatively, the centers of the spectrograms appear very similar in Fig. 4.11. The reconstructed field has a slightly asymmetric intensity envelope with a FWHM of 74fs and a quadratic phase chirp.
These results highlight a critical challenge in using the PCGPA to reconstruct cFROG data. The algorithm minimizes the error between the reconstructed spectrogram and FROG spectrogram. When performing the collinear acquisition, the iterative comparison is between the reconstruction and the filtered FROG, not the original cFROG. As a result, it is difficult to obtain a true quantitative comparison between the ground truth and the reconstruction. To make some sort of quantitative comparison between the reconstruction and known entities, it is helpful to compare the reconstructed results with the autocorrelation. We calculate the temporal FWHM
\[ \Delta t \] to be 98fs from the autocorrelation. For a Gaussian pulse, the autocorrelation width scales by a factor of \( \sqrt{2} \) which estimates the pulse width at 69fs. This is slightly shorter than the predicted 74fs from the iterative retrieval algorithm and suggests the filtering and reconstruction sequence may be slightly broadening the field.

![Raw and Filtered Autocorrelation](image)

**Fig. 4.12:** Raw and filtered autocorrelation corresponding to the previous cFROG spectrogram. The autocorrelation predicts a temporal FWHM of 69fs compared to 74fs from the full iterative pulse retrieval.

To make a true quantitative error comparison, it is necessary to go back one step and generate the cFROG trace from the reconstructed pulse. The iterative reconstruction algorithm retrieves the envelope of the pulse \( E(t) \), but not the full pulse with the associated carrier \( \mathcal{E}(t) = E(t)e^{i\omega(t+t_0)} \). The carrier contains two terms that must be accounted for to reconstruct the ground truth. These are the time offset and the carrier frequency. To examine this, we analytically construct the full cFROG
spectrogram from the retrieved field and compare it to the experimentally collected cFROG spectrogram. We attempt to minimize the error between the two by detuning the center wavelength and adding a time shift. After spectrally centering the cFROG, we first detune the carrier frequency which results in an expansion or contraction of the cFROG spectrogram along the delay axis. We determine the minimum error corresponds to a center wavelength of 800.2nm as expected. A similar procedure is used to resolve the time shift. In both situations though, the weighted MSE is still over 50% between the true cFROG and the reconstructed cFROG compared to an error of 2.5% between the filtered FROG and reconstructed FROG. Due to the periodic interference pattern, slight shifts or imperfections in the interpolation steps can lead to very high errors, even while appearing visually similar. These effects introduce misalignments between the true and reconstructed cFROG spectrograms and make it difficult to make a valid error comparison.

Fig. 4.13: Weighted mean squared error between the original experimentally collected cFROG and the cFROG created after reconstruction. A) Weighted MSE as a function of spectral detuning. B) Weighted MSE as a function of time shift.
4.6 - Neural Network Reconstruction of Ultrafast Optical Pulses

We have seen how the iterative reconstruction algorithm can retrieve the unknown complex field from a cFROG spectrogram. This approach has serious limitations though. There exists an intrinsic loss of information in the filtering step and it is challenging to make a quantitative comparison between the cFROG reconstruction and the ground truth. These limitations are very relevant when applied to full spatio-temporal microscopy of nanostructures as it requires acquisition in the collinear geometry. Furthermore, the time for reconstruction scales linearly with the number of pulses to characterize. This is a nonissue for single point measurements, but becomes a serious constraint if it is necessary to retrieve a large number of pulses, for example, a spatio-temporal raster scan of a nanostructure. After filtering, the iterative algorithm must be applied at every spatial location to determine the associated temporal response. Even for a 10 x 10 pixel scan, this requires 100x the computational time.

To address these issues, we utilize machine learning. Machine learning applications can broadly be separated into two categories, classification and regression [107], [108]. Complex field retrieval falls into the latter category of supervised training and has recently been used for ultrashort pulse characterization under certain circumstances [20], [109]. Similarly, we will train a convolutional neural network (CNN) with many known pulses and known cFROG spectrograms. After sufficient training, the CNN should be able to ingest a never before seen spectrogram and retrieve the associated complex field. In this approach, there is no analytical requirement that the cFROG spectrograms must be first be filtered. Furthermore, pulse retrieval is nearly instantaneous following a baseline training time for the CNN.
Our method is similar to the one originally performed by Zahavy et. al. [20]. In their project, a CNN was trained with 60,000 simulated pulses and the corresponding FROG spectrograms. After training, their model was able to reconstruct the complex field of an unknown experimental pulse with an error comparable to the iterative reconstruction algorithm. Furthermore, their method outperforms the PCGPA and ptychography when retrieving fields from low SNR data. These factors make machine learning a prime candidate for attacking this problem.

![Image](image.png)


In our initial CNN prototyping, we reproduce some of the same results presented by Zahavy et. al. by reconstructing the complex field from noncollinear FROG data. We construct a sequential CNN using Keras/TensorFlow with a sample set of 10,000 unique pulses. These pulses were randomly generated with a temporal FWHM ranging between 30 – 150fs with added first, second, third, and fourth order dispersion terms. A FROG spectrogram is then generated from each pulse. To incorporate noise and limit overfitting, we add 30dB of additive white Gaussian noise
(WGN) to each FROG spectrogram. We then run the training procedure using the mean squared error (MSE) as the base optimizer. After 10,000 training epochs, which takes approximately 48 hours running on a dedicated NVIDIA RTX 2080Ti graphics card, the average MSE drops to less than 1e-4. For a visual comparison, we reconstruct four pulses unknown to the training model, separated into the absolute value and phase components. In Fig. 4.15, the phase is scaled from $[0, 2\pi]$ to $[0, 1]$ and the amplitude is normalized due to the CNN architecture. The phase data is truncated outside the range where the absolute value of the pulse drops to less than 0.001 because the MSE optimizer will attempt to optimize all of the phase, and optimizing the phase where there is no absolute value will compromise the reconstruction quality.

![Graphs showing reconstruction results](image)

**Fig. 4.15:** Reconstruction of 4 random pulses unknown to the training set. True pulses in blue and reconstructions in orange. The absolute value reconstructions are near exact while the phase reconstructions show more error.

We then turn our attention to a much more difficult task, reconstructing the complex field from unfiltered cFROG spectrograms. This goal required an entire
rewrite of our CNN architecture. We initially attempted the same approach of training a neural network with a set of cFROG traces and the corresponding absolute field values and phases as the labels, with poor results. Instead, we modify our CNN to minimize the training error not between the true and reconstructed fields, but between the true and reconstructed cFROG spectrograms. This is a valid approach as the FROG and cFROG spectrograms are a unique analytical mapping to a specific electric field minus time shift, time-reversal, and constant phase ambiguities [110]. This is a more robust approach than comparing the pulse profiles as it makes a quantitative comparison directly to the ground truth.

Our DeepcFROG neural network consists of three main stages. The first stage consists of 15 sequential hidden convolutional layers. This layer convolves over the initial cFROG spectrogram with windows of various sizes. Since the goal of our model is reconstruction not classification, each convolutional layer features a dropout to prevent overfitting to the training data. The output of the convolution layers is flattened into an array which contains the real and imaginary components of the complex field of interest. We use the real and imaginary components instead of the absolute value and phase to shield the network from any potential ambiguities that may arise from phase wrapping.

Fig. 4.16: Overview of the DeepcFROG neural network structure.
After the convolutional layers, we add two custom tensor-based layers. The reconstructed field at this point often displays high frequency ripples which are not physically realizable. To counter this effect, we incorporate a filtering layer which convolves the pulse with a super-Gaussian window function. The denoised field is then passed into a custom FROG layer which calculates the corresponding cFROG trace and compares the error between the output of this layer and the ground truth cFROG spectrogram. The resultant field reconstruction is pulled from the intermediate layer between filtering and FROG layers.

Fig. 4.17: DeePCFROG convolutional neural network consisting of several hidden convolutional layers along with the custom filtering and FROG layers.

We train this model with a set of 5,000 simulated pulses with temporal widths ranging between 40 – 150fs which corresponds to the expected pulse width of our Ti-Sapphire laser. Each pulse has first, second, third, and fourth order dispersion terms added to increase the pulse complexity. After training for approximately 4 days, we evaluate the performance of the model by reconstructing another set of 500 simulated pulses not included in the original training set. The true and reconstructed cFROG spectrograms along with the true and reconstructed fields of one of these pulses are in Fig. 4.18.
After reconstruction, the weighted mean square error between the simulated cFROG spectrograms and the reconstructed cFROG spectrograms is 0.08%. Compare this to the earlier results using the iterative reconstruction algorithm. The PCGPA reconstructs the spectrograms exactly for simulated noncollinear FROG data but not for collinear cFROG data. The PCGPA results show a reconstruction error of 2.5% in addition to 0.08% lost in the filtering process for simulated pulses. At best, this yields a cumulative error of 3.3% compared to 0.08% for the DeepcFROG neural network. In experimental reconstructions, this disparity should be even more pronounced as the neural network is shown to be more noise tolerant, and furthermore, the error is calculated all the way back up to the ground truth, eliminating the filtering step.

Fig. 4.18: True cFROG and reconstructed cFROG spectrograms after propagating through the neural net.
After demonstrating that our neural network can recover the complex field for simulated pulses, we turn our attention to the final problem: reconstructing real experimentally collected cFROG data via the DeepcFROG neural network. In this procedure, we experimentally collect cFROG data, interpolate the experimental trace onto the same frequency and delay grid used in the CNN, and pass the interpolated experimental cFROG spectrogram through the trained CNN. While the neural network is effective at recovering the field from simulated data, it performs very poorly for experimental data at this stage. Again, this is a result of fringe misalignment between the experimentally collected cFROG and the reconstruction generated through the CNN. The CNN prioritizes edge detection over envelope determination when it comes to feature extraction. The result of this is that any slight
misalignments between the experimentally collected cFROG and the ones generated through the CNN drive the error through the roof and result in a very poor reconstruction.

The most obvious solution to this problem would be to train the model at the exact sampling as the experimental collection. According to Nyquist, experimentally resolving each fringe, theoretically eliminating any aliasing in the experimental capture, requires sampling below 0.68fs. For a window of 500fs, this would require in excess of 830 points. In a 2D neural network such as ours, the training time scales linearly with the number of pixels in each element of the training set. This is simply too large of an image to sufficiently train our model using only a single 12GB GPU. Faced with this computational constraint, we must find a new approach. It is definitely worth revisiting this method though if one has access to a dedicated machine learning cluster.

Our next approach must address the three constraints. It must be able to reconstruct the complex field for experimentally captured cFROG data. It must limit the amount of information lost in filtering cFROG to FROG data. And it must be computationally efficient to the point where it can be run only on a GPU, not a cluster. Under these constraints, we modify the final FROG layer and loss function in our CNN. This is the benefit of building a modular CNN, it is possible to modify only certain layers without having to rewrite the entire network. Originally, the FROG layer computes the full cFROG spectrogram from the denoised complex field in the previous layer and uses this to minimize error in the network. We now modify the network to compare the cFROG after Fourier transforming along the delay axis, \( I(\omega, k) \). Since an ideal cFROG should be symmetric along the delay axis, the Fourier transform along the delay axis is purely real valued. Obviously though, experimentally collected cFROG data will not be perfectly symmetric due to noise,
system instability, and sampling error. By considering only the magnitude of the
Fourier transform and not the phase component, we enforce the constraint that the
cFROG should be symmetric and to ignore components which produce asymmetry.

It is evident in Fig. 4.20 that the most of the information of the Fourier
transformed spectrogram is now contained in a small region. Similar to the filtering
step used prior to the iterative reconstruction, we crop the center of $I(\omega,k)$ to isolate
the baseband and DC terms. The resultant structure is now a factor of 16 smaller
which significantly speeds up the training time by a proportional amount. This enables us to both quickly prototype the CNN and train with more samples. At this point it may be worth asking, “If you are isolating the baseband and DC terms, how is this different than the previous filtering process?” Good question. In this process, we are not Fourier transforming back from $I(\omega,k)$ to $I'(\omega,\tau)$. Spectral filtering always leaves a residue due to the shape of the filter, be it square, Gaussian, raised Cosine, etc. By not Fourier transforming back, we eliminate any artifacts caused by this. When reconstructing the field using either an iterative reconstruction or a DeepFROG
CNN like the one constructed in Zahavy, et. al., it is also necessary to remove the
background term $E(t)^2 + E(t-\tau)^2$. In our CNN, we retain the background and do not
Fourier transform back to $I'(\omega,k)$, eliminating possible residues that may originate
from this process. We admit though, this CNN is not true DeepcFROG as it does
ignore the higher order nonlinear terms. This is a hybridized CNN which retains
more information than traditional FROG (via iterative or CNN reconstruction) but less
than true DeepcFROG.
After training our model with the updated FROG layer, we calculate a weighted MSE of approximately $10^{-6}$ over 1,000 samples. After training, we reconstruct an experimental pulse from our Ti-Sapphire laser using both our hybridized CNN and the iterative PCGPA algorithm. We measure a FWHM of 63fs for the iteratively reconstructed pulse, 60fs for the CNN reconstruction, and 61fs from the autocorrelation. Qualitatively, the iteratively reconstructed pulse has a smooth Gaussian envelope with a quadratic chirp in the phase term. The CNN reconstruction exhibits temporal broadening in the bottom portion of the envelope compared to the iterative reconstruction as well as broadening of the phase.
To quantitively compare the error between the CNN and PCGPA retrievals, we compute the weighted mean squared error of the center $I(\omega, k)$ term between the experimental capture, the CNN reconstruction, and the one generated from the iteratively reconstructed complex field. To correct for any slight subpixel offsets between CNN and the experimental terms, we shift ±2 pixels in 0.2 pixel increments along the $\omega$ and $k$ axes and reconstruct a unique field at each point via the CNN. At a shift of 0 pixels along $\omega$ and 1.0 pixels along $k$, we compute a minimum WMSE of 7.1%. Unexpectedly, the reconstructed pulses are nearly identical for each shifted spectrogram fed into the CNN, even for ones with high errors between the experimental and reconstructed $I(\omega, k)$ terms. This implies that our model is approaching shift invariance and can compensate for small misalignments along both axes.
Fig. 4.22: WMSEs calculated between the experimental and reconstructed $I(\omega, k)$ terms for several shifts along both axes. The shifts strongly affect the error, but have minimal correlation to the pulse shape.

To compare the error against the PCGPA reconstruction, we must fully generate the cFROG spectrogram from the reconstructed field then Fourier transform to recover the $I(\omega, k)$ term. Between the experimental term and this term, we compute an error of almost 90% and visually, the PCGPA reconstruction term appears much broader than both the CNN and experimental terms. This is unexpected as the error between the experimentally filtered FROG and iteratively reconstructed FROG is only 2.7%. This is similar to the effect we observed previously when comparing the original cFROG to the cFROG generated from the reconstructed field following the iterative reconstruction. It is difficult to make a valid error comparison between the experimental capture and the reconstruction after the filtering process. Experimentally, it can only be assumed that this effect is magnified due to background noise, vibrations in the interferometer arm, laser fluctuations, and the
interpolation in preprocessing. Further work is needed to explore the impact of these factors in the resultant error. This result highlights the critical need to make an error determination with as little filtering and interpolation as possible.

Fig. 4.23: A) Experimental $I(\omega, k)$ generated by Fourier transforming along the $\tau$ axis and cropping the baseband and DC terms. B) Reconstructed $I(\omega, k)$ generated from the CNN. C) Reconstructed $I(\omega, k)$ generated by first iteratively retrieving the pulse via the PCGPA, constructing the full eFROG trace, Fourier transforming along the delay axis, then cropping.
4.7 - Neural Network Replication of cFROG Noise

To more accurately replicate the experimental parameters, we trace down the sources of noise in the resulting $I(\omega,k)$ spectrograms. We determine that there are three separate noise sources present in the experimental spectrograms that are not present when generating noise free spectrograms. This can be observed faintly in the normalized spectrogram, but it helps to plot these on a logarithm scale for better visualization in Fig. 4.24. The first obvious source of noise is background speckle noise present in the spectrometer and is marked by the black oval. To account for this, we add white Gaussian noise to the generated $I(\omega,\tau)$ before Fourier transforming over to $I(\omega,k)$. The second source of noise occurs at the baseband and DC location at the center of the $k$ axis and spreads to the outer edges of the $\omega$ axis which is marked by the white oval. This is due to both the noise floor of the spectrometer and the background speckle noise when Fourier transformed along the delay axis, as the DC component dominates. Even when subtracting the minimum value in the cFROG spectrogram, this term persists. The third noise term appears as a blurry region along the entire delay axis, marked by the red circle. This term is caused by instability in the interferometer via vibrations, air currents, and thermal drift. We saw previously in Chapter 3 how these effects had to be mitigated in spatial holography. In temporal gating, these effects are compounded by imperfect translation of the delay stage. In an ideal world, the delay stage would translate perfectly from one location to the next, say 500nm to 600nm. In reality though, it is more likely that the stage translates from 504nm to 598nm, for instance. This effect is accumulated over hundreds of data points and results in imperfect sampling which is smeared out in the frequency domain.
Fig. 4.24: A) Experimental $I(\omega,k)$ term cropped to only the DC and baseband terms. B) Same term plotted logarithmically to better visualize the sources of error. Error due to background speckle circled in black. Error due to background speckle and the spectrometer noise floor is circled in white. Error due to imperfect stage translation and sampling is circled in red.

Experimentally, it is desirable to limit these effects by reducing background light, shielding and damping the interferometer, and using a closed-loop translation stage. However, it is impossible to experimentally limit all sources of noise and we must model this in our CNN. To accurately replicate the experimental parameters, we now introduce these sources of error into the training set of our CNN while leaving the final FROG layer and loss function unchanged. From Fig. 4.24, we determine that each noise source has a different SNR. The background speckle noise is approximately -30dB, the interferometer and stage translation noise source is approximately -20dB, and the DC floor noise is approximately -10dB. Each of these terms are added to the generating code to simulate the experimental noise and interferometer instability. In principle, these parameters can be tuned to closely match and replicate the sources of noise in the experimental acquisition process. In Fig. 4.25, the noisy generated spectrogram now much more closely resembles the experimental one.
We now retrain our CNN with this noisy data tuned to match the experimental parameters. We observe that the training error in our model now decays more slowly compared to the noise free training. After training with 5,000 samples, we reprocess the experimental cFROG data through the CNN. Similar to the noise free case, we measure a FWHM of 60fs for the CNN reconstruction and 63fs for the PCGPA reconstruction. In Fig. 4.26, we observe broadening at the bottom of the pulse, but much less compared to the noise free CNN. Also, the phase profiles between the CNN and PCGPA now track much more closely. We measure an error of 7.1% between the $I(\omega,k)$ terms compared to 7.2% percent for the noise free CNN reconstruction. Because the CNN does not output noise within the model itself, it suggests that the error is now dominated by the noise contained in the experimental spectrogram.

Fig. 4.25: A) Simulated noise free $I(\omega,k)$ spectrogram plotted logarithmically. B) Simulated noisy $I(\omega,k)$ spectrogram with the 3 noise sources tuned to match the experimental parameters.
Based on these results, we conclude that accurately replicating experimental noise is critical in the CNN reconstruction. Qualitatively, the CNN results now closely track the PCGPA reconstruction, although they are slightly more structured. Considering the substantial filtering that occurs before applying the PCGPA, it is quite possible that the fine structure of the CNN is more representative of the true pulse while the PCGPA is displaying a smoothed-out version. It should be noted in this case that the noise parameters were carefully chosen to replicate the experimentally observed ones. In practice though, it is not efficient to first characterize the noise parameters of the experimental cFROG and then retrain the model for every occasion. The solution to this would be to train the CNN with a wide range of noise parameters. While this would increase the initial training time, it
would not require retraining. Similarly, adding more complex pulse profiles to the training set would generalize the model to a much broader range of potential pulse reconstructions.

In summary, we have demonstrated how machine learning can be applied to reconstruct ultrashort optical pulses with certain distinct advantages over the traditional iterative methods. While the iterative approach works well in the noncollinear geometry, filtering cFROG spectrograms to FROG spectrograms results in a loss of information. Furthermore, this loss of information is irrecoverable and it is difficult make a quantitative comparison all the way back to the ground truth. In comparison, reconstruction with a noise trained CNN is both faster and is fully quantitative for both FROG and cFROG.
5: Near-Field Microscopy of Nanomaterials

5.1 – Near-Field Scanning Optical Microscopy (NSOM)

In this chapter, we will discuss how near-field scanning optical microscopy (NSOM) can be utilized to characterize nanostructures on the subwavelength scale \cite{23}, \cite{112}. Using a homebuilt NSOM designed and constructed with Prof. Hans Hallen of NC State and Dr. William Murray of Penn State, we demonstrate how NSOM can be used to generate super-resolved second harmonic images of TMDs with a spatial resolution approaching 100nm. The same approach can also be applied to characterize the near-field behavior of metasurfaces. We also show how NSOM can be used in conjunction with ultrafast optical pulse reconstruction via FROG and machine learning to recover the full spatio-temporal near-field on the femtosecond scale. For a detailed explanation of NSOM construction and operation, see the 2020 Ph.D. dissertation by Dr. William Murray \cite{113}.

Acquiring a comprehensive understanding of optically active nanomaterials and nanostructures requires optical characterization below the diffraction limit. For example, WS\textsubscript{2} exhibits photoluminescent enhancement along edges and grain boundaries for monolayer depositions \cite{39}, \cite{61}, \cite{81}, \cite{114}. The exact physical mechanisms of these effects are still not thoroughly understood, nor are the spatial scales as evidenced in Fig. 5.1. Similarly, engineered metasurfaces operate by modifying the phase of an electromagnetic field on the subwavelength scale and thorough characterization requires subwavelength resolution via super-resolved optical microscopy \cite{2}, \cite{9}, \cite{46}.
While super resolution techniques such as stimulated emission and depletion (STED) or stochastic optical reconstruction microscopy (STORM) have been critical in understanding cellular and biological processes in the deep subwavelength regime, they both require fluorescent labeling and are primarily used for biological applications [15], [115]. Super-resolved biological imaging methods such as STED operate by superimposing two wavefronts to localize the excitation region on a sample to below the diffraction limit by taking advantage of nonlinear fluorescent bleaching [16], [116]. While fluorescent markers such as the green fluorescent protein (GFP) or quantum dots are widely used to label biological samples, they are not suitable for all photonic nanostructures. Because NSOM requires no fluorescent markers for operation, it is a viable technique for the super-resolved characterization of 2D materials and metasurfaces.

Super-resolved imaging without fluorophore attachment operates via a completely different mechanism [23], [45], [117], [118]. In this realm, breaking the diffraction limit requires collecting the near-field. Far-field optical microscopy contains the vast majority of optical techniques, and the theoretical transverse
resolution is given as $\frac{\lambda_0}{2NA}$ where $\lambda$ is the vacuum wavelength and $NA$ is the effective numeric aperture used for imaging. Practically, this is usually limited to $\frac{\lambda_0}{2}$ although liquid or solid immersion objectives can push the numeric aperture past 1 and obtain resolutions approaching $\frac{\lambda_0}{4}$ [119]. Imaging in the near-field is much more difficult, as it requires collecting the evanescent field which decays exponentially from surface, practically limiting it to 10s of nanometers. As such, it was mostly ignored until the creation of near-field microscopes in the 1980s [23].

![Diagram of near-field and far-field propagation. The near-field decays exponentially away from the sample surface.](image)

The mathematics of near-field optics make no simplifications or approximations of Maxwell’s equations and are far more complicated than their far-field counterparts [120], [121]. This predicts nanoscale effects such as polariton coupling and single molecule interactions which are not readily apparent in the far-field [22], [112], [122]. Since its practical realization, near-field microscopy has been
used to probe polariton interactions, quantum effects, semiconductor properties of materials at the nanoscale, and its functionality can be even further extended by combining it with techniques such as Raman [123]–[126].

There are three basic geometries which can be implemented for NSOM, one apertureless and two with apertures. Apertureless NSOM, also referred to as scattering scanning near-field optical microscopy (sSNOM), is perhaps the most common implementation and relatively turn-key commercial systems are available for purchase [22], [127]–[129]. In this geometry, a resonant tip is either held in close proximity to the surface of a sample, or makes slight contact with the sample if operating in tapping mode. Perturbations in the mechanical resonance of the tuning fork are used in conjunction with the scattered field to infer the optical response of the sample [127], [130]. Recently, sSNOM has been used to probe the polariton response of 2D materials on the nanoscale, providing valuable insight into the nanoscale material physics [129], [131], [132]. Furthermore, sSNOM can be combined with pump-probe spectroscopy to examine temporal material properties on the nanoscale [133]. While this geometry provides information on the near-field response of the sample, it has several limitations. sSNOM only enhances the z-polarized field, and therefore, is challenging to probe effects which operate on the x and y polarizations. It is also susceptible to field distortions, scattering from the imperfections in the substrate, and contains limited spectrographic information.

In other experimental geometries, also referred to as apertured NSOM, a tapered optical fiber is used to either illuminate or collect the signal generated off the sample. In the tip illumination geometry shown in Fig. 5.3B, the incident laser source is fiber-coupled into the probe, delivered through the tip, then evanescently coupled onto the sample and the signal is collected by a far-field detector or back through the tip. This geometry makes alignment and signal detection experimentally simple, but
is limited by how much power can be pushed through the tip before burning either the probe or the sample. This geometry is also not desirable when using an ultrafast illumination source due to temporal broadening through the tip. In our experiments, we illuminate in the far-field and collect evanescently through the tip. While this geometry is more challenging to align properly, it limits the risk of burning the tip and allows us to perform microscopy with minimal broadening effects that would be encountered by illuminating via an optical fiber. Originally, our NSOM was constructed in reflection mode (Fig. 5.3D), but we were forced to switch to transmission mode (Fig. 5.3C) due to burning issues. When oriented in transmission mode, we are able to evanescently collect both the fundamental and second harmonic signal without damaging the sample.

Using an ultrafast excitation source in the near-field collection geometry shown in Fig. 5.3C enables the collection of the ultrashort pulse absent temporal broadening. In this geometry, any second harmonic signal generated outside the tip preserves the temporal information of the ultrashort pulse. Because apertured NSOM permits simultaneous spectrographic measurement, a spectrally resolved technique such as FROG can be used in conjunction with NSOM to retrieve the temporal field of an ultrashort pulse at the nanoscale. Furthermore, we will show later that the presence of an NSOM probe and evanescent near-field collection does not result in significant temporal distortion of the original pulse.
In our experiments, we use a tapered optical fiber for near-field signal collection. After fabrication, the tips are coated with a thin (approx. 100nm) film of aluminum to form the aperture. A smaller, more well defined aperture produces a higher spatial resolution down to a practical limit of about 12nm due to the finite skin depth of the metal coating [112]. This fiber tip is cantilevered out and glued onto an electrically driven tuning fork [134], [135]. The tuning fork assembly is connected first to a dedicated pre-amp PCB then to the main PCB for the NSOM which further amplifies the signal and compensates for the capacitance and resistance of the tuning fork assembly. The signal is then sent back to a lock-in amplifier that sweeps to find the resonant frequency and measures any shifts in the resonant electrical behavior. The fiber tip probe can be modeled as a driven damped harmonic oscillator with a particular resonant frequency [134], [135]. When far away
from the sample surface, the tip and tuning fork assembly will resonant freely. When close, interactions between the tip and sample produce damped oscillations and a resonance shift to higher frequencies [134], [135]. This is analogous to pulling a vibrating string taut.

To measure in the near-field, the sample must be maintained in close proximity a few nanometers away from the sample surface in a state known as feedback. This is performed by first measuring the resonant amplitude and frequency of the tip assembly when not close to the sample. Based on this, we set a reference signal corresponding to about 80% of the maximum amplitude. Using a piezo driver, the sample is then pushed towards the tip assembly and this new signal is monitored. When this signal reaches the value for reference in, the sample is said to be "in feedback". The system continuously monitors this state and translates the piezo forwards or backwards (positive or negative voltage) to compensate. Near-field signal can only be collected while in feedback. Too far away will yield no evanescent coupling and too close will crash the probe, resulting in irreparable destruction of the tip. When monitored on an oscilloscope, 1V of this error signal corresponds to piezo high voltage signal of 20V and a movement of 95nm.
Fig. 5.4: A) SEM image of an NSOM tip that produces super-resolved SHG images. B) Mechanical setup at the core of the NSOM. The tip oscillates at a fixed position while the sample (e.g., a 2D material) is placed on a transparent substrate which is kept in feedback by a piezo tube. Far-field illumination is provided via the other side in transmission mode. C) Resonance curve for an NSOM tip not yet in feedback. Peak resonance is slightly below 61kHz. D) Monitoring the feedback in signal over 500ms. Roughly 20mV of noise on the feedback in signal corresponds to a spatial instability of 19nm. (A) SEM image courtesy of Yao Duan. (B,D) Figures courtesy of Dr. William Murray.

While feedback eats noise, too much instability in the system via electrical spikes, thermal drift, acoustics, or mechanical vibrations can easily knock the tip out of feedback, sometimes catastrophically. To limit these effects, the NSOM system was completely redesigned and rebuilt. To reduce electrical problems, the tip pre-amp and main NSOM PCBs were redesigned and rebuilt. The entire electrical supply to the amplifiers and probe was floated with two 12V car batteries to eliminate electrical spikes from the power grid. Eight piezo driven picomotor actuators were added to attain nanoscale alignment between the tip, sample, and laser spot. A Styrofoam and metal enclosure was constructed to increase the thermal stability and block air currents, and the inside of the entire enclosure was wrapped in DynaMat (commonly used for car audio installations) to soak up outside acoustics and
vibrations. Sandbags were also added at various locations on the main optical table to limit vibrations into the NSOM. Finally, the entire mechanical loop was shortened by placing the piezo sample stage and tip stage all on top of a large Newport 562 translation stage. A thorough discussion on the electrical and mechanical redesign of the NSOM can be found in the Ph.D. dissertation by Dr. William Murray of Penn State [113].

![NSOM System Diagram](image)

Fig. 5.5: A) Homebuilt microscope designed by Prof. Hans Hallen of NC State and Dr. William Murray of Penn State. The entire NSOM system is enclosed in a Styrofoam and metal enclosure to minimize thermal fluctuations and air currents. The inside of the enclosure is wrapped in DynaMat and sandbags are placed on the table to dampen acoustic and mechanical vibrations. 1) Transparent substrate on which the sample is mounted. 2) Extender tube for the piezo scanner. 3) Clamp for attaching the piezo tube. 4) Aluminum base plate for attaching the 5-axis piezo driven stage to the main piezo tube. 5) 5-axis New Focus stage with picomotor piezo drivers. 6) Aluminum baseplate. 7) Block for mounting the tip assembly. 8) Pre-amp for the tip assembly. 9) NSOM probe tip. 10) Newport 562 3-axis translation stage. 11) Picomotors for translating the Newport 562 3-axis stage. 12) Main base plate. 13) PCB to amplify and send signal from tip pre-amp to the lock-in amplifier.
5.2 – Algorithm Design of SNR Enhancement

NSOM has many promising applications as it is super-resolved, can probe all polarizations, and collects spectrographic information. Coupling NSOM with an ultrafast source also opens the door for possible spatio-temporal microscopy. However, this presents a significant challenge for signal collection and is compounded when attempting to characterize nonlinear behavior as effects such as SHG are much weaker. While NSOM probes often exhibit field enhancement effects due to the vector nature of the fields, signal collection at the nanoscale is still far weaker compared to far-field techniques [136], [137].

The two obvious solutions to this problem are to either increase the incident laser power or increase the spectrometer integration time, but both have practical limitations. The power of the excitation beam can only be set so high before damaging the sample. With the NSOM tip nearby, this threshold is even lower due to heating of the metallized tip and manifests as a lightning-rod burn pattern on the sample. This burn pattern was the primary reason why we changed geometries from reflection mode to transmission mode NSOM. In reflection mode, we were unable to generate sufficient SHG without burning the substrate. In transmission mode, we are able to collect sufficient SHG without burning, but there is still a threshold on how much power can be focused onto the sample before burning occurs.
Increasing the integration time of the spectrometer also has practical limits. In other low signal optical measurements such as Raman, these limitations are due to possible laser fluctuations, thermal stability, and dark noise on the spectrometer. In NSOM, we are challenged by the nanometer scale stability needed to maintain feedback during the entire duration of a scan. Any vibrations or thermal drift during this time can cause the tip to either move out of feedback or crash into the sample. The challenge then is to run a scan in the shortest possible time while also collecting high SNR data. Experimentally, these two factors are at odds with one another so we instead look to a numerical solution.

The challenge of accurate weak signal detection is two-fold. Typically, measuring the nonlinear response of a 2D material involves measuring the peak of the second harmonic spectrum. When the signal peak is thousands of counts above the spectrometer floor, this can be determined reliably. However, it is far more difficult to make an accurate determination for a signal that is only 3 or 4 counts above the noise floor. This is not even accounting for noise spikes due to factors
such as cosmic rays. This can be visualized in Fig. 5.7 where we perform a conventional nonlinear raster scan of a WS$_2$. The low signal measurement uses 250μW average power of the fundamental beam with 300ms spectrometer integration and corresponds to an SNR of approximately 1dB. The high signal measurement uses 2mW of average power with a 1000ms integration time and corresponds to an SNR of approximately 18dB. We define the SNR in decibels by comparing the max signal power to the RMS power of the background noise. The RMS power of the background is calculated after subtracting the floor of the spectrometer which is approximately 795 counts and thresholding above the max SHG signal to eliminate any cosmic rays.

\[
SNR_{dB} = 10\log_{10} \left( \frac{P_{sig}}{P_{noise}} \right)
\]  

(24)

As we observe in the low signal scan in Fig. 5.7B and Fig. 5.8B, the SHG region is difficult to visualize relative to the background and making an accurate determination of the peak SHG spectral intensity is difficult.

Fig. 5.7: A) 1000ms exposure raster scan of a WS$_2$ monolayer at 2mW of incident power. B) 300ms exposure raster scan of a WS$_2$ monolayer at 250μW of incident power.
In certain situations, it is desirable to preserve the entire SHG spectrum. However, for the purpose of merely determining peak SHG intensity, this is not required. Instead, we can utilize the entire SHG spectrum to make an accurate, decimal valued, single point evaluation. Often, this can be accomplished by simply taking the area under the curve of the SHG spectrum. To show this as a comparison, we process the low signal (250μW incident power, 300ms integration time) by integrating under the curve. We first threshold above the max SHG signal ceiling to eliminate any hot pixels and then sum along the spectral dimension. When comparing the background region to the SHG region, we calculate an SNR of approximately 4dB and can qualitatively visualize the enhancement in Fig. 5.9. This is certainly an improvement over the original low signal scan, but we can do better.
Our first approach uses an artificial neural network (ANN) to reconstruct the denoised SHG spectrum from low SNR raw spectrums. This is similar to the process used in Chapter 4, except an ANN uses only fully connected layers with no convolutional layers. Our model is trained with approximately 4,000 experimentally collected spectrums. Roughly half contain SHG spectrums and the other half contain background regions. The high signal spectrums were matched up to the low signal spectrums for both the SHG and the background spectrums.
While training, we observe the network validation loss plateaus very early, after only 10 iterations. This is not a good sign. When examining the reconstructions, it is evident that the neural network minimizes the model by fitting a spectrum that is between the background and SHG region for all spectrums. The model is unable to differentiate between low SNR background and SHG spectrums. It may be possible for the CNN to accurately reconstruct the spectrums from low SNR data by training with many more samples, however, collecting this much experimental data is not practical for our application. It is possible though to achieve good performance out of neural networks trained with limited or sparse data sets [138]–[140]. Similarly, neural networks have been used to denoise images and other signals [141], [142]. Such methods though require extensive modelling and control over the network parameters such as the number of neurons, weights, layer activations, and are usually used to extract features from multi-channel images or hyperspectral data [140], [141]. For our purpose though, we step back and modify our preprocessing algorithm before training with the ANN. In the future, it is worth developing a more comprehensive neural network to process the raw spectrums without any necessary preprocessing.

![Graph](image)

Fig. 5.11: Reconstructions of the background and SHG regions for the SNR boosting regression-based CNN are virtually identical.
We now preprocess the spectral data before the ANN by applying the constraint that the SHG spectrums are Gaussian shaped. This is a valid assumption for using a Ti-Sapphire laser source. A similar process could be used for hyperbolic secant pulse profiles. The first step is to go through the spectrums with a windowing function which removes any single pixel noise spikes followed by spectral filtering which removes much of the high frequency noise present in the spectrum. After this sequence, most of the pulse envelope is recovered. We then fit the filtered spectrum to Gaussian function. Fitting to a function, such as a Gaussian, is commonly done using least squares minimization, or the $L_2$ norm. In our situation, however, we need a fit to a smoothly varying envelope, and fitting to outlying data points is undesirable. For this reason, we use the $L_1$ norm. The Gaussian fit in (26) is applied to the spectrum $y(\lambda)$ in (25) with the free parameters $A$, $b$, $\lambda_0$, and $\Delta \lambda$.

$$\text{Argmin} \sum_j |y(\lambda_j) - f(\lambda_j)|$$  (25)
\[ f(\lambda) = A e^{\frac{(\lambda - \lambda_0)^2}{2\Delta \lambda^2}} + b \]  

(26)

In theory, the free parameters in (26) can be constrained specifically to the expected values for the SHG spectrum with \( A \) freely varying. The results of this preprocessing sequence are shown in Fig. 5.13.

![Spectral preprocessing sequence](image)

Fig. 5.13: Spectral preprocessing sequence. A) Raw low signal experimental SHG spectrum. B) SHG spectrum after removing hot pixels and spectral filtering. C) SHG spectrum with Gaussian fit applied.

We observe the preprocessing algorithm effectively recovers a decimal valued SHG spectrum at only 4 counts above the noise floor. From this step, it is easy to make a single point power determination. The obvious method forward would be to make a single point power measurement for every spectrum and use this to generate the denoised raster scan image. There is a flaw though. The Gaussian fit will attempt to fit a Gaussian profile to every spectrum. Ideally, this would produce a smooth Gaussian spectrum for a SHG region and a flat (zero amplitude or infinite width) fit
to background data. Numerically though, it is difficult to fully constrain this Gaussian fitting in a computationally efficient manner and when processing thousands of spectrums, this routine is almost guaranteed to create some false fits to noise as evidenced in Fig. 5.14.

![Fig. 5.14: Example of a false fit to noise in a background region containing no SHG signal.](image)

Visually, most of these false fits are easy to identify. Manually separating true SHG spectrums from false fits though, is not a practical approach when processing thousands of individual spectrums contained in a raster scan. We need a turn-key method for automating this process, and turn our attention back to machine learning. Previously, we observed that a regression-based ANN was unable to reconstruct a high-fidelity spectrum from low SNR data. We now task the ANN with a much simpler task, separating true SHG spectrums from false ones.

Our new ANN consists of 15 full connected layers of size 1340 by 1. It ingests the preprocessed spectrums and separates them into two categories, either SHG or background, using a binary cross entropy loss function at the final layer. To train the model, we use a set of approximately 2,000 experimentally collected, high signal SHG and background spectrums. These high signal spectrums were first separated into SHG and background categories by taking the area under the curve and thresholding. While effective for high signal spectrums, this separation procedure cannot be used for low signal spectrums as they contain too much noise.
After training, our model correctly classifies background vs. SHG at 99.4% accuracy for data included in the training set and 98.6% for data in the validation set (unknown to the training model). Using a larger set of experimental data for training would likely increase the accuracy even further. Following classification by the ANN, the spectrums are converted into single point power measurements by taking the peak intensity of the SHG spectrums and an average value for background spectrums. The results of this fitting routine are visualized in Fig. 5.16 where we show the same low signal (250μW incident power and 300ms integration time) raster scan before and after the denoising algorithm described above. The final SNR is approximately 9dB compared to 1dB for the raw scan and 4dB for summing the spectrum. These results suggest it is possible to collect meaningful SHG data at only 3 or 4 counts above the floor of the spectrometer. While this algorithm was designed specifically to handle Gaussian-like spectrums for processing NSOM data, it could easily be adapted to other applications in low signal microscopy.
Fig. 5.16: A) Original low signal (250μW incident power and 300ms integration time) raster scan of a WS$_2$ monolayer. B) Low signal raster scan after classifying SHG vs background spectrums from the ANN. C) Low signal WS$_2$ monolayer raster scan after denoising and spectral integration. D) Low signal raster scan after total denoising algorithm and ANN classification with peak second harmonic signal 3-4 counts above the background.
5.3 - Super-Resolved Nonlinear Microscopy of 2D Materials

After creating the denoising algorithm to process low SNR data, we turn our attention back to the NSOM. To demonstrate super-resolved imaging through the NSOM, we conduct a 2D raster scan of a sample consisting of WS$_2$ monolayers deposited on a quartz substrate. Using a Ti-Sapphire model-locked excitation source with a fundamental wavelength centered around 800nm, we collect the second harmonic signal generated off the sample. We first collect the reflected signal in the far-field across a 2µm x 2µm square region scanned at 20nm increments using the piezo drivers on the primary NSOM stage. In Fig. 5.17, we display the 2D raster scan, both before and after the SNR enhancing algorithm, with a line scan taken at x = 600nm. We observe a maximal second harmonic intensity of about 70 counts above the spectrometer floor. At this level of signal relative to the background noise, the effects of the denoising algorithm are not as pronounced as those demonstrated in the previous section, although this still assists in accurately determining the resolution. To determine the resolution, we compare the SHG intensity between the 10% and 90% points along the line scan at x = 600nm and estimate a 708nm resolution for the far-field scan and a 630nm resolution when correcting for the angular offset of the line scan relative to the monolayer edge.
We then perform a smaller 1μm x 1μm raster scan along the same WS$_2$ edge and collect the near-field second harmonic through the NSOM tip in transmission mode. Our resolution should now be limited by the aperture of the tip and, as shown in Fig. 5.18, the collected signal is now much weaker. While it is possible to increase the integration time on the spectrometer, a longer scan increases the risk of thermal drift, laser drift, and unstable feedback. Fortunately, our denoising algorithm can process the data to accurately determine the system resolution. Using the same 10 – 90 estimate, we calculate an optical resolution of 142nm. Our true resolution is even higher as the monolayer edge is oriented at 27 degrees relative to the line scan direction. Correcting for this angle, we calculate a resolution of 126nm which is
conclusively into super-resolved territory. Using a better probe with a tighter aperture would yield an even higher resolution.

Fig. 5.18: Near-field second harmonic raster scan of the same WS$_2$ monolayer deposited on a quartz substrate. A) Raw SHG raster scan. B) Raster scan after denoising and fitting algorithm. C) Raw line scan taken at $x = 300$nm. D) Line scan taken at $x = 300$nm from the fitted 2D raster scan.

During the course of measurement, we also observe a polarization dependence between the coupling of both the fundamental and the second harmonic signal into the NSOM tip. This polarization dependence may be due to the probe tip geometry or to some plasmonic resonant interaction between the sample and the tip when in feedback. To further examine this effect, we place a half-wave polarizer in the path of the NSOM laser and record the coupled signal intensity as the waveplate is turned. Interestingly, the second harmonic signal is offset angularly from the fundamental
and exhibits secondary side lobes. At this time, the exact mechanisms of this effect are unknown and further investigation is needed.

Fig. 5.19: Polar plot (in degrees) as a function of incident polarization for the coupled fundamental and second harmonic near-field signals.
5.4 - Near-Field Optical Pulse Reconstruction

We have demonstrated that NSOM can optically characterize materials below the diffraction limit, in our case approaching 100nm. We are not limited to only characterizing the spatial field though. NSOM probes have previously been used to collect time-gated data for pulse retrieval, but not while spatially super-resolved [143]. Temporal near-field characterization is a much more difficult problem and requires evanescently collecting a cFROG spectrogram through a tightly apertured NSOM probe. To our knowledge, we are the first people to accomplish this feat. This is particularly relevant as many nanostructures are theorized to have distinct temporal response and spatio-temporal field coupling which is only observable on the nanoscale [11].

Before examining any potential temporal effects generated from a patterned nanostructure, we must first determine if there are any temporal distortion effects resulting from near-field evanescent coupling. To examine this, we place a Michelson interferometer in the laser excitation path before the NSOM. A prism pair is used prior to the interferometer to precompensate for broadening. In this geometry, the nonlinear medium used for cFROG data generation is a nonlinear 2D material. We use this system to collect cFROG data both in the far-field, which epi-collected, and in the near-field by collecting cFROG data directly through the tip in feedback. In the near-field, it is necessary to collect the full cFROG as there is no phase matching. Even if one could set up the off-axis FROG acquisition geometry within the NSOM environment, any near-field collected signal would contain other terms in addition to the FROG term.
After collection, we numerically filter out the FROG term of interest from the full cFROG trace. In this situation, the FROG data was sampled at 1fs intervals over a 400fs range to shorten the acquisition time and mitigate the risk of the NSOM falling out of feedback. In this scenario, we cannot use the previously described fitting algorithm for raster scans as the spectral domain of the cFROG trace is structured due to the interference pattern and we can no longer assume a Gaussian profile for fitting. After filtering out the baseband term $E(t)E(t - \tau)$ and removing the DC component, we interpolate the resultant FROG onto a square grid and apply the PCGPA iterative retrieval algorithm to recover the complex field.
Fig. 5.21: Reconstruction of the complex field from the cFROG spectrogram collected off a WS$_2$ monolayer and captured in the far-field. Top) FROG spectrogram after filtering and the reconstructed FROG spectrogram after the PCGPA. Middle) Reconstructed time domain field separated into absolute value and phase. Bottom) Reconstructed spectral domain field separated into absolute value and phase.
Fig. 5.22: Reconstruction of the complex near-field from the cFROG spectrogram collected off a WS₂ monolayer and captured evanescently through an NSOM probe in feedback. Top) FROG spectrogram after filtering and the reconstructed FROG spectrogram after the PCGPA. Middle) Reconstructed time domain field separated into absolute value and phase. Bottom) Reconstructed spectral domain field separated into absolute value and phase.

In Fig. 5.21 and Fig. 5.22, we calculate a weighted MSE of 1.0% and 4.0% between the reconstructed and filtered experimental FROGs for the far-field and near-field, respectively. The lower error in the far-field collection is likely due to having more signal in the far-field collection compared to the near-field collection. To better compare the near-field and far-field reconstructions, we manually correct for the ambiguities which cannot be resolved by FROG (time-reversal, time shift, and constant phase.) We calculate a FWHM of 69fs for the far-field collected pulse and 74fs for near-field collected pulse. When comparing the phase profiles, they overlap very closely inside the main pulse envelope. Outside of this region, there is almost no field amplitude and the phases walk off from each other as expected.
After comparing the temporal results via the iterative PCGPA algorithm, we now compare them with noise trained neural network discussed in Chapter 4. In this situation, the full cFROG spectrograms were passed into the CNN without any filtering or smoothing applied. From the reconstructions shown in Fig. 5.24, we measure a FWHM of 69fs for the far-field and 68fs for the near-field. The far-field FWHM calculated from the CNN is the same as the PCGPA reconstruction while the near-field reconstruction is 5fs narrower. When comparing the weighted MSEs between the experimental and CNN reconstructed $I(\omega, k)$ terms, we measure 13.2% error for the far-field reconstruction and 8.9% for the near-field reconstruction. This is interesting as we would assume the near-field reconstructions contain more noise which should lead to a higher relative error. Since the filtering process leads to
broadening, these results may suggest that the CNN and iterative reconstructions are similar for the high-signal far-field data, but the CNN does a better job handling the low-signal near-field data. Based on these results, it appears that evanescently collecting through an apertured NSOM probe has minimal (if any) effect on the temporal pulse profile. Obviously, more work is needed to thoroughly explore any potential temporal broadening effects due to the probe as all of our NSOM results were collected with a singular probe.

Fig. 5.24: Comparison between the far-field and near-field CNN pulse reconstructions showing the overlaid absolute values (top) and phases (bottom).

After comparing the far-field vs. near-field temporal behavior, we turn our attention back to the polarization effect. As shown earlier in Fig. 5.19, there is polarization dependence between the incident field, the coupled near-field fundamental, and the coupled near-field second harmonic. We investigate whether this polarization effect also perturbs the temporal behavior. We first use the iterative PCGPA to reconstruct the coupled near-fields at incident polarizations ranging from
25 – 86 degrees, going from the null to the maximum in the polarization polar plot. In Fig. 5.25, the reconstructed fields are similar for all polarizations with a FWHM ranging between 81 – 84fs and a similar phase structure throughout the main pulse envelope.

![Graphs showing reconstructed fields for different polarizations.](image)

**Fig. 5.25:** Comparison between the PCGPA iteratively reconstructed near-fields at incident polarizations of 25, 38, 51, 65, 76, and 85 degrees showing the absolute values (top) and phases (bottom).

We then use the CNN to reconstruct this set of pulses and compare it to the PCGPA iterative algorithm in Fig. 5.26. Once again, the intensity and phase envelopes track each other, although less so compared to the PCGPA reconstruction. It is difficult to say if the fine structure in Fig. 5.26 is a real effect or is an artifact from the CNN.
Interestingly, the range of pulse widths for the CNN near-field polarization reconstructions vary between 67 – 75fs compared to 81 – 84fs from the PCGPA reconstruction and 63 – 68fs from the autocorrelation calculations. Interestingly, the FWHM of the CNN reconstructions are closer to the autocorrelation widths than the PCGPA. Since the near-field scans are relatively weak, it is possible that filtering the experimental cFROG spectrograms in preparation for the iterative reconstruction introduces significant broadening that is not present in the CNN retrievals. When comparing the weighted MSEs between the experimentally captured and reconstructed cFROGs, we observe CNN significantly outperforms the PCGPA with errors ranging from 6.3 – 13% compared to 52 – 93%, respectively. This is better visualized in Fig. 5.27.
Based on the results in this Chapter, we conclude that it is possible to characterize nanomaterials in the near-field using apertured NSOM with a resolution approaching 100nm. Given an NSOM probe with a tighter aperture, we could likely improve this resolution by a factor of 2 or more. We demonstrate that super-resolved optical characterization via NSOM can be applied to investigate not only the linear, but also the nonlinear optical properties of nanomaterials and nanostructures.

Furthermore, we demonstrate that it is possible to recover the temporal near-field using both the PCGPA and the DeepcFROG CNN described in Chapter 4. To the best of our knowledge, this is the first instance of retrieving the temporal pulse profile
of the near-field. Based on our results, we conclude that the presence of an NSOM probe in close proximity to the sample does not introduce any significant temporal distortions. When applied to nanostructure characterization, this is critical as any temporal perturbations can then be attributed solely to the material response. The synthesis of these techniques creates a powerful tool for characterizing the spatio-temporal response of nanomaterials and nanostructures on the nanometer-femtosecond scale.
6: Conclusions and Future Work

We have developed and demonstrated a set of new and novel tools for the efficient characterization of optical nanostructures and nanomaterials. Two-dimensional materials, specifically transition metal dichalcogenides (TMDs), exhibit widely varying optical and electronic properties as they are thinned down from their bulk counterparts to single and few layered depositions [4], [5]. Current tools such as AFM and ellipsometry can accurately quantify their surface topography and refractive index, but not both simultaneously. They are both scanning methods which makes them cumbersome and inefficient for measuring particularly large or heterogenous samples. In Chapter 3, we construct a holographic microscope to simultaneously image and quantify samples containing TMDs of varying thicknesses. Such a method is non-invasive and single capture with a field of view exceeding 100μm. We further extend the system capabilities by numerically modelling the multi-layer sample structure to predict both the 2D material thickness and the complex refractive index. The values for the predicted layer thickness are in good agreement with measurements obtained through AFM, and the predicted refractive indices for WS<sub>2</sub> and MoS<sub>2</sub> are in agreement with those reported in the literature [8], [83], [95], [98]. This work is summarized and published in [90]. Practically, such an approach can be applied to other materials which require simultaneous topographic and optical characterization.

Many 2D materials also exhibit nanoscale optical effects, such as photoluminescent edge enhancement and defect based optical perturbations which cannot be fully understood through far-field characterization [60], [61], [81]. Systematic probing of these effects is critical to advance our knowledge of material interactions at the nanoscale. A similar set of challenges exist for characterizing
artificially engineered optical nanostructures, commonly known as metamaterials and metasurfaces, which modify the phase of an electromagnetic wave. Metasurfaces are created by lithographically patterning metallic or dielectric structures with specific dimensions well below the diffraction limit [10], [22], [48]. As such, their optical response is often measured in the far-field. In Chapter 5, we use apertured near-field scanning optical microscopy (NSOM) to optically characterize nanomaterials with a resolution approaching 100nm. This was following a complete electromechanical redesign by Dr. William Murray and the development of a machine learning based fitting routine that enables the extraction of meaningful results from low SNR data. A nearly turn-key system is now operational and can be used to probe the near-field of more 2D materials, nanostructures, and metasurfaces.

Some optical nanostructures also exhibit strong coupling between the spatial and temporal domains [11]. This phenomenon is predicted through simulations as currently there are few methods which can probe the optical field on the required nanometer-femtosecond scale. In Chapter 5, we demonstrate that collinear frequency resolved optical gating (cFROG) can be performed within the NSOM environment to retrieve the temporal near-field. Due to the absence of phase matching in the near-field, pulse retrieval must be obtained through cFROG. In Chapter 4, we explore the limitations in reconstructing ultrafast pulses in this geometry due to a loss of information in the required filtering step. Inspired by the work of Zahavy, et. al., we construct a modified convolutional neural network to retrieve the complex field from the full cFROG spectrogram (DeepcFROG) [20]. We discover that replicating the noise inherent in experimental pulse retrieval is critical in effective CNN pulse retrieval. Our CNN outperforms the iterative principal components generalized projections (PCGPA) iterative algorithm in error, noise tolerance, and computation speed after the model has been sufficiently trained.
These improvements speed up both the data acquisition and processing time. Near term work on the CNN will incorporate a more generalized noise model and a training set consisting of more complicated and structured pulses. This will generalize the CNN to work for many more experimental situations. Potential far-term work may include modifying the model to retrieve attosecond scale pulses. In the attosecond regime, conventional FROG retrieval methods fail as the pulse duration is now shorter than the carrier. Recently, machine learning has been applied to this problem, but only for characterizing pure laser pulses [109], [144]. Attosecond scale temporal resolution together with nanometer spatial resolution would probe fundamental material physics never before seen.

Using both the DeepcFROG neural network and the PCGPA iterative retrieval, we compare the far-field and near-field temporal structure. We conclude that evanescently collecting the near-field via an NSOM probe does not introduce appreciable temporal distortion, and therefore, any temporal response could then be directly attributed to the sample under test. We further show that the polarization dependence of the evanescently collected SHG is not due to any temporal effects. When retrieving the temporal near-field, the CNN reconstruction consistently produces a slightly narrower pulse than the one retrieved iteratively and exhibits a smaller error between the reconstructed and experimentally collected cFROG spectrograms. These results may be attributed to broadening in the filtering step prior to PCGPA retrieval.

These techniques outlined in this dissertation can potentially be synthesized into scanning ultrafast near-field second harmonic optical microscopy (SUNSHOM), a method which recovers the spatio-temporal electromagnetic field on the nanometer-femtosecond scale. This is analogous to a method in biological optical
characterization of combining super-resolved stimulated emission depletion (STED) microscopy with temporal fluorescent lifetime imaging (FLIM) [15].

<table>
<thead>
<tr>
<th>Application</th>
<th>AFM</th>
<th>FLIM</th>
<th>STED</th>
<th>FLIM + STED</th>
<th>FROG</th>
<th>NSOM</th>
<th>SUNSHOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lateral Resolution</td>
<td>&gt; 1nm</td>
<td>-</td>
<td>&gt; 20nm</td>
<td>&gt; 20nm</td>
<td>-</td>
<td>&gt; 12nm</td>
<td>&gt; 12nm</td>
</tr>
<tr>
<td>Axial Resolution</td>
<td>&gt; 1nm</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&gt; 1nm</td>
<td>&gt; 1nm</td>
</tr>
<tr>
<td>Temporal Resolution</td>
<td>-</td>
<td>ns</td>
<td>-</td>
<td>ns</td>
<td>fs</td>
<td>-</td>
<td>fs, as?</td>
</tr>
</tbody>
</table>

Table 6.1: Comparison between the lateral, axial, and temporal resolution of select microscopy techniques [15], [16], [76], [77], [102], [109], [122].

At this stage, collecting temporal information requires a nonlinear medium on the substrate. However, it is possible to functionalize an NSOM probe tip with nanoparticles to modify the near-field response and obtain specific functionalities [145], [146]. Our next step would be to functionalize an NSOM probe tip with a nonlinear medium, such as ZnO nanoparticles [147]. This would move the nonlinear interaction necessary for pulse retrieval directly to the tip and enable temporal reconstruction to be performed on samples which do not intrinsically possess a $\chi^{(2)}$.

Fig. 6.1: Diagram of measuring a nanostructure in the NSOM. Moving the nonlinear interaction from the substrate to the tip permits temporal characterization of structures that do not intrinsically possess a second order nonlinear susceptibility.
NSOM can be used to extract even further information beyond the temporal response. For example, it is possible to collect Raman in the near-field, furthering the spectroscopic understanding of materials and molecular interactions at the nanoscale [98], [108]. It may also be possible to perform true optical holography in the near-field which has been postulated and numerically simulated, but never experimentally realized [148].

In summary, we have discussed how certain techniques in nonlinear and ultrafast optics can be merged with applied numerical methods to create a new caliber of tools for understanding nanomaterials and nanostructures on the nanometer-femtosecond scale. It has been said that to conduct research is to stand on the shoulders of giants. It is my hope that this work is just the tip of the iceberg and will expanded upon to gain more understanding into the physical world at the nanoscale.
References


VITA

Joshua Alton Noble

Education

Ph.D., Electrical Engineering
Focus – Optics and Numerical Methods
Pennsylvania State University
Dec. 2020

M.S., Electrical, Computer, and Energy Engineering
Focus - Optics
University of Colorado – Boulder
Aug. 2016

B.S., Electrical Engineering
B.S., Physics
Minor, Mathematics
Pennsylvania State University
May. 2014

Experience

Graduate Research Assistant
Pennsylvania State University
2016 – 2020

Teaching Assistant, Senior Design Lab
Pennsylvania State University
2016 – 2017

Graduate Research Assistant
University of Colorado – Boulder
2014 – 2016

Undergraduate Teaching Intern
Pennsylvania State University
2012 – 2014

Awards and Accomplishments

First to retrieve the temporal near-field

Renaissance Graduate Fellowship, GAANN Graduate Fellowship, 2014 James M. Barnak Outstanding Senior in Electrical Engineering, McNair Scholar, Erdős number: 4

Selected Publications and Presentations


