Deep Learning for Glaucoma Diagnosis and Monitoring and for Video Processing

DISSERTATION

Submitted in Partial Fulfillment of

the Requirements for

the Degree of

DOCTOR OF PHILOSOPHY (Electrical Engineering)

at the

NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

by

Zhiqi Chen

January 2024

Approved:

Ivan Selesnick

Department Chair Signature

Dec 30, 2023

Date

University ID: N12463724

Net ID: zc1337

Approved by the Guidance Committee:

Major: Electrical Engineering

Wang Yao

Yao Wang
Professor
NYU Tandon School of Engineering

12/29/2023
Date

Joel S Schuman

Joel S. Schuman
Professor
Wills Eye Hospital

12-22-2023
Date

Anna Choromanska
Assistant Professor
NYU Tandon School of Engineering

12/22/2023
Date

Hiroshi Ishikawa
Professor
OHSU School of Medicine

12/29/2023
Date

G. Wollstein

Gadi Wollstein
Professor
NYU Grossman School of Medicine

12/26/2023
Date

Microfilm or copies of this dissertation may be obtained from

UMI Dissertation Publishing
ProQuest CSA
789 E. Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Vita

Zhiqi Chen was born in Putian, Fujian, China, in 1997. She received her B.S. degree in Biomedical Engineering from Beihang University (previously known as the Beijing University of Aeronautics and Astronautics) in Beijing, China, in 2018. She then started her doctoral training at the Tandon School of Engineering of New York University in Fall 2018. Her Ph.D. research focuses on computer vision, deep learning, and its applications in medical image analysis, especially video processing and retinal Optical Coherence Tomography image understanding in the context of glaucoma. During her PhD years, she also interned at Philips, Amazon Robotics, and Meta Reality Lab on various research projects related to machine learning and computer vision.

Acknowledgements

During the remarkable journey of pursuing my Ph.D., I find myself profoundly thankful for the abundant support I've received, both academically and emotionally, from a multitude of individuals. Without their unwavering assistance, the completion of this dissertation would have been an insurmountable task.

Foremost among these contributors are my esteemed advisors, Dr. Joel S. Schuman and Prof. Yao Wang. Throughout the years, their dedication and support have been instrumental in shaping every facet of my research. Dr. Schuman's generous guidance illuminated the field of Ophthalmology for me, despite my initial lack of experience. Prof. Wang's motivation, extensive knowledge, and unwavering patience have also played a pivotal role in my academic endeavors. Additionally, I extend my heartfelt gratitude to Dr. Hiroshi Ishikawa, my co-advisor during the first two years of my Ph.D. training, for his invaluable insights and mentorship. Their collective support has been indispensable, making the realization of this dissertation and its associated work possible. I extend my sincere thanks to the remaining members of my dissertation committee, Dr. Gadi Wollstein and Prof. Anna Choromanska, for their insightful comments and constructive suggestions, which have significantly enriched the quality of my work.

Special appreciation is also reserved for all the current and former members of the Advanced Ophthalmic Imaging Lab at NYU Langone and Video Lab at NYU. Their collaboration and assistance have played a pivotal role in various aspects of my research and doctoral life.

Lastly, I extend heartfelt thanks to my family and friends, both in the United States and China, whose unwavering presence sustained me throughout these prolonged years. Special gratitude is reserved for Dr. Jing Zhang, future Dr. Feiran Yang, future Dr. Chang Liu, and future Dr. Weihao Xu for their generous emotional support. This achievement would have remained a distant goal without their invaluable contributions to my journey.

Zhiqi Chen

December 18, 2023

ABSTRACT

Deep Learning for Glaucoma Diagnosis and Monitoring and for Video Processing

by

Zhiqi Chen

Advisors: Prof. Yao Wang, Dr. Joel S. Schuman

Submitted in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy (Electrical Engineering)

January 2024

Deep learning (DL)'s success in computer vision tasks, driven by robust parallel computing and extensive data, has expanded to higher-dimensional data like 3-dimensional (3D) medical images and spatiotemporal sequences. This thesis focuses on employing DL algorithms for two critical vision tasks: comprehending 3D volumetric retinal optical coherence tomography (OCT), especially in glaucoma applications, and understanding spatiotemporal sequences, including longitudinal glaucomatous retinal image prediction and natural video processing.

Glaucoma, a significant cause of global blindness, involves the gradual and irreversible loss of retinal ganglion cells and their axons, leading to functional defects. The structural changes observed through OCT and functional abnormality measured by standard automated perimetry (SAP) are crucial for comprehensive glaucoma diagnosis. However, SAP's subjectivity and susceptibility to fluctuations pose challenges, while OCT exhibits excellent reproducibility. To bridge this gap, a 3D Convolutional Neural Network (CNN) is proposed to estimate point-wise visual field (VF) sensitivities from segmentation-free 3D OCT volumes, outperforming its two-dimensional (2D) counterpart relying on segmentation-dependent 2D OCT thickness maps. This innovative 3D model enhances the understanding of structure-function relationships, overcoming limitations of prior segmentation-dependent measurements brought by the minimal measurable level (floor effect), and potentially aiding in VF surrogate derivation without SAP's inherent limitations.

With a robust model capturing the interplay between structural and functional measurements without domain knowledge or segmentations, this research opens avenues for discovering unexpected anatomical or structural features highly associated with function through model visualization. The establishment of a general point-wise spatial mapping between structure and function, as demonstrated by occlusion analysis on the above 3D CNN model, provides insights consistent with manually derived maps. This approach enables the exploration of new findings from machine learning models, potentially offering robust and unbiased insights. Given the diverse progression patterns in glaucoma, the detailed spatial correlation map holds promise for identifying personalized progression patterns, improving assessment, and enhancing the forecasting of progression.

Monitoring glaucoma progression is crucial for effective clinical management, where ganglion cell inner plexiform layer (GCIPL) thickness serves as a critical biomarker. Despite deriving from 3D OCT scans, the conventional clinical progression analysis uses summarized GCIPL measurements, neglecting its 2D nature and potentially missing subtle changes and spatial patterns. To address this limitation, this thesis introduces a Time-aware Convolutional Long Short-Term Memory (TC-LSTM) model. This novel model predicts future 2D GCIPL thickness maps by leveraging spatial and temporal correlations in irregularly sampled longitudinal sequences. Experimental results demonstrate the superiority of the proposed TC-LSTM over traditional methods.

The final part of this thesis explores another example of spatiotemporal sequences, natural videos, and develops an efficient video frame interpolation algorithm, Pyramid Deformable Warping Network (PDWN). By integrating a pyramid structure and deformable convolution in its design, PDWN effectively merges the advantages of optical flow and kernel methods, surpassing state-of-the-art models in accuracy across various datasets for video interpolation, while reducing the model parameters and inference time.

Contents

List of Figures

List of Tables

List of Abbreviations

2D two dimensional.

3D three dimensional.

BMO Bruch's membrane opening.

cLSTM Convolutional Long Short-Term Memory.

CNN Convolutional Neural Network.

DConv deformable convolution.

DL deep learning.

GCIPL Ganglion Cell–Inner Plexiform layer.

HRT Heidelberg retina tomography.

IE interpolation error.

LR linear regression.

LSTM Long Short-Term Memory.

MAE mean absolute error.

MD mean deviation.

MSE mean square error.

OCT optical coherence tomography.

ONH optic nerve head.

PCA principle component analysis.

PDWN Pyramid Deformable Warping Network.

PSNR peak signal noise ratio.

RMSE root mean square error.

RNFL retinal nerve fiber layer.

RNN Recurrent Neural Network.

SAP standard automated perimetry.

SD-OCT spectral-domain OCT.

SS-OCT swept-source OCT.

SSIM structure similarity index.

TC-LSTM time-aware Convolutional Long Short-Term Memory.

VF visual field.

VFI visual field index.

Chapter 1

Introduction

1.1 Background

This dissertation delves into the realm of employing deep learning (DL) algorithms for critical vision tasks in light of recent advancements in high-performance parallel computing and the availability of big data. DL has emerged as a powerhouse in computer vision, demonstrating remarkable success across various tasks like classification, segmentation, and detection. As attention turns toward higher dimensional data, particularly spatiotemporal sequences and three dimensional (3D) medical images, their relevance to real-world applications becomes increasingly apparent. This dissertation focuses on employing DL algorithms for two critical vision tasks: comprehending 3D volumetric retinal optical coherence tomography (OCT), especially in glaucoma applications, and understanding spatiotemporal sequences, including longitudinal glaucomatous retinal image prediction and natural video interpolation.

Glaucoma, the second leading cause of blindness, is characterized by the progressive loss of retinal ganglion cells and their axons [85, 86, 88, 91, 104, 106]. This degeneration may lead to changes in the optic nerve head (ONH) and retinal nerve fiber layer (RNFL), eventually resulting in vision impairment and irreversible blindness. Early diagnosis and meticulous progression monitoring are crucial for effective glaucoma management. The integration of DL into ophthalmology has marked a significant shift due to its capacity to extract pertinent features from complex, high-dimensional data [15–17,19,20,24,37,58,70, 87,107]. This infusion of DL has the potential to revolutionize glaucoma diagnosis, management, and understanding by interpreting structural and functional information, identifying phenotypes, and deciphering progression patterns.

OCT stands as a non-invasive imaging technique providing micrometer resolution cross-sectional and volumetric retinal images. The scheme of the OCT scanner is shown in Figure 1.1. It has become the de facto standard for objectively quantifying structural damage in glaucoma. Commercially available Cirrus OCT (spectral-domain OCT (SD-OCT)) acquires OCT data with 5 μm axial resolution in tissue. It allows for visualization of 3D retinal structure and quantitative assessment of RNFL thickness within a very short time span at a single visit. Figure 1.2 shows examples of Cirrus OCT reports. Early exploration into DL for glaucoma applications has revolved around segmentation algorithms applied to OCT volumes to derive standard measurements like RNFL or macular Ganglion Cell–Inner Plexiform layer (GCIPL) thickness [21, 69, 94, 102, 105, 109]. However, these algorithms, while accurate, are limited by relying on segmentation, making them prone to errors and neglecting potential 3D features related to the disease. Consequently, the quest continues for more intricate, segmentation-free methods capable of handling volumetric data effectively.

Apart from structural measurements, functional abnormality measurements are critical for comprehensive glaucoma diagnosis and management. Figure 1.3 shows examples of visual field (VF) reports generated by standard automated perimetry (SAP). Understanding the relationship between structural loss and functional impairment in glaucoma remains a topic of debate [27, 30, 46, 50, 54, 66, 84, 99, 108, 126]. DL's success in identifying and predicting glaucoma progression holds promise in unraveling this complex relationship. Moreover, given the subjectivity, time consumption, and noise associated with VF tests, accurate estimation of VF from OCT data could potentially reduce unnecessary testing in

Figure 1.1. Schematic diagram of imaging optics comprising the OCT scanner and an example of a tomographic image of the retina obtained along the papillomacular axis. Cross-sectional images of optical reflectivity v.s. depth are created in a manner similar to ultrasound scans. The axial resolution is determined by the coherence length of the superluminescent diode source and is 10 μm (full width at half-maximum) in the retina. (images created by and borrowed from [44])

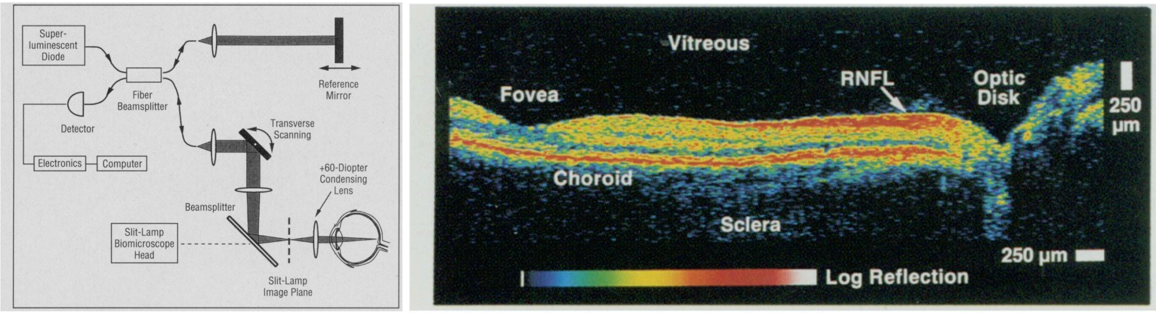
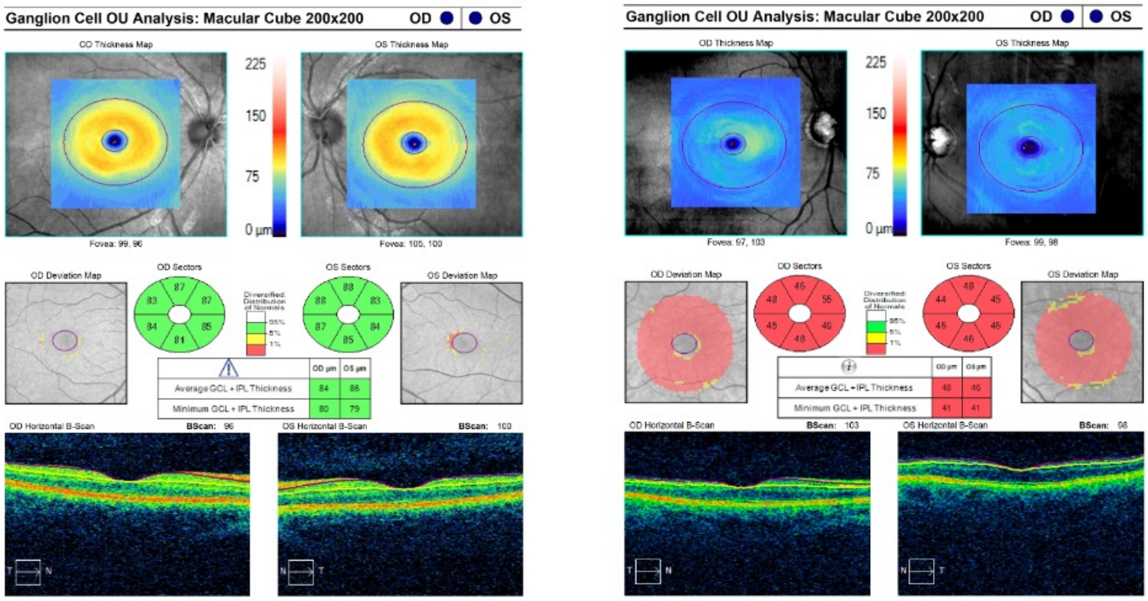


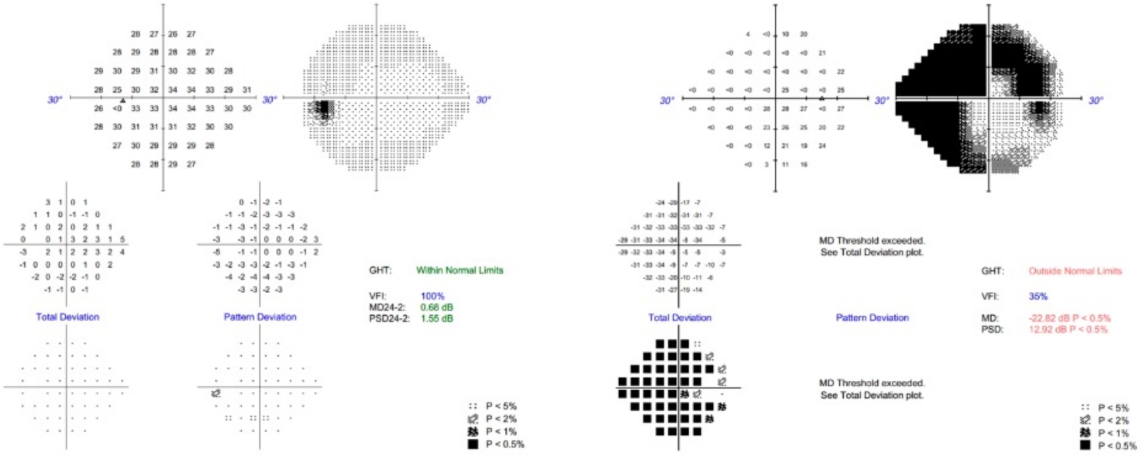
Figure 1.2. Examples of Cirrus OCT report from a healthy case (a) and a glaucomatous case (b). The image is color-coded (red, orange, and yellow represent thicker areas while green and blue represent thinner areas).



(a) A healthy case

(b) A glaucomatous case

Figure 1.3. Examples of Humphrey 24-2 VF report from a healthy eye (a) and a glaucomatous eye (b). Figure (b) shows advanced visual field damages with superior hemifield and nasal step.



(a) A healthy eye

(b) A glaucomatous eye

stable eyes.

The accelerated loss of retinal ganglion cells characterizes glaucoma progression alongside functional damages. Identifying and estimating the rate of this progression, whether structurally or functionally, remain crucial in managing glaucoma. While current clinical standards like Guided Progression Analysis enable clinicians to detect progression and evaluate losses compared to baselines, they do not forecast future progressions [61, 72]. DL's application in predictive medicine presents an intriguing avenue for glaucoma management, particularly in predicting future findings, an area that has seen limited investigation so far [9, 13, 80, 94, 99, 111, 115]. The high-dimensional feature space of retinal progression falls within the realm of understanding spatiotemporal sequences. The development of tailored DL algorithms for this specific application presents a considerable challenge.

In addition to medical sequences, natural videos present another form of spatiotemporal sequences ubiquitous in daily life. Video processing forms the backbone of multimedia, and spatiotemporal modeling plays a pivotal role in advancing video processing technologies. Unlike medical sequences, natural videos encompass rigid and non-rigid deformations and possess more diverse appearances with uniform, frequent sampling rates. Tailoring DL algorithms to meet the distinct demands of video processing, characterized by these varying attributes, poses a significant challenge in this domain.

1.2 Problem statement

This dissertation is primarily focused on two vision tasks:

- **3D volumetric OCT analysis in relating structure and function of glaucomatous eyes.** Analyzing unprocessed 3D OCT volumes circumvents segmentation procedures needed to derive clinical features like retinal layer thicknesses and rim volume. This approach is free from segmentation errors and offers richer information compared to summarized segmentation measurements.

  - **Segmentation-free point-wise VF estimation**. The structural changes measured by OCT are closely related to functional changes in VF. Prior studies in characterizing structural-functional relationships have focused on correlating VF outcomes with segmentation-dependent OCT measurements, such as RNFL thickness. However, segmentation-dependent OCT measurements have floor effects that will affect the structural-functional relationship learning for patients with advanced disease. To overcome the floor effect of segmentation-dependent OCT measurements, we aim to develop a 3D model to estimate the functional deterioration directly from segmentation-free 3D OCT volumes and compare it to the model trained with segmentation-dependent two dimensional (2D) OCT thickness maps.

  - **Generalized point-wise spatial mapping of structure to function**. With a robust model capturing the interplay between structural and functional measurements without domain knowledge or segmentations, it is possible to discover unexpected anatomical or structural features highly associated with function through model visualization. Thus, in this study, we aim to establish a generalized point-wise spatial mapping between structure and function by conducting occlusion analysis on a DL model trained on an extensive clinical cohort of patients to predict point-wise VF sensitivities from 3D OCT volumes.

- **Spatiotemporal sequence generation** aims to generate intermediate or future time points given a spatiotemporal sequence. Here, we addressed two specific problems in two different domains:

  - **Retinal GCIPL thickness map prediction**. GCIPL thickness is an important biomarker for the clinical management of glaucoma. Clinical analysis of GCIPL progression uses averaged thickness only, which easily washes out small changes and reveals no spatial patterns. 2D thickness maps may allow clinicians to pick up small changes. Also, the spatial pattern of GCIPL thickness often contains useful features to detect subtle potential progression. So in this study, we aim to utilize both spatial and temporal information to predict the progression of glaucoma regarding 2D GCIPL damages. Following the clinical convention, we used GCIPL thickness maps from the past 4 visits as baselines to predict the map of the future 5th visit.

  - **Video interpolation**. Video interpolation aims to generate intermediate frames between given prior and post frames. However, natural videos include complicated appearance and motion dynamics, e.g., various object scales, different viewpoints, varied motion patterns, object occlusions, and dis-occlusions, making interpolation of realistic frames a significant challenge. To alleviate the above issues, we aim to design an efficient algorithm capable of generating realistic intermediate frames by modeling non-linear motion dynamics and complicated appearances.

1.3 Contributions

This dissertation focuses on crafting efficient algorithms tailored for four significant applications: segmentation-free point-wise VF estimation, retinal structure-to-function spatial mapping, GCIPL thickness map prediction, and video interpolation. The main contributions are detailed below.

- **Segmentation-free point-wise VF estimation in glaucoma**. We introduce a novel 3D Convolutional Neural Network (CNN) designed to directly estimate point-wise VF sensitivities from segmentation-free 3D OCT volumes. This approach mitigates floor effects (the minimal measurable level) inherent in segmentation-dependent OCT measurements. Comparing its performance to a model trained with segmentation-dependent 2D OCT thickness maps in a substantial clinical dataset, we demonstrate the 3D model's superior performance both globally and point-wise. The analysis indicates reduced influence from floor effects in the 3D model, resulting in more accurate estimations. This offers a potentially valuable avenue for developing substitutes for VF tests from OCT retinal scans, aiding clinicians and patients unable to undergo conventional VF examinations.

- **Generalized point-wise spatial mapping of structure to function in glaucoma**. Through occlusion analysis of the aforementioned 3D model, we establish a generalizable point-wise spatial relationship between OCT-based structure and VF-based function. These derived maps visualize global trends in point-by-point spatial relationships between structure and function, presenting a bias-free learning opportunity from trained machine learning models without prior knowledge or segmentation of OCT volumes. The revealed spatial correlations align with previously published mappings, opening avenues for robust and bias-free learning.

- **Retinal GCIPL thickness map prediction**. This work introduces the first attempt to predict 2D GCIPL thickness maps. We propose a novel time-aware Convolutional Long Short-Term Memory (TC-LSTM) unit within an auto-encoder-decoder framework. This model effectively handles irregular sampling intervals of longitudinal GCIPL thickness map sequences, capturing both spatial and temporal correlations. Experiments demonstrate the superiority of our proposed model over traditional methods.

- **Video interpolation**. We present Pyramid Deformable Warping Network (PDWN), a light yet efficient model for video interpolation. Using a pyramid structure, PDWN generates deformable convolution (DConv) offsets for predicting the unknown middle frame relative to known frames through successive refinements. Ablation studies validate the effectiveness of coarse-to-fine offset refinement, cost volumes, and DConv. PDWN achieves comparable accuracy to state-of-the-art models across multiple datasets while significantly reducing model parameters and inference time. Additionally, our extended framework utilizing longer input sequences substantially improves performance with a minimal increase in model size and inference time.

1.4 Outline and Organization

In Chapter 2, a novel 3D framework for estimating functional measurement e.g. VF sensitivities directly from segmentation-free 3D OCT volumes is explored. Chapter 3 extends the work in Chapter 2 by establishing a generalizable spatial mapping from structure to function in glaucoma using occlusion analysis. The derived map identifies statistically significant ONH regions for predicting VF test points for specific patient groups, not only demonstrating the correctness of the model in Chapter 2 but also opening up the possibility of learning from trained machine learning models, potentially robust and free from bias. Chapter 4 introduces a TC-LSTM model to predict future 2D GCIPL thickness maps, enabling the prediction of the progression of glaucoma. Chapter 5 presents a video interpolation algorithm to generate realistic intermediate frames for video frame rate conversion. Finally, Chapter 6 concludes the thesis and provides remarks on the possible directions for future research.

Chapter 2

Segmentation-Free Point-wise VF Estimation

2.1 Introduction

Glaucoma is a slowly progressive disease accompanied by characteristic loss of retinal ganglion cells and their axons and functional defects, which can greatly impact the quality of life [40, 74]. In practice, visual field (VF) test is an essential examination to identify and monitor functional abnormalities. However, VF test is highly subjective and highly depends on patient compliance. It suffers from random errors and fluctuations due to various factors such as patient attention, fatigue, learning curve, and rim artifacts [26, 67, 116]. Moreover, the fluctuations are more severe in glaucomatous patients than in healthy subjects [11, 56]. On the other hand, optical coherence tomography (OCT) which is widely used for visualizing and measuring retinal structures has very good Reproducibility in both healthy and glaucomatous subjects [12, 28, 45]. Previous studies have shown that the structural changes measured by OCT are closely related to the functional changes measured in VF [32, 60, 90, 92, 117]. Thus, the surrogates of VF test outcomes could be derived from OCT retinal scans via accurate quantitative models encapsulating structural-functional relationships. It can not only benefit the patients by reducing the long testing time of VFs but also allow clinicians to make clinical judgments without the inherent limitations of VFs such as subjective nature and high test-to-test variability.

Prior attempts to characterize structural-functional relationships have focused on correlating VF outcomes and structural measurements [25, 27, 30, 32]. Garway et.al. [32] proposed the widely used and validated Garway-Heath map to map localized retinal nerve fiber layer (RNFL) defects measured by red-free RNFL photographs to the location of points on SAP. With spectral-domain OCT (SD-OCT) becoming popular and able to provide a better assessment of RNFL, many previous researches have modeled the structural-functional relationship between OCT measurements like peripapillary RNFL thickness and standard automated perimetry (SAP) using statistical tools [25,27,30]. However, these studies relied on small samples and oversmoothed summarized thickness measurements.

Recent developments of artificial intelligence have shown the potential of deep learning (DL) algorithms in modeling complex nonlinear relationships and learning task-specific features automatically from high-dimensional data in various medical sectors [19, 37, 64, 95]. Recent research has attempted to use DL methods to estimate VF from higher-dimensional SD-OCT measurements such as two dimensional (2D) SD-OCT thickness maps [22, 39, 65, 66, 81, 97]. The promising results proved the advantages of 2D thickness maps over summarized measurements. However, 2D thickness maps are prone to segmentation errors introduced by the adopted segmentation algorithms, leading to inaccurate estimation of VF. Furthermore, the presence of segmentation errors is associated with macular diseases [2] and hence impedes the practical application of directly estimating function from structure. In addition to segmentation errors, segmentation-based OCT measurements have floor effects [71]. The floor effect is the point at which no further structural loss can be detected by segmentation-based OCT measurements. So it will affect the structural-functional relationship learning for patients with advanced disease. Alternatively, the three dimensional (3D) OCT volume not only is segmentation-free but also can provide more abundant information than 2D thickness maps. Therefore, to overcome the above limitations, we developed a DL model to estimate pointwise functional outcome directly from segmentation-free 3D OCT volumes and compared the performance with the model trained with segmentation-dependent 2D OCT thickness maps. We also proposed a gradient loss term to utilize spatial information in VF by reshaping VF into 2D arrays and calculating gradients between adjacent VF points.

2.2 Methods

2.2.1 Model architecture

A Convolutional Neural Network (CNN) was developed to take one optic nerve head (ONH) OCT as input to predict a 52-dimensional VF vector. We adopted ResNet18 [43] as the backbone of the feature extractor and replaced the last fully connected layer with 2 convolutional layers to output 52-point VF sensitivities. The 3D version replaces all 2D Convolutional layers in the 2D ResNet18 to be 3D convolutions. Model details are shown in Fig 2.1.

2.2.2 Loss function

To train the network, we used mean square error as the reconstruction loss:

$$\mathrm{L}_{\text{reconstruction}} = \frac{1}{N \times 52} \sum_{n=1,i=1}^{N,52} (y_i^n - \hat{y}_i^n)^2, \tag{2.1}$$

where y_i^n and $\hat{y}_i^n$ were the ground-truth and estimated value, respectively, for the i^{th} component of the 52-point VF vector for the n^{th} sample. We also experimented on mean absolute error loss and got similar results as mean square error loss. Therefore, we only reported results trained with mean square error loss in the study.

Typical glaucomatous visual field loss is characterized by arcuate defects, nasal steps,

Figure 2.1. Model details. (A): 3D model taking a 3D OCT volume as input ($1 \times 128 \times 64 \times 64$, $channel \times depth \times width \times height$). (B): 2D model taking a 2D RNFL thickness map as input ($3 \times 200 \times 200$, $channel \times width \times height$).

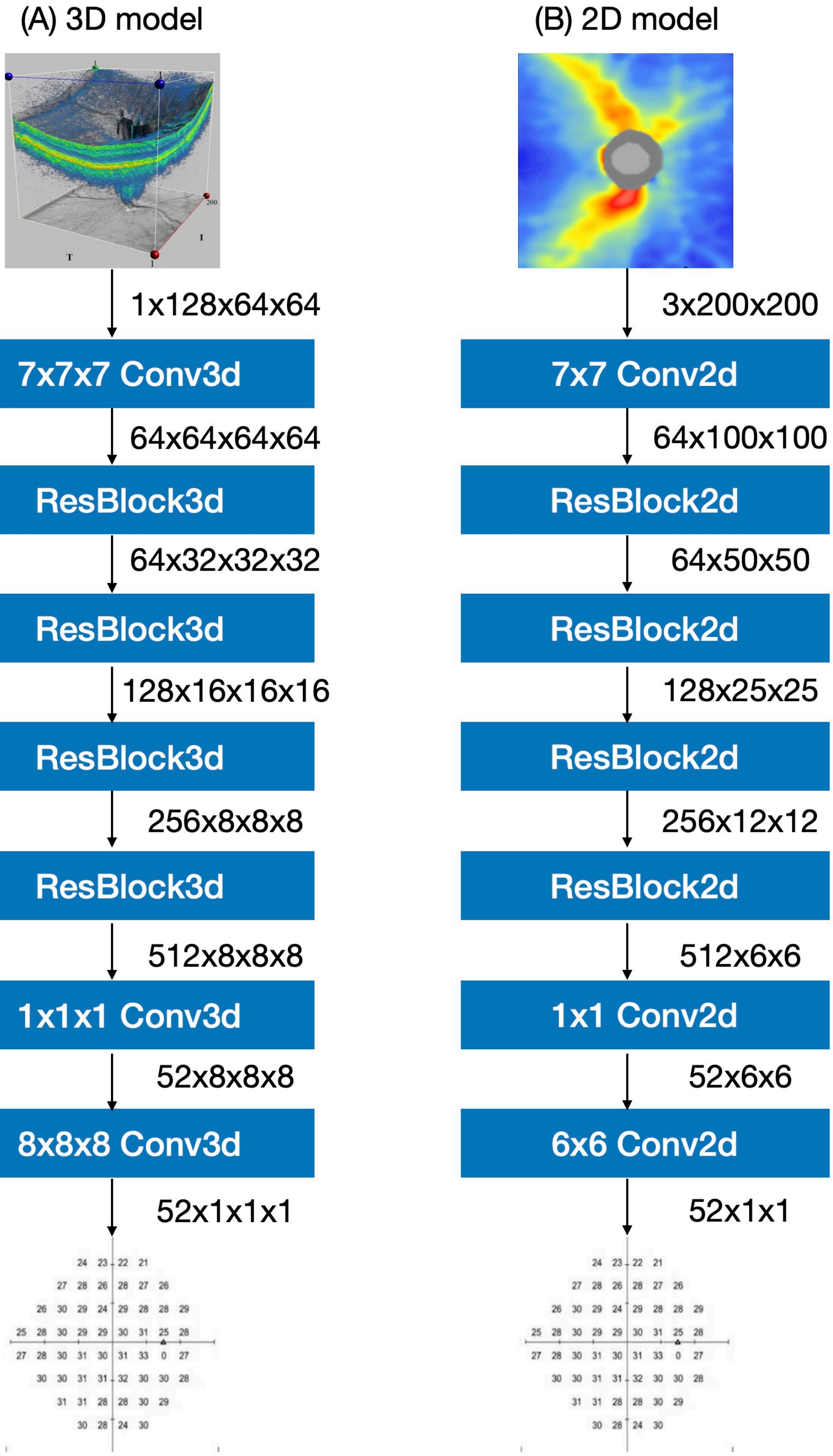
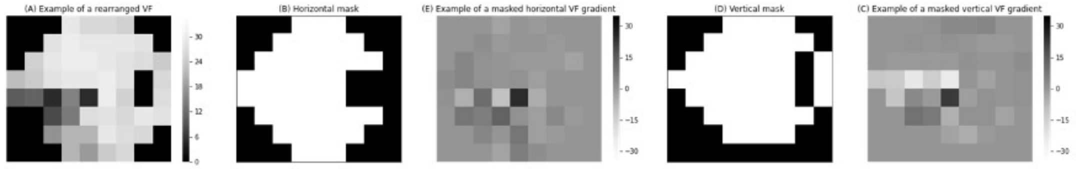


Figure 2.2. An example of rearrange 2D VF and gradients.



and other patterns on rectangular grids. To better utilize the spatial correlation in nearby VF points, we rearranged the output VF vector into a 8×9 2D array and filled in the boundary with zeros as depicted in Figure 2.2. Then a gradient loss term was proposed to minimize the differences of the horizontal and vertical gradients between the estimated and ground truth VF array as follows:

$$\mathrm{L}_{\text{gradient}}^{\text{horizontal}} = \|M_h * (y_{i,j} - y_{i-1,j}) - M_h * (\hat{y}_{i,j} - \hat{y}_{i-1,j})\|_1, \tag{2.2}$$

$$\mathrm{L}_{\text{gradient}}^{\text{vertical}} = \|M_v * (y_{i,j} - y_{i,j-1}) - M_v * (\hat{y}_{i,j} - \hat{y}_{i,j-1})\|_1, \tag{2.3}$$

where $y_{i,j}$ denotes value of the point at i^{th} row and j^{th} column in the ground truth 2D VF array. M_h and M_v are the binary masks to exclude gradients for blind spots and boundary points for horizontal and vertical gradients as shown in Figure 2.2 (B) and (D). With the gradient loss, the 52 points of the VF vectors were not independent to one another anymore. The model was enforced to not only reconstruct the individual points faithfully but also to match with the change pattern of ground truth visual field defects. Thus, the gradient loss emphasized on the learning of the spatial changes between adjacent VF points, which is essentially the spatial patterns of VF defects.

Finally, we set the training loss as:

$$\mathrm{L}_{\text{total}} = \mathrm{L}_{\text{reconstruction}} + \lambda \mathrm{L}_{\text{gradient}}^{\text{horizontal}} + \lambda \mathrm{L}_{\text{gradient}}^{\text{vertical}}, \tag{2.4}$$

where λ is set to be 10.

The model was trained with stochastic gradient descent, optimized by the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ [55]. The initial learning rate was 2×10^{-4} and decayed every 100 epochs by 10^{-1}. We trained the model for 200 epochs.

2.2.3 Statistical analysis

We used mean absolute error (MAE) and Pearson's correlation coefficients between the measured and estimated VF to evaluate the model performance. Pearson's correlation coefficients were tested with the Williams test for equality of correlations and MAEs were tested with the Wilcoxon Signed-rank test.

2.3 Results

2.3.1 Data collection

This retrospective study was performed in accordance with the tenets of the Declaration of Helsinki. The study was approved by the Institutional Review Board of Langone Health Center of New York University.

Subjects that had at least one Humphrey VF test and one SD-OCT visit within 90 days of each other were included in the study. VF tests were performed using SAP with the 24-2 Swedish Interactive Threshold Algorithm (Carl Zeiss Meditec, Inc., Dublin, CA, USA) protocol. Tests with more than 33% fixation losses, 15% false positive errors, or 15% false negative errors were excluded. SD-OCT tests were acquired using Cirrus SD-OCT instrument (Carl Zeiss Meditec). RNFL thickness maps were obtained from the $6mm \times 6mm$ ONH scan 200×200 protocol. Tests with signal strength less than 6 dB were also excluded.

The final dataset contains 8387 VF tests and 15026 ONH OCT scans from 1129 patients over multiple visits. Table 2.1 summarizes the demographic characteristics of the dataset. The dataset was randomly split into the training and testing datasets at a ratio of 9:1 with no patient overlap between datasets. To reduce variances caused during OCT imaging, the Bruch's membrane surface was flattened and aligned by adjusting A-scans in the z-direction. The 3D scans were centrally cropped around the optic nerve head to be $128 \times 128 \times 512$ voxels. The 3D scans were then downsampled to be $64 \times 64 \times 128$ voxels with Gaussian denoising to reduce the memory consumption during model training. Of the 54 points of 24-2 VF tests, 2 blind spot points were excluded. The sensitivity values of the remaining 52 test points were temporally smoothed using point-wise linear regression to reduce random fluctuations. All left eye visits were flipped horizontally to match the right eye format for both OCTs and VFs. During training, every VF visit was randomly associated with a corresponding OCT visit within 90 days to form a training pair.

Table 2.1. Demographic characteristics of the dataset.

	Mean ± standard deviation	Range
Number of patients	1129	-
Number of eyes	2151	-
Age at the time of visit (year)	64.2 ± 12.6	18 - 94
Number of VF visits per patient	3.9 ± 2.9	1 - 20
Number of OCT visits per patient	6.6 ± 5.5	1 - 63
VF mean deviation	-4.59 ± 6.80	-32.78 - 5.78
Peripapillary RNFL thickness (μm)	76.76 ± 14.95	14 - 80

2.3.2 3D model vs 2D model

The test dataset has 996 VF-OCT pairs from 147 patients. Table 2.2 summarizes the global performance comparison between the model trained with 3D OCT volumes and the model trained with 2D thickness maps. The MAE is significantly lower in 3D than 2D models (3.11 vs. 3.47 dB, $p < 0.001$, Wilcoxon Signed-rank test). Pearson's correlation coefficient is also significantly better in 3D than 2D models (0.80 vs. 0.75, respectively, $p < 0.001$, Williams test for equality of correlations). Both metrics demonstrate that the overall performance of the 3D model is better than that of the 2D model.

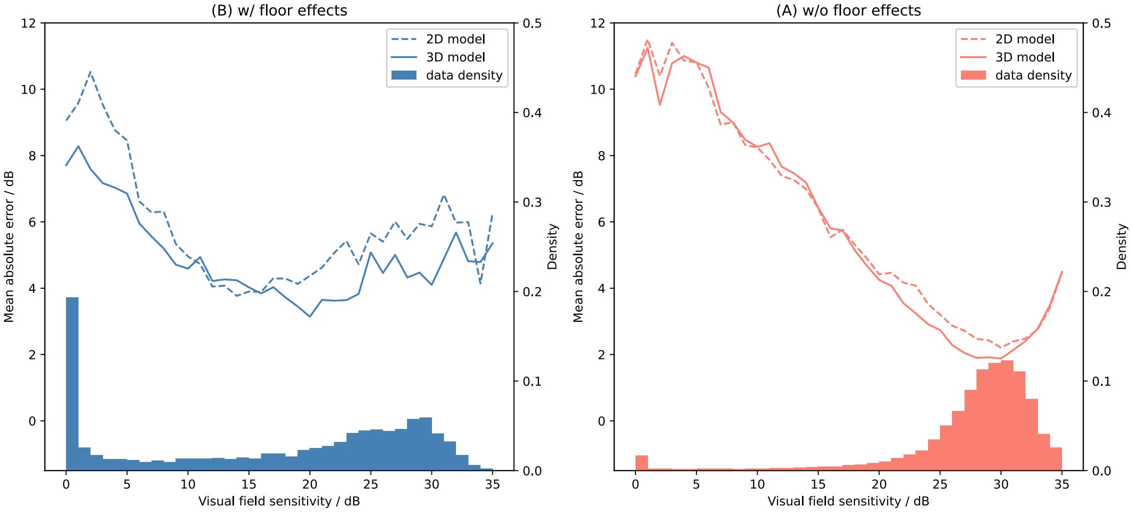
Table 2.2. Comparison between 2D-thickness-map-based model and 3D-volume-based model.

Model	All test data		With floor effects		Without floor effects	
	MAE	Correlation	MAE	Correlation	MAE	Correlation
2D model	3.47 ± 3.75	0.75	6.34 ± 4.58	0.74	3.09 ± 3.45	0.67
3D model	3.11 ± 3.54	0.80	5.24 ± 3.99	0.83	2.82 ± 3.37	0.70
	p < 0.001	p < 0.001	p < 0.001	p = 0.001	p < 0.001	p = 0.004

To further investigate model performances when RNFL reaches the measurement floor, we split the test dataset into two parts, one with and the other without floor effects. The set with floor effects consists of visits that have an average RNFL thickness less than or equal to 57 μm according to [71]. The floored test set has 117 VF-OCT pairs from 20 patients while the other test set has 879 pairs from 127 patients. Global results are shown in table 2.2. Globally, both MAE and correlation coefficients of the 3D model are significantly better than that of the 2D model in terms of MAE. The Pearson's correlation coefficient's gain of the 3D model is significant when floor effects are present while the gain is marginal in data without floor effects. Moreover, the performance gain of the 3D model is much greater in the floored test data than that in test data without floor effects (4 times for MAE, 3 times for correlation).

Fig 2.3 shows the error trends of both models at every VF sensitivity level. The error trends of 3D and 2D models do not differ much for data without floor effects in subfigure (B). Conversely, the MAE of the 3D model clearly shows a better trend than that of the 2D model for data with floor effects in subfigure (A). The gap between the two lines demonstrates that the model using 3D volumes has less influence from the floor effects. Regardless of floor

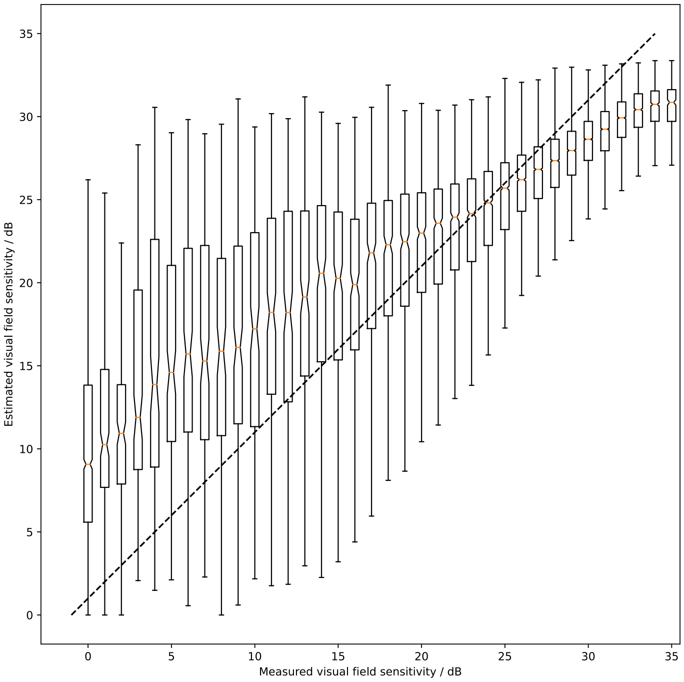
Figure 2.3. Error trend comparison on test data with and without floor effects. (A): with floor effects. (B): without floor effects. The dashed line and the solid line represent the MAE of the 2D thickness-map-based model and the 3D volume-based model, respectively. Histograms show the VF density for data with and without floor effects.



effects, both models perform better for VF sensitivity between 20 and 35 dB, which are most frequently sampled in our dataset, than values under 20 dB. As a result, a plateau effect is presented in Fig 2.4 when the measured sensitivity is less than 20 dB. A similar pattern was also presented in [66]. In addition, the high test-to-test variability of VF sensitivity values below 20 dB may also contribute to the large estimation error in the low sensitivity end [29, 113].

In Fig 2.5 we plot the difference between evaluation metrics of 3D and 2D models for each VF point. Subplot (A) is the point-wise mean absolute error map, i.e., $MAE^i_{2D} - MAE^i_{3D}$ where i represents one of the 52 points. Red in the map means that the 3D model has lower/better MAE than the 2D model and blue means the opposite. Subplot (B) shows the p value of the MAE difference. White cells have p value ≥ 0.05 and black and greyish cells have p value < 0.05. The MAE difference map (subplot (A)) is almost all red except for points in the central region. And the significance map is almost greyish. So the 3D model is significantly better than the 2D model in most VF positions in terms of MAE. Similarly, the point-wise Pearson's correlation coefficients of the 3D model are significantly better than those of 2D models in most VF positions. Therefore, the point-wise analysis again

Figure 2.4. Box plot of the 3D-volume-based model estimations.



demonstrates the supremacy of 3D OCT volumes versus 2D thickness maps.

Table 2.3 summarizes the sectional results of the 3D model. We use the sectors defined in Garway-Heath map [32]. The observation that the superior sector has smaller MAEs compared to the corresponding inferior sector (3.14 vs 3.67 dB temporally, 2.75 vs 3.23 dB nasally) coincides with [38, 81]. [38] claimed that it is due to the superior retina having a higher structural-functional correlation than the inferior retina. [81] suggested another reason for the observation. Glaucomatous damage occurs sequentially from the inferotemporal sector to the superotemporal sectors [93]. As a result, the inferotemporal ONH sector could have a larger error since the inferotemporal ONH sector progressed more than other sectors. Our results may support their hypothesis. The superior-inferior MAE gaps narrow as glaucoma progresses. For example, the superior-inferior MAE gaps narrow from 0.49 dB ($|2.96 - 3.47|$ dB) temporally and 0.43 dB ($|2.47 - 2.90|$ dB) nasally in data without floor effects to 0.12 dB ($|5.06 - 5.18|$ dB) temporally and 0.09 dB ($|5.26 - 5.35|$ dB) nasally in data with floor effects. Nevertheless, the pattern of Pearson's correlation coefficients does not agree with the pattern of MAE. The superior correlations are not better than the inferior correlations.

Figure 2.5. Point-wise analysis on floored test data. (A): MAE difference map, $MAE_{2D} - MAE_{3D}$. The red cell means the 3D model has a lower MAE than that of the 2D model at the point while the blue cell means the opposite. (B): Significance map of MAE difference. (C): Correlation difference map, $correlation_{3D} - correlation_{2D}$. The red cell means the 3D model has a higher Pearson's correlation coefficient than that of the 2D model at the point while the blue cell means the opposite. (D): Significance map of correlation difference. In the significance maps, white cells have p value ≥ 0.05 and black and gray cells have p value < 0.05.

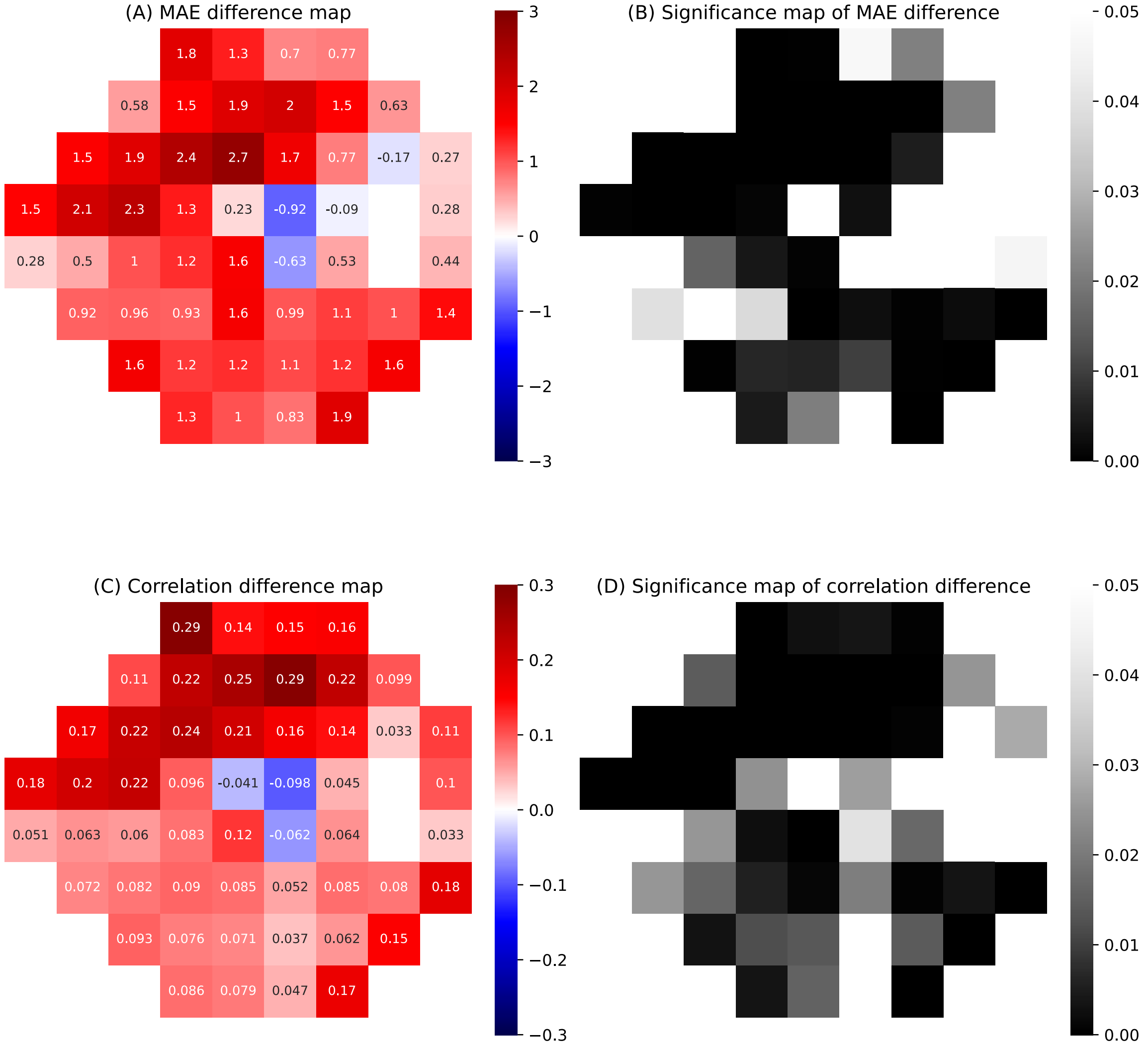


Table 2.3. Sectional results of the 3D model.

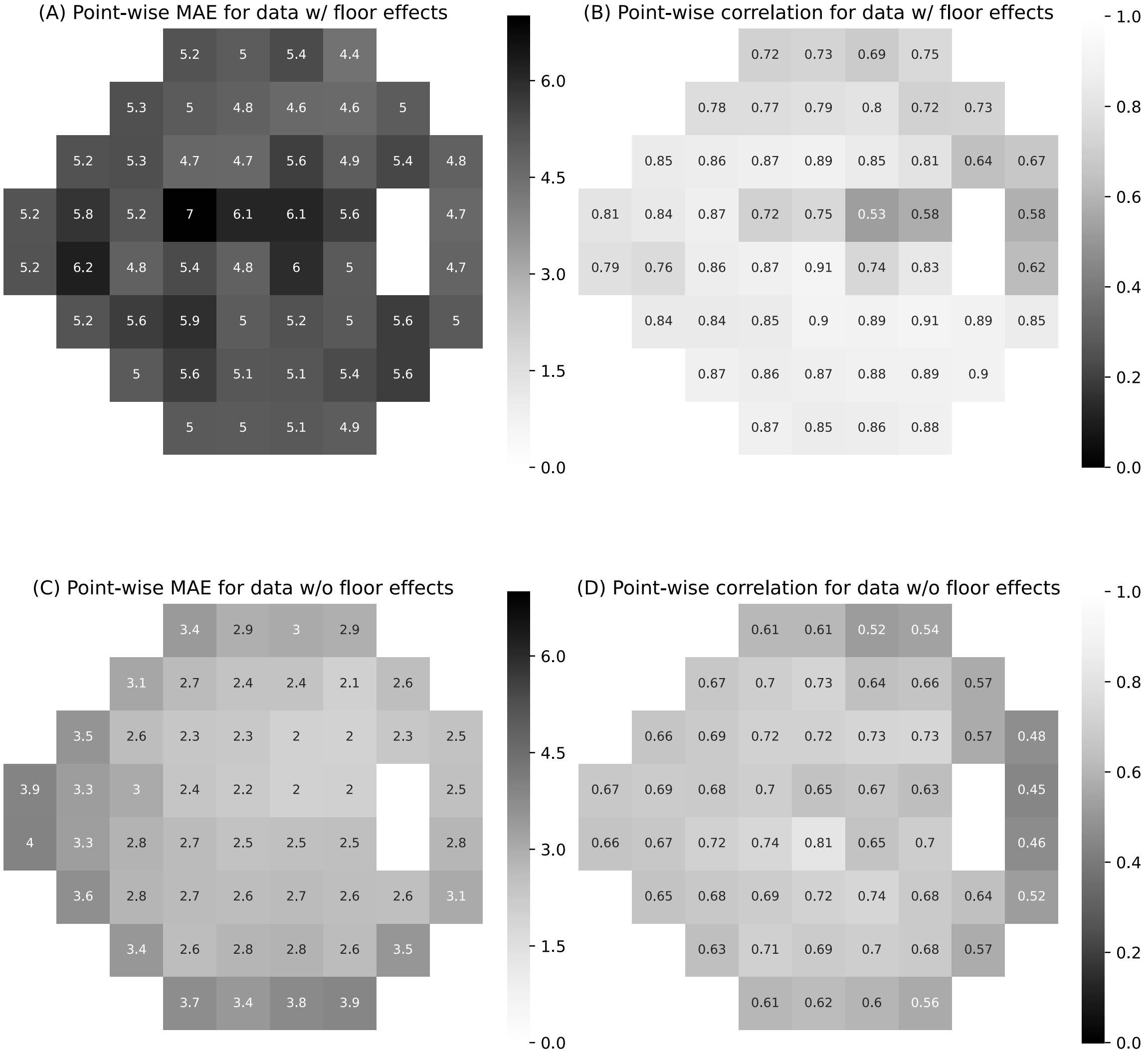
Sector	All test data		With floor effects		Without floor effects	
	MAE	Correlation	MAE	Correlation	MAE	Correlation
Superotemporal	3.14 ± 3.59	0.77	5.06 ± 3.87	0.77	2.96 ± 3.53	0.67
Temporal	2.93 ± 3.33	0.65	4.82 ± 3.91	0.71	2.71 ± 3.11	0.49
Inferotemporal	3.67 ± 3.44	0.78	5.18 ± 3.62	0.87	3.47 ± 3.35	0.63
Inferonasal	3.23 ± 3.65	0.82	5.35 ± 4.04	0.85	2.90 ± 3.45	0.72
Nasal	2.72 ± 3.43	0.76	5.62 ± 4.62	0.77	2.30 ± 2.95	0.71
Superonasal	2.75 ± 3.46	0.82	5.26 ± 3.87	0.83	2.47 ± 3.35	0.71

Fig 2.6 shows a point-wise analysis of the 3D model performance for data with and without floor effects. For data without floor effects, the 3D model performs better in central locations than in boundary locations probably due to rim artifacts of VF. However, the performance in central locations is worse than that in boundary locations for data with floor effects.

2.3.3 Comparison with prior studies

Previous studies have also used DL to learn structure-function relationships. Commonly, segmentation-based thickness measurements by OCT devices are used as inputs to predict VF outcomes. Shin et. al. [97] compared 24-2 VFs outcomes from 2D RNFL and Ganglion Cell–Inner Plexiform layer (GCIPL) thickness maps measured by SD-OCT and by swept-source OCT (SS-OCT). They showed that their model estimated VFs better with SS-OCT (root mean square error (RMSE) = 4.51 ± 2.54 dB) than did with SD-OCT (RMSE = 5.29 ± 2.68 dB). Though we cannot directly compare with their results, we achieve similar RMSE (4.22 ± 2.88 dB) for our 2D model which also utilized RNFL map of SD-OCT and better RMSE (3.83 ± 2.74 dB) for our 3D model. Park et. al. [81] developed an InceptionV3-based model to predict 24-2 VFs from combined GCIPL and RNFL thickness maps and achieved RMSE of 4.79 ± 2.56 dB, which is also similar to our 2D model (4.22 ± 2.88 dB). Mariottoni et. al. [66] used CNN to predict 24-2 VFs from 768 peripapillary RNFL thickness points in SD-OCT. They reported an average correlation coefficient of 0.60 and an MAE of 4.25 dB. In our case, the correlation coefficients are 0.75 and 0.80 for

Figure 2.6. Point-wise results of the 3D model for data with and without floor effects. (A): Point-wise MAE for data with floor effects. (B): Point-wise correlation for data without floor effects. (C): Point-wise MAE for data without floor effects. (D): Point-wise correlation for data without floor effects.

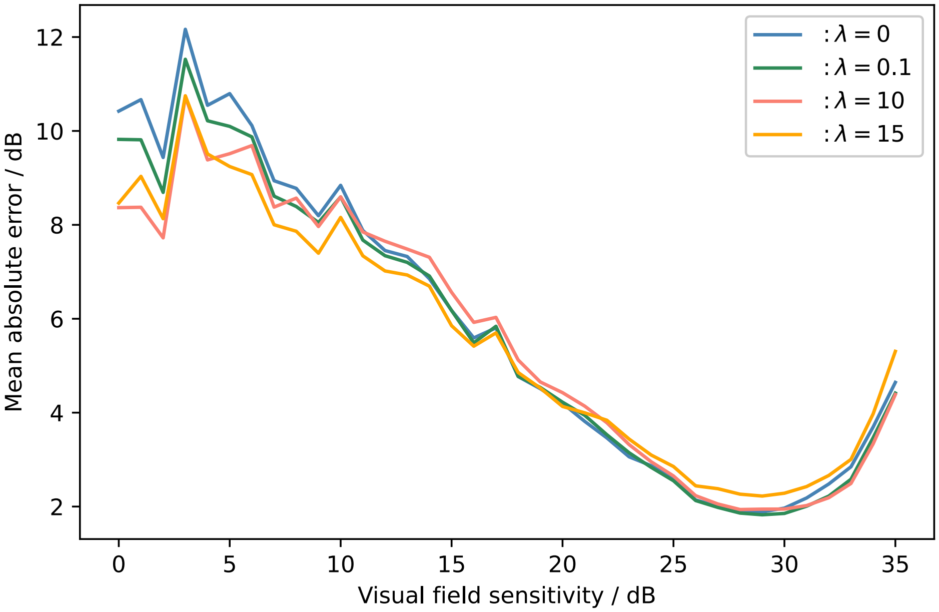


2D and 3D models respectively. Overall, our 2D model has similar performance with previous segmentation-dependent methods, but our 3D model significantly outperforms previous segmentation-dependent methods. Comparisons are summarized in Table 2.4.

Table 2.4. Comparison with prior studies using SD-OCT.

Method	MAE	RMSE	Correlation
Shin, et al [97]	-	5.29 ± 2.68	-
Park, et al [81]	-	4.79 ± 2.56	-
Mariottoni et al [66]	4.25	-	0.60
Ours (2D)	3.47 ± 3.75	4.22 ± 2.88	0.75
Ours (3D)	3.11 ± 3.54	3.83 ± 2.74	0.80

Figure 2.7. Effectiveness of the gradient loss.



2.3.4 Gradient loss

Figure 2.7 shows performances corresponding to different settings of λ in the training loss function (2.4). Introducing the gradient loss clearly gives a boost to the performance in the lower VF sensitivity end. We chose $\lambda = 10$ in our experiments because it provides the lowest estimation errors in low-sensitivity regions in the training set.

2.4 Conclusion

In conclusion, we investigate a DL model to estimate point-wise VF sensitivities directly from segmentation-free 3D OCT volumes to overcome the floor effects of segmentation-dependent OCT measurements. We compare the performance with the model trained with segmentation-dependent 2D OCT thickness maps in a large clinical dataset. We show that the 3D model is significantly better than the 2D model both globally and point-wisely. Further analysis on a subset of the test dataset with floor effects demonstrates that the 3D model had less influence from the floor effects and thus generated more accurate results than the 2D model. Moreover, we propose a gradient loss function to be combined with mean square error loss to utilize the spatial information of VFs. The proposed loss improves the estimation error for low-sensitivity values. Our study provides a better quantitative model to encapsulate the structural-functional relationship more accurately. Our study could offer new insights into developing surrogates of VF test outcomes from OCT retinal scans. This may help both clinicians and patients who are unable to undergo real VF examinations.

Despite the improvement introduced by segmentation-free OCT volumes, this study still has limitations. First, the dataset is imbalanced in terms of VF sensitivity values. A relatively large error is present due to under-represented low and very high sensitivities. Though we demonstrate that predicting VFs directly from 3D OCT volumes and using gradient loss could alleviate the issue of low sensitivity values, the problem persists. Further investigation is needed with additional low and very high sensitivity data. Second, VF tests are prone to errors and variability, leading to difficulties in model training and evaluations, which imposes a lower bound on the achievable predictive performance. In addition, the test-retest variability of VF is even higher for sensitivity values under 19 dB [31], further limiting the model's predictive performance for low VF sensitivity values. Repeated tests may help suppress noise in VF to construct a cleaner dataset. Finally, despite the advantage of feature agnosticism, using non-segmented OCT volumes is inefficient in terms of memory and computation since the OCT volumes contain a substantial area without tissue information. Although flattening, cropping, and downsampling have been applied in preprocessing steps to improve the efficiency of memory and computation, more advanced methods combining the segmentation masks can be explored in future work.

Chapter 3

Generalized Point-Wise Spatial Mapping of Structure to Function

In Chapter 2, we have introduced a three dimensional (3D) model that is able to predict visual field (VF)-based functional damage from 3D segmentation-free optical coherence tomography (OCT)-based structural damage. Although there is a strong correlation between structural and functional measurements, there are many clinical cases that cannot be clearly explained with such a simple correlation. Since glaucoma progression patterns widely vary from individual to individual, a detailed spatial correlation map may help identify personalized progression patterns and provide better assessment and forecasting of progression. Built upon the work in Chapter 2, this chapter introduces how to derive spatial mapping from structure to function using model visualization.

3.1 Introduction

Structure-function spatial correlation has been widely investigated [25, 30, 32, 49, 108]. Perhaps the most well-known depiction is the Garway-Heath map that associates clusters of VF test points with sectors of the optic disc by superimposing 24-2 VF test grid on retinal photographs and manually tracing visible retinal nerve fiber layer (RNFL) defects or prominent nerve fiber bundles to note their point of intersection [32]. The derived map divides the optic nerve head (ONH) and 24-2 VF into six corresponding sectors. Jansonius et. al. [49] later proposed a mathematical model fitting hand-traced retinal nerve fiber trajectories to reduce variabilities in hand-tracing, leading to a more robust portrayal of the structure-function relationship. Alternative approaches used statistical methods to produce the structure-function correspondence. Gardiner et. al. [30] utilized the maximum correlation between the normalized rim area of 36 sectors measured by Heidelberg retina tomography (HRT) and 24-2 VF sensitivities. Turpin et. al. [108] further constrained the correlation between HRT measurements and VF to be anatomically plausible with a computational model of the axon growth of retinal ganglion cells. Ferreras et. al. [25] used factor analysis to divide 24-2 VF grid into 10 sectors. Then, a similar correlation approach was applied to relate predefined 10 VF sectors to clock-hour sectors of peripapillary RNFL thickness measured by OCT.

All previous studies are based on prior knowledge regarding anatomical structures and their functions or require segmentations to get ONH measurements. It can certainly be a good way of establishing structure-function relationships. However, it is possible to discover unexpected anatomical or structural features that are highly associated with function using artificial intelligence. Recent advances in deep learning (DL) approaches achieve unprecedented performance – sometimes better than human experts – in many medical applications. While DL models are known to be black boxes, recently many techniques to reveal which location within the input image contributed the most to reach the output have been developed. In other words, it is now possible to learn from well-trained DL models.

Several previous studies have attempted to predict VF outcomes using OCT measurements through DL algorithms [3, 18, 21, 41, 52, 53, 57, 66, 79, 81–83]. Although these studies have shown promising results in approximating VF metrics from OCT data, the precise spatial relationship between structural damage and functional damage remains less well-established. Mariottoni et. al. [66] created a mapping between the 768-point RNFL thickness profile obtained from a spectral-domain OCT (SD-OCT) circumpapillary scan and the 24-2 standard automated perimetry (SAP) VF loss by simulating localized RNFL defects of varying locations and characteristics. They observed the impact of these defects on VF outcomes using a CNN designed to predict VF sensitivities from RNFL thickness profiles. The derived map offers a more detailed spatial structure-function relationship compared to the Garway-Heath map, but their method depends on the segmentation outcomes, which can be affected by image quality and segmentation errors [2]. Kihara et. al. [53] proposed a multimodal policy DL system that directly predicts VF from unsegmented circumpapillary OCT and Scanning Laser Ophthalmoscopy image of the ONH. Thus, a circumpapillary sector structure-function mapping was derived in a data-driven, feature-agnostic fashion. Nonetheless, all prior mappings remained limited to sector representations, which is suboptimal as they fail to fully exploit 3D nature of the retinal structure. A more comprehensive spatial mapping, derived from 3D structure measurements (e.g., 3D OCT volume) and independent of domain-specific knowledge (e.g., segmented RNFL thickness), is desired to enhance our understanding of the spatial relationship between structure and function.

Recently, DL algorithms have ventured into analyzing higher-dimensional data to leverage 3D information that may not be readily discernible through conventional methods [18]. Consequently, in this study, we aim to establish a generalized point-wise spatial mapping between structure and function by conducting occlusion analysis on a DL model trained on an extensive clinical cohort of patients to predict point-wise VF sensitivities from 3D OCT volumes. The revealed spatial correlations were consistent with previously manually derived maps. This may provide a gateway to the discovery of new findings from machine learning models, potentially robust and free from bias.

3.2 Methods

3.2.1 Model architecture and training

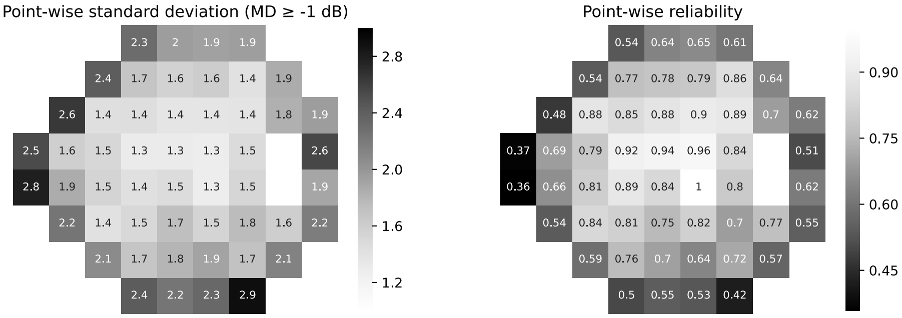
Since the purpose of this study was to derive the spatial relationship between structure and function, rather than developing a new model to predict function from structure, we adopted the same model architecture, a 3D CNN, from Section 2.2.1, which showed promising performance in predicting SAP VF sensitives from 3D SD-OCT volumes of the ONH.

We adopted the same training strategy as Section 2.2.1, i.e., we trained the model with the Adam optimizer [55] for 200 epochs and a batch size of 16. The learning rate was initially set to 2 x 10-4 and linear decayed every 100 epochs by 10-1. Different from 2.2.1, we used a reliability-weighted mean square error loss function instead of standard mean square error loss to compensate for larger noises in peripheral VF test points. The reliability was defined as the inverse of the point-wise standard deviation of sensitivities for subjects with VF mean deviation (MD) larger than -1 dB, with the highest reliability normalized to 1. The weight for each VF test point is shown in Figure 3.1.

3.2.2 Structure-function mapping

Once an accurate model is established to quantify the relationship between OCT volumes and point-wise VF sensitivities, it becomes possible to derive spatial correlations between each test point of the VF and the corresponding regions of the ONH using the model. To establish this spatial mapping, we first applied occlusion analysis on the trained model, generating 52 point-wise 3D saliency volumes for every sample in the test set. This allowed us to evaluate the contribution of individual regions in the input OCT volume to the model's predictions. To ensure consistency, all 3D saliency volumes were registered using the geometric center of the optic disc. Additionally, to account for variations along

Figure 3.1. Point-wise reliabilities.



the depth dimension, we averaged each 3D saliency volume across depth to generate a two dimensional (2D) individual saliency map, analogous to OCT en face images.

Next, we divided the test set into two groups based on VF MD values (cutoff at MD -6 dB) and performed a point-wise t-test separately for each small ONH region within each group. This enabled us to generate group t-statistic maps, revealing the detailed spatial relationship between each VF test point and the corresponding statistically significant and relevant regions of the ONH for a specific group.

Individual saliency map by occlusion analysis

Occlusion analysis is widely used to visualize the decision-making process of black box models. In this study, we utilized occlusion analysis to quantify the contribution, also known as saliency, of each small region of the ONH on the model's prediction for each VF test point. The underlying assumption is that if a region of the ONH is related to a VF test point, removing information from that region will significantly alter the DL model's prediction for the corresponding point. Conversely, the model's prediction should remain consistent when removing information from irrelevant ONH regions.

To implement this, we replaced a small region $(4 \times 4 \times 4$ voxels, $240 \times 240 \times 31.25\mu m)$

within an input volume with a gray patch (mean intensity of the input) and calculated the saliency by comparing the model's prediction with the original input to its prediction with the occluded input. The saliency was defined as the absolute difference between the model predictions for the original and occluded inputs. By repeating this process for all locations throughout the entire input volume for all VF test points of each ONH volume in the dataset, we generated 52 saliency volumes for every OCT-VF pair in the dataset. Specifically, for each OCT volume V_i in the test set $\{V_1, V_2, \ldots, V_N\}$, we obtained 52 saliency volumes $\{S_{i,1}, S_{i,2}, \ldots, S_{i,52}\}$ corresponding to the 52 VF test points. Each voxel $S^j_{i,pt}$ within a saliency volume $S_{i,pt}$ represented the absolute difference between the model predictions when occluding the j-th small region of the ONH volume:

$$S^j_{i,pt} = |f(V_i) - f(T^j(V_i))| \tag{3.1}$$

where V_i denotes the original ONH volume, $T^j(\cdot)$ denotes the operation that replaces the j-th patch of V_i with a value equal to the volume's mean intensity $\bar{V}_i$. An example of a 3D saliency volume was presented in Figure 3.2 and Figure 3.3.

Figure 3.2. The cross-sectional view of an individual saliency volume. (A) A VF test. (B) The associated en-face OCT image. (C) Cross-sectional B-scan associated with the red line in (B), overlaid with the corresponding saliency of point 21 highlighted with the red bounding box in (A). (D) Cross-sectional B-scan associated with the green line in (B), overlaid with the corresponding saliency of point 21 highlighted with the red bounding box in (A).

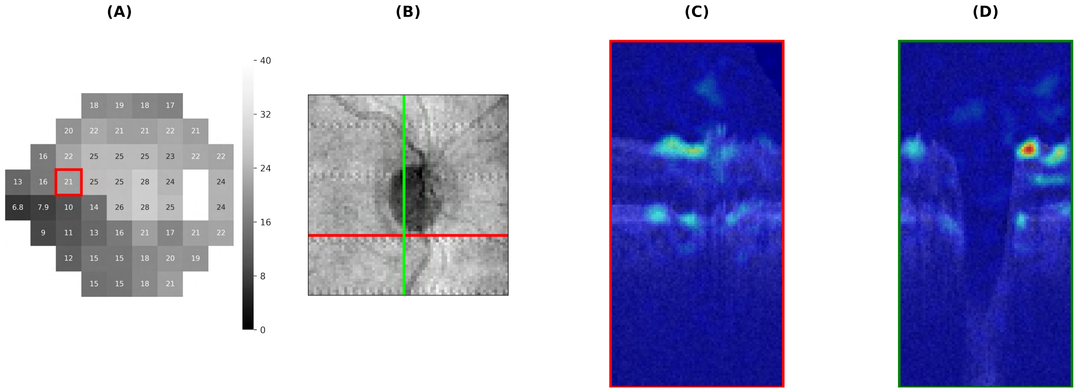
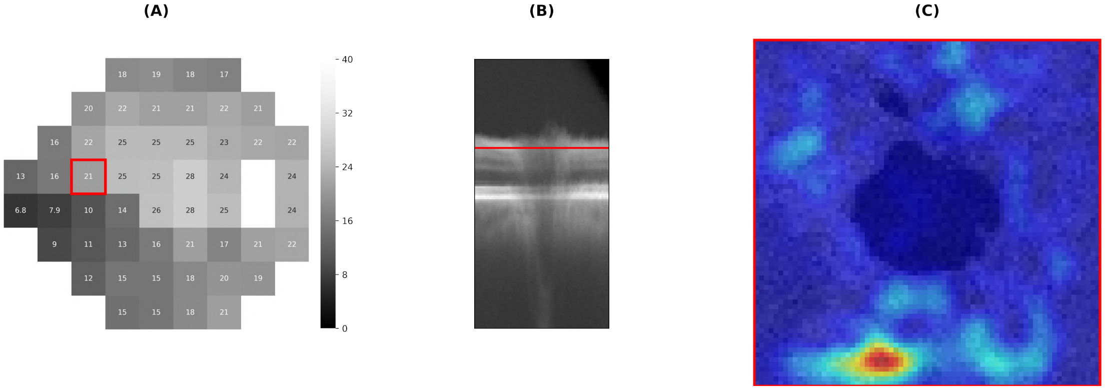


Figure 3.3. The en-face view of an individual saliency volume. (A) A VF test. (B) The associated cross-sectional OCT B-scan. (C) A scan associated with the red line in (B), overlaid with the corresponding saliency of point 21 highlighted with the red bounding box in (A).



To account for variation across depth due to the adopted coarse registration, the 3D individual saliency volumes

$$\{\{S_{1,1}, S_{1,2}, \ldots, S_{1,52}\}, \{S_{2,1}, S_{2,2}, \ldots, S_{2,52}\}, \ldots, \{S_{N,1}, S_{N,2}, \ldots, S_{N,52}\}\}$$

were projected onto the en face plane, generating 2D individual saliency maps

$$\{\{S'_{1,1}, S'_{1,2}, \ldots, S'_{1,52}\}, \{S'_{2,1}, S'_{2,2}, \ldots, S'_{2,52}\}, \ldots, \{S'_{N,1}, S'_{N,2}, \ldots, {S'}_{N,52}\}\}$$

to address depth-related variations and to ease visualization.

Group saliency map by t-test

To identify statistically significant and relevant ONH regions for each VF test point in a specific group, we divided the test set into two groups as described above and separately conducted t-tests within each group. This process yielded the corresponding group t-statistic maps, which encode the group-specific spatial relationships between structure and function.

The group t-statistic maps were generated by separately performing t-tests for each small ONH region, comparing its saliency with the point-wise group-averaged saliency. For a particular 4×4 region k in the en face plane and a particular VF test point pt, we had

$\{S'^{k}_{1,pt}, S'^{k}_{2,pt}, \ldots, S'^{k}_{N,pt}\}$ representing the saliency of region k for model prediction at VF test point pt for each subject in a group. The group t-statistic map T_{pt} of a particular VF test point pt was created by conducting t-test separately across all regions of ONH:

$$T^{k}_{pt} = \begin{cases} \dfrac{\bar{S}^{k}_{pt} - \mu_{pt}}{(\sigma^{k}_{pt}/\sqrt{N})}, & \alpha \leq 0.05 \\ 0, & \alpha > 0.05 \end{cases} \tag{3.2}$$

where $\bar{S}^{k}_{pt}$ and σ^{k}_{pt} denote the sample mean and sample standard deviation of saliencies for a particular patch k, and μ_{pt} denotes the hypothesis mean. We set μ_{pt} to be $\bar{S}_{pt} + \lambda\sigma_{pt}$, where $\lambda = 0.75$. In other words, a one-sample t-test was conducted to determine the extent to which the mean saliency of a particular patch k for a particular test point pt exceeds the mean saliency averaged over all patches $\bar{S}_{pt}$ (adjusted by standard deviation σ_{pt} as well) within the same group for that test point pt.

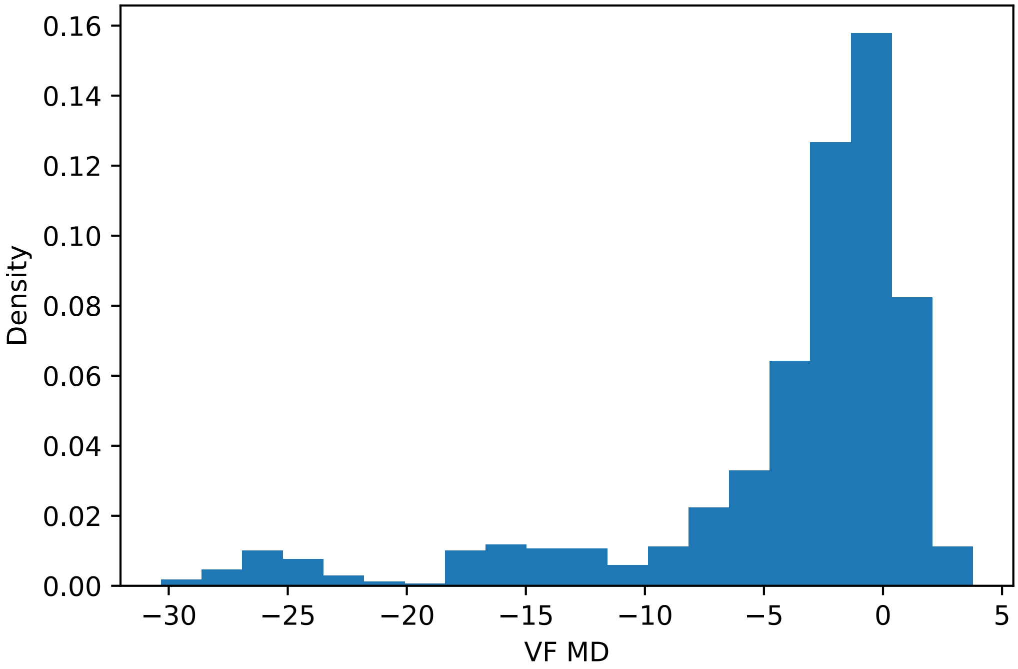
As a result, this study generated a new map, namely the group t-statistic map, which establishes the spatial relationship between each VF test point and the corresponding significantly relevant regions of the ONH.

3.3 Results

3.3.1 Data collection and VF estimation accuracy

We use the same dataset as described in Section 2.3.1. The distribution of VF MD is shown in Figure 3.4. The test set contained 996 OCT-VF pairs from 247 eyes. All participants were clinically diagnosed with glaucoma, glaucoma suspect, or healthy after undergoing a comprehensive ophthalmic evaluation that included a clinical exam, VF testing (Humphrey Field Analyzer; Zeiss, Dublin, CA), and an SD-OCT (Cirrus HD-OCT, Zeiss, Dublin, CA). Among them, 121 eyes have glaucoma, 108 eyes are glaucoma suspects, and 18 eyes are healthy. Every OCT and VF visit in the test set was unique, i.e., one OCT was

Figure 3.4. Distribution of VF MD.



only associated with one VF test.

To preprocess the OCT volumes and VFs, we adopted the same preprocessing steps as in Section 2.3.1. In brief, for OCT, we detected the ONH region and flattened Bruch's membrane opening (BMO) surface by segmenting BMO with smoothing and moving each A-scan along the z direction. Then central cropping and downsampling with Gaussian antialiasing filtering were applied to reduce memory consumption during model training. After preprocessing, all ONH volumes were flattened by BMO, centrally cropped at the optic disc center to $144 \times 144 \times 576$ voxels (covers a $4.32 \times 4.32 \times 1.125mm$ region of ONH), and downsampled to $72 \times 72 \times 144$ voxels. For VF, the 2 blind spot points were excluded. The sensitivities were temporally smoothed over 5 consecutive VF visits of the same eye using pointwise linear regression to reduce random fluctuations. All left eye visits were flipped horizontally to match the right eye format for both OCTs and VFs.

The resulting model achieved 2.92 dB of mean absolute error in estimating VFs, compared to 3.11 dB in Section 2.3.2.

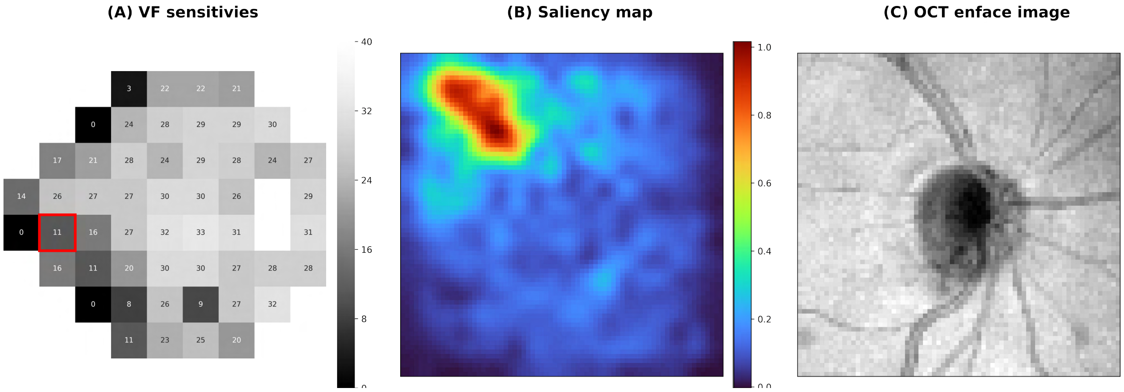
3.3.2 Group t-statistic maps

MD $>$ -6 dB V.S. MD $\leq$ -6dB

All results shown in this study were generated on the test set. There were 792 OCT-VF pairs from 207 eyes for MD $>$ -6 dB (-1.32 ± 1.90 dB) group and 204 pairs from 66 eyes for MD $\leq$ -6 dB (-17.93 ± 7.68 dB) group. Figure 3.5 shows an example of an individual saliency map that corresponds to a particular VF test point. Figure 3.6 shows the mean VF sensitivity maps and the group t-statistic maps for both groups. ONH sectors proposed by the Garway-Heath map were overlaid on top of the group t-statistic maps for comparison. As illustrated in Figure 3.6, the group t-statistic maps indicated that the structural locations with the most significant impact on VF sensitivity prediction were largely consistent with the Garway-Heath map.

However, some minor deviations were observed. For instance, in both the MD $>$ -6 dB and MD $\leq$ -6 dB groups, points 28 and 37 were slightly closer to the temporal aspect in our derived map, while point 43 was slightly closer to the nasal aspect. These discrepancies primarily occurred at the edge points of VF clusters defined in the Garway-Heath map, suggesting the existence of finer clusters that were not captured by the coarse mapping. Additionally, several factors might have contributed to the observed differences. Firstly, despite temporal smoothing of VF tests to mitigate noise, the remaining sensitivity variability of VF measurements could still hinder the model's ability to accurately characterize the structure-function relationship, leading to potential inaccuracies in the spatial mapping. Secondly, the coarse individual image registration method to generate group maps did not account for morphological variables such as disc to foveola angle. While we mitigated interindividual variation by using a relatively large sample size, these variations could still contribute to the observed discrepancies between our map and the Garway-Heath map. Finally, differences in the sample population, including demographic and clinical characteristics, could also contribute to slight variations in the structure-function mapping.

Figure 3.5. An example of an individual saliency map of a particular VF test point. (A) VF sensitivities of a subject in the test set. (B) The saliency map of a particular test point (highlighted with a red bounding box in (A)). (C) the corresponding en-face OCT image.



Overall, despite these minor discrepancies and factors potentially influencing the results, our study provides valuable insights into the spatial relationship between structure and function in the context of glaucoma. We successfully generated a generalized 2D mapping that establishes the group spatial relationship between VF test points and regions of the ONH at a fine resolution. Importantly, our algorithm relied solely on the data without any prior knowledge about the structure-function relationship, free from potential bias, segmentation errors, and/or floor effect. Despite the absence of explicit domain knowledge, the derived mapping captured spatial relationships that align with clinical expectations.

Superior defect V.S. inferior defect

We further divided the MD $\leq$ -6 dB group into superior and inferior defect subgroups by manually observing the 24-2 VF defect patterns. The corresponding VF sensitivity maps and group t-statistic maps are shown in Figure 3.7. Similarly, Garway-Heath sectors were overlaid on the t-statistic maps for comparison. The two groups' t-statistic maps showed symmetric patterns for superior and inferior damage groups. That is, the groups' t-statistic map of superior damages highlighted the inferior part of the retina and vice versa for the map of inferior damages.

Figure 3.6. Group t-statistic maps for MD ≥ -6 dB and MD ≤ -6dB. ONH sectors proposed by the Garway-Heath map were overlaid on top for comparison. Different colors represent different VF clusters defined in the Garway-Heath map.

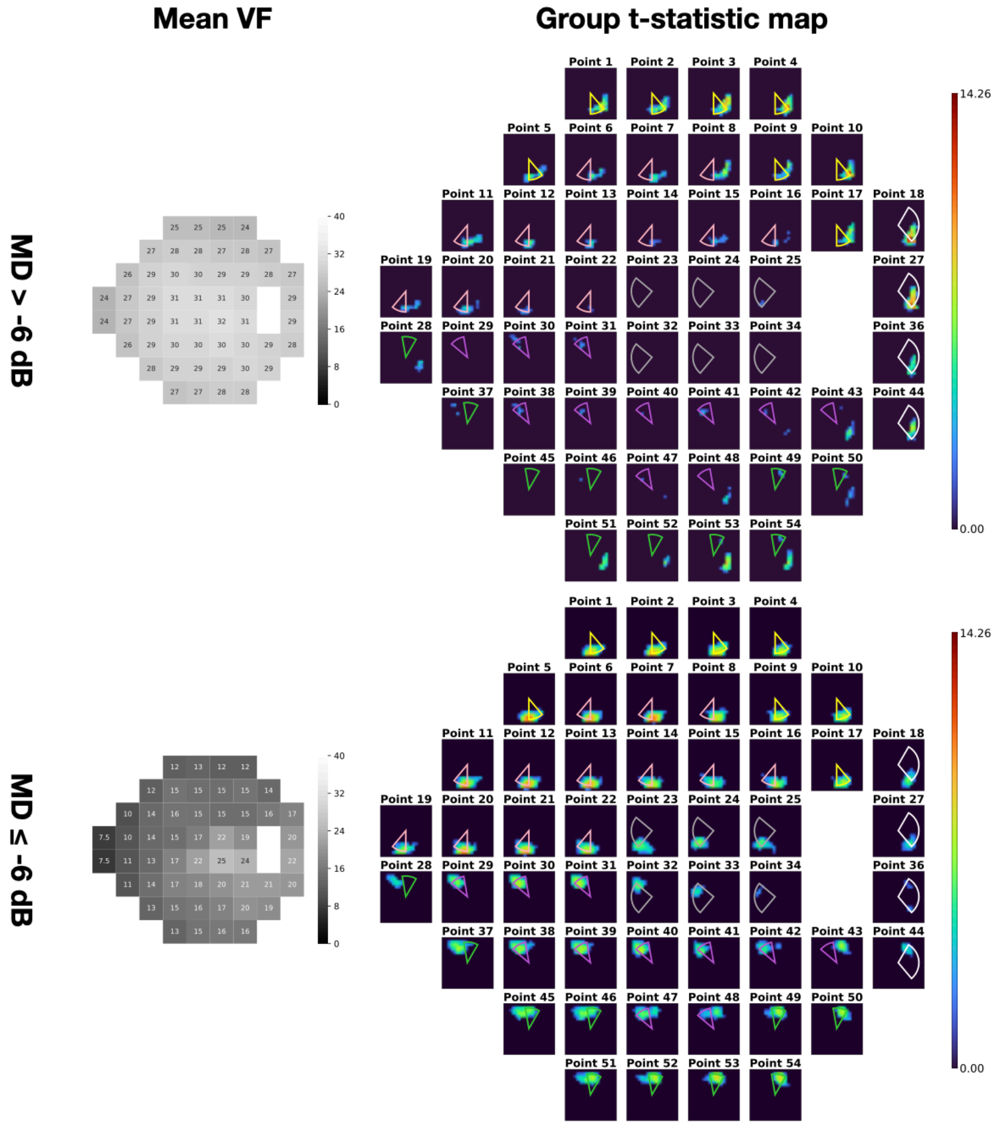


Figure 3.7. Group t-statistic maps of superior and inferior defects. ONH sectors proposed by the Garway-Heath map were overlaid on top for comparison. Different colors represent different VF clusters defined in the Garway-Heath map.

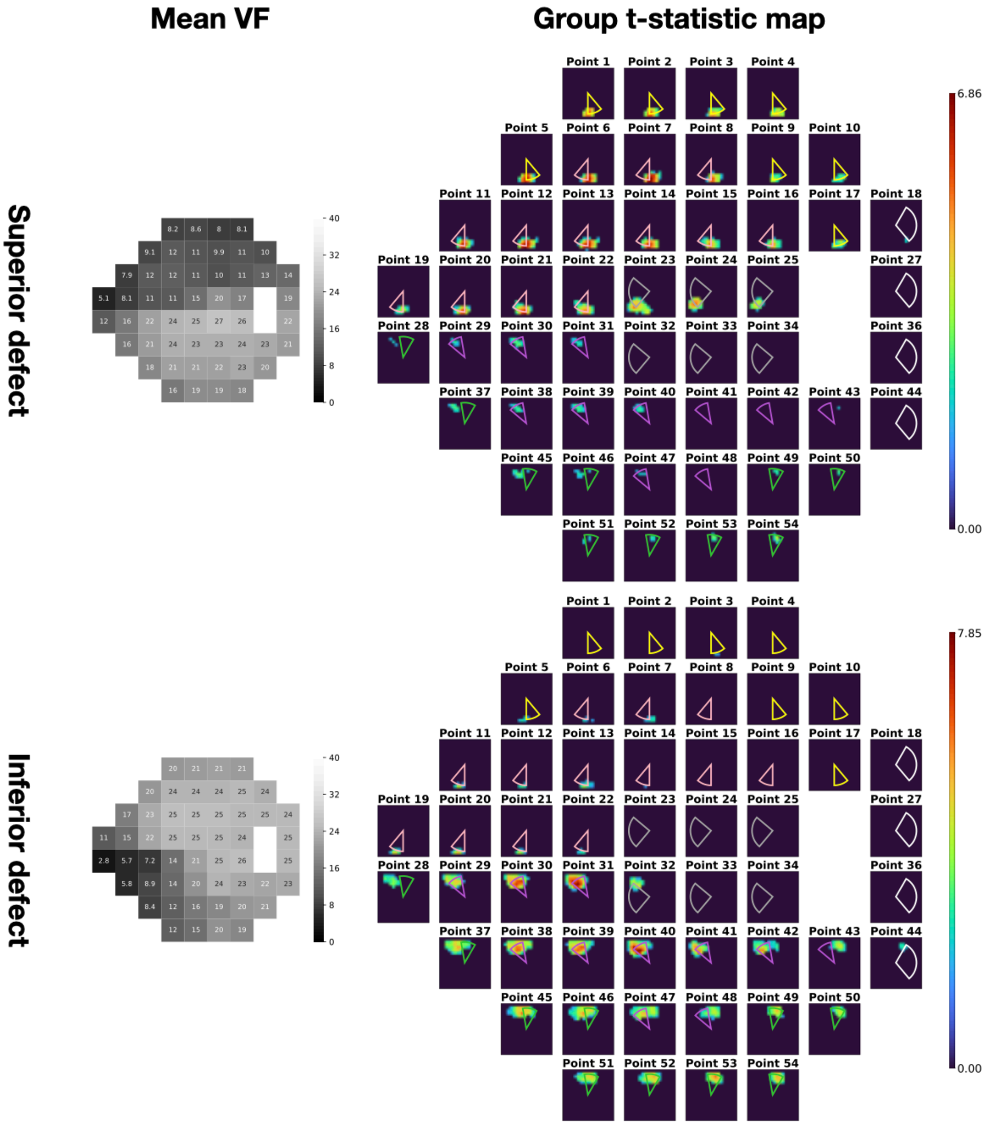
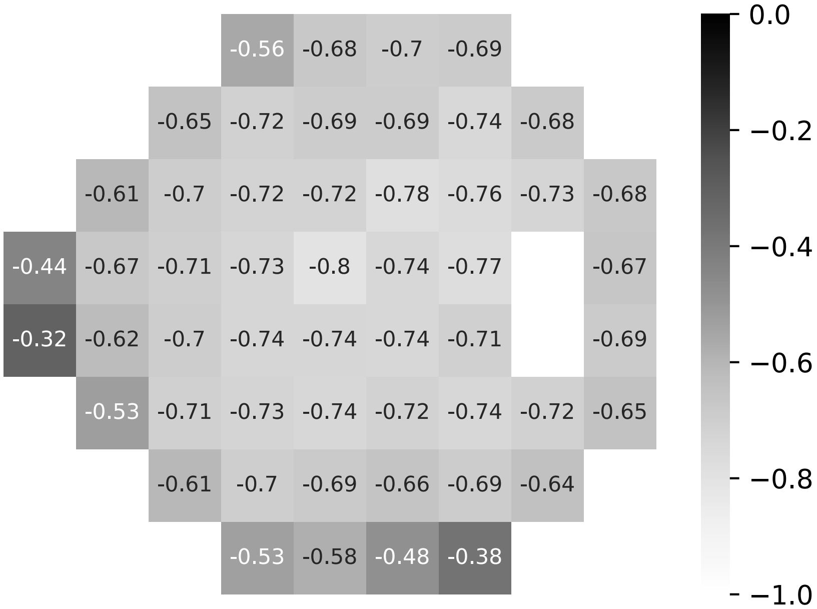


Figure 3.8. Point-wise Pearson's correlation between saliency magnitude and VF MD.



3.3.3 Effect of disease severity

Previous studies have reported that the correlation between structure and function varies with the severity of the disease [1, 33, 50, 54]. In line with these findings, our study also demonstrated a connection between the averaged saliency across the entire volume and the severity of VF damage. Figure 3.8 illustrates the point-wise Pearson's correlation between saliency and VF MD for all subjects in the test set. It was observed that the saliency exhibited a negative correlation with VF MD, indicating a stronger association between saliency and MD when defects are more severe. As a result, the group t-statistic map of the MD $>$ -6 dB group appears less representative compared to that of the MD $\leq$ -6 dB group. Furthermore, this correlation leads to symmetric spatial patterns for the subgroups with superior and inferior defects, as depicted in Figure 3.7.

3.3.4 Comparison with prior studies

While the application of occlusion analysis to visualize the effects of OCT on VF prediction is not groundbreaking [21,53], the majority of these studies primarily center around confirming the accuracy of the proposed DL model. For instance, Christopher et. al. [21] employed a DL model to predict averaged function measurements of VF sectors as defined in the Garway-Heath map. They utilized occlusion analysis to generate a structure-function map for individual cases. Although their map demonstrated specific sectoral structure-function relationships, such as the model's emphasis on superior ONH structures to predict function in the inferior and inferior nasal VF sectors, and vice versa, it was not specifically tailored to assess the broader trend of spatial mapping between structure and function. Its primary objective was to establish the validity of the proposed DL model on an individual case basis. Similarly, Kihara et. al. [53] managed to derive a more refined occlusion-based structure-function map with more advanced DL methods that were able to predict point-wiseVF. However, the resulting map remained specific to individual cases and thus could not represent the overarching structure-to-function trend as effectively as our group saliency map did.

In another study, Mariottoni et. al. [66] developed a DL-based spatial structure-function mapping by simulating localized peripapillary RNFL defects and feeding the resulting thickness profile into a pre-trained CNN model. The identified pattern exhibited agreement with previous maps such that the RNFL defects simulated on the temporal superior and temporal inferior regions led to arcuate VF defects in the inferior and superior hemifield, respectively. However, it is important to note that the derived map remained sectoral in nature and was confined to the peripapillary sampling circle, lacking a comprehensive representation of the entire structure-function relationship. Also, their method relied on RNFL segmentation.

3.4 Conclusion

In summary, this study utilized novel DL methods to explore point-by-point spatial relationships between structure and function without relying on prior knowledge or segmentation of OCT volumes, potentially free from bias. The revealed spatial correlations offer detailed and specific mapping that is consistent with previous studies, highlighting the potential of machine learning in establishing intricate structure-function relationships.

However, there are several limitations to consider in this study. Firstly, the accuracy of the saliency maps relies on the CNN's ability to accurately predict VF sensitivities from SDOCT volumes. Uncommon defect patterns, such as early onset patterns, may not be well captured by the network during training, leading to underappreciation of certain structure-function relationships. Additionally, VF test results in glaucoma patients are susceptible to random noise and subjectivity [26, 56, 67], which inherently reduces the accuracy of the model in predicting VF sensitivities. Another limitation is the use of naive registration in this study. Although the derived map demonstrates high resolution and aligns well with known clinical knowledge and understanding of structure-function relationships, the naive registration does not account for refractive errors or other morphological variables like disc-foveola angle. These variations can contribute to differences in the spatial relationship between VF test points and corresponding ONH regions. Advanced registration techniques are needed to uncover more subtle spatial relationships and fully leverage the three-dimensional nature of the ONH.

Chapter 4

GCIPL Thickness Map Prediction for Glaucoma Progression

Chapter 2 and Chapter 3 have covered two applications of applying deep learning (DL) algorithms to understand three dimensional (3D) volumetric optical coherence tomography (OCT) data for improving glaucoma diagnosis and management. In this chapter, we will investigate another data modality, temporal sequences with spatial dependencies, in another important glaucoma application, forecasting glaucomatous progression.

4.1 Introduction

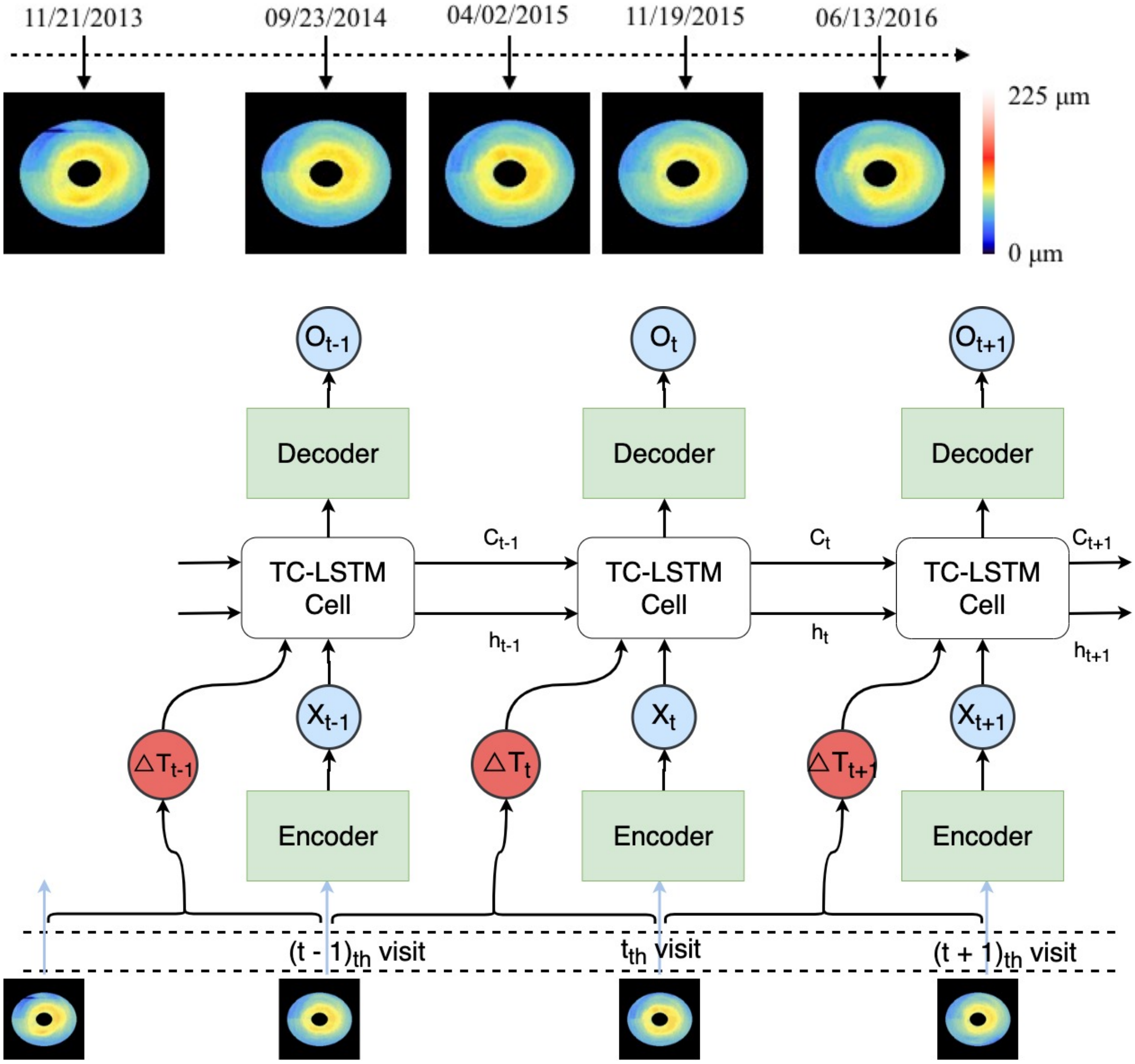
For clinical management of slowly progressing diseases such as glaucoma, early diagnosis and longitudinal progression monitoring are essential [100]. The accelerated retinal ganglion cell loss is a characteristic feature of glaucoma progression and is often associated with functional damages [98]. Measurements performed on OCT scans, especially the macular Ganglion Cell–Inner Plexiform layer (GCIPL) thickness, are clinically used as a biomarker for diagnosis and monitoring of glaucoma [10]. Clinical progression analysis on GCIPL uses summarized numbers (mean values on global or sectoral measurements) only.

two dimensional (2D) thickness maps are used just for subjective assessment as a supplement. 2D thickness maps may allow clinicians to pick up small changes earlier because summarized numbers can easily wash out such small changes. Also, the spatial pattern of GCIPL thickness often contains useful features to detect subtle potential progression [98]. So in this study, we aim to utilize both spatial and temporal information to predict the progression of glaucoma regarding 2D GCIPL damages. Following the clinical convention, we use 4 visits of GCIPL thickness maps as baselines to predict the map of the 5th visit based on those baselines.

Machine learning has been widely used in the diagnosis and monitoring of glaucoma. Song et al. [100] predicted mean circumpapillary retinal nerve fiber layer (RNFL) thickness and visual field index (VFI) via a 2D continuous-time hidden Markov model. Yoshida et al. [121] trained a support vector machine to predict monocular visual field (VF) from RNFL, GCIPL, and RNFL + GCIPL thickness, then calculatedvisual field index (VFI) by merging bilateral simulated monocular VF. Yousefi et al. [123] detected glaucoma progression by clustering VFI into normal, early glaucoma, and advanced glaucoma via a Gaussian mixture model with expectation maximization. Yousefi et al. [122] fed longitudinal feature vectors of the L1 norm of the data at the baseline and at each follow-up visit to classifiers to classify each eye as stable or progressed over time. Maetschke et al. [64] proposed a 3D CNN to classify eyes directly from raw OCT scans. These explorations, however, lie only in utilizing and predicting summarized numbers of measurements, which reveal no spatial information, or simple classification.

Recent advances in Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) provide useful insights into understanding temporal sequences [4, 35, 112]. Convolutional Long Short-Term Memory (cLSTM) was used to better capture spatiotemporal correlations [96]. However, RNN and LSTM implicitly assume equally sampled inputs, which might cause problems when dealing with unequally sampled longitudinal medical data. Neil et al. [73] extended the LSTM unit by adding a time gate to model timestamps,

Figure 4.1. Top: an example of a GCIPL thickness map sequence in false color. Hot color represents thick GCIPL while cool color indicates thin GCIPL. **Bottom: our pipeline.** The inputs are the 2D GCIPL thickness map at each visit and the time interval between 2 successive visits. TC-LSTM cell contains a learnable time penalty function that handles irregular time intervals.



called Phased LSTM. Though Phased LSTM has advantages over standard LSTM in applications that require precise timing of updates, it might fail to model very sparse samples because the model is only updated by samples lying in the model's active state during training. Aware of the problem, Zhu et al. [128] proposed two different versions of time gates to model users' actions in recommendation systems. The first version exploited time intervals to capture the short-term and long-term interests. The second version used two time gates, one for capturing short-term interests for current recommendations and the other for modeling long-term interests for the latter. Baytas et al. [8] proposed Time-Aware LSTM which added a fixed scalar function to handle irregular elapsed time in longitudinal patient records. However, this model has limitations in modeling the relationship between sequence elements since the time penalty function is prefixed. Inspired by [8] and [96], we propose a model which leverages the time-aware Convolutional Long Short-Term Memory (TC-LSTM) in an auto-encoder-decoder to make full use of both spatial and temporal information in GCIPL thickness maps, as shown in Figure 4.1 (*bottom*). We modify the Time-Aware LSTM unit with a learnable time penalty function and replace the gates with convolutional gates so that it can also model spatial information. The contributions of this study are as follows:

- A novel application of LSTM in predicting glaucoma progression. This is the first work to predict 2D GCIPL thickness maps from multiple past maps.

- Our proposed model equips LSTM with carefully designed time gates and convolutional operations. So it can not only utilize the interval information between objects to better model temporal correlations but also learn spatial features in spatial-temporal sequences. This model is not limited to GCIPL thickness map prediction, but can be applied to all unequally sampled spatiotemporal sequences.

- Our proposed models are trained and evaluated on both real-world data and synthesized data. Experiments show the superiority of the proposed method over traditional methods.

4.2 Methods

4.2.1 Time-aware convolutional LSTM

cLSTM [96] has been proven to be powerful in modeling spatiotemporal relations. cLSTM replaces full connections in input-to-state and state-to-state transitions in standard LSTM with convolution operators so that it can encode the spatial information as well as the temporal correlation. The key equations are shown below:

$$i_t = \sigma(W_{xi} * \mathscr{X}_t + W_{hi} * \mathscr{H}_{t-1} + b_i) \tag{4.1}$$

$$f_t = \sigma(W_{xf} * \mathscr{X}_t + W_{hf} * \mathscr{H}_{t-1} + b_f) \tag{4.2}$$

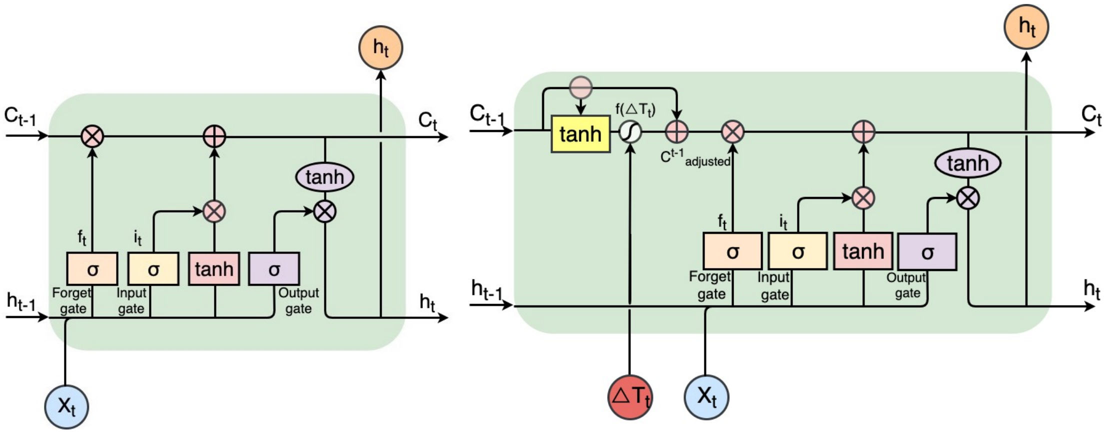
$$\mathscr{C}_t = f_t \circ \mathscr{C}_{t-1} + i_t \circ tanh(W_{xc} * \mathscr{X}_t + W_{hc} * \mathscr{H}_{t-1} + b_c) \tag{4.3}$$

$$\mathscr{H}_t = \sigma(W_{xo} * \mathscr{X}_t + W_{ho} * \mathscr{H}_{t-1} + W_{co} \circ \mathscr{C}_t + b_o) \tag{4.4}$$

where $*$ denotes convolution operation and $\circ$ denotes the Hadamard product. However, cLSTM implicitly assumes that the input sequences are uniformly sampled in time.

TC-LSTM extends cLSTM to nonuniformly sampled 2D sequence data by incorporating the sampling intervals into the model. TC-LSTM first decomposes memory into long-term memory and short-term memory. Then a time penalty is applied to the decomposed short-term memory which transforms the elapsed time between two sampling points as a scalar to penalize the short-term memory without losing the global profile of changes. In other words, the longer time elapsed, the smaller the effect of previous memory is on the current output while the long-term memory plays a more significant role in predicting current output. The

Figure 4.2. Architecture of (a) Standard cLSTM cell and (b) TC-LSTM cell.



(a)

(b)

key equations are:

$$\mathscr{C}^{t-1}_{short} = tanh(W_{cs} \circ \mathscr{C}^{t-1} + b_s) \tag{4.5}$$

$$\mathscr{C}^{t-1}_{long} = \mathscr{C}^{t-1} - \mathscr{C}^{t-1}_{short} \tag{4.6}$$

$$\mathscr{C}^{t-1}_{adjusted} = f(\Delta \mathscr{T}_t) \mathscr{C}^{t-1}_{short} + \mathscr{C}^{t-1}_{long} \tag{4.7}$$

$$i^t = \sigma(W_{xi} * \mathscr{X}^t + W_{hi} * \mathscr{H}^{t-1} + b_i) \tag{4.8}$$

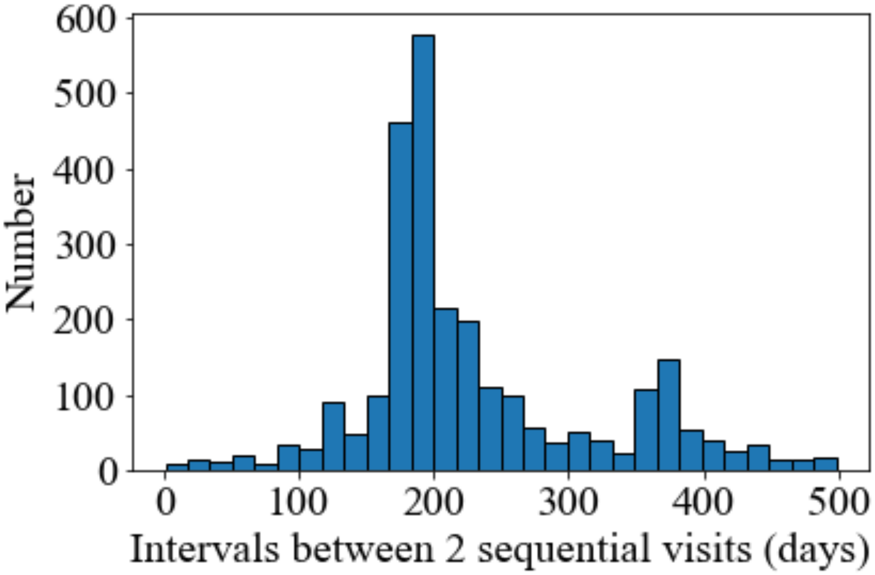
$$f^t = \sigma(W_{xf} * \mathscr{X}^t + W_{hf} * \mathscr{H}^{t-1} + b_f) \tag{4.9}$$

$$\mathscr{C}^t = f^t \circ \mathscr{C}^{t-1}_{adjusted} + i^t \circ tanh(W_{xc} * \mathscr{X}^t + W_{hc} * \mathscr{H}^{t-1} + b_c) \tag{4.10}$$

$$\mathscr{H}^t = \sigma(W_{xo} * \mathscr{X}^t + W_{ho} * \mathscr{H}^{t-1} + W_{co} \circ \mathscr{C}^t + b_o) \tag{4.11}$$

Equation 4.5 extracts the short-term memory from the previous memory, while equation 4.6 gets the long-term memory by subtracting short-term memory from the previous memory. $f(\Delta \mathscr{T}_t)$ in equation 4.7 is a scalar function that maps sampling intervals $\Delta \mathscr{T}_t$ between samples at t-1 and t to [0, 1] and monotonically decreases as sampling intervals increase. We call this scalar function a time penalty function. Hence, equation 4.7 adjusts the previous memory by penalizing the short-term memory while preserving the long-term memory. Equation 4.8 - 4.11 are the same as cLSTM but operate on the adjusted previous memory. Figure 4.2 shows the architecture of standard cLSTM cell and TC-LSTM cell.

Figure 4.3. Histogram of intervals between 2 sequential visits.



4.2.2 Time penalty function

Different time penalty functions can be chosen according to typical sampling intervals of time in the underlying application. In our case, the time interval varies from 1 month to 2 years as shown in Figure 4.3. Therefore we use a unit of day. We experiment with three types of penalty functions, given in 4.12 - 4.14. The function in 4.12 is fixed, and the inverse of log mapping is used instead of simple inverse $(f(\Delta\mathscr{T}_t) = 1/\Delta\mathscr{T}_t)$ because the time intervals are numerically large. In order to introduce more flexibility in the model, we also consider 4.13 and 4.14, whose parameters are learned during training.

$$f(\Delta\mathscr{T}_t) = \frac{1}{log(e+\Delta\mathscr{T}_t)} \tag{4.12}$$

$$f(\Delta\mathscr{T}_t) = \frac{1}{a\Delta\mathscr{T}_t+b} \tag{4.13}$$

$$f(\Delta\mathscr{T}_t) = \frac{1}{e^{a\Delta\mathscr{T}_t+b}} \tag{4.14}$$

4.3 Experiments

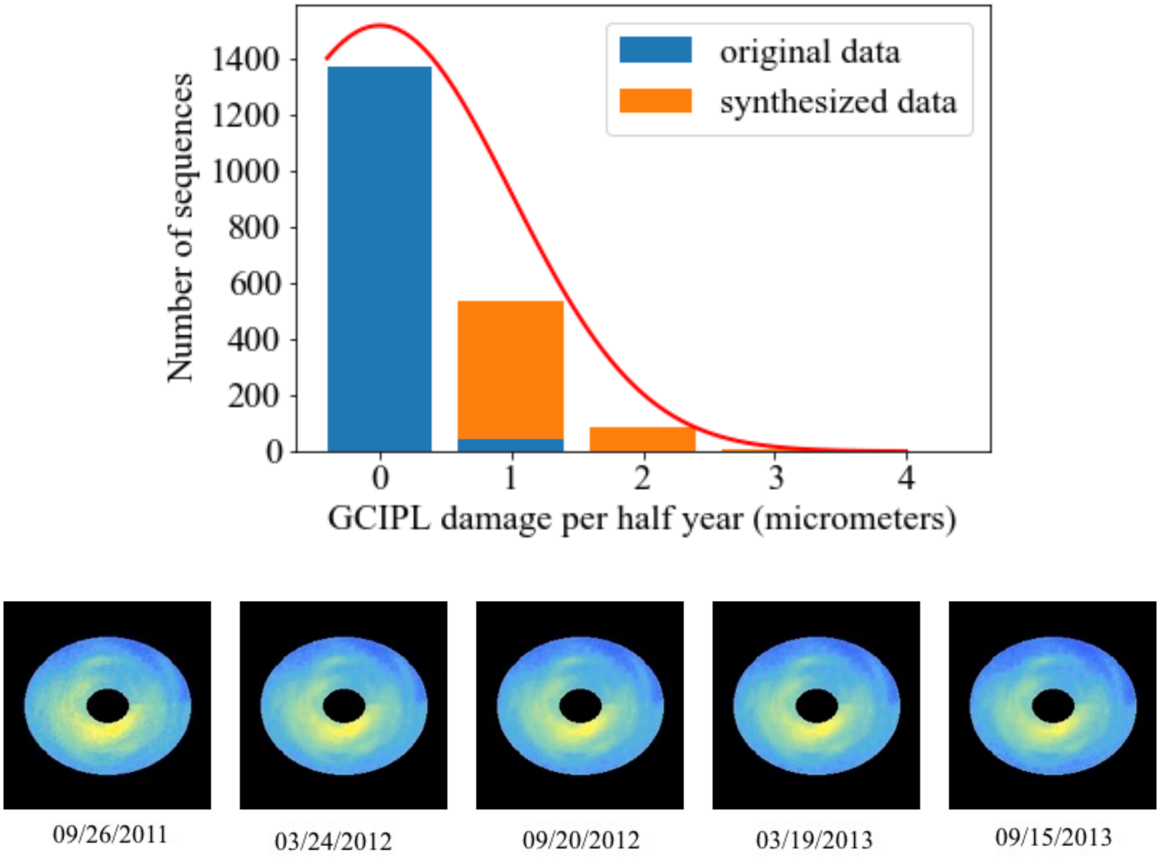
4.3.1 Dataset and experiment settings

The dataset contains 346 eyes from 191 patients. The average number of visits is 9.49 ± 3.39, and the average follow-up period is 5.85 ± 2.01 years. Each eye is imaged by a commercially available spectral-domain OCT (SD-OCT) device (Cirrus OCT; Zeiss). The GCIPL thickness map is 200×200 pixels and is generated by measuring GCIPL thickness across a $6 \times 6 mm^2$ area centered at the fovea. We randomly split the patients into the train set and the test set with a 9:1 ratio. By setting the max time interval to 500 days and extracting 5-visit sequences for every patient, we get 1648 5-visit sequences in total. Each map is registered to the same fovea position and normalized to [0,1]. We trained the network using the ADAM optimizer [55] with a learning rate of 0.0002 and a batch size of 16. We experimented with 3 different loss functions, L2 loss, L1 loss, and L2 + L1 loss.

4.3.2 Data simulation

83.2% of patients are under a quite stable state with the average GCIPL damage less than $2\mu m$ per year. The average damage over the whole dataset is $0.84\mu m$ per year. So we simulated progressing cases according to the GCIPL damage patterns of glaucoma, diffuse damage, and hemifield damage. The progressive GCIPL thinning is detected most frequently at 2.08 mm from the fovea and extends in an arcuate shape [98]. For diffuse damage, GCIPL around the fovea gets thinning more or less equally. For hemifield damage, usually temporal side usually gets more pronounced thinning while the nasal side is preserved. We set the time interval to 6 months and randomly applied diffuse or hemifield damage (both superior and inferior cases) with equal probability to randomly selected real maps 4 times to create 5-visit sequences. The average thickness losses per half year match a standard normal distribution. In total, 576 progressing sequences are generated. Though

Figure 4.4. Top: histogram of average GCIPL damage per half year. Bottom: example of a synthesized sequence.



these data are artificially generated, it is confirmed by expert ophthalmologists that the synthesized data contains similar characteristics as real data, and cannot be distinguished from the real patient data. Figure 4.4 shows an example of a synthesized sequence.

4.3.3 Compared methods

Given the lack of prior published work for the prediction of 2D GCIPL thickness maps, we compare TC-LSTM to a linear regression (LR) baseline and a regular cLSTM without considering time interval differences. LR also takes time into account as time is the variant of LR. For the LR baseline, we first use principle component analysis (PCA) directly to all maps in the training data to identify a set of principal components so that the mean square error (MSE) is 1% of the original signal variance. Each 200x200 map is then represented by the projection coefficients associated with these 962 principal components. Finally, we

Table 4.1. Method comparison.

Method	MSE	PSNR	SSIM
Copy last	0.00106	30.27	0.947
LR	0.00061	32.52	0.967
cLSTM (L2)	0.00053	33.93	0.939
TC-LSTM (L2 & Eq. (13))	**0.00049**	34.08	0.966
TC-LSTM (L1 & Eq. (13))	0.00050	34.18	**0.973**
TC-LSTM (L2+L1 & Eq. (13))	**0.00049**	**34.45**	0.972
TC-LSTM (L2 & Eq. (12))	0.00052	33.83	0.972
TC-LSTM (L2 & Eq. (14))	0.00050	34.10	0.965

apply LR to predict the PCA coefficients of the future image from the coefficients of the past 4 images. Then, the predicted 5th-visit GCIPL thickness map is reconstructed from the predicted PCA coefficients.

MSE, peak signal noise ratio (PSNR), and structure similarity index (SSIM) [114] are used to quantitatively evaluate the quality of predicted GCIPL thickness map; lower MSE and higher PSNR and SSIM indicate better results. However, it is well known that existing numerical measures cannot reflect human perception well, and temporal coherence cannot be evaluated from these numerical metrics. From our observation, the visual difference in the quality of the predicted maps via our method and those via linear regression is more significant than the numerical difference. The measurements do not represent subjective impressions well (the differences among different methods are small) because the subtle thinning area changes among different time points (as shown in Figure 4.5) cannot be evaluated by those summarized metrics well. Therefore, we also perform a subjective evaluation. Three independent expert ophthalmologists are asked to choose the best prediction out of three methods based on the whole map sequence and the similarity between the ground truth and the predicted map.

Table 4.2. Subjective Rating Results. The number in a column indicates the percentage of maps predicted by a particular method that is rated the best. TC-LSTMs are trained with a penalty function of Eq. (13). There are 2 sub-methods for the TC-LSTM approach, one trained with L2 loss and the other trained with L2 + L1 loss. The preferences of the particular TC-LSTM method for each rater are 50% and 42.8%, 86.2% and 7.2%, and 59.6% and 35.8% respectively.

Rater	TC-LSTM approach	Linear reg. approach
Rater 1	**92.8%**	7.2%
Rater 2	**93.4%**	6.6%
Rater 3	**94.8%**	5.2%

4.3.4 Quantitative results

As shown in Table 4.1, our method (TC-LSTM trained with L2+L1 loss and time penalty function of Eq.(13)) outperforms both LR and cLSTM baselines regarding all three metrics. By properly handling the time interval difference, the TC-LSTM approach significantly improves the SSIM of the prediction compared to cLSTM, while the SSIM of cLSTM is the worst among all methods, even worse than that of directly copying from the last visit. Results of different time penalty functions show that by introducing learnable parameters into the time penalty function, the SSIM of the prediction is significantly improved, while PSNR is also improved. The Wilcoxon signed-rank test shows a significant difference between our best TC-LSTM results and LR results (MSE 0.00049 vs. 0.00061, $p<0.001$, and PSNR 34.45 vs. 32.52 dB, $p=0.035$).

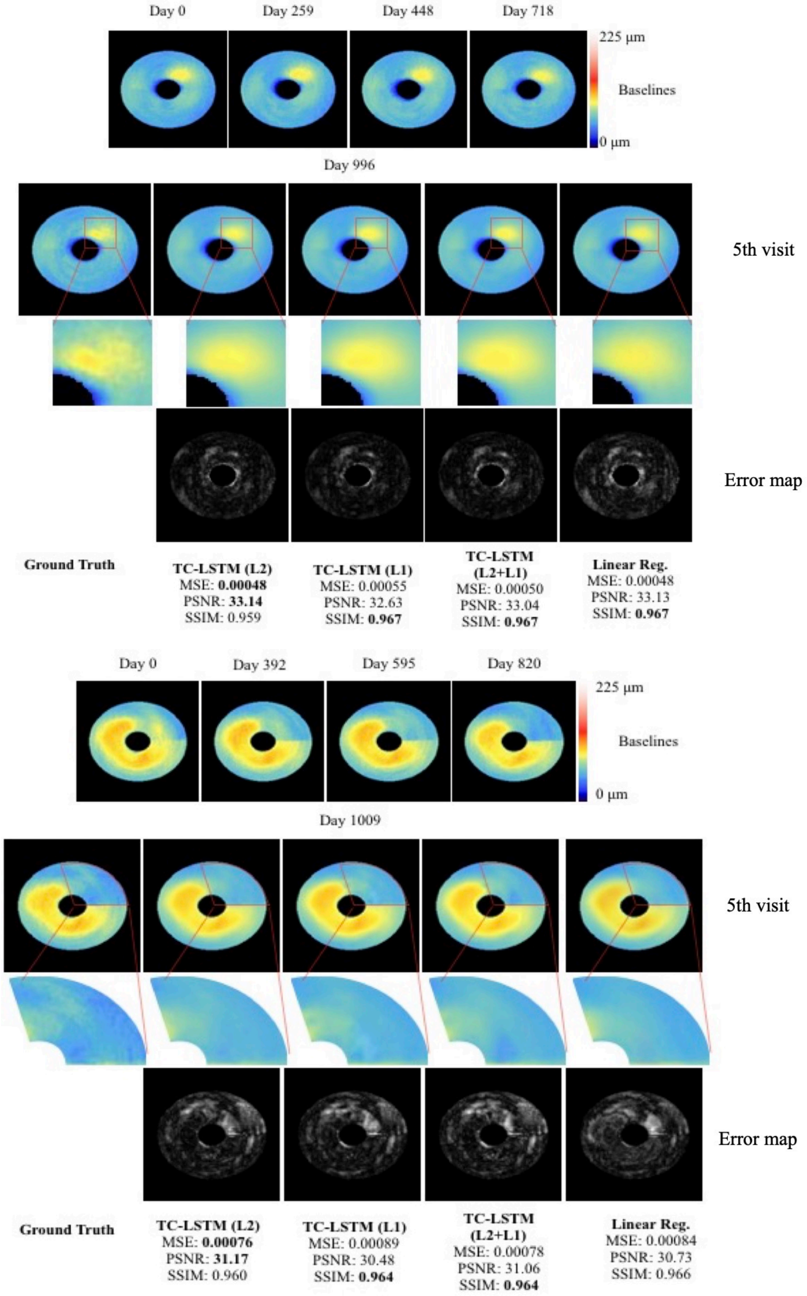
4.3.5 Subjective evaluations

As Table 4.2 shows, TC-LSTM-based methods are significantly preferred over LR. All raters chose the TC-LSTM approach as the best predictor over 90% of real test sequences. They agree on TC-LSTM (L2) as the best predictor in 42 out of 151 cases and TC-LSTM (L2 + L1) as the best predictor in 2 out of 151 cases. Figure 4.5 shows 2 real examples, one for a progressing case with moderate glaucoma and the other for a stable case with advanced glaucoma. Both TC-LSTM and LR can learn reasonable overall thinning patterns. However, TC-LSTM contains more details while LR produces blurry results. Moreover, the thinning area is more accurate in the prediction of TC-LSTM compared to that of LR. For the top example, TC-LSTM trained with L2+L1 loss is picked by the expert as the best. In the superior temporal region (red box area), the pattern of the thinned layer in TC-LSTM (L2+L1) resembles the ground truth most closely, while LR does not capture the subtle thinning area changes. For the bottom example, TC-LSTM trained with L1 loss looks the best because it shows the tiny heated area properly without showing an overly smoothed appearance. However, at the same time, none of the methods shows the notch in the middle of the ground truth image.

4.4 Conclusion

In this study, we propose an end-to-end model, TC-LSTM, for a novel application of predicting 2D GCIPL thickness maps. Our model is designed to handle irregularly sampled spatiotemporal sequence modeling. Experiments show that this approach is able to predict reasonable GCIPL thinning patterns of glaucoma and outperforms linear regression for GCIPL thickness map prediction.

Figure 4.5. Examples results. *Top:* a progressing case with moderate glaucoma. *Bottom:* a stable case with advanced glaucoma. Images in the last rows are difference maps between predictions and ground truth.



Chapter 5

Video Interpolation

In Chapter 4, we discussed spatiotemporal sequence prediction in medical scenes. In this chapter, we will explore a different example of spatiotemporal sequence, natural videos, which is ubiquitous in daily life.

5.1 Introduction

Video interpolation, which aims to generate intermediate frames between given prior (or left) and post (or right) frames, is widely applied in video coding [118] and video frame rate conversion [14]. However, natural videos include complicated appearance and motion dynamics, e.g., various object scales, different viewpoints, varied motion patterns, object occlusions, and dis-occlusions, making interpolation of realistic frames a significant challenge.

Flow-based methods have been proven to work well in video interpolation [6, 7, 51, 63, 119]. Many state-of-the-art methods first use an optical flow estimator to obtain optical flow between given frames, and then infer the optical flow between the missing middle frame and the left and right known frames, respectively, by prefixed motion assumptions such as linear motion [6, 51, 75] or quadratic motion [119]. The middle frame is then obtained by backward warping input frames using the estimated optical flows. Such approaches are prone to flow errors caused by adopted flow estimators and errors in the motion assumption. Thus, additional flow correction networks [119] or additional information such as depth [6] are usually required to refine the initial interpolated optical flows, leading to sophisticated models. Moreover, training such models requires ground truth optical flow or depth information, which is expensive to obtain in large quantities.

Though flow-based methods have achieved great success in video interpolation, they are prone to errors and face the challenge of complicated dynamic scenes including nonlinear motions, lighting changes, and occlusions. Recently, deformable convolution (DConv) has been investigated in video interpolation to warp features and frames [36,59]. DConv produces multiple offsets for each pixel to be interpolated with respect to each input frame, and uses a weighted average of these offset pixels in the previous (or future) frame to predict the target pixel. When the filter size of DConv is 1x1 and the filter coefficient is 1, DConv offset is the same as optical flow. When the filter size is larger than 1, DConv performs many-to-one weighted warping, and thus the offsets can be considered as many-to-one flows. Generally, DConv offsets are more robust than single optical flow. Furthermore, DConv filter coefficients enable the model to produce more complex transformations. However, the increased degree of freedom of DConv makes the model hard to train.

To alleviate the above issues, we propose a Pyramid Deformable Warping Network (PDWN) to perform coarse-to-fine frame warping. The coarse-to-fine structure has been proven to be powerful in optical flow estimation [48, 89, 103]. In video interpolation, however, relatively few approaches explored the coarse-to-fine strategy. Amersfoort and Shi [110] proposed a multi-scale generative adversarial network to generate the predicted flow and the synthesized frame in a coarse-to-fine fashion. Zhang et. al. [125] designed a recurrent residual pyramid architecture to refine optical flow using a shared network across pyramid levels. Other methods, despite the usage of multi-scale features, only generate one-stage optical flow [6,36,63]. In our work, we exploit the advantages of the warping strategy and cost volume in addition to the pyramid structure to estimate DConv offsets from coarse to fine.

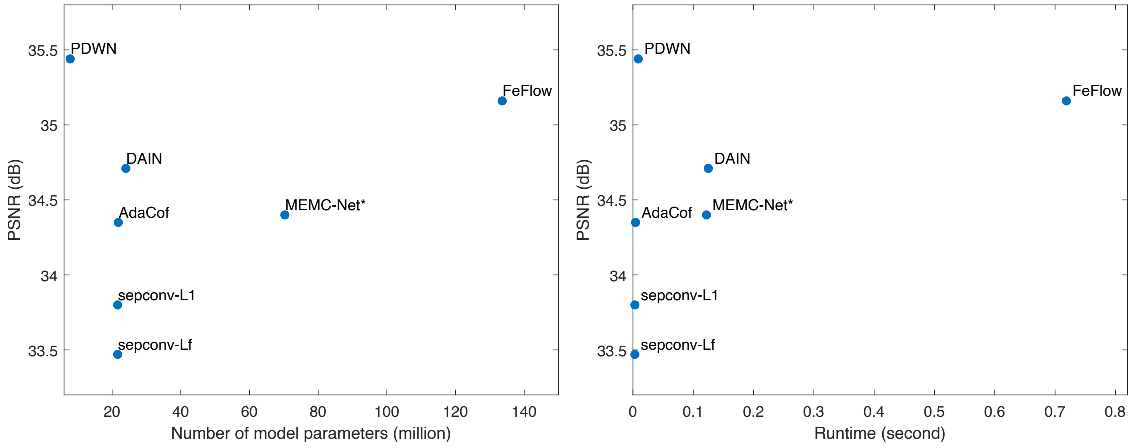
The proposed network follows a pyramid structure that extracts features at various resolution scales from each input frame. At every pyramid level, DConv is adopted to warp features from the past and future frames towards the middle frame, and a matching cost volume under different additional offsets between two warped features is constructed and exploited to infer residual DConv offsets. By warping features with the obtained offsets and passing the cost volume to the next pyramid level, the network refines the estimated offsets from coarse to fine. We demonstrate that such a methodology for video interpolation generates more realistic frames without requiring additional information such as ground truth optical flow information or depth during training. Our proposed network greatly reduces the number of model parameters and the inference time, while achieving better or on-par performance compared to state-of-the-art models as shown in Figure 5.1. Furthermore, our proposed approach can be extended to using multiple input frames easily, and using four instead of two frames as input leads to significantly improved interpolation results.

5.2 Related work

5.2.1 Video interpolation

Video interpolation has been extensively explored in the literature [6, 7, 36, 51, 59, 63, 75, 77, 119, 120]. Prior methods can be grouped into two categories: kernel-based approach and flow-based approach. Kernel-based approaches [59, 76, 77] estimate convolution kernel parameters to hallucinate intermediate frame. However, kernel-based approaches typically fail in cases with large motions unless very large filter kernels are used, and suffer from large computational loads. Flow-based approaches estimate the optical flow to warp pixels to synthesize the target frame. Super SloMo [51] adopted one UNet to estimate optical flow between two input frames, and another UNet to correct the linearly interpolated

Figure 5.1. Accuracy / efficiency tradeoff for video interpolation on Vimeo-90K dataset: PDWN in comparison to previous work. Left: PDWN outperforms state-of-the-art methods in both accuracy and model size. Right: PDWN reaches the best balance between accuracy and runtime. PDWN++ is the enhanced PDWN model with input normalization, network improvements, and self-ensembling. PDWN++ further improves the performance with a small cost of model size and nearly 8 times of the runtime. The runtime is the time needed to interpolate one frame on GeForce RTX 2080 Ti GPU card.



(a) Accuracy / size tradeoff

(b) Accuracy / runtime tradeoff

flow vector. Beyond linear motion assumptions, QuaFlow [119] adopted PWC-Net [103] to estimate optical flow between input frames. Then the quadratically interpolated flow was refined through a UNet. MEMC-Net [7] estimated both motion vectors and compensation filters through Convolutional Neural Network (CNN). Note that four input frames are required for QuaFlow to construct a quadratic model. Instead of bilinear interpolation, MEMC proposed an adaptive warping layer based on optical flow and compensation filters to reduce blur. Based on MEMC-Net, DAIN [6] used depth information estimated by a pre-trained hourglass architecture [62] to detect occlusions. Different from the above methods, Softmax Splatting [75] estimated forward flow using an off-the-shelf optical flow estimator and designed a differentiable way to do forward warping. Though flow-based methods can generate sharp frames, inaccurate flow estimation often leads to severe artifacts. Unlike the methods described above, our method directly estimates the ”flow” between given input frames and the unknown middle frame without assuming the trajectory is linear or quadratic or has other parametric forms. And we estimate many-to-one ”flow” which is more robust compared to single optical flow. Furthermore, we estimate the flows in a coarse-to-fine manner, to efficiently handle large motions.

5.2.2 Pyramid structure and the cost volume

Pyramid structure has been proven to be powerful in optical flow estimation. Ilg et al. [48] achieved state-of-the-art performance by stacking several UNets into a large model, called FlowNet2. To reduce the over-fitting problem caused by large models, SpyNet [89] incorporated two classical principles, pyramid structure, and warping, into deep learning. A spatial pyramid network was constructed for each of the two frames, and it estimated the flow in each scale and warped the second image to the first one at each scale repeatedly to reduce motion between the two images. PWC-Net [103] further explored the trade-off between accuracy and model size. Instead of image pyramids, PWC-Net constructed feature pyramids that are invariant to shadows and lighting change. Partial cost volume is used to represent matching costs associated with different disparities. Inspired by classical pyramid energy minimization in optical flow algorithms, RRPN [125] designed a recurrent residual pyramid architecture for video frame interpolation to refine optical flow using a shared network for every pyramid level. Following the above methods, we also exploit the advantages of classical principles of optical flow – the pyramid structure, multi-scale warping, and cost volume. Different from RRPN, we replace the flow estimation in each scale with the estimation of many-to-one offset maps through the use of deformable convolution filters, significantly reducing artifacts that are associated with occasional wrong flow estimates. Furthermore, cost volume is incorporated into our model non-trivially. We demonstrate that the cost volume between the warped features of the two known frames can provide useful information for estimating the flow between the unknown middle frame and the known prior and post frames.

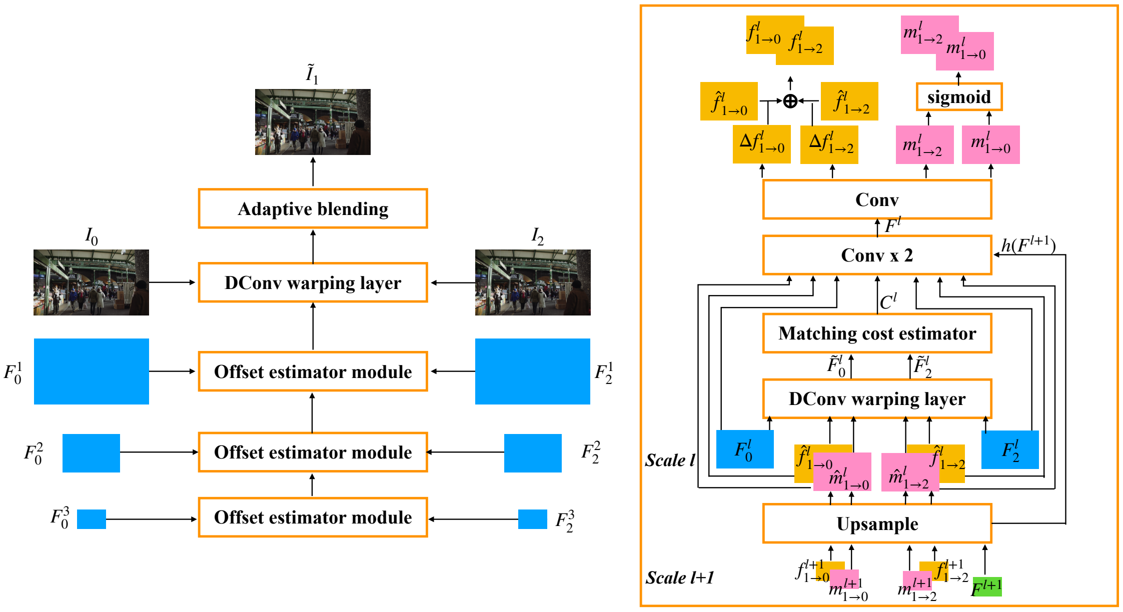
5.2.3 Deformable convolution

DConv operation [23] is originally proposed to overcome the limitation of CNN due to fixed filter support configuration and to enhance the transformation modeling capacity of CNN. It estimates a set of K offsets at each pixel and a global filter (non-spatially varying) with K coefficients to be applied for the K offset pixels. Zhu et. al. [127] further improved DConv by adding spatially adaptive modulation weights to modulate the global filter coefficient associated with each offset. The improved DConv thus has the ability to vary the attention to different offset pixels. Recognizing that DConv can be viewed as many-to-one weighted backward warping, FeFlow [36] used DConv to align input features from two known frames and fused aligned features to synthesize the middle frames. AdaCoF [59] constructed a UNet to estimate both local filter weights and offsets for each target pixel to synthesize output frames. We have found that learning a global filter plus spatially-varying modulation weights as in [127] is better than directly estimating locally adaptive filters. Different from FeFlow and AdaCoF [59] that estimate the DConv offsets directly in the original image resolution, we perform offset estimation and feature alignment in a coarse-to-fine successive refinement manner. Specifically, we successively refine DConv offsets from the coarser scales to the finer scales. We further utilize the cost volume computed from two aligned features at each scale to improve the accuracy of the offset update.

5.3 Methods

The structure of PDWN is shown in Figure 5.2. Given two input frames I_0 and I_2, we aim to synthesize the intermediate frame I_1 by gradually warping features of input frames to the middle frame using DConv. First, we construct a feature pyramid for each input frame using a shared feature extractor. Second, we generate the offsets and the associated modulation weights between each input frame and the middle frame, and then warp features of both input images toward the middle frame. This operation is taken at every level of the

Figure 5.2. (a) The architecture of PDWN. Given the past frame I_0 and the future frame I_2, PDWN first generates two feature pyramids. Then DConv offsets $f^l_{1\to0}, f^l_{1\to2}$ and associated modulation weights $m^l_{1\to0}, m^l_{1\to2}$ are generated from the coarsest scale to the finest scale. Finally, two warped frames are adaptively blended to synthesize the middle frame. **(b) Offset estimator module**. Features of scale l are warped producing $\tilde{F}^l_0$ and $\tilde{F}^l_2$ via DConv with generated offsets and associated modulation weights. The cost volume between $\tilde{F}^l_0$ and $\tilde{F}^l_2$, together with input features, are fed to 2 convolutional layers to refine the next-scale DConv offsets and the associated modulation weights. The above process is repeated until it gets to the finest level.



(a) PDWN

(b) Offset estimator module

feature pyramid to refine the motion. Thus, the estimated DConv offsets, which can be considered as many-to-one flow, are refined from the coarse level to the fine level. Third, at the finest resolution level (same as the input frame), interpolation weight maps between the warped left and right frames are generated to handle occlusions. Finally, following the post-processing scheme of DAIN, we adopt a context enhance network to further enhance the interpolated frame, shown in Figure 4.2.

5.3.1 Shared feature pyramid encoder

A multi-layer CNN is used to construct L-scale pyramids of feature representations for both input frames $\{F_i^l \mid i \in \{0, 2\}, l \in \{1, 2, ..., L\}$. The features at the first scale, F_i^1, have the same spatial resolution as the input frames. The lth scale feature F_i^l is downsampled by a factor of 2 both horizontally and vertically from the $(l-1)$-th scale feature F_i^{l-1}. Each scale consists of two convolution blocks.

5.3.2 Offset estimator module

The Offset estimator module is used in every scale of PDWN. It jointly predicts the DConv offsets from the unknown intermediate frame to the given input frames and the associated modulation weights for each offset in order to warp input frames and features to the intermediate frame.

Deformable warping with spatially-varying modulation coefficients

A deformable convolution filter is specified by a global filter $w(j)$, a set of spatially-varying offsets $f(j, x)$, and modulation coefficients $m(j, x)$, where j denotes j-th location in a filter support $\mathscr{R}$ and x indicates pixel location. The global filter $w(j)$ here is the same convolution filter as regular convolutions except that the sampling is irregular. The support

$$\mathscr{R} = \{(-1, -1), (-1, 0), ..., (0, 1), (1, 1)\}$$

specifies a 3×3 filter in our model. The offset is defined by horizontal and vertical displacements. And every sampling point is associated with a modulation weight. Thus the offset tensor and the modulation tensor have channel dimensions of 18 and 9, respectively. The global filter has the size of 3×3. To use DConv for video interpolation at multiple scales, we generate two sets of offsets and modulation coefficients at scale l, $f_{1\to i}^{l}(j,x), m_{1\to i}^{l}(j,x)$ with $i = 0,2$ indicating the known prior and post frame and $i = 1$ the unknown middle frame. The global filter weights $w^{l}(j)$ are learnt and stay fixed after training for each scale and shared for known input features. Specifically, we generate the warped feature at scale l at pixel x from the original features for frame i as follows:

$$\tilde{F}_i^l(x) = \sum_{j=1}^{|R|} w^l(j) m_{1\to i}^l(j,x) F_i^l(x + R(j) + f_{1\to i}^l(j,x)) \tag{5.1}$$

Cost volume between features warped towards the middle frame

The notion of cost volumes has been widely used in optical flow methods [47,103,124] to provide explicit representation of matching cost under different displacements between two given frames for each pixel. In the PWC method for optical flow estimation, the cost volume is constructed between a warped image and a fixed image. Typically, for each pixel x in one frame, the correlation between the feature at x in this frame and the feature at a displaced location $x + d$ in the other frame is computed, for a finite set of displacements $d \in \mathscr{D}_k(x)$. $\mathscr{D}_k(x)$ is a square neighborhood of pixel x with neighborhood size $k \times k$. In our case, however, a cost volume is calculated between two sets of warped features $\tilde{F}_0^l$ and $\tilde{F}_2^l$ based on the estimated offsets from each known frame to the middle frame, determined in a lower scale. The cost volume indicates the correlation between the features for corresponding pixels in the left and the right warped features under different displacements. Specifically, given $\tilde{F}_0^l$ and $\tilde{F}_2^l$, a cost volume C^l is constructed based on

$$C^l(x_1, x_2) = \frac{1}{k^2} \tilde{F}_0^l(x_1)^T \tilde{F}_2^l(x_2), \quad x_2 \in \mathscr{D}_k(x_1) \tag{5.2}$$

where x_1 and x_2 are pixel indexes. We set $k = 9$, including displacement from -4 to 4 in both horizontal and vertical directions. Thus the cost volume has a channel dimension of 81.

Instead of using a pre-determined way to calculate the matching cost, one can also train a small network (learnt as part of the entire network) $v(\cdot)$ that takes the two warped features and outputs the cost volume:

$$C^l = v(\tilde{F}_0^l, \tilde{F}_2^l) \tag{5.3}$$

We experimented with both approaches, where we used a network with two conv layers for the network $v(\cdot)$.

Multi-scale offset estimation

As shown in Figure 5.2, we estimate the offsets between the middle frame and each of the two input frames from coarse to fine scales with a total of L scales ($L = 3$ in Figure 5.2). DConv offsets are generated within each scale to gradually reduce the distance between two sets of features warped towards the middle frame.

At l-th scale, the offset estimation block first upsamples the estimated offsets $f_{1\to i}^{l+1}$ and modulation weights $m_{1\to i}^{l+1}$ at the lower scale $l + 1$ to the current resolution using bilinear interpolater $h(\cdot)$, yielding

$$\hat{f}_{1\to i}^l = 2 * h(f_{1\to i}^{l+1}) \tag{5.4}$$

$$\hat{m}_{1\to i}^l = h(m_{1\to i}^{l+1}) \tag{5.5}$$

Then it warps the original features F_i^l towards the middle frame based on $\hat{f}_{1\to i}^l$, $\hat{m}_{1\to i}^l$, and the learnt global filter w^l, generating the warped features $\tilde{F}_i^l$ using Eq. (5.1). Then, the offset estimator computes the cost volume C^l between the two warped features using Eq. (5.2). Next, it generates two sets of DConv offsets residuals $\Delta f_{1\to i}^l$ and two sets of modulation weight $m_{1\to i}^l$ from C^l, $\hat{f}_{1\to i}^l$, $\hat{m}_{1\to i}^l$, the original features F_i^l, and the upsampled features

$h(F^{l+1})$ from the features F^{l+1} generated by the offset estimator in the previous scale:

$$\Delta f^l_{1\to i}, m^l_{1\to i} = g(C^l, F^l_i, \hat{f}^l_{1\to i}, \hat{m}^l_{1\to i}, h(F^{l+1})), i = 0, 2 \tag{5.6}$$

where $g(\cdot)$ denotes a three-layer CNN. The final offsets and modulation weights are obtained by

$$f^l_{1\to i} = \hat{f}^l_{1\to i} + \Delta f^l_{1\to i} \tag{5.7}$$

$$m^l_{1\to i} = \sigma(m^l_{1\to i}) \tag{5.8}$$

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{5.9}$$

where $\sigma(\cdot)$ denotes a sigmoid activation function. We can use a small subnetwork (consisting of three conv layers) to estimate the offset fields because the motion between two warped features is usually small. The same process repeats until we complete scale 1.

For the coarsest scale L, the offset estimator only takes the original features in that scale F^L_0 and F^L_2 as input and generates $f^L_{1\to i}$ and $m^L_{1\to i}$ directly.

To summarize, the offset estimator at each scale needs to generate two sets of offset tensors and two sets of modulation tensors, with a total channel dimension of 54.

5.3.3 Adaptive frame blending

Using the estimated offset $f^1_{1\to i}$, modulation weights $m^1_{1\to i}$, and global filter w^1 at scale 1, we warp frame i towards the middle frame, generating two candidate estimates of the middle frames $\tilde{I}_i, i \in 0, 2$. Occlusions often happen due to the movement of objects. Therefore, in order to select valid pixels from two warped reference frames, we design a blending layer that generates a weight map $\alpha(x)$ to average the two transformed frames at position x. The layer is constructed by a three-layer CNN. See Table I, the network takes two warped frames, $\tilde{I}_0$ and $\tilde{I}_2$, and two warped features, $\tilde{F}^1_0$ and $\tilde{F}^1_2$, at first scale of the feature pyramids as input and generates the weight map with a softmax activation applied on the output layer. At position x, the blended frame is

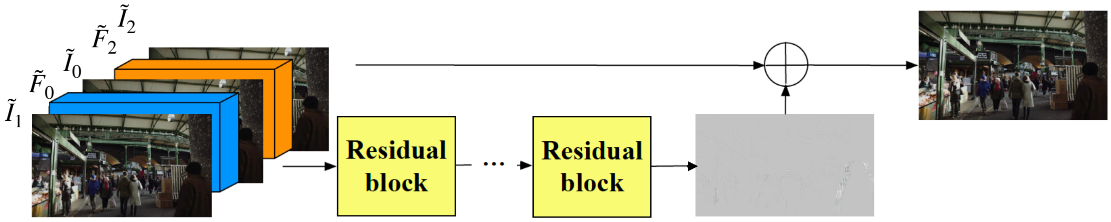
$$\tilde{I}_1(x) = \alpha(x) * \tilde{I}_0(x) + (1 - \alpha(x)) * \tilde{I}_2(x) \tag{5.10}$$

The warped features provide contextual information to estimate the weight map.

5.3.4 Context enhancement network

To generate the final output, we construct a context enhancement network which takes warped images and features at scale 1 as input and outputs a residual image between the unknown ground truth intermediate frame and the blended frame. The network consists of five residual blocks, shown in Figure 5.3.

Figure 5.3. Context enhancement network. After getting the initial synthesized middle frame, warped input frames and warped 1st-level features are fed into five residual blocks to further enhance the contextual details of the synthesized frame.



5.3.5 Extending to four input frames

Quadratic flow [119] shows an improvement in moving trajectory estimation by estimating acceleration information from four input frames. We also extend our model to exploit the information in additional input frames and to estimate the motion more accurately. Our extended model takes four input frames (two previous and two following frames). A pyramid feature encoder is shared between four input frames to generate four feature pyramids.

In the offset estimator, we input four feature maps of the four input frames instead of two in the first conv layer in Figure 5.2. (b). This allows the network to recognize the motion trajectory over a longer temporal scope and yield more accurate offset estimation. In higher scales, we still generate the warped feature maps for the two closest past and future frames using the estimated offsets and modulation weights from the lower scale and determine the cost volume from these two warped features. Then the cost volume is concatenated with four original features of input frames as well as offsets and modulation weights and fed into the next scale to refine offsets and modulation weights in the next scale. Note that even though the input consists of four frames, the network only generates two sets of offsets, between the middle frame and its left and right neighboring frames, respectively. The final interpolated frame is the adaptively weighted average of these two closest frames warped by deformable convolution.

5.3.6 Implementation detail

Architecture configuration

The configurations of PDWN with 6 scales and predefined matching cost calculations, are evaluated in this study. Detailed configurations are shown in Table 5.1

Loss function

L1 norm has been proven to generate less blurry results in image synthesis tasks [34,68]. Thus, L1 Reconstruction loss between the reconstructed frame and the ground truth frame is used to train the model:

$$\mathscr{L} = ||\tilde{I}_1 - I_1||_1 \tag{5.11}$$

We also explore a multi-scale L1 reconstruction loss for training. Specifically, we downsample the input frames and the ground truth middle frame. Then, we apply the estimated offsets and modulation weights to the downsampled input images to generate the interpo-

Table 5.1. Architecture of PDWN

Module	Scale	Output size	Configuration
Feature extractor	1	$H \times W$	Conv 7 - 3 - 16 Conv 5 - 16 - 16
	2	$H/2 \times W/2$	Conv 3 - 16 - 32 Conv 3 - 32 - 32
	3	$H/4 \times W/4$	Conv 3 - 32 - 64 Conv 3 - 64 - 64
	4	$H/8 \times W/8$	Conv 3 - 64 - 96 Conv 3 - 96 -96
	5	$H/16 \times W/16$	Conv 3 - 96 - 128 Conv 3 - 128 - 128
	6	$H/32 \times W/32$	Conv 3 - 128 - 196 Conv 3 - 196 - 196
Offset estimator	6	$H/32 \times W/32$	Conv 3 - 473 - 256 Conv 3 - 256 - 256 Conv 3 - 256 - 54
	5	$H/16 \times W/16$	DConv 3 - 128 - 128 Conv 3 - 647 - 196 Conv 3 - 196 - 196 Conv 3 - 196 - 54
	4	$H/8 \times W/8$	DConv 3 - 96 - 96 Conv 3 - 523 - 128 Conv 3 - 128 - 128 Conv 3 - 128 - 54
	3	$H/4 \times W/4$	DConv 3 - 64 - 64 Conv 3 - 391 - 64 Conv 3 - 64 - 64 Conv 3 - 64 - 54
	2	$H/2 \times W/2$	DConv 3 - 32 - 32 Conv 3 - 295 - 64 Conv 3- 64 - 64 Conv 3 - 64 - 54
	1	$H \times W$	DConv 3 - 16 - 16 Conv 3 - 231 - 64 Conv 3 - 64 - 64 Conv 3 - 64 - 54
Adaptive frame blending	1	$H \times W$	DConv 3 - 3 - 3 DConv 3 - 16 - 16 Conv 3 - 38 - 16 Conv 3 - 16 - 16 Conv 3 - 16 - 2
Context enhancement	1	$H \times W$	Conv 3 - 41 - 64 Conv 3 - 64 - 64 $\times 2 \times 4$ Conv 3 - 64 - 3

[*] The convolutional and deformable convolutional layer parameters are denoted as “Conv/DConv <filter size> - <number of input channels> - <number of output channels>”. The leakyReLU activation function, max pool layer, bilinear upsample layer, and matching cost layer are not shown for brevity.

lated frame at each scale. Finally, the L1 reconstruction losses between the reconstructed frame and the ground truth frame for all scales are combined. Through our experiment, we find that the multi-scale loss does not improve the final results compared to simple L1 reconstruction loss at the finest scale. However, we do observe that the multi-scale loss could speed up the convergence during training. For simplicity, all results reported in this paper are obtained by using the simple L1 reconstruction loss at the finest scale.

Training dataset

We use Vimeo-90k training set [120], which has 51312 triplets, to train our model. Each triplet has 3 consecutive frames and each frame has a resolution of 448×256. Horizontal flipping and temporal reversing are adopted as data augmentation.

Training strategy

We train PDWN sequentially. In other words, we first train PDWN without context enhance network for 80 epochs, then finetune the whole system end-to-end for another 20 epochs. We use Adam [55] with $\beta_1 = 0.9$ and $\beta 2 = 0.999$ to optimize our model. The initial learning rate is set to 0.0002. Mini-batch size is set to 20. Following the techniques introduced in [78], we also train a variant of PDWN, called PDWN++, with input normalization, network improvements, and self-ensembling. Specifically, each color channel of the input frames is normalized independently to have zero mean and unit variance. Then, we replace the two-layer convolution with residual blocks. Moreover, the global filter of the deformable convolution that warps frames at level 1 is shared not only between input frames but also across RGB color channels. Finally, 7 transforms, including reverse, flipping, mirroring, reverse and flipping, and rotation by 90, 180, and 270 respectively, are applied during the inference phase for self-ensembling.

5.4 Results

In this section, we first introduce evaluation datasets. Then, we conduct ablation studies to evaluate the contribution of each component and to compare our proposed model with state-of-the-art on two input frames. Finally, we compare the performance of our models using two vs. four input frames and also compare with other models using four input frames.

5.4.1 Evaluation datasets and metrics

Evaluation Datasets Our model is trained on a single dataset (Vimeo-90K training set) but validated on multiple datasets including Vimeo-90K [120] test dataset (448 × 256), UCF [63, 101] dataset (25 FPS, 256 × 256), and the Middlebury dataset [5] (typically 640 × 480). The Middlebury dataset has two subsets. The OTHER set provides the ground-truth middle frames while the EVALUATION set hides the ground-truth and can only be evaluated by uploading the results to the benchmark website.

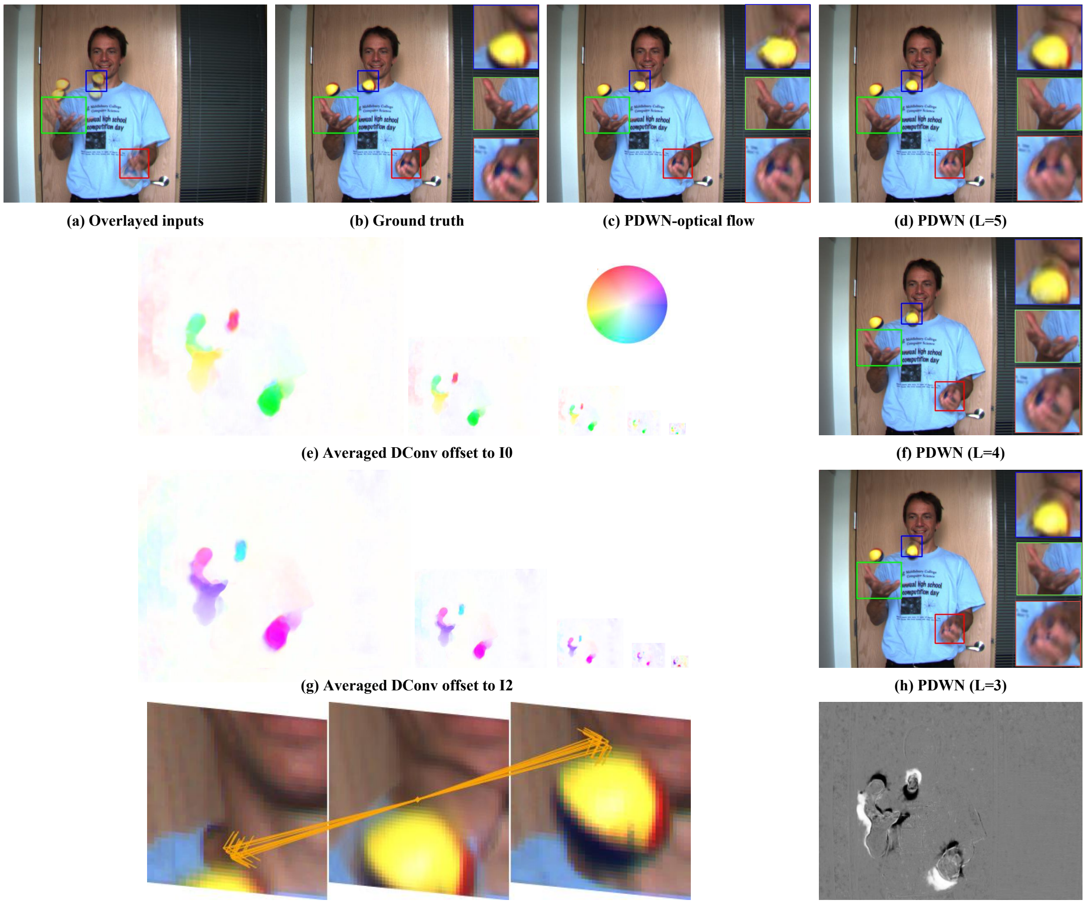
Evaluation Metrics We report peak signal noise ratio (PSNR), structure similarity index (SSIM) [114], and interpolation error (IE) for model comparison on multiple datasets with various resolutions and contents. IE is the average absolute color error. Higher PSNR or SSIM and lower IE indicate better performance.

5.4.2 Ablation studies

Optical flow V.S. DConv

To analyze how well the proposed framework performs with different image warping techniques, we train two variants of our approach, one using optical flow and the other using DConv at each scale. To integrate optical flow into our model, PDWN-optical flow generates and refines two sets of optical flow in every pyramid level instead of deformable offsets and

Figure 5.4. Analysis on warping operations and network scales & visualization of DConv offsets and adaptive blending weights. (c)-(d) Optical flow V.S. DConv. (d), (f), and (h) compares models with different number of scales. The model with larger scales is able to generate more accurate and sharper contents. (i) visualizes the sampling points of DConv in the past and future frames respectively. (e) and (g) show weighted averaged offsets to the past and future frame respectively at each scale, calculated based on 5.12. (j) is the adaptive weight map α for the warped past frame, i.e., the weight for the future frame is $1 - \alpha$. Thus, the black regions around the hand and ball show PDWN's capacity to handle occlusion.



(a) Overlayed inputs

(b) Ground truth

(c) PDWN-optical flow

(d) PDWN (L=5)

(e) Averaged DConv offset to I0

(f) PDWN (L=4)

(g) Averaged DConv offset to I2

(h) PDWN (L=3)

(i) DConv offset visualization

(j) Adaptive weight map

Table 5.2. Ablation studies on different components of PDWN

Model	Vimeo-90k		Middlebury OTHER		
	PSNR	SSIM	PSNR	SSIM	IE
PDWN-optical flow	34.59	0.961	35.35	0.957	2.47
PDWN w/o modulation	35.23	0.965	**37.00**	**0.966**	2.02
PDWN w/ modulation	**35.38**	**0.966**	**37.00**	**0.966**	**2.00**
PDWN w/o CV	35.13	0.964	37.09	0.966	1.99
PDWN w/ CV	35.38	0.966	37.00	0.966	2.00
PDWN w/ learnt CV	**35.42**	**0.966**	**37.17**	**0.967**	**1.98**
PDWN w/o coarse-to-fine	34.54	0.959	35.95	0.961	2.19
PDWN w/ coarse-to-fine	**35.42**	**0.966**	**37.17**	**0.967**	**1.98**
PDWN w/o c. e.	35.42	**0.966**	37.17	**0.967**	**1.98**
PDWN w/ c. e.	**35.44**	**0.966**	**37.20**	**0.967**	**1.98**

* CV denotes cost volume and c.e. denotes context enhancement. All models presented here use 6 scales. Models in section 1, 2, and 3 are trained without context enhancement.

modulation weights. Features and frames are backward warped by optical flow in PDWN-optical flow to replace deformable convolution in PDWN. As shown in Table 5.2 (section 1), DConv outperforms optical flow in terms of all performance metrics, which demonstrates the effectiveness of DConv. In Figure 5.4. (i), we visualize the DConv sampling points in the past and future frame respectively of an occluded point. We observe that the proposed model is able to point to locations in the left frame where the color is similar to the occluded region. As discussed above, DConv offsets can be considered as many-to-one backward warping flow. The redundancy of many-to-one flow makes the model more robust. In Figure 5.4.(e) and 5.4.(g), we visualize the weighted averaged DConv offsets by:

$$\bar{f}_{1\to i}(x) = \frac{\sum_{j=1}^{|R|}(R(j) + f_{1\to i}(j,x))m_{1\to i}(j,x)}{\sum_{j=1}^{|R|} m_{1\to i}(j,x)} \tag{5.12}$$

.

Cost volume

To analyze the effectiveness of using cost volumes, we consider three variants of our approach. The first model takes warped features only as input to the first conv layer in the offset estimator in Figure 5.2.(b). The second model first computes the cost volume between two warped features, then concatenates the cost volume and the original features to estimate DConv offset residuals. The third model replaces the cost volume layer with a two-layer CNN to learn the matching cost between two warped features. As shown in Table 5.2, cost volumes bring additional improvements without adding more parameters on the Vimeo-90K dataset. Replacing the predefined cost with the learned cost further improves the results for both datasets.

Coarse-to-fine successive refinement manner

In the proposed model, we warp features and construct the matching cost between warped features to estimate DConv offset *residuals* $\Delta f^{l}_{1\to i}$ at every pyramid level in a coarse-to-fine manner. It reduces the distance between two input frames gradually and is particularly important when the ground truth motion is large. We investigate the contribution of this coarse-to-fine structure via training another variant of our model, without the coarse-to-fine structure. In other words, this model is simply a UNet structure with 6 spatial scales that takes two images I_0 and I_2 as input and directly outputs DConv offsets and modulation weights in the finest scale. We show the quantitative results in Table 5.2 and qualitative results in Figure 5.5. By introducing the coarse-to-fine structure, the performance is significantly improved, demonstrating the effectiveness of our successive coarse-to-fine successive refinement approach.

Figure 5.5. Effect of the coarse-to-fine structure. By introducing the coarse-to-fine structure, PDWN generates more realistic interpolation results.

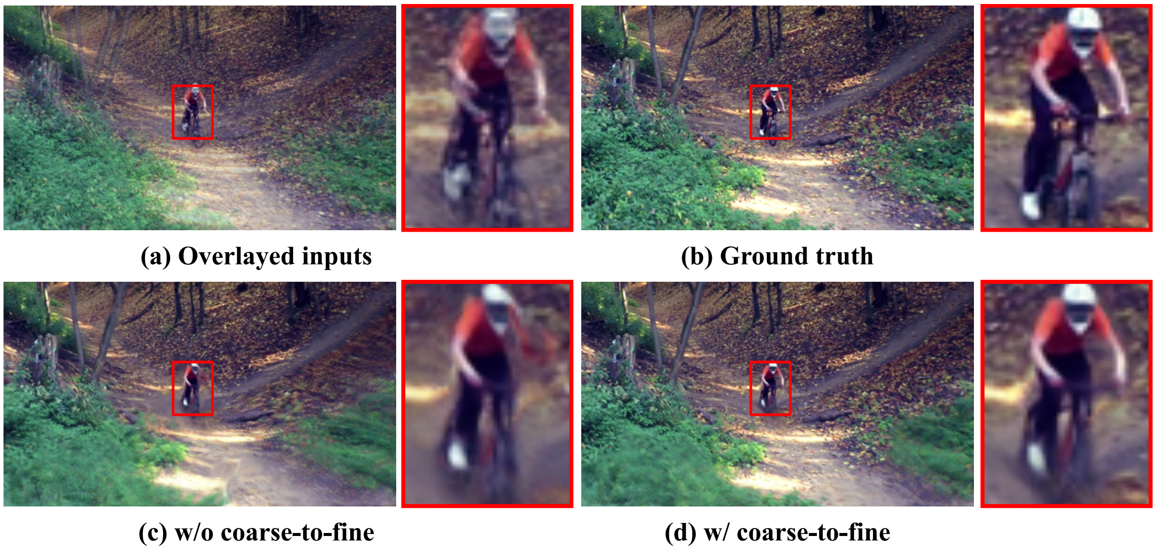

(a) Overlayed inputs
(b) Ground truth
(c) w/o coarse-to-fine
(d) w/ coarse-to-fine

Table 5.3. Effect of the number of scales.

Scale	Runtime	Param.	Vimeo-90k		Middlebury OTHER		
	(second)	(million)	PSNR	SSIM	PSNR	SSIM	IE
L=4	**0.0056**	**1.7**	35.02	0.963	36.63	0.964	2.07
L=5	0.0068	3.4	35.19	**0.965**	36.85	0.965	2.04
L=6	0.0086	6.6	**35.23**	**0.965**	**37.00**	**0.966**	**2.02**

* L denotes the number of scales. Note that in this experiment we use a simpler version of DConv where the modulation weights are all set to 1 and the cost volume is predefined. The models trained here are all without context enhancement. The feature size is downsampled 8, 16, 32 times for L = 4, 5, 6, respectively. The runtime is evaluated for interpolating one middle frame of ”DogDance” from Middlebury OTHER dataset, with a size of 640×480, on GeForce RTX 2080 Ti.

Impact of the number of scales

To analyze the impact of the number of scales on the performance, we investigate three different pyramid scales (L = 4, 5, 6). Quantitative results are shown in Table 5.3, and the visual comparison is provided in Figure 5.4. We find that with model size increasing from 1.7, 3.4, to 6.6 million, the PSNR steadily gets better from 36.63, 36.85, to 37.00 dB on the Middlebury OTHER dataset. The example in Figure 5.4 also shows that the model using more scales generates sharper outcomes. The gain on Vimeo-90K, however, is not as significant as that on the Middlebury OTHER dataset. That is probably because the Middlebury OTHER dataset has a larger image size (and hence larger motion in terms of pixels) than the Vimeo-90K dataset. Even though the model size almost doubles with each additional scale, the runtime only increases slightly, as the lower-scale images and features have a smaller spatial dimension.

Adaptive Blending Weight

Figure 5.4.(j) shows an example of adaptive blending weight map. As discussed in section 3.3, $\alpha(x) = 0$ means pixel x from I_0 is occluded and pixels x from I_1 is fully trusted. The black region around the ball in the weight map indicates that our model can detect and solve occlusion by selecting pixels from the previous and following frames softly.

Context enhancement network

To analyze the contribution of the context enhancement module, we train a variant of PDWN without context enhancement and show the results in Table 5.2. Though DAIN gains significantly from adding the context enhancement module (0.27 dB on Vimeo-90k in terms of PSNR) [6], the context enhancement network has little contribution to PDWN. By adding the context enhancement network, the number of model parameters increases from 7.4 million to 7.8 million and the runtime increases from 0.0082 to 0.0086 for interpolating ”DogDance” image (640×480) in the Middlebury-OTHER dataset, using an NVIDIA RTX 8000 GPU card.

5.4.3 Comparison with state-of-the-arts

We compare our model with state-of-the-art video interpolation models both quantitatively and qualitatively, including deep voxel flow (DVF) [63], SepConv [77], SepConv++ [78], SuperSloMo [51], MEMC-Net* [7], DAIN [6], AdaCof [59], FeFlow [36], on three different datasets, Vimeo-90K, UCF, and Middlebury dataset. Note that we only compare with methods that use backward optical flow or DConv for backward image warping. For SepConv, AdaCof, and FeFlow, we download their published models and test on the testing datasets. For DVF, SuperSloMo, MEMC-Net*, and DAIN, we calculate the numbers from their published interpolated data. For RRPN and SepConv++, we directly report their published numbers.

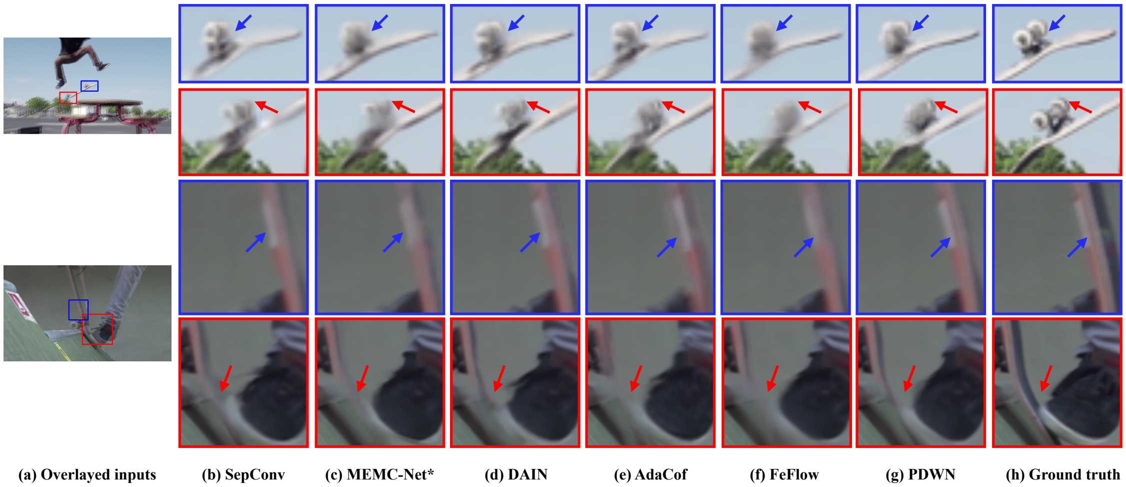


Figure 5.6. Visualized examples on Vimeo-90k test dataset.

As shown in Table 5.4, our proposed model outperforms all methods on the Vimeo-90k dataset and Middlebury OTHER dataset except SepConv++. Using similar techniques applied to SepConv++, PDWN++ surpasses SepConv++ for 0.88 dB on the Middlebury OTHER dataset with respect to PSNR. Meanwhile, the number of model parameters increases from 7.8 million to 8.6 million and the runtime increases nearly 8 times. On the UCF dataset, our model achieves on par performance with state-of-the-art methods. Note that DAIN uses additional depth information to detect occlusion in order to compensate for errors in the linear interpolated optical flow. DAIN relies on the accuracy of depth information, i.e., their model cannot learn meaningful depth information without a good initialization of (pretrained) depth estimation network and thus yields lower quality results than MEMC-Net. Our model does not need depth information for training information but still achieves 0.73 dB higher PSNR than DAIN on Vimeo-90K. FeFlow uses multiple groups of DConv offsets in every layer to avoid occlusion and edge maps generated by BDCN [42] as structure guidance. Compared to FeFlow, our model performs better on Vimeo-90K without edge maps and with only a single group of DConv offsets, which demonstrates the supremacy of using DConv in a coarse-to-fine manner. Moreover, our model size is only 5.8% of that of FeFlow. Figure 5.6 presents two examples from the Vimeo-90k dataset. Notably, our model generates the sharpest results among all compared methods.

Table 5.5 shows the comparison on the Middlebury EVALUATION dataset. Our pro-

Table 5.4. Comparison with state-of-the-arts

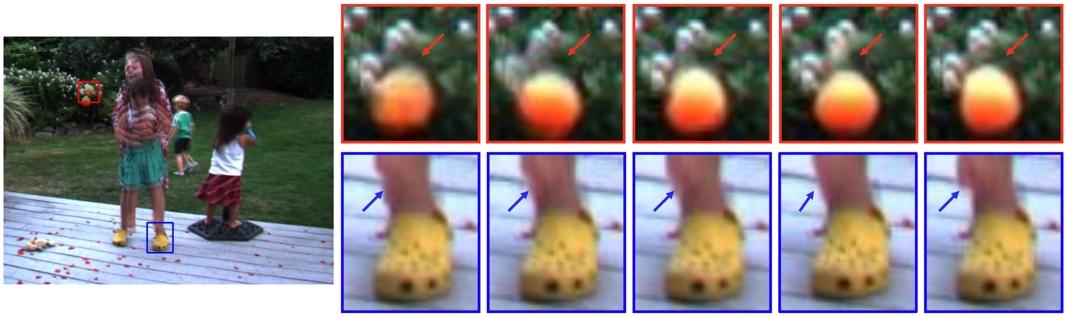
Method	Runtime	Param.	Vimeo-90k		Middlebury OTHER			UCF	
	(second)	(million)	PSNR	SSIM	PSNR	SSIM	IE	PSNR	SSIM
DVF [63]	-	**3.8**	-	-	-	-	-	34.12	0.942
SepConv-L1 [77]	**0.0032**	21.6	33.80	0.956	35.89	0.959	2.24	34.69	0.945
SepConv++ [78]	-	13.6	34.98	-	**37.47**	-	-	**35.29**	-
SuperSlowMo [51]	-	39.6	-	-	-	-	-	34.75	0.947
MEMC-Net* [7]	0.122	70.3	34.40	0.962	36.48	0.964	2.12	35.01	**0.949**
DAIN [6]	0.125	24.0	34.71	0.964	36.70	0.965	2.04	34.99	**0.949**
RRPN [125]	-	-	-	-	-	-	-	34.76	-
AdaCof [59]	**0.0043**	21.8	34.35	0.956	35.69	0.958	2.26	34.90	**0.949**
FeFlow [36]	0.7188	133.6	35.16	0.963	36.61	0.965	2.14	34.89	**0.949**
PDWN (L=6)	0.0089	**7.8**	**35.44**	**0.966**	37.20	**0.967**	**1.98**	35.00	**0.950**
PDWN++ (L=6)	0.0669	8.6	**35.69**	**0.968**	**38.35**	**0.971**	**1.81**	**35.10**	**0.950**

[*] PDWN achieves on-par performance with much fewer parameters compared to previous methods.

[*] The runtime of DAIN and MEMC-Net* is reported in their paper on a 640x480 image using an NVIDIA Titan X (Pascal) GPU card. Other runtime numbers reported are estimated for ”DogDance” image on an Nvidia RTX 2080 Ti GPU card.

[*] The number in red and blue represents the best and second best performance.

Figure 5.7. Visualized examples on Middlebury EVALUATION dataset. PDWN generates high-quality details on the foot and girl's toe while other methods produce blurry output. Moreover, PDWN shows its capacity to deal with occlusion and semantic shape distortion on the ball and white flowers.



(a) Overlayed inputs (b) MEMC-Net* (c) DAIN (d) AdaCof (e) FeFlow (f) PDWN

Table 5.5. Results on Middlebury EVALUATION datase

Method	Average		Mequon		Schefflera		Urban		Teddy		Backyard		Basketball		Dumptruck		Evergreen	
	IE	NIE	IE	NIE	IE	NIE	IE	NIE	IE	NIE	IE	NIE	IE	NIE	IE	NIE	IE	NIE
SuperSlowMo [51]	5.31	0.78	2.51	0.59	3.66	0.72	2.91	0.74	5.05	0.98	9.56	0.94	5.37	0.96	6.69	0.60	6.73	0.69
MEMC-Net* [7]	4.99	0.74	2.39	0.59	3.36	0.64	3.37	0.80	4.84	0.88	8.55	0.88	4.70	**0.85**	6.40	0.64	6.37	0.63
DAIN [6]	4.85	**0.71**	2.38	0.58	3.28	0.60	3.32	**0.69**	4.65	**0.86**	7.88	0.87	4.73	**0.85**	6.36	**0.59**	6.25	0.66
AdaCof [59]	4.75	0.73	2.41	0.60	**3.10**	0.59	3.48	0.84	4.84	0.92	8.68	0.90	**4.13**	**0.84**	**5.77**	**0.58**	**5.60**	**0.57**
FeFlow [36]	4.82	**0.71**	**2.28**	**0.51**	3.50	0.66	**2.82**	0.70	4.75	0.87	**7.62**	**0.84**	4.74	0.86	6.07	0.64	6.78	0.67
RRPN [125]	4.93	0.75	2.38	0.53	3.70	0.69	3.29	0.87	5.05	0.94	8.20	0.88	4.38	0.88	6.50	0.65	6.00	0.62
SepConv++ [78]	**3.88**	0.73	2.39	0.58	**2.98**	**0.56**	3.34	0.95	**4.49**	0.87	**7.64**	**0.85**	**3.77**	**0.84**	**5.26**	**0.59**	**5.71**	**0.59**
PDWN (L=6)	**4.71**	**0.69**	**2.09**	**0.46**	3.12	**0.58**	**2.38**	**0.64**	**4.29**	**0.85**	8.61	0.87	4.80	0.88	6.24	0.60	6.18	0.62

[*] NIE: normalized interpolation error.

posed method performs favorably against state-of-the-art methods. Our model performs well quantitatively on sequences with small motion or fine textures such as *Mequon, Teddy* and *Schefflera*. For videos with complicated motions, Figure 5.7 shows a visualized example. Our model produces more details at the girl's toe in the *backyard* example while other methods output blurry results. And our model handles occlusion well around the boundary of the orange ball.

5.4.4 Extending to four input frames

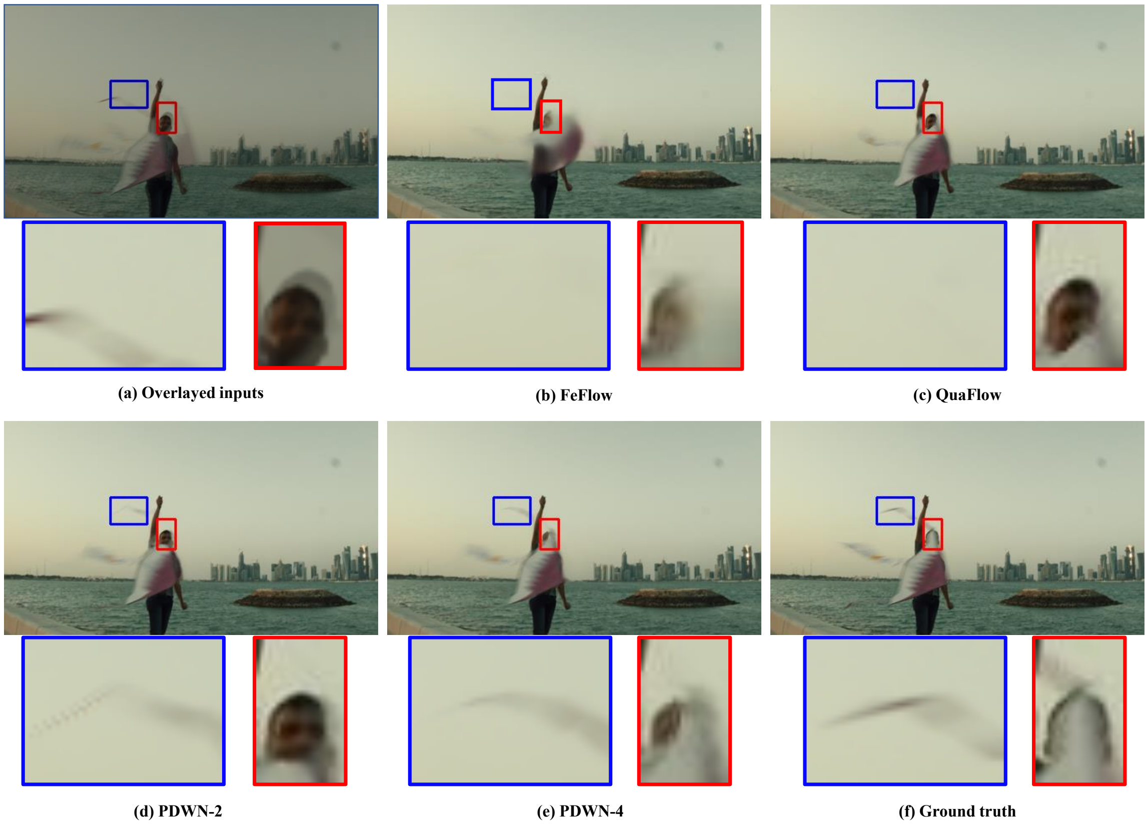
Vimeo-90K septuplet dataset is used to train and test our extended model PDWN-4 which takes four input frames as input and has 6 pyramid levels. We use frame 1, 3, 5, and 7 to interpolate frame 4 and compare the interpolated frame 4 with the original frame 4 for every sequence in the Vimeo-90K septuplet dataset. We compare the results with our two-input model PDWN-2 and state-of-the-art methods including FeFlow [36] and QuaFlow [119]. PDWN-2 is pretrained on the Vimeo-90K triplet dataset and finetuned on the Vimeo-90K septuplet dataset. Results are given in Table 5.6. Figure 5.8 shows visualized results on the Vimeo-90K septuplet test dataset. Both the quantitative and visual evaluations demonstrate that the extended PDWN with four input frames can significantly improve the interpolation accuracy over using two input frames, with only modest increases in the model size and the runtime. Furthermore, both PDWN-2 and PDWN-4 yield better results than QuaFlow which uses four input frames.

5.5 Conclusion

In this work, we propose a pyramid video interpolation model that estimates the many-to-one flows with modulation maps of the middle frame to the left and right input frames. We show that the offset estimator can benefit from using the cost volumes computed from the aligned features, compared to using the aligned features directly. Our model is significantly smaller in model size and requires substantially less inference time compared to state-of-the-art models and yet achieves better or on-par interpolation accuracy. Besides, our model does not rely on additional information (e.g. ground truth depth information or optical flow) for training. Moreover, our model that uses two input frames can be extended to use four input frames easily, with only a small increase in the model size and the inference time, and yet the extended model significantly improves the interpolation accuracy.

A recent work [75], which proposes a differentiable forward warping operation using

Figure 5.8. Visualized examples of the extended PDWN with 4 input frames. PDWN-2 takes only 1 past frame and 1 future frame as input. QuaFlow and PDWN-4 take 2 past frames and 2 future frames as input.



forward optical flow to handle occlusion and dis-occlusion regions directly, outperforms all backward-flow-based methods. It shows a promising direction for video interpolation. In future work, we will also explore how to combine forward warping with a coarse-to-fine structure. Furthermore, we will explore the integration of PDWN in video coding, where the encoder can encode every other frame; Skipped frames will be interpolated by the PDWN method and the interpolation error images can be additionally coded.

Table 5.6. Results of the extended PDWN with four input frames on the Vimeo-septuplet dataset

Method	Runtime	Param.	Vimeo-septuplet	
	(second)	(million)	PSNR	SSIM
FeFlow [36]	0.221	133.6	33.88	0.946
PDWN-2	**0.010**	**7.4**	35.53	0.958
QuaFlow [119]	0.090	19.6	34.28	0.950
PDWN-4	0.012	8.3	**35.93**	**0.960**

[*] FeFlow and PDWN-2 take only 1 past frame and 1 future frame as input. QuaFlow and PDWN-4 take 2 past frames and 2 future frames as input.

[*] Both PDWN-2 and PDWN-4 have 6 pyramid levels and no contextual enhancement module.

[*] The runtime reported is the average runtime for the Vimeo-septuplet dataset with image size 448×256 on an Nvidia Tesla V100 GPU card.

Chapter 6

Conclusion and Future Work

In this dissertation, we've covered diverse applications of three dimensional (3D) volumetric image analysis and spatiotemporal sequence understanding. Our goals encompass enhancing glaucoma diagnosis and management, alongside advancements in video processing. In this chapter, we'll provide a concise overview of our primary contributions and offer insights into potential future research directions.

To overcome limitations in segmentation algorithms applied to 3D optical coherence tomography (OCT) and harness the volumetric nature of OCT images, Chapter 2 introduces a novel 3D Convolutional Neural Network (CNN). This novel approach directly estimates point-wise visual field (VF) sensitivities from segmentation-free 3D OCT images, aiming to overcome the constraints of prior segmentation-dependent measurements. It provides a VF surrogate potentially without standard automated perimetry (SAP)'s inherent limitations.

Building upon the 3D CNN model developed in Chapter 2, Chapter 3 employs occlusion analysis to establish a generalized spatial structure-to-function mapping, visualizing significant optic nerve head (ONH) regions in predicting point-wise VF sensitivities. The derived maps are consistent with existing knowledge and understanding of structure-function spatial relationships. This presents possibilities of learning from trained machine learning models without applying any prior knowledge, potentially robust and free from bias. While offering insights aligned with existing knowledge of structure-function relationships, this work has limitations due to the use of naive registration. Advanced registration techniques are necessary to uncover subtler spatial relationships and fully exploit the 3D nature of the ONH.

In Chapter 4, we delve into spatiotemporal sequence modeling for glaucoma progression. We propose a time-aware Convolutional Long Short-Term Memory (TC-LSTM) model to predict future two dimensional (2D) Ganglion Cell–Inner Plexiform layer (GCIPL) thickness maps – a vital biomarker for monitoring glaucoma progression. This novel model leverages spatial and temporal correlations in irregularly sampled longitudinal sequences. Experimental results demonstrate the superiority of the proposed TC-LSTM over traditional methods.

In Chapter 5, we explore another example of spatiotemporal sequences, natural videos, and develop an efficient video interpolation algorithm, Pyramid Deformable Warping Network (PDWN). By integrating a pyramid structure and deformable convolution in its design, PDWN effectively merges the advantages of optical flow and kernel methods, surpassing state-of-the-art models in accuracy across various datasets for video interpolation, while reducing the model parameters and inference time. While excelling in modeling large motions, PDWN might encounter challenges in scenarios involving tiny objects with extensive movements. Further investigating Transformer approaches might provide benefits. Transformer models' attention mechanism enables extended correlation modeling across distant regions of video frames, preventing the omission of small objects. Additionally, exploring video prediction through similar methods holds promise for broader applications in video coding.

Bibliography

[1] Tarek Alasil, Kaidi Wang, Fei Yu, Matthew G Field, Hang Lee, Neda Baniasadi, Johannes F de Boer, Anne L Coleman, and Teresa C Chen. Correlation of retinal nerve fiber layer thickness and visual fields in glaucoma: a broken stick model. *American journal of ophthalmology*, 157(5):953–959, 2014.

[2] Rayan A Alshareef, Abhilash Goud, Mikel Mikhail, Hady Saheb, Hari Kumar Peguda, Sunila Dumpala, Shruthi Rapole, and Jay Chhablani. Segmentation errors in macular ganglion cell analysis as determined by optical coherence tomography in eyes with macular pathology. *International journal of retina and vitreous*, 3(1):1–8, 2017.

[3] Shotaro Asano, Ryo Asaoka, Hiroshi Murata, Yohei Hashimoto, Atsuya Miki, Kazuhiko Mori, Yoko Ikeda, Takashi Kanamoto, Junkichi Yamagami, and Kenji Inoue. Predicting the central 10 degrees visual field in glaucoma by applying a deep learning algorithm to optical coherence tomography images. *Scientific Reports*, 11(1):2214, 2021.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011.

[6] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019.

[7] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[8] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 65–74, 2017.

[9] Brigid D Betz-Stablein, William H Morgan, Philip H House, and Martin L Hazelton. Spatial modeling of visual field data for assessing glaucoma progression. *Investigative ophthalmology & visual science*, 54(2):1544–1553, 2013.

[10] Christopher Bowd, Robert N. Weinreb, Julia M. Williams, and Linda M. Zangwill. The Retinal Nerve Fiber Layer Thickness in Ocular Hypertensive, Normal, and Glaucomatous Eyes With Optical Coherence Tomography. *Archives of Ophthalmology*, 118(1):22–26, 01 2000.

[11] RS Brenton and William A Argus. Fluctuations on the humphrey and octopus perimeters. *Investigative ophthalmology & visual science*, 28(5):767–771, 1987.

[12] Donald L Budenz, Marie-Josée Fredette, William J Feuer, and Douglas R Anderson. Reproducibility of peripapillary retinal nerve fiber thickness measurements with stratus oct in glaucomatous eyes. *Ophthalmology*, 115(4):661–666, 2008.

[13] Joseph Caprioli, Dennis Mock, Elena Bitrian, Abdelmonem A Afifi, Fei Yu, Kouros Nouri-Mahdavi, and Anne L Coleman. A method to measure and predict rates of re-

gional visual field decay in glaucoma. *Investigative ophthalmology & visual science*, 52(7):4765–4773, 2011.

[14] Roberto Castagno, Petri Haavisto, and Giovanni Ramponi. A method for motion adaptive frame rate up-conversion. *IEEE Transactions on circuits and Systems for Video Technology*, 6(5):436–446, 1996.

[15] Xiangyu Chen, Yanwu Xu, Damon Wing Kee Wong, Tien Yin Wong, and Jiang Liu. Glaucoma detection based on deep convolutional neural network. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 715–718. IEEE, 2015.

[16] Zhiqi Chen and Hiroshi Ishikawa. Technical aspects of deep learning in ophthalmology. *Artificial Intelligence in Ophthalmology*, pages 69–75, 2021.

[17] Zhiqi Chen, Eitan Shemuelian, Gadi Wollstein, Yao Wang, Hiroshi Ishikawa, and Joel S. Schuman. Segmentation-Free OCT-Volume-Based Deep Learning Model Improves Pointwise Visual Field Sensitivity Estimation. *Translational Vision Science & Technology*, 12(6):28–28, 06 2023.

[18] Zhiqi Chen, Eitan Shemuelian, Gadi Wollstein, Yao Wang, Hiroshi Ishikawa, and Joel S Schuman. Segmentation-free oct-volume-based deep learning model improves pointwise visual field sensitivity estimation. *Translational Vision Science & Technology*, 12(6):28–28, 2023.

[19] Zhiqi Chen, Yao Wang, Gadi Wollstein, María de los Angeles Ramos-Cadena, Joel Schuman, and Hiroshi Ishikawa. Macular gcipl thickness map prediction via time-aware convolutional lstm. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2020.

[20] Zhiqi Chen, Gadi Wollstein, Joel S Schuman, and Hiroshi Ishikawa. Ai and glaucoma. *Artificial Intelligence in Ophthalmology*, pages 113–125, 2021.

[21] Mark Christopher, Christopher Bowd, Akram Belghith, Michael H Goldbaum, Robert N Weinreb, Massimo A Fazio, Christopher A Girkin, Jeffrey M Liebmann, and Linda M Zangwill. Deep learning approaches predict glaucomatous visual field damage from oct optic nerve head en face images and retinal nerve fiber layer thickness maps. *Ophthalmology*, 127(3):346–356, 2020.

[22] Mark Christopher, Christopher Bowd, James A Proudfoot, Akram Belghith, Michael H Goldbaum, Jasmin Rezapour, Massimo A Fazio, Christopher A Girkin, Gustavo De Moraes, Jeffrey M Liebmann, et al. Deep learning estimation of 10-2 and 24-2 visual field metrics based on thickness maps from macula oct. *Ophthalmology*, 128(11):1534–1548, 2021.

[23] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[24] Sripad Krishna Devalla, Zhang Liang, Tan Hung Pham, Craig Boote, Nicholas G Strouthidis, Alexandre H Thiery, and Michael JA Girard. Glaucoma management in the era of artificial intelligence. *British Journal of Ophthalmology*, 104(3):301–311, 2020.

[25] Antonio Ferreras, Luis E Pablo, David F Garway-Heath, Paolo Fogagnolo, and Julian Garcia-Feijoo. Mapping standard automated perimetry to the peripapillary retinal nerve fiber layer in glaucoma. *Investigative ophthalmology & visual science*, 49(7):3018–3025, 2008.

[26] Paolo Fogagnolo, Chiara Sangermani, Francesco Oddone, Paolo Frezzotti, Michele Iester, Michele Figus, Antonio Ferreras, Simona Romano, Stefano Gandolfi, Marco Centofanti, et al. Long-term perimetric fluctuation in patients with different stages of glaucoma. *British Journal of Ophthalmology*, 95(2):189–193, 2011.

[27] Yuri Fujino, Hiroshi Murata, Masato Matsuura, Mieko Yanagisawa, Nobuyuki Shoji, Kenji Inoue, Junkichi Yamagami, and Ryo Asaoka. Mapping the central 10° visual field to the optic nerve head using the structure–function relationship. *Investigative ophthalmology & visual science*, 59(7):2801–2807, 2018.

[28] Elena Garcia-Martin, Isabel Pinilla, Miriam Idoipe, Isabel Fuertes, and Victoria Pueyo. Intra and interoperator reproducibility of retinal nerve fibre and macular thickness measurements using cirrus fourier-domain oct. *Acta ophthalmologica*, 89(1):e23–e29, 2011.

[29] Stuart K. Gardiner, Shaban Demirel, Deborah Goren, Steven L. Mansberger, and William H. Swanson. The Effect of Stimulus Size on the Reliable Stimulus Range of Perimetry. *Translational Vision Science & Technology*, 4(2):10–10, 04 2015.

[30] Stuart K Gardiner, Chris A Johnson, and George A Cioffi. Evaluation of the structure-function relationship in glaucoma. *Investigative ophthalmology & visual science*, 46(10):3712–3717, 2005.

[31] Stuart K Gardiner, William H Swanson, Deborah Goren, Steven L Mansberger, and Shaban Demirel. Assessment of the reliability of standard automated perimetry in regions of glaucomatous damage. *Ophthalmology*, 121(7):1359–1369, 2014.

[32] David F Garway-Heath, Darmalingum Poinoosawmy, Frederick W Fitzke, and Roger A Hitchings. Mapping the visual field to the optic disc in normal tension glaucoma eyes. *Ophthalmology*, 107(10):1809–1815, 2000.

[33] Marta Gonzalez-Hernandez, Luis E Pablo, K Armas-Dominguez, R Rodriguez de La Vega, Antonio Ferreras, and M Gonzalez De La Rosa. Structure–function relationship depends on glaucoma severity. *British Journal of Ophthalmology*, 93(9):1195–1199, 2009.

[34] Ross Goroshin, Michaël Mathieu, and Yann LeCun. Learning to linearize under uncertainty. *CoRR*, abs/1506.03011, 2015.

[35] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

[36] Shurui Gui, Chaoyue Wang, Qihua Chen, and Dacheng Tao. Featureflow: Robust video interpolation via structure-to-texture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14004–14013, 2020.

[37] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[38] Zhihui Guo, Young H. Kwon, Kyungmoo Lee, Kai Wang, Andreas Wahle, Wallace L. M. Alward, John H. Fingert, Daniel I. Bettis, Chris A. Johnson, Mona K. Garvin, Milan Sonka, and Michael D. Abràmoff. Optical Coherence Tomography Analysis Based Prediction of Humphrey 24-2 Visual Field Thresholds in Patients With Glaucoma. *Investigative Ophthalmology & Visual Science*, 58(10):3975–3985, 08 2017.

[39] Zhihui Guo, Young H Kwon, Kyungmoo Lee, Kai Wang, Andreas Wahle, Wallace LM Alward, John H Fingert, Daniel I Bettis, Chris A Johnson, Mona K Garvin, et al. Optical coherence tomography analysis based prediction of humphrey 24-2 visual field thresholds in patients with glaucoma. *Investigative ophthalmology & visual science*, 58(10):3975–3985, 2017.

[40] Peter Gutierrez, M Roy Wilson, Chris Johnson, Mae Gordon, George A Cioffi, Robert Ritch, Mark Sherwood, Karen Meng, and Carol M Mangione. Influence of glaucomatous visual field loss on health-related quality of life. *Archives of ophthalmology*, 115(6):777–784, 1997.

[41] Yohei Hashimoto, Ryo Asaoka, Taichi Kiwaki, Hiroki Sugiura, Shotaro Asano, Hiroshi Murata, Yuri Fujino, Masato Matsuura, Atsuya Miki, Kazuhiko Mori, et al. Deep learning model to predict visual field in central 10 from optical coherence tomography measurement in glaucoma. *British Journal of Ophthalmology*, 105(4):507–513, 2021.

[42] Jianzhong He, Shiliang Zhang, Ming Yang, Yanhu Shan, and Tiejun Huang. Bi-directional cascade network for perceptual edge detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[44] Michael R Hee, Joseph A Izatt, Eric A Swanson, David Huang, Joel S Schuman, Charles P Lin, Carmen A Puliafito, and James G Fujimoto. Optical coherence tomography of the human retina. *Archives of ophthalmology*, 113(3):325–332, 1995.

[45] Samin Hong, Chan Yun Kim, Won Seok Lee, and Gong Je Seong. Reproducibility of peripapillary retinal nerve fiber layer thickness with spectral domain cirrus high-definition optical coherence tomography in normal eyes. *Japanese journal of ophthalmology*, 54(1):43–47, 2010.

[46] Donald C Hood and Randy H Kardon. A framework for comparing structural and functional measures of glaucomatous damage. *Progress in retinal and eye research*, 26(6):688–710, 2007.

[47] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):504–511, 2012.

[48] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[49] Nomdo M Jansonius, Jukka Nevalainen, Bettina Selig, LM Zangwill, PA Sample, WM Budde, JB Jonas, Wolf Alexander Lagrèze, PJ Airaksinen, Reinhard Vonthein, et al. A mathematical description of nerve fiber bundle trajectories and their variability in the human retina. *Vision research*, 49(17):2157–2163, 2009.

[50] Laia Jaumandreu, Francisco J Muñoz-Negrete, Noelia Oblanca, and Gema Rebolleda. Mapping the structure-function relationship in glaucoma and healthy patients measured with spectralis oct and humphrey perimetry. *Journal of Ophthalmology*, 2018, 2018.

[51] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018.

[52] Alireza Kamalipour, Sasan Moghimi, Pooya Khosravi, Mohammad Sadegh Jazayeri, Takashi Nishida, Golnoush Mahmoudinezhad, Elizabeth H Li, Mark Christopher, Jeffrey M Liebmann, Massimo A Fazio, et al. Deep learning estimation of 10-2 visual field map based on circumpapillary retinal nerve fiber layer thickness measurements. *American Journal of Ophthalmology*, 246:163–173, 2023.

[53] Yuka Kihara, Giovanni Montesano, Andrew Chen, Nishani Amerasinghe, Chrysostomos Dimitriou, Aby Jacob, Almira Chabi, David P Crabb, and Aaron Y Lee. Policy-driven, multimodal deep learning for predicting visual fields from the optic disc and oct imaging. *Ophthalmology*, 129(7):781–791, 2022.

[54] Soa Kim, Jin Young Lee, Seon-Ok Kim, and Michael S Kook. Macular structure–function relationship at various spatial locations in glaucoma. *British Journal of Ophthalmology*, 99(10):1412–1418, 2015.

[55] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[56] CT Langerhorst, TJTP Van den Berg, R Van Spronsen, and EL Greve. Results of a fluctuation analysis and defect volume program for automated static threshold perimetry with the scoperimeter. In *Sixth International Visual Field Symposium*, pages 1–6. Springer, 1985.

[57] Georgios Lazaridis, Giovanni Montesano, Saman Sadeghi Afgeh, Jibran Mohamed-Noriega, Sebastien Ourselin, Marco Lorenzi, and David F Garway-Heath. Predicting visual fields from optical coherence tomography via an ensemble of deep representation learners. *American journal of ophthalmology*, 238:52–65, 2022.

[58] Aaron Lee, Paul Taylor, Jayashree Kalpathy-Cramer, and Adnan Tufail. Machine learning has arrived! *Ophthalmology*, 124(12):1726–1728, 2017.

[59] Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. Adacof: Adaptive collaboration of flows for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5316–5325, 2020.

[60] Ji-Woong Lee, Esteban Morales, Farideh Sharifipour, Navid Amini, Fei Yu, Abdelmonem A Afifi, Anne L Coleman, Joseph Caprioli, and Kouros Nouri-Mahdavi.

The relationship between central visual field sensitivity and macular ganglion cell/inner plexiform layer thickness in glaucoma. *British Journal of Ophthalmology*, 101(8):1052–1058, 2017.

[61] Christopher Kai-shun Leung, Carol Yim Lui Cheung, Robert N Weinreb, Kunliang Qiu, Shu Liu, Haitao Li, Guihua Xu, Ning Fan, Chi Pui Pang, Kwok Kay Tse, et al. Evaluation of retinal nerve fiber layer progression in glaucoma: a study on optical coherence tomography guided progression analysis. *Investigative ophthalmology & visual science*, 51(1):217–222, 2010.

[62] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

[63] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4463–4471, 2017.

[64] Stefan Maetschke, Bhavna Antony, Hiroshi Ishikawa, Gadi Wollstein, Joel Schuman, and Rahil Garnavi. A feature agnostic approach for glaucoma detection in oct volumes. *PloS one*, 14(7):e0219126, 2019.

[65] Stefan Maetschke, Bhavna Antony, Hiroshi Ishikawa, Gadi Wollstein, Joel Schuman, and Rahil Garnavi. Inference of visual field test performance from oct volumes using deep learning. *arXiv preprint arXiv:1908.01428*, 2019.

[66] Eduardo B Mariottoni, Shounak Datta, David Dov, Alessandro A Jammal, Samuel I Berchuck, Ivan M Tavares, Lawrence Carin, and Felipe A Medeiros. Artificial intelligence mapping of structure to function in glaucoma. *Translational vision science & technology*, 9(2):19–19, 2020.

[67] Gianni Marra and Josef Flammer. The learning and fatigue effect in automated perimetry. *Graefe's archive for clinical and experimental ophthalmology*, 229(6):501–504, 1991.

[68] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error, 2015.

[69] Hassan Muhammad, Thomas J Fuchs, Nicole De Cuir, Carlos G De Moraes, Dana M Blumberg, Jeffrey M Liebmann, Robert Ritch, and Donald C Hood. Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects. *Journal of glaucoma*, 26(12):1086, 2017.

[70] Anna S Mursch-Edlmayr, Wai Siene Ng, Alberto Diniz-Filho, David C Sousa, Louis Arnould, Matthew B Schlenker, Karla Duenas-Angeles, Pearse A Keane, Jonathan G Crowston, and Hari Jayaram. Artificial intelligence algorithms to diagnose glaucoma and detect glaucoma progression: translation to clinical practice. *Translational vision science & technology*, 9(2):55–55, 2020.

[71] Jean-Claude Mwanza, Hanna Y Kim, Donald L Budenz, Joshua L Warren, Michael Margolis, Scott D Lawrence, Pooja D Jani, Garrett S Thompson, and Richard K Lee. Residual and dynamic range of retinal nerve fiber layer thickness in glaucoma: comparison of three oct platforms. *Investigative ophthalmology & visual science*, 56(11):6344–6351, 2015.

[72] Jung Hwa Na, Kyung Rim Sung, Seunghee Baek, Jin Young Lee, and Soa Kim. Progression of retinal nerve fiber layer thinning in glaucoma assessed by cirrus optical coherence tomography-guided progression analysis. *Current eye research*, 38(3):386–395, 2013.

[73] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased lstm: Accelerating recurrent network training for long or event-based sequences. *Advances in neural information processing systems*, 29, 2016.

[74] Patricia Nelson, Peter Aspinall, Orestis Papasouliotis, Bruce Worton, and Colm O'Brien. Quality of life in glaucoma and its relationship with visual function. *Journal of glaucoma*, 12(2):139–150, 2003.

[75] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *IEEE International Conference on Computer Vision*, 2020.

[76] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 670–679, 2017.

[77] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017.

[78] Simon Niklaus, Long Mai, and Oliver Wang. Revisiting adaptive convolutions for video frame interpolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1099–1109, 2021.

[79] Keunheung Park, Jinmi Kim, Sangyoon Kim, and Jonghoon Shin. Prediction of visual field from swept-source optical coherence tomography using deep learning algorithms. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 258:2489–2499, 2020.

[80] Keunheung Park, Jinmi Kim, and Jiwoong Lee. Visual field prediction using recurrent neural network. *Scientific reports*, 9(1):8385, 2019.

[81] Keunheung Park, Jinmi Kim, and Jiwoong Lee. A deep learning approach to predict visual field using optical coherence tomography. *PloS one*, 15(7):e0234902, 2020.

[82] Quang TM Pham, Jong Chul Han, and Jitae Shin. Visual field prediction with missing and noisy data based on distance-based loss. In *Workshop on Medical Image Learning with Limited and Noisy Data*, pages 199–205. Springer, 2022.

[83] Quang TM Pham, Jong Chul Han, Jitae Shin, et al. Multimodal deep learning model of predicting future visual field for glaucoma patients. *IEEE Access*, 11:19049–19058, 2023.

[84] Frédéric Pollet-Villard, Christophe Chiquet, Jean-Paul Romanet, Christian Noel, and Florent Aptel. Structure-function relationships with spectral-domain optical coherence tomography retinal nerve fiber layer and optic nerve head measurements. *Investigative ophthalmology & visual science*, 55(5):2953–2962, 2014.

[85] H A Quigley and A T Broman. The number of people with glaucoma worldwide in 2010 and 2020. *British Journal of Ophthalmology*, 90(3):262–267, 2006.

[86] Harry A Quigley and Aimee T Broman. The number of people with glaucoma worldwide in 2010 and 2020. *British journal of ophthalmology*, 90(3):262–267, 2006.

[87] Ehsan Rahimy. Deep learning applications in ophthalmology. *Current opinion in ophthalmology*, 29(3):254–260, 2018.

[88] Pradeep Ramulu. Glaucoma and disability: which tasks are affected, and at what stage of disease? *Current opinion in ophthalmology*, 20(2):92, 2009.

[89] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4161–4170, 2017.

[90] Ali S Raza, Jungsuk Cho, Carlos GV de Moraes, Min Wang, Xian Zhang, Randy H Kardon, Jeffrey M Liebmann, Robert Ritch, and Donald C Hood. Retinal ganglion cell layer thickness and local visual field sensitivity in glaucoma. *Archives of ophthalmology*, 129(12):1529–1536, 2011.

[91] Serge Resnikoff, Donatella Pascolini, Daniel Etya'Ale, Ivo Kocur, Ramachandra Pararajasegaram, Gopal P Pokharel, and Silvio P Mariotti. Global data on visual impairment in the year 2002. *Bulletin of the world health organization*, 82(11):844–851, 2004.

[92] Shino Sato, Kazuyuki Hirooka, Tetsuya Baba, Kaori Tenkumo, Eri Nitta, and Fumio Shiraga. Correlation between the ganglion cell-inner plexiform layer thickness measured with cirrus hd-oct and macular visual field sensitivity measured with microperimetry. *Investigative ophthalmology & visual science*, 54(4):3046–3051, 2013.

[93] Giacomo Savini, Piero Barboni, Vincenzo Parisi, and Michele Carbonelli. The influence of axial length on retinal nerve fibre layer thickness and optic-disc size measurements by spectral-domain oct. *British Journal of Ophthalmology*, 96(1):57–61, 2012.

[94] Suman Sedai, Bhavna Antony, Hiroshi Ishikawa, Gadi Wollstein, Joel S Schuman, and Rahil Garnavi. Forecasting retinal nerve fiber layer thickness from multimodal temporal data incorporating oct volumes. *Ophthalmology Glaucoma*, 3(1):14–24, 2020.

[95] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221, 2017.

[96] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for

precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[97] Jonghoon Shin, Sungjoon Kim, Jinmi Kim, and Keunheung Park. Visual field inference from optical coherence tomography using deep learning algorithms: a comparison between devices. *Translational vision science & technology*, 10(7):4–4, 2021.

[98] Joong Won Shin, Kyung Rim Sung, and Sun-Won Park. Patterns of progressive ganglion cell–inner plexiform layer thinning in glaucoma detected by oct. *Ophthalmology*, 125(10):1515–1525, 2018.

[99] Youngseok Song, Hiroshi Ishikawa, Mengfei Wu, Yu-Ying Liu, Katie A Lucy, Fabio Lavinsky, Mengling Liu, Gadi Wollstein, and Joel S Schuman. Clinical prediction performance of glaucoma progression using a 2-dimensional continuous-time hidden markov model with structural and functional measurements. *Ophthalmology*, 125(9):1354–1361, 2018.

[100] Wu M Liu YY Lucy KA Lavinsky F Liu M Wollstein G Schuman JS Song Y, Ishikawa H. Clinical prediction performance of glaucoma progression using a 2-dimensional continuous-time hidden markov model with structural and functional measurements. *Invest. Ophthalmol. Vis. Sci.*, 59, 2018.

[101] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[102] Hiroki Sugiura, Taichi Kiwaki, Siamak Yousefi, Hiroshi Murata, Ryo Asaoka, and Kenji Yamanishi. Estimating glaucomatous visual sensitivity from retinal thickness with pattern-based regularization and visualization. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 783–792, 2018.

[103] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[104] Ou Tan, Vikas Chopra, Ake Tzu-Hui Lu, Joel S Schuman, Hiroshi Ishikawa, Gadi Wollstein, Rohit Varma, and David Huang. Detection of macular ganglion cell loss in glaucoma by fourier-domain optical coherence tomography. *Ophthalmology*, 116(12):2305–2314, 2009.

[105] Kaveri A Thakoor, Xinhui Li, Emmanouil Tsamis, Paul Sajda, and Donald C Hood. Enhancing the accuracy of glaucoma detection from oct probability maps using convolutional neural networks. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 2036–2040. IEEE, 2019.

[106] Yih-Chung Tham, Xiang Li, Tien Y Wong, Harry A Quigley, Tin Aung, and Ching-Yu Cheng. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*, 121(11):2081–2090, 2014.

[107] Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, Gavin Siew Wei Tan, Leopold Schmetterer, Pearse A Keane, and Tien Yin Wong. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175, 2019.

[108] Andrew Turpin, Geoff P Sampson, and Allison M McKendrick. Combining ganglion cell topology and data of patients with glaucoma to determine a structure–function map. *Investigative ophthalmology & visual science*, 50(7):3249–3256, 2009.

[109] Toshimitsu Uesaka, Kai Morino, Hiroki Sugiura, Taichi Kiwaki, Hiroshi Murata, Ryo Asaoka, and Kenji Yamanishi. Multi-view learning over retinal thickness and visual sensitivity on glaucomatous eyes. In *Proceedings of the 23rd ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, pages 2041–2050, 2017.

[110] Joost van Amersfoort, Wenzhe Shi, Alejandro Acosta, Francisco Massa, Johannes Totz, Zehan Wang, and Jose Caballero. Frame interpolation with multi-scale deep loss functions and generative adversarial networks. *arXiv preprint arXiv:1711.06045*, 2017.

[111] John VanBuren, Jacob J Oleson, Gideon KD Zamba, and Michael Wall. Integrating independent spatio-temporal replications to assess population trends in disease spread. *Statistics in medicine*, 35(28):5210–5221, 2016.

[112] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *international conference on machine learning*, pages 3560–3569. PMLR, 2017.

[113] Michael Wall, Kimberly R Woodward, Carrie K Doyle, and Gideon Zamba. The effective dynamic ranges of standard automated perimetry sizes iii and v and motion and matrix perimetry. *Archives of ophthalmology*, 128(5):570–576, 2010.

[114] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[115] Joanne C Wen, Cecilia S Lee, Pearse A Keane, Sa Xiao, Ariel S Rokem, Philip P Chen, Yue Wu, and Aaron Y Lee. Forecasting future humphrey visual fields using deep learning. *PloS one*, 14(4):e0214875, 2019.

[116] JM Wild, M Dengler-Harles, AET Searle, EC O'Neill, and SJ Crews. The influence of the learning effect on automated perimetry in patients with suspected glaucoma. *Acta ophthalmologica*, 67(5):537–545, 1989.

[117] Gadi Wollstein, Joel S Schuman, Lori L Price, Ali Aydin, Siobahn A Beaton, Paul C Stark, James G Fujimoto, and Hiroshi Ishikawa. Optical coherence tomography (oct) macular and peripapillary retinal nerve fiber layer measurements and automated visual fields. *American journal of ophthalmology*, 138(2):218–225, 2004.

[118] Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 416–431, 2018.

[119] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *Advances in Neural Information Processing Systems*, pages 1645–1654, 2019.

[120] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

[121] Masaaki Yoshida, Shiho Kunimatsu-Sanuki, Kazuko Omodaka, and Toru Nakazawa. Predicting the integrated visual field with wide-scan optical coherence tomography in glaucoma patients. *Current eye research*, 43(6):754–761, 2018.

[122] Siamak Yousefi, Michael H Goldbaum, Madhusudhanan Balasubramanian, Tzyy-Ping Jung, Robert N Weinreb, Felipe A Medeiros, Linda M Zangwill, Jeffrey M Liebmann, Christopher A Girkin, and Christopher Bowd. Glaucoma progression detection using structural retinal nerve fiber layer measurements and functional visual field points. *IEEE Transactions on Biomedical Engineering*, 61(4):1143–1154, 2013.

[123] Siamak Yousefi, Taichi Kiwaki, Yuhui Zheng, Hiroki Sugiura, Ryo Asaoka, Hiroshi Murata, Hans Lemij, and Kenji Yamanishi. Detection of longitudinal visual field progression in glaucoma using machine learning. *American journal of ophthalmology*, 193:71–79, 2018.

[124] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *The journal of machine learning research*, 17(1):2287–2318, 2016.

[125] Haoxian Zhang, Yang Zhao, and Ronggang Wang. A flexible recurrent residual pyramid network for video frame interpolation. In *European Conference on Computer Vision*, pages 474–491. Springer, 2020.

[126] Haogang Zhu, David P Crabb, Patricio G Schlottmann, Hans G Lemij, Nicolaas J Reus, Paul R Healey, Paul Mitchell, Tuan Ho, and David F Garway-Heath. Predicting visual function from the measurements of retinal nerve fiber layer structure. *Investigative ophthalmology & visual science*, 51(11):5657–5666, 2010.

[127] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results, 2018.

[128] Yu Zhu, Hao Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. What to do next: Modeling user behaviors by time-lstm. In *IJCAI*, volume 17, pages 3602–3608, 2017.

List of Publications

Journals

- **Chen, Z.**, Ishikawa, H., Wollstein, G., Wang, Y., & Schuman, J. S. (2023). Deep-Learning-Based Group Point-Wise Spatial Mapping of Structure to Function in Glaucoma. (manuscript submitted to Ophthalmology Science)
- **Chen, Z.**, Shemuelian, E., Wollstein, G., Wang, Y., Ishikawa, H., & Schuman, J. S. (2023). Segmentation-Free OCT-Volume-Based Deep Learning Model Improves Pointwise Visual Field Sensitivity Estimation. Translational Vision Science & Technology, 12(6), 28-28.
- Liu, H., Lu, M., **Chen, Z.**, Cao, X., Ma, Z., & Wang, Y. (2022). End-to-end neural video coding using a compound spatiotemporal representation. IEEE Transactions on Circuits and Systems for Video Technology, 32(8), 5650-5662.
- **Chen, Z.**, Wang, R., Liu, H., & Wang, Y. (2021). PDWN: Pyramid deformable warping network for video interpolation. IEEE Open Journal of Signal Processing, 2, 413-424.

Conferences

- **Chen, Z.**, Ishikawa, H., Wollstein, G., Wang, Y., & Schuman, J. S. (2023). Deep-Learning-Based Group Point-Wise Spatial Mapping of Structure to Function in Glaucoma. Investigative Ophthalmology & Visual Science, 64(8), 344-344. (conference abstract)

- **Chen, Z.**, Shemuelian, E., Zheng, L., Wollstein, G., Wang, Y., Ishikawa, H., & Schuman, J. S. (2022). Segmentation-Free OCT-Volume-Based Deep Learning Model Improves Point-Wise Visual Field Threshold Estimation. Investigative Ophthalmology & Visual Science, 63(7), 852-852. (conference abstract)
- **Chen, Z.**, Zambrano, R., Wollstein, G., Schuman, J. S., & Ishikawa, H. (2021). OCT Denoising Performance Comparison on 2D and 1D Approaches. Investigative Ophthalmology & Visual Science, 62(8), 1785-1785. (conference abstract)
- **Chen, Z.**, Wang, Y., de los Angeles Ramos-Cadena, M., Wollstein, G., Schuman, J. S., & Ishikawa, H. (2020). Predicting Macular Progression Map Using Deep Learning. Investigative Ophthalmology & Visual Science, 61(7), 4532-4532. (conference abstract)
- **Chen, Z.**, Wang, Y., Wollstein, G., de los Angeles Ramos-Cadena, M., Schuman, J., & Ishikawa, H. (2020, April). Macular GCIPL thickness map prediction via time-aware convolutional LSTM. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE.

Book chapters

- **Chen, Z.**, Wollstein, G., Schuman, J. S., & Ishikawa, H. (2021). AI and Glaucoma. Artificial Intelligence in Ophthalmology, 113-125.
- **Chen, Z.**, & Ishikawa, H. (2021). Technical Aspects of Deep Learning in Ophthalmology. Artificial Intelligence in Ophthalmology, 69-75.