HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÙI THỊ THÙY

NGHIÊN CỨU KỸ THUẬT XỬ LÝ ẢNH DỰA VÀO CÔNG NGHỆ VI MẠCH QUANG TỬ TÍCH HỢP

LUẬN ÁN TIẾN SĨ KỸ THUẬT MÁY TÍNH

HÀ NỘI - 2023

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÙI THỊ THÙY

NGHIÊN CỨU KỸ THUẬT XỬ LÝ ẢNH DỰA VÀO CÔNG NGHỆ VI MẠCH QUANG TỬ TÍCH HỢP

Chuyển ngành : Kỹ thuật máy tính
Mã số : 9.18.01.06

LUẬN ÁN TIẾN SĨ KỸ THUẬT MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC
PGS.TS Lê Trung Thành
PGS.TS Đặng Thế Ngọc

HÀ NỘI - 2023

LỜI CAM ĐOAN

Tôi xin cam đoan rằng các kết quả khoa học được trình bày trong Luận án này là thành quả nghiên cứu của tôi trong suốt thời gian làm nghiên cứu sinh và chưa từng xuất hiện trong các công bố của các tác giả khác. Các kết quả đạt được là hoàn toàn chính xác và trung thực.

Nghiên cứu sinh

LỜI CẢM ƠN

Trong quá trình nghiên cứu, triển khai và hoàn thành Luận án, nghiên cứu sinh đã nhận được nhiều sự giúp đỡ, động viên quý báu của các thầy cô giáo, các nhà khoa học và bạn bè đồng nghiệp. Nghiên cứu sinh xin được bày tỏ lòng biết ơn sâu sắc nhất đến **PGS.TS. Lê Trung Thành** và **PGS.TS. Đặng Thế Ngọc** đã hướng dẫn, giúp đỡ tận tình, tạo mọi điều kiện thuận lợi cho nghiên cứu sinh trong học tập, nghiên cứu hoàn thành Luận án.

Nghiên cứu sinh cũng xin bày tỏ sự cảm ơn sâu sắc đến các thầy, cô trong Học viện Công nghệ Bưu chính Viễn thông; các thầy cô, cán bộ tại Khoa Đào tạo Sau đại học, Khoa Công nghệ Thông tin, Kỹ thuật Điện tử đã giảng dạy, giúp đỡ cho nghiên cứu sinh trong quá trình học tập và nghiên cứu. Nghiên cứu sinh xin trân trọng gửi lời cảm ơn đến các đồng nghiệp trong Trường Đại học Tài nguyên và Môi trường Hà Nội, Trường Đại học FPT – nơi nghiên cứu sinh mới chuyển công tác về và Trường Quốc tế, ĐH Quốc gia Hà Nội đã giúp đỡ, tạo điều kiện cho nghiên cứu sinh trong học tập và nghiên cứu để hoàn thành tốt Luận án này.

Cuối cùng, nghiên cứu sinh cũng xin được cảm ơn gia đình, bố mẹ, bạn bè, đồng nghiệp, đã cộng tác góp ý trao đổi để nghiên cứu sinh có điều kiện hoàn thành kết quả nghiên cứu của mình. Do vấn đề nghiên cứu có tính liên ngành, là vấn đề mới, đang phát triển và do kiến thức còn hạn chế, thời gian có hạn nên chắc rằng không tránh khỏi thiếu sót. Nghiên cứu sinh mong rằng sẽ nhận được nhiều sự quan tâm góp ý của các thầy, cô, các bạn bè đồng nghiệp trong và ngoài Trường để luận án được hoàn thiện hơn và tiếp tục được mở rộng nghiên cứu với những kết quả thu được trong giai đoạn sau này.

Hà Nội, tháng 5 năm 2023

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ VIẾT TẮT

TT	Từ viết tắt	Tiếng Anh	Tiếng Việt
1	JPEG	Joint Photographic Experts Group	Định dạng ảnh JPEG
2	CCD	Charge Coupled Device	Cảm biến CCD
3	CPU	Central Processing Unit	Đơn vị xử lý trung tâm
4	CS	Compressed sensing	Cảm biến nén
5	MMD	Micro Mirror Devices	Thiết bị vi gương kỹ thuật số
6	OCNN	Optical Convolutional Neural networks	Mạng nơ – ron nhân chập quang
7	ASP	Angle Sensitive Pixels	Camera ASP
8	CNN	Convolutional Neural Network	Mạng nơ – ron tích chập
9	GPU	Graphic Processing Unit	Đơn vị xử lý đồ họa
10	ANN	Artificial Neural Network	Mạng nơ – ron nhân tạo
11	ASIC	Application-specific integrated circuit	Mạch tích hợp cho ứng dụng cụ thể
12	FPGA	Field Programmable Gate Array	Vi mạch dùng cấu trúc mảng phần tử logic có thể lập trình được
13	ONN	Optical Neural Network	Mạng nơ – ron toàn quang học
14	WDM	Wavelength Division Multplexer	Bộ phân chia bước sóng
15	OCU	Optical Convolutional Uint	Đơn vị tích chập quang học
16	OEO	Optical – Electronic – Optical	Các bước chuyển đổi quang điện – điện quang
17	DWT	Discrete Wavelet Transform	Biến đổi Wavelet rời rạc
18	DHT	Discrete Haar Transform	Biến đổi Haar rời rạc
19	PLC	Programmable Logic Controller	Bộ điều khiển logic khả trình
20	PIC	Photonic Integrated-Circuits	Mạch tích hợp quang tử
21	MMI	Multimode interference	Bộ ghép giao thoa đa mode
22	DCT	Discrete Cosine Transform	Biến đổi Cosine rời rạc
23	DST	Discrete Sine Transform	Biến đổi Since rời rạc

TT	Từ viết tắt	Tiếng Anh	Tiếng Việt
24	KLT	Karhunen–Loève Transform	Biến đổi Karhunen–Loève
25	CMOS	Complementary Metal-Oxide Semiconductor	Công nghệ chế tạo vi mạch CMOS
26	FDTD	Finite Difference Time Domain	Miền thời gian chênh lệch hữu hạn
27	EME	Eigen-Mode Expansion	Mở rộng chế độ Eigen
28	BPM	Beam Propagation Method	Phương pháp truyền dẫn chùm
29	MNIST	Modified National Institute of Standards and Technology database	Cơ sở dữ liệu lớn chứa các chữ số viết tay
30	RGB	Red – Green – Blue	Hệ màu Đỏ - Xanh – Lục
31	ADC	Analog-to-Digital Converter	Bộ chuyển đổi Analog sang kỹ thuật số
32	DFT	Discrete Fourier Transform	Biến đổi Fourier rời rạc
33	SLM	Spatial light modulator	Bộ điều biến ánh sáng không gian
34	ReLU	Rectified Linear Unit	Đơn vị tuyến tính chỉnh lưu
35	ELU	Exponential Linear Unit	Đơn vị tuyến tính hàm mũ
36	OR	Or	Phép toán logic Hoặc
37	AND	And	Phép toán logic Và
38	NAND	NOT AND	Nghịch đảo của AND
39	MLP	Multiple Layer Perceptron	Mạng nơ-ron đa lớp
40	RNN	Recurrent Neural Network	Mạng nơ-ron tái diễn
41	TPU	Tensor Processing Unit	Bộ xử lý Tensor
42	MZI		Giao thoa kế Mach-Zehnder
43	MRR	Micro-Ring Resonators	Cấu trúc vi cộng hưởng MRR
44	SOA	Semiconductor Optical Amplifier	Khuếch đại quang bán dẫn SOA
45	CR	Compressed ratio	Tỷ lệ nén
46	MSE	Mean square error	Sai số bình phương trung bình
47	PSNR	Peak Signal to Noise Ratio	Tỷ số tín hiệu trên tạp âm đỉnh
48	AI	Artifical Intelligence	Trí tuệ nhân tạo
49	ARM	Acorn RISC Machine	Máy Acorn RISC

TT	Từ viết tắt	Tiếng Anh	Tiếng Việt
50	VR	Virtual Reality	Công nghệ hực tế ảo
51	AR	Reality	Thực tế tăng cường
52	VLSI	Very Large-Scale Integration	Rất thích hợp với quy mô lớn
53	OVMM	Optical Vector Matrix Multiplication	Phép nhân ma trận vectơ quang
54	OONN	On Chip Optical Neural Networks	mạng nơ-ron quang học trên chip
55	MVM	Multi Vector Matrix	Vecto ma trận quang
56	WDM	Wavelength Division Multiplexing	Phương thức ghép kênh quang theo bước sóng
57	GSW	Graphene Silicon Nitride Waveguide	Ống dẫn sóng Graphene Silicon Nitride

DANH MỤC CÁC KÝ HIỆU

STT	Ký hiệu	Ý nghĩa
1	x_i	Dữ liệu ảnh đầu vào
2	w_i	Hệ số bộ lọc Kernel
3	b	Hằng số bias
4	w_{ij}	Hệ số ma trận bộ lọc nhân chập
5	L_π	Chiều dài phách của bộ MMI
6	W_{MMI}	Độ rộng MMI
7	L_{MMI}	Chiều dài MMI
8	λ	Bước sóng
9	n_{eff}	Chiết suất hiệu dụng
10	$x(i,j)$	Pixel tại (i,j)
11	T_p	Công suất ra chuẩn hóa tại cổng “pass”
12	T_d	Công suất ra chuẩn hóa tại cổng “drop”
13	V_g	Điện áp cổng đặt vào graphene
14	ϕ	Pha tín hiệu
15	α	Hệ số suy hao ống dẫn sóng
16	R	Bán kính vi cộng hưởng
17	a_{ij}	Hệ số biên độ phức của ma trận
18	δ	Sai số
19	k	Hằng số lan truyền
20	E_m	Biên độ phức tín hiệu truyền trong MMI
21	T_{uv}	Ma trận trung gian
22	M_{DST}	Ma trận DST
23	M_{DCT}	Ma trận DCT

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ

MỞ ĐẦU

1. Sự cần thiết của đề tài nghiên cứu

Trong kỷ nguyên của Internet, yêu cầu về lưu trữ, xử lý, truyền dẫn dữ liệu ngày càng tăng. Theo ước tính, dữ liệu tăng trung bình 40% một năm, trong đó khoảng 90% dung lượng dữ liệu ảnh và video [1]. Một trong những mục tiêu quan trọng của kỹ thuật xử lý ảnh là thực hiện một số phân tích cụ thể và xử lý thông tin ảnh để đáp ứng nhu cầu của ứng dụng thực tế của con người và tâm lý học trực quan. Có hai loại công nghệ chính để thu nhận, xử lý ảnh là xử lý ảnh số và xử lý ảnh quang học. Bản thân các ảnh số được chuyển đổi từ tín hiệu quang. Do vậy, xử lý được trực tiếp tín hiệu ảnh trong miền toàn quang là mong muốn từ lâu.

Xử lý hình quang hay toàn quang là một công nghệ sử dụng mạch quang để xử lý, lưu trữ và truyền dẫn trực tiếp thông tin trong miền quang. Trước đây, quang học Fourier thường được sử dụng để thu nhận, tách biên, nhận dạng và bảo mật ảnh. Xử lý ảnh trực tiếp trong miền quang đặc biệt có ưu điểm là tốc độ cao (lên đến tốc độ ánh sáng), có khả năng xử lý thời gian thực và xử lý song song [2].

Ảnh số thường được biểu diễn bởi ma trận các điểm ảnh. Các ảnh số được số hóa từ ảnh quang và ảnh tương tự. Bản chất của ảnh số là một ma trận lưu trữ các số hay một chuỗi dữ liệu đã được số hóa. Do đó, xử lý ảnh số thường phải kết hợp với các thuật toán phần mềm và phần cứng. Nó có ưu điểm là độ chính xác xử lý cao, linh hoạt, dễ dàng điều chỉnh các bộ phận và khả năng xử lý phi tuyến phức tạp. Tuy nhiên, công nghệ này có nhược điểm là yêu cầu phần cứng cao và tốc độ tương đối chậm. Đặc biệt xử lý ảnh dữ liệu lớn thì rất khó khả thi và khó có khả năng xử lý trong thời gian thực. Hoặc ở mức độ nào đó, để xử lý thời gian thực đáp ứng các yêu cầu nhận dạng, lưu trữ và truyền dẫn, yêu cầu về phần cứng và phần mềm, các hệ thống tính toán rất phức tạp và đắt tiền.

Thêm vào đó, công suất tiêu thụ là một vấn đề lớn với hệ thống tính toán này do sự giới hạn về kích thước và khả năng tích hợp của các hệ thống máy tính hiện tại và vi mạch điện tử. Các nghiên cứu về tính toán, xử lý ảnh trực tiếp trong miền quang do đó là một chủ đề nghiên cứu mới của lĩnh vực kỹ thuật máy tính, xử lý thông tin, công nghệ thông tin để thay thế vượt qua các giới hạn của kỹ thuật xử lý ảnh số hiện tại, đặc biệt trong điều kiện xử lý một khối lượng lớn dữ liệu ảnh [3].

Sự phát triển nhanh chóng của công nghệ nano và chế tạo vi mạch quang tử cho các hệ thống tính toán và máy tính quang đã thúc đẩy nghiên cứu, thiết kế và ứng dụng các hệ thông quang tích hợp. Việc nghiên cứu về máy tính quang và hệ thống xử lý thông tin quang đang phát triển và được xem như sự phát triển của máy tính những năm 80 của thế kỷ trước. Theo dự báo, trong khoảng 10-15 năm nữa các hệ thống tính toán quang và lượng tử sẽ thay thế dần các hệ thống máy tính sử dụng công nghệ vi mạch điện tử hiện tại. Các vi mạch quang tử dần thay thế các thiết bị xử lý tín hiệu quang sử dụng các linh kiện quang hình và quang sợi có kích thước lớn, không có khả năng tích hợp.

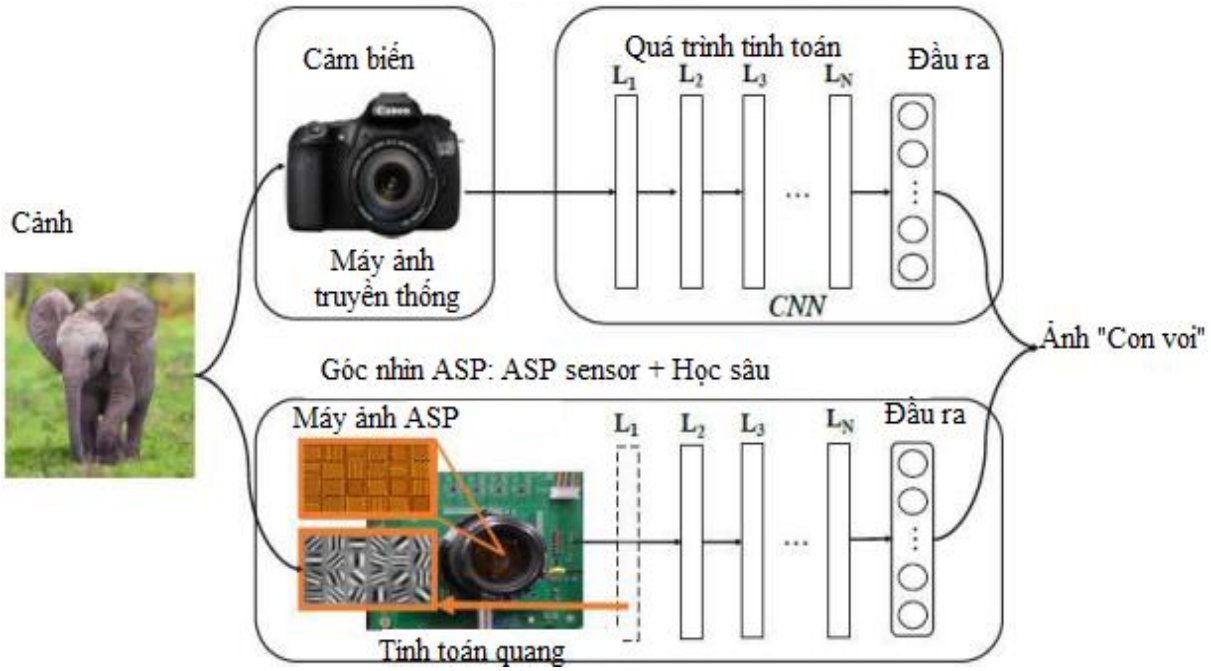
Khi lượng thông tin được truyền tải ngày càng lớn và tốc độ truyền tải trở nên nhanh hơn, nén dữ liệu đang trở thành một thách thức quan trọng trong ảnh video. Mục tiêu của nén ảnh là giảm sự không liên quan và dư thừa của dữ liệu ảnh để có thể lưu trữ hoặc truyền dữ liệu ở dạng hiệu quả hơn. Có hai chủ đề nghiên cứu chính trong lĩnh vực xử lý ảnh số là nén ảnh và mã hóa ảnh. Mục tiêu chung trong lĩnh vực này là giảm số lượng dữ liệu được truyền (nén) và bảo vệ việc sử dụng dữ liệu chống lại truy cập trái phép (mã hóa). Nén dữ liệu đề cập đến quá trình giảm lượng dữ liệu cần thiết để biểu diễn, lưu trữ và truyền đi một lượng thông tin nhất định. Hiện nay có nhiều kỹ thuật nén ảnh, nhưng phân làm hai loại chính là nén có tổn hao (mất mát thông tin) và không tổn hao. Cả hai phương pháp đều liên quan đến 3 loại thông tin về phổ, không gian và thời gian.

Nén không tổn hao, ví dụ, kỹ thuật Lempel-Ziv-Welch, được ưu tiên cho mục đích lưu trữ và thường được sử dụng cho hình ảnh y tế. Các phương pháp nén tổn hao, ví dụ JPEG, đặc biệt khi được sử dụng ở tốc độ bit thấp. Phương pháp suy hao đặc biệt thích hợp cho các ảnh tự nhiên [2]. Chụp, phân tích và mô tả đặc điểm ảnh tốc độ cao đã biến đổi các lĩnh vực như kính hiển vi thông lượng cao và thị giác máy tính. Sử dụng các kỹ thuật truyền thống, việc thu nhận hình ảnh được thực hiện trong miền điện tử bằng cách sử dụng cảm biến hình ảnh hoặc CCD. Tuy nhiên, các thiết bị này có hai hạn chế lớn: Thứ nhất, tốc độ khung hình cho các máy dò dựa trên mảng bị giới hạn ở một vài MHz đọc liên tục do tốc độ truyền dữ liệu điện tử chậm. Thứ hai, thời gian phơi sáng pixel là một hàm của thời gian sạc thiết bị và không thể giảm tùy ý, do đó dẫn đến hiện tượng nhòe hình ảnh.

Các nghiên cứu gần đây đã tập trung vào việc giảm bớt những thiếu sót này bằng cách khai thác các công nghệ cáp quang. Trước đây các hệ thống xử lý ảnh dùng biến

đổi ảnh được thực hiện trong miền điện qua phần cứng và phần mềm. Việc xử lý dữ liệu ảnh như kỹ thuật nén ảnh trực tiếp trong miền quang sẽ giảm được thời gian, dung lượng lưu trữ và tăng băng thông hệ thống truyền dẫn. Do đó, việc xử lý dữ liệu ảnh trực tiếp trong miền quang đang trở thành chủ đề nghiên cứu hấp dẫn do có khả năng xử lý dữ liệu lớn thời gian thực và có thể trực tiếp truyền qua mạng thông tin quang tốc độ cao. Đã có một số nghiên cứu gần đây xử lý ảnh trong miền quang sử dụng sợi quang, cấu trúc ghép có hướng, các cấu trúc siêu vật liệu bề mặt,... Mặc dù các hệ thống này xử lý tốc độ cao nhưng khó có thể tích hợp để hướng đến máy tính toàn quang trong tương lai [4, 5, 6].

Với nhu cầu gia tăng về tốc độ xử lý ảnh, việc thu thập, lưu trữ và xử lý dữ liệu hình ảnh trong lĩnh vực hiện nay có một nút thắt cổ chai nghiêm trọng. Bằng cách chuyển một số tác vụ xử lý tín hiệu thông thường như đệm, số hóa, biến đổi và nén dữ liệu sang miền quang tử, có thể giảm đáng kể khối lượng công việc của máy tính điện tử. Đặc biệt, các phép biến đổi tuyến tính thời gian thực, là một trong những tác vụ xử lý tín hiệu cơ bản nhất, chiếm một lượng đáng kể sức mạnh xử lý trên CPU... Cảm biến nén (CS) là một lĩnh vực khác đã thu hút nhiều sự chú ý gần đây. Hầu hết các công việc ban đầu trong lĩnh vực này đều dựa trên máy ảnh pixel đơn kết hợp các thiết bị vi gương kỹ thuật số (MMD-micro mirror devices) [7]. Kể từ đó, CS đã được áp dụng cho các lĩnh vực như kính hiển vi huỳnh quang, hình ảnh 3D, hình ảnh siêu kính, và thu thập video tốc độ cao. Gần đây, một nghiên cứu về máy ảnh CS tốc độ cao có khả năng chụp ảnh ở 39,6 Giga megapixel/s với hình ảnh được nén xuống 2% so với kích thước ban đầu của chúng [8]. Mặc dù tốc độ thu thập và tốc độ nén ấn tượng, việc tạo lại hình ảnh bằng CS đòi hỏi các thuật toán tốn nhiều thời gian, điều này gây ra thách thức khi mong muốn xử lý tín hiệu theo thời gian thực. Đặc biệt, năm 2016 lần đầu tiên các nhà khoa học tại Đại học Rice và Cornell đã tích hợp hệ thống mạng nơ-ron nhân chập quang trực tiếp với camera ASP để xứ lý ảnh trong các cảm biến hình ảnh [9] như chỉ ra ở Hình 1 dưới đây:



Hình 1. Hệ thống mạng nơron tích hợp với camera ASP

Các hệ thống camera thế hệ cũ thường có một số nhược điểm:

(1) Về công suất tiêu thụ yêu cầu cao: Thường chiếm hơn 50% tiêu thụ điện năng trong nhiều ứng dụng thị giác nhúng. Ngoài ra, cảm biến hình ảnh hiện tại không được tối ưu hóa để tiết kiệm đáng kể điện năng cho tầm nhìn máy tính;

(2) Về công suất tính toán: Mạng CNN cung cấp rất nhiều lợi ích hiệu suất, cũng làm tăng đáng kể độ phức tạp tính toán. Đơn vị xử lý đồ họa và các bộ xử lý đa lõi yêu cầu công suất tiêu thụ cao;

(3) Về băng thông dữ liệu: Yêu cầu rất nghiêm ngặt với các hệ thống kiến trúc camera truyền thống. Độ phân giải hình ảnh vừa phải 1 megapixel ở 30 fps (khung hình/giây) dẫn đến yêu cầu băng thông trên 0,5 Gbps. Điều này tạo ra các nghẽn khi truyền hình ảnh từ camera, các sensor đến CPU và làm tăng công suất, tăng bộ nhớ và độ phức tạp hệ thống. Hệ thống tích hợp mạng CNN giải quyết được các nhược điểm trên của hệ thống camera hình ảnh truyền thống.

Bên cạnh đó, máy tính có thể học, kết hợp và phân tích lượng lớn thông tin một cách nhanh chóng, hiệu quả và không cần hướng dẫn rõ ràng đang nổi lên như một công cụ mạnh mẽ để xử lý các tập dữ liệu lớn. Các thuật toán học sâu đã nhận được sự quan tâm bùng nổ trong cả giới học thuật và công nghiệp vì tiện ích của chúng trong nhận dạng hình ảnh, dịch ngôn ngữ, các vấn đề ra quyết định. Các đơn vị xử lý trung

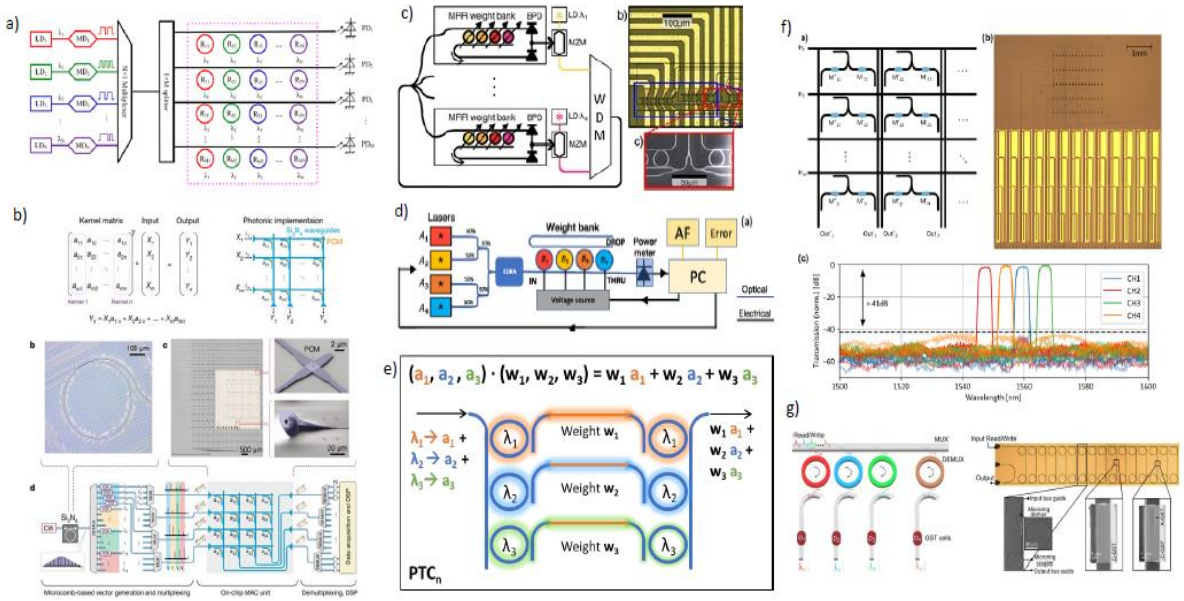
tâm truyền thống (CPU) là không tối ưu để triển khai các thuật toán này và nỗ lực ngày càng tăng trong giới học thuật và công nghiệp đã hướng tới việc phát triển các kiến trúc phần cứng mới phù hợp với các ứng dụng trong mạng nơ-ron nhân tạo (ANN) và học sâu. Các đơn vị xử lý đồ họa (GPU), mạch tích hợp ASIC và FPGA đã cải thiện cả hiệu quả năng lượng và tăng cường tốc độ cho các tác vụ. Luận án đưa ra một kiến trúc thực hiện mạng nơ-ron quang tử mới thực hiện các chức năng xử lý ảnh như phân loại và tách biên ảnh.

Gần đây, học máy (Machine Learning-ML) đã được quan tâm đặc biệt trở lại do sự gia tăng theo cấp số nhân của các hệ thống máy tính hiệu suất cao, tạo ra một môi trường nơi các mạng nơ-ron sâu DNN (Deep Neural Network) có thể có hàng chục lớp và hàng triệu tham số. Một ví dụ có cho thấy tất cả tiềm năng của phương pháp này được gọi là DALL E2, một trong những DNN chuyển văn bản thành hình ảnh tiên tiến nhất, với hơn 3,5 tỷ tham số [10]. Các mạng lớn và mở rộng như vậy đặt ra một yêu cầu rất lớn về sức mạnh tính toán [11]. Kéo theo đó là sự thách thức của công nghệ hiện tại về phần cứng, độ trễ và điện năng tiêu thụ. Tính linh hoạt và khả năng mở rộng của thiết bị điện tử kỹ thuật số đã cho phép tạo ra một khuôn mẫu nơi các mạng nơron (Neural Networks-NN) có thể được mã hóa, thử nghiệm và sử dụng [12].

Hiện nay yêu cầu về mạng nơron ngày càng lớn hơn, do vậy các các nhà khoa học trong và ngoài nước trong 1-2 năm trở lại đây đang tìm kiếm các giải pháp mới để theo kịp và cung cấp đủ mức hiệu suất để chạy NN [13]. Những giải pháp đó là dựa trên quy mô, bằng cách sử dụng phần cứng được kết nối với nhau trong dữ liệu trung tâm hoặc thay đổi kiến trúc mới, ví dụ như di chuyển từ CPU chung cho ứng dụng cụ thể, chẳng hạn dưới dạng FPGA, GPU hoặc ASIC, được gọi là lõi Tensor [14, 15, 16]. Tuy nhiên, các hệ thống hiện nay còn tồn tại một số hạn chế rất lớn do có nhiều lý do giới hạn vật lý, chẳng hạn như tiêu thụ năng lượng và độ trễ [17]. Vì những lý do này, các nhà khoa học đã bắt đầu tìm kiếm các công nghệ có thể cung cấp một bộ tăng tốc phần cứng tốt hơn cho mạng nơ-ron. Trong đó, quang học (hay quang tử-optics) đã được xem như một giải pháp thay thế cách tiếp cận để triển khai phần cứng NN hiệu quả, nhờ vào độ trễ của tốc độ ánh sáng và mức tiêu thụ năng lượng thấp [18, 19]. Hơn nữa, công nghệ quang tử silic (Silicon Photonics) đã bắt đầu trở thành một công nghệ đáng tin cậy và phổ biến, cho phép chế tạo hàng loạt mạch quang tử dùng công nghệ vi điện tử, thực hiện của máy gia tốc phần cứng mạng thần kinh quang tử (Photonic Neural Networks-PNN) tại quy mô chíp, để phù hợp hơn với nhu cầu của người dùng đầu cuối [20].

Mạng nơ-ron toàn quang (ONN-optical neural networks) cung cấp một cách tiếp cận thay thế đầy hứa hẹn cho việc triển khai vi điện tử và quang điện tử lai. Việc thiết kế thành công các mạng nơ-ron quang tử giải quyết được vấn đề tốc độ tính toán và công suất tiêu thụ của các hệ thống máy tính hiện tại. Năm 2017 [21], Shen và các nhà khoa học tại MIT và Stanford đã thành công trong việc thiết kế mạng nơ-ron toàn quang cho các thuật toán học sâu và ứng dụng trong nhận dạng âm thanh, hình ảnh. Từ đó, đã có nhiều công trình nghiên cứu về mạng nơ-ron quang tử ứng dụng trong nhận dạng, xử lý ảnh. Hầu hết các hệ thống này sử dụng cấu trúc vi cộng hưởng quang với bộ ghép có hướng và các cấu trúc giao thoa Mach Zehnder [22, 23, 24, 25]. Một số kiến trúc mạch tích hợp quang tử (Photonic Integrated Circuits-PIC) đã được đề xuất trên những năm trước để thực hiện các nhiệm vụ lõi Tensor cho PNN [26], [27].

Bằng cách cho phép điều khiển ánh sáng sử dụng ống dẫn sóng kích thước nhỏ, các mạch quang tử tích hợp có thể tích hợp một số lượng lớn trên một chip. Hoạt động của phép toán nhân và cộng tích lũy (Multiplication and Accumulation-MAC) được thực hiện trên quy mô nhỏ, sử dụng nhiều đầu vào, bộ điều chế tốc độ cao và bộ tách sóng quang. Kiến trúc sử dụng bộ ghép kênh theo bước sóng (Wavelength Division Multiplexing-WDM) để thực hiện phép nhân Ma trận-Vector được đưa ra gần đây trên Hình 2 trong đó: Hình 2(a) là kiến trúc đầu tiên được đề xuất bởi Yang et al. sử dụng vi cộng hưởng nối tiếp dùng bộ ghép có hướng [28]. Hình 2(b) kiến trúc khai thác các bộ ghép suy giảm thanh ngang, được đưa ra bởi Feldmann et al. [29].

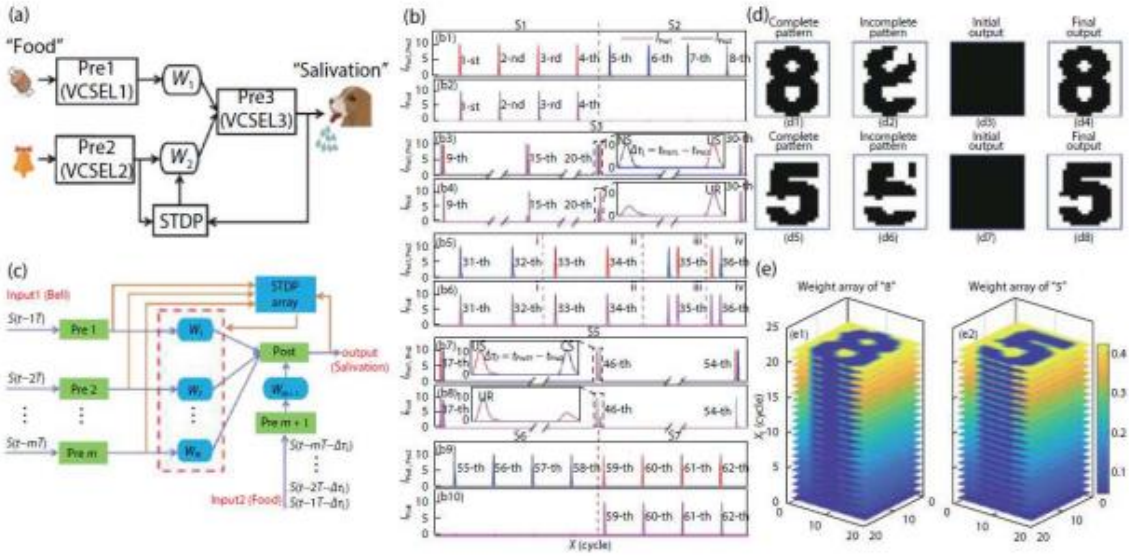


Hình 2. Kiến trúc thực hiện mạng noron quang tử [28], [29], [32], [33]

Hình 2(c) thực hiện đầu tiên của cách tiếp cận "quảng bá và trọng lượng" (broadcast-weight) từ Tait et al.[30] để thực hiện kiến trúc nhân và cộng ma trận cho mạng nơron. Hình 2(d) là cách tiếp cận "quảng bá và trọng lượng" tương tự, có thể thực hiện đào tạo và kiểm tra mạng Hopfield [31]. Hình 2(e) Triển khai phép nhân ma trận WDM bằng cách sử dụng các bộ cộng hưởng vi vòng bổ sung, được thực hiện bởi Ma et al. [32]. Hình 2(f) là kiến trúc dùng cách tử Bragg để thực hiện nơron [33]. Cuối cùng Hình 2(g) là phương pháp sử dụng kiến trúc vi cộng hưởng kết hợp vật liệu thay đổi pha để thực hiện mạng nơron quang tử [34].

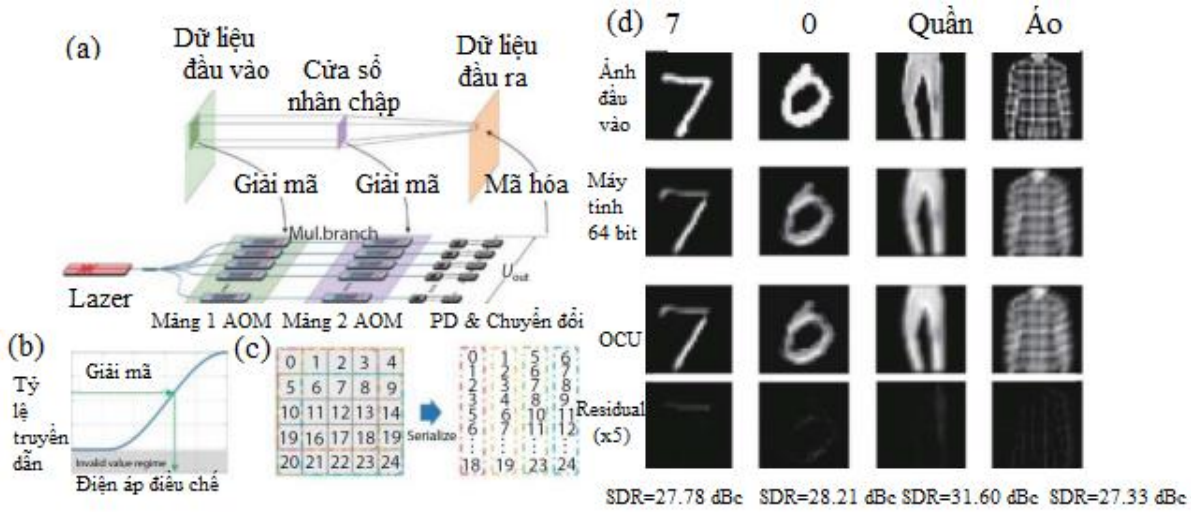
Nhược điểm của các hệ thống này là yêu cầu các hệ thống off-chip như bộ phận chia bước sóng WDM, làm việc với hệ số dương, yêu cầu có một hệ thống điều khiển phản hồi phức tạp để đạt được các hệ số nhân (kernel) mong muốn. Do vậy việc nghiên cứu, thiết kế được các kiến trúc mạng nơ-ron nhân chập trong miền toàn quang giải quyết các nhược điểm trên là một chủ đề nghiên cứu đang được các nhà khoa học rất quan tâm. Luận án tập trung nghiên cứu để tìm giải pháp xử lý ảnh trong miền toàn quang, các hệ thống có khả năng tích hợp với camera và các hệ thống máy tính trong tương lai, đặc biệt là các hệ thống máy tính nhúng với khả năng xử lý dữ liệu lớn và tốc độ cao trong miền toàn quang.

Lấy ví dụ gần đây, trình học tập được mô phỏng trong nơron quang tử đơn được đưa ra [35]. Sơ đồ của mạng học liên kết quang tử được thể hiện trong Hình 3 trong đó Hình 3(b) cho thấy rằng cả quá trình học và quên kết hợp đều có thể đạt được nhờ quy tắc STDP quang tử. Sự nhớ lại mẫu dựa trên học tập kết hợp đã được chứng minh thêm trong SNN quang tử được trình bày trong Hình 3 (c). Mẫu hoàn chỉnh và mẫu không hoàn chỉnh của số 8 được thể hiện trong Hình 3(d1) và 3(d2), tương ứng. Hình 3 (d3) hiển thị đầu ra ban đầu (đầu ra cuối cùng) của số 8 trước [sau] quá trình học liên kết. Sự phát triển của trọng lượng khớp thần kinh tương ứng với việc nhớ lại mẫu số 8 được trình bày trong Hình 3(e1). Không mất tính tổng quát, Hình. 3 (d5, d6, d7, d8) và 14 (e2) cho thấy quá trình nhớ lại mẫu của số 5 và sự phát triển cân nặng tương ứng. Rõ ràng, mẫu không hoàn chỉnh có thể được phục hồi và việc nhớ lại mẫu được thực hiện dựa trên mạng học liên kết quang tử.



Hình 3. Sơ đồ về quá trình học dựa trên VCSEL quang tử [35]

Việc triển khai quang học của CNN với tốc độ hoạt động nhanh và hiệu quả năng lượng cao rất hấp dẫn do khả năng khai thác tính năng vượt trội của nó. Đơn vị tích chập quang học (OCU) có độ chính xác cao với các mảng bộ điều chế quang acousto xếp tầng được minh họa trong Hình 4 [36]. Dữ liệu đầu vào và hạt nhân tích chập được đưa vào các mảng bộ điều chế để thực hiện hoạt động. Với kế hoạch tái sử dụng phần cứng, các CNN phức tạp có thể được các đơn vị tiến hành. Trong Hình 4 kết quả tích chập trên máy tính kỹ thuật số và OCU được đề xuất được hiển thị để hỗ trợ tính khả thi.



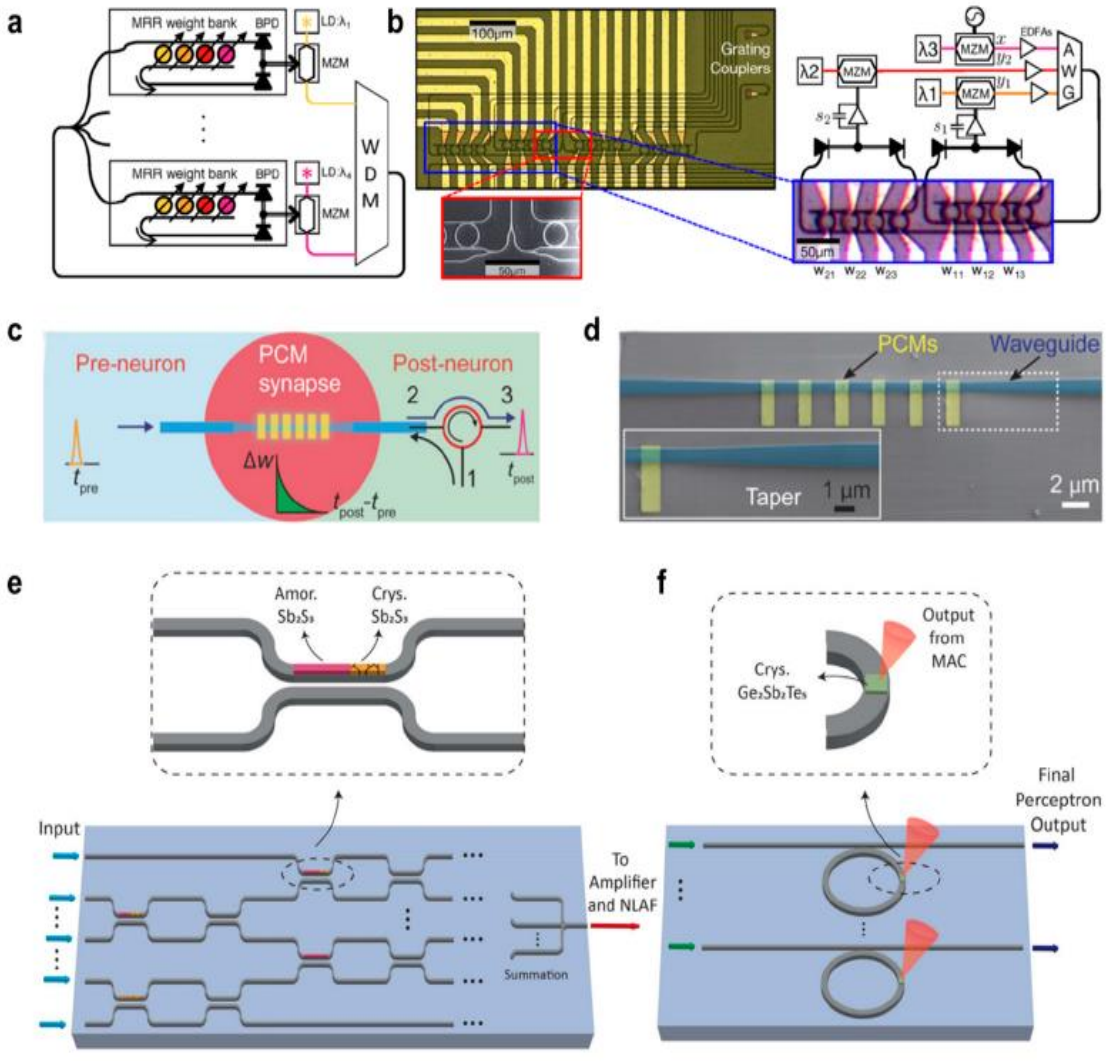
Hình 4. Kiến trúc mạng nơron quang dùng mảng điều chế [36]

Ảnh có độ phân giải cao cần thiết trong nhiều lĩnh vực, ví dụ: sinh học, chẩn đoán y tế, giám sát môi trường, v.v… tạo ra một lượng lớn dữ liệu cần được nén do giới hạn về dung lượng lưu trữ và băng thông, đặc biệt là đối với các ứng dụng giao tiếp thời gian thực [37]. Gần đây, mối quan tâm cao hơn đang được dành cho việc xử lý và nén dữ liệu bằng các mạch quang tử toàn phần, vì chúng có lợi ích trong việc khắc phục các hạn chế về độ trễ xuất phát từ các bước chuyển đổi quang điện-điện quang (OEO) và góp phần vào cải thiện thông tin liên lạc thời gian thực và tiêu thụ điện năng với chi phí thấp hơn [38]. Hơn nữa, trong các ứng dụng bao gồm một lượng lớn dữ liệu và cần tốc độ cao, các tác vụ xử lý thời gian thực chiếm một lượng đáng kể hiệu năng xử lý trong miền điện. Bằng cách chuyển một số tác vụ này sang miền quang, chẳng hạn như biến đổi và nén dữ liệu sang miền quang học, có thể đạt được yêu cầu xử lý và thời gian tính toán thấp hơn [39]. Nén dữ liệu được thực hiện thông qua việc giảm hoặc loại bỏ các thông tin dư thừa hoặc không đáng kể với sự hỗ trợ của hoạt động ngưỡng trên các phép biến đổi không gian toán học [40].

Các phép biến đổi không gian có thể đạt được sự phân rã tần số không gian của tín hiệu và cô lập các thành phần tần số cao, ít thiết yếu hơn đối với chất lượng nhận thức, trong một tập hợp các hệ số biến đổi riêng biệt. Biến đổi phổ biến nhất để xử lý và nén tín hiệu không cố định là biến đổi wavelet rời rạc (DWT). Trong số các phép biến đổi wavelet khác nhau, Biến đổi Haar (HT) [41] được chọn do tính ứng dụng cao và sức mạnh tính toán nhanh trong xử lý và / hoặc nén dữ liệu, hình ảnh. Nó cũng có thiết kế đơn giản và hiệu quả cao. Hơn nữa, nó còn có thêm một lợi thế nữa là dễ dàng nhận ra bởi các mạch sóng ánh sáng phẳng (PLC) hoặc mạch tích hợp quang tử (PIC), cung cấp một phương pháp suy hao toàn quang để nén ảnh theo thời gian thực [42]. Đối với những ứng dụng như đa phương tiện, yêu cầu mức độ nén ảnh cao hơn, các phương pháp nén mất dữ liệu thường được sử dụng. Các phương pháp không tổn hao không được xem xét trong nghiên cứu này vì chúng thường đạt được tỷ lệ nén thấp hơn, hiếm khi được sử dụng để nén ảnh. Các đơn đặt hàng khác nhau của HT có thể được thiết kế và triển khai thành PIC bằng cách sử dụng các cấu trúc quang tử khác nhau như bộ ghép định hướng không đối xứng [43, 6] và bộ ghép giao thoa đa mode (MMI) [44, 45]. Các mạch tích hợp quang tử dựa trên cấu trúc MMI có một số ưu điểm so với các mạch dựa trên bộ ghép bất đối xứng quang học, chẳng hạn như giảm kích thước, tổn hao quang học thấp hơn, băng thông cao hơn, tăng dung sai chế tạo và độ nhạy phân cực. Tuy nhiên, có thể đạt được mạch HT quang dựa trên bộ ghép bất

đối xứng quang học mà không cần bộ dịch pha và có thể có sai số pha thấp hơn mạch dựa trên bộ ghép MMI.

Quang tử silicon sử dụng các kỹ thuật chế tạo CMOS cơ bản và kết hợp điện tử và mạch quang tử. Hầu hết các công việc ban đầu về mạng thần kinh quang học trong quang tử silicon sử dụng cả quang học và điện tử. Trong phần này, Luận án khảo sát các cách tiếp cận khác nhau để thực hiện chức năng kiến trúc vi mô với các thiết bị quang tử silicon và thảo luận về sự khác biệt giữa đồng thiết kế quang tử điện tử và điện toán thần kinh, neuron toàn quang.

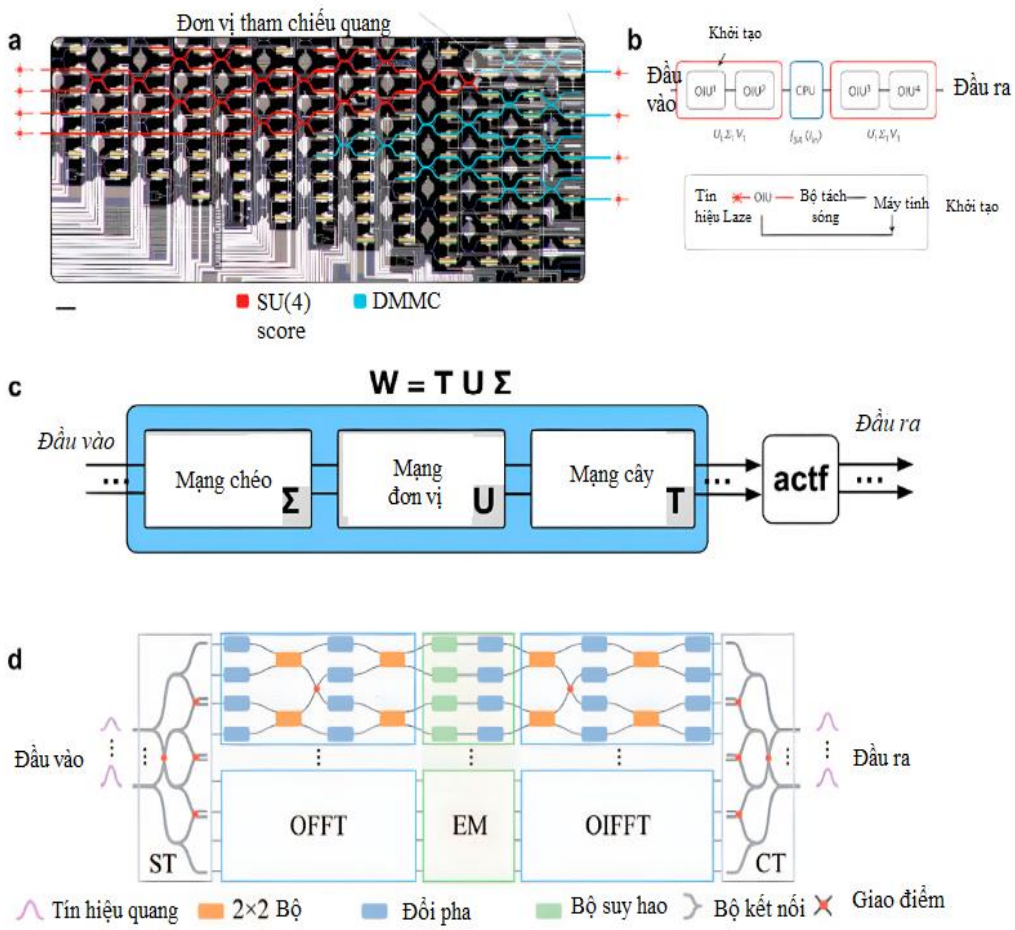


Hình 5. Các phương pháp tạo trọng số quang (weight) cho mạng nơ-ron quang tử [46] [47] [48]

+ Trọng số (weight): Chức năng trọng số là điều cần thiết để bắt chước một khớp thần kinh sinh học kể từ khi thay đổi trọng số là chức năng chính của việc học trong một mạng lưới thần kinh. Khi việc học tiếp tục, các tham số này được điều chỉnh theo các giá trị tạo ra đầu ra chính xác. vi cộng hưởng là một phương pháp phổ biến để điều chỉnh giá trị trọng lượng và lần đầu tiên được sử dụng để thực hiện hàm trọng số và phép nhân ma trận [46]. Hình 5 tổng kết các phương pháp tạo trọng số trong miền quang đến nay, trong đó: Hình 5a, 5b [47] sử dụng cấu trúc vi cộng hưởng nối tiếp dạng add-drop, Hình 5c sử dụng cấu trúc vật liệu thay đổi pha PCM (phase change material) và kết nối thành bộ ghép có hướng và cấu trúc Mach Zehnder như ở Hình 5e và 5f [48].

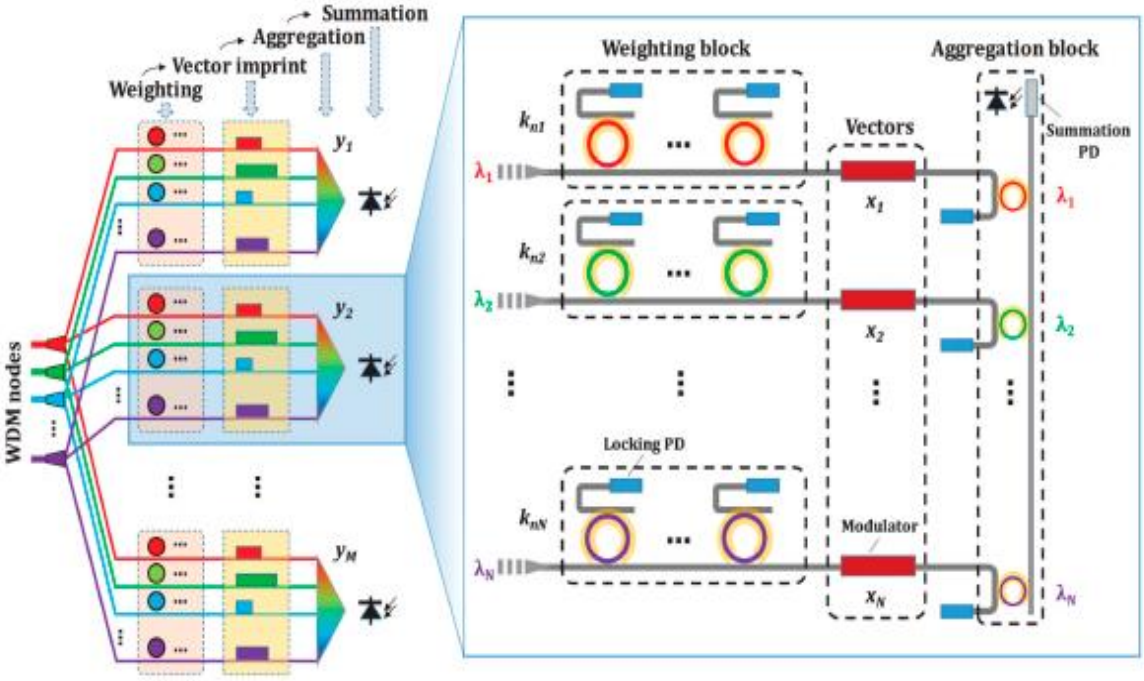
+ Thực hiện kiến trúc mạng nơ-ron: có 2 phương pháp chính được công bố đến nay là dùng MZI và vi cộng hưởng.

Giao thoa kế Mach–Zehnder (MZIs), được sử dụng rộng rãi trong các bộ điều chế quang học, giao tiếp quang và điện toán quang tử, là những thành phần quan trọng của mạng dựa trên giao thoa quang. Bằng cách sử dụng bộ suy hao và bộ dịch pha trên các nhánh MZI để điều khiển các trọng số nhằm thay đổi pha và biên độ của quang tín hiệu, MZI có thể hoạt động như một đơn vị nhân ma trận tự nhiên, từ đó có thể thực hiện chức năng nhân và cộng ma trận ứng dụng trong mạng nơ-ron như chỉ ra ở Hình 6 [21].



Hình 6. Mạng nơ-ron bằng kết nối MZI [21]

Kiến trúc thực hiện mạng nơ-ron dùng kết nối các bộ vi cộng hưởng được đưa ra gần đây như sơ đồ ở Hình 7 kết hợp miền điện và miền quang, gọi là đồng thiết kế quang-điện (co-design) [49]. Các tác giả đã đồng thiết kế kiến trúc quang tử silicon microring với FPGA, cung cấp một cách để xây dựng phép nhân ma trận quy mô lớn bằng cách sử dụng MRR trong miền bước sóng và giảm độ phức tạp của việc phân tách hệ thống. Trong trường hợp không có bộ điều khiển điện tử, mạng thần kinh đồng thiết kế quang tử điện tử là một lộ trình thiết thực hơn cho các ANN hiện tại cho đến khi hiệu quả của điều khiển điện tử có thể được tìm thấy như một ứng cử viên cạnh tranh trong miền quang. Mặc dù chế tạo nguyên khối mang lại cơ hội tốt để tích hợp điện tử và quang tử trên cùng một đế bán dẫn, độ trễ cao và mức tiêu thụ điện năng do các thành phần điện tử đặt ra những thách thức đối với bộ điều khiển điện tử. Trong ONN, bộ điều khiển sẽ quản lý các thiết bị quang tử và duy trì hoạt động ổn định của các nơ-ron trong thời gian thực, ở tốc độ cao và hiệu quả.



Hình 7. Mạng nơ-ron bằng kết nối vi cộng hưởng [49]

Các cấu trúc đề xuất phụ thuộc vào bộ ghép có hướng. Trong khi đó, bộ ghép này khó điều khiển để đạt được hệ số ghép mong muốn và có dung sai chế tạo nhỏ, kích thước lớn. Do vậy, Luận án này tìm cách giải quyết các bài toán xử lý ảnh trong miền quang sử dụng các bộ biến đổi và mạng nơ-ron sử dụng các cấu trúc giao thoa đa mode. Các cấu trúc mới được tạo thành có ưu điểm dựa vào các đặc tính của giao thoa đa mode như suy hao thấp, kích thước nhỏ, có độ chính xác cao, băng thông lớn và có thể điều khiển được.

2. Mục tiêu nghiên cứu của Luận án

Mục tiêu nghiên cứu của Luận án là thiết kế hệ thống xử lý ảnh trong miền toàn quang nhằm giải quyết bài toán tăng tốc độ tính toán, tích hợp với các hệ thống máy tính toàn quang trong tương lai, có kích thước nhỏ, độ suy hao thấp, băng thông lớn và độ chính xác cao. Luận án tập trung 2 mục tiêu chính:

- Thiết kế được các bộ biến đổi toàn quang tích hợp ứng dụng trong nén dữ liệu ảnh.

- Thiết kế được hệ thống mạng nơ-ron quang tích hợp khả trình ứng dụng cho tách biên và nhận dạng ảnh.

3. Nội dung nghiên cứu của Luận án

Luận án nghiên cứu về kỹ thuật xử lý ảnh số trong miền toàn quang, tập trung vào kỹ thuật nén ảnh sử dụng các bộ biến đổi ảnh như biến đổi Haar rời rạc (DHT), biến đổi cosine rời rạc (DCT) và biến đổi sine rời rạc (DST) và biến đổi KLT (KLT); nghiên cứu về mạng nơ-ron toàn quang và ứng dụng mạng nơ-ron toàn quang trong tách biên và nhận dạng ảnh. Các hệ thống được thiết kế sử dụng công nghệ chế tạo vi mạch CMOS hiện thời nhằm có khả năng tương thích với vi mạch điện tử hiện tại và thiết kế các hệ thống máy tính quang trong tương lai.

4. Đối tượng, phạm vi nghiên cứu và phương pháp nghiên cứu

Đối tượng nghiên cứu là các bộ biến đổi tín hiệu trong miền toàn quang, mạng nơ-ron quang tử tích hợp, kỹ thuật xử lý ảnh như nén ảnh, tách biên ảnh và nhận dạng ảnh. Luận án quan tâm đến thiết kế phần cứng cho các thế hệ máy tính quang.

Luận án sử dụng các mô hình toán học, phân tích giải tích để thiết kế lý thuyết các hệ thống biến đổi ảnh ứng dụng cho nén ảnh và các hệ thống mạng nơ-ron quang tử để tách biên, nhận dạng ảnh. Các kết quả lý thuyết sau đó được mô phỏng, phân tích, đánh giá và so sánh trong miền quang sử dụng phương pháp số như FDTD, EME, BPM

5. Các đóng góp của Luận án

Luận án đã có 2 nhóm đóng góp chính sau đây:

1. Thiết kế được các bộ biến đổi toàn quang DHT, DCT, KLT ứng dụng cho nén ảnh. Cấu trúc mới có khả năng tích hợp với hệ thống camera thông minh, xử lý dữ liệu tốc độ cao, băng thông lớn, thời gian thực. Các cấu trúc đề xuất được thiết kế đơn giản, có độ chính xác cao so với công nghệ vi mạch hiện nay.

2. Thiết kế được nơ-ron quang mới, từ đó đề xuất kiến trúc và thuật toán mạng nơ-ron quang ứng dụng cho tách biên ảnh và phân loại ảnh trong miền quang. Luận án thiết kế mới các bộ biến đổi trong miền quang có khả năng tích hợp với các hệ thống cảm biến và camera nhúng; thiết kế mạng nơ-ron quang ứng dụng trong tách biên và nhận dạng ảnh, có đóng góp cho các lĩnh vực kỹ thuật máy tính, công nghệ thông tin, xử lý dữ liệu và hệ thống máy tính hiệu năng cao.

6. Bố cục của Luận án

Luận án gồm 3 chương:

Chương 1: Trình bày tổng quan và cơ sở lý thuyết về xử lý ảnh số, nén ảnh sử dụng các biến đổi tín hiệu; lý thuyết về mạch quang và nguyên lý của mạng nơ-ron quang.

Chương 2: Trình bày các kết quả thiết kế bộ biến đổi tín hiệu DHT, DCT, KLT sử dụng các cấu trúc tích hợp quang mới dựa vào cấu trúc giao thoa đa mode 4×4 và 6×6 đầu vào/ra ứng dụng cho nén ảnh trong miền toàn quang. Các kết quả được thiết kế trên vật liệu Si_3N_4 phù hợp với công nghệ CMOS hiện tại và hoạt động trong dải tần nhìn thấy của các màu R, G và B.

Chương 3: Trình bày thiết kế nơ-ron quang mới, kiến trúc thực hiện tích chập trong miền quang (kernel) và mạng nơ-ron quang. Dựa vào kiến trúc mới kỹ thuật tách biên ảnh sử dụng toán tử Roberts, Sobel và Prewitt được thiết kế trong miền quang. Đồng thời, chương 3 mô phỏng, đánh giá mạng nơ-ron quang ứng dụng cho nhận dạng tập dữ liệu viết tay MNIST.

Chương 1. TỔNG QUAN VỀ TÌNH HÌNH NGHIÊN CỨU

Chương 1 trình bày một số cơ sở lý thuyết về xử lý ảnh số, biến đổi ảnh, mạng nơ-ron quang tử, vi mạch quang tử. Các nghiên cứu tập trung vào nguyên lý để thiết kế các phần cứng xử lý ảnh.

1.1 Tổng quan

Ở Việt Nam, xử lý ảnh là một lĩnh vực nghiên cứu khá mở. Có nhiều nhóm nghiên cứu, đề tài và các công trình nghiên cứu về các phương pháp xử lý ảnh trong miền điện. Tuy nhiên, nghiên cứu xử lý ảnh dùng quang tử tích hợp còn hạn chế và chưa có công trình, kết quả nghiên cứu được công bố.

Hiện vẫn còn chưa nhiều công trình nghiên cứu về thiết kế cấu trúc vi mạch quang ứng dụng trong xử lý tín hiệu, đặc biệt là ứng dụng trong thông tin lượng tử. Sự phát triển nhanh chóng của các dịch vụ mạng băng rộng là động lực thúc đẩy sự phát triển của mạng quang thế hệ kế tiếp dựa trên nền tảng các công nghệ ghép kênh phân chia bước sóng (xWDM). Trong tiến trình quang hóa mạng truyền thông, các nối chéo quang OXC (Optical Cross-connects) với chức năng chuyển mạch tuyến quang là công nghệ quan trọng cốt lõi, cho phép tăng cường khả năng đáp ứng của mạng với các biến động lưu lượng và tối ưu cấu hình mạng.

Một số nhóm nghiên cứu về xử lý tín hiệu quang tại Việt Nam như nhóm nghiên cứu của PGS.TS. Nguyễn Hoàng Hải – ĐH Bách Khoa Hà Nội và các cộng sự (Yoshinori Namihira, Shubi Kaijage, Feroza Begum, S. M. Abdur Razzak and K. Miyagi). Vào năm 2009, nhóm công bố kết quả nghiên cứu *"Dispersion Compensating Square Photonic Crystal Fiber for Optical Communication Systems*" và "*Broadband Nearly-Zero Ultra-Flattened Dispersion Single Mode Index Guiding Holey Fiber*" . Trọng tâm của các nghiên cứu này là nghiên cứu, thiết kế và ứng dụng các sợi quang trong truyền thông mạng quang.

Nhóm nghiên cứu của PGS.TS. Ngô Quang Minh đã nghiên cứu về quang tử tích hợp cấu trúc tinh thể (crystal). Luận án tiến sĩ của Hoàng Thu Trang, "Nghiên cứu, thiết kế cấu trúc tinh thể quang tử 1D và 2D ứng dụng cho linh kiện lưỡng trạng thái

ổn định" năm 2020 đã thiết kế cấu trúc quang tử tích hợp cho ứng dụng tạo trạng thái lưỡng ổn để xử lý tín hiệu.

Nhóm nghiên cứu của PGS.TS Lê Trung Thành – ĐH Quốc Gia Hà Nội và các cộng sự cũng tập trung nghiên cứu và có một số kết quả nghiên cứu về lĩnh vực này. Cụ thể, nhóm đã công bố kết quả chế tạo bộ biến đổi tín hiệu HAAR trong xử lý tín hiệu quang. Nhóm cũng đã công bố bài báo *"All-Optical Signal Processing Circuits Using Multimode Interference Structures on Silicon Waveguides*" cung cấp những lý thuyết nền tảng về giao thoa đa mode để xây dựng các mạch tích hợp xử lý tín hiệu toàn quang, đặc biệt là xử lý ảnh.

Năm 2012, nhóm tác giả Lê Trung Thành cũng đã công bố bài *báo The Design of Optical Signal Transforms Based on Planar Waveguides on a Silicon on Insulator Platform*. Trong đó có đưa ra những thiết kế mới cho các thành phần xử lý tín hiệu quang dựa trên nền quang tử Silic.

Cần nhấn mạnh rằng, các bộ xử lý tín hiệu trong miền quang như DHT, DCT, DFT và biến đổi wavelet,…đã được thiết kế và chứng tỏ ưu việt trong các ứng dụng phân tích phổ, lọc và mã hóa. Tuy nhiên các thiết kế này dựa vào công nghệ sợi quang, không có khả năng tích hợp và tiến đến tích hợp trên một bộ

Như vậy, có thể thấy công nghệ vi mạch quang tử tích hợp ngày càng có nhiều ứng dụng trong xử lý tín hiệu, đặc biệt trong xử lý ảnh. Đây cũng là một trong những vấn đề hiện nay đang được nhiều nhóm nghiên cứu tại Việt Nam bắt đầu quan tâm và đi sâu nghiên cứu.

Việc nghiên cứu xử lý tín hiệu toàn quang dùng sợi quang, phi tuyến sợi, bộ khuếch đại quang bán dẫn, thấu kính,… đã được quan tâm nghiên cứu trong nhiều năm qua trên khắp thế giới. Trong đó, nhiều thiết bị toàn quang và bộ xử lý toàn quang đã được thiết kế và chế tạo thành công. Tuy nhiên, hầu như chúng không có khả năng tích hợp, có kích thước lớn và do vậy cũng không sử dụng được công nghệ vi mạch hiện thời. Trong khoảng 10 năm trở lại đây, công nghệ vi mạch quang tử trên vật liệu silic

đã được phát triển thành công, điều này đã mở ra hướng mới về khả năng thiết kế, chế tạo các bộ xử lý, các bộ vi xử lý toàn quang sử dụng công nghệ VLSI hiện nay.

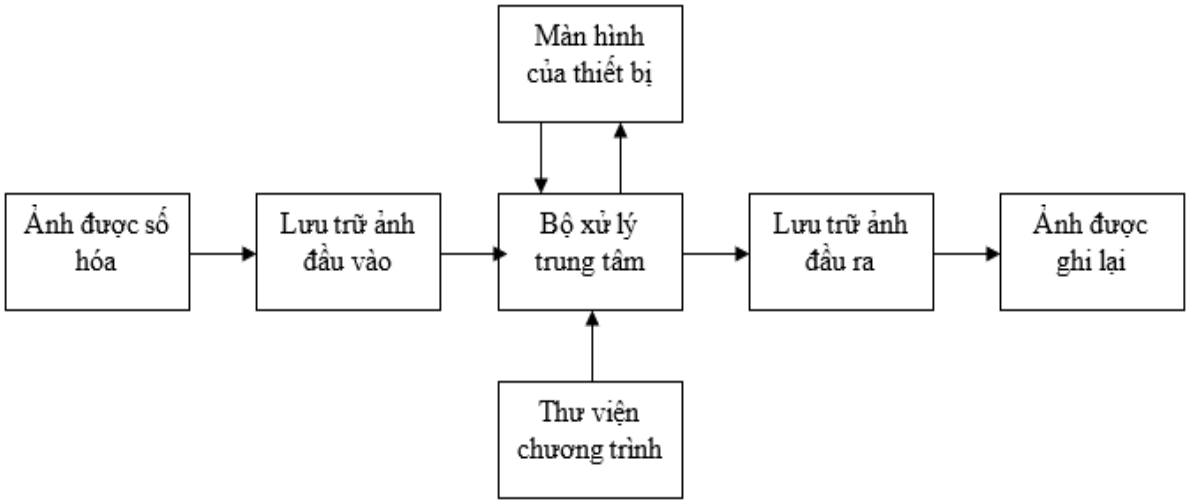
Đặc biệt, năm 2012, nhóm nghiên cứu Giorgia Parca, Pedro Teixeira, António Teixeira công bố bài báo "All-Optical Integrated System for 2D Data Wavelet Transform and Compression" là kết quả nghiên cứu đầu tiên về ứng dụng của các hệ thống quang tử trong xử lý ảnh. Trọng tâm của nghiên cứu này là thiết kế đưa ra hệ thống tích hợp toàn quang cho môi trường sóng nhằm nén và hủy dữ liệu quang 2D. Cũng trong năm 2012, nhóm còn công bố 1 kết quả nghiên cứu quan trọng trong vấn đề xử lý ảnh dựa vào truyền dẫn toàn quang để xử lý ảnh 3D. Các thiết kế cấu trúc MMI tham khảo từ cấu trúc MMI của nhóm do PGS.TS. Lê Trung Thành chủ trì.

Năm 2014, nhóm nghiên cứu gồm L. Almeida, N. Kumar, G. Parca, A. Tavares, A. Lopes, A. Teixeira công bố kết quả nghiên cứu thông qua bài báo "All-Optical image processing based on Integrated Optics" – "Xử lý ảnh toàn quang dựa trên các mạch quang tích hợp". Trọng tâm của nghiên cứu này là một bộ xử lý toàn quang dựa trên mạch quang tích hợp. Biến đổi sóng gián đoạn (Discrete Wavelet Transform – DWT) trong miền 2 chiều được áp dụng khóa dữ liệu của 1 bức ảnh có thể được thực thi bằng nhiều phương pháp nén. Do đó, thiết bị biến đổi sóng HAAR toàn quang được thiết kế có thể áp dụng DWT. Việc nén và xử lý ảnh bằng phương pháp sử dụng bộ biến đổi toàn quang HAAR cho tỷ lệ nén thấp hơn nhiều so với các phương pháp khác.

Công nghệ vi mạch quang tử tích hợp với những ưu điểm vượt trội về tốc độ xử lý đang là xu hướng nghiên cứu và phát triển trên thế giới, được kỳ vọng là sẽ thay thế công nghệ điện tử hiện nay. Bên cạnh đó, xử lý dữ liệu lớn, xử lý ảnh sử dụng các vi mạch quang đang là chủ đề nghiên cứu được quan tâm. Hiện chưa có đề xuất nào giải quyết được toàn bộ vấn đề đặt ra, và đó cũng là động lực để đặt mục tiêu nghiên cứu cho NCS về lĩnh vực này.

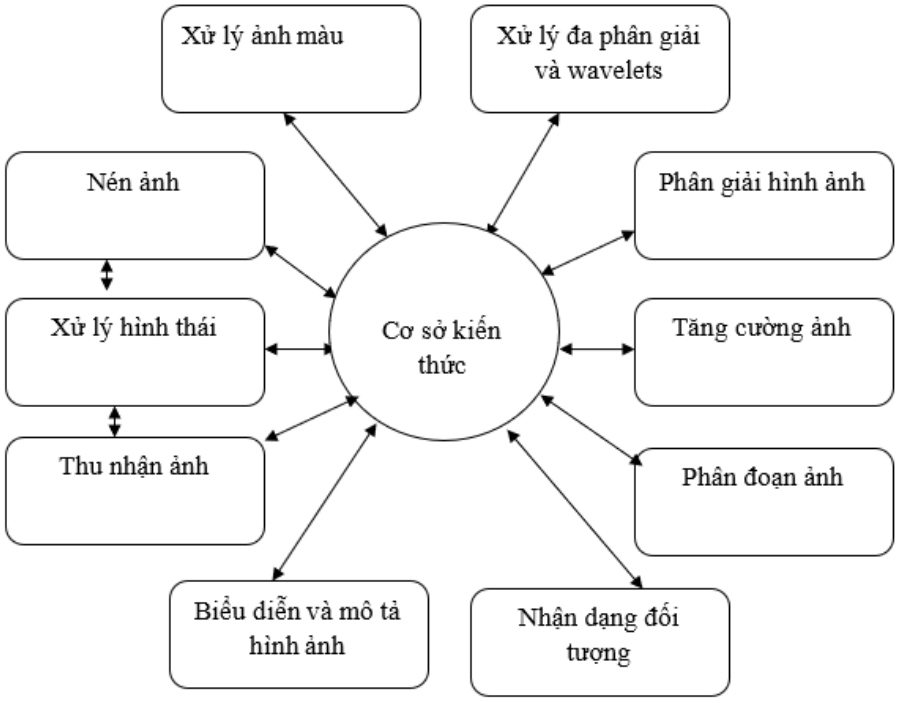
Hệ thống xử lý ảnh được chỉ ra ở Hình 1.1 [50]. Xử lý ảnh yêu cầu thao tác dữ liệu ảnh bằng cách sử dụng nhiều thiết bị điện tử và phần mềm. Cùng với các thiết bị, xử lý ảnh kỹ thuật số yêu cầu áp dụng các thuật toán khác nhau theo yêu cầu để

chuyển đổi hình ảnh vật lý thành ảnh kỹ thuật số để tìm nạp thông tin hoặc tính năng mong muốn.



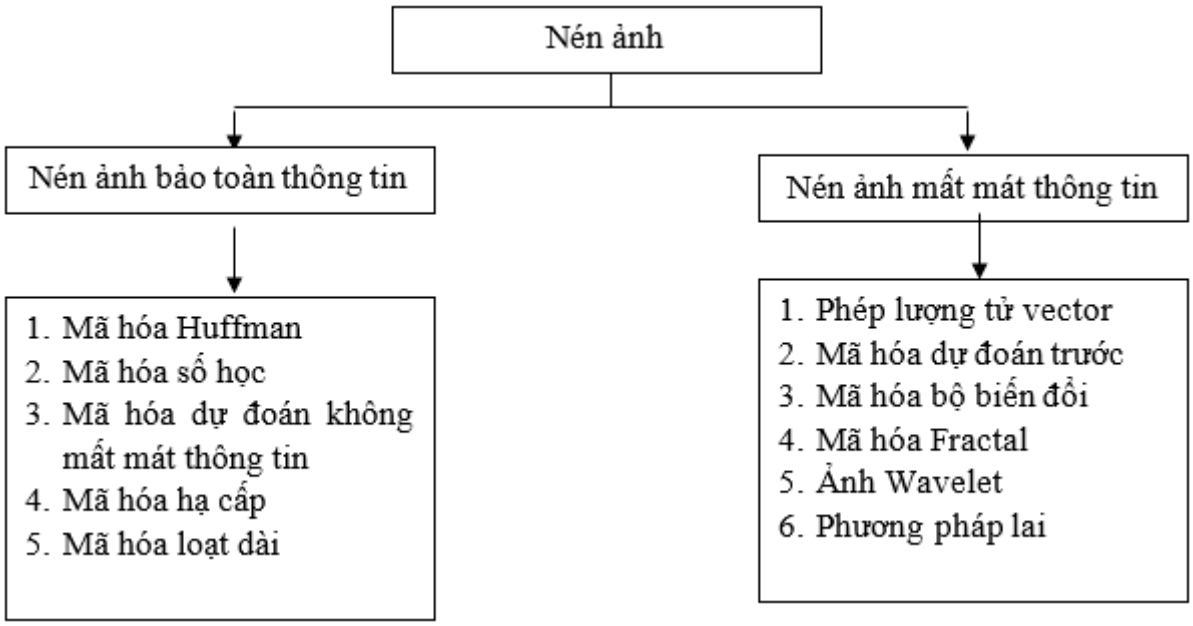
Hình 1.1: Quá trình xử lý ảnh số

Các khâu của xử lý ảnh được chỉ ra ở Hình 1.2, gồm thu nhận ảnh, nén ảnh, xử lý màu, xử lý độ phân dải, nâng cao ảnh, tách biên ảnh, nhận dạng đối tượng, mô tả ảnh, khôi phục ảnh... Tất cả dữ liệu ảnh tạo thành đầu vào hoặc đầu ra của một hệ tri thức. Luận án tập trung nghiên cứu 2 vấn đề chính là nén ảnh có suy hao sử dụng biến đổi tín hiệu và nhận dạng ảnh trong miền quang.



Hình 1.2: Các bài toán xử lý ảnh

Nén ảnh có thể chia ra làm nén có tổn hao và nén không tổn hao. Việc phân loại các phương pháp nén ảnh được chỉ ra ở Hình 1.3. Phương pháp nén ảnh sử dụng biến đổi tín hiệu chuyển các pixel trong miền ảnh thành một miền khác để chuẩn bị một tập hợp các hệ số với cách biểu diễn tự nhiên và nhỏ gọn hơn. Để đạt được điều này, trước đây mã hóa biến đổi sử dụng biến đổi Fourier ánh xạ hình ảnh thành một tập hợp các hệ số mà sau này được lượng tử hóa và mã hóa. Phép biến đổi tốt hơn kết hợp càng nhiều dữ liệu càng tốt thành một số lượng nhỏ các hệ số biến đổi. Sau quá trình này, quá trình lượng tử hóa loại bỏ những hệ số mang ít thông tin nhất.



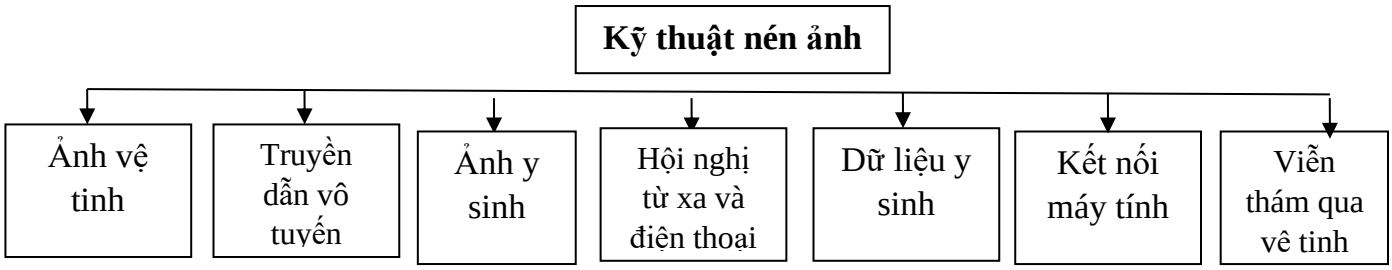
Hình 1.3: Kỹ thuật nén ảnh

Trong phương pháp mã hóa biến đổi, hình ảnh đầu vào $N \times N$ đầu tiên được chia thành một số (khối) $n \times n$ không trùng lặp, sau đó được chuyển đổi để tạo ra $\frac{N}{2} \times \frac{N}{2}$ mảng biến đổi hình ảnh con, mỗi mảng có kích thước $n \times n$, và phép biến đổi được áp dụng riêng cho từng khối này. Ba cơ chế liên quan đến mã hóa biến đổi làm cho phương pháp này trở thành một phương pháp nén cao.

Ba cơ chế này hoạt động như sau: trong giai đoạn đầu tiên, quá trình mã hóa biến đổi một khối dữ liệu chứ không phải là một phần tử duy nhất của ảnh. Trong giai đoạn thứ hai, quá trình lượng tử hóa các hệ số được biến đổi dẫn đến việc loại bỏ mối tương quan được xác định giữa các pixel của mỗi ảnh con. Trong giai đoạn thứ ba, tất cả các hệ số được biến đổi không được lượng tử hóa hoặc không được truyền đến máy thu để tạo ra tốc độ nén cao.

Hệ thống mã hóa biến đổi cũng bao gồm hai phần, đó là bộ mã hóa và bộ giải mã, trong đó bộ mã hóa hoạt động trong bốn giai đoạn là phân rã ảnh con, biến đổi, lượng tử hóa và mã hóa. Chúng được sử dụng để chuyển đổi các giá trị của mức xám trong mỗi khối. Các giá trị lớn hơn có thể chịu trách nhiệm ảnh hưởng đến năng lượng của hệ thống sẽ được lượng tử hóa, trong khi các giá trị khác được đặt bằng 0. Tất cả các quá trình của bộ mã hóa theo thứ tự ngược lại ngoại trừ quá trình lượng tử hóa đều được giải mã.

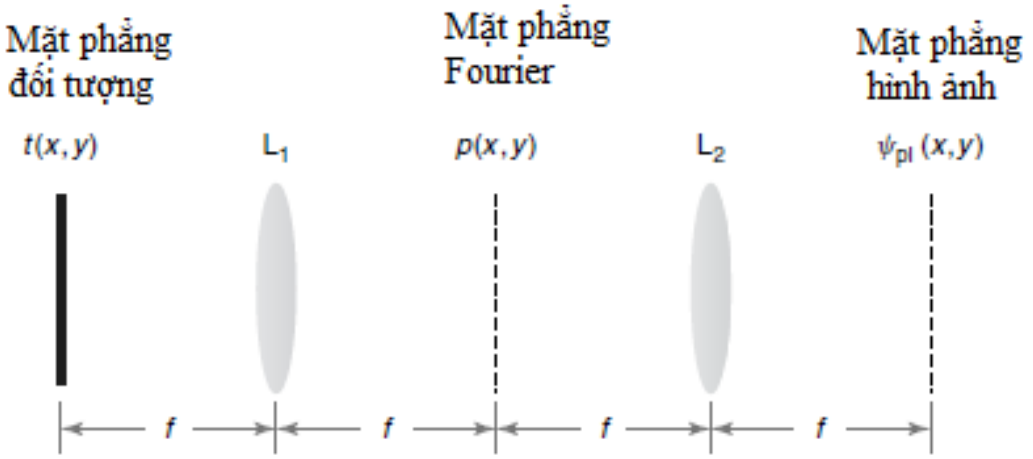
Nén ảnh có nhiều ứng dụng trong thực tiễn như trong thông tin máy tính, xử lý ảnh vệ tinh, ảnh viễn thám với lượng dữ liệu lớn, xử lý các dữ liệu video, các dữ liệu trực tuyến từ xa, các dữ liệu y sinh,... Hình 1.4 chỉ ra một vài ứng dụng của nén ảnh.



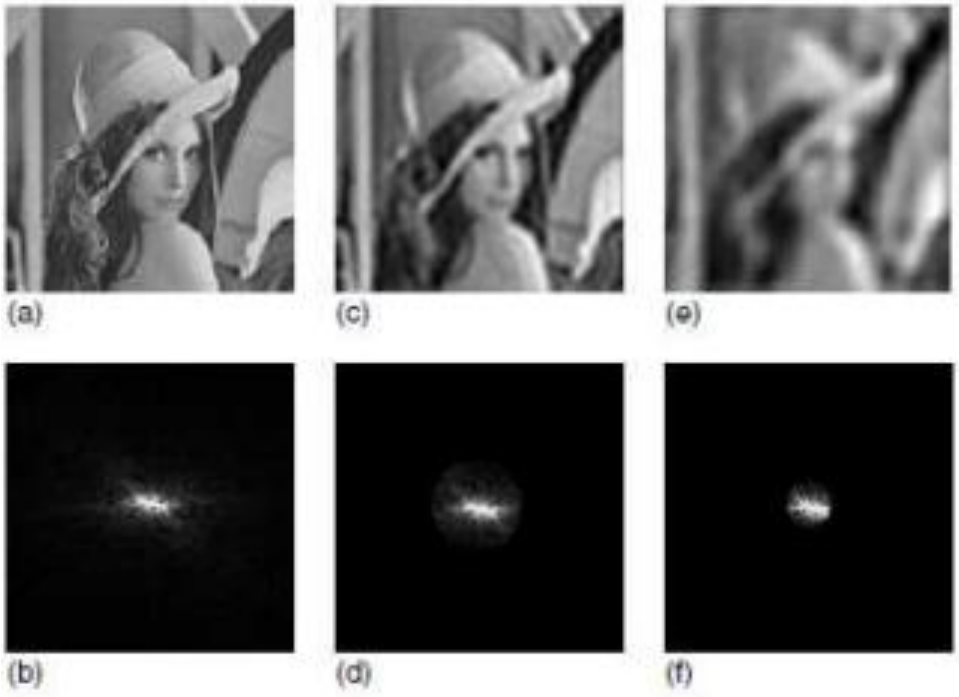
Hình 1.4: Ứng dụng của nén ảnh

Trước đây, việc xử lý ảnh quang thường dùng các hệ thống thấu kính như chỉ ra ở Hình 1.5. Các hệ thống này không thể tích hợp và phát triển thành các cấu trúc máy tính quang trong tương lai do rời rạc, kích thước lớn và không tương thích với các vi mạch tích hợp.

Năm 2013, lần đầu tiên kỹ thuật xử lý ảnh sử dụng biến đổi Haar trên mạch tích hợp được thiết kế thành công [5]. Cấu trúc bộ ghép có hướng được sử dụng để thực hiện các ma trận Haar. Bằng cách kết nối nhiều cấu trúc ghép có hướng với nhau, Haar bậc cao có thể được thực hiện. Tuy nhiên, hệ thống này có nhược điểm là suy hao cao, kích thước lớn và đặc biệt rất khó thực hiện chính xác các hệ số ma trận do các tham số của bộ ghép phụ thuộc nhiều yếu tố như yêu cầu về sự điều khiển chính xác khoảng cách giữa hai ống dẫn sóng [J1], độ nhạy theo bước sóng hoạt động trong dải màu nhìn thấy RGB, các tham số kích thước ống dẫn sóng, các vị trí lấy tín hiệu vào ra của ống dẫn sóng,...



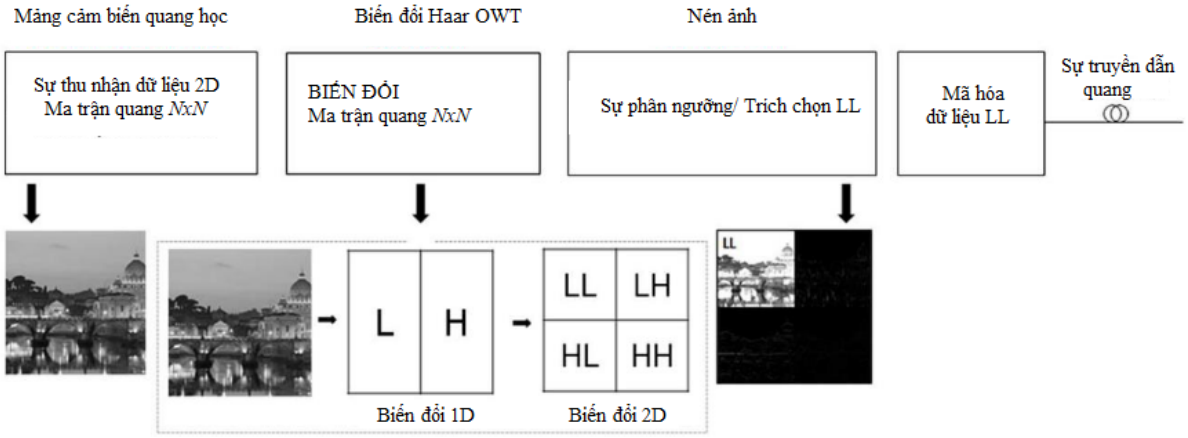
(a)



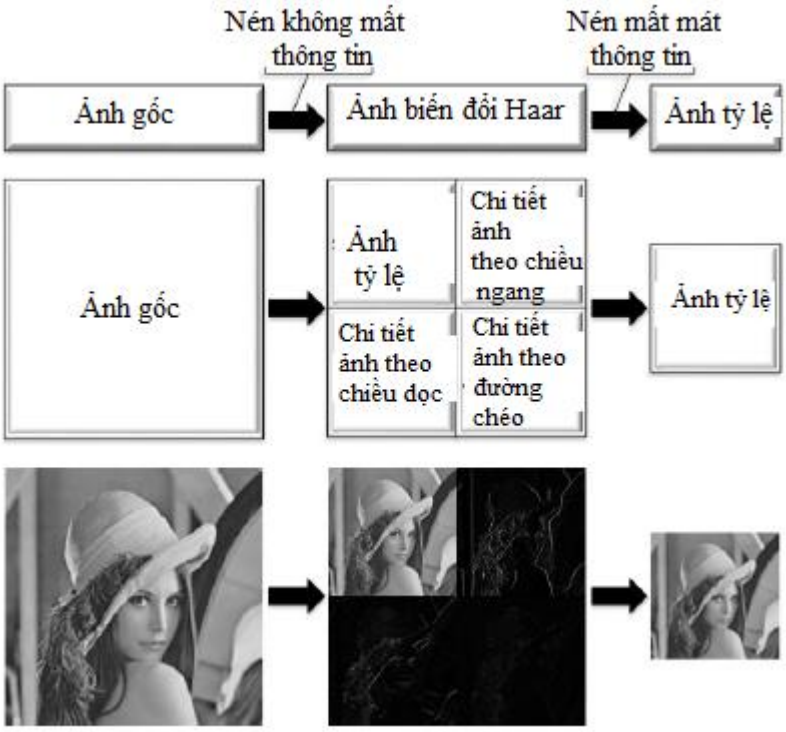
(b)

Hình 1.5: (a) Kỹ thuật xử lý ảnh quang truyền thống, (b) Biến đổi Fourier quang

Hình 1.6 trình bày về quá trình biến đổi nén ảnh sử dụng biến đổi Haar quang, trong đó Hình 1.6 (a) mô tả quá trình tổng quát, Hình 1.6 (b) mô tả chi tiết nén ảnh dùng biến đổi Haar và 1 ví dụ minh họa nén ảnh dùng biến đổi Haar trên ảnh mẫu Lena.



(a)



(b)

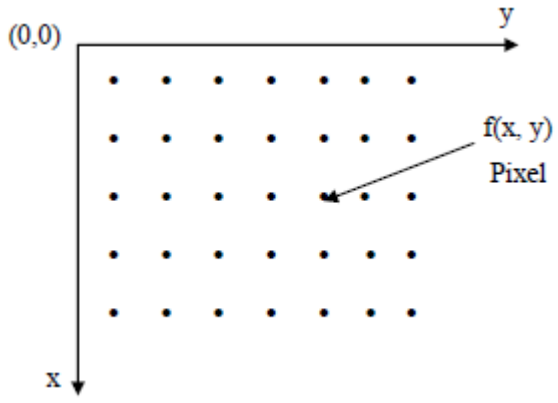
Hình 1.6: (a) Biến đổi Haar quang và (b) nén ảnh dùng biến đổi Haar

1.2 Nén ảnh số dùng biến đổi tín hiệu

Một ảnh số được biểu diễn bằng ma trận các pixel tại vị trí *(x,y)* trong không gian 2 chiều. Có nhiều loại hình ảnh khác nhau tùy thuộc vào số lượng bit dữ liệu khác nhau trên mỗi pixel để biểu diễn chúng. Chất lượng ảnh có thể được đánh giá bằng trực quan hoặc bằng công thức toán học. Một số liệu đánh giá chất lượng khách quan phổ biến cho ảnh thu được sau khi giải nén là PSNR (tỷ lệ nhiễu tín hiệu đỉnh). Nén ảnh mất mát dựa trên chuyển đổi rất linh hoạt vì nó có thể nén ảnh ở các chất lượng

khác nhau tùy thuộc vào ứng dụng của ảnh. JPEG sử dụng DCT 2-D khối 8x8 làm biến đổi. DCT có năng lượng nén rất cao và hiệu suất của nó gần như tương tự như phép biến đổi KLT với ưu điểm là nhân không đổi và tính toán ít phức tạp hơn. Tuy nhiên, đối với việc triển khai phần cứng, loại biến đổi tương tự sẽ ít phức tạp hơn về tính toán và do đó yêu cầu phần cứng ít hơn với hiệu suất gần như tương tự như DCT có thể là lựa chọn ưu tiên.

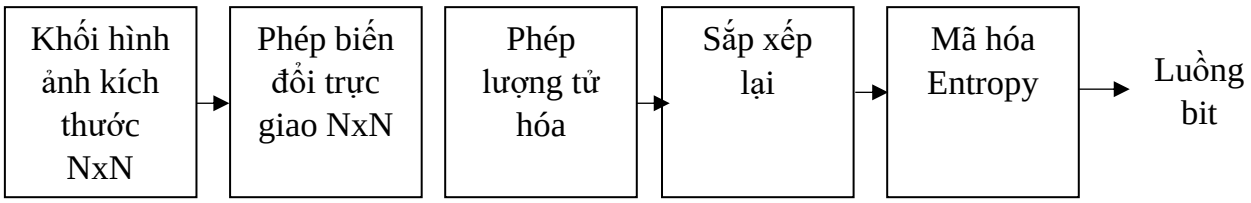
Ảnh của một cảnh tự nhiên có mức độ sáng và cường độ màu khác nhau vô hạn. Ngoài cường độ, chúng là hàm liên tục trong không gian hai chiều. Để xử lý ảnh cho các ứng dụng khác bằng các bộ vi xử lý cùng với việc lưu trữ trong bộ nhớ, dữ liệu ảnh thu được ghi từ cảm biến hình ảnh điện tử (CCD hoặc CMOS) trong máy ảnh kỹ thuật số, máy quét hoặc bất kỳ thiết bị tương tự nào được chuyển đổi thành dạng ảnh số bằng bộ chuyển đổi tương tự sang số (ADC). Các bước lấy mẫu và lượng tử hóa được sử dụng. Tính liên tục trong không gian, bản thân nó được lấy mẫu bởi các điểm cố định có trên cảm biến, được chuyển thành rời rạc. Giờ đây, tín hiệu ảnh liên tục (cảnh tự nhiên) là một hàm số hai chiều, được biểu thị bằng *f(x, y)*, trong đó độ lớn của hàm *f* thể hiện cường độ trong số các mức cường độ hữu hạn tại bất kỳ điểm nào *(x, y)* trong không gian 2 chiều. Tọa độ *(x, y)* là rời rạc như trong Hình 1.7 [51]. Các cường độ tại các điểm khác nhau trong không gian được gọi là phần tử pixel hoặc pixel của ảnh. Một ví dụ về mức cường độ hữu hạn có thể là tất cả các giá trị từ 0 đến 255. Nói chung, bất kỳ ảnh số nào sẽ có số lượng phần tử pixel cố định theo hướng ngang cũng như dọc. Thuật ngữ kích thước của ảnh được sử dụng cho tổng số phần tử pixel trong một ảnh. Nó được biểu diễn bằng MxN, trong đó M là số hàng và N là số cột dữ liệu ảnh.



Hình 1.7: Biểu diễn ảnh số trong không gian 2 chiều

Thường là các pixel lân cận tương quan với nhau và dư thừa trong ảnh. Sự dư thừa chiếm không gian lưu trữ không cần thiết, làm giảm tốc độ truyền và băng thông

của hệ thống. Do đó, mục đích của nén ảnh là giảm độ dư thừa của ảnh. Điều này có thể đạt được bằng kỹ thuật nén ảnh. Các ý tưởng chính đằng sau kỹ thuật nén ảnh là sử dụng phép biến đổi trực tiếp làm cho giá trị pixel nhỏ hơn giá trị ban đầu. Sự biến đổi của ảnh cũng làm cho các hệ số của ma trận được biến đổi không tương quan với mỗi cái khác. Có nhiều phép biến đổi khác nhau đang được sử dụng để nén dữ liệu như DHT, DCT, KLT và biến đổi wavelet DWT. Phương pháp nén tổn hao tạo ra biến dạng không thể phục hồi. Các phương pháp mã hóa biến đổi phổ biến nhất dựa trên biến đổi Fourier (DFT) và cosine rời rạc (DCT) và ánh xạ ảnh thành một tập hợp các hệ số biến đổi sau đó được lượng tử hóa và mã hóa. Mục tiêu của phép biến đổi là sắp xếp lại các pixel của một khối hình ảnh nhất định sao cho hầu hết thông tin được đóng gói thành một số hệ số biến đổi nhỏ nhất. Việc lựa chọn một phép chuyển đổi trong một ứng dụng nhất định phụ thuộc vào số lượng lỗi xây dựng lại có thể được chấp nhận và các tài nguyên tính toán có sẵn. Hình 1.8 mô tả sơ đồ nguyên lý chung của hệ thống nén ảnh dùng biến đổi ảnh.



Hình 1.8: Sơ đồ nén ảnh

Khi sử dụng các phép biến đổi trong nén ảnh, tách tần số mang lại dữ liệu được biến đổi được tạo thành từ các hệ số tần số khác nhau. Ảnh chứa thông tin trực quan cao thường nằm ở miền tần số thấp, trong khi các chi tiết rất nhỏ được thể hiện bằng nội dung tần số cao của ảnh. Trong thực tế, các ứng dụng không cần đến các chi tiết nhỏ (cũng trong nhiều trường hợp, các chi tiết này không quan trọng vì mắt người không nhìn thấy được). Do đó, nếu biết thứ tự tần số rõ ràng, các hệ số tần số cao có thể được bỏ qua (lượng tử hóa bằng 0) trong giai đoạn mã hóa và do đó đạt được sự nén.

Tính trực giao là một đặc tính quan trọng đối với phân tích đa độ phân giải, trong đó tín hiệu ảnh gốc có thể được tách thành các thành phần tần số thấp và cao mà không bị trùng lặp thông tin. Các hàm này chỉ yêu cầu các phép trừ và phép cộng cho các phép biến đổi thuận và nghịch của chúng. Ví dụ về các phép biến đổi này là biến đổi Fourier rời rạc (DFT), biến đổi Cosin rời rạc (DCT) và biến đổi Wavelet rời rạc (DWT) [52]. Một phép biến đổi ảnh lý tưởng phải có hai đặc tính là nén với năng lượng lớn và

độ phức tạp tính toán giảm. Bằng cách nén năng lượng, rất ít hệ số có thể có giá trị cao trong miền biến đổi. Do đó, giá trị hệ số càng nhỏ thì độ nén càng cao. Nén ảnh nhanh được yêu cầu trong nhiều hệ thống nén và biến đổi phức tạp dẫn đến thời gian tính toán cao làm cho quá trình chậm hơn. Ngoài ra, trong trường hợp thực hiện nhanh hơn, phần cứng chuyên dụng sẽ được sử dụng. Hơn nữa, thuật toán phức tạp cao đòi hỏi nhiều diện tích phần cứng hơn, làm cho việc thiết kế bộ mã hóa trở nên tốn kém và cũng tiêu tốn nhiều điện năng hơn. Do đó việc thiết kế được các biến đổi ảnh đơn giản, thực hiện nhanh trong miền quang là hết sức cần thiết.

1.3 Biểu diễn tín hiệu ảnh trong miền quang

Đầu tiên, thu nhận ảnh sử dụng mảng cảm biến quang học để phát hiện ánh sáng và lấy mẫu dữ liệu 2D để có được ma trận dữ liệu đầu vào quang có cùng kích thước $N \times N$ của hình ảnh gốc. Sau đó dữ liệu này được qua bộ biến đổi ảnh trực tiếp trong miền quang để xử lý tín hiệu mà không cần thông qua số hóa. Các ảnh số có các mức xám được mã hóa bằng mức công suất hay cường độ quang. Do vậy các điểm ảnh *(x,y)* trong ma trận ảnh số tương ứng với các mức công suất quang khác nhau.

Đối với ảnh 3 chiều, xử lý tín hiệu quang đã và đang cung cấp các giải pháp liên quan để chuyển đổi dữ liệu thành tín hiệu quang kết hợp được điều chế không gian với các thiết bị SLM [53], cho phép thực hiện hiệu quả ảnh ba chiều kỹ thuật số [54]. Một trong những đặc tính hữu ích nhất của ảnh ba chiều là khả năng kiểm soát pha và biên độ ánh sáng trong trường xa. Biến đổi Fourier mô tả mối quan hệ giữa hình ba chiều (trường gần) và trường phát lại tương ứng của nó (trường xa). Hình ảnh ba chiều có thể tái tạo dạng sóng từ một đối tượng hiện có. Với những tiến bộ kỹ thuật số và xử lý tín hiệu quang học, có thể tính toán số lượng các mẫu giao thoa để tạo ra các mặt trận sóng tổng hợp hoàn toàn có dạng tùy ý. SLM là một thiết bị có thể được sử dụng để điều chế ánh sáng phù hợp với các pixel cố định.

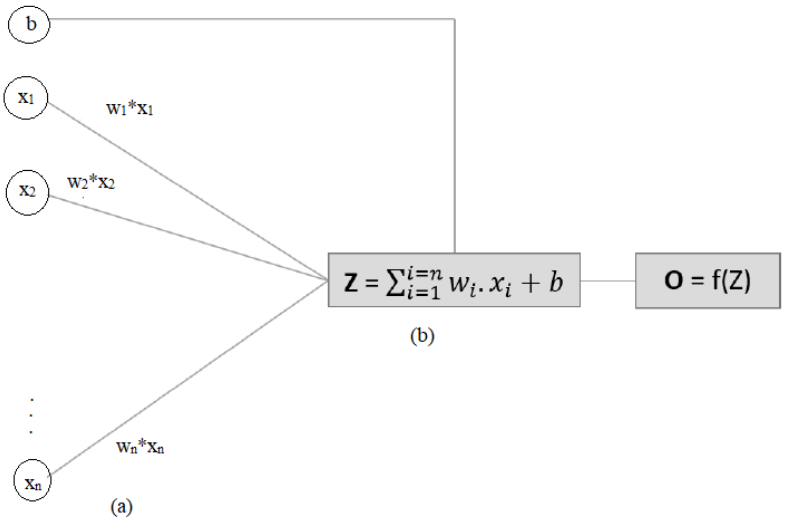
1.4 Mạng nơ – ron

1.4.1 Neuron

Tế bào thần kinh (neuron) là một hệ thống phi tuyến lấy cảm hứng từ sinh học có thể được sử dụng như khối cơ bản cho các mạng nơ-ron và máy học phức tạp hơn [55]. Về mặt toán học, một nơron tương đương với một ánh xạ từ N đầu vào đến một đầu ra, với một đầu vào-đầu ra quan hệ được đưa ra bởi phương trình:

$$O = f(\sum_{i=1}^{N} w_i x_i + b) \tag{1.1}$$

trong đó *f(z)* là hàm kích hoạt nơ-ron, có thể là hàm signmoid, softmax, ReLU hoặc ELU; w_i là hàm trọng số, cần được thiết kế phù hợp để thực hiện xử lý tín hiệu, có thể điều chỉnh được; *b* là hệ số bias. Khả năng điều chỉnh này là một yêu cầu thiết yếu nếu tế bào thần kinh để thực hiện một nhiệm vụ học tập, trong đó nó điều chỉnh chức năng truyền tải của mình theo một tập dữ liệu đào tạo. Chương tiếp theo Luận án sẽ thiết kế hệ thống để thực hiện được ma trận hàm trọng số trong miền quang.

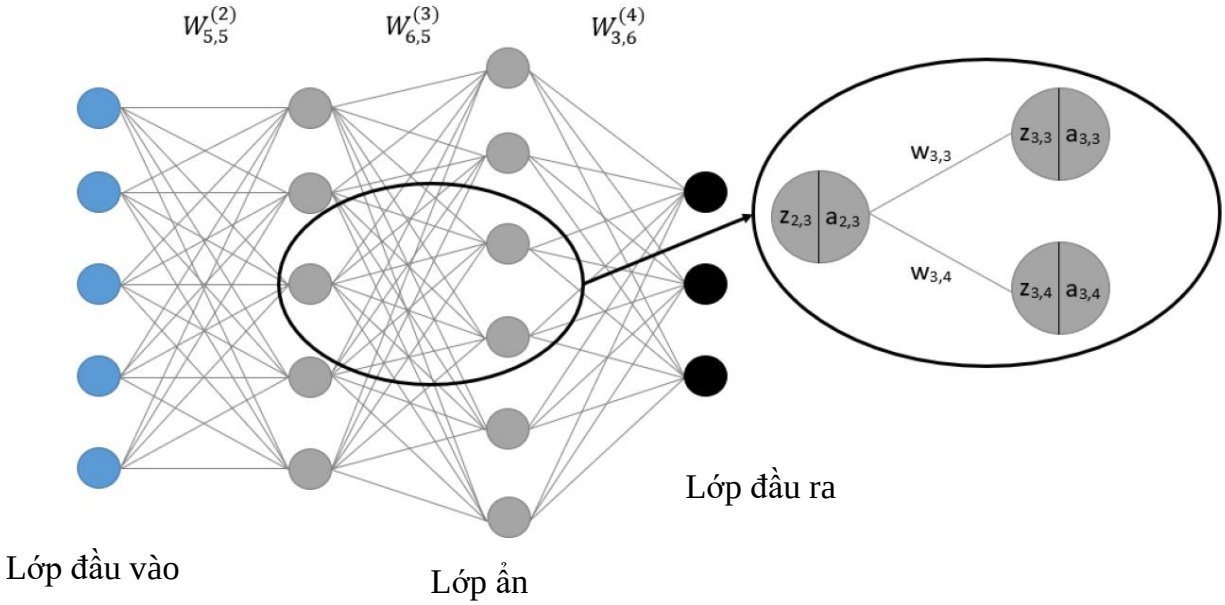


Hình 1.9: (a) Sơ đồ của nơ-ron với tín hiệu vào xi , (b) hàm kích hoạt phi tuyến

Trong trường hợp của nơ-ron, nó biểu diễn sự biến đổi của vectơ đầu vào thành đầu ra bằng cách thực hiện phép nhân vectơ và một hàm kích hoạt như ở Hình 1.9. Nơ-ron cũng cần một thuật toán học. Mỗi khi một mục dữ liệu trải qua quá trình chuyển đổi truyền tiến, nơ-ron sẽ tính toán lỗi đầu ra và cập nhật trọng số của nơ-ron để giảm thiểu lỗi này. Tốc độ cập nhật vectơ trọng lượng phụ thuộc vào một biến, được gọi là tốc độ học. Hàm để cập nhật trọng số *là* $w_i = w_i + \alpha(O_t - O)x_i$. Một nơ-ron thường giải quyết được bài toán tuyến tính, ví dụ để thực hiện các chức năng cổng logic OR, AND hay NAND. Để thực giải quyết được các bài toán phức tạp như nhận dạng thì cần các mạng nơ-ron đa lớp MLP (Multiple Layer Perceptron).

1.4.2 Cấu trúc mạng nơ – ron đa lớp

Nơ-ron nhiều lớp (MLP) là nơ-ron bao gồm ít nhất ba lớp [56]. Một lớp đầu vào, một lớp đầu ra và một hoặc nhiều lớp ẩn (hidden layers). Các lớp này được cấu tạo bởi cái mà chúng ta gọi là tế bào thần kinh nhân tạo hay nói một cách đơn giản hơn là tế bào thần kinh. Hình 1.10 được gọi là mạng nơ-ron được kết nối đầy đủ vì tất cả các nơ-ron của một lớp được kết nối với tất cả các nơ-ron của lớp tiếp theo. Mỗi tế bào thần kinh của một lớp, ngoại trừ các tế bào thần kinh của lớp đầu vào, được xác định bởi trọng lượng và chức năng kích hoạt của nó.



Hình 1.10: Mạng nơ-ron kết nhiều lớp kết nối đầy đủ

Tương tự như trong nơ-ron đơn, sự lan truyền thuận của MLP tương ứng với việc biến đổi vectơ đầu vào X thành vectơ đầu ra Y bằng một chuỗi phép nhân ma trận và hàm kích hoạt liên tiếp. Nếu thuật toán được áp dụng cho một phân loại vấn đề, vectơ đầu ra đại diện cho xác suất của mỗi lớp là đúng. Đây là hai bước xảy ra trong nơ-ron thứ j của lớp thứ i trong quá trình truyền thuận (forward propagation):

• **Bước 1**: tổng trọng số đầu ra của lớp trước $z_{ij} = \sum_{j=1}^{N} w_{ij} x_j$ hoặc $z_{ij} = \sum_{j=1}^{N} w_{ij} a_{i-1,j}$

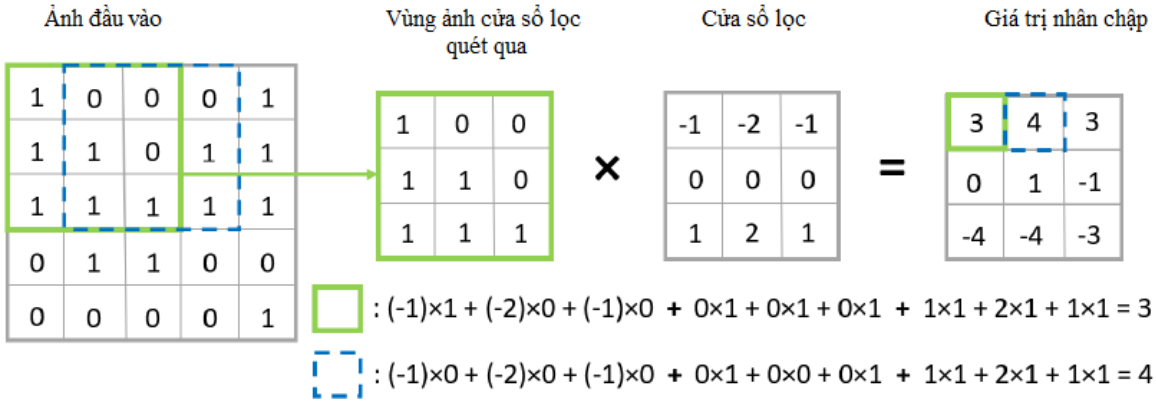
• **Bước 2:** là áp dụng hàm kích hoạt σ_i: $a_{ij} = \sigma_i(z_{ij})$ Trong đó x_j là thành phần j của đầu vào X, w_{ij} là trọng số cho nơ-ron j của lớp (j-1) và $a_{i-1,j}$ là nơ-ron này.

Quá trình hoạt động như vậy được biểu diễn bằng một ma trận, trong đó trọng số w_{ij} có thể được biểu diễn bằng ma trận $M \times N$, trong đó M là kích cỡ của lớp i và N là kích cỡ của lớp j.

Có 2 kiến trúc mạng nơ-ron chính: Mạng nơ-ron nhân chập CNN và mạng rơ-ron hồi quy RNN.

Mạng nơ-ron CNN: chủ yếu dùng để phân tích ảnh. Kiến trúc của chúng bao gồm các lớp được kết nối đầy đủ, giống như lớp được sử dụng trong MLP, được đặt ở cuối mô hình để phân loại tập dữ liệu. Tuy nhiên, CNN được định nghĩa bằng cách bổ sung hai loại lớp khác có khả năng trích xuất các tính năng phù hợp hơn và đồng thời

giảm số lượng đầu vào. Hai loại lớp mới được sử dụng trong CNN là: lớp chập và lớp chèn. Mục tiêu của lớp chập là trích xuất các tính năng mới, được gọi là các tính năng tích hợp, từ một ảnh đầu vào. Ảnh được thể hiện bằng một ma trận các pixel. Một bộ lọc, còn được gọi là nhân, được trượt trên ảnh đầu vào với một khoảng cách là n, nghĩa là bộ lọc được di chuyển n pixel tại một thời điểm. Tại mỗi vị trí, một phép nhân ma trận được thực hiện giữa một tập hợp con của hình ảnh đầu vào và bộ lọc nhân. Các phần tử của ma trận kết quả được tổng hợp và đầu ra được thêm vào ma trận cuối cùng của các đối tượng tích hợp. Hình 1.11 minh họa cách trích xuất các tính năng này. Trong trường hợp cụ thể này, kernel có thể làm nổi bật các cạnh ngang của ảnh. Các giá trị của bộ lọc có thể được xác định trước để thực hiện một tác vụ cụ thể như phát hiện cạnh ngang, hoặc theo cách giống như ma trận trọng số trong mạng CNN cổ điển, có thể học được trong giai đoạn đào tạo.

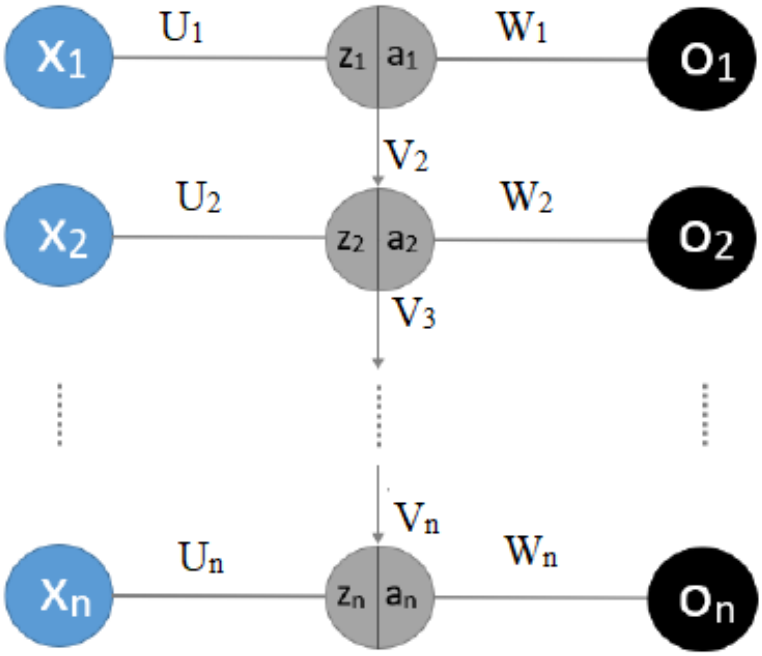


Hình 1.11: Ví dụ về lớp chập dùng ma trận 3×3 tách biên ảnh

Với lớp chèn (gộp), mục tiêu của nó là giảm kích thước của mảng pixel chỉ lưu giữ những thông tin quan trọng nhất. Quá trình này tương tự như hoạt động của tích chập và thực sự có thể được coi là một trường hợp đặc biệt của tích chập. Hai hoạt động phổ biến nhất của gộp chung là gộp tối đa và gộp trung bình. Trong trường hợp đầu tiên, bộ lọc truy xuất pixel có giá trị cao nhất trong tập hợp con của hình ảnh, trong trường hợp thứ hai, bộ lọc tính giá trị trung bình của các pixel trong tập hợp con. Trong mỗi trường hợp, giá trị đầu ra được chuyển thành ma trận mà trở thành hình ảnh đầu vào mới của lớp tiếp theo.

CNN bao gồm một số lớp tích chập xen kẽ, tiếp theo là một hàm kích hoạt và tổng hợp. Khi hình ảnh được coi là đủ nhỏ hoặc trích xuất tính năng có vẻ đủ, hình ảnh được chuyển đổi thành một vectơ đơn giản để cuối cùng được xử lý và phân loại theo các lớp được kết nối đầy đủ.

- Mạng nơ - ron hồi quy RNN: được sử dụng để nghiên cứu, dịch hoặc hiểu các chuỗi logic của từ, số hoặc ký hiệu được in hoặc nghe, chẳng hạn như câu hoặc bản nhạc [57]. Mỗi đầu vào có thể là một từ hoặc một nốt nhạc. Khi dịch một văn bản hoặc một bài phát biểu, cần phải xem xét toàn bộ từng câu và ngữ cảnh của những câu này. Vì vậy, ý tưởng của một RNN là ở mỗi từ của một câu sẽ được tính đến để xử lý câu sau. Hình 1.12 mô tả cấu trúc của một mạng RNN. Giả sử ta muốn thực hiện dịch văn bản gồm n từ, mỗi đầu vào x_i và đầu ra o_i thể hiện dịch của i từ đầu tiên. Đầu vào đầu tiên được đi qua mạng RNN với $\sigma_1 = \sigma_1(U_1 x_1)$, trong đó σ_1 là hàm kích hoạt của lớp 1. Kết quả hoạt động của các lớp tiếp theo là: $z_i = U_i x_i + V_i a_{i-1}$ và $a_i = \sigma_i(W_i z_i)$



Hình 1.12: Sơ đồ mạng RNN

1.5 Mạng nơ – ron quang

Với sự bùng nổ của dữ liệu lớn, việc tạo ra các kiến trúc nhanh và hiệu quả năng lượng cho các mạng trí tuệ nhân tạo trên chip là một thách thức thực sự. Các nền tảng tính toán khác nhau được sử dụng để tích hợp các này và tối ưu hóa tốc độ cũng như mức tiêu thụ điện năng [58, 59]:

- Mạng nơ-ron dựa trên FPGA: FPGA là một thiết bị bán dẫn dựa trên ma trận các Configurable Logic Blocks (CLB), theo đó phần lớn chức năng điện bên trong thiết bị có thể được thay đổi bởi kỹ sư thiết kế. Nghiên cứu cho thấy FPGA có thể đạt được tốc độ và hiệu quả năng lượng tốt hơn gấp 10 lần so với GPU hiện đại.

- Bộ xử lý Tensor (TPU): Là một mạch tích hợp dành riêng cho ứng dụng (ASIC) tùy chỉnh, được phát triển bởi Google. Trong một nghiên cứu, Google phát hiện ra rằng TPU mang lại hiệu suất cao hơn 15–30 lần và hiệu suất trên mỗi watt cao hơn 30–80 so với các CPU và GPU hiện đại.

Quang tử dùng công nghệ CMOS cho phép tích hợp các thiết bị quang tử và điện tử trên cùng một nền tảng. Nó được cho là công nghệ có tiềm năng nhất liên quan đến việc sản xuất các mạch quang tử, hứa hẹn hiệu suất khả quan với chi phí thấp. Một đặc điểm quan trọng của mạng quang tử là quá trình giao tiếp của nó là tính song song, có nghĩa là một số gói thông tin có thể qua mạng đồng thời mà không gây nhiễu lẫn nhau. Trong trường hợp ANN, nó cho phép xử lý tất cả dữ liệu cùng một lúc. Một ưu điểm khác là mạng nơ-ron quang có tốc độ tính toán và hiệu suất năng lượng hơn hẳn các máy tính điện tử. Một phép toán nặng đối với một máy tính thông thường, chẳng hạn như phép nhân ma trận, có thể được tính với tốc độ ánh sáng nếu được thực hiện trong mạng quang tử. Giao thoa kế Mach-Zehnder (MZI) và MRR đều là những công nghệ đầy hứa hẹn và thích hợp để xử lý và tích hợp quang học.

Một trong những thách thức lớn nhất trong việc tái tạo ANN với ONN là có thể triển khai tối ưu từng bước của ANN cổ điển. Việc thực hiện phép nhân ma trận quang đã được đưa ra nhưng chức năng kích hoạt, là một chức năng thiết yếu trong ANN, không được giải quyết đầy đủ trong miền quang. Các đóng góp hiện tại, hoặc là các triển khai ONN trong đó chức năng kích hoạt được thực hiện bằng điện hoặc khi được triển khai về mặt quang học, sử dụng các điểm không tuyến tính. Trong trường hợp đầu tiên, việc chuyển đổi thông tin từ mạch quang được thực hiện với chức năng kích hoạt trên máy tính và sau đó đầu ra được chuyển đổi trở lại mạch quang.

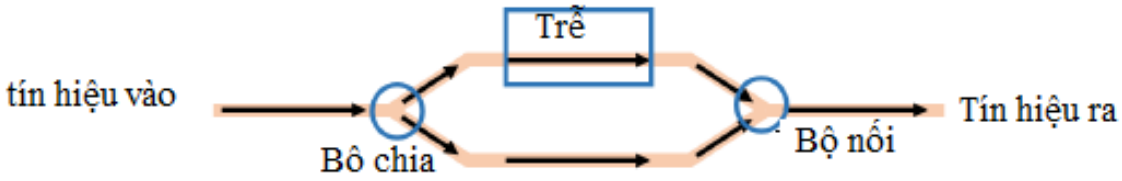
Tuy nhiên, điều này có nghĩa là mạch có thể bị giới hạn bởi tốc độ đồng hồ điện tử. Trong trường hợp thứ hai, những điểm phi tuyến tính phức tạp này không thể so sánh với những điểm không tuyến tính được sử dụng trong lĩnh vực AI. Thật vậy, có một sự khác biệt chung trong việc liên kết các lĩnh vực AI và quang tử. Để vượt qua thách thức này, chúng ta cần giảm khoảng cách giữa những gì được tạo ra bằng quang học và những gì có thể được thực hiện trong ANN. Luận án này nhằm mục đích thu hẹp khoảng cách giữa các lĩnh vực quang tử và AI và xây dựng một mô hình mà từ đó các dự án liên ngành trong tương lai có thể phát triển. Luận án đề xuất một giải pháp để triển khai thiết kế mạch quang học tương tự như các chức năng đã được sử dụng trong ANN.

1.5.1. Thành phần mạng nơ-ron quang

Trong một mạng quang, ánh sáng được sử dụng để truyền thông tin. Thay vì sử dụng dây dẫn điện, như trong các thiết bị điện, ánh sáng truyền qua ống dẫn sóng. Có

hai loại linh kiện chính xây dựng mạch quang gồm linh kiện thụ động và linh kiện chủ động. Các linh kiện thụ động như bộ chia quang, ống dẫn sóng quang. Các mạch chủ động có thể thay đổi được pha của tín hiệu truyền qua như cấu trúc giao thoa MZI, bộ vi cộng hưởng.

• Cấu trúc giao thoa MZI [60]: Sơ đồ của cấu trúc giao thoa MZI được chỉ ra ở Hình 1.13. Trong đó bộ di pha Δφ được điều khiển qua các hiệu ứng điện quang, nhiệt quang, plasma hoặc thay thế hóa trị của vật liệu graphene và phi tuyến [61]. Trong luận án này, tác giả sử dụng điều khiển dùng graphene để tăng cường tốc độ tính toán cho các hệ thống mạng nơ-ron nhân tạo quang.



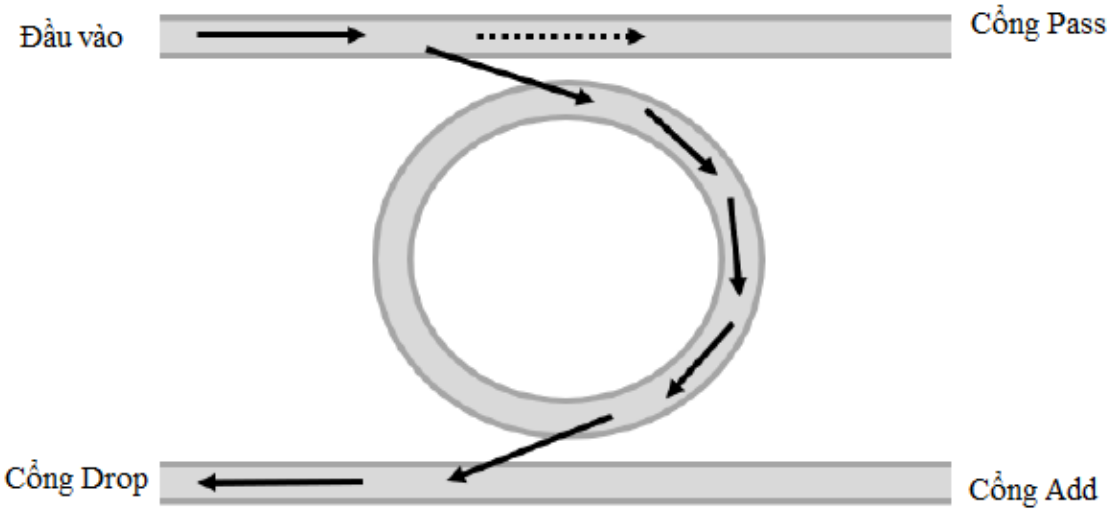
Hình 1.13: Giao thoa MZI

Công suất tín hiệu ra tại cổng 1 và 2 của bộ MZI được tính theo công thức:

$$P_{out1} = sin^2(\frac{\Delta\phi}{2}) \quad (1.2)$$

$$P_{out2} = cos^2(\frac{\Delta\phi}{2}) \quad (1.3)$$

• Cấu trúc vi cộng hưởng MRR [62]: Cấu trúc vi cộng hưởng được dùng như một bộ lọc tại tần số cộng hưởng. Cấu trúc của một bộ MRR chỉ ra ở Hình 1.14. Tín hiệu vào đi qua cấu trúc ra ở cổng đầu ra Pass (hoặc Through) và cổng Drop. Cổng Add được dùng để thêm tín hiệu vào hệ thống. Sơ đồ trên dùng cấu trúc bộ ghép có hướng.



Hình 1.14: Cấu trúc vi cộng hưởng

Cấu trúc giao thoa đa mode MMI: Cấu trúc giao thoa đa mode gồm một ống dẫn sóng hỗ trợ truyền dẫn đa mode kết nối với các cổng đơn mode đầu vào và đầu ra [63, 64]. Bằng thiết kế phù hợp qua chọn vị trí ống dẫn sóng đầu vào, ra và kích thước của MMI, MMI có thể được biểu diễn bằng một ma trận đặc biệt M, thể hiện tín hiệu ra và tín hiệu vào quan hệ qua phương trình ma trận $b = Ma$, trong đó a và b là các tín hiệu tại các đầu vào và ra, M là ma trận đặc tính của MMI, $M = m_{ij} = a_{ij}exp(\phi_{ij})$; a_{ij}, ϕ_{ij} là biên độ và pha của các hệ số ma trận M.

Tín hiệu quang truyền dẫn trong các mạch quang được phân tích giải tích, sau đó dùng phương pháp mô phỏng số để tối ưu hóa. Thông thường phương pháp phân tích truyền mode (MPA- Mode Propagation Analysis) được sử dụng. Lấy ví dụ trong các cấu trúc giao thoa đa mode MMI hoạt động dựa vào nguyên tắc tự tạo ảnh, tức là sau một khoảng cách truyền dẫn nào đó tín hiệu ra sẽ được tái tạo chính xác tín hiệu vào. Xét một ống dẫn sóng phẳng đa mode có chiều rộng $W = W_{MMI}$. Giả sử ống dẫn sóng hỗ trợ M mode $\upsilon = 0,..,M-1$ có các profile $\phi_\upsilon(y)$ và hằng số truyền lan là β_υ. Tín hiệu có profile $\psi(y,0)$ được đưa vào ống dẫn sóng đa mode có thể được phân tích thành tổng các phân bố trường $\phi_\upsilon(y)$ của các mode như sau [90]:

$$\psi(y,0) = \sum_{\upsilon=0}^{M-1} c_\upsilon \phi_\upsilon(y) \tag{1.4}$$

trong đó c_υ là hệ số kích thích trường, được tính theo công thức:

$$c_\upsilon = \frac{\int \psi(y,0)\phi_\upsilon^*(y)dy}{\int \left|\phi_\upsilon(y)\right|^2 dy} \tag{1.5}$$

Do vậy, tại vị trí z=L trong ống dẫn sóng, trường được tính theo công thức:

$$\psi(y, z=L) = \sum_{\upsilon=0}^{M-1} c_\upsilon \phi_\upsilon(y) e^{-j\beta_\upsilon L} \tag{1.6}$$

Đồng thời, áp dụng xấp xỉ Euler, hằng số truyền lan trong ống dẫn sóng đa mode là :

$$\beta_\upsilon \approx k_0 n_f - \frac{(\upsilon+1)^2 \pi\lambda}{4 n_f W_e^2} \quad (1.7)$$

trong đó n_f là chiết suất lõi và W_e là độ rộng hiệu dụng của ống dẫn sóng cho mode cơ bản (bậc nhất). Sự sai khác hằng số truyền lan của mode cơ bản ($\upsilon=0$) và mode υ là:

$$\beta_0 - \beta_\upsilon \approx \frac{\upsilon(\upsilon+2)\pi\lambda}{4 n_f W_e^2} \quad (1.8)$$

Đặt $L_\pi = \frac{\pi}{\beta_0 - \beta_1} \approx \frac{4 n_f W_e^2}{3\lambda}$, L_π gọi là chiều dài phách của hai mode bậc thấp nhất; phương trình **Error! Reference source not found.** được viết lại thành:

$$\beta_0 - \beta_\upsilon \approx \frac{\upsilon(\upsilon+2)\pi}{3 L_\pi} \quad (1.9)$$

Kết quả là, trường trong ống dẫn sóng tại vị trí z=L được tính theo công thức:

$$\Psi(y, z=L) = e^{-j\beta_0 L} \sum_{\upsilon=0}^{M-1} c_\upsilon \phi_\upsilon(y) \exp[j \frac{\upsilon(\upsilon+2)}{3 L_\pi} L] \quad (1.10)$$

Việc mô phỏng linh kiện quang tích hợp là việc giải phương trình Maxwell bằng số. Có hai phương pháp cơ bản để tiếp cận giải phương trình Maxwell là tiếp cận giải trực tiếp trong miền thời gian hoặc thực hiện trong miền tần số dùng biến đổi Fourier. Trong phần này luận án trình bày hai phương pháp được dùng rộng rãi nhất hiện nay để mô phỏng linh kiện quang tích hợp là phương pháp BPM (Beam propagation method), FDTD (Finite difference time domain) và EME (Eigenmode Expansion). Các phương pháp mô phỏng này sử dụng các phần mềm thương mại thiết kế công nghiệp chuyên dụng như Omnisim của Photon Design, OptiFDTD của Optiwave.

1.5.2 Thực hiện mạng nơ – ron quang

Mạng nơ-ron quang có thể thực hiện với hiệu năng rất cao, tốc độ lớn và công suất nhỏ [65]. Bộ xử lý quang có thể thực hiện với tốc độ cao gấp nghìn lần so với các hệ thống máy tính hiện tại với công suất tiêu thụ thấp hơn [66]. Mạng nơ-ron quang đã được thiết kế để thực hiện các thuật toán học sâu gồm 2 phương pháp chính:

- Mạng nơ-ron dựa vào khuếch đại quang bán dẫn SOA (Semiconductor Optical Amplifier) [67]: Mạng này đã được thiết kế để phân loại tập dữ liệu Iris, cấu trúc gồm lớp đầu vào, 2 lớp ẩn và lớp đầu ra để phân loại 150 loại hoa với 4 đặc tính. SOA được dùng để làm bộ kernel các hệ số trọng số. Hàm kích hoạt phi tuyến được thực hiện trong miền điện. Qua thử nghiệm hệ thống này đạt chính xác 95% trên tập dữ liệu Iris.

- Mạch quang đồng bộ: Gần đây năm 2017 các nhà khoa học đã thiết kế thành công cấu trúc mạng quang tích hợp thực hiện học sâu [21]. Phép tích chập trong miền quang được thực hiện thông mạch gồm 53 MZIs và 213 bộ dịch pha. Các nhà nghiên cứu đề xuất sử dụng các lớp graphene để thực hiện các điểm phi tuyến tính. Tuy nhiên, đối với mô phỏng của họ, hàm phi tuyến tính được thực hiện trong miền điện tử. Họ đã áp dụng kiến trúc của mình vào nhiệm vụ nhận dạng nguyên âm. Các kết quả mô phỏng chỉ ra chính xác đạt được 76,7%, so với 91,7% trên máy tính và độ chính xác 95% trên tập dữ liệu MNIST.

1.6 Các tham số hiệu năng

Tỷ lệ nén:

• Nén ảnh làm giảm lượng dữ liệu từ biểu diễn ảnh gốc. Dữ liệu ảnh được nén bằng phương pháp nén không mất dữ liệu có thể được truy xuất trở lại một cách chính xác trong quá trình ngược lại được gọi là giải nén. Phương pháp nén không tổn hao có nhược điểm là hình ảnh có thể được nén theo tỷ lệ nén tối đa khoảng 3 đến 4 (nén rất thấp), trong đó tỷ lệ nén (CR-compressed ratio) được đưa ra bởi:

$$CR=\frac{n_1}{n_2} \tag{1.4}$$

Trong đó n_1 là tổng số bít trong ảnh gốc và n_2 là tổng số bit trong ảnh nén.

• Sai số bình phương trung bình (MSE-Mean square error) [68]: Sai số bình phương trung bình là một cách để đánh giá sự khác biệt giữa giá trị thu được và giá trị thực của pixel. MSE đo mức trung bình của bình phương sai số. MSE giữa hai ảnh *f* và g được xác định bởi công thức tính bằng pixel như sau:

$$MSE=\frac{1}{MxN}\sum\sum[f(i,k)-g(j,k)]^2 \tag{1.5}$$

trong đó (i,k) và (j,k) là điểm ảnh của ảnh f và ảnh g.

• Tỷ số tín hiệu trên tạp âm đỉnh PSNR (Peak Signal to Noise Ratio), đơn vị dB [69]:

Là bình phương của giá trị đỉnh của ảnh (trong trường hợp ảnh 8 bit, giá trị cao nhất là 255) và được xác định bởi công thức:

$$PSNR=10Lg(\frac{255^2}{MSE}) \quad (1.6)$$

1.7 Kết luận Chương 1

Chương 1 trình bày các khái niệm và các vấn đề cơ bản về xử lý tín hiệu số, tập trung vào nén ảnh và thực hiện mạng nơ-ron trong miền toàn quang. Các tham số hiệu năng để đánh giá kỹ thuật nén ảnh như sai số bình phương trung bình, tỷ số tín hiệu trên tạp âm đỉnh, tỷ lệ nén. Đồng thời chương 1 trình bày các nội dung về mạng CNN và RNN để làm cơ sở nghiên cứu cho các nội dung tiếp theo của Luận án.

Chương 2: NÉN ẢNH DỰA VÀO BIẾN ĐỔI TÍN HIỆU TOÀN QUANG

Chương 2 trình bày thiết kế, đánh giá, mô phỏng các bộ biến đổi DHT, DCT, KLT trong miền toàn quang. Kết quả được ứng dụng trong nén ảnh toàn quang. Các bộ biến đổi ảnh có ưu điểm nhỏ gọn, hoạt động tốc độ cao và suy hao thấp, phù hợp với tích hợp với các máy tính trong tương lai.

2.1. Nén ảnh sửa dụng biến đổi Haar (DHT) toàn quang

Những năm gần đây, yêu cầu về nén dữ liệu tốc độ cao ngày càng tăng do lượng thông tin hình ảnh và video tăng lên nhanh chóng. Các hệ thống camera và các mạng cảm biến không dây ở khắp mọi nơi. Việc thực hiện các giải thuật trí tuệ nhân tạo, đặc biệt là các mô hình học sâu phần lớn dựa vào các máy tính điện tử và tính toán trong miền điện [70]. Để tăng tốc độ xử lý, các vi mạch đặc biệt như ASIC, FPGA nhằm thực hiện hỗ trợ tính toán theo cấu trúc máy Von Neumann. Tuy nhiên rất khó để thực hiện được các bài toán lớn. GPU là một trong những giải pháp để thực hiện các thuật toán học sâu. Yêu cầu về công suất rất cao do giới hạn của định luật Moore trong vi mạch, nhiễu giữa các tín hiệu và giới hạn về băng thông. Với yêu cầu tăng lên về tốc độ xử lý, lưu trữ, truyền đi của các dữ liệu ảnh và video, đã tạo ra sự nghẽn trong việc thực hiện các thuật toán AI và học sâu trên các hệ thống máy tính hiện tại.

Do vậy, bằng cách thực hiện dần các khối của hệ thống máy tính CPU như đệm, số hóa, biến đổi tín hiệu, nén dữ liệu sang miền quang có thể giải quyết được vấn đề trên. Đặc biệt các bộ biến đổi tín hiệu thực hiện được trong thời gian thực là một trong những mong muốn lâu nay của lĩnh vực xử lý tín hiệu và xử lý ảnh số. Ưu điểm chính của các hệ thống xử lý toàn quang là tốc độ cao, suy hao thấp, có khả năng tính toán song song và thực hiện được trong thời gian thực. Vì vậy, tính toán trong miền toàn quang được phát triển, trong đó thay vì các bit 0/1 được biểu diễn bằng tín hiệu điện trong các bộ nhớ máy tính, nay biểu diễn trong miền quang [71].

Một trong các đặc điểm của tính toán quang là khả năng thể hiện các mảng 2 chiều của các dữ liệu nhị phân qua trạng thái on/off của ánh sáng. Điều này tạo ra hướng nghiên cứu mới để thay thế các kiến trúc hệ thống máy tính hiện tại bằng các hệ thống máy tính toàn quang. Đặc biệt gần đây các nhà khoa học tại Khoa Công nghệ thông tin của Đại học Ghent và Đại học Valencia đã thiết kế thành công các mạch logic quang khả trình [72], mở ra hướng mới cho lĩnh vực kỹ thuật máy tính.

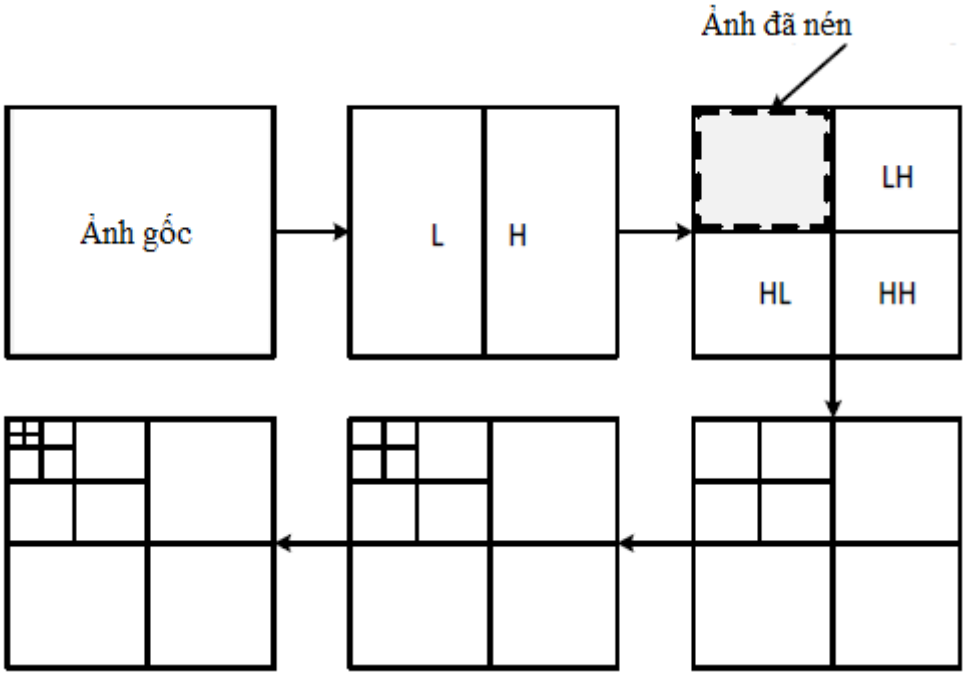
Hiện tại các ảnh được thu nhận qua các cảm biến CMOS và CCD trong miền điện. Tuy nhiên, có hai nhược điểm lớn là tốc độ khung hình chậm chỉ vài MHz do giới hạn về truyền dẫn điện và thời gian xử lý pixel phụ thuộc vào nạp của thiết bị nên tốc độ chậm, tạo hình nhòe và không thể di chuyển tốc độ cao. Việc xử lý trong miền quang giải quyết được 2 vấn đề trên. Trong đó nén dữ liệu và nén ảnh sử dụng biến đổi Haar gần đây đã được nghiên cứu, phát triển trong miền quang [6, 4, 42]. Bộ biến đổi Haar có nhược điểm là sử dụng các cấu trúc rời rạc, sai số chế tạo cho phép phải rất nhỏ để thực hiện được chính xác thuật toán mong muốn [43].

Với tiền đề rằng việc mất độ chính xác là có thể chấp nhận được, máy nén ảnh là một kỹ thuật quan trọng có thể hỗ trợ đáng kể trong việc giảm kích thước tệp và sử dụng băng thông. Kích thước của mảng được chỉ định dưới dạng lũy thừa của hai. Độ phân giải ban đầu của các bức ảnh được biến đổi theo toán học thành công suất lớn hơn của hai bức ảnh tiếp theo và kích thước mảng được khởi tạo tương ứng. Biến đổi Haar chia một hình ảnh thành các thành phần của tần số cao và tần số thấp. Nén ảnh không bị mất và bị mất có thể được thực hiện một cách hiệu quả bằng cách sử dụng nén Haar. Nó phụ thuộc vào việc lấy trung bình và phân biệt các giá trị ma trận ảnh để xây dựng một ma trận thưa hoặc gần như thưa thớt. Ma trận thưa thớt là ma trận trong đó phần lớn các phần tử của nó bằng không. Ma trận thưa thớt có thể được lưu trữ hiệu quả, dẫn đến giảm kích thước tệp.

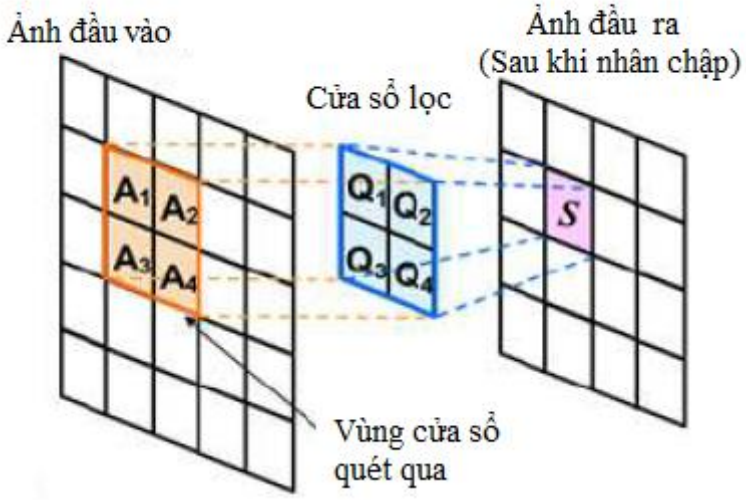
Luận án đề xuất 2 phương pháp thực hiện biến đổi Haar trong miền quang ứng dụng cho nén ảnh. Phương pháp thứ nhất sử dụng cấu trúc giao thoa 4x4 kết hợp với 2×2 MMI và phương pháp thứ hai dùng một bộ 6×6 MMI. Các phương pháp này đều có ưu điểm kích thước nhỏ, suy hao thấp, có khả năng tích hợp trên một nền với các vi mạch hiện thời nhờ sử dụng công nghệ chế tạo vi mạch CMOS.

2.1.1 Nguyên lý nén ảnh sử dụng DHT

Mảng hình ảnh được chia thành hai nửa chứa dữ liệu được biến đổi và các hệ số chi tiết. Hệ số dữ liệu được biến đổi là kết quả của bộ lọc thông thấp trong khi hệ số chi tiết là kết quả của bộ lọc thông cao. Sau khi biến đổi hình ảnh trong hàng, hình ảnh sau đó được chuyển đổi dọc theo cột. Hình 2.1 cho thấy nguyên lý hoạt động của nén ảnh dựa trên biến đổi Wavelet Haar rời rạc (HT). Các ảnh được xử lý theo pixel trong miền quang được chỉ ra ở Hình 2.2 thể hiện xử lý dữ liệu pixel qua biến đổi Haar 2×2, trong đó $S = A_1Q_1 + A_2Q_2 + A_3Q_3 + A_4Q_4$. Trong đó Ai là các giá trị pixel của ảnh vào, Qi là giá trị mặt nạ.



Hình 2.1: Nguyên lý nén ảnh dùng DHT



Hình 2.2: Xử lý dữ liệu pixel qua biến đổi Haar

Luận án thiết kế các bộ biến đổi Haar trên vật liệu Si_3N_4 hoạt động ở bước sóng đỏ (632nm), xanh lam (405nm) và xanh lục (532nm). Cấu trúc này hữu ích cho việc xử lý hình ảnh tốc độ cao và nén dữ liệu lớn. Phương pháp mới được đề xuất có ưu điểm là tốc độ cao, tổn hao thấp và tương thích với công nghệ CMOS.

DHT phân dải tín hiệu rời rạc thành hai tín hiệu con có độ dài bằng nửa độ dài tín hiệu ban đầu. Một tín hiệu con là đường trung bình hoặc xu hướng và tín hiệu còn lại

là sự chênh lệch hoặc dao động đang chạy. Biến đổi Haar hữu ích trong các ứng dụng yêu cầu thực hiện phát hiện cạnh hoặc trích đường viền theo thời gian thực [9]. Trong luận án này, tác giả đề xuất một phương pháp tổng hợp để thực hiện các phép biến đổi Haar. Phương pháp này phù hợp để thực hiện đường ống và song song.

Biến đổi wavelet của tín hiệu liên tục *f(t)* được biểu diễn theo công thức:

$$CWT_f(\tau, a) = \frac{1}{\sqrt{a}} \int f(t) h^*(\frac{\tau - a}{a}) dt \tag{2.1}$$

Trong đó, $h^*(\frac{\tau-a}{a})$ là wavelet mẹ với dịch hệ số a và tỷ lệ τ. Để rời rạc hóa biến đổi wavelet thành nhị phân, ta có $a = 2^j$, khi đó wavelet mẹ được tính theo công thức: $h_{j,k} = 2^{-\frac{j}{2}} h(2^{-j}t - k)$. Biến đổi wavelet của tín hiệu rời rạc f(n) khi đó trở thành:

$$DWT_f(j, k) = \sum f(n) h_{j,k}(n) \tag{2.2}$$

Biến đổi Haar phân dải tín hiệu rời rạc thành hai tín hiệu con có chiều dài N như sau: $a^1 = (a_1, a_2, a_3, \dots, a_{N/2})$, $d^1 = (d_1, d_2, d_3, \dots, d_{N/2})$. Trong đó: $a_m = \frac{f_{2m-1} + f_{2m}}{\sqrt{2}}$ và $d_m = \frac{f_{2m-1} - f_{2m}}{\sqrt{2}}$ với *m=1,2, …, N/2*.

Biến đổi Haar bậc 1 là ánh xạ $f^{H_1} \mapsto H_1 = (a^1, d^1)$ được biểu diễn dưới dạng ma trận kernel:

$$H_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{2.3}$$

2.1.2 Biến đổi Haar dùng 4x4 MMI và 2x2 MMI

Trong nghiên cứu này, tác giả thiết kế các bộ biến đổi Haar dùng cấu trúc giao thoa đa mode MMI. Cấu trúc được đề xuất mới thể hiện ở Hình 2.3, trong đó các vị trí ống dẫn sóng đầu vào và ra được đặt tại $x_i = (i + 0.5)\frac{W_{MMI}}{N}$. Bằng cách thiết kế, chọn chiều dài MMI thích hợp tại $L_{MMI} = 1.5L_\pi$, trong đó L_π là chiều dài phách [63], cấu trúc trên được tính toán bằng ma trận:

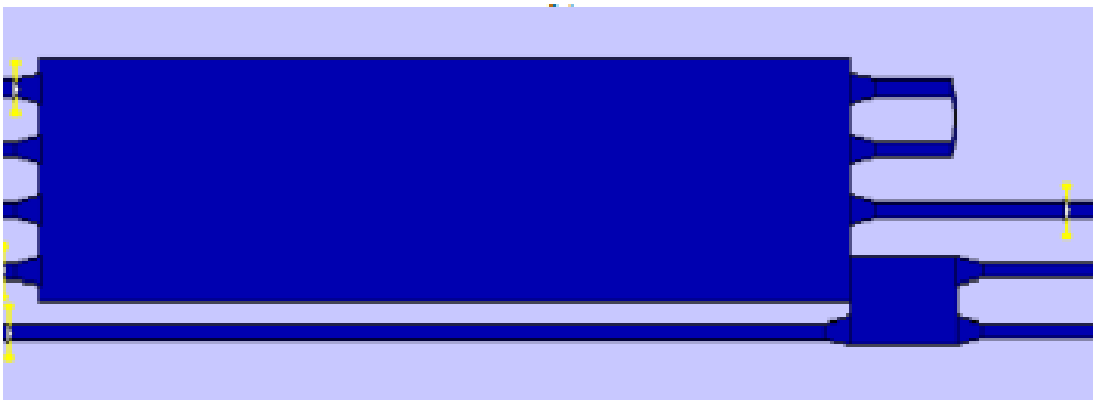
$$M = \begin{bmatrix} 1-j & 0 & 0 & 1+j \\ 0 & 1-j & 1+j & 0 \\ 0 & 1+j & 1-j & 0 \\ 1+j & 0 & 0 & 1-j \end{bmatrix} \tag{2.4}$$

Biến đổi Haar bậc 2 được tính theo ma trận từ phương trình (2.4):

$$H_2 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 0 & -1 \end{bmatrix} \tag{2.5}$$

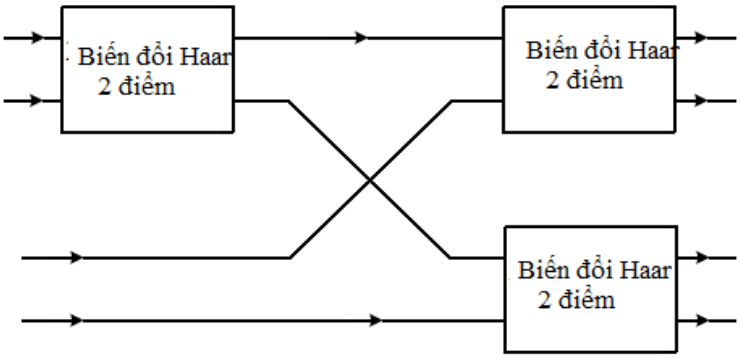
Bằng cách sử dụng các bộ di pha tại các đầu vào và ra $\pm\frac{\pi}{2}$, ma trận H_2 được tạo thành từ việc kết kết hợp 4×4 và 2×2 MMI như Hình 2.3. Ma trận của 4×4 MMI có thể được viết lại dưới dạng:

$$M = \frac{1}{\sqrt{2}}\exp\left(-j\frac{\pi}{4}\right)\begin{bmatrix} 1 & 0 & 0 & \exp\left(j\frac{\pi}{2}\right) \\ 0 & 1 & \exp\left(j\frac{\pi}{2}\right) & 0 \\ 0 & \exp\left(j\frac{\pi}{2}\right) & 1 & 0 \\ \exp\left(j\frac{\pi}{2}\right) & 0 & 0 & 1 \end{bmatrix} \tag{2.6}$$



Hình 2.3: Biến đổi Haar dùng 2×2 và 4×4 MMI

Kết quả là thu được Haar 4 điểm như trong Hình 2.4. Cấu trúc Haar 4 điểm được thiết kế từ Haar 2 điểm bằng cách kết nối như sơ đồ. Trong luận án này, tác giả đề xuất một cấu trúc mới chỉ dùng 4×4 và 2×2 MMI kết nối không cần ghép các ống dẫn sóng vào ra, tạo thành hệ thống có thể tích hợp trên cùng một chip đơn. Thiết kế sử dụng cấu trúc dẫn tín hiệu quang Si_3N_4 có thể hoạt động được ở các vùng bước sóng nhìn thấy ứng với các màu RGB.

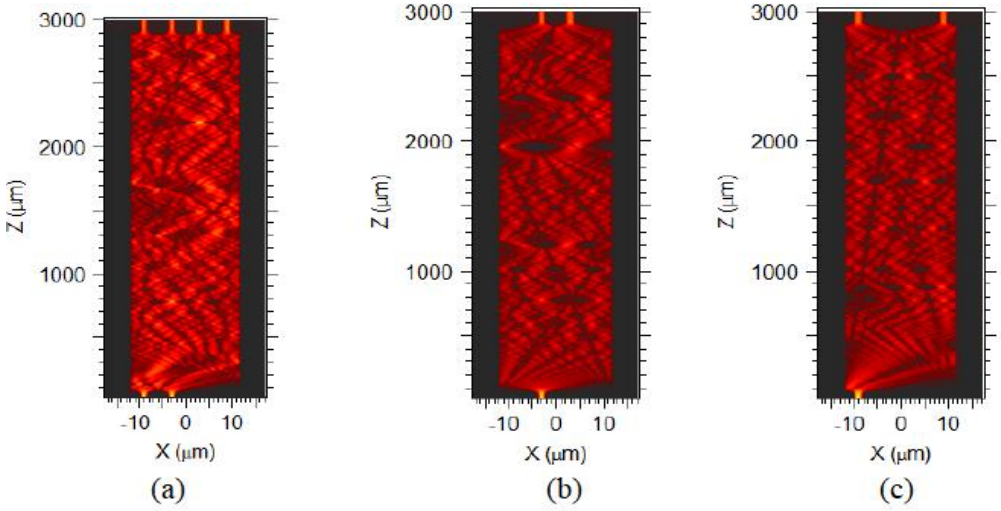


Hình 2.4: Biến đổi Haar 4 điểm từ Haar 2 điểm

Cấu trúc ống dẫn sóng được mô tả ở Hình 2.5, trong đó kích thước ống dẫn sóng là 1600nm chiều rộng và 170nm chiều cao cho các cổng tín hiệu vào và ra. Đối với 4×4 MMI, Luận án chọn chiều rộng là 24µm để hỗ trợ được 4 cổng ra và kết nối nối tiếp được với cấu trúc 2×2 MMI đi sau đó như ở Hình 2.3.

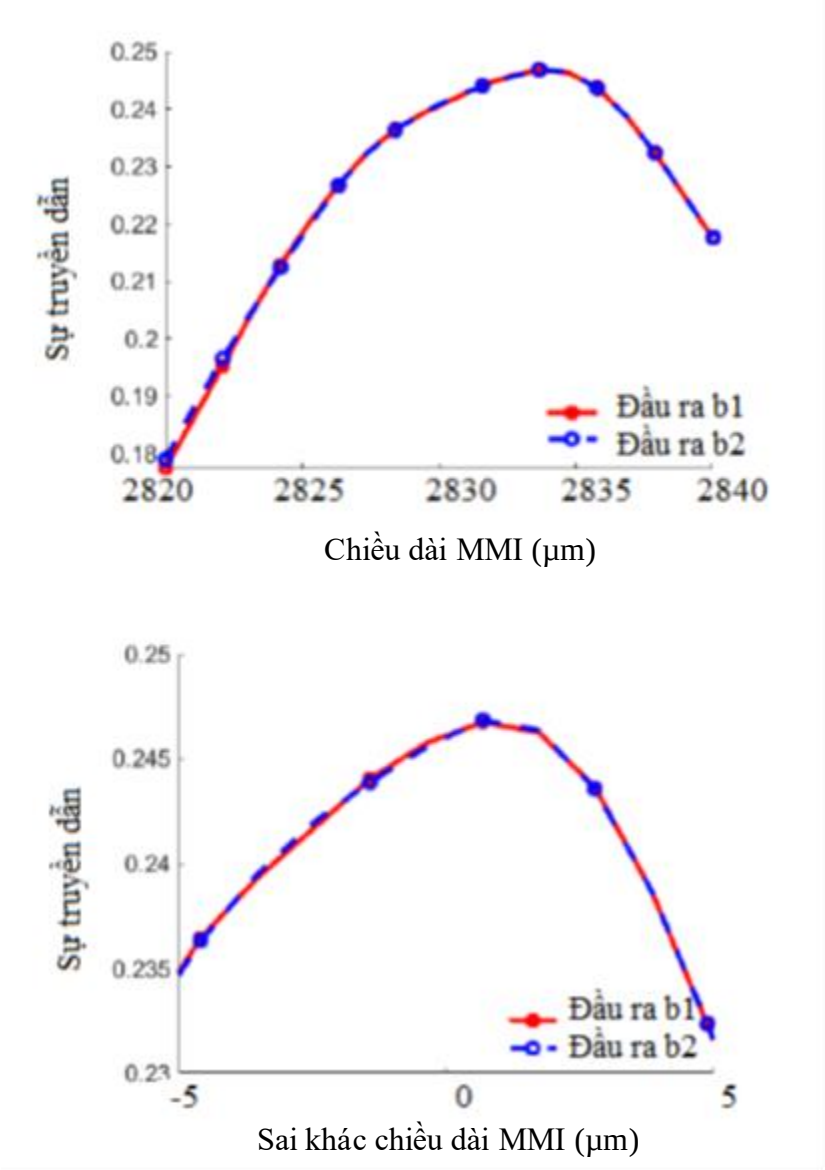
Hình 2.5: Cấu trúc ống dẫn sóng

Kết quả mô phỏng tín hiệu quang tương ứng với giá trị mức xám của ảnh vào cổng 1, 2 và cả cổng 1 và 2 được chỉ ra ở Hình 2.6. Kết quả mô phỏng cho thấy cấu trúc 4×4 MMI đã thực hiện được theo ma trận thiết kế ở trên để có thể thực hiện được Haar 4×4.



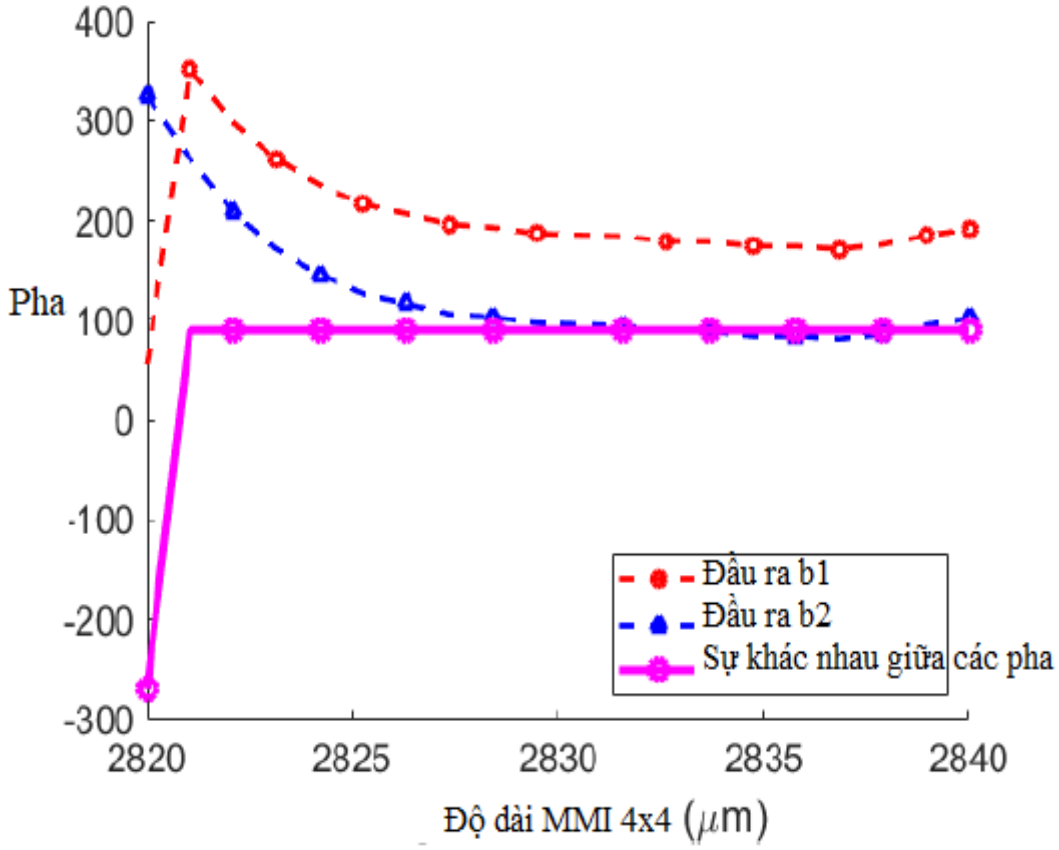
Hình 2.6: Kết quả mô phỏng tín hiệu vào tại cổng (a) 1, 2, (b) 2 và (c) 1

Việc thực hiện được chính xác ma trận Haar là rất quan trọng. Vị trí ống dẫn sóng tín hiệu đầu vào và ra có ảnh hưởng lên độ chính xác này. Công nghệ CMOS thường có dung sai chế tạo cỡ ±5nm [73] cho quang tử silic. Nếu biên độ và pha của tín hiệu ảnh ra sau cấu trúc sai số lớn trong khoảng dung sai này thì bộ biến đổi không khả thi trong thực tế. Do vậy Luận án đã nghiên cứu dung sai chế tạo dùng công nghệ CMOS cho bộ biến đổi Haar. Kết quả mô phỏng cho sự thay đổi của công suất ánh sáng ra tại các chiều dài MMI khác nhau được chỉ ra ở Hình 2.7. Kết quả mô phỏng cho thấy trong khoảng ±2µm công suất ảnh đầu ra chỉ thay đổi 1%. Điều này cho phép cấu trúc được thiết kế thực hiện bộ biến đổi Haar rất chính xác với công nghệ hiện nay.



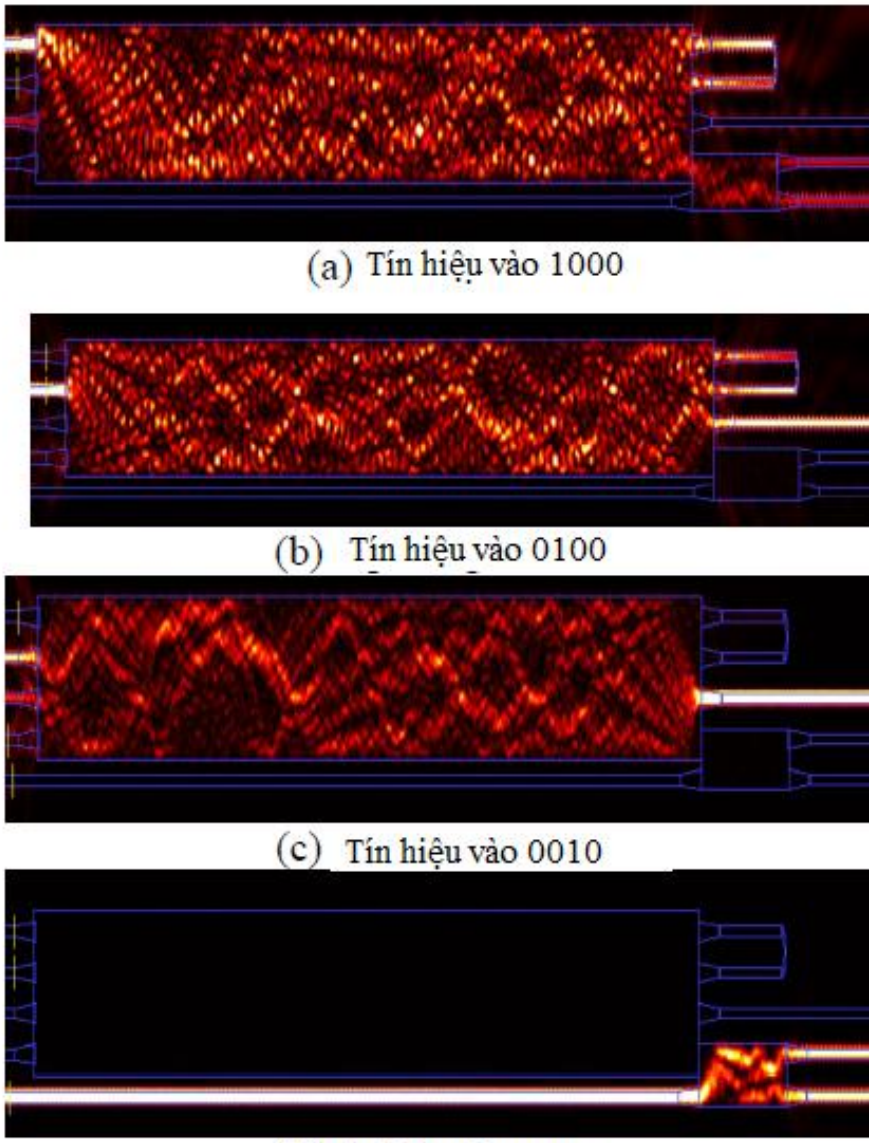
Hình 2.7: Cường độ mức pixel ra tại cổng 1, 2 với chiều dài MMI khác nhau

Tiếp theo, pha của tín hiệu ra được phân tích. Kết quả mô phỏng pha của tín hiệu tại các cổng ra 1 và 4 khi tín hiệu ảnh vào cổng 1 được chỉ ra ở Hình 2.8. Trên hình cũng chỉ ra sai pha giữa 2 cổng. Kết quả cho thấy sai pha là 90^0 trong 1 dải từ 2825 đến 2840μm, cho phép thực hiện bộ biến đổi Haar toàn quang rất chính xác.



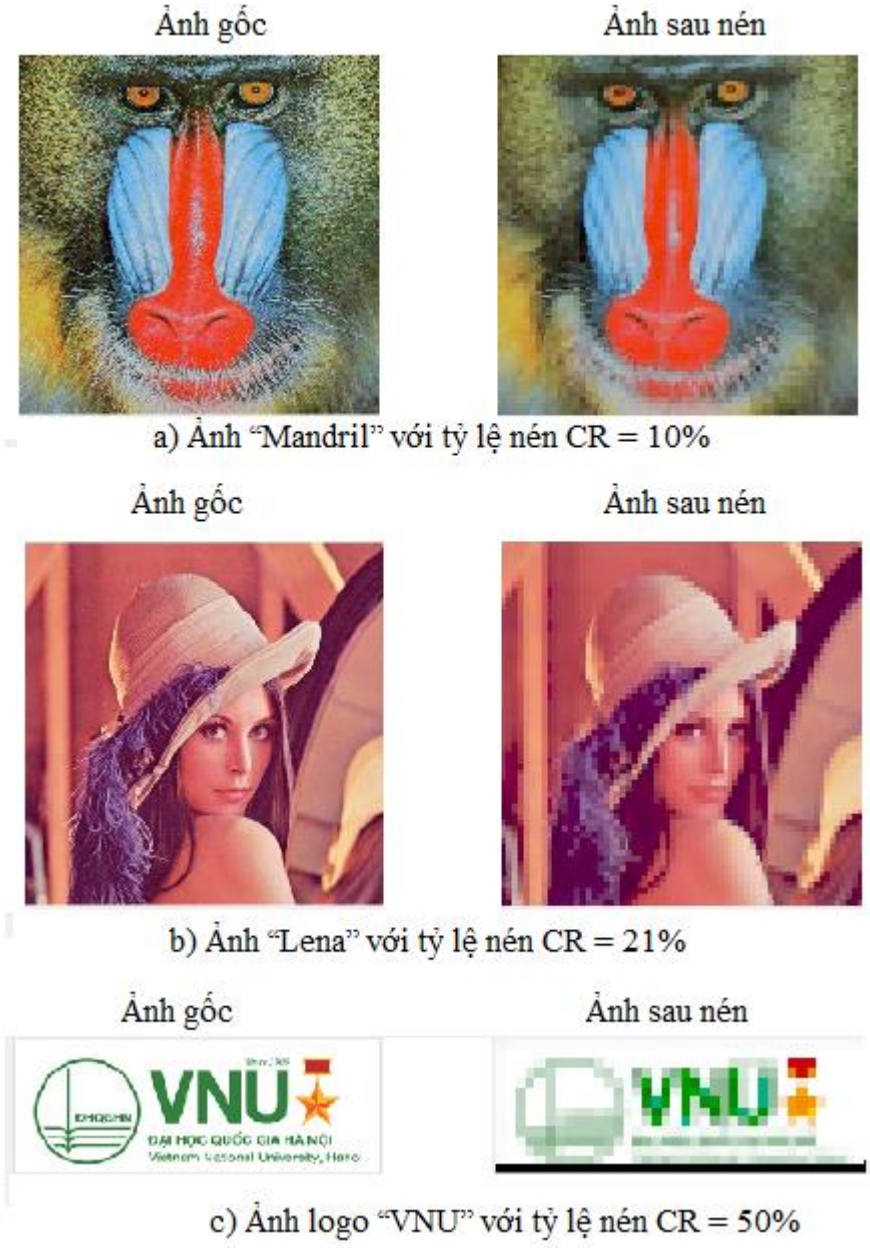
Hình 2.8: Pha tín hiệu tại cổng 1 và 4 với chiều dài MMI khác nhau

Kết quả xử lý tín hiệu ảnh truyền qua bộ biến đổi Haar khi tín hiệu vào các cổng 1, 2, 3, 4 tương ứng được chỉ ra ở Hình 2.9. Kết quả này phù hợp với lý thuyết đã phân tích ở trên. Kết quả mô phỏng số cho thấy suy hao của toàn bộ cấu trúc rất thấp khoảng 0.95dB.



Hình 2.9: Tín hiệu ảnh truyền qua cấu trúc Haar 4×4 tại các đầu vào khác nhau

Tiếp theo Luận án mã hóa ma trận Haar được thiết kế từ kết quả trên để mô phỏng ở mức hệ thống. Các tín hiệu ảnh được đọc dưới dạng ma trận các mức cường độ. Kết quả mô phỏng nén ảnh đầu vào với 3 ảnh khác nhau được chỉ ra ở Hình 2.10. Kết quả cho thấy ma trận Haar toàn quang đã thực hiện thành công việc nén ảnh với các kết quả như Bảng 2.1.



Hình 2.10: Ảnh gốc và ảnh nén sau bộ biến đổi Haar 4×4 MMI toàn quang

Bảng 2.1: Kết quả MSE và PSNR của ảnh gốc và ảnh nén dùng Haar 4×4 MMI

Ảnh	**Kích thước ảnh gốc**	**Kích thước ảnh nén**	**Tỷ lệ nén CR**	**MSE**	**PSNR dB**
Madril	787KB	875KB	10%	0,37	131
Lena	787KB	378KB	21%	0,23	140
Logo VNU	19KB	10KB	50%	34	40

2.1.3 Biến đổi Haar dùng 6x6 MMI

Trong phần này Luận án đề xuất bộ biến đổi Haar sử dụng duy nhất một cấu trúc giao thoa đa mode 6×6, với 6 đầu vào và 6 cổng ra. Bằng cách lựa chọn vị trí cổng đầu vào và cổng đầu ra thích hợp tại $x_i = (i + 0.5)\frac{W_{MMI}}{6}$, chiều dài của 6x6 MMI là $L_{MMI} = 1.5L_\pi$.

Hình 2.11: Bộ biến đổi Haar dùng duy nhất 6×6 MMI

Kết quả tính toán cho thấy, ma trận của 6×6 MMI trong trường hợp này được tính theo công thức:

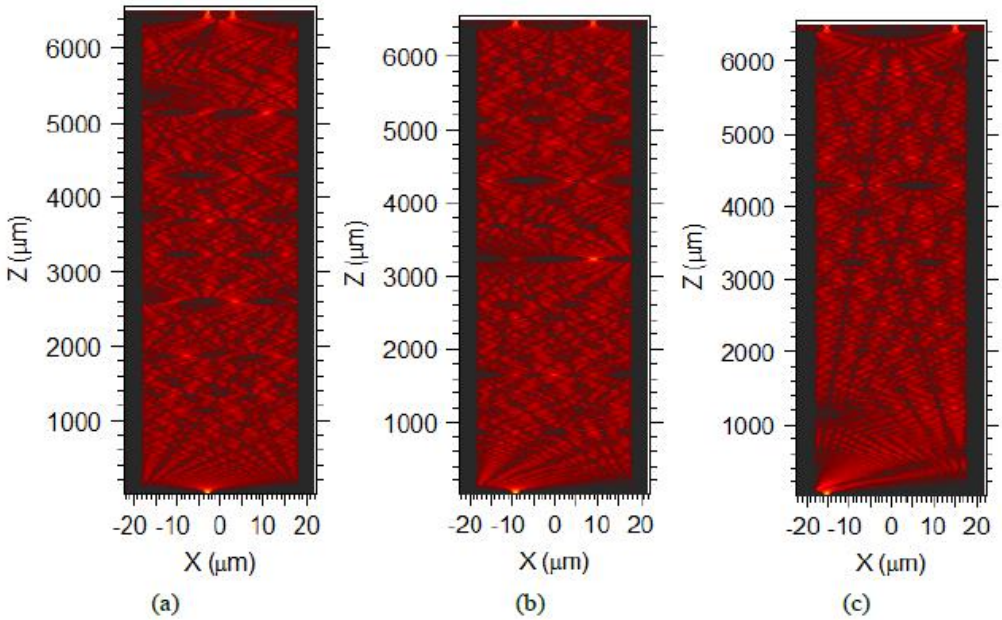
$$S = \frac{1}{\sqrt{2}}\begin{bmatrix} \exp(j\frac{\pi}{4}) & 0 & 0 & 0 & 0 & \exp(j\frac{\pi}{4}) \\ 0 & \exp(j\frac{\pi}{4}) & 0 & 0 & \exp(j\frac{\pi}{4}) & 0 \\ 0 & 0 & \exp(j\frac{\pi}{4}) & \exp(j\frac{\pi}{4}) & 0 & 0 \\ 0 & 0 & \exp(j\frac{\pi}{4}) & \exp(j\frac{\pi}{4}) & 0 & 0 \\ 0 & \exp(j\frac{\pi}{4}) & 0 & 0 & 0 & 0 \\ \exp(j\frac{\pi}{4}) & 0 & 0 & 0 & 0 & \exp(j\frac{\pi}{4}) \end{bmatrix} \quad (2.7)$$

Bằng cách sử dụng các ống dẫn sóng phản hồi như ở Hình 2.11, cổng ra 1 nối với cổng ra 4, cổng ra 2 nối với cổng ra 3, ta có được bộ biến đổi Haar 4 điểm, trong đó ma trận Haar lúc này được viết lại thành:

$$H_4 = \begin{bmatrix} \pm a_{11}e^{\pm j\delta} & \pm a_{12}e^{\pm j\delta} & \pm a_{13}e^{\pm j\delta} & 0 \\ \pm a_{21}e^{\pm j\delta} & \pm a_{22}e^{\pm j\delta} & \pm a_{23}e^{\pm j\delta} & 0 \\ \pm a_{31}e^{\pm j\delta} & \pm a_{32}e^{\pm j\delta} & \pm a_{33}e^{\pm j\delta} & \pm a_{34}e^{\pm j\delta} \\ \pm a_{41}e^{\pm j\delta} & \pm a_{42}e^{\pm j\delta} & 0 & \pm a_{44}e^{\pm j\delta} \end{bmatrix} \quad (2.8)$$

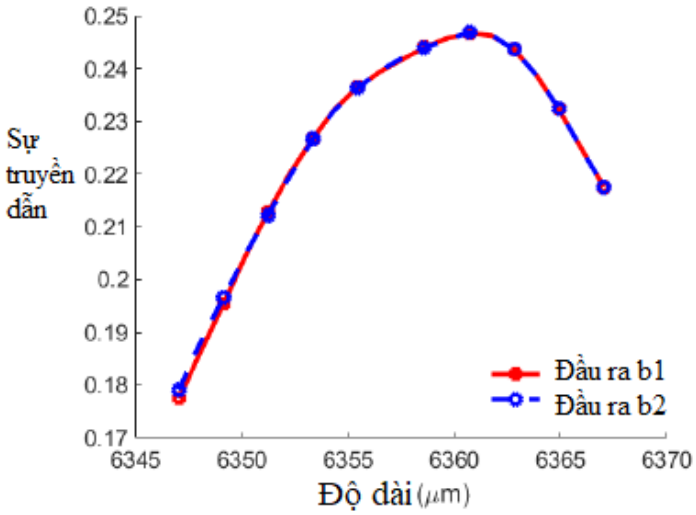
Trong đó các giá trị a_{ij} và δ là sai số xung quanh giá trị 1 và 90 độ. Qua mô phỏng, thiết kế cấu trúc 6×6 ở phần sau cho thấy các giá trị sai số này trong khoảng chấp nhận được và rất gần với các giá trị trung tâm +1 và 90^0. Do vậy thực hiện biến đổi Haar trong miền quang dùng cấu trúc này có độ chính xác cao, đặc biệt phù hợp với nén ảnh có tổn hao.

Tiếp theo kết quả xử lý tín hiệu quang truyền qua cấu trúc 6×6 MMI được mô phỏng ở Hình 2.12. Sử dụng phương pháp mô phỏng số, chiều dài tối ưu của 6×6 MMI được tính toán tại 6360 µm với chiều rộng của MMI là 36 µm.



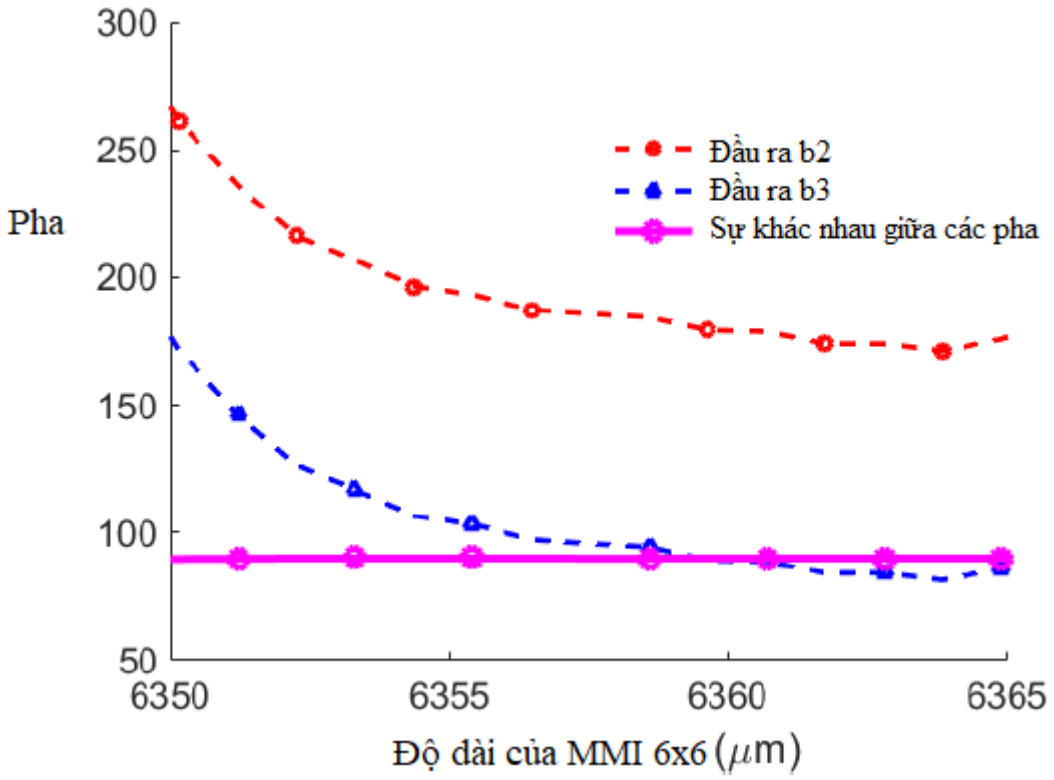
Hình 2.12: Tín hiệu ảnh truyền qua 6×6 MMI tại các đầu vào khác nhau

Cường độ mức pixel đầu ra xung quanh chiều dài tối ưu 6360 µm được chỉ ra ở Hình 2.13. Kết quả mô phỏng cho thấy trong khoảng ±2µm công suất ảnh đầu ra chỉ thay đổi 1% như cấu trúc 4×4 MMI. Điều này cho phép cấu trúc được thiết kế thực hiện bộ biến đổi Haar 6×6 MMI cũng rất chính xác với công nghệ hiện nay.



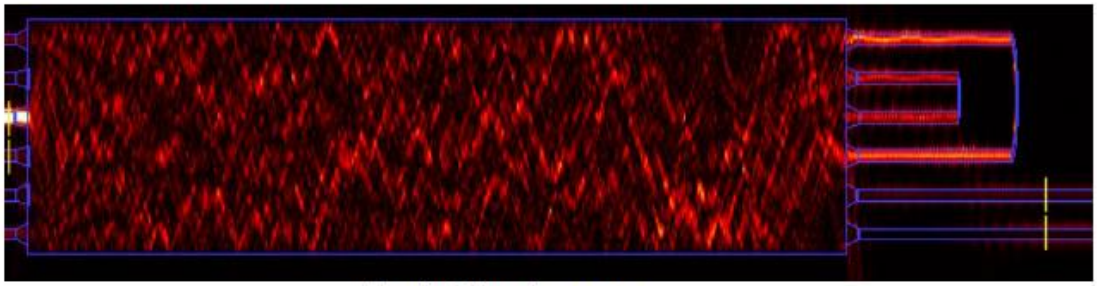
Hình 2.13: Cường độ mức pixel ra tại cổng 1 với chiều dài 6×6 MMI khác nhau

Tiếp theo, pha của tín hiệu ra được phân tích. Kết quả mô phỏng pha của tín hiệu tại các cổng ra 1 và 4 khi tín hiệu ảnh vào cổng 1 được chỉ ra ở Hình 2.14. Trên hình cũng chỉ ra sai pha giữa 2 cổng. Kết quả cho thấy sai pha là 90^0 trong 1 dải từ 2825 đến 2840µm, cho phép thực hiện bộ biến đổi Haar toàn quang rất chính xác.

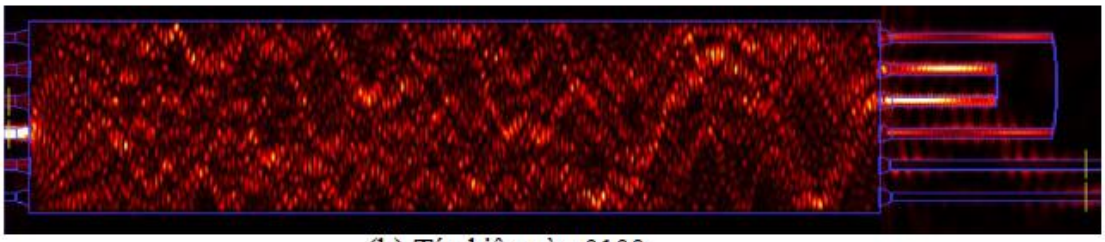


Hình 2.14: Pha (độ) tín hiệu tại cổng 1 và 4 với chiều dài 6×6 MMI khác nhau

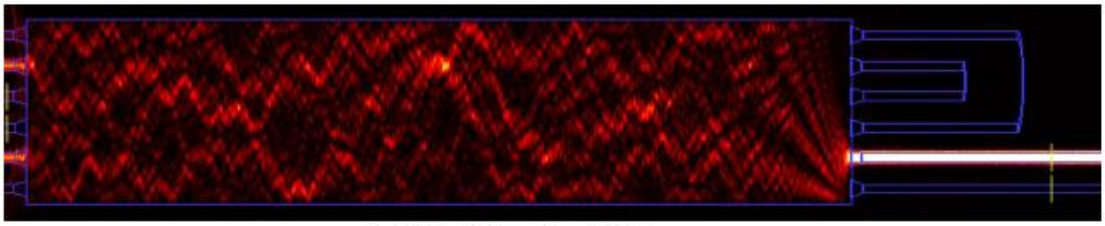
Kết quả xử lý tín hiệu ảnh truyền qua bộ biến đổi Haar khi tín hiệu vào các cổng 1, 2, 3, 4 tương ứng được chỉ ra ở Hình 2.15. Kết quả này phù hợp với lý thuyết đã phân tích ở trên.



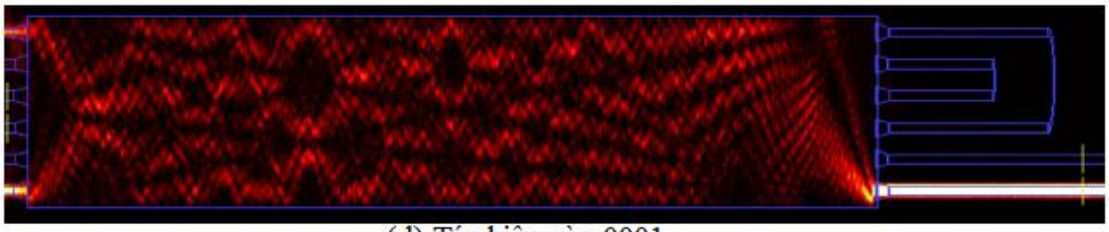
(a) Tín hiệu vào 1000



(b) Tín hiệu vào 0100



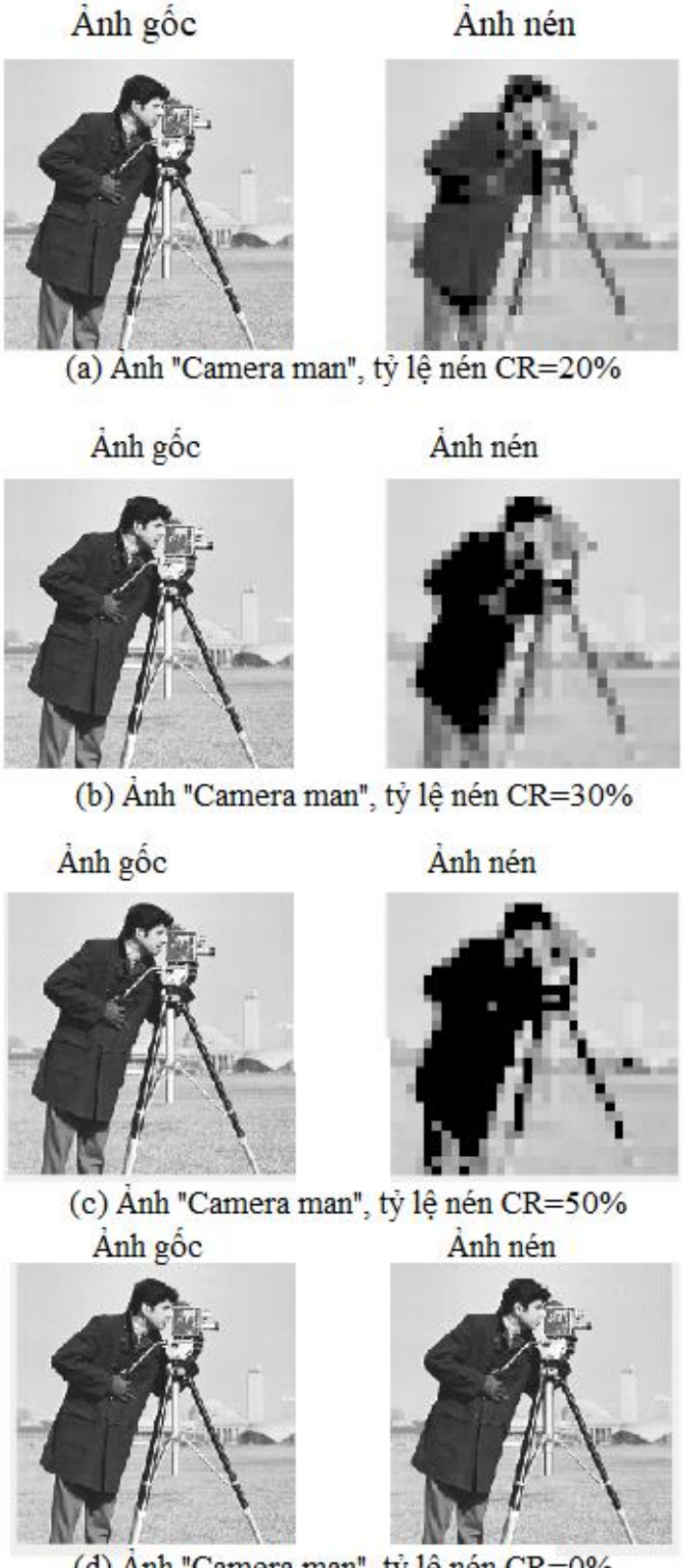
(c) Tín hiệu vào 0010



(d) Tín hiệu vào 0001

Hình 2.15: Tín hiệu ảnh truyền qua 6x6 MMI tại các đầu vào khác nhau

Tiếp theo Luận án mô phỏng thực hiện nén ảnh sử dụng 6×6 MMI cho ảnh "camera man" làm ví dụ. Ảnh "camera man" có kích thước 256×256 là ảnh mức xám 8 bit. Giả sử mong muốn nén với các tỷ lệ 0%, 20%, 30% và 50%. Kết quả mô phỏng được chỉ ra ở Hình 2.16.



Hình 2.16: Ảnh gốc và ảnh nén sau bộ biến đổi Haar 6×6 MMI toàn quang

Kết quả tính toán MSE và PSNR với các tỷ lệ nén khác nhau được chỉ ra ở Bảng 2.2.

Bảng 2.2: Kết quả MSE và PSNR của ảnh gốc và ảnh nén dùng Haar 6×6 MMI

Tỷ lệ nén CR	**PSNR (dB)**	**MSE**
20%	67	0.0126
30%	65	0.0223
50%	62	0.0333
0%	inf	0

Như vậy phần này đã thiết kế thành công kỹ thuật nén ảnh dựa trên biến đổi Haar bằng cách sử dụng bộ ghép MMI 6×6 trong miền toàn quang. Cách tiếp cận được đề xuất rất hữu ích cho việc xử lý ảnh và xử lý dữ liệu lớn ở tốc độ cực cao. Phương pháp được đề xuất có thể được tích hợp với camera AI để thực hiện xử lý ảnh trực tiếp trong camera ASP.

2.2 Nén ảnh sử dụng g biến đổi cosine (DCT) toàn quang

Các kỹ thuật quang học đã được sử dụng cho một loạt các xử lý tín hiệu như nhận dạng mẫu, tạo bề mặt không rõ ràng cho các ứng dụng xử lý tín hiệu radar và xử lý hình ảnh [74]. Lý do chính để sử dụng bộ xử lý tín hiệu quang là lợi thế về băng thông cao hơn bộ vi xử lý điện tử. Do thông lượng cao, ứng dụng của hệ thống tính toán hiệu suất cao xử lý tín hiệu quang rất hấp dẫn. Các phép biến đổi xử lý tín hiệu quang tử như biến đổi Fourier rời rạc (DFT) [75], biến đổi cosin rời rạc (DCT) và biến đổi wavelet rời rạc (DWT) rất hữu ích cho việc xử lý tín hiệu không gian và tính toán quang học như phân tích phổ, lọc và mã hóa, v.v.

Trong những năm gần đây, các nhiệm vụ xử lý tín hiệu trong miền quang học sử dụng các hệ thống thấu kính, bộ ghép định hướng và mạng sao đơn mode. Tuy nhiên, các hệ thống dựa trên các công nghệ này thường khá lớn, thiếu độ chính xác và yêu cầu đặt cơ khí chính xác cao. Ngoài ra, cấu trúc để thực hiện các chuyển đổi dựa trên công nghệ sợi quang đòi hỏi các cáp sợi quang chéo cồng kềnh. Gần đây, thiết kế của biến đổi DFT và DCT sử dụng bộ ghép hướng sợi quang đã được đề xuất [76, 77]. Trong tài liệu [78] một số phép biến đổi như phép biến đổi Hadamard và phép biến đổi đơn thể rời rạc đã sử dụng cấu trúc MMI và hình ba chiều ống dẫn sóng đa mode. Tuy

nhiên, các thiết bị này được thiết kế cho hệ thống vật liệu InP [42]. Đối với thiết bị sử dụng hình ảnh ba chiều, cần phải có một quy trình chế tạo phức tạp. Sự hiện diện của hình ảnh ba chiều trong ống dẫn sóng đa mode có xu hướng dẫn đến tổn hao bổ sung. Gần đây, các phương pháp thực hiện phép biến đổi Fourier toàn quang và DCT dựa trên giao thoa đa mode MMI đã được đề xuất [79]. Việc thiết kế các hiết bị này đã được thực hiện trên hệ thống vật liệu silica. Tuy nhiên các cấu trúc trên sử dụng các hệ thống dịch pha và ghép mode phức tạp, cồng kềnh và khó áp dụng cho thiết kế tích hợp xử lý ảnh.

Ngoài ra, nghiên cứu gần đây về xử lý ảnh toàn quang dựa trên quang học tích hợp đã được thực hiện [37, 24]. Tuy nhiên, chỉ có phép biến đổi Haar sử dụng bộ ghép định hướng để nén ảnh mới được nghiên cứu. Thêm vào đó, hỗ trợ nén dữ liệu để giảm việc sử dụng nhiều tài nguyên đắt tiền. Trong những năm gần đây, việc triển khai thuật toán trí tuệ nhân tạo (AI) tiếp tục dựa vào các hệ thống máy tính điện tử. Các mạng nơ-ron ban đầu dựa trên các thiết kế CPU tiêu chuẩn để tính toán, không có khả năng đáp ứng các yêu cầu của các tập dữ liệu cực lớn. Hiệu suất tính toán song song thấp, cần phải thay thế bằng GPU có khả năng tính toán song song. GPU tạo điều kiện cho sự phát triển của học sâu và AI trong thực tế. Tuy nhiên, học sâu cổ điển có một điểm nghẽn do xử lý tốc độ cao của các đầu vào điện. Các nhà nghiên cứu đang cố gắng tìm ra các phương pháp khác để giải quyết các khiếm khuyết điện tử và các hệ thống tính toán hiện tại. Một trong những câu trả lời hứa hẹn nhất để giải quyết các vấn đề về vận chuyển dữ liệu là liên kết quang tử hoặc hệ thống tính toán toàn quang [24].

Trong khi DST và DCT được sử dụng như một tổng các hàm cosine dao động với các tần số khác nhau để thể hiện các điểm dữ liệu trong một chuỗi hữu hạn. Nó có các ứng dụng rộng rãi trong kỹ thuật máy tính như nén âm thanh MP3 và nén ảnh JPEG với cosine so với các hàm sin, nhưng một số hàm cosine có thể được sử dụng để xấp xỉ tín hiệu nhằm thể hiện hiệu quả các điều kiện biên cho các phương trình vi phân [80]. DCT được sử dụng trong nén video và ảnh vì các thuật toán khác nhau được phát triển để làm cho DCT trở nên hiệu quả về mặt tính toán và được triển khai với DCT để giảm độ phức tạp bằng cách loại bỏ phép nhân để tính gần đúng. Nhiều ứng dụng xử lý hình ảnh, chẳng hạn như theo dõi, yêu cầu DCT lớn hơn. Tính gần đúng DCT có các đặc điểm như độ phức tạp thấp, năng lượng lỗi thấp và độ dài DCT cao để hỗ trợ các phương pháp mã hóa video mới nhất cũng như các ứng dụng khác như giám sát, mã hóa, theo dõi, mã hóa và nén. Tuy nhiên, các thuật toán DCT hiện tại không thể hoạt động hiệu quả cho tất cả các tính năng đã đề cập do giới hạn của các hệ thống tính toán

trong miền điện tử. Việc triển khai DST và DCT cho các ứng dụng dữ liệu lớn là rất hấp dẫn, đặc biệt là khi chúng áp dụng cho các tập dữ liệu khổng lồ Việc sử dụng bộ xử lý ARM trong FPGA và chuyển FPGA sang các nền tảng tính toán không đồng nhất [81]. Do đó, trong nghiên cứu này, tác giả đề xuất một phương pháp mới để hiện thực hóa DST và DCT toàn quang dựa trên cấu trúc giao thoa đa mode (MMI) sử dụng ống dẫn sóng silicon cho các ứng dụng xử lý hình ảnh. Cấu trúc vi mạch được đề xuất có ưu điểm của tổn hao thấp, kích thước nhỏ gọn và dung sai chế tạo cao, phù hợp với công nghệ CMOS hiện thời.

2.2.1. Nguyên lý thiết kế DCT và DST toàn quang

Cấu trúc bộ biến đổi DCT và DST toàn quang sử dụng 4×4 MMI được đề xuất ở Hình 2.17. Chiều rộng của bộ ghép 4×4 MMI là W_{MMI} và chiều dài là L_{MMI}. Trường thông tin trong cấu trúc MMI được diễn dưới dạng [82]:

$$E(x,z) = exp\,(-jkz)\sum_{m=1}^{M} E_m exp\,(j\frac{m^2\pi}{4\Lambda})sin\,(\frac{m\pi}{W_{MMI}}x) \tag{2.9}$$

Trong đó $k = \frac{2\pi n}{\lambda}$, λ là bước sóng hoạt động. Trong nghiên cứu này sử dụng bước sóng của các màu R, G, B tương ứng với ảnh màu R, G, B; n là chiết suất của ống dẫn tín hiệu; M là tổng số mode trong MMI.

Hình 2.17: Biến đổi DCT và DST dùng 4x4 MMI

Trường mode bên trong cấu trúc MMI được tính theo công thức:

$$V_{ir} = \begin{cases} \sqrt{\frac{2}{N}}\sin\,(\frac{r\pi}{N}(i+0.5) & r \neq N \\ \sqrt{\frac{1}{N}}\sin\,(\pi(i+0.5) & r = N \end{cases} \tag{2.10}$$

Trong đó V_{ir} là các phần tử dòng i và hàng r của ma trận V_N thể hiện mối quan hệ giữa các mode trong cổng tín hiệu ra và trường đầu ra. Trong nghiên cứu này chọn chiều dài MMI phù hợp tại $L_{MMI} = 0.5\Lambda$,trong đó $\Lambda = n\frac{W_{MMI}^2}{\lambda}$. Ma trận truyền mode qua MMI lúc này được tính như sau:

$$M = VBV^T \tag{2.11}$$

Trong đó: B là ma trận chéo của các phần tử $b_{rr} = \exp\left(j\frac{r^2\pi}{8}\right)$ cho cấu trúc 4x4 MMI. Phương trình trên có thể được viết lại thành:

$$M_{uv} = je^{j\frac{\pi}{4}}\sqrt{\frac{2}{N}}sin\left(\frac{\pi(u+0.5)(v+0.5)}{N}\right)exp\ (-j\pi\frac{(u+0.5)(v+0.5)}{2N}) \tag{2.12}$$

Nếu các bộ dịch pha được sử dụng tại các cổng đầu vào và ra của MMI thì ma trận đặc tính của MMI được tính theo công thức:

$$T = D_{out}MD_{in} = D_{out}(VBV^T)D_{in} \tag{2.13}$$

Bằng cách thiết lập các bộ di pha thụ động nhờ sử dụng cấu trúc ống dẫn sóng rộng như được chỉ ra dưới đây, ta có thể thiết kế được MMI có hàm truyền:

$$T_{uv} = je^{j\frac{\pi}{4}}\sqrt{\frac{2}{N}}\sin\ (\frac{\pi(u+0.5)(v+0.5)}{N} \tag{2.14}$$

Do vậy, DST và DCT được thiết kế thành công. Ví dụ với *N=4* thì ma trận của DST là:

$$M_{DST} = \begin{bmatrix} 0.1379 & 0.3928 & 0.5879 & 0.6935 \\ 0.3928 & 0.6935 & 0.1379 & -0.5879 \\ 0.5879 & 0.1379 & -0.6935 & 0.3928 \\ 0.6935 & -0.5879 & 0.3928 & -0.1379 \end{bmatrix} \tag{2.15}$$

Ma trận của DCT là:

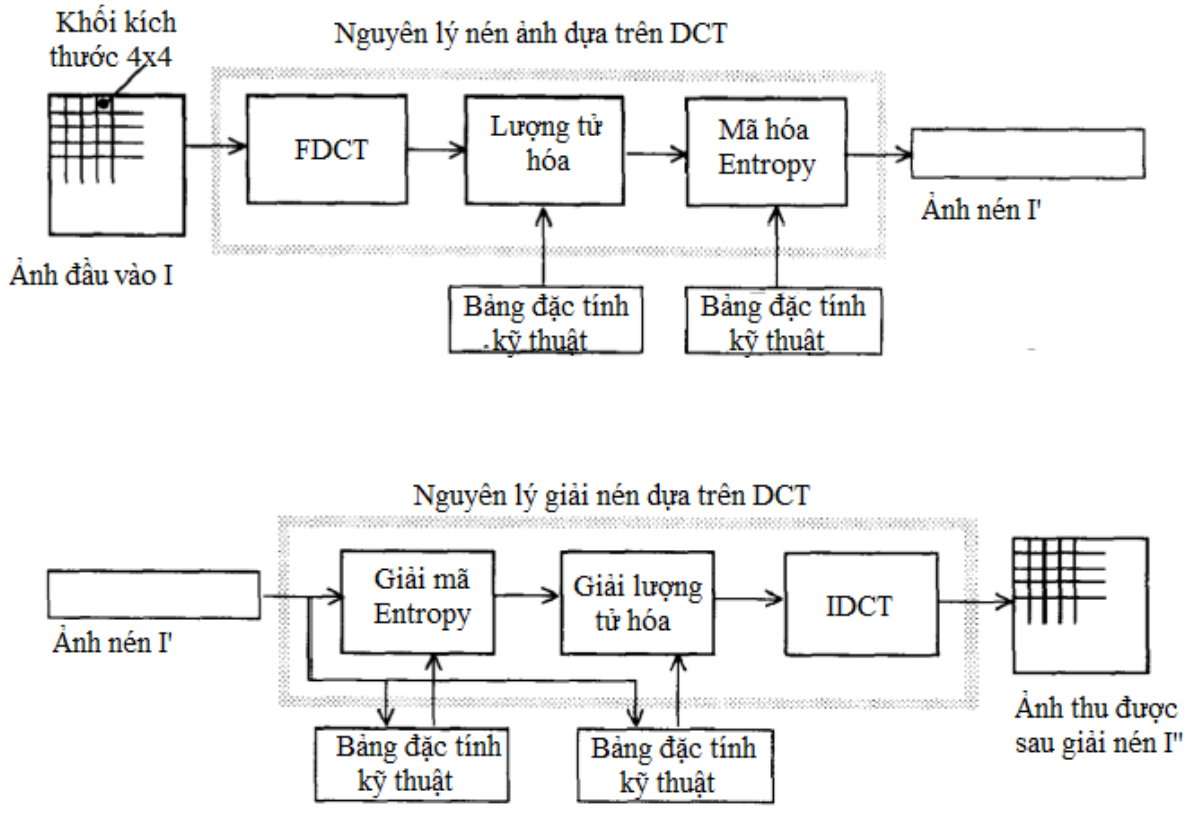
$$M_{DST} = \begin{bmatrix} 0.6935 & 0.5879 & 0.2928 & 0.1379 \\ 0.5879 & -0.1379 & -0.6935 & -0.2928 \\ 0.2928 & -0.6935 & 0.1379 & 0.5879 \\ 0.1379 & -0.1379 & 0.5879 & -0.6935 \end{bmatrix} \tag{2.16}$$

Ở định dạng số, ảnh hai chiều là một tập hợp của hàng triệu pixel, mỗi pixel được biểu diễn bằng một chuỗi các bit. Những ảnh này được raster hoặc bitmapped trái ngược với ảnh vectơ, sử dụng các công thức toán học để tạo ra các đối tượng hình học. Khi nhu cầu về ảnh và phim chất lượng cao ngày càng tăng, người ta đã cố gắng nâng cao độ phân giải lưu trữ ảnh [83]. Ví dụ: 45.000 ảnh sẽ chiếm khoảng 1 TB dung lượng lưu trữ, làm tăng băng thông truyền tải và thời gian truyền tải khi chia sẻ nhiều ảnh cùng một lúc. Hơn nữa, lưu trữ ảnh trên máy chủ web cũng sẽ tốn dung lượng lưu

trữ đồng thời làm tăng thời gian tải, khiến trải nghiệm của khách hàng trở nên kém tối ưu. Nén ảnh loại bỏ nhu cầu không gian lưu trữ lớn này bằng cách cung cấp các giải pháp hiệu quả để chia sẻ, xem và lưu trữ một số lượng lớn ảnh.

Do đó, DCT và DST đã được sử dụng trong xử lý giọng nói và hình ảnh để nén, lọc và trích xuất tính năng trong suốt thập kỷ qua. Sử dụng DCT, một hình ảnh có thể được chia nhỏ thành các phần thành phần của nó. DCT thể hiện một chuỗi gồm vô số điểm dữ liệu thực và rời rạc có tính đối xứng bằng nhau bằng cách tính tổng các hàm cosin dao động ở các tần số khác nhau.

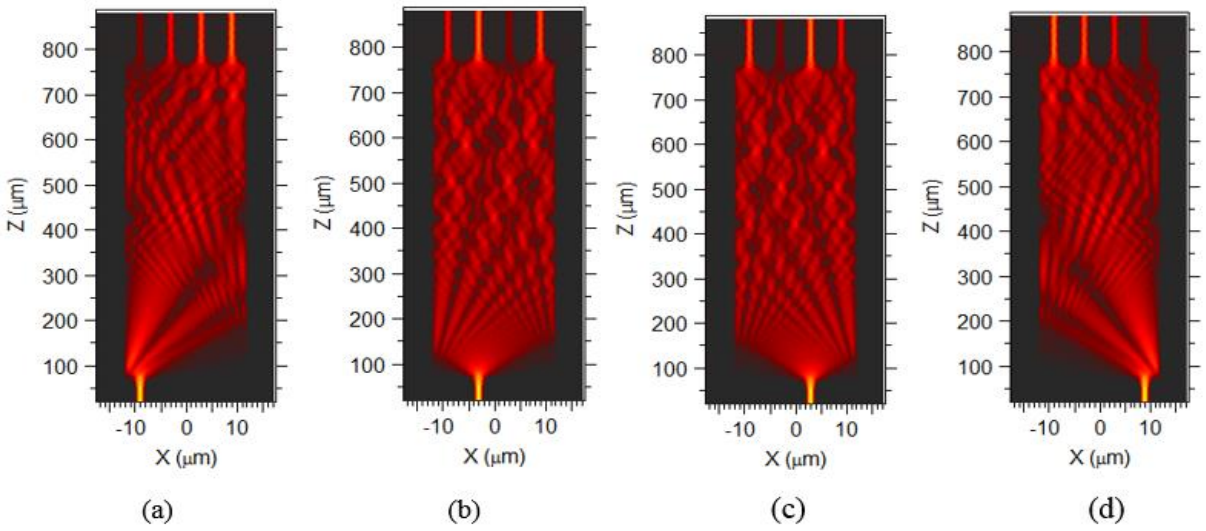
Hình 2.18 cho thấy nguyên tắc nén và giải nén ảnh dựa trên DCT và DST. DST và DCT chuyển đổi dữ liệu hình ảnh thành miền tần số tương đương của nó bằng cách phân vùng ma trận pixel hình ảnh thành các khối có kích thước $N \times N$, N tùy thuộc vào loại hình ảnh. Ví dụ: nếu chúng ta sử dụng hình ảnh đen trắng 8 bit thì tất cả các bóng của màu đen và trắng có thể được thể hiện thành 8 bit do đó tác giả sử dụng $N=8$, tương tự đối với hình ảnh màu 24 bit, tác giả có thể sử dụng $N=24$ nhưng sử dụng khối kích thước $N=24$, độ phức tạp về thời gian có thể tăng lên. Trong nghiên cứu này, tác giả xem xét khối 4×4 bằng cách sử dụng bộ ghép 4×4 MMI.



Hình 2.18: Nguyên lý nén ảnh dùng DCT

Luận án mô phỏng nguyên lý hoạt động của cấu trúc DCT và DST sử dụng 4x4 MMI toàn quang nhờ kỹ thuật mô phỏng số. Kết quả được chỉ ra ở Hình 2.19 với các dữ liệu pixel đầu vào tại các cổng 1, 2, 3 và 4 tương ứng với các tín hiệu $(x_0x_1x_2x_3)^T$ = *(1000),(0100),(0010),(0001)*. Ở đây các tín hiệu màu thể hiện mức xám của ảnh.

Biên độ và pha tương ứng với mức xám đầu vào được tính toán bằng phương pháp mô phỏng số.



Hình 2.19: Mô phỏng DCT dùng 4×4 MMI

Cho xử lý ảnh toàn quang, các ma trận DCT và DST khi sử dụng 4×4 MMI được biểu diễn dưới dạng:

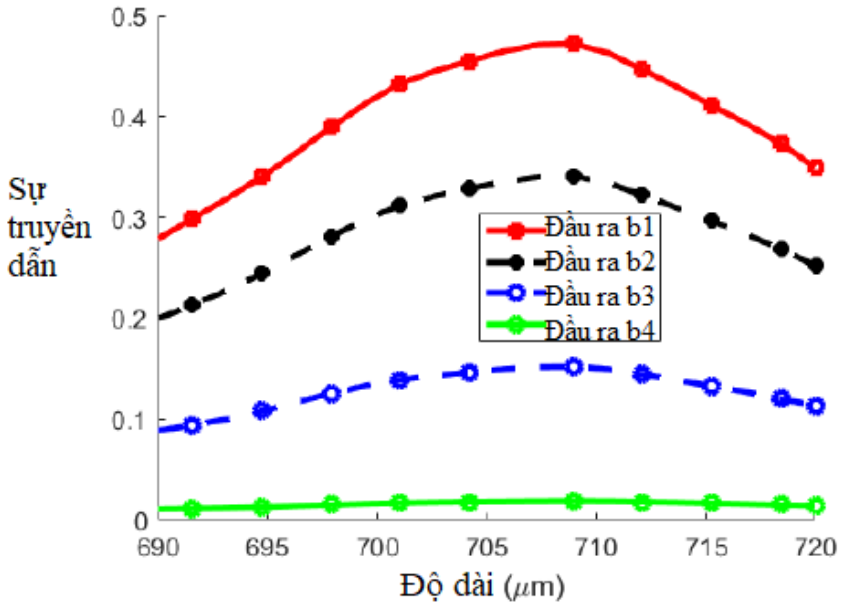
$$H_{4x4} = \begin{bmatrix} \pm a_{11}e^{\pm j\delta} & \pm a_{12}e^{\pm j\delta} & \pm a_{13}e^{\pm j\delta} & 0 \\ \pm a_{21}e^{\pm j\delta} & \pm a_{22}e^{\pm j\delta} & \pm a_{23}e^{\pm j\delta} & 0 \\ \pm a_{31}e^{\pm j\delta} & \pm a_{32}e^{\pm j\delta} & \pm a_{33}e^{\pm j\delta} & \pm a_{34}e^{\pm j\delta} \\ \pm a_{41}e^{\pm j\delta} & \pm a_{42}e^{\pm j\delta} & 0 & \pm a_{44}e^{\pm j\delta} \end{bmatrix} \quad (2.17)$$

Công suất chuẩn hóa tương ứng với cường độ mức xám của ma trận DST và DCT biểu diễn trong công suất quang được tính theo công thức:

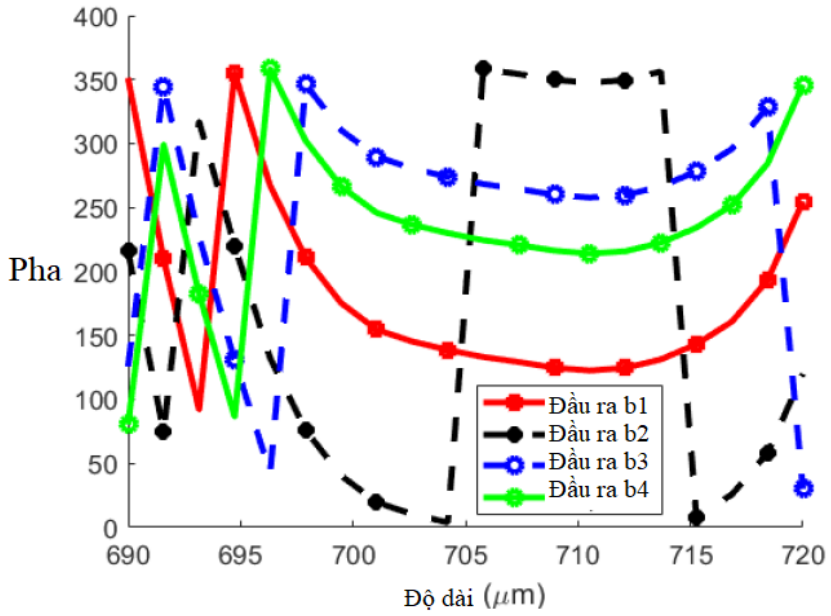
$$M_{DCT} = \begin{bmatrix} 0.4810 & 0.3457 & 0.1543 & 0.0190 \\ 0.3457 & 0.0190 & 0.4810 & 0.1543 \\ 0.1543 & 0.4810 & 0.0190 & 0.3457 \\ 0.0190 & 0.1543 & 0.3457 & 0.4810 \end{bmatrix} \quad (2.18)$$

$$M_{DST} = \begin{bmatrix} 0.0190 & 0.5143 & 0.3457 & 0.4810 \\ 0.5143 & 0.4810 & 0.0190 & 0.4810 \\ 0.3457 & 0.0190 & 0.4810 & 0.5143 \\ 0.4810 & 0.3457 & 0.5143 & 0.0190 \end{bmatrix} \quad (2.19)$$

Qua mô phỏng số tối ưu, tác giả tìm được chiều dài tối ưu của MMI là 706µm. Công suất tín hiệu ra tại các cổng 1, 2, 3 và 4 khi tín hiệu vào tại cổng 1 quanh giá trị tối ưu này được mô phỏng ở Hình 2.20. Pha của tín hiệu ra được mô phỏng ở Hình 2.21. Các kết quả mô phỏng cho thấy chiều rộng của MMI có thể thay đổi trong khoảng *±2µm* và chiều dài thay đổi trong khoảng *±18µm* là không làm ảnh hưởng đến pha và biên độ của tín hiệu ảnh đầu ra. Do vậy cấu trúc DCT và DST thiết kế cho phép sai số chế tạo cao so với công nghệ CMOS hiện nay.

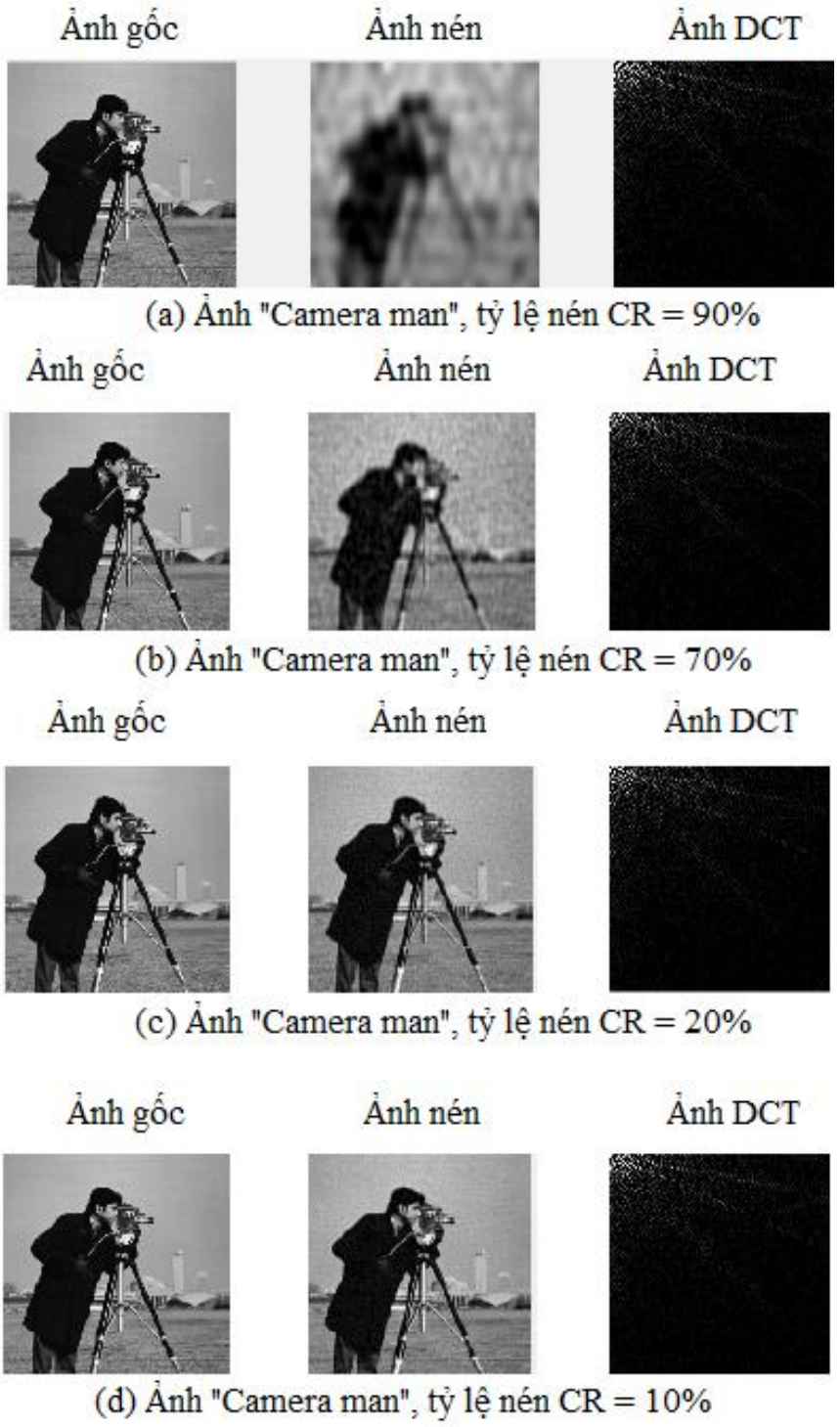


Hình 2.20: Công suất ra của bộ biến đổi DCT và DST theo chiều dài MMI



Hình 2.21: Pha đầu ra của bộ biến đổi DCT và DST theo chiều dài MMI

Cuối cùng ma trận DCT và DST toàn quang được đưa vào mô phỏng ở mức hệ thống với ảnh đầu vào camera man kích cỡ 256×256 với tỷ lệ nén 10%, 20%, 70% và 90% làm ví dụ. Kết quả ảnh đầu và và các tham số MSE và PSNR được tính theo Bảng 2.3.



Hình 2.22: Kết quả mô phỏng nén ảnh sử dụng DCT toàn quang

Kết quả tính toán MSE và PSNR với các tỷ lệ nén khác nhau được chỉ ra ở Bảng 2.3.

Bảng 2.3: Kết quả MSE và PSNR của ảnh gốc và ảnh nén dùng DCT toàn quang

Tỷ lệ nén CR	**PSNR dB**	**MSE**
90%	77	0.0012
70%	75	0.002
20%	70	0.062
10%	66	0.075

Như vậy tác giả đã thiết kế thành công kỹ thuật nén ảnh toàn quang sử dụng DCT và DST dùng cấu trúc tích hợp MMI.

2.3. Nén ảnh sử dụng biến đổi Karhunen–Loève (KLT) toàn quang

Trong nghiên cứu này, tác giả đề xuất một phương pháp mới để nén ảnh trong miền toàn quang sử dụng kỹ thuật biến đổi KLT. Biến đổi KLT được thực hiện chỉ dùng một cấu trúc giao thoa đa mode MMI trong miền toàn quang. Cấu trúc giao thoa sử dụng hiệu ứng giao thoa giới hạn với việc thiết kế vị trí đầu vào và ra phù hợp để tạo được biến đổi KLT. Kỹ thuật nén ảnh toàn quang sử dụng KLT đã được thiết kế thành công trong dải bước sóng RGB nhìn thấy với độ chính xác cao, dải sai số chế tạo $\pm 2\mu m$ trong chiều dài MMI.

Các phép biến đổi tín hiệu đã thu hút được sự quan tâm đáng kể để sử dụng trong các ứng dụng nén dữ liệu, xử lý hình ảnh và các ứng dụng xử lý tín hiệu khác. Trong số một số phép biến đổi tín hiệu, phép biến đổi KLT được coi là tốt nhất do hiệu quả tính toán, tương quan phần dư và lợi ích tiêu chí biến dạng tốc độ của nó. Các phép biến đổi tín hiệu trực giao như biến đổi Fourier, biến đổi cosin rời rạc và biến đổi sin rời rạc hữu ích trong các hệ thống truyền thông và xử lý tín hiệu [84].

Trong số nhiều phép biến đổi, phép biến đổi Karhunen-Loeve nổi tiếng với khả năng nén và lọc dữ liệu và được biết là tối ưu theo nghĩa chúng mang lại dữ liệu không liên quan, đơn giản hóa các hoạt động thành công. Trong khi biến đổi wavelet rời rạc (DWT) được sử dụng để nén ảnh, thì KLT được sử dụng để trang trí hình ảnh; có nghĩa là, KLT được sử dụng trong các phương pháp nén của nhiều hình ảnh có mức độ

tương quan lẫn nhau cao, chẳng hạn như khung hình ảnh y tế và hình ảnh siêu kính video [85].

Trong những năm gần đây, một số nỗ lực đã được thực hiện để nén các tập dữ liệu như vậy một cách hiệu quả nhất có thể. Mục đích là tạo ra một biểu diễn dữ liệu đồng thời xem xét cả lợi ích và nhược điểm của KLT để nén hiệu quả nhất dựa trên tương quan trang trí tối ưu. Trong mọi trường hợp, KLT được sử dụng để trang trí lại trong miền quang phổ. Trước tiên, tất cả các hình ảnh được phân tách thành các khối và mỗi khối sử dụng KLT của riêng nó thay vì một ma trận duy nhất cho toàn bộ hình ảnh. Mục tiêu của nén ảnh là lưu ảnh ở dạng sử dụng ít bit hơn để mã hóa so với ảnh gốc. Điều này có thể hình dung được vì ảnh ở dạng "thô" bao gồm một lượng dữ liệu trùng lặp đáng kể. Phần lớn các hình ảnh không bao gồm các thay đổi cường độ ngẫu nhiên. Mỗi bức tranh trực quan đều có một số kiểu cấu trúc. Do đó, có một số liên kết giữa các pixel liền kề. Nếu có thể phát hiện ra một phép biến đổi thuận nghịch giúp loại bỏ sự trùng lặp bằng cách sắp xếp lại các dữ liệu, thì một bức ảnh có thể được lưu trữ hiệu quả hơn. Phép biến đổi tuyến tính thực hiện điều này là phép biến đổi KLT.

Trong trường hợp độ phân giải thấp và nén tốc độ bit thấp, cách tiếp cận này kém hơn JPEG tiêu chuẩn. Tuy nhiên, trong trường hợp ảnh chất lượng cao ở tốc độ bit cao, số lượng thông tin bên trở nên tương đối nhỏ so với lượng thông tin chính. Ngoài ra, việc chụp những bức ảnh đa chiều ngày càng trở nên quan trọng hơn trong thời đại ngày nay, đặc biệt là để nâng cao giá trị của hệ thống thực tế ảo (VR) và thực tế tăng cường (AR). PCA là một phương pháp thường được sử dụng để giảm các tập dữ liệu đa chiều xuống các kích thước thấp hơn cho các mục đích phân tích, nén hoặc phân loại. PCA yêu cầu tính toán phân tách giá trị riêng hoặc phân tách giá trị đơn lẻ của một bộ sưu tập dữ liệu, thường là sau khi căn giữa trung bình. Tuy nhiên, cần nhấn mạnh rằng việc sử dụng KLT cho các tác vụ như nhận dạng mẫu hoặc xử lý hình ảnh có thể khó khăn vì nó xử lý dữ liệu là một chiều trong khi chúng là hai chiều trong thực tế. Do đó, hầu hết các phương pháp đã thiết lập đều sử dụng một số loại giảm trước kích thước, trong nhiều trường hợp, bỏ qua các mối quan hệ không gian giữa các pixel.

Đối với bộ xử lý tín hiệu tốc độ cao, xử lý tín hiệu được mong đợi thực hiện trong miền toàn quang [86]. Có nhiều phương pháp xử lý tín hiệu toàn quang, hầu hết chúng đều dựa trên sợi quang học hoặc thấu kính [6]. Một cách tiếp cận quan trọng

khác để thực hiện các phép biến đổi trực giao toàn quang là sử dụng các cấu trúc giao thoa đa mode do ưu điểm của chúng là nhỏ gọn, dung sai chế tạo tốt và dễ tích hợp.

Trong Luận án này, tác giả tập trung vào việc thực hiện KLT để xử lý hình ảnh có thể ứng dụng trong miền toàn quang. Luận án sử dụng cấu trúc MMI cấu tạo đặc biệt để nhận biết các KLT. Dữ liệu hình ảnh được xử lý trực tiếp trong miền toàn quang và nó không cần chuyển đổi sang tín hiệu số. Bằng thiết kế MMI đặc biệt, biến đổi KLT được thiết kế để ứng dụng trong nén ảnh. Đề xuất thiết bị sau đó được xác minh và thiết kế tối ưu bằng cách sử dụng các công cụ mô phỏng số.

Cấu trúc bộ biến đổi DCT và DST toàn quang sử dụng 4×4 MMI được đề xuất ở Hình 2.23. Chiều rộng của bộ ghép 4×4 MMI là W_{MMI}và chiều dài là $L_{MMI} = 2\Lambda/(N+1)$.

Hình 2.23: Biến đổi DCT và DST dùng 4x4 MMI

Trong thiết kế này, tác giả chọn các vị trí dẫn tín hiệu vào và ra của MMI tại các vị trí $x_i = W_{MMI}/(N+1)$, với 4×4 MMI ta có N=4. Trường mode bên trong cấu trúc MMI được tính theo công thức:

$$V_{ir} = \frac{2}{\sqrt{2N+2}} sin\left(\frac{ir\pi}{N+1}\right) \quad (2.20)$$

Trong đó V_{ir} là các phần tử dòng i và hàng r của ma trận V_N thể hiện mối quan hệ giữa các mode trong cổng tín hiệu ra và trường đầu ra. Trong nghiên cứu này chọn chiều dài MMI phù hợp tại $L_{MMI} = 0.5\Lambda$, trong đó $\Lambda = \frac{nW_{MMI}^2}{\lambda}$. Ma trận truyền mode qua MMI lúc này được tính như sau:

$$M = VBV^T \quad (2.21)$$

Trong đó, B là ma trận chéo của các phần tử $b_{rr} = \exp\left(j\frac{r^2\pi}{2N+2}\right)$ cho cấu trúc 4x4 MMI. Phương trình trên có thể được viết lại thành:

$$M_{uv} = 2j\frac{e^{j\frac{\pi}{4}}}{\sqrt{2N+2}}\sin\left(\frac{\pi(uv)}{N+1}\right)\exp\left(-j\pi\frac{(u^2+v^2)}{N}\right) \quad (2.22)$$

Nếu các bộ dịch pha được sử dụng tại các cổng đầu vào và ra của MMI thì ma trận đặc tính của MMI được tính theo công thức:

$$T = D_{out}MD_{in} = D_{out}(VBV^T)D_{in} \quad (2.23)$$

Bằng cách thiết lập các bộ di pha thụ động nhờ sử dụng cấu trúc ống dẫn sóng rộng như được chỉ ra dưới đây, ta có thể thiết kế được MMI có hàm truyền:

$$T_{uv} = 2j\frac{e^{j\frac{\pi}{4}}}{\sqrt{2N+2}}\sin\left(\frac{\pi(uv)}{N+1}\right) \quad (2.24)$$

Do vậy, KLT được thiết kế thành công. Ví dụ với N=4 thì ma trận của KLT là:

$$M_{KLT} = \begin{bmatrix} 0.3717 & 0.6015 & 0.6015 & 0.3717 \\ 0.6015 & 0.3717 & -0.3717 & -0.6015 \\ 0.6015 & -0.3717 & -0.3717 & 0.6015 \\ 0.3717 & -0.6015 & 0.6015 & -0.3717 \end{bmatrix} \quad (2.25)$$

Ma trận công suất quang của KLT là:

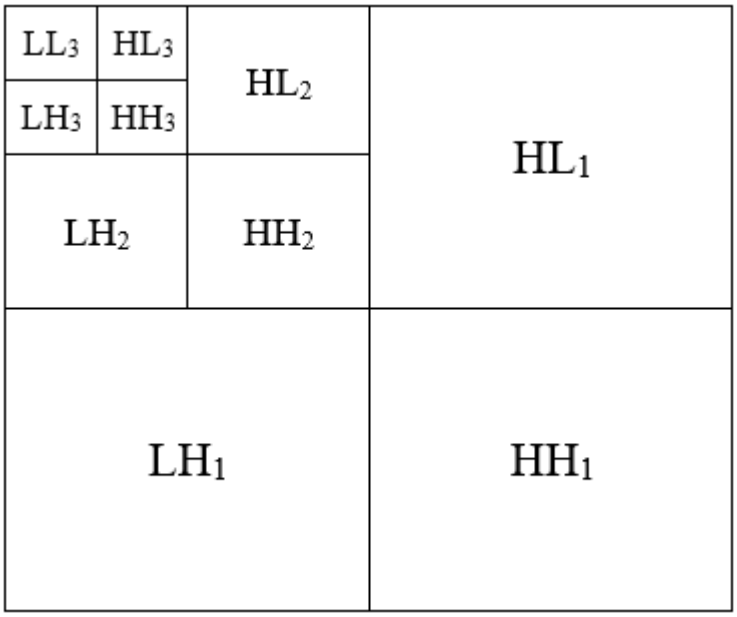
$$M_{KLT} = \begin{bmatrix} 0.1382 & 0.3618 & 0.3618 & 0.1382 \\ 0.3618 & 0.1382 & 0.1382 & 0.3618 \\ 0.3618 & 0.1382 & 0.1382 & 0.3618 \\ 0.1382 & 0.3618 & 0.3618 & 0.1382 \end{bmatrix} \quad (2.26)$$

Phép biến đổi KLT đề cập đến các biểu thức xấp xỉ đa phân. Trong thực tế, phân tích đa phân giải được thực hiện bằng cách sử dụng 4 ngân hàng bộ lọc kênh (cho mỗi cấp độ phân hủy) bao gồm một bộ lọc thông thấp và một bộ lọc thông cao, và mỗi ngân hàng bộ lọc được lấy mẫu ở một nửa tỷ lệ (1/2 lấy mẫu xuống) của tần số trước đó. Bằng cách lặp lại phương pháp này, bất kỳ thứ tự biến đổi wavelet nào cũng có thể đạt được. Phương pháp lấy mẫu giảm duy trì tham số tỷ lệ (bằng 1/2) trong các lần biến đổi wavelet liên tiếp, giúp cải thiện việc thực thi máy tính. Trong trường hợp là hình ảnh, việc lọc được thực hiện độc lập bằng cách lọc các dòng và cột. Các phần ở mỗi tỷ lệ được phân rã một cách đệ quy và được minh họa trong Hình 2.24.

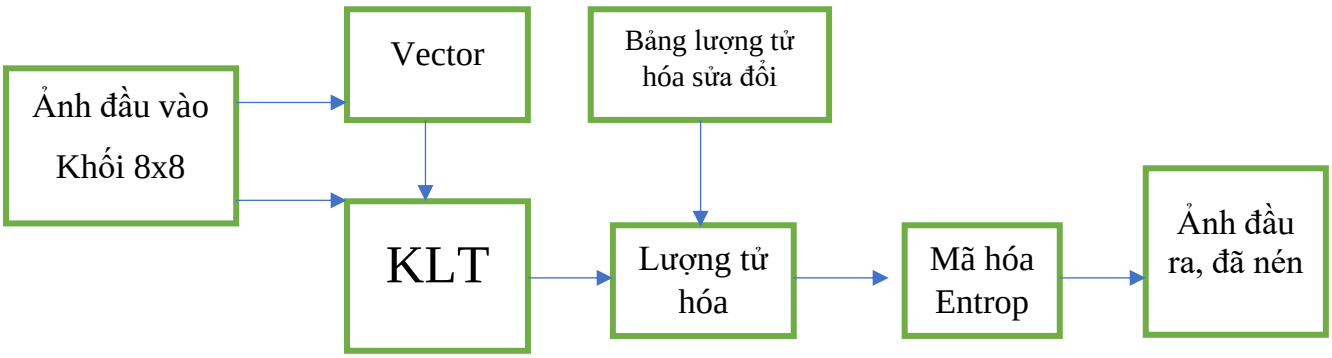
Biến đổi KLT bắt đầu với ma trận phương sai của vec tơ $x = (x_1,x_2,...,x_n)^T$ được tạo từ các điểm ảnh lân cận được sắp xếp theo từng khối như ở Hình 2.25. Ma trận phương sai $C_x = E((x-m_x)(x-m_x)^T)$, m_x là vec tơ trung bình và E là giá trị tham số. Kết quả là KLT được biểu diễn dưới dạng:

$$X = V^T(x - m_x) \quad (2.27)$$

Trong đó $X = (X_1, X_2, \ldots, X_n)^T$ là tập vec tơ được biến đổi, C_x là vec tơ riêng.

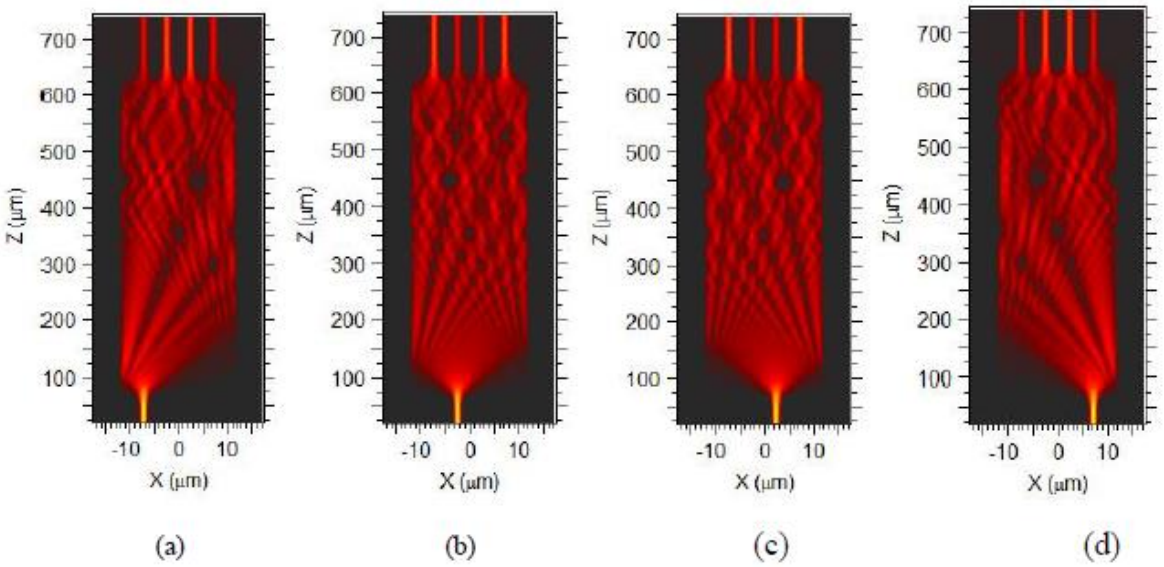


Hình 2.24: Thể hiện dữ liệu ảnh theo thông cao và thấp



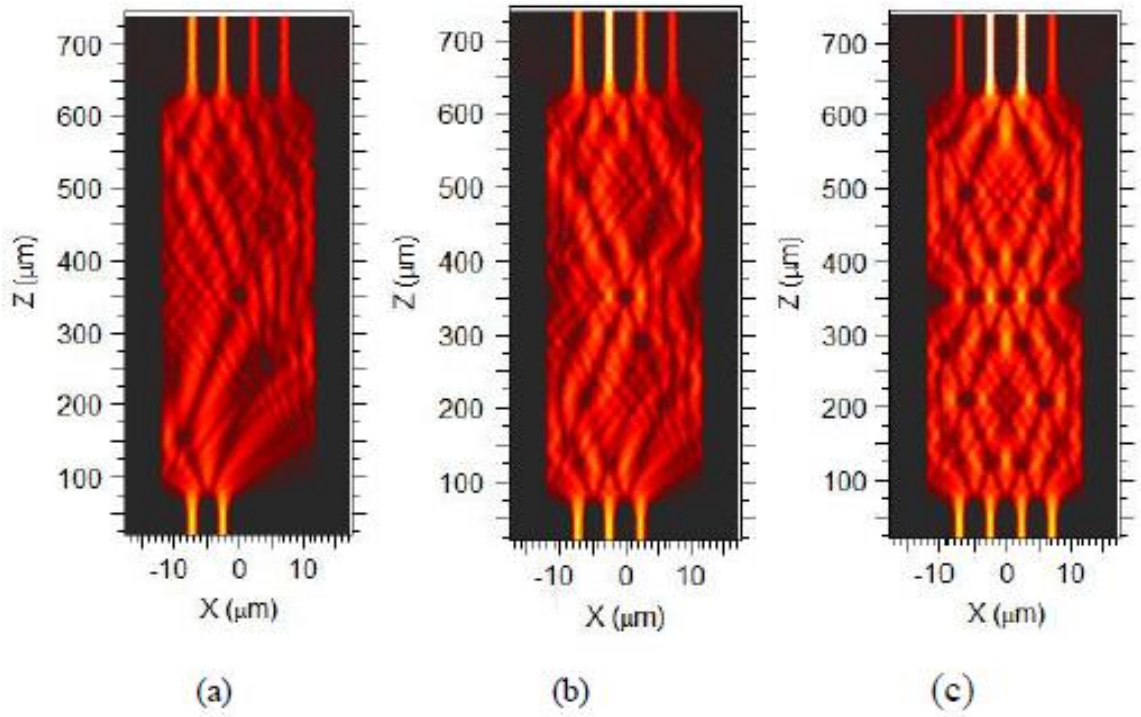
Hình 2.25: Nguyên lý nén ảnh dùng KLT

Luận án mô phỏng nguyên lý hoạt động của cấu trúc KLT sử dụng 4×4 MMI toàn quang nhờ kỹ thuật mô phỏng số. Kết quả được chỉ ra ở Hình 2.26 với các dữ liệu pixel đầu vào tại các cổng 1, 2, 3 và 4 tương ứng với các tín hiệu $(x_0x_1x_2x_3)^T$ = *(1000),(0100),(0010),(0001).* Ở đây các tín hiệu màu thể hiện mức xám của ảnh. Biên độ và pha tương ứng với mức xám đầu vào được tính toán của mô phỏng số.



Hình 2.26: Mô phỏng nguyên lý hoạt động của cấu trúc KLT dùng 4x4 MMI

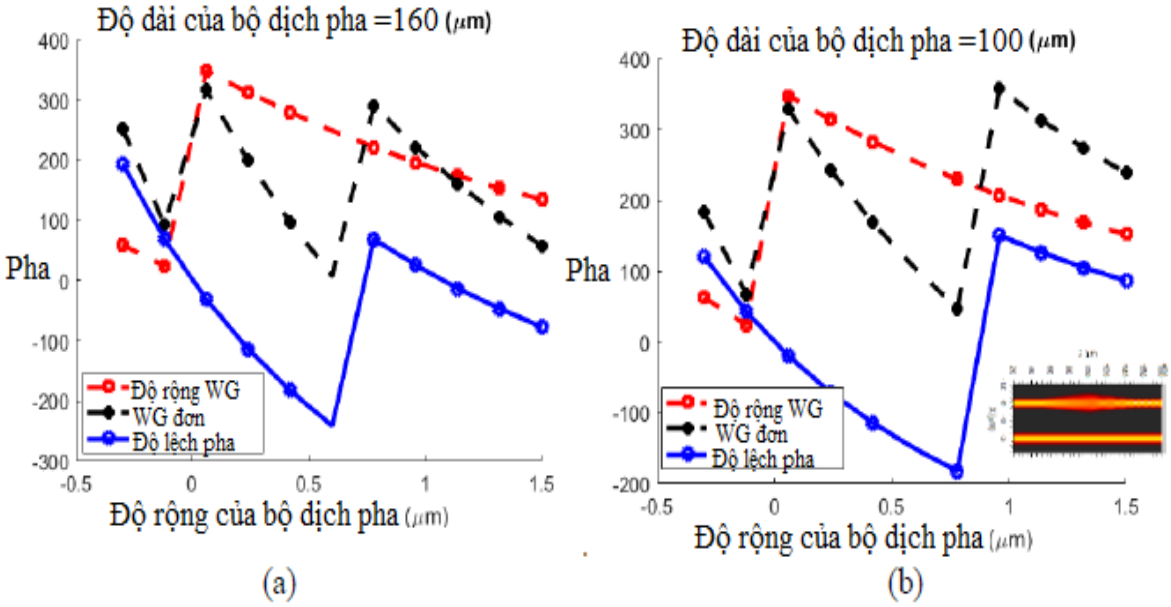
Tiếp theo kết quả với 2 điểm ảnh cùng truyền qua KLT toàn quang được mô phỏng ở Hình 2.27 với các dữ liệu pixel đầu vào tại các cổng 1, 2, 3 và 4 tương ứng với các tín hiệu $(x_0x_1x_2x_3)^T = (1100),(1110),(1111)$.



Hình 2.27: Mức xám ảnh truyền qua KLT với 2 điểm ảnh đầu vào

Trong nghiên cứu này, tác giả sử dụng cấu trúc ống dẫn sóng rộng để tạo ra pha mong muốn bất kỳ. Pha như vậy hoàn toàn thụ động và đơn giản, có thể tích hợp được

với MMI mà không cần cấu trúc phức tạp như chỉ ra ở nghiên cứu gần đây [75]. Kết quả mô phỏng dịch pha dùng ống dẫn sóng rộng tại các cổng đầu vào và ra của bộ biến đổi KLT được chỉ ra ở Hình 2.28.

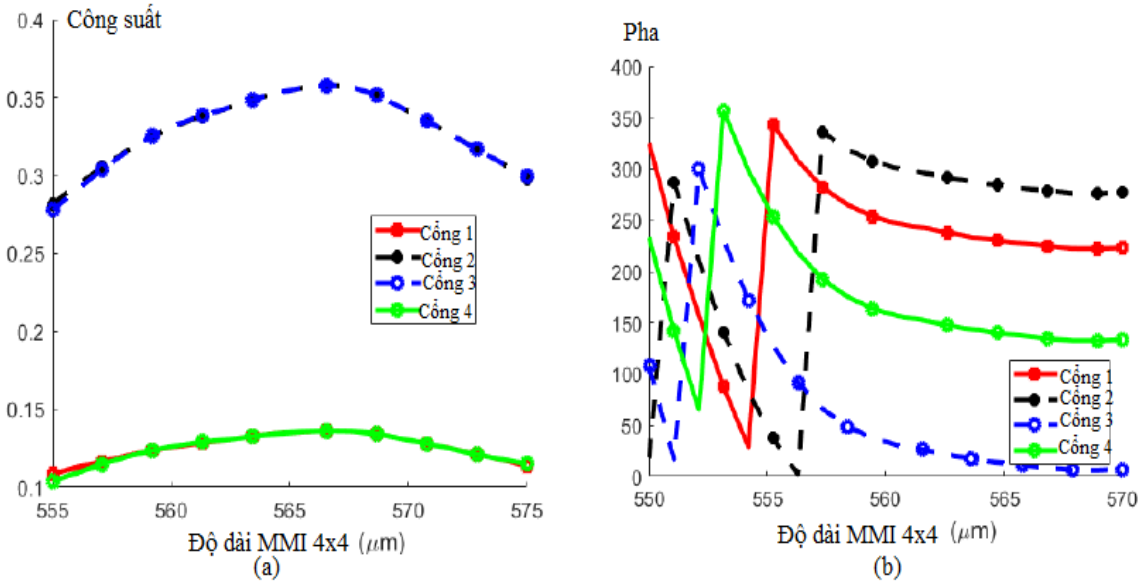


Hình 2.28: Bộ dịch pha tín hiệu đạt được từ sử dụng ống dẫn sóng rộng

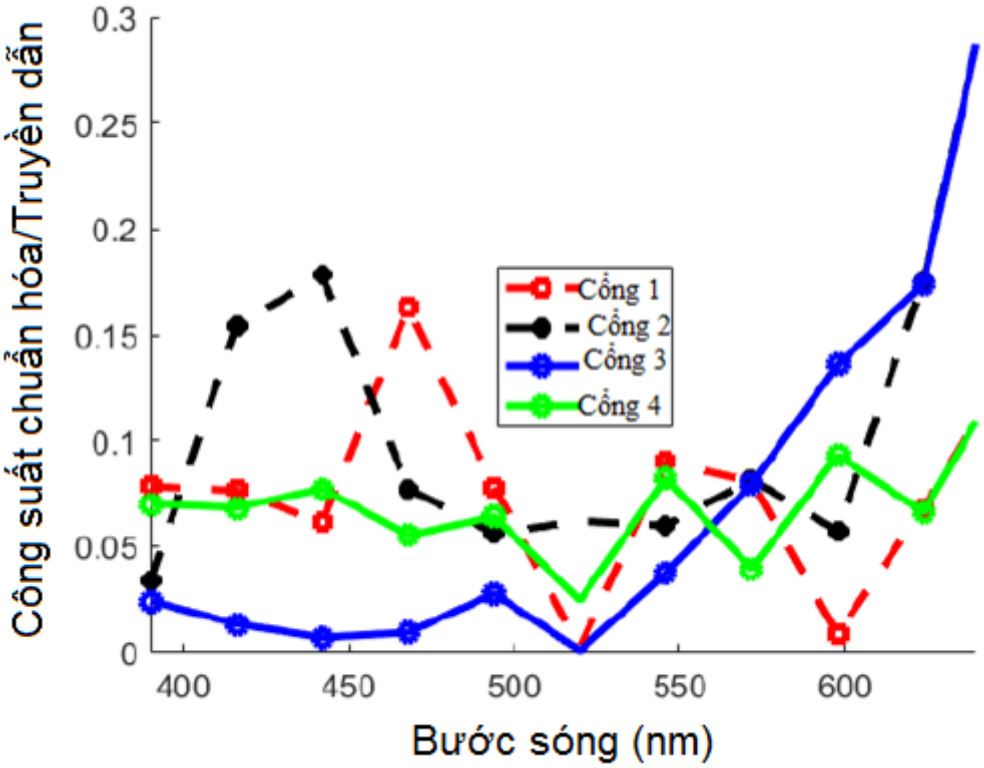
Với ma trận KLT đạt được từ cấu trúc MMI, ta có thể viết dưới dạng tổng quát:

$$H_{4x4} = \begin{bmatrix} \pm a_{11}e^{\pm j\delta} & \pm a_{12}e^{\pm j\delta} & \pm a_{13}e^{\pm j\delta} & \pm a_{14}e^{\pm j\delta} \\ \pm a_{21}e^{\pm j\delta} & \pm a_{22}e^{\pm j\delta} & \pm a_{23}e^{\pm j\delta} & \pm a_{24}e^{\pm j\delta} \\ \pm a_{31}e^{\pm j\delta} & \pm a_{32}e^{\pm j\delta} & \pm a_{33}e^{\pm j\delta} & \pm a_{34}e^{\pm j\delta} \\ \pm a_{41}e^{\pm j\delta} & \pm a_{42}e^{\pm j\delta} & \pm a_{43}e^{\pm j\delta} & \pm a_{44}e^{\pm j\delta} \end{bmatrix} \quad (2.28)$$

Qua mô phỏng số tối ưu, tác giả tìm được chiều dài tối ưu của MMI là 566μm. Công suất tín hiệu ra tại các cổng 1, 2, 3 và 4 khi tín hiệu vào tại cổng 1 quanh giá trị tối ưu này được mô phỏng ở Hình 2.29. Pha của tín hiệu ra được mô phỏng ở Hình 2.30. Hình 2.30 thể hiện công suất ra thay đổi xung quanh các bước sóng màu RGB. Dải công suất chuẩn hóa thay đổi từ 0-0.2. Điều này cho phép xử lý tín hiệu ảnh màu chính xác.



Hình 2.29: Công suất ra và pha của KLT dùng MMI quanh giá trị tối ưu



Hình 2.30: Công suất đầu ra tại các cổng 1-4 trong dải ánh sáng RGB

Cuối cùng tác giả dùng thuật toán máy tính để chuyển ma trận KLT toàn quang được đưa vào mô phỏng ở mức hệ thống với ảnh đầu vào "camera man kích cỡ 256×256 với tỷ lệ nén 10%, 20%, 70% làm ví dụ. Kết quả cho thấy đã thực hiện thành công nén ảnh toàn quang dùng biến đổi KLT toàn quang sử dụng 1 cấu trúc MMI duy nhất.

Ảnh gốc Ảnh nén

(a) Ảnh “Camera man”, CR=70%

Ảnh gốc Ảnh nén

(b) Ảnh “Camera man”, CR=20%

Ảnh gốc Ảnh nén

(c) Ảnh “Camera man”, CR=10%

Hình 2.31: Kết quả mô phỏng nén ảnh sử dụng KLT toàn quang

2.4. Kết luận Chương 2

Chương 2 đã đề xuất phương pháp mới để hiện thực hóa phép biến đổi DHT, DCT/DST và KLT sử dụng cấu trúc 4x4 MMI và 6x6 MMI cho nén ảnh trực tiếp trong miền quang. Các bộ biến đổi toàn quang được thiết kế thành công trên nền tảng vật liệu Si3N4 phù hợp với các mạch VLSI và FPGA. Phương pháp tiếp cận toàn quang học này để thực hiện KLT có thể hữu ích cho các ứng dụng xử lý hình ảnh thời gian thực và tốc độ cao toàn quang học như nén, lọc và mã hóa dữ liệu. Phương pháp này cũng có thể hữu ích cho việc tích hợp xử lý hình ảnh nhanh vào camera AI trong tương lai. Các kết quả có liên quan đến Chương 2 được công bố trong các công trình [J2-J5] và [C1, C2].

Chương 3. TÁCH BIÊN ẢNH VÀ NHẬN DẠNG ẢNH SỬ DỤNG MẠNG NƠ - RON TOÀN QUANG

Chương 3 đề xuất một bộ xử lý quang học mới thực hiện phép nhân ma trận vectơ quang (OVMM) sử dụng cấu trúc bộ cộng hưởng vi mạch dựa trên giao thoa đa mode (MMI). Cấu trúc này chỉ có thể được tích hợp vào một chip duy nhất mà không cần đến bộ ghép kênh phân chia bước sóng (WDM). Việc điều khiển các trọng số và tín hiệu ảnh dựa trên vật liệu graphene đạt tốc độ cao. Kiến trúc được đề xuất cung cấp độ chính xác cao, có thể điều khiển được tín hiệu đầu vào và trọng số bộ lọc, có tốc độ cao, tính nhỏ gọn và khả năng làm việc với các giá trị âm. Toàn bộ thiết bị được thiết kế và mô phỏng trên nền tảng Si3N4 có thể cung cấp mức suy hao thấp và hoạt động trực tiếp với dải bước sóng nhìn thấy của hình ảnh. Kiến trúc mới có thể được áp dụng cho các mạng nơ-ron quang với nhiều lớp các nơ-ron cho mạng nơ-ron nhân tạo. Kết quả được mô phỏng, đánh giá với việc tách biên ảnh dùng các toán tử Roberts, Prewitt và Sobel [87]. Tốc độ xử lý lên đến 28GHz khi sử dụng bộ điều chế dùng graphene và sự khác biệt của MSE với phương pháp thông thường là 0,1. Kết quả mô phỏng cũng cho thấy áp dụng OONN trên tập dữ liệu MNIST cho tốc độ cao gấp 2.8 đến 14 lần so với các phương pháp dùng GPU hiện nay. Không mất tính tổng quát, Luận án thử nghiệm tách biên ảnh trong miền toàn quang sử dụng các toán tử trên.

Đồng thời chương này đề xuất thiết kế một mạng nơ-ron quang học mới trên chip (OONN) dựa trên bộ cộng hưởng vi mạch giao thoa đa mode (MMI-RRs). Kết cấu này có ưu điểm là tốc độ tính toán cao, tổn hao thấp và khả năng làm việc với cả giá trị âm và dương. OONN đã được áp dụng cho tập dữ liệu MNIST với tốc độ nhanh hơn gấp 5-6 lần so với các phương pháp thông thường.

3.1. Thiết kế bộ nhân chập quang tử

Trong những năm gần đây, để đáp ứng nhu cầu ngày càng tăng về tính toán nhanh hơn, các bộ xử lý tính toán như bộ xử lý trung tâm (CPU), bộ xử lý đồ họa (GPU) và bộ xử lý tensor (TPU) đã được phát triển rộng rãi [88]. Tuy nhiên, định luật Moore trong lĩnh vực điện tử đang tiến gần đến giới hạn và làm chậm tốc độ của các cải tiến liên quan đến xử lý dữ liệu. Phương thức xử lý thông tin trong miền quang gần đây đã được thiết lập như một phương tiện truyền thông cho các trung tâm tính toán và dữ liệu lớn trong nhiều năm, nhưng vẫn chưa được sử dụng rộng rãi trong xử lý thông tin và tính toán. Các mạch tích hợp quang tử (PIC), thao tác tín hiệu ánh sáng bằng cách sử dụng ống dẫn sóng quang học trên chip, bộ ghép chùm và bộ tách, bộ điều

biến điện quang, máy dò ảnh và laser, v.v., đã được sử dụng làm nền tảng mới cho xử lý thông tin tốc độ cực cao để đối phó với các giới hạn của định luật Moore [22]. Tận dụng các photon thay vì các electron để tính toán, bộ xử lý quang có thể cung cấp hiệu suất tính toán thông lượng cao, tiết kiệm điện và độ trễ thấp bằng cách khắc phục những hạn chế vốn có của thiết bị điện tử. Nhiều bộ xử lý quang học ứng dụng cụ thể đã được khai thác để giải quyết các nhiệm vụ xử lý tín hiệu và toán học với hiệu suất vượt xa các bộ xử lý điện tử hiện có theo hoặc-ders độ lớn. Các thành phần quang điện tử trên nền tảng PIC đã phát triển mạnh mẽ do khả năng chuyển đổi tín hiệu giữa ánh sáng và điện.

Tuy nhiên, đối với xử lý thông tin quang học trên chip, tồn tại rất ít khối xây dựng cơ bản tương đương với khối được sử dụng trong mạch điện tử. Bộ xử lý quang học với nhiệm vụ của sản phẩm chính là một hoạt động cơ bản trong các lĩnh vực xử lý tín hiệu số hiện đại như xử lý ảnh số, xử lý tín hiệu radar và giao tiếp quang mạch nhất quán. Vào năm 2012, một phương pháp tính toán song song mới cho đa vectơ ma trận quang (MVM) bằng cách thay thế quạt ra và quạt vào bằng thấu kính quang học bằng cách phân tách công suất và ghép bước sóng đã được đề xuất. Phương pháp này cho phép cải thiện tính năng ổn định và hiệu quả sử dụng điện của hệ thống [28].

Những năm gần đây, có một số phương pháp quang học để thực hiện phép nhân tích số chấm hoặc phép nhân vectơ ma trận. Có thể sử dụng bộ cộng hưởng (MRR), bộ cộng hưởng vi đĩa (MDR), cách tử Bragg hoặc giao thoa kế Mach – Zehnder (MZI) [89, 90]. Hai phương pháp điều khiển chính dựa trên hiệu ứng quang nhiệt hoặc hiệu ứng phân tán plasma trong bộ điều biến quang học. So với MZI, kích thước thu nhỏ của MRR khiến chúng trở thành ứng cử viên tốt hơn cho các hệ thống quang tử quy mô lớn vì chúng cho phép tích hợp trên chip dày đặc để giảm dấu chân, tiêu thụ điện năng và chi phí cũng như các hoạt động song song với các nguồn sáng không mạch lạc. Tuy nhiên, đối với các bộ cộng hưởng microring, cần sử dụng bộ ghép trực tiếp quang học [91]. Nhược điểm của cấu trúc này là khó đạt được truyền dẫn mong muốn do độ nhạy của các tỷ số ghép đối với chế tạo. Điều chế tín hiệu quang được thực hiện bằng cách làm trôi đỉnh cộng hưởng của bộ điều chế MRR. Hơn nữa, các cấu trúc trước đây rất khó để tích hợp trong một chip duy nhất do yêu cầu của các phần tử WDM. Do đó, trong nghiên cứu này, tác giả đề xuất một bộ cộng hưởng vi mạch mới dựa trên bộ ghép MMI mà không sử dụng bộ ghép định hướng như trong nghiên cứu trước.

Với sự phát triển của quang tử silicon, kiến trúc dựa trên quang tử silicon đã được chứng minh là thực hiện các hoạt động tích lũy nhân ở tần số nhanh hơn gấp 5 lần so với điện tử thông thường [30]. Phương pháp này sử dụng một ngân hàng các MRR silicon có thể điều chỉnh được để tạo lại trọng lượng khớp thần kinh trên chip. Tuy nhiên, vật liệu graphene là chất hấp dẫn đặc biệt để tạo ra các thiết bị quang học tốc độ cao. Tốc độ điều chỉnh hiện đại nhất của vi mạch graphene là 28–80 GHz do chiết suất của tấm graphene bị thay đổi bởi điện áp đặt vào tấm graphene [92, 93]. Mặt khác, các bộ xử lý điện tử có giới hạn xung nhịp ở khoảng 4–5 GHz khi chúng đạt đến giới hạn phân tán nhiệt [94]. Do đó, việc nghiên cứu cách thức quang tử có thể được sử dụng để thực hiện tích chập và nhân ma trận là cần thiết và quan trọng. Việc triển khai quang học của mạng nơ-ron tích tụ với tốc độ hoạt động nhanh và hiệu quả năng lượng cao đang hấp dẫn do khả năng khai thác tính năng và xử lý dữ liệu tốc độ cao vượt trội. Đặc biệt, xử lý tích hợp dựa trên MVM, là một hoạt động tổng hợp tính toán trong thiết bị điện tử, chiếm hơn 80% tổng thời gian xử lý trong mạng nơ-ron tích chập, do đó có thể đạt được tăng tốc tính toán cho mạng nơ-ron tích hợp bằng cách kết hợp phần cứng và các phép toán MVM [95].

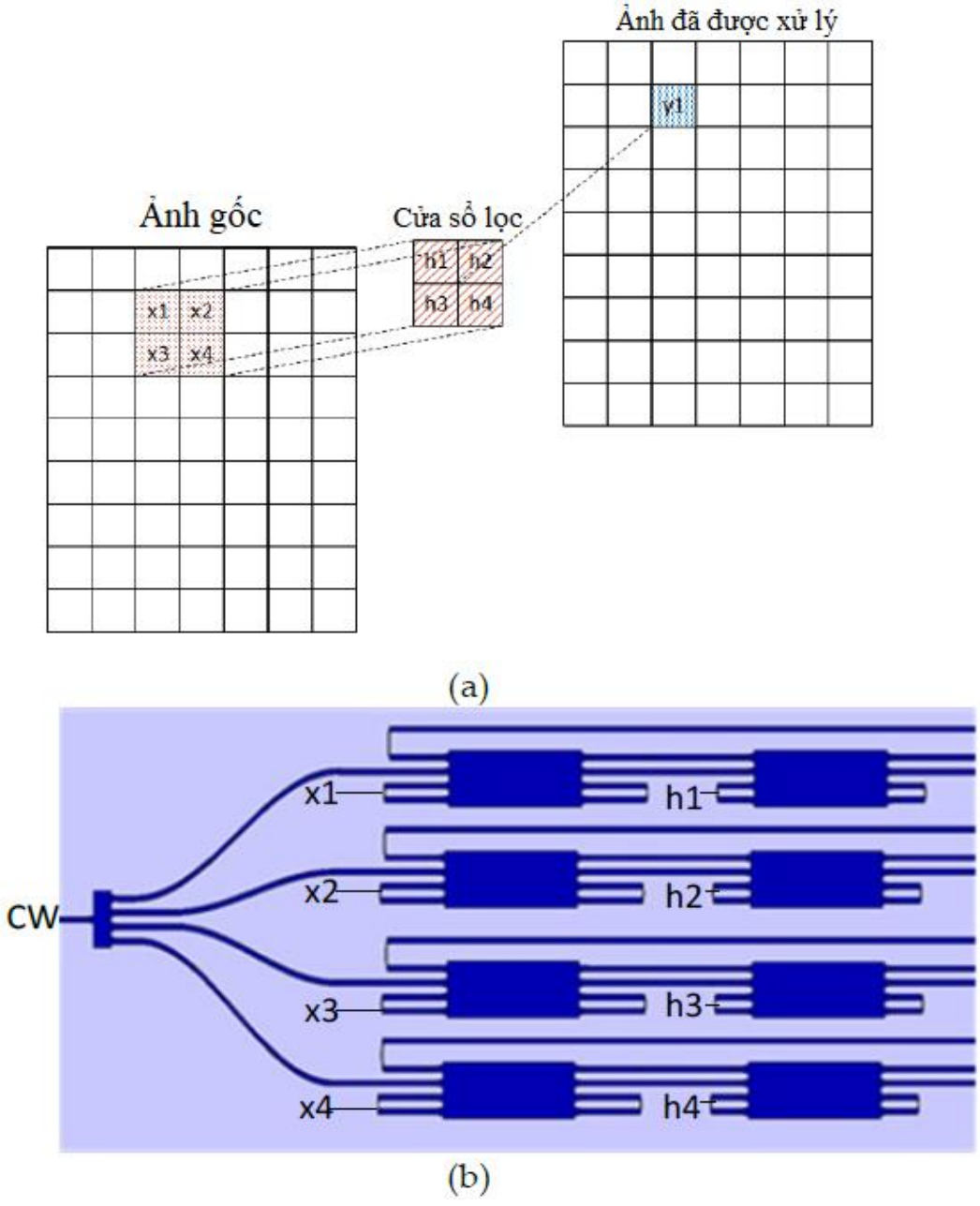
3.1.1. Nguyên lý thiết kế

Cấu trúc đề xuất cho nhân vectơ ma trận hoặc nhân tích chập được thể hiện trong Hình 3.1. Tín hiệu đầu vào được mã hóa bằng cách sử dụng mảng cột đầu tiên của bộ cộng hưởng vi mạch dựa trên MMI và các hệ số trọng số hoặc bộ lọc hạt nhân không cần điều chỉnh vì không cần thay đổi nhân nếu áp dụng cho 1 bộ lọc. Chỉ sử dụng cấu trúc này, bất kỳ bộ lọc nào cũng có thể được tạo ra bằng cách thay đổi hệ số trọng lượng thông qua việc kiểm soát bước sóng cộng hưởng như được phân tích trong phần tiếp theo. Hình 3.1(a) trình bày một ví dụ đơn giản về tính toán tích chập trong xử lý hình ảnh. Ma trận bộ lọc được thiết lập để trích xuất một đối tượng địa lý từ hình ảnh đầu vào và áp dụng cho một cửa sổ trong hình ảnh. Bộ lọc và cửa sổ có cùng kích thước ma trận và tổng các sản phẩm chấm của chúng được tính như sau:

$$y_1 = x_1 h_1 + x_2 h_2 + x_3 h_3 + x_4 h_4 \tag{3.1}$$

Trong đó x_i là các tín hiệu của ảnh đầu vào và hi là các giá trị hệ số trọng số của cửa sổ hay bộ lọc nhân kernel. Bộ lọc h sẽ quét khắp các ảnh theo công thức trên để có pixel ảnh đầu ra. Ví dụ với bộ lọc 2×2, toán tử Roberts 2×2 để tách biên ảnh có $H_x = \begin{bmatrix} +1 & 0 \\ 0 & -1 \end{bmatrix}$, $H_y = \begin{bmatrix} 0 & +1 \\ -1 & 0 \end{bmatrix}$. Trong hình 3.1(b) thể hiện cho ma trận cửa sổ 2×2 sử

dụng 4 bộ vi cộng hưởng dựa vào MMI cho mã hóa 4 pixel ảnh đầu vào và 4 hệ số của cửa sổ bộ lọc. Ở đây chỉ cần dùng một tín hiệu laser nguồn CW. Các hệ số h_i có thể được lập trình nhờ sự thay đổi điện áp đặt vào graphene ở các ống dẫn sóng của MMI cột thứ 2.



Hình 3.1: Cấu trúc nơ-ron nhân chập mới dùng MMI và vi cộng hưởng

Cấu trúc vi mạch add-drop được áp dụng rộng rãi trong tính toán quang học trên chíp do khả năng xử lý rất đặc biệt. Vì giá trị công suất là không âm, công việc ban đầu chỉ sử dụng cổng "thông qua", khi đó ma trận truyền và vectơ đầu ra là không âm,

do đó hoạt động của ma trận bị giới hạn trong miền số không âm. Tuy nhiên, các phép toán cơ bản như phép nhân ma trận thường được thực hiện trong miền số thực trong thực tế. Để mở rộng phép toán ma trận sang miền số thực đầy đủ, kết quả cuối cùng cần thu được thông qua xử lý vi phân giữa các giá trị công suất của cổng "thả" và cổng "thông qua"; theo cách này, ma trận truyền và vectơ đầu ra cuối cùng đều có thể chứa miền âm. Cường độ đầu ra tại hai cổng của đầu ra trong các phát hiện cân bằng có thể được biểu thị qua công thức:

$$I_{out1} = \alpha cos^2(\frac{\Delta\phi}{2})I_{in} \tag{3.2}$$

$$I_{out2} = \alpha sin^2(\frac{\Delta\phi}{2})I_{in} \tag{3.3}$$

Trong đó α là hệ số suy hao. Kết quả là cường độ sau bộ tách sóng cân bằng là

$$\Delta I = I_{out1} - I_{out2} = \alpha I_{in} |cos(\Delta\phi)|;$$

Với $\Delta\phi$ là sự sau khác pha giữa 2 tín hiệu. Dùng phương pháp này cả hệ số âm và dương đều được thực hiện.

Trong xử lý ảnh kỹ thuật số, tích chập của một ảnh X với nhân *h* tạo ra một ảnh chập O. Một ảnh được biểu diễn dưới dạng ma trận các số với chiều $M \times N$, trong đó M và N lần lượt là chiều cao và chiều rộng của hình ảnh. Mỗi phần tử của ma trận đại diện cho cường độ của một pixel tại vị trí không gian cụ thể đó. Bộ lọc nhân là một ma trận gồm các số dương hoặc số âm với kích thước $R \times R$. Giá trị của một pixel đối ngẫu cụ thể được xác định bởi:

$$y = (x_o, x_1, \ldots x_R)(h_o, h_1, \ldots h_R)^T = \sum_{k=1}^{R}\sum_{l=1}^{R} x_{k,l}h_{k,l} = \sum_{k=1}^{R}\frac{\eta P}{R}T(x_i)\,T(h_i) \tag{3.4}$$

Trong đó P là công suất ra của mạch cân bằng, T là hàm truyền, η là hiệu suất của photodiode. Ma trận MVM được mô tả bằng phương trình toán học sau:

$$y = hX = \begin{bmatrix} h_{11} & h_{12} & \ldots & h_{1R} \\ h_{21} & h_{22} & \ldots & h_{2R} \\ \ldots & \ldots & \ldots & \ldots \\ h_{R1} & h_{R2} & \ldots & h_{RR} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_R \end{bmatrix} \tag{3.5}$$

Một bộ cộng hưởng vi mạch quang học mới chỉ dựa trên một ống dẫn sóng đa mode với bốn cổng được thể hiện trong Hình 3.2. Luận án sử dụng ống dẫn sóng Si_3N_4 với chiều rộng 1600nm và chiều cao 180nm cho ống dẫn sóng đầu vào và đầu ra. Đối với ống dẫn sóng đa mode, tác giả sử dụng chiều rộng rộng hơn. Trong cấu trúc này,

tác giả sử dụng một ống dẫn sóng phản hồi cho ống dẫn sóng vòng và tạo thành bộ cộng hưởng vi mạch bổ sung. Việc thả và thông qua cổng T_p và T_d được thể hiện trong Hình 3.2. Trong cấu trúc này tác giả sử dụng chiều dài MMI là $L_{MMI} = 1.5L_\pi$.

Trong thiết kế này, tác giả thiết kế tại bước sóng 1550nm để phù hợp với các cấu trúc điều chế graphene. Ảnh quang được số hóa và chuyển đổi thành tín hiệu quang qua bộ điều chế như chỉ ra ở Hình 3.1.

Hình 3.2: Cấu trúc vi cộng hưởng dùng MMI

Tín hiệu truyền qua cấu trúc cộng hưởng trên được tính ở cổng T_p và T_d theo công thức [96]:

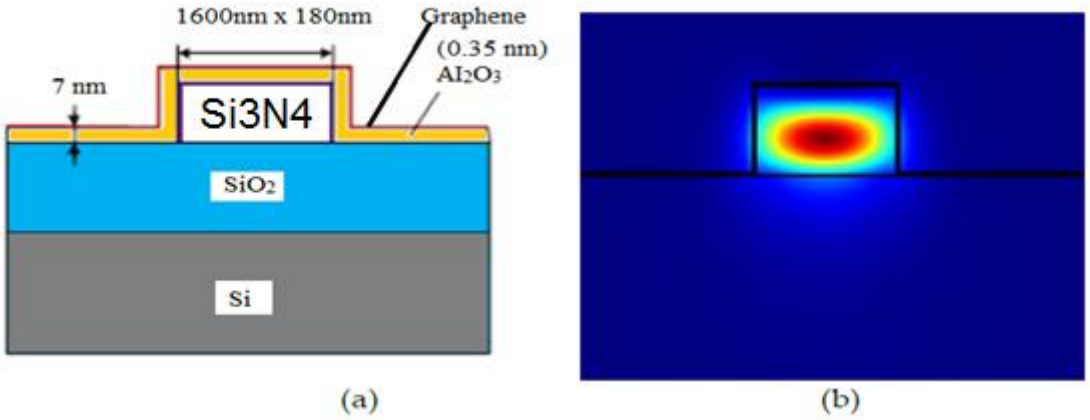
$$T_p = \frac{0.5\alpha^2 - \alpha cos\phi + 0.5}{1 - \alpha cos\phi + (0.5\alpha)^2} = I_{out1} \tag{3.6}$$

$$T_p = \frac{0.25\alpha}{1 - \alpha cos\phi + (0.5\alpha)^2} = I_{out2} \tag{3.7}$$

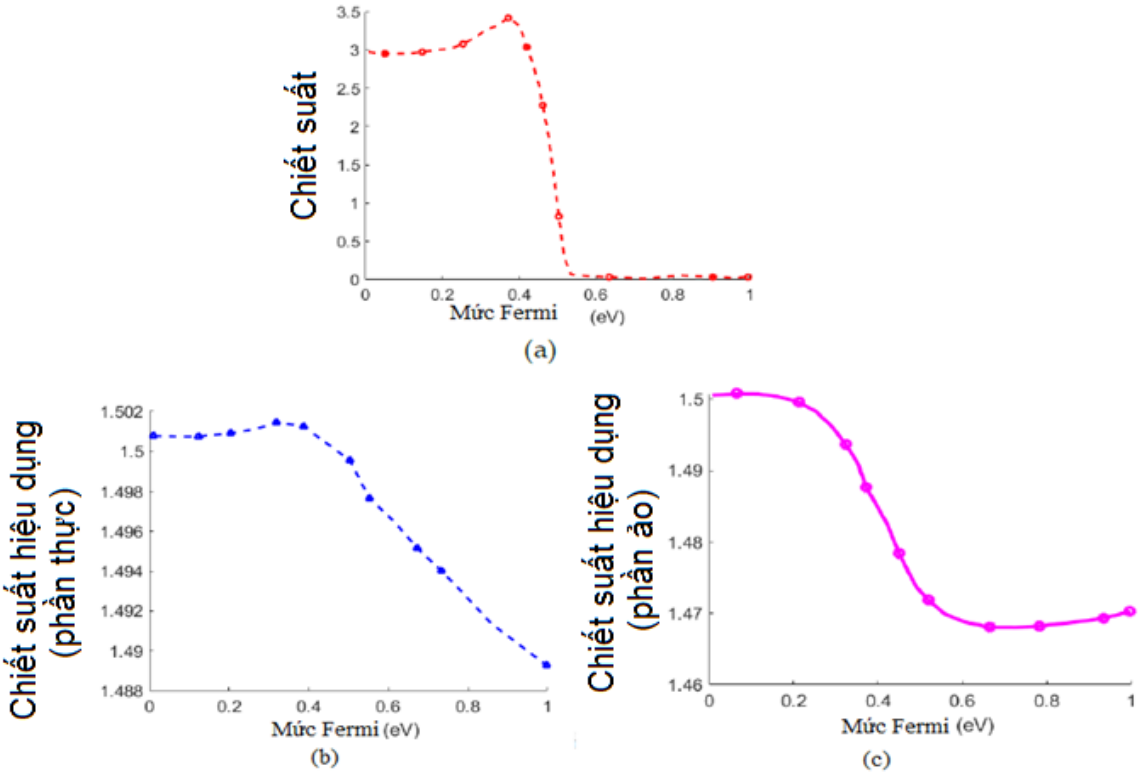
Để thực hiện được các hệ số x_i và h_i , Luận án sử dụng tấm graphene trên ống dẫn sóng Si_3N_4 để điều chế tín hiệu. Graphene được tích hợp với Si_3N_4 tạo thành một cấu trúc mới gọi là cấu trúc Si_3N_4 graphene GSW (Graphene Silicon Nitride Waveguide). Chiều dài của graphene tương ứng với chiều dài của bộ di pha.

3.1.2. Kết quả mô phỏng, đánh giá

Cấu trúc bộ di pha và điều chế tín hiệu được chỉ ra ở Hình 3.3. Sự xuất hiện graphene sẽ tạo ra sự thay đổi đặc tính truyền tín hiệu qua ống dẫn sóng. Bằng cách thay đổi điện áp vào graphene với một điện áp phù hợp V_g điện thế hóa của graphene thay đổi. Phần thực và phần ảo của chiết suất graphene phụ thuộc vào thế hóa trị được mô phỏng ở Hình 3.3. Hình 3.3(a) là cấu trúc của ống dẫn sóng với lớp phủ graphene, Hình 3.3(b) là mode tín hiệu của ống dẫn sóng Si3N4. Khi cung cấp 1 điện áp Vg vào lớp graphene sẽ làm thay đổi chiết suất của ống dẫn sóng được thể hiện ở Hình 3.4. Hình 3.4(a) là chiết suất của graphene theo mức điện áp Vg hay mức Fermi. Hình 3.4(b) và (c) là chiết suất hiệu dụng (phần thực và phần ảo).



Hình 3.3: Điều khiển dùng graphene mode trong ống dẫn sóng



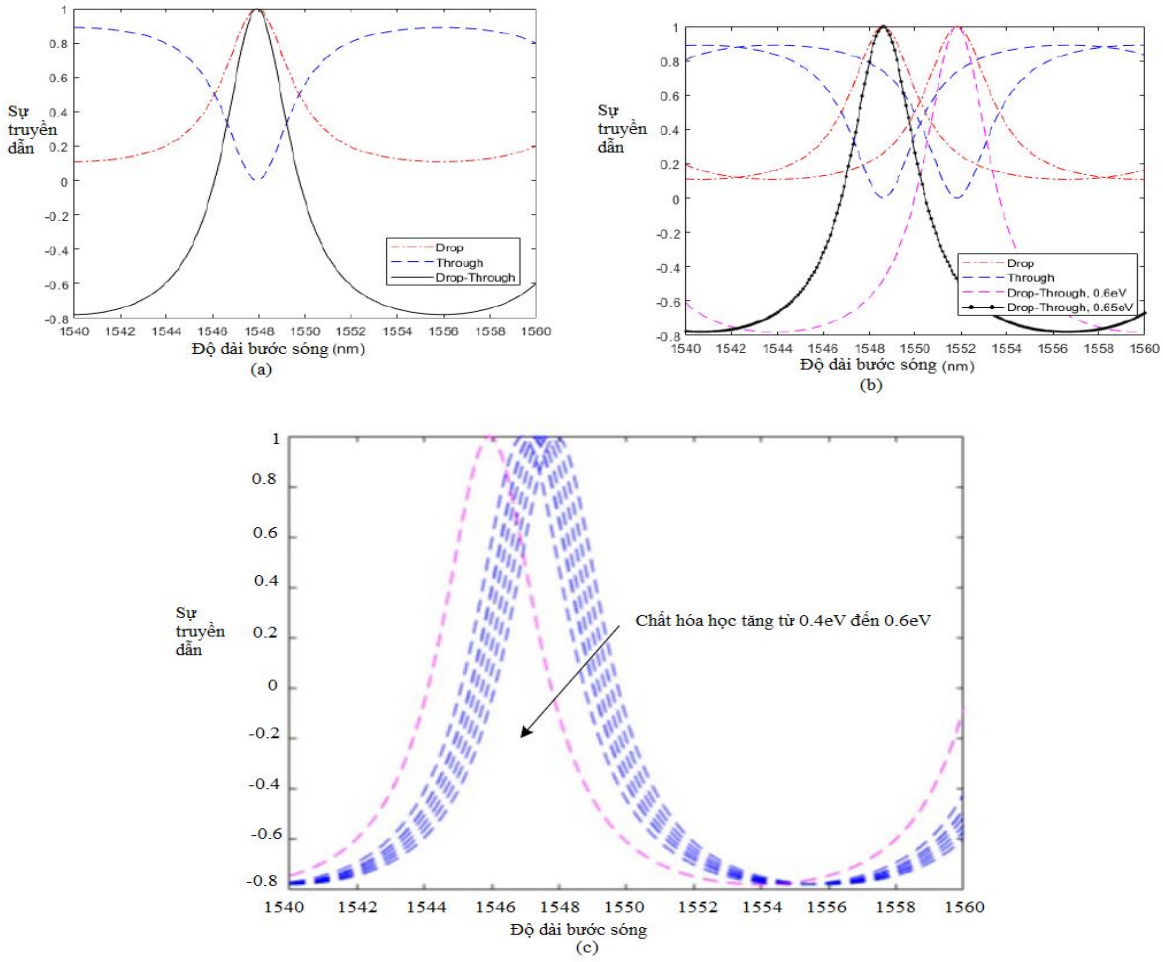
Hình 3.4: Chiết suất của graphene và chiết suất hiệu dụng theo V_g

Graphene được mô tả theo lý thuyết Kubo về độ dẫn trong băng và ngoài băng theo công thức [97]:

$$\delta(\omega) = \sigma_{intra}(\omega) + \sigma_{inter}(\omega) \tag{3.8}$$

Công suất chuẩn hóa tại các cổng T_p và T_d của bộ cộng hưởng dựa trên MMI được thể hiện trong Hình 3.5. Công suất chênh lệch (T_p-T_d) giữa hai cổng nằm trong

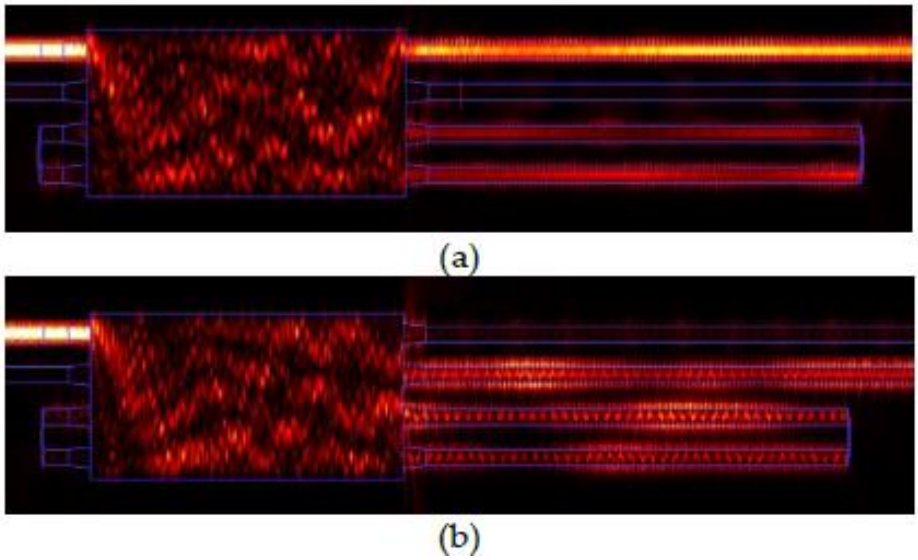
khoảng (-1, +1) đối với các giá trị âm của bộ lọc nhân kernel. Trong mô phỏng này, điện thế hóa học tại graphene là 0,6eV. Bằng cách điều khiển điện áp vào graphene, chúng ta có thể điều khiển được đặc tính truyền dẫn tại T_p và T_d, từ đó có thể đạt được các hệ số trọng số bộ lọc tương ứng. Do graphene cho phép tốc độ điều chế cao, từ 30-80GHz (trong khi hệ thống máy tính hiện tại khoảng 3-5GHz), nên cho phép xử lý dữ liệu với tốc độ cao. Trong cấu trúc của Luận án này đề xuất, tính toán cho thấy tốc độ điều chế đạt đến 28GHz, tức gấp 5-6 lần so với hệ thống máy tính hiện nay. Bởi giá trị của thế hóa trị được tính theo công thức [97]: $\left|\mu_c(V_g)\right| = hV_f\sqrt{\pi\eta|V_g - V_0|}$.



Hình 3.5: Hàm T_p và T_d dùng cho hệ số trọng số và tín hiệu

Kết quả mô phỏng số cho tín hiệu lan truyền qua bộ cộng hưởng dựa trên MMI với tín hiệu đầu vào tại cổng 1 được thể hiện trong Hình 3.6. Kết quả cho thấy sự truyền tín hiệu đối với cộng hưởng (ON) và Hình 6 (b) cho thấy sự truyền tín hiệu cho tắt cộng hưởng (OFF). Trong nghiên cứu này, bằng cách thiết kế độ dài của ống dẫn sóng phản hồi L_r tương ứng, chúng ta có thể đạt được sự thay đổi cộng hưởng cơ bản.

Sau đó, bằng cách thay đổi điện thế hóa học thông qua điện áp đặt trên tấm graphene, chúng ta có thể thu được sự thay đổi cộng hưởng mong muốn. Bước sóng cộng hưởng thu được ở điều kiện cộng hưởng $m\lambda_r = n_{eff}L_r$, với m là các số nguyên.



Hình 3.6: Tín hiệu ảnh truyền qua vi cộng hưởng ở ON và OFF

3.2. Tách biên ảnh sử dụng nơ-ron quang tử

Trong phần này, OVMM được thiết kế ở trên được ứng dụng để tách biên ảnh trong miền quang. Các hệ số bộ lọc được thiết kế qua điều chỉnh các điện áp trên graphene. Kết quả mô phỏng tín hiệu ảnh truyền qua hệ thống với tín hiệu các mức xám x_1,..., x_4 giữ nguyên và thay đổi hệ số bộ lọc h_i được thể hiện ở Hình 3.7. Tiếp theo hệ thống trên được thiết kế để thực hiện các thuật toán tách biên ảnh dùng mặt nạ Roberts, Sobel và Prewitt.

Gradient của ảnh f tại vị trí (x,y) được xác định theo công thức [51]:

$$\Delta f = grad(f) = \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{3.9}$$

Gradient ∇f là vector thể hiện hướng thay đổi cực đại của ảnh f tại vị trí (x,y). Kích thước của gradient f được tính theo công thức [51]:

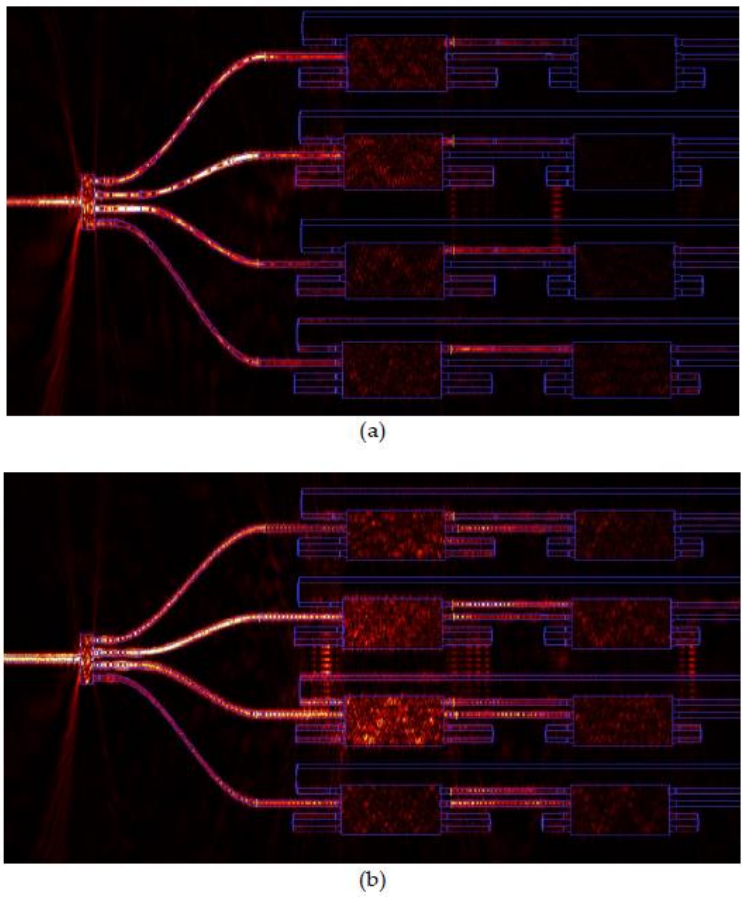
$$M(x,y) = mag(\nabla f) = \sqrt{g_x^2 + g_y^2} \tag{3.10}$$

Ở đây hệ số toán tử Sobel được thể hiện nhỏ hơn 1 với ma trận hệ số kernel là:

$$H_x = \frac{1}{2}\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, H_y = \frac{1}{2}\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad (3.11)$$

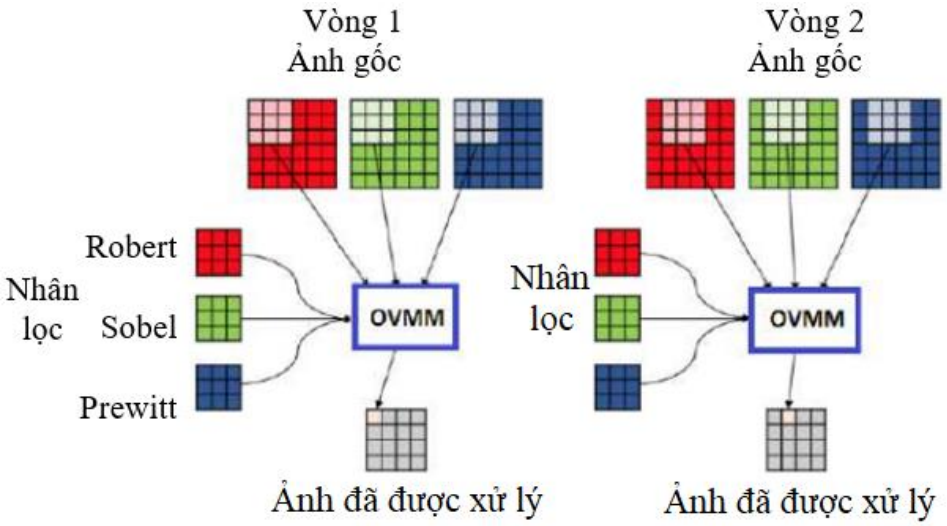
Toán tử Prewitt có các hệ số kernel:

$$H_x = \frac{1}{2}\begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}, H_y = \frac{1}{2}\begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad (3.12)$$



Hình 3.7: Tín hiệu mức xám ảnh truyền qua hệ thống

Thuật toán tách biên được thực hiện như sau: Bộ lọc nhân *h* sẽ chạy từ trái sang phải, từ trên xuống dưới của hình ảnh đầu vào như trong Hình 3.8 để thực hiện tích chập của toàn bộ hình ảnh. Một ưu điểm nổi trội của cấu trúc đề xuất là không cần thay đổi phần cứng để thực hiện đồng thời 3 toán tử lọc biên ảnh tốc độ cao.



Hình 3.8: Thuật toán tách biên ảnh dùng cùng một phần cứng OVMM

Thuật toán thực hiện nhân chập và tách biên trên Python được chỉ ra ở Hình 3.9.

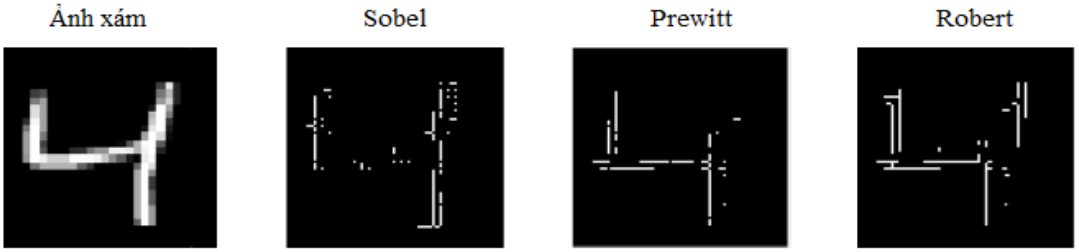
```
Algorithm – Convolutions for OONN using MMI resonators

function convolve(D, F, R, O, S, H, W) do
for (k = 0; k < K; k = k + 1) do
load kernel weights from F[:, :, :, k]
for (h = 0; h < H – R + 1; h = h + S) do
for (w = 0; w < W – R + 1; w = w + S) do
load inputs from D[h:min(h+R, H), w:min(w+R, W), :]
perform convolution
store results in O[h/S, w/S, k]
end
end
end
end
```

Hình 3.9: Thuật toán tách biên ảnh dùng quang

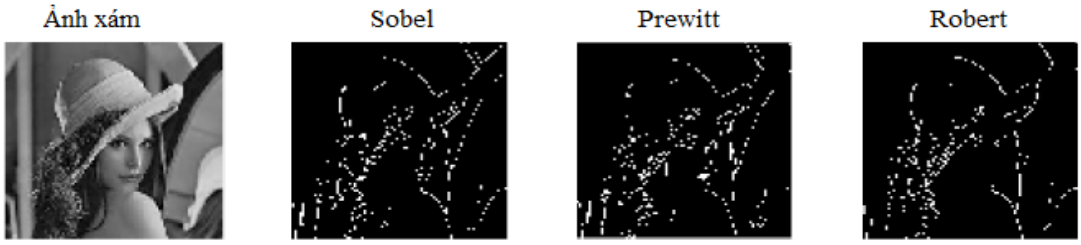
Kết quả thực hiện thuật toán tách biên ảnh đối với tập dữ liệu MNIST và Lena được chỉ ra ở Hình 3.10. Kết quả mô phỏng ở lớp trên cùng dùng môi trường Python, sau đó được đánh giá với kết quả tách biên ảnh sử dụng thư viện NumPy và Scipy như chỉ ra ở Hình 3.11. NumPy là một mô-đun mã nguồn mở cho Python cung cấp các phép toán toán học và số học tổng quát dưới dạng các chức năng được biên dịch sẵn và nhanh chóng. Chúng được kết hợp vào các gói cấp cao. Chúng cung cấp các chức năng có thể được so sánh với MatLab. NumPy (Numeric Python) cung cấp các phương pháp cơ bản để xử lý các mảng lớn và ma trận. SciPy (Scientific Python) với một bộ sưu tập lớn các thuật toán hữu ích như tối thiểu hóa, biến đổi Fourier, hồi quy và các kỹ thuật

toán học khác. Kết quả cho thấy sai số giữa OVMM và Scipy từ 0.05-0.12. Ở đây Luận án sử dụng sai lệc MSE giữa hai thuật toán qua phương trình:

$$\Delta MSE = \left| \frac{1}{MxN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (x(i,j) - y_{OVMM}(i,j))^2 \right| \quad (3.13)$$

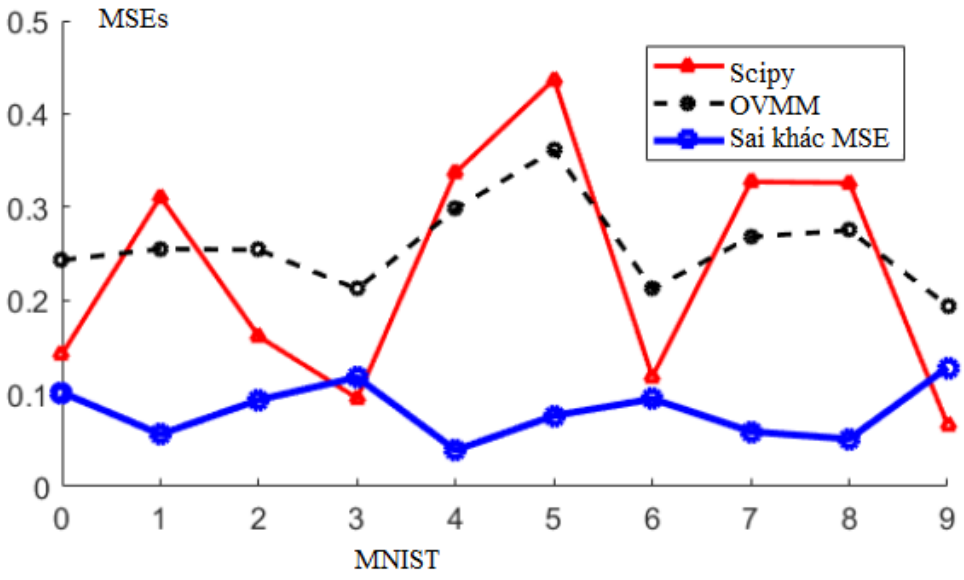


(a) Phát hiện biên, tập dữ liệu MNIST



(b) Phát hiện biên, ảnh Lena

Hình 3.10: Kết quả đánh giá tách biên ảnh sử dụng OVMM



Hình 3.11: Đánh giá sai số MSE, so sánh OVMM và Scipy

Như vậy, bộ xử lý quang học mới thực hiện mạch nhân ma trận vectơ quang (OVMM) sử dụng cấu trúc giao thoa đa MMI đã được đề xuất trong nghiên cứu này. Tốc độ cao 28GHz có được bằng cách sử dụng graphene trên ống dẫn sóng Si_3N_4. Cấu trúc đề xuất chỉ sử dụng bộ ghép MMI với ống dẫn sóng phản hồi. Ưu điểm của cấu trúc đề xuất là không cần các phần tử WDM, dung sai chế tạo cao để tính toán chính xác cao, nhỏ gọn, tổn hao thấp. Kiến trúc mới có thể được áp dụng cho các mạng nơ-ron quang với thứ tự kết nối cao của các nơ-ron cho mạng nơ-ron nhân tạo với suy hao thấp và băng thông cực rộng. Tác giả chứng minh sự tách biệt cạnh bằng cách sử dụng các toán tử Roberts, Prewitt và Sobel với sự khác biệt của MSE với Scipy theo thứ tự 0,1.

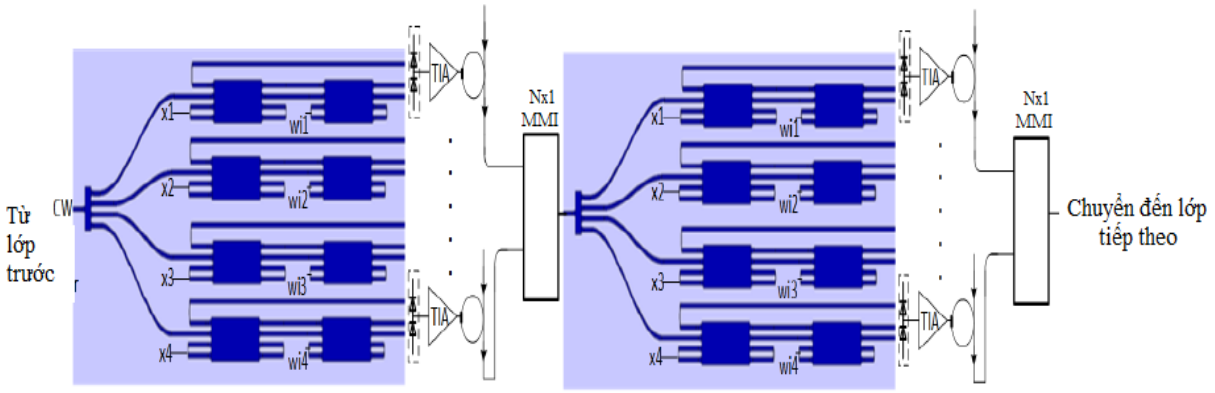
3.3. Thiết kế mạng nơ-ron quang tử ứng dụng cho nhận dạng ảnh

3.3.1. Nguyên lý

Có thể xem OVMM như một neuron trong mạng nơ-ron quang. Nếu thay các kernel h_i bằng các hệ số w_{ij} thì ta có thể biểu diễn tín hiệu ra y_1 ở neuron lớp thứ nhất trong sơ đồ Hình 3.1 bằng biểu thức:

$$y_1 = x_1 w_{11} + x_2 w_{12} + x_3 w_{13} + x_4 w_{14} \quad (3.14)$$

Hình 3.12 đề xuất 1 cấu trúc thực hiện mạng nơ-ron quang sử dụng các OVMM kết nối nhiều lớp. Mạng này áp dụng cho các bài toán phi tuyến như nhận dạng, phân loại ảnh. Đặc điểm của cấu trúc này là có khả năng tích hợp trên chip đơn nhờ các cấu trúc tích hợp, ta gọi là mạng nơ-ron quang tử tích hợp chip OONN đề xuất được huấn luyện bằng cách thay đổi các giá trị của nhân lõi, tương tự như cách mạng nơ-ron dữ liệu chuyển tiếp được huấn luyện bằng cách thay đổi các kết nối có trọng số. Giá trị nhân và trọng số ước tính được yêu cầu trong giai đoạn thử nghiệm. Tín hiệu đầu vào được mã hóa bằng cách sử dụng mảng cột đầu tiên của bộ cộng hưởng vi mạch dựa trên MMI. Chỉ sử dụng cấu trúc này, bất kỳ bộ lọc nào cũng có thể được tạo ra bằng cách thay đổi hệ số trọng lượng thông qua việc kiểm soát bước sóng cộng hưởng như được trình bày trong phần tiếp theo.



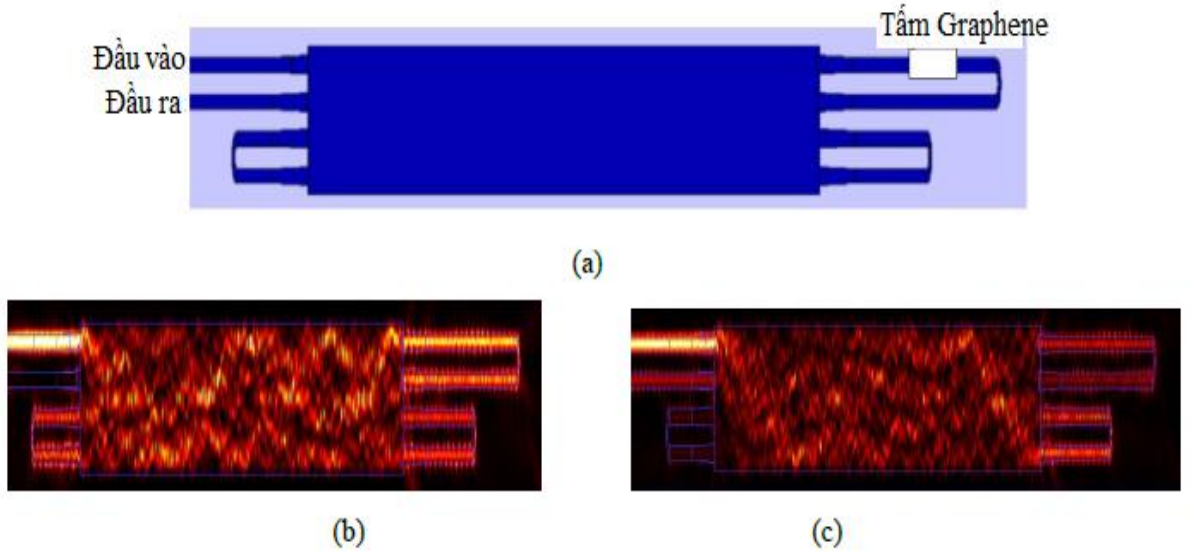
Hình 3.12: Cấu trúc mạng nơ-ron quang nhân chập dùng neuron OVMM

Qua tính toán các phương trình ở phần trên, ta có được tín hiệu tại cổng T_d và T_p được biểu diễn dưới dạng ma trận như sau, với nhân có kích cỡ NxN:

$$Y_d = T_d = (y_{d1}, y_{d2}, \dots, y_{dN})^T = WX = \begin{bmatrix} w_{11} & w_{12} & \dots & h_{1N} \\ w_{21} & w_{22} & \dots & h_{2N} \\ \dots & \dots & \dots & \dots \\ w_{N1} & h_{N2} & \dots & h_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_N \end{bmatrix} \quad (3.15)$$

$$Y_p = T_p = (y_{p1}, y_{p2}, \dots, y_{pN})^T = \begin{bmatrix} 1-w_{11} & 1-w_{12} & \dots & 1-h_{1N} \\ 1-w_{21} & 1-w_{22} & \dots & 1-h_{2N} \\ \dots & \dots & \dots & \dots \\ 1-w_{N1} & 1-h_{N2} & \dots & 1-h_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_N \end{bmatrix} \quad (3.16)$$

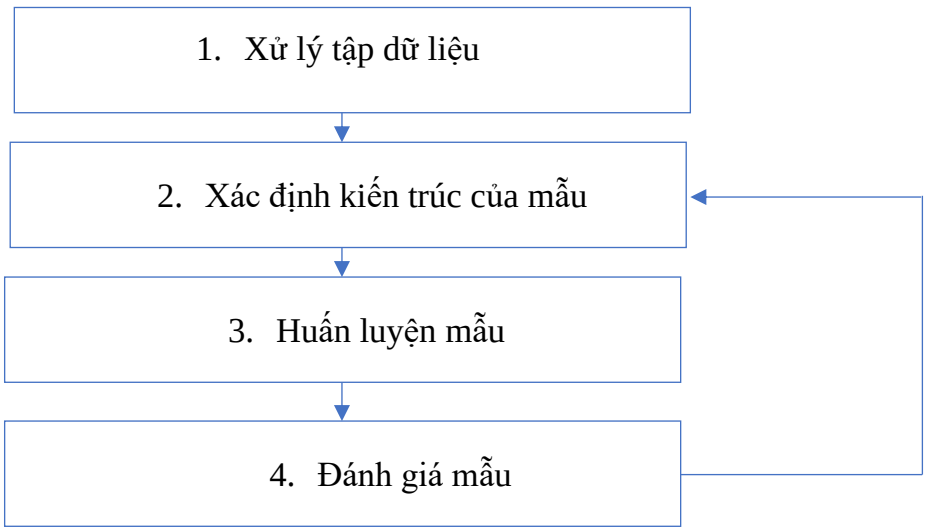
Đặc biệt, trong nghiên cứu này tác giả đề xuất sử dụng cấu trúc vi cộng hưởng MMI làm bộ điều chế sau bộ tách sóng cân bằng với vật liệu graphene để tăng tốc độ xử lý thông tin. Cấu trúc dựa trên vi cộng hưởng MMI được chỉ ra ở Hình 3.13. Cấu trúc này cho phép giảm kích thước, có khả năng tích hợp, điều khiển chính xác và tốc độ cao. Đặc biệt do sử dụng ống dẫn sóng phản hồi với kích thước nhỏ nên băng thông của hệ thống cao, cho phép xử lý dữ liệu với băng thông lớn.



Hình 3.13: Bộ điều chế mới sử dụng vi cộng hưởng MMI

3.3.2. Kết quả mô phỏng

Trong phần này, tác giả trình bày phương pháp thực hiện mô hình AI sử dụng cấu trúc OONN đề xuất. Phương pháp xử lý dữ liệu được làm tương tự như trình bày trong [98], để xây dựng một mạng nơ-ron sâu cho học có giám sát. Sơ đồ tổng thể được thể hiện ở Hình 3.14.



Hình 3.14: Sơ đồ thực nghiệm tổng quát

Sau khi xử lý tập dữ liệu, các kiến trúc ANN khác nhau có thể được ưu tiên, chẳng hạn như RNN để xử lý dữ liệu tuần tự [44]. Giá trị gia tăng của ONN là xử lý lớn cơ sở dữ liệu nhanh chóng, có khả năng tiêu thụ ít năng lượng hơn so với một máy

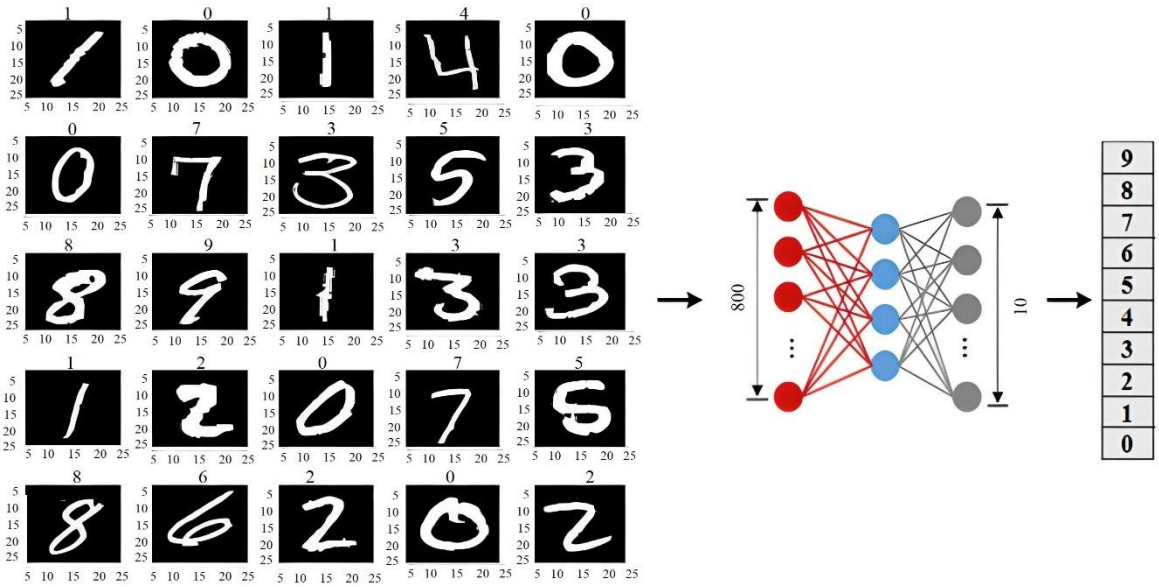
tính cổ điển. Do đó, nó có thể thú vị để xem nó có thể thực hiện như thế nào đối với một tác vụ cần được chạy trong thời gian thực và quản lý một lượng lớn dữ liệu. Trong luận án này, tác giả tập trung vào khả năng sử dụng ONN cho nhiệm vụ phân loại hình ảnh, vì chúng thường yêu cầu phản hồi nhanh và xử lý lượng dữ liệu ngày càng tăng.

Dữ liệu thử nghiệm: dữ liệu MNIST bao gồm dữ liệu viết tay từ 250 người, mỗi bức ảnh có *28×28* pixels tương ứng với 60.000 ảnh đào tạo và 10.000 ảnh kiểm thử nghiệm. MNIST được sử dụng rộng rãi để đánh giá hiệu suất tổng thể của kiến trúc mạng nơ-ron. Kết quả tốt nhất trên MNIST do CNN đạt được với độ chính xác là 99,77% [99].

Luận án đã áp dụng OONN được đề xuất để thực hiện nhận dạng hình ảnh trên tập dữ liệu MNIST. Các tham số được tối ưu hóa để giải MNIST có thể được phân loại thành hai nhóm [100]: hai nhân *5×5×8* khác nhau và hai lớp được kết nối đầy đủ có kích thước *800×1* và *10×1* như được trình bày trong Hình 10 (a). Tác giả sử dụng bộ lọc hạt nhân *5×5* để mô phỏng

$$W = \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} & w_{15} \\ w_{21} & w_{22} & w_{23} & w_{24} & w_{25} \\ w_{31} & w_{32} & w_{33} & w_{34} & w_{35} \\ w_{41} & w_{42} & w_{43} & w_{44} & w_{45} \\ w_{51} & w_{52} & w_{53} & w_{54} & w_{55} \end{bmatrix} \tag{3.17}$$

Sơ đồ nhận dạng tập dữ liệu viết tay MNIST được mô hình trên Hình 3.15.



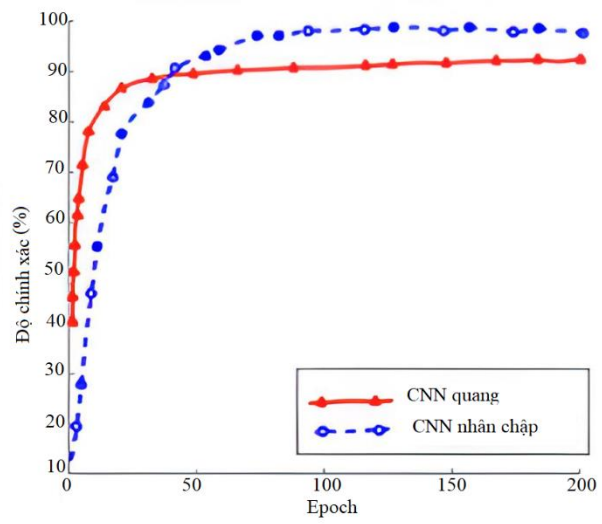
Hình 3.15: Sơ đồ thực hiện nhận dạng chữ viết tay

Thuật toán thực hiện xử lý ảnh, nhận dạng MNIST dùng nhân quang được đề xuất thực hiện trên Python được chỉ ra ở Hình 3.16.

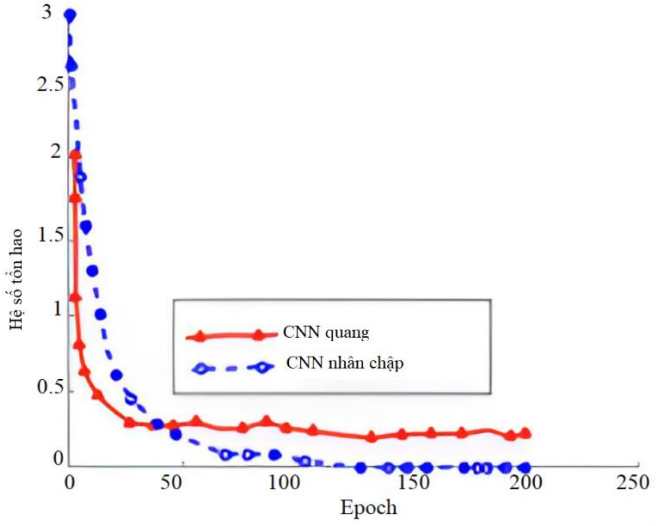
Thuật toán 2: Nhân chập dùng cho xử lý ảnh sử dụng cấu trúc đã đề xuất

```
function convolve(D, F, R, O, N, H, W, Hm, Wm) do
load kernel weights from F[:, :, 0]
let Hout = H - R + 1
let Wout = W - R = 1
for (n = 0; n < N; n = n + 1) do
for (h = 0; h < H - R + 1; h = h + Hm - R + 1) do
for (w = 0; w < W - R + 1; w = w + Wm - R + 1) do
load inputs from
D[h:min(h + Hm, H), w:min(w + Wm, W), 0, n]
perform convolution
store results in
O[h:min(h + Hm, H) - R + 1,
w:min(w + Wm, W) - R + 1, 0, n]
end
end
end
end
```

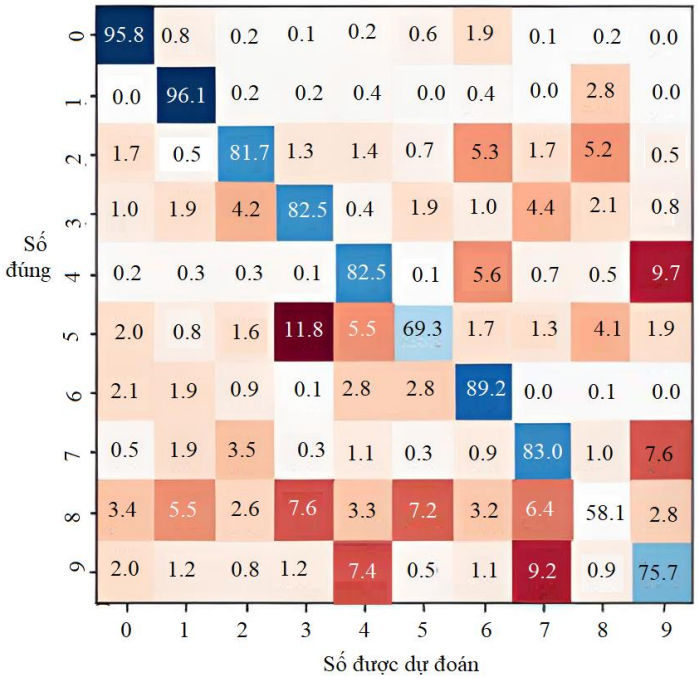
Hình 3.16: Thuật toán xử lý ảnh dùng cấu trúc quang MMI trên Python



(a)



(b)



(c)

Hình 3.17: So sánh độ chính xác và hệ số tổn hao

Bằng cách xếp tầng các bộ cộng hưởng vi vòng dựa trên 4x4 MMI để triển khai mạng nơ-ron như trong Hình 3.12, ví dụ như với 100 bộ cộng hưởng vi vòng có cùng bán kính 5µm, thời gian truyền qua nơ-ron là:

$$t_p = \frac{N(2\pi R)}{c} \tag{3.18}$$

Trong đó c là tốc độ ánh sáng, R là bán kính của ống dẫn sóng vòng, N là số lượng bộ cộng hưởng microring. Khi N=100, thời gian lan truyền là và thông lượng là $t_p = 11ps$. Bộ kích hoạt vi vòng được bao phủ bởi tấm graphene có thể được điều chế ở tốc độ 130 GS/s, nghĩa là tần số điều chế của MRR không gây tắc nghẽn thông lượng của nơ-ron.

Tiếp theo, Luận án so sánh hiệu suất của cấu trúc được đề xuất với DeepBench. DeepBench là một tập dữ liệu chứa các loại GPU khác nhau mất bao lâu để thực hiện phép tích chập cho một tập tham số tích chập nhất định. Mức sử dụng năng lượng của hai trong số các điểm chuẩn tích chập cho GPU từ bộ dữ liệu DeepBench là: AMD MI25 với 300W, Nvidia GTX 1080Ti với 250W [101]. Bằng cách sử dụng cấu trúc được đề xuất, đơn vị biến đổi có thể tạo ra một pixel của hình ảnh trong 100ps để thực

hiện phép tích chập với K bộ lọc sử dụng hai pixel trên mỗi vòng tròn. Tốc độ của tích chập có thể được ước tính bằng:

$$t_{runtime} = 50ps \times K(H - R + 1)(W - R + 1) \quad (3.19)$$

Trong đó R là chiều dài cạnh của kernel không có phần đệm, H và W là chiều cao và chiều rộng của hình ảnh đầu vào. Mức tiêu thụ điện năng của tổ hợp được đề xuất có thể được ước tính vào khoảng 110W so với mức tiêu thụ điện năng trung bình của GPU là 295W. Ngoài ra, tốc độ của tích chập được đề xuất nhanh hơn từ 2,8 đến 14 lần so với thời gian chạy GPU trung bình.

Trong mô phỏng này, tác giả sử dụng OONN với hai lớp và chức năng kích hoạt phi tuyến ReLU được sử dụng. Kết quả của nhiệm vụ MNIST được giải quyết bởi OONN được thể hiện trong Hình 3.17. Độ chính xác tổng thể là 92,4% thu được sau 10 lần tương tác. Tác giả cũng so sánh kết quả với CNN thông thường bao gồm hai lớp với độ chính xác là 99,2% sau 50 lần tương tác. Mặc dù độ chính xác của OONN được đề xuất thấp hơn CNN dùng hệ thống máy tính hiện nay, cấu trúc được đề xuất nhanh hơn 5 lần và yêu cầu mức tiêu thụ điện năng thấp hơn. Điều này phù hợp để kết nối cao hơn với nhiều lớp trong các ứng dụng phức tạp khác. Ngoài ra, CNN hiện tại sử dụng dấu phẩy động 32 bit trong khi cấu trúc đề xuất chỉ sử dụng độ chính xác 6 bit. Độ chính xác có thể được cải thiện nhiều hơn nếu chúng ta sử dụng độ chính xác bit cao hơn như 9 bit vừa được thiết kế gần đây trong miền toàn quang [102].

3.4. Kết luận Chương 3

Chương 3 đã thiết kế thành công cấu trúc nơ-ron quang tích hợp trên 1 chip đơn có khả năng tính toán tốc độ cao gấp 5 lần so với các cấu trúc trước đây. Cấu trúc mới được sử dụng thử nghiệm với tách biên ảnh dùng toán tử Roberts, Sobel, Prewitt có sai số MSE so với dùng Scipy khoảng 0.05-0.12. Chương 3 cũng trình bày một cấu trúc mạng nơ-ron quang tích hợp mới (OONN) có khả năng tính toán tốc độ cao, nhỏ gọn. Kết quả được so sánh, mô phỏng để nhận dạng tập dữ liệu chữ viết tay MNIST. Dù sử dụng số bit thấp hơn nhưng tốc độ tính toán cao hơn máy tính truyền thống và khả năng nhận dạng chính xác thấp hơn 10% so với truyền thống. OONN ước tính sẽ thực hiện các phép chập nhanh hơn GPU từ 2,8 đến 14 lần trong khi sử dụng gần như mức tiêu thụ điện năng thấp hơn. Ngoài ra, cấu trúc được đề xuất có thể xử lý cả số dương và số âm cũng như các công việc phức tạp hơn cho các ứng dụng sắp tới sử dụng OONN được đề xuất.Các kết quả có liên quan đến Chương 3 được công bố trong các công trình [J1, J6, J7] và [C3].

KẾT LUẬN

I. Những kết quả của Luận án

Luận án đã nghiên cứu, thiết kế thành công bộ biến đổi DHT, DCT và KLT trong miền quang, ứng dụng cho xử lý ảnh tốc độ cao. Đồng thời Luận án đã đề xuất và thiết kế cấu trúc nơ-ron mới toàn quang có khả năng tính toán tích chập trong miền quang tốc độ cao. Từ đó ứng dụng cho tách biên ảnh sử dụng toán tử Roberts, Prewitt và Sobel trong miền quang. Luận án đã đề xuất và thiết kế thành công mạng nơ-ron quang tử và thử nghiệm cho phân loại dữ liệu ảnh. Các kết quả Luận án là nghiên cứu liên ngành hướng đến thiết kế các hệ thống tính toán, máy tính toàn quang trong tương lai không xa.

1. Thiết kế được các bộ biến đổi toàn quang DHT, DCT, KLT ứng dụng cho nén ảnh

Xử lý ảnh trong miền quang trước đây được thực hiện thông qua các hệ thống thấu kính, Fourier quang và sợi quang [38]. Từ năm 2013 lần đầu tiên xử lý ảnh trong miền quang được thực hiện trong cấu trúc quang tích hợp sử dụng ống dẫn sóng quang trên vật liệu polymer [5, 3], trong đó hệ thống xử lý ảnh được thiết kế dựa vào cấu trúc giao thoa đa mode kết hợp với bộ ghép có hướng. Nhược điểm của các phương pháp này là kích thước lớn, cần ghép nhiều bộ cấu trúc có hướng với nhau nên suy hao lớn. Đồng thời để đạt độ chính xác của bộ ghép cần giải pháp chế tạo chính xác. Băng thông hay tốc độ dữ liệu cũng bị hạn chế khi sử dụng cấu trúc ghép có hướng vì hệ số ghép thay đổi nhanh khi thay đổi bước sóng hoạt động, đặc biệt hoạt động trong dải bước sóng của ảnh màu RGB.

Luận án đã thiết kế, phân tích kỹ thuật nén ảnh sử dụng các biến đổi DHT, DCT và KLT chỉ sử dụng cấu trúc giao thoa đa mode MMI. Ưu điểm giải pháp mới này là có khả năng tích hợp toàn bộ hệ thống trên một vi mạch đơn chiếc, có khả năng tích hợp với các hệ thống xử lý thông tin trong các node cảm biến, máy tính với hệ điều hảnh nhỏ gọn, tiêu thụ ít năng lượng và yêu cầu tài nguyên thấp. Bên cạnh đó, các cấu trúc do Luận án đề xuất có ưu điểm thực hiện chính xác các phép biến đổi mà với sai số chế tạo cho phép lớn đến ±18μm, băng thông và tốc độ dữ liệu xử lý cao. Cấu trúc mới có khả năng tích hợp với hệ thống camera thông minh, xử lý dữ liệu tốc độ cao, băng thông lớn, thời gian thực. Các cấu trúc đề xuất được thiết kế đơn giản, có độ chính xác cao so với công nghệ vi mạch hiện tại.

2. Thiết kế được nơ-ron quang mới, từ đó thiết kế mạng nơ-ron quang ứng dụng cho tách biên ảnh và phân loại ảnh trong miền quang. Cấu trúc mới có khả năng tích hợp, tốc độ cao gấp 5 lần so với hệ thống hiện tại.

Mặc dù mạng nơron quang đã được nghiên cứu từ những năm 1991 [[103], các nghiên cứu trước đây dựa vào quang hình học và các thiết bị sợi quang. Từ năm 2017 [21], thuật toán học sâu lần đầu tiên đã được thực hiện thành công trên cấu trúc vi mạch quang, tạo ra hướng nghiên cứu mới cho thiết kế các hệ thống mạng nơron cho các bài toán học sâu, hồi quy phức tạp [104]. Tuy nhiên, các giải pháp thiết kế mạng nơron quang sử dụng chủ yếu các cấu trúc vi cộng hưởng dựa vào bộ ghép có hướng. Điều này hạn chế xây dựng các mạng nơron có nhiều node mạng với khả năng xử lý các bài toán dữ liệu lớn, cần lưu trữ các giá trị trọng số trung gian trong quá trình học. Bên cạnh đó, rất khó để điều khiển hàng chục nút mạng một lúc với độ chính xác cao nếu sử dụng cấu trúc vi cộng hưởng đó.

Do vậy, Luận án đã đề xuất kiến trúc và thuật toán mới hiết kế mạng nơron sử dụng cấu trúc giao thoa đa mode kết hợp với ống dẫn sóng vòng tạo ra vi cộng hưởng nhỏ gọn, băng thông lớn, tốc độ cao, có thể điều khiển chính xác hệ số bộ lọc kernel tương ứng. Luận án đã thiết kế, mô phỏng, đánh giá các thuật toán tách biên ảnh và nhận dạng chữ viết tay trên cấu trúc mới này. Mặc dù độ chính xác của nhận dạng chưa đạt được như thực hiện qua hệ thống máy tính hiện nay do hạn chế về số bit mã hóa trong miền quang so với 32 và 64 bit, nhưng tốc độ xử lý dữ liệu trong miền quang cao gấp hàng chục lần so với miền điện.

II. Hướng phát triển của Luận án

Trên cơ sở kết quả của Luận án, có một số vấn đề và hướng nghiên cứu mới như:

• Thiết kế hệ thống tích hợp bộ biến đổi ảnh trong miền quang với các bộ nhớ quang trong các hệ thống camera thông minh và xử lý dữ liệu ảnh thời gian thực. Đồng thời thiết kế các hệ thống toàn quang xử lý dữ liệu AR/VR.

• Phát triển mô hình mạng OONN cho các ứng dụng AI thời gian thực, đặc biệt thiết kế các hàm kích hoạt hoàn toàn trong miền quang.

• Cải tiến cấu trúc ống dẫn sóng cấu trúc graphene để tăng tốc độ xử lý dữ liệu và tốc độ học, từ đó thực hiện các bài toán phân tích dữ liệu lớn.

• Nghiên cứu đối sánh giữa xử lý ảnh trong miền quang và miền điện.

DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ CỦA NGHIÊN CỨU SINH

[J1]. Le Trung Thanh, Nguyen Canh Minh, Nguyen Van Khoi, **Bui Thi Thuy**, Nguyen Thi Hong Loan, "*Design of silicon wires based directional couplers for microring resonators*", The University of Danang, Journal Of Science and Technology, No. 12(97), vol. 1, 2015

[J2]. **Thi Thuy Bui**, The Ngoc Dang and Trung Thanh Le, "*All-Optical Karhunen Loeve Transform Using MMI Couplers For Image Processing Applications"*, Tạp chí Khoa học và Công nghệ, Đại học Thái Nguyên, T.227, S.15 (2022), 66-74. DOI: https://doi.org/10.34238/tnu-jst.6360

[J3]. **Thi Thuy Bui**, The Ngoc Dang and Trung Thanh Le, "*On-chip All-optical Haar Transform based on a 4x4 MMI coupler cascaded with a 2x2 MMI coupler for Image Compression*", "*On-chip All-optical Haar Transform based on a 4x4 MMI coupler cascaded with a 2x2 MMI coupler for Image Compression*", Tạp chí Khoa học Máy tính và Kỹ thuật truyền thông, Tạp chí Khoa học Đại học Quốc gia Hà Nội, VNU Journal of Science: Comp. Science & Com. Eng, **Published** Dec 16, 2022, **DOI:** https://doi.org/10.25073/2588-1086/vnucsce.446.

[J4]. **Bui Thi Thuy**, Le Trung Thanh, "*Image Compression in All-Optical Domain Using One 6x6 Multimode Interference Coupler*", Tạp chí Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Received: 6 August 2022; Accepted for publication: 21 September 2022, Vol.61, No.2(2023) : (2) (2023), 347 -357, doi:10.15625/2525-2518/17417 (Scopus)

[J5]. **Thi Thuy Bui**, The Ngoc Dang and Trung Thanh Le, "*Image Compression using All-optical DCT and DST*", Tạp chí Nghiên cứu Khoa học và Công nghệ quân sự, số 82, ngày 28 tháng 10 năm 2022, 159-166, DOI: https://doi.org/10.54939/1859-1043.j.mst.82.2022.159-166

[J6]. **Bui Thi Thuy**, The Ngoc Dang and Le Trung Thanh, "*On-chip Processor based on MMI Microring Resonators for Image Edge Detection in All-optical Domain*", Tạp

chí Khoa học công nghệ Thông tin và Truyền thông, Học viện Công nghệ Bưu chính Viễn thông, ISSN 2525 – 2224, Số 02 (CS.01) 2022, p. 31-37.

[J7]. **Thi Thuy Bui**, Duy Tien Le, Thi Hong Loan Nguyen, Trung Thanh Le, "*On Chip Optical Neural Networks Based on MMI Microring Resonators for Image Classification*", Computer Optics, ISSN 0134-2452(print) ISSN 2412-6179 (online),2023, Issue Vol. 47(4), DOI: 10.18287/2412-6179 (Q1 ISI)

[C1]. **Thi-Thuy Bui**; Trung-Thanh Le, "*Glucose sensor based on 4×4 multimode interference coupler with microring resonators*", 2017 International Conference on Information and Communications (ICIC), Doi: 10.1109/INFOC.2017.8001679, 07 August 2017 (Scopus)

[C2]. **Thi-Thuy Bui**; Trung-Thanh Le, "*Two channel highly sensitive sensors based on 4×4 multimode interference coupler"*, International Conference on Information and Communications (ICIC), Doi: 10.1109/INFOC.2017.8001687, 07 August 2017 (Scopus)

DANH MỤC TÀI LIỆU THAM KHẢO

[1] L. T. Y. Kua Ching, LiBeniamino DiMartino and Q. Zhang, Smart Data Stateof-the-Art Perspectives in Computing and Applications. CRC Press, 2019.

[2] A. Alfalou and C. Brosseau, Recent Advances in Optical Image Processing. Elsevier B.V, 2015.

[3] S. He, R. Wang, and H. Luo, "Computing metasurfaces for all-optical image processing: a brief review," Nanophotonics, vol. 11, no. 6, pp. 1083–1108, 2022. [Online]. Available: https://doi.org/10.1515/nanoph-2021-0823

[4] M. Alemohammad, J. R. Stroud, B. T. Bosworth, and M. A. Foster, "Highspeed all-optical haar wavelet transform for real-time image compression," Opt. Express, vol. 25, no. 9, pp. 9802–9811, May 2017. [Online]. Available: http://opg.optica.org/oe/abstract.cfm?URI=oe-25-9-9802

[5] G. Parca, P. Teixeira, and A. Teixeira, "All-optical image processing and compression based on haar wavelet transform," Appl. Opt., vol. 52, no. 12, pp. 2932–2939, Apr 2013. [Online]. Available: http://opg.optica.org/ao/abstract. cfm?URI=ao-52-12-2932

[6] A. A. Fashi, M. H. V. Samiei, C. Pinho, and A. L. Teixeira, "Photonic integrated chip on triplex platform for realizing optical haar transform and compression in the visible spectrum," IEEE Journal of Quantum Electronics, vol. 57, no. 5, pp. 1–10, 2021.

[7] M. P. Edgar, G. M. Gibson, R. W. Bowman, B. Sun, N. Radwell, K. J. Mitchell, S. S. Welsh, and M. J. Padgett, "Simultaneous real-time visible and infrared video with single-pixel detectors," Scientific Reports, vol. 5, no. 1, p. 10669, 2015. [Online]. Available: https://doi.org/10.1038/srep10669

[8] B. T. Bosworth, J. R. Stroud, D. N. Tran, T. D. Tran, S. Chin, and M. A. Foster, "High-speed flow microscopy using compressed sensing with ultrafast laser 86 pulses," Opt. Express, vol. 23, no. 8, pp. 10 521–10 532, Apr 2015. [Online]. Available: http://opg.optica.org/oe/abstract.cfm?URI=oe-23-8-10521

[9] H. G. Chen, S. Jayasuriya, J. Yang, J. Stephen, S. Sivaramakrishnan, A. Veeraraghavan, and A. Molnar, "Asp vision: Optically computing the first layer of

convolutional neural networks using angle sensitive pixels," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 903– 912.

[10] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljaciˇ c, "Deep learning ´ with coherent nanophotonic circuits," Nature Photonics, vol. 11, no. 7, pp. 441–446, 2017. [Online]. Available: https://doi.org/10.1038/nphoton.2017.93

[11] X. Sui, Q. Wu, J. Liu, Q. Chen, and G. Gu, "A review of optical neural networks," IEEE Access, vol. 8, pp. 70 773–70 783, 2020.

[12] D. Zhang and Z. Tan, "A review of optical neural networks," Applied Sciences, vol. 12, no. 11, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/ 12/11/5338

[13] F. Ashtiani, A. J. Geers, and F. Aflatouni, "An on-chip photonic deep neural network for image classification," Nature, vol. 606, no. 7914, pp. 501–506, 2022. [Online]. Available: https://doi.org/10.1038/s41586-022-04714-0

[14] C. Huang, V. J. Sorger, M. Miscuglio, M. Al-Qadasi, A. Mukherjee, L. Lampe, M. Nichols, A. N. Tait, T. F. de Lima, B. A. Marquez, J. Wang, L. Chrostowski, M. P. Fok, D. Brunner, S. Fan, S. Shekhar, P. R. Prucnal, and B. J. Shastri, "Prospects and applications of photonic neural networks," Advances in Physics: X, vol. 7, no. 1, p. 1981155, 2022. [Online]. Available: https://doi.org/ 10.1080/23746149.2021.1981155

[15] S. Wang, S. Xiang, G. Han, Z. Song, Z. Ren, A. Wen, and Y. Hao, "Photonic associative learning neural network based on vcsels and stdp," J. Lightwave Technol., vol. 38, no. 17, pp. 4691–4698, 2020. [Online]. Available: http://opg. optica.org/jlt/abstract.cfm?URI=jlt-38-17-4691

[16] S. Xu, J. Wang, R. Wang, J. Chen, and W. Zou, "High-accuracy optical convolution unit architecture for convolutional neural networks by cascaded acousto-optical modulator arrays," Opt. Express, vol. 27, no. 14, pp. 19 778–19 787, 2019. [Online]. Available: http://opg.optica.org/oe/abstract. cfm?URI=oe-27-14-19778 87

[17] L. Almeida, N. Kumar, G. Parca, A. Tavares, A. Lopes, and A. Teixeira, "Alloptical image processing based on integrated optics," in 2014 16th International Conference on Transparent Optical Networks (ICTON), 2014, pp. 1–5.

[18] H. T. Gabriel Cristobal, Peter Schelkens, Optical and Digital Image Processing: Fundamentals and Applications. Wiley-VCH, 2011.

[19] G. Parca, P. Teixeira, and A. Teixeira, "3d interferometric integrated passive scheme for all optical transform," in 2012 14th International Conference on Transparent Optical Networks (ICTON), 2012, pp. 1–4.

[20] U. Jayasankar, V. Thirumal, and D. Ponnurangam, "A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications," Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 2, pp. 119–140, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157818301101

[21] S. Sarkar and S. S. Bhairannawar, "Efficient fpga architecture of optimized haar wavelet transform for image and video processing applications," Multidimensional Systems and Signal Processing, vol. 32, no. 2, pp. 821–844, 2021. [Online]. Available: https://doi.org/10.1007/s11045-020-00759-4

[22] C. Pinho, B. Neto, T. Morgado, H. Neto, M. Lima, and A. Teixeira, "Inp aac for data compression applications," IET Optoelectronics, vol. 13, pp. 67–71(4), April 2019. [Online]. Available: https://digital-library.theiet.org/ content/journals/10.1049/iet-opt.2018.5084

[23] A. Azimi Fashi, M. Vadjed Samiei, and A. Teixeira, "Design of a visible light photonic chip for haar transform based optical compression," Optik, vol. 217, p. 164929, 2020. [Online]. Available: https://www.sciencedirect.com/science/ article/pii/S0030402620307658

[24] T.-T. Le and L. Cahill, "Generation of two fano resonances using 4x4 multimode interference structures on silicon waveguides," Optics Communications, vol. 301-302, pp. 100–105, 2013.

[25] L. Cahill and T. Le, "The design of signal processing devices employing soi mmi couplers," in Paper 7220-2, Integrated optoelectronic devices (OPTO 2009), Photonics West, Proceedings of the SPIE, San Jose Convention Center, San Jose, California, USA, 24 - 29 January 2009. [

26] A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight 88

banks," Scientific Reports, vol. 7, no. 1, p. 7430, 2017. [Online]. Available: https://doi.org/10.1038/s41598-017-07754-z

[27] Z. Cheng, C. Ríos, W. H. P. Pernice, C. D. Wright, and H. Bhaskaran, "On-chip photonic synapse," Science Advances, vol. 3, no. 9, p. e1700160, 2017. [Online]. Available: https://www.science.org/doi/abs/10.1126/sciadv.1700160

[28] T. Y. Teo, X. Ma, E. Pastor, H. Wang, J. K. George, J. K. W. Yang, S. Wall, M. Miscuglio, R. E. Simpson, and V. J. Sorger, "Programmable chalcogenide-based all-optical deep neural networks," Nanophotonics, vol. 11, no. 17, pp. 4073–4088, 2022. [Online]. Available: https://doi.org/10.1515/nanoph-2022-0099

[29] Q. Cheng, J. Kwon, M. Glick, M. Bahadori, L. P. Carloni, and K. Bergman, "Silicon photonics codesign for deep learning," Proceedings of the IEEE, vol. 108, no. 8, pp. 1261–1282, 2020. DOI: 10.1109/JPROC.2020.2968184

[30] A. Dumka and A. Ashok, "Advanced Digital ImageProcessing and ItsApplications in Big Data", Taylor and Francis, 2020. https://doi.org/10.1201/9780429351310

[31] R. Gonzalez and R. Woods, Digital Image Processing, 4th. Pearson, 2017.

[32] J. Carpenter, "Holographic Mode Division Multiplexing in Optical Fibres", University of Cambridge, 2012.

[33] A. Kallepalli, J. Innes, and M. J. Padgett, "Compressed sensing in the far-field of the spatial light modulator in high noise conditions," Scientific Reports, vol. 11, no. 1, p. 17460, 2021. [Online]. Available: https://doi.org/10.1038/ s41598-021-97072-2

[34] M. L. Catia Pinho, Isiaka Alimi, Spatial Light Modulation as a Flexible Platform for Optical Systems. Intech Publisher, 2018.

[35] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 7436–7456, 2022.

[36] H. Rhee, Y. I. Jang, S. Kim, and N. I. Cho, "Lossless image compression by joint prediction of pixel and context using duplex neural networks," IEEE Access, vol. 9, pp. 86 632–86 645, 2021.

[37] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach. Prentice Hall, 2009. 89

[38] K. Guo, S. Zeng, J. Yu, Y. Wang, and H. Yang, "[dl] a survey of fpga-based neural network inference accelerators," ACM Trans. Reconfigurable Technol. Syst., vol. 12, no. 1, mar 2019. [Online]. Available: https://doi.org/10.1145/ 3289185

[39] M. S. Akhoon, S. A. Suandi, A. Alshahrani, A.-M. H. Y. Saad, F. R. Albogamy, M. Z. B. Abdullah, and S. A. Loan, "High performance accelerators for deep neural networks: A review," Expert Systems, vol. 39, no. 1, p. e12831, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10. 1111/exsy.12831

[40] C. Qiu, H. Xiao, L. Wang, and Y. Tian, "Recent advances in integrated optical directed logic operations for high performance optical computing: a review," Frontiers of Optoelectronics, vol. 15, no. 1, p. 1, 2022. [Online]. Available: https://doi.org/10.1007/s12200-022-00001-y

[41] T. Chattopadhyay and D. K. Gayen, "Optical half and full adders using the nonlinear mach-zehnder interferometer," Journal of Optics, vol. 50, no. 2, pp. 314–321, 2021. [Online]. Available: https://doi.org/10.1007/ s12596-021-00692-0

[42] M. Hossain, K. E. Zoiros, T. Chattopadhyay, and J. K. Rakshit, "Speed enhancement of all-optical pseudo random binary sequence (prbs) generator using microring resonator," Optical and Quantum Electronics, vol. 53, no. 12, p. 670, 2021. [Online]. Available: https://doi.org/10.1007/s11082-021-03329

[43] T.-T. Le, Multimode Interference Structures for Photonic Signal Processing: Modeling and Design. Lambert Academic Publishing, Germany, ISBN 3838361199, 2010.

[44] Q. Yang, X. Mou, Y. Wang, A. Yan, Y. Yang, and T. ling Ren, "Reconfigurable analog computing architecture based on planar optical waveguides," in Seventh Asia Pacific Conference on Optics Manufacture and 2021 International Forum of Young Scientists on Advanced Optical Manufacturing (APCOM and YSAOM 2021), J. Tan, X. Luo, M. Huang, L. Kong, and D. Zhang, Eds., vol. 12166, International Society for Optics and Photonics. SPIE, 2022, p. 121667Q. [Online]. Available: https://doi.org/10.1117/12.2618002

[45] L. De Marinis, M. Cococcioni, P. Castoldi, and N. Andriolli, "Photonic neural networks: A survey," IEEE Access, vol. 7, pp. 175 827–175 841, 2019. 90

[46] T. F. de Lima, H.-T. Peng, A. N. Tait, M. A. Nahmias, H. B. Miller, B. J. Shastri, and P. R. Prucnal, “Machine learning with neuromorphic photonics,” J. Lightwave Technol., vol. 37, no. 5, pp. 1515–1534, Mar 2019. [Online]. Available: http://opg.optica.org/jlt/abstract.cfm?URI=jlt-37-5-1515

[47] M. Connelly, Semiconductor Optical Amplifiers, 2004.

[48] A. E. Ilesanmi and T. O. Ilesanmi, “Methods for image denoising using convolutional neural network: a review,” Complex and Intelligent Systems, vol. 7, no. 5, pp. 2179–2198, 2021. [Online]. Available: https://doi.org/10. 1007/s40747-021-00428-4

[49] S. Rawat, K. Rana, and V. Kumar, “A novel complex-valued convolutional neural network for medical image denoising,” Biomedical Signal Processing and Control, vol. 69, p. 102859, 2021. [Online]. Available: https://www. sciencedirect.com/science/article/pii/S1746809421004560

[50] L. Deligiannidis and H. Arabnia, Emerging Trends in Image Processing, Computer Vision and Pattern Recognition. Morgan Kaufmann, 2014.

[51] M. Papaioannou, E. Plum, and N. I. Zheludev, “All-optical pattern recognition and image processing on a metamaterial beam splitter,” ACS Photonics, vol. 4, no. 2, pp. 217–222, Feb. 2017. [Online]. Available: https://doi.org/10.1021/ acsphotonics.6b00921

[52] W. Bogaerts, D. Pérez, J. Capmany, D. A. B. Miller, J. Poon, D. Englund, F. Morichetti, and A. Melloni, “Programmable photonic circuits,” Nature, vol. 586, no. 7828, pp. 207–216, 2020. [Online]. Available: https://doi.org/10.1038/ s41586-020-2764-0

[53] W. Bogaerts, R. Baets, and P. D. e. al., Nanophotonic waveguides in silicon-oninsulator fabricated with CMOS technology, 2005, vol. 23, no. 1.

[54] N. Le, Photonic signal processing : techniques and applications. CRC Press, 2007.

[55] J. Zhou, “Realization of discrete fourier transform and inverse discrete fourier transform on one single multimode interference coupler,” IEEE Photonics Technology Letters, vol. 23, no. 5, pp. 302–304, 2011.

[56] M. Kamalian, J. E. Prilepsky, S. T. Le, and S. K. Turitsyn, “Periodic nonlinear fourier transform for fiber-optic communications, part i: theory 91 and numerical methods,” Opt. Express, vol. 24, no. 16, pp. 18 353– 18 369, Aug 2016. [Online]. Available: http://opg.optica.org/oe/abstract.cfm? URI=oe-24-16-18353

[57] M. S. Moreolo and G. Cincotti, “Fiber optics transforms,” in 2008 10th Anniversary International Conference on Transparent Optical Networks, vol. 1, 2008, pp. 136–139.

[58] A. R. Gupta, K. Tsutsumi, and J. Nakayama, “Synthesis of hadamard transformers by use of multimode interference optical waveguides,” Applied Optics, vol. 42, pp. 2730–2738, 2003.

[59] J. Zhou and M. Zhang, “All-optical discrete sine transform and discrete cosine transform based on multimode interference couplers,” IEEE Photonics Technology Letters, vol. 22, no. 5, pp. 317–319, 2010.

[60] M. Deivakani, S. S. Kumar, N. U. Kumar, E. F. I. Raj, and V. Ramakrishna, “VLSI implementation of discrete cosine transform approximation recursive algorithm,” Journal of Physics: Conference Series, vol. 1817, no. 1, p. 012017, mar 2021. [Online]. Available: https://doi.org/10.1088/1742-6596/ 1817/1/012017

[61] G. L. R. Woods, J. McAllister and Y. Yi, FPGA-based Implementation of Signal Processing Systems. Wiley, 2017.

[62] J. Heaton and R. Jenkins, “General matrix theory of self-imaging in multimode interference(mmi) couplers,” IEEE Photonics Technology Letters, vol. 11, no. 2, pp. 212–214, 1999.

[63] J. John, “Discrete cosine transform in jpeg compression,” 2021. [Online]. Available: https://arxiv.org/abs/2102.06968

[64] K. Rao and P. Yip, Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic Press, 2007.

[65] A. P. Radunz, F. M. Bayer, and R. J. Cintra, “Low-complexity rounded ¨ klt approximation for image compression,” Journal of Real-Time Image Processing, vol. 19, no. 1, pp. 173–183, 2022. [Online]. Available: https://doi. org/10.1007/s11554-021-01173-0

[66] F. P. Sunny, E. Taheri, M. Nikdast, and S. Pasricha, “A survey on silicon photonics for deep learning,” J. Emerg. Technol. Comput. Syst., vol. 17, no. 4, jun 2021. [Online]. Available: https://doi.org/10.1145/3459009 92

[67] J. Jing, S. Liu, G. Wang, W. Zhang, and C. Sun, “Recent advances on image edge detection: A comprehensive review,” Neurocomputing, vol. 503, pp. 259–271, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222008141

[68] S. Xiang, Y. Han, Z. Song, X. Guo, Y. Zhang, Z. Ren, S. Wang, Y. Ma, W. Zou, B. Ma, S. Xu, J. Dong, H. Zhou, Q. Ren, T. Deng, Y. Liu, G. Han, and Y. Hao, “A review: Photonics devices, architectures, and algorithms for optical neural computing,” Journal of Semiconductors, vol. 42, no. 2, p. 023105, feb 2021. [Online]. Available: https://doi.org/10.1088/1674-4926/42/2/023105

[69] L. Yang, R. Ji, L. Zhang, J. Ding, and Q. Xu, “On-chip cmos-compatible optical signal processor,” Opt. Express, vol. 20, no. 12, pp. 13 560– 13 565, Jun 2012. [Online]. Available: http://opg.optica.org/oe/abstract.cfm? URI=oe-20-12-13560

[70] M. Salmani, A. Eshaghi, E. Luan, and S. Saha, “Photonic computing to accelerate data processing in wireless communications,” Opt. Express, vol. 29, no. 14, pp. 22 299–22 314, Jul 2021. [Online]. Available: http://opg.optica.org/oe/abstract.cfm?URI=oe-29-14-22299

[71] N. C. Harris, J. Carolan, D. Bunandar, M. Prabhu, M. Hochberg, T. BaehrJones, M. L. Fanto, A. M. Smith, C. C. Tison, P. M. Alsing, and D. Englund, “Linear programmable nanophotonic processors,” Optica, vol. 5, no. 12, pp. 1623–1631, Dec 2018. [Online]. Available: http://opg.optica.org/optica/abstract.cfm?URI=optica-5-12-1623

[72] T. Le, L. Cahill, and D. Elton, “The design of 2x2 soi mmi couplers with arbitrary power coupling ratios,” Electronics Letters, vol. 45, no. 22, pp. 1118– 1119, 2009.

[73] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Broadcast and weight: An integrated network for scalable photonic spike processing,” J. Lightwave Technol., vol. 32, no. 21, pp. 3427–3439, Nov 2014. [Online]. Available: http://opg.optica.org/jlt/abstract.cfm?URI=jlt-32-21-3427

[74] H. Shu, Z. Su, L. Huang, Z. Wu, X. Wang, Z. Zhang, and Z. Zhou, "Significantly high modulation efficiency of compact graphene modulator based on silicon waveguide," Scientific Reports, vol. 8, no. 1, p. 991, 2018. [Online]. Available: https://doi.org/10.1038/s41598-018-19171-x 93

[75] L. Wu, H. Liu, J. Li, S. Wang, S. Qu, and L. Dong, "A 130 ghz electro-optic ring modulator with double-layer graphene," Crystals, vol. 7, no. 3, 2017. [Online]. Available: https://www.mdpi.com/2073-4352/7/3/65

[76] J. Liu, Z. U. Khan, C. Wang, H. Zhang, and S. Sarjoghian, "Review of graphene modulators from the low to the high figure of merits," Journal of Physics D: Applied Physics, vol. 53, no. 23, p. 233002, apr 2020. [Online]. Available: https://doi.org/10.1088/1361-6463/ab7cf6

[77] X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, and D. J. Moss, "11 tops photonic convolutional accelerator for optical neural networks," Nature, vol. 589, no. 7840, pp. 44–51, 2021. [Online]. Available: https://doi.org/10.1038/ s41586-020-03063-0

[78] C. Huang, S. Bilodeau, T. Ferreira de Lima, A. N. Tait, P. Y. Ma, E. C. Blow, A. Jha, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits," APL Photonics, vol. 5, no. 4, p. 040803, 2020. [Online]. Available: https://doi. org/10.1063/1.5144121

[79] P. Xing, K. J. A. Ooi, and D. T. H. Tan, Ultra-broadband and compact graphene-on-silicon integrated waveguide mode filters, 2018, vol. 8, no. 1. [Online]. Available: https://doi.org/10.1038/s41598-018-28076-8

[80] J. Moolayil, Learn Keras for Deep Neural Networks: A Fast-Track Approach to Modern Deep Learning with Python. Apress Springer, 2019.

[81] D. Cires¸an, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," 2012. [Online]. Available: https://arxiv.org/ abs/1202.2745

[82] V. Bangari, B. A. Marquez, H. Miller, A. N. Tait, M. A. Nahmias, T. F. de Lima, H.-T. Peng, P. R. Prucnal, and B. J. Shastri, "Digital electronics and analog photonics for convolutional neural networks (deap-cnns)," IEEE Journal of Selected Topics in Quantum Electronics, vol. 26, no. 1, pp. 1–13, 2020.

[83] AMD, https://www.amd.com/en/products/professional-graphics/instinct-mi25, accessed date 20 Nov. 2022.

[84] W. Zhang, C. Huang, H.-T. Peng, S. Bilodeau, A. Jha, E. Blow, T. F. de Lima, B. J. Shastri, and P. Prucnal, “Silicon microring synapses enable photonic deep learning beyond 9-bit precision,” Optica, vol. 9, no. 5, pp. 94 579–584, May 2022. [Online]. Available: http://opg.optica.org/optica/abstract. cfm?URI=optica-9-5-579

[85] C. Denz, Optical Neural Networks. Springer, 1998.

[86] S. Xu, J. Wang, H. Shu, Z. Zhang, S. Yi, B. Bai, X. Wang, J. Liu, and W. Zou, “Optical coherent dot-product chip for sophisticated deep learning regression,” Light: Science and Applications, vol. 10, no. 1, p. 221, 2021. [Online]. Available: https://doi.org/10.1038/s41377-021-00666-8