Fourier Phase Retrieval: A Survey

March 17, 2021

Zeyang Huang

1. Introduction

Foundations of holography started in 1948 when a two-step lensless imaging process was proposed by Dennis Gabor, who considered objects being illuminated under an inline point light source [27]. Given a proper coherent illuminated object, phase and amplitude can be preserved, and we can obtain the wavefront reconstruction of the original object [29]. There are two steps involved in this process: wavefront detection and reconstruction. For the detection or recording step, phase information needs to correspond to light intensity information with the interferometry method. Similar to stereo matching, where pairs of images corresponding to left and right eyes are considered in recovering the true depth of 3d object, for interferometry, a reference wavefront that is mutually coherent with the original, has a uniform intensity, and providing a linear mapping of intensity is introduced as the recording medium. For the second reconstruction step, since the amplitude and phase information of the original wave is recorded, we can obtain a copy of the original or the conjugate of the reference wave. Books and papers like [8, 27, 29] provide a theoretical foundation of holography.

In 1966, Brown and Lohmann wrote a milestone paper on Computer Generated Holograms (CGH). Since it was not practical to control both the amplitude and phase of transparency, the “detour-phase” method was proposed to encode complex amplitude and plot binary patterns [7, 8]. Starting then, many approaches emerged consisting mainly of iterative optimization methods. Analogous to the differences between integration methods and numerical integration, CGH works towards an easier way of obtaining phase and amplitude instead of the two-step wavefront reconstruction by using computational cues and approximations [6]. Recently, because of the huge increase in computational power, the theoretically structured idea has found its applications in various fields besides microscopy including VR, AR, and other dynamic holographic displays. Neural networks are used in replacing some steps in the pipeline of traditional iterative methods, or wrapped with end-to-end methods. An Increasing number of deep learning-based methods are proposed during these years for getting more accurate and fast results.

The wavefront reconstruction process is based on regenerating a complex field in the hologram plane, and in CGH, the field is sampled with each sample point associated with complex values. We wish to create a hologram field that is either the Fourier transform in the far-field or the fresnel transform in the near field of the desired object [12, 29]. The existence of a spatial light modulator at the Fourier plane enables dynamic presentations by reading out interference intensities, with an extra cost in configuration [34]. However, phase information may need to be recovered to know the original wavefront. In this survey, we focused on Fourier phase retrieval methods. The goal of Fourier phase retrieval is to recover an unknown signal whose Fourier transform leads to the observation signal; in other words, we wish to recover the original signal given the nonlinear magnitude measurements of it.

Similar to stereo matching, Fourier phase retrieval is a sensing problem, and unlike the passive stereo matching problem, the cost of its measurements is high. Via methods like spatial light modulator, we need to find ways to rectify the image From a theoretical perspective, we would like to know if there is enough information to reconstruct the original wave given the summation of the reference wave and the original [29]. Practically, similar to many inverse problems such as scene reconstruction in computer vision, common practical challenges of CGH are accuracy and speed, and it suffers from missing or corrupted data, and noises such as aberration, and works towards higher resolution and faster speed. More specifically, the common limitations of methods include noise sensitivity, stagnation, and lack of generalizability.

Moreover, the phase retrieval problem desires results that are consistent with observation measurements, and constraints. It is impossible to solve without a priori information [22]. The challenge of phase trivial from an optimization perspective is that it is ill-posed in one dimension. The mapping function is symmetric and remains invariant after shifts or conjugation. Because of its nonconvex nonlinear nature from rotational or permutational symmetry, the convergence may only be guaranteed to local minimizers [23]. We are unsure if the local minima encode any global information, or which of multiple local minimizers it converges to [13]. Many methods are introduced in the following sections to tackle the above challenges from various aspects.

2. Quantitative Methods

Both the intensity and amplitude-based loss function in minimization suffer from non-concavity, and the original gradient algorithm does not guarantee convergence to the target solution with random initialization. Since non-convex optimization is difficult to solve, there are two main categories of tackling the question: using random samples represented by high dimensional matrices to do alternative projection so that it spans in the measurement matrices, and applying convex relaxations to approximate the problem so that linear operators will be applied [20, 27].

2.1 Iterative Optimization with Sampling

Iterative optimization algorithms play an important role in coherent diffraction imaging, where Fourier transforms relate each far-field wavefront with object distribution. This relationship can be replaced by integral transformations such as the Fresnel principle to be expanded in hologram [26]. The error reducing Gerchberg-Saxton (GS) algorithm, falling into the first category of alternating projection, is probably one of the best-known algorithms in getting a phase-only complex value wavefront using backward and forward propagation between the object and Fourier planes based on intensity inputs from sample plane and detector plane [23].

Similar to other iterative methods such as expected maximization, the procedures of GS algorithms also make an initial estimation, optimizing the model, and repeating the steps. Specifically, the algorithm assigns random phase distributions with the desired modulus in the object field, transforms into Fourier diffraction domain, updates the field with new modules in the diffraction plane, applies inverse Fourier transformation in the object plane, and then replace with measured modulus resulting in an updated object distribution to start over [24, 26]. This GS method marks as a template that enables recovering phases that are quantitatively correct as well as absorption distribution from the objects [26]. However, practically, the absolute error of each bit remains high even though global variance will be reduced. Since from an input-output perspective, we could subtract the current input certain times to drive a positive pixel value to be zero, similarly, Fienup pointed out one important observation in solving this by adding a small perturbation of the input resulting in a constant in the output that relates to the distribution of the Fourier transform moduli [24].

2.1.1 Gradient Descent

The GS algorithm is closely related to gradient descent [24]. If we take a closer look at the convergence of the GS algorithm by computing the square error in the Fourier domain between the first and second step, then we need to take the sum of the square error of estimations that violate the Fourier domain constraints. A new estimation can be formed by computing its gradients in N2 dimensions with enforcing satisfactions of object-domain constraint, for N2 representing the 2 dimensions of sampled data, while the step size in the direction of gradient descent can be estimated by a first-order Taylor series of the error metric [23]. Single intensity measurements of the gradient descent and GS algorithms look identical and the error reduction algorithm converges as error monotonically decreases. However, Fienup also pointed out that the slow convergence rate can be improved using a hybrid input-output (HIO) algorithm that avoids stagnation of outputs in successive iterations by adding additional correction terms.

The improvements of the GS algorithms are either from the perspective of input by introducing terms and constraints in the input real domain, or alternatively, change the error metrics so the different feedback would result in different convergence [29]. More methods are proposed to handle the stagnation of the convergence rate. This convex projection HIO method is explored further by Bauschke et al. with a hybrid projection–reflection (HPR) algorithm that uses Douglas-Rachford projections and specific nonnegativity and support constraints. Bauschke et al. increase the flexibility of the original algorithm by adding a relaxation parameter. A common performance improvement practice is by coupling methods together such as iterating the error-reduction algorithms after HIO [6].

The initialization process is further refined by Wyrowski and Bryngdahl to avoid stagnation as well as the speckle problem [63]. A sampling distance based on bandwidth is specified and after a sinc interpolation, a continuous distribution can be obtained in the Fourier domain resulting in a band-limit constraint. To avoid the remaining first-order zeros in amplitude that cannot be eliminated by the iterative Fourier transform algorithm, they formulated the initial distribution by first getting 1d samples from Fourier transform with consideration of band limit, and then using it to synthesize the 2d phase distribution and superpose the result phrase onto the desired domain. Because light is being spread within a finite area in the Fourier domain, the diffraction patterns could maintain speckle-free [52].

Fienup discussed the relationship between GS and gradient descent, and Wang et al. further modified the GS algorithm by incorporating gradient descent in the iterative steps by modulating the phase with an additional gradient directional term that is calculated by the difference between the current iteration and the previous iteration. Also, to prevent stagnation, they introduced weight terms and applied accordingly based on the root mean square error between two calculated patterns in consecutive iterations [61].

Wirtinger Flow (WF) is another non-convex gradient descent method in an iterative greedy refining fashion that can obtain a linear convergence rate with proper initialization for coded diffraction patterns [10]. Even though for most of its derivative methods, the convergence analysis is based on Gaussian measurements, it can achieve a relatively low computational cost and optimal sample complexity when the initial point is within a bound of the true underlying vector for Gaussian measurements [3]. Chen et al. gives a theoretical guarantee for WF method's general global convergence without a sophisticated sampling method during initialization or saddle-point escaping scheme [15].

However practically, WF methods' performance depends on initialization with outlier measurements and tuning, since its non-convex nature requires stronger conditions. Chen and Candès proposed a linear-time Truncated Wirtinger Flow algorithm in addition to their previous work with differences in the initialization step of regularizing the gradient flow with respect to iteration-varying subsets of data [14]. And Zhang et al. further resolved the outlier sensitivity issue in TWF by considering sample median instead of a sample mean in the truncation rule leading to a more robust response [69].

2.1.2 Bayesian Model

The GS algorithm starts with an initial random guess and imposes observed Fourier and temporal magnitude throughout the iteration, which analogously relates to the EM algorithm, and those iterative optimization algorithms can also be viewed from a Bayesian perspective. For the GS method, around four times of intensity measurements are required to necessarily reconstruct a signal from its linear measurements, while a probabilistic approach could reduce the sufficient measurement need to be lower. Schniter proposed adaptation of a generalized approximate message passing algorithm that applies to compressive phase retrieval, while the missing observation of phases modeled by a measurement likelihood function, and a signal prior sparsity reducing probability density function [54].

Dremeau and Krzakala expanded the probabilistic setup with complex-valued observations and marginalized maximum a posteriori estimates [20]. The goal is then minimizing the Kullback-Leibler divergence with variational mean-field approximation resulting in immediate and faster convergence, while the learning is done similarly by the variational Bayes EM algorithm. The probabilistic approach utilized the sparse structure to reduce sample complexities in recovering the phases, while approaches, like random initialization and restarting, are used to avoid the pitfall of bad local minima. To improve the robustness of random initialization, Liu et al. proposed a Gaussian distribution layer prior and a Wishart distribution layer hyperprior to ensure low-rankness [41].

2.1.3 Uniqueness and Off-Axis

In-line holography is easier to set up practically because it does not need extra optical elements to split the beam [45]. However, general inline phase retrieval is challenging due to zero-order diffractions and the symmetric structure of the mapping function. Even though the ambiguities brought by symmetry in rotation, translation, and conjugate reflection are considered to be trivial ambiguities, there are up to 2^{N-2} non-trivial solutions in one dimension where the resulting ambiguities can be calculated based on the zero sets of true roots of the autocorrelation polynomial [74]. Like how multiview stereo matching generally increases overall accuracy, additional information can be used to inference the unknown signal, as ambiguity cannot be resolved based on algebraic variety from lower dimension [74]. Bendory et al summarized the approaches that guarantee uniqueness from a theoretical perspective.

Wolf in 1962 discovered that some spectral emission functions with a blackbody radiator source might naturally have minimum-phase functions, while one needs to apply nonlinear filtering to ensure a function is in minimum phase, which is often an assumption for maximum entropy methods [6]. If zeros lie outside the unit circle, minimum phase constraint enforces the uniqueness, and it can be ensured by augmenting each complex signal to a minimum phase. One can also construct two independent masks or reconstruct from a special case of it, the phaseless STFT measurements, to ensure uniqueness, then the problem can be transformed into a sequence of original formulation with a non-vanishing constraint on the window size. To get the STFT model, a measurement technique X-FROG can be employed by introducing a reference window to gate the observed signal [74]. Since it is impossible to decide relative uniqueness a priori and stagnation often occurs due to the trivial ambiguities, Fanjiang uses random illumination under coded aperture imaging to ensure absolute uniqueness [21].

Inline phase retrieval ensures higher resolution without a need for large coherence lengths of light sources. A common practice is to use additional information from the corresponding off-axis holograms by applying a carrier wave modulation for in-line phase retrieval, firstly introduced by Leith and Upatnieks [39]. However, the reconstruction bandwidth is limited and it is sensitive to vibrations. Orzo overcomes the former issue with an integration of methods by combining off-axis low spatial frequency information with the GS inline retrieval process, in which the estimated wave field spectrum in the last step is replaced by the off-axis reconstruction data [45].

For the latter, Leith and Upatnieks mentioned in a later paper that with diffused illumination algorithms, for two-beam hologram, besides the simple dispersion encoding where each resolution element is recorded to occupy the entire plate, the diffused coherent illumination ensures the spatial-frequency to be greater than object transparency which introduces redundancy as well as insensitivity, with potential superimposing the two transparency-different diffraction patterns to a single hologram [39]. The artifact twin image from superposition can be eliminated by inverse filtering or some further modification of the GS algorithm. Liu and Scott initialized the GS algorithm with digitally reconstructed images and corrected input similar to HIO, and used the deconvolutional kernel to retrieve the phase, and get the field domain from the Fourier domain [41].

Off-axis setups come with a compromise of resolutions, while additional measuring methods can be introduced as a combination such as coherent aperture synthesis that reduces diffraction noise and therefore enhances resolution [38], or phase-diverse phase retrieval leading to a faster convergence with optimized diversity selection functions [17].

2.1.4 Sequences and Higher Dimensions

Unlike in 1d, multidimensional phase retrieval's non-trivial ambiguities rarely occur because most multivariate polynomials cannot be reduced into linear factors resulting in zero terms [74]. With a sequence of intensity measurements, one can obtain a complete wavefront reconstruction even without introducing a reference wave. The GS algorithm can infer phase information from the intensity from a known plane as well as additional information in another plane, and Pedrini et al. extend the additional information to be the intensity patterns of multiple positions instead of the overlapping with reference wave [48]. In this approach, the number of intensity patterns increases while the complexity decreases. Similar to some multiview stereo approaches, they move the sensor with small distance increments, and calculate the amplitude of the wavefront throughout the propagation that eventually adjusts the phase of the object in each step [48].

Rodenburg et al. demonstrated a technique that delivers a wavelength-limited resolution that is able to locate areas of interest without sophisticated camera performance requirements by collecting sequences of data. With a moving illumination of different but overlapping regions across the beam, Fraunhofer diffraction patterns can be collected to analyze intensity at different overlapping areas and eliminate ambiguity with oversampling [50].

Bao et al. further expanded this idea and resolved the 2pi phase modulo limitation by measuring sequences of the diffraction patterns under multiple wavelengths, resulting in a more robust algorithm [5]. To tackle the ill-posed phase retrieval problem, the previous algorithm either needs applying constraints or getting diffraction patterns at multiple planes, for Bao et al. they only need additional illumination wavelength information from tuning with the only constraint on predictable dispersions. The algorithm procedure is performed in a similar iterative fashion, but the initial guess is made based on the current wavelength's intensity and before propagating the wavefront to the recording plane, the phase is converted for the next wavelength with respect to phase retardance from the transmitting geometry of the object model [5].

Polygonal mesh with a collection of plane transformations can represent 3d Points, and Fast Fourier Transform can be used to get its diffraction patterns, so that there is no need to compute wave propagation for large point sets [12]. The previous GS algorithms are limited to 2d planes, while 3d samples can be recovered from two holograms or more. The transition from a Fourier transform of CDI to Fresnel regime in holography needs more flexibility to have different object-to-detector distances, while with different distance sample holograms, limitation on the sample distribution can be ignored as the intensity measures are outside of the sample domain [36]. Latychevskaia showed that GS-based iterative algorithms can be easily adapted with multiple intensity measurements, as a completely recovered complex-valued wavefront at one of the planes can propagate backward to recover the exit wave with thick samples [36]. Moreover, self-extrapolation can be applied when part of the hologram is missing during the iterative steps to enhance resolution with 3d samples [37].

2.2 Convex Relaxations and SDP

The methods above more or less suffer from a lack of understanding of gradient convergence due to the symmetries in the objective functions, and stagnation with respect to heuristics such as gradient descent and other minimization is persistently the main issue to solve [33]. Another approach to solving non-convex optimization problems of phase retrieval is to relax the problem to convex which is manageable to solve. Even though the relaxation gaps may lead to distortions of original formulation and large scale optimization is particularly difficult, the phase retrieval optimization's geometrical low-rank structure and other properties such as linearities in matrix transformation make this kind of approach challenging as well as fruitful. The methods that fall in this category try to develop convex optimization that is as efficient as non-convex approaches.

2.2.1 PhaseLift and PhaseCut

There exist linearities in the matrix transformation with respect to the feasibility problem of phase retrieval optimization. For methods like semidefinite programming (SDP) relaxation, the measurement model can be reformulated by a linear Hermitian low-rank matrix [74]. One important approach that deals with the invariant original Fourier mapping is to replace it with a linear operator that has simpler geometries. Even though the latter does not capture coded diffraction patterns, the method itself is much easier to solve.

Chai et al. minimized the rank of a reflective semidefinite decision matrix to recover the exact result of narrowband array imaging [11]. Rank minimization problem is further reduced into minimizing the nuclear norm in the semidefinite matrix with l1 norm minimization leading to a polynomial-time exact recovery of scatterer positions and reflectivities. This matrix completion method managed to reduce into trackable optimization with an assumption on random Gaussian measurements, but did not consider different noise models. Candès et al. used a similar relaxation approach that transforms combinatorial problems into convex, but also incorporating structured illuminations of collected diffraction patterns with the lifting step referring to a shift from quadratic constraints for vectors, into affine constraints for rank one matrices [9]. This systematic approach handles multiple statistical noise models by introducing a regularized maximum likelihood term as a penalty [10].

Compared to the SDP-based methods which twitch large unknown matrix dimensions, the iterative gradient descent methods seem to be more scalable and practical. Doelman et al. combine those two approaches by iteratively solving a sequence of convex relaxation subproblems, and within each iteration, alternating direction methods of multipliers are used to replace the expensive nuclear norm minimization. Via affine coefficient estimation, a sparse solution can be obtained with a better computational complexity compared to phaseLift [19].

Combining phaseLift and MaxCut relaxation and removing the rank constraint, Waldspurger et al. proposed the PhaseCut method that interprets phase retrieval as quadratic optimization on a unit complex torus resulting in a larger but simple SDP to solve. PhaseCut is more stable to noise scenarios than PhaseLift such that the initialization is within a certain size of neighborhood and perturbations remain in an explicit Dikin's ellipsoid so that the interior point methods within the torus topology would converge back to the solution in a full Newton iteration step [59]. There are some similarities between PhaseCut and Wirtinger flow that they all consider a smaller set of neighbor samples from spectral methods for initialization and perform low-rank estimations, but the lifting nature of PhaseCut enables a more accurate and efficient estimation of quadratic phaseless measurements with limited-scale random measurement models [3].

2.2.2 Scalability

Both the phaseCut and phaseLift methods come with guarantees of the recovery results since the problem is lifted into a higher dimensional space to deal with the rotational symmetric structure of the original mapping, and many approaches are done to resolve the scalability issue to solve large-scale SDP [57]. To minimize the nuclear norm and recover the low-rank matrix, we need to measure also the minimum sum of coefficients as a superposition, and then the atomic gauge function can be formed as convex relaxations. The gauge dual problem for lifting then becomes maximization of eigenvalue over a linear constraint [26]. Such formulation and simple constraint can be computationally efficient for developing dual-descent solvers on a large scale by applying projected first-order methods in each iteration [25]. However, there are flexibility issues for gauge duality solvers which require specifications on gauge functions, and

Aravkin et al. tackles this issue by introducing an equal footing between gauge duality and Fenchel-Rockafellar duality with sensitivity measures [2]. The symmetric properties that Gauge duality shares with Lagrange duality make it effective in phase recovery as a low-rank spatial optimization problem.

Sketching algorithms can also be used in this convex low-rank matrix optimization setting as ways to economize computational costs. Yurtsever et al. developed a numerical SketchyCGAL method to solve large SDPs while randomized sketches along with MaxCut are combined with a primal-dual optimization technique. Instead of storing the positive-semidefinite matrix variable from one coarse eigenvector computation, a matrix sketching technique is introduced to obtain a compressed representation of it, which will be extracted after optimization to retrieve a low-rank approximating solution [68].

Many more approaches have been made from a "non-lifting" relaxation perspective to ensure scalability. Bahmani and Romberg proposed relaxation in signal's natural domain by relaxing quadratic equations of phaseless measurements to inequality constraints with symmetric structure [3]. Goldstein and Studer proposed a PhaseMax method that is applied on the original dimension and performs a basis pursuit to its dual problem with a stronger condition assumption [28].

To avoid increasing the dimension with lifting, one can also use methods like re-weighted Wirtinger Flow. It is similar to PhaseMax that they all require special initializers that produce accurate vectors to begin with [60, 66, 67]. For the re-weighted WF algorithm, it searches the global minimum with respect to a series of sub phase retrieval problems that have changing weights. The RWF method is also based on a high-degree model, but with a lower sampling complexity from iteratively calculation of weight with least square metrics of previous inputs [67].

2.3 Maximum Entropy Methods

Entropy in a general sense can also be used to describe the highly textured versus more uniform regions, while maximum entropy methods (MEM) is a useful method for inverse problems that have insufficient information with assumptions that the maximized entropy should be consistent with different incomplete data inference [16, 66]. It wishes to find the most probable distribution with experimental data. Often, Lagrange multipliers are used to constraint the distribution. To facilitate the weak signal recovery in nonlinear optical spectroscopy and especially for coherent anti-Stokes Raman spectroscopy, quantitative spectral phase retrieval approaches such as MEM and Kramers-Kronig (KK) transforms are often used to transform frequency-domain data to the time domain to differentiate spectrally varying nonresonant background in CARS [16, 42]. Also, Saldin et al. demonstrated that MEM can be considered as a special form of object-domain operations for input-output phasing algorithms [53].

Without theoretical model knowledge, MEM can still get phase information via intensity spectrums with data from measured range, while the KK relations method requires extrapolated out-of-range measurement data. Incorporating MEM and polynomial interpolation with phase retrieval leads to a smooth error phase estimation [66].

3. Toward End to End Methods

Deep learning methods provide a new approach for phase retrieval with twin-image eliminating and self-interference-related spatial artifacts [48]. There are two main categories of models for the phase retrieval problem: the data-driven approach with end-to-end methods as a black box in solving a specific formulated mapping problem from large-scale data, and the model driven approach especially deep unfolding methods that construct network topology with respect to domain knowledge. The former is easier to set up with often satisfying results but may lack interpretability and generalizability, while the latter requires high computational complexity with limited performance guarantee [4].

3.1 Data Driven

Because of the densely connected nature of samples, Riverson et al. proposed a CNN that can be trained to learn the spatial features of real images based on features from twin images as well as other interference terms [48]. ResNet uses residual learning and skip connections to prevent gradient vanishing or exploding. Sinha et al. used the ResNet architecture but with the objective of acting as the non-linear inverse transform matrix allowing phase object inputs [56]. Nishizaki et al. also proposed a CNN for noisy measurements by feeding noise data to improve robustness. Specifically, they designed ResNets with one or multiple down and up sampling processes [44]. Metzler et al. integrated a DnCNN structure that acts as a denoiser with plug-and-play regularization in the traditional optimization algorithm to provide robustness to noises [62].

Ju et al. proposed a feature-based machine learning approach for phase retrieval. The problem is formulated as a non-linear fitting problem and features can be extracted by orthogonal Tchebichef moments from the point spread functions. With the features extraction, data can be compressed so that 2d intensity information is encoded in a 1d feature vector that will be used as input for the neural network to establish non-linear mapping to the wavefront aberration coefficients. Approaches like this, compared with traditional iterative approaches like GS algorithm, run faster with lower computational load. However, the accuracy of those feature based methods are typically lower [35].

CNN needs relatively little pre-processing of input but the training result seems to lack practical feasibility and generalizability with not particularly high accuracy on image-based wavefront sensing, so that a feature-based neural network proposed by Ju et al. could help address this. The data-driven Convolution Neural networks in general seem to solve the stagnation problem in iterative error reduction algorithms, but with lower accuracy.

When predicting sequential data from a machine learning perspective, it is natural to think about models like RNN, LSTM, and Transformers. Houhou et al. explored the LSTM approach by building a network that maps between the squared modulus and its imaginary component with artificially constructed data, with advantages on independence with non-resonant background [32]. Similarly, Xin et al. extracted object-independent features with complex pupil functions from focal plane images, and decomposed them into sequences of patches to feed into LSTM networks, which would output aberration coefficients via non-linear mapping [64]. Kim and Chung also used the LSTM network for phase retrieval that maps measurement vectors to support in the Fourier domain but has only an extended support estimation that outputs a relatively small set of indices for optimization [73].

The data driven methods' results are highly dependent on the input data distribution. Deng et al. pointed out that particularly for quantitative phase retrieval, the training database's spatial frequencies are underrepresented [18]. Even though one can try to improve the quality and resolution of results using ad hoc solutions and pre-amplifying the high spatial frequency data, unwanting distortion and artifacts might also appear. As a result, they proposed learning to synthesize (LS) methods that process data with different frequencies separately and will only be synthesized during the final round. The process is done by three DNNs to separately produce high-pass filtered outputs, and synthesize the final image, which makes the LS scheme perform well under high noise conditions [18].

3.2 Model Driven

Deep unfolding methods integrate iterative approaches with neural networks. Each iteration can be decomposed into a layer-wise structure by introducing trainable parameters [4]. And the resulting algorithms have more explainability and lower complexity than the data-driven methods.

Naimipour et al. presented a deep neural network where each layer will imitate one iteration of the iterative optimization unfolding the Incremental Reshaped Wirtinger Flow algorithm. As a result, only a small dataset is needed, and the L-layer network will be predicting the estimated results from the reformulated iterative optimization algorithm after L iterations, and specifically, it will be learning a fixed number of preconditioning matrices with a diagonal structure that accelerates the first-order gradient method [72].

To improve resolution and scalability, Wang et al. proposed a trainable decentralized generalized expectation consistent signal recovery network (deGEC-SR-Net). In their modular-based approach, the original GEC-SR algorithm is divided into three subtasks: a de-nonlinear process, an unbiased estimation of the linear transformation process, and a denoising process, and each task in one iteration is unfolded into a network layer of a deep neural network. To improve scalability, one can do data partitioning and split measurements into L independent clusters, do local inference, and then get a global estimation based on the local results [71].

For the two-deep unfolding setup above, the original iterative algorithm's mathematical formulation is kept, while the adjustable variables like regularization parameters are optimized through DNN to get faster convergence, greater accuracy, with a small training time and preserves interpretability. However, those approaches lack flexibility since iteration numbers are fixed, and often, the network is trained under task-specific settings which makes it not generalized well.

Wang et al., try to resolve the above issue by adopting strategies such as introducing a new hypernetwork unit to generate parameters for the original unfolding method. Also, to make the number of layers flexible, they trained the hypernetwork with RNN as a controller that observes convergence status and alters the damping factors in an online fashion. Specifically, LSTM with a self-attention function is used for generating optimal factors for signal recovery to ensure the adaptivity of their algorithm in different settings [21].

4. Trends and Conclusions

This survey provides an overview of phase retrieval methods over these years. Because of the rapid development in computing power, there are more methods involving deep phase retrieval and it is worthwhile revisiting previous approaches decades ago for new inspirations and adaptations with respect to current computational resources. There is also steady progress of non-convex optimization methods. From a theoretical perspective, there is still a large room for convergence analysis of some primal-dual schemes, Wirtinger flows with different thresholds, and deep unfolding methods. Also, the design of deep unfolding networks might be able to obtain inspiration from previous optimization progress on traditional iterative methods. For the convex relaxation approach, we might take a closer look at and utilize more of the rotational geometry and other symmetries. Also, maximization entropy methods might be explored further besides their main usage in spectroscopy-specific applications. Regarding the bayesian formation mentioned above, one might be able to experiment on adapting the existing framework into deep bayesian networks. Also, it might be interesting to use LSTM or Transformers with not only the passive decomposed sequences, but also the sequences of distance or wavelength increments mentioned in section 2.1.4.

5. Course Feedback

Throughout this quarter, I get a more in-depth view of various topics in computer vision and benefit a lot from the theoretical explanation of those methods during lectures as well as reading the assigned papers. I especially like how the lecture is structured. For example, for the visual object detection topic, introducing the classifier and discriminant function and Adaboost really helps in having a better understanding of those integrated methods such as DPM-based detection. I also especially enjoy the lecture about the l0 and l1 norms in sparse representation. For the project assignment, I benefited a lot from the process of collecting papers and formalized them into a survey, not just getting more information about the topic I choose, but the whole process of getting started for a new topic and how to search and organize relevant papers. However, I did a lot of unnecessary and redundant work at the beginning with little organization on the papers I've read. And at this moment, I wish that I could know the overall methodology of reading papers systematically earlier.

References

[1] I. P. Androulakis, C. D. Maranas, and C. A. Floudas. 1995. A global optimization method for general constrained nonconvex problems. *J Glob Optim* 7, 4 (December 1995), 337–363. DOI:https://doi.org/10.1007/BF01099647

[2] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and K. J. MacPhee. 2018. Foundations of Gauge and Perspective Duality. *SIAM J. Optim.* 28, 3 (January 2018), 2406–2434. DOI:https://doi.org/10.1137/17M1119020

[3] Sohail Bahmani and Justin Romberg. 2017. A flexible convex relaxation for phase retrieval. *Electron. J. Statist.* 11, 2 (2017), 5254–5281. DOI:https://doi.org/10.1214/17-EJS1378SI

[4] Alexios Balatsoukas-Stimming and Christoph Studer. 2019. Deep Unfolding for Communications Systems: A Survey and Some New Directions. In *2019 IEEE International Workshop on Signal Processing Systems (SiPS)*, IEEE, Nanjing, China, 266–271. DOI:https://doi.org/10.1109/SiPS47522.2019.9020494

[5] Peng Bao, Fucai Zhang, Giancarlo Pedrini, and Wolfgang Osten. 2008. Phase retrieval using multiple illumination wavelengths. *Opt. Lett.* 33, 4 (February 2008), 309. DOI:https://doi.org/10.1364/OL.33.000309

[6] Heinz H. Bauschke, Patrick L. Combettes, and D. Russell Luke. 2003. Hybrid projection–reflection method for phase retrieval. *J. Opt. Soc. Am. A* 20, 6 (June 2003), 1025. DOI:https://doi.org/10.1364/JOSAA.20.001025

[7] Ralf Bräuer, Frank Wyrowski, and Olof Bryngdahl. 1991. Diffusers in digital holography. *J. Opt. Soc. Am. A* 8, 3 (March 1991), 572. DOI:https://doi.org/10.1364/JOSAA.8.000572

[8] B. R. Brown and A. W. Lohmann. 1966. Complex Spatial Filtering with Binary Masks. *Appl. Opt.* 5, 6 (June 1966), 967. DOI:https://doi.org/10.1364/AO.5.000967

[9] Emmanuel J. Candès, Yonina C. Eldar, Thomas Strohmer, and Vladislav Voroninski. 2015. Phase Retrieval via Matrix Completion. *SIAM Rev.* 57, 2 (January 2015), 225–251. DOI:https://doi.org/10.1137/151005099

[10] Emmanuel J. Candes, Xiaodong Li, and Mahdi Soltanolkotabi. 2015. Phase Retrieval via Wirtinger Flow: Theory and Algorithms. *IEEE Trans. Inform. Theory* 61, 4 (April 2015), 1985–2007. DOI:https://doi.org/10.1109/TIT.2015.2399924

[11] Anwei Chai, Miguel Moscoso, and George Papanicolaou. 2011. Array imaging using intensity-only measurements. *Inverse Problems* 27, 1 (January 2011), 015005. DOI:https://doi.org/10.1088/0266-5611/27/1/015005

[12] Praneeth Chakravarthula, Ethan Tseng, Tarun Srivastava, Henry Fuchs, and Felix Heide. 2020. Learned hardware-in-the-loop phase retrieval for holographic near-eye displays. *ACM Trans. Graph.* 39, 6 (November 2020), 1–18. DOI:https://doi.org/10.1145/3414685.3417846

[13] Xiaojun Chen. 2012. Smoothing methods for nonsmooth, nonconvex minimization. *Math. Program.* 134, 1 (August 2012), 71–99. DOI:https://doi.org/10.1007/s10107-012-0569-0

[14]Yuxin Chen and Emmanuel J. Candès. 2017. Solving Random Quadratic Systems of Equations Is Nearly as Easy as Solving Linear Systems. *Comm. Pure Appl. Math.* 70, 5 (May 2017), 822–883. DOI:https://doi.org/10.1002/cpa.21638

[15] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. 2019. Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Math. Program.* 176, 1–2 (July 2019), 5–37. DOI:https://doi.org/10.1007/s10107-019-01363-6

[16] Marcus T. Cicerone, Khaled A. Aamer, Young Jong Lee, and Erik Vartiainen. 2012. Maximum entropy and time-domain Kramers-Kronig phase retrieval approaches are functionally equivalent for CARS microspectroscopy: Maximum entropy and TDKK phase retrieval approaches are functionally equivalent for CARS microspectroscopy. *J. Raman Spectrosc.* 43, 5 (May 2012), 637–643. DOI:https://doi.org/10.1002/jrs.3169

[17] Bruce H. Dean and Charles W. Bowers. 2003. Diversity selection for phase-diverse phase retrieval. *J. Opt. Soc. Am. A* 20, 8 (August 2003), 1490. DOI:https://doi.org/10.1364/JOSAA.20.001490

[18] Mo Deng, Shuai Li, Alexandre Goy, Iksung Kang, and George Barbastathis. 2020. Learning to synthesize: robust phase retrieval at low photon counts. *Light Sci Appl* 9, 1 (December 2020), 36. DOI:https://doi.org/10.1038/s41377-020-0267-2

[19] Reinier Doelman, Nguyen H. Thao, and Michel Verhaegen. 2018. Solving large-scale general phase retrieval problems via a sequence of convex relaxations. *J. Opt. Soc. Am. A* 35, 8 (August 2018), 1410. DOI:https://doi.org/10.1364/JOSAA.35.001410

[20] Angelique Dremeau and Florent Krzakala. 2015. Phase recovery from a Bayesian point of view: The variational approach. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, South Brisbane, Queensland, Australia, 3661–3665. DOI:https://doi.org/10.1109/ICASSP.2015.7178654

[21] Albert Fannjiang. 2012. Absolute uniqueness of phase retrieval with random illumination. *Inverse Problems* 28, 7 (July 2012), 075008. DOI:https://doi.org/10.1088/0266-5611/28/7/075008

[22] M.A. Fiddy and U. Shahid. 2013. Legacies of the Gerchberg–Saxton algorithm. *Ultramicroscopy* 134, (November 2013), 48–54. DOI:https://doi.org/10.1016/j.ultramic.2013.05.009

[23] J. R. Fienup. 1982. Phase retrieval algorithms: a comparison. *Appl. Opt.* 21, 15 (August 1982), 2758. DOI:https://doi.org/10.1364/AO.21.002758

[24]James R. Fienup. 2013. Phase retrieval algorithms: a personal tour [Invited]. *Appl. Opt.* 52, 1 (January 2013), 45. DOI:https://doi.org/10.1364/AO.52.000045

[25] Michael P. Friedlander and Ives Macêdo. 2016. Low-Rank Spectral Optimization via Gauge Duality. *SIAM J. Sci. Comput.* 38, 3 (January 2016), A1616–A1638. DOI:https://doi.org/10.1137/15M1034283

[26] Michael P. Friedlander, Ives Macêdo, and Ting Kei Pong. 2014. Gauge Optimization and Duality. *SIAM J. Optim.* 24, 4 (January 2014), 1999–2022. DOI:https://doi.org/10.1137/130940785

[27] D. Gabor. 1948. A New Microscopic Principle. *Nature* 161, 4098 (May 1948), 777–778. DOI:https://doi.org/10.1038/161777a0

[28] Tom Goldstein and Christoph Studer. 2018. PhaseMax: Convex Phase Retrieval via Basis Pursuit. *IEEE Trans. Inform. Theory* 64, 4 (April 2018), 2675–2689. DOI:https://doi.org/10.1109/TIT.2018.2800768

[29] Joseph W. Goodman. 2005. *Introduction to Fourier optics* (3rd ed ed.). Roberts & Co, Englewood, Colo.

[30] Philipp Grohs, Sarah Koppensteiner, and Martin Rathmair. 2020. Phase Retrieval: Uniqueness and Stability. *SIAM Rev.* 62, 2 (January 2020), 301–350. DOI:https://doi.org/10.1137/19M1256865

[31] Min Gu, Xinyuan Fang, Haoran Ren, and Elena Goi. 2019. Optically Digitalized Holography: A Perspective for All-Optical Machine Learning. *Engineering* 5, 3 (June 2019), 363–365. DOI:https://doi.org/10.1016/j.eng.2019.04.002

[32] Rola Houhou, Parijat Barman, Micheal Schmitt, Tobias Meyer, Juergen Popp, and Thomas Bocklitz. 2020. Deep learning as phase retrieval tool for CARS spectra. *Opt. Express* 28, 14 (July 2020), 21002. DOI:https://doi.org/10.1364/OE.390413

[33]Prateek Jain and Purshottam Kar. 2017. *Non-convex optimization for machine learning*. Now, Boston.

[34] Bahram Javidi and Chung-Jung Kuo. 1988. Joint transform image correlation using a binary spatial light modulator at the Fourier plane. *Appl. Opt.* 27, 4 (February 1988), 663. DOI:https://doi.org/10.1364/AO.27.000663

[35] Guohao Ju, Xin Qi, Hongcai Ma, and Changxiang Yan. 2018. Feature-based phase retrieval wavefront sensing approach using machine learning. *Opt. Express* 26, 24 (November 2018), 31767. DOI:https://doi.org/10.1364/OE.26.031767

[36] Tatiana Latychevskaia. 2020. Iterative phase retrieval for digital holography: tutorial: publisher's note. *J. Opt. Soc. Am. A* 37, 1 (January 2020), 45. DOI:https://doi.org/10.1364/JOSAA.37.000045

[37] Tatiana Latychevskaia and Hans-Werner Fink. 2013. Resolution enhancement in digital holography by self-extrapolation of holograms. *Opt. Express* 21, 6 (March 2013), 7726. DOI:https://doi.org/10.1364/OE.21.007726

[38] Dennis J. Lee and Andrew M. Weiner. 2014. Optical phase imaging using a synthetic aperture phase retrieval technique. *Opt. Express* 22, 8 (April 2014), 9380. DOI:https://doi.org/10.1364/OE.22.009380

[39] Emmett N. Leith and Juris Upatnieks. 1962. Reconstructed Wavefronts and Communication Theory*. *J. Opt. Soc. Am.* 52, 10 (October 1962), 1123. DOI:https://doi.org/10.1364/JOSA.52.001123

[40] G. Liu and P. D. Scott. 1987. Phase retrieval and twin-image elimination for in-line Fresnel holograms. *J. Opt. Soc. Am. A* 4, 1 (January 1987), 159. DOI:https://doi.org/10.1364/JOSAA.4.000159

[41] Kaihui Liu, Jiayi Wang, Zhengli Xing, Linxiao Yang, and Jun Fang. 2019. Low-Rank Phase Retrieval via Variational Bayesian Learning. *IEEE Access* 7, (2019), 5642–5648. DOI:https://doi.org/10.1109/ACCESS.2018.2889518

[42] Yuexin Liu, Young Jong Lee, and Marcus T. Cicerone. 2009. Broadband CARS spectral phase retrieval using a time-domain Kramers–Kronig transform. *Opt. Lett.* 34, 9 (May 2009), 1363. DOI:https://doi.org/10.1364/OL.34.001363

[43] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. 2015. Phase Retrieval Using Alternating Minimization. *IEEE Trans. Signal Process.* 63, 18 (September 2015), 4814–4826. DOI:https://doi.org/10.1109/TSP.2015.2448516

[44] Yohei Nishizaki, Ryoichi Horisaki, Katsuhisa Kitaguchi, Mamoru Saito, and Jun Tanida. 2020. Analysis of non-iterative phase retrieval based on machine learning. *Opt Rev* 27, 1 (February 2020), 136–141. DOI:https://doi.org/10.1007/s10043-019-00574-8

[45] László Orzó. 2015. High speed phase retrieval of in-line holograms by the assistance of corresponding off-axis holograms. *Opt. Express* 23, 13 (June 2015), 16638. DOI:https://doi.org/10.1364/OE.23.016638

[46] Scott W. Paine and James R. Fienup. 2018. Machine learning for improved image-based wavefront sensing. *Opt. Lett.* 43, 6 (March 2018), 1235. DOI:https://doi.org/10.1364/OL.43.001235

[47] Ramtin Pedarsani, Dong Yin, Kangwook Lee, and Kannan Ramchandran. 2017. PhaseCode: Fast and Efficient Compressive Phase Retrieval Based on Sparse-Graph Codes. *IEEE Trans. Inform. Theory* 63, 6 (June 2017), 3663–3691. DOI:https://doi.org/10.1109/TIT.2017.2693287

[48] Giancarlo Pedrini, Wolfgang Osten, and Yan Zhang. 2005. Wave-front reconstruction from a sequence of interferograms recorded at different planes. *Opt. Lett.* 30, 8 (April 2005), 833. DOI:https://doi.org/10.1364/OL.30.000833

[49] Yair Rivenson, Yibo Zhang, Harun Günaydın, Da Teng, and Aydogan Ozcan. 2018. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light Sci Appl* 7, 2 (February 2018), 17141–17141. DOI:https://doi.org/10.1038/lsa.2017.141

[50] J. M. Rodenburg, A. C. Hurst, A. G. Cullis, B. R. Dobson, F. Pfeiffer, O. Bunk, C. David, K. Jefimovs, and I. Johnson. 2007. Hard-X-Ray Lensless Imaging of Extended Objects. *Phys. Rev. Lett.* 98, 3 (January 2007), 034801. DOI:https://doi.org/10.1103/PhysRevLett.98.034801

[51] Francesca Rossi, Peter Van Beek, and Toby Walsh (Eds.). 2006. *Handbook of constraint programming* (1st ed ed.). Elsevier, Amsterdam ; Boston.

[52] Erdem Sahin, Elena Stoykova, Jani Mäkinen, and Atanas Gotchev. 2020. Computer-Generated Holograms for 3D Imaging: A Survey. *ACM Comput. Surv.* 53, 2 (July 2020), 1–35. DOI:https://doi.org/10.1145/3378444

[53] D K Saldin, R J Harder, V L Shneerson, and W Moritz. 2001. Phase retrieval methods for surface x-ray diffraction. *J. Phys.: Condens. Matter* 13, 47 (November 2001), 10689–10707. DOI:https://doi.org/10.1088/0953-8984/13/47/311

[54] Philip Schniter and Rangan. 2014. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing 63.4 (2014): 1043-1055* (2014).

[55] Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. 2015. Phase Retrieval with Application to Optical Imaging: A contemporary overview. *IEEE Signal Process. Mag.* 32, 3 (May 2015), 87–109. DOI:https://doi.org/10.1109/MSP.2014.2352673

[56] Ayan Sinha, Justin Lee, Shuai Li, and George Barbastathis. 2017. Lensless computational imaging through deep learning. *Optica* 4, 9 (September 2017), 1117. DOI:https://doi.org/10.1364/OPTICA.4.001117

[57] Ju Sun, Qing Qu, and John Wright. 2018. A Geometric Analysis of Phase Retrieval. *Found Comput Math* 18, 5 (October 2018), 1131–1198. DOI:https://doi.org/10.1007/s10208-017-9365-9

[58] Joel A. Tropp. 2017. Sketchy decisions: convex optimization with optimal storage (Conference Presentation). In *Wavelets and Sparsity XVII*, SPIE, San Diego, United States, 75. DOI:https://doi.org/10.1117/12.2281058

[59] Irène Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. 2015. Phase recovery, MaxCut and complex semidefinite programming. *Math. Program.* 149, 1–2 (February 2015), 47–81. DOI:https://doi.org/10.1007/s10107-013-0738-9

[60] Chang-Jen Wang, Chao-Kai Wen, Shang-Ho Tsai, and Shi Jin. 2020. Decentralized Expectation Consistent Signal Recovery for Phase Retrieval. *IEEE Trans. Signal Process.* 68, (2020), 1484–1499. DOI:https://doi.org/10.1109/TSP.2020.2974711

[61] Haichao Wang, Weirui Yue, Qiang Song, Jingdan Liu, and Guohai Situ. 2017. A hybrid Gerchberg–Saxton-like algorithm for DOE and CGH calculation. *Optics and Lasers in Engineering* 89, (February 2017), 109–115. DOI:https://doi.org/10.1016/j.optlaseng.2016.04.005

[62] Daniel S. Weller. 2016. Robust phase retrieval with sparsity under nonnegativity constraints. In *2016 50th Asilomar Conference on Signals, Systems and Computers*, IEEE, Pacific Grove, CA, USA, 1043–1047. DOI:https://doi.org/10.1109/ACSSC.2016.7869528

[63] Frank Wyrowski and Olof Bryngdahl. 1988. Iterative Fourier-transform algorithm applied to computer holography. *J. Opt. Soc. Am. A* 5, 7 (July 1988), 1058. DOI:https://doi.org/10.1364/JOSAA.5.001058

[64] Qi Xin, Guohao Ju, Chunyue Zhang, and Shuyan Xu. 2019. Object-independent image-based wavefront sensing approach using phase diversity images and deep learning. *Opt. Express* 27, 18 (September 2019), 26102. DOI:https://doi.org/10.1364/OE.27.026102

[65] Cheng Yang, Risheng Liu, Long Ma, Xin Fan, Haojie Li, and Miao Zhang. 2018. Unrolled Optimization with Deep Priors for Intrinsic Image Decomposition. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, IEEE, Xi'an, 1–7. DOI:https://doi.org/10.1109/BigMM.2018.8499478

[66] Pao-Keng Yang and Jung Y. Huang. 1997. Phase-retrieval problems in infrared–visible sum-frequency generation spectroscopy by the maximum-entropy method. *J. Opt. Soc. Am. B* 14, 10 (October 1997), 2443. DOI:https://doi.org/10.1364/JOSAB.14.002443

[67] Ziyang Yuan and Hongxia Wang. 2017. Phase retrieval via reweighted Wirtinger flow. *Appl. Opt.* 56, 9 (March 2017), 2418. DOI:https://doi.org/10.1364/AO.56.002418

[68] Alp Yurtsever, Joel A. Tropp, Olivier Fercoq, Madeleine Udell, and Volkan Cevher. 2021. Scalable Semidefinite Programming. *SIAM Journal on Mathematics of Data Science* 3, 1 (January 2021), 171–200. DOI:https://doi.org/10.1137/19M1305045

[69] Huishuai Zhang, Yuejie Chi, and Yingbin Liang. 2018. Median-Truncated Nonconvex Approach for Phase Retrieval With Outliers. *IEEE Trans. Inform. Theory* 64, 11 (November 2018), 7287–7310. DOI:https://doi.org/10.1109/TIT.2018.2847695

[70] Quanbing Zhang, Feihang Hu, Dequn Liu, and Yufan Yuan. 2020. Phase retrieval with outliers via median truncated amplitude flow. In *Proceedings of the 5th International Conference on Multimedia and Image Processing*, ACM, Nanjing China, 115–118. DOI:https://doi.org/10.1145/3381271.3381287

[71] C. J. Wang, C. K. Wen, S. H. Tsai, and S. Jin. 2020. Phase retrieval with learning unfolded expectation consistent signal recovery algorithm. *IEEE Signal Process*. Lett., vol. 27, pp. 780–784. DOI:https://doi.org/10.1109/LSP.2019.2935814

[72] Naveed Naimipour, Shahin Khobahi, and Mojtaba Soltanalian. 2020. Upr: A model-driven architecture for deep phase retrieval. arXiv preprint arXiv:2003.04396 (2020).

[73] Kyung-Su Kim, and Sae-Young Chung. 2019. Fourier phase retrieval with extended support estimation via deep neural network. *IEEE Signal Processing.* Letters 26.10 (2019): 1506-1510. DOI:https://doi.org/110.1109/LSP.2019.2935814

[74] Tamir Bendory, Robert Beinert, and Yonina C. Eldar. 2017. Fourier phase retrieval: Uniqueness and algorithms. *Compressed Sensing and its Applications*. Birkhäuser, Cham, 2017. 55-91.