Graph Mining Seminar

Vision GNN: An Image is Worth Graph of Nodes

Thành viên

- Đoàn Ngọc Mai - 21127104
- Lê Nguyễn Kiều Oanh - 21127129
- Dương Trường Bình - 21127229
- Lê Phước Quang Huy - 21127616

Giảng viên

- Lê Ngọc Thành
- Lê Nhựt Nam

Thành phố Hồ Chí Minh, Tháng 8 2024

Introduction

01 Overview

02 Problem Statements

03 Related Works

04 Plan

01

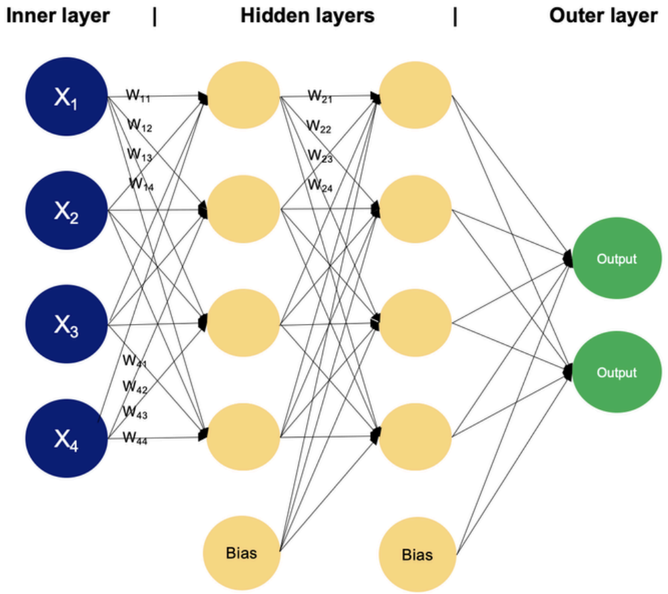
Overview

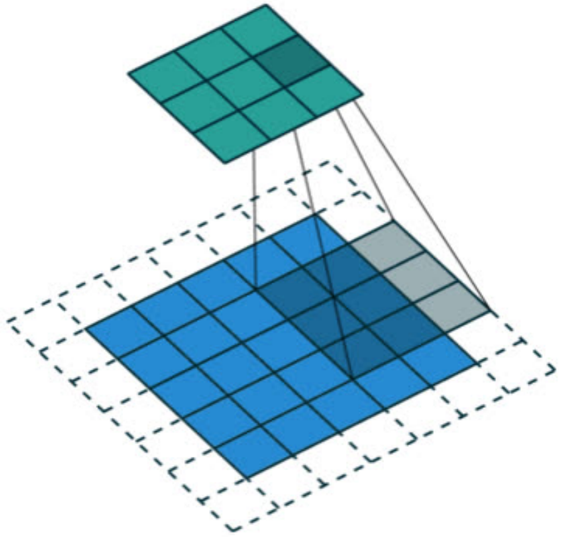
01 Bối cảnh & động lực

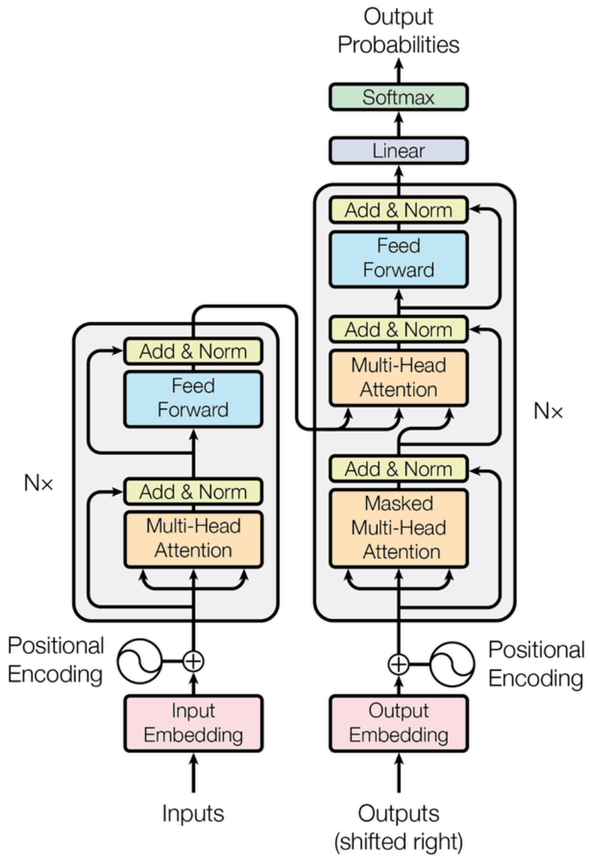
02 Ý nghĩa của việc giải quyết vấn đề

01

Bối cảnh & động lực

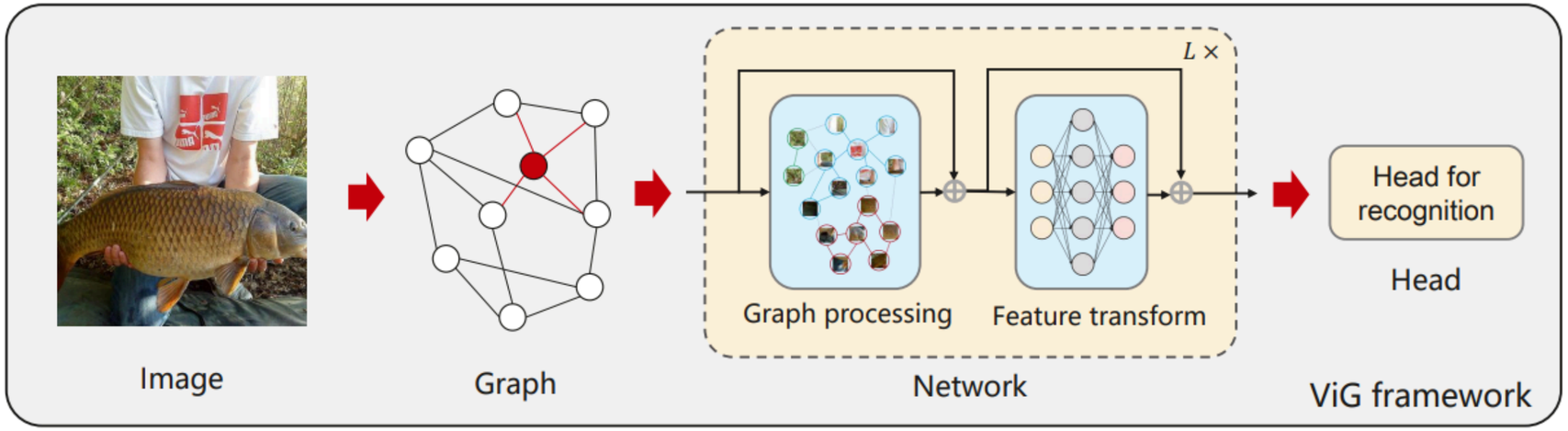






01 Ý nghĩa của nghiên cứu

- ViG: Kiến trúc mạng mới, biểu diễn hình ảnh bằng đồ thị.
- Mô hình hóa đối tượng phức tạp, nắm bắt mối quan hệ hiệu quả.
- Khắc phục hạn chế của CNN và Transformer.
- Nghiên cứu tiên phong áp dụng GNN trực tiếp cho thị giác máy tính



01

Ứng dụng thực tế

- Nhận dạng hình ảnh: Độ chính xác cao trên ImageNet.
- Phát hiện đối tượng: Hiệu quả làm backbone cho RetinaNet và Mask R-CNN.

Tiềm năng khác của ViG:

- Phân đoạn hình ảnh
- Tạo chú thích hình ảnh

"...our Pyramid ViG-S achieves 82.1% top-1 accuracy on ImageNet classification task, which outperforms the representative CNN (ResNet), MLP (CycleMLP) and transformer (Swin-T) with similar FLOPs (about 4.5G)"

02

Problem Statement

- Hạn chế của CNN và Transformer khi xử lý các đối tượng phức tạp

(a) Grid structure.

(b) Sequence structure.

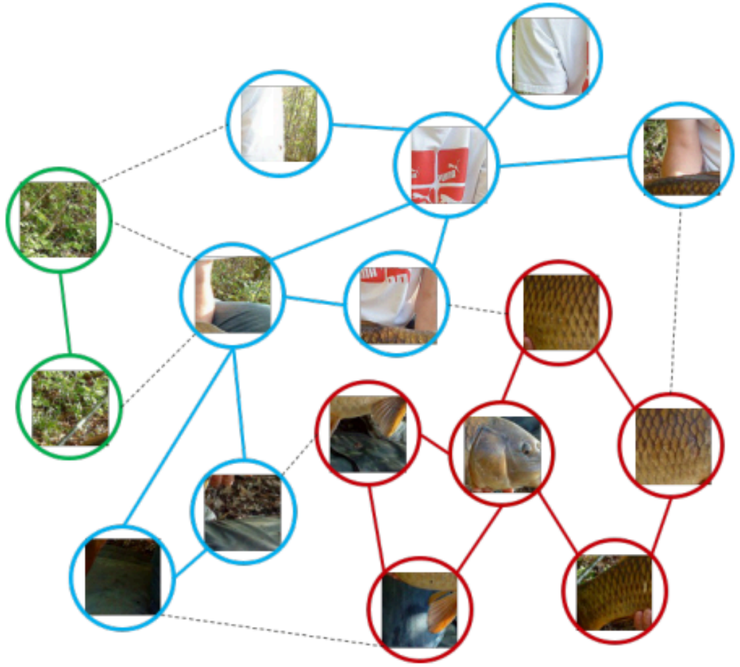
"Since the objects are usually not quadrate whose shape is irregular, the commonly-used grid or sequence structures in previous networks like ResNet and ViT are redundant and inflexible to process them"

02

Problem Statement

- Tính linh hoạt của cấu trúc đồ thị

An object can be viewed as a composition of parts, e.g., a human can be roughly divided into head, upper body, arms and legs. These parts linked by joints naturally form a graph structure. By analyzing the graph, we are able to recognize the human. Moreover, ***graph is a generalized data structure that grid and sequence can be viewed as a special case of graph.*** *Viewing an image as a graph is more flexible and effective for visual perception*



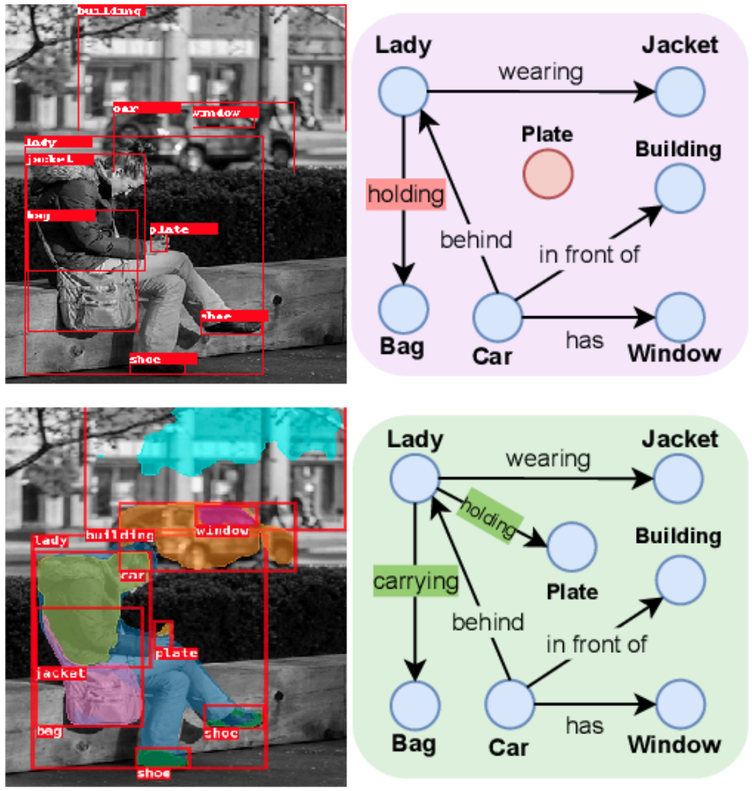
(b) Graph structure.

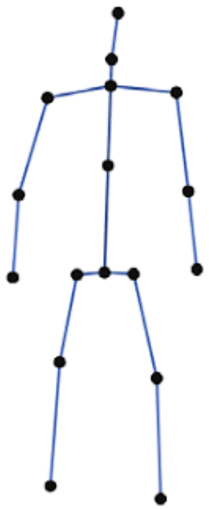
02

Problem Statement

- Tiềm năng của GNN trong thị giác máy tính

"GCN can only tackle specific visual tasks with naturally constructed graph. For general applications in computer vision, we need a GCN-based backbone network that directly processes the image data."





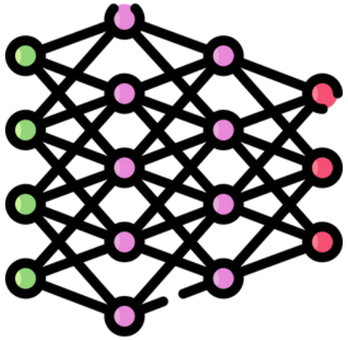
03

Framework

Ảnh

Trích xuất đặc trưng

Sử dụng các mô hình học sâu như CNN để trích xuất các đặc trưng từ hình ảnh.



Bộ phân loại

Sử dụng các mô hình như MLP để đưa ra các dự đoán dựa trên các đặc trưng đã trích xuất.

Cat

Nhãn

03

Related Work

- VGGNet
- ResNet
- Inception
- ...

2014-2015

- CycleMLP
- ResMLP
- ...

2020-2022

2012

CNN

2017

- GCN
- Transformer

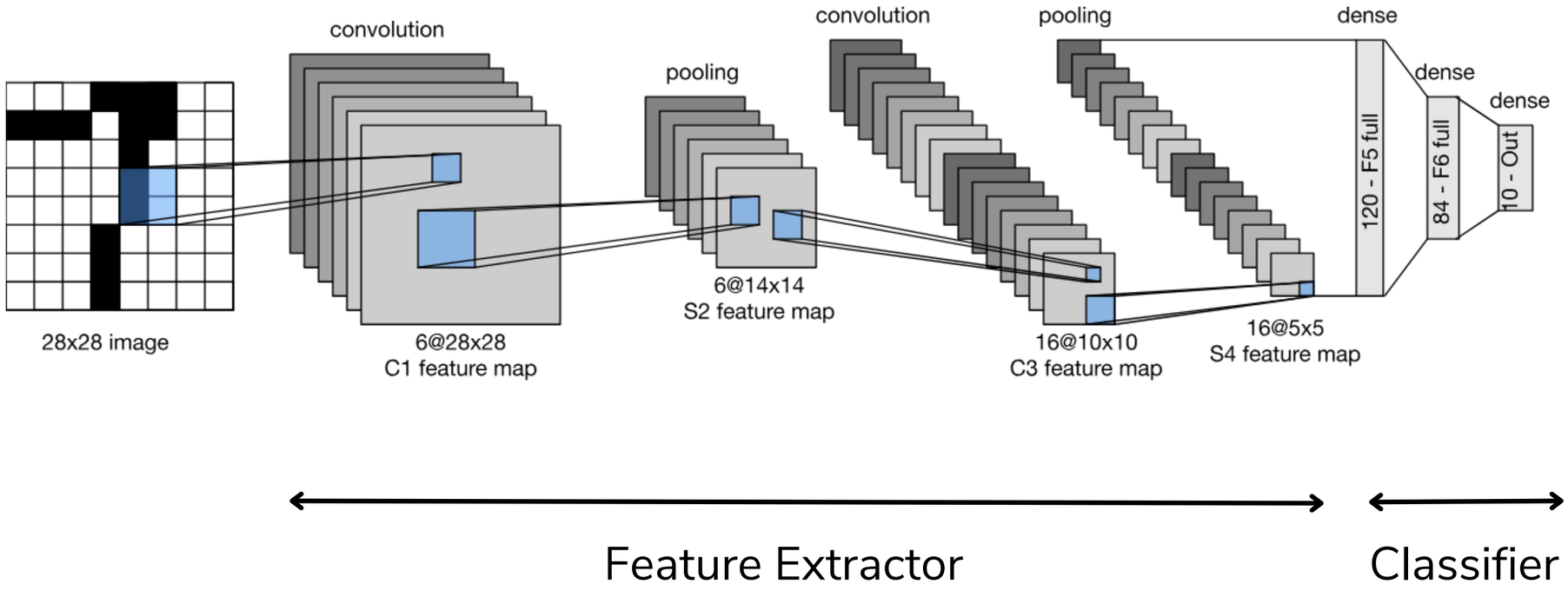
2020

ViT

2022

ViG

CNN


convolution
pooling
convolution
pooling
dense
dense
dense
28x28 image
6@28x28
C1 feature map
6@14x14
S2 feature map
16@10x10
C3 feature map
16@5x5
S4 feature map
120 - F5 full
84 - F6 full
10 - Out
Feature Extractor
Classifier

03

Related Work

- VGGNet
- ResNet
- Inception
- ...

2014-2015

- CycleMLP
- ResMLP
- ...

2020-2022

2012

CNN

2017

- GCN
- Transformer

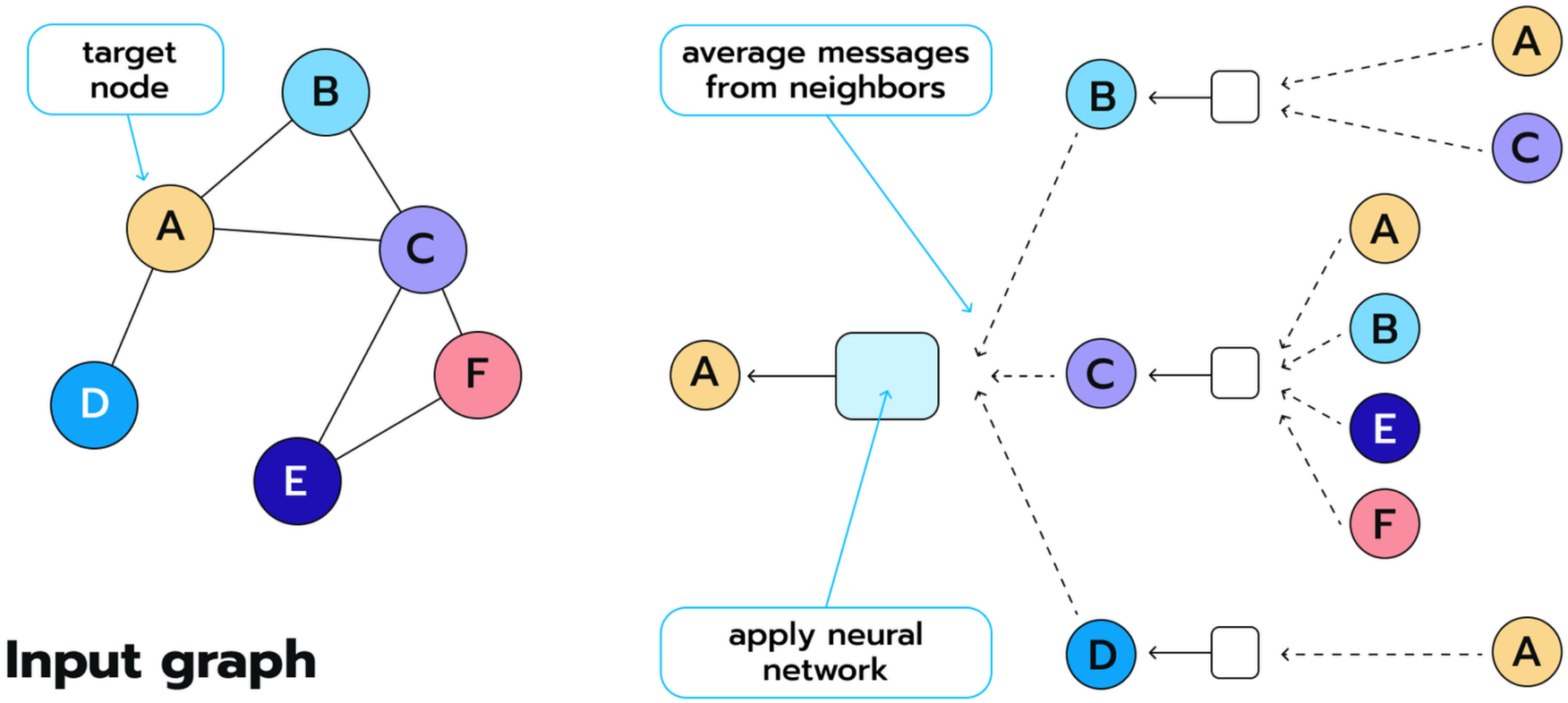
2020

ViT

2022

ViG

GNN


target node
A
B
C
D
E
F
Input graph
average messages from neighbors
apply neural network
A
B
C
D
A
C
A
B
E
F
A

03

Related Work

- VGGNet
- ResNet
- Inception
- ...

2014-2015

- CycleMLP
- ResMLP
- ...

2020-2022

2012

CNN

2017

- GCN
- Transformer

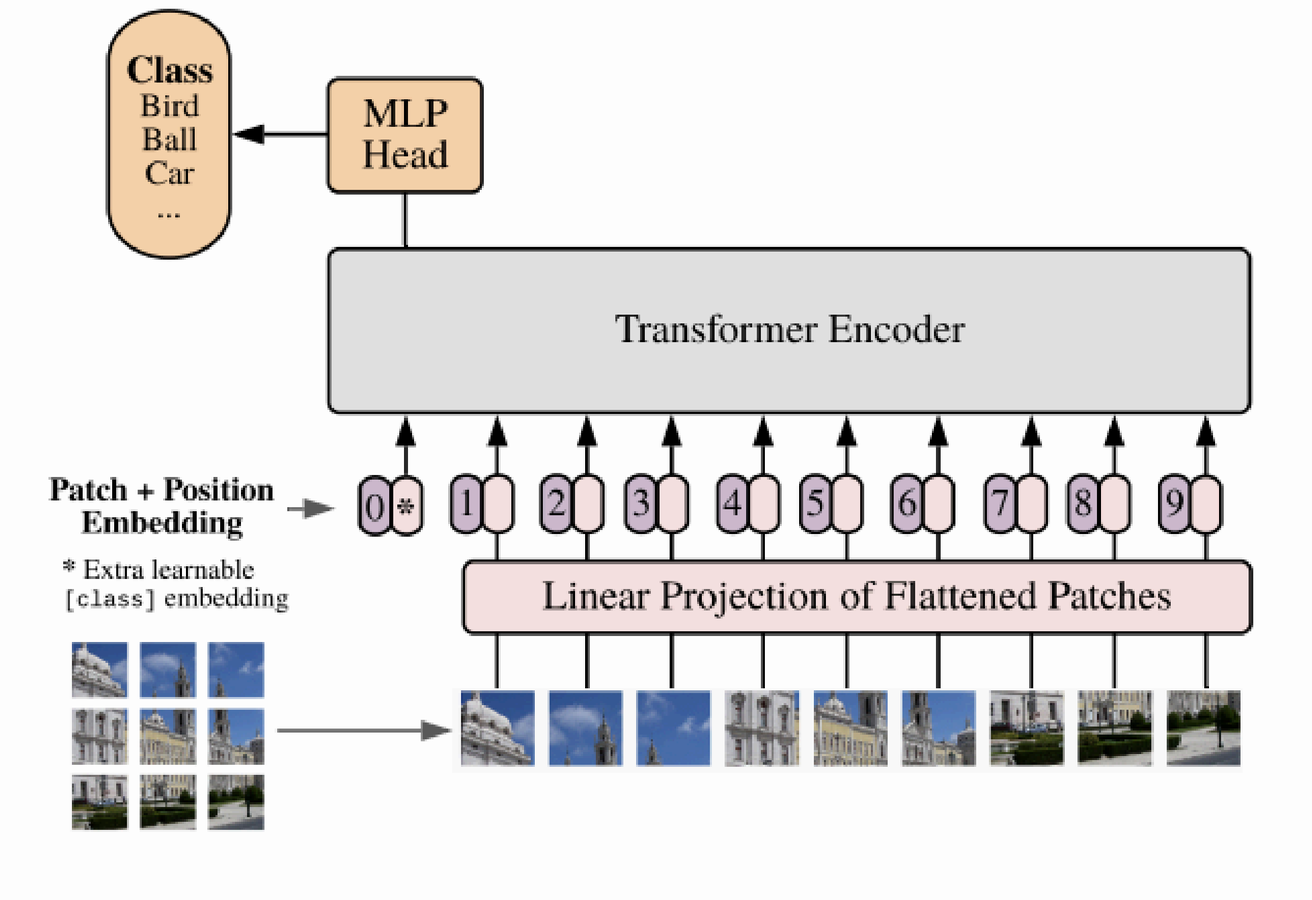
2020

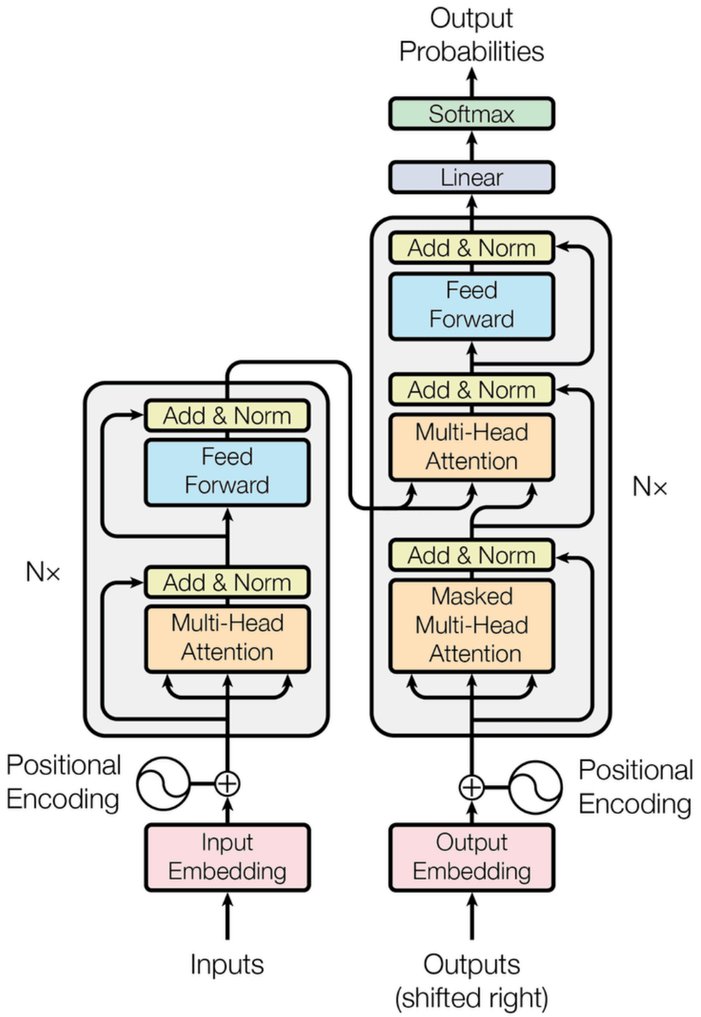
ViT

2022

ViG

Vision Transformer (ViT)





03

Related Work

- VGGNet
- ResNet
- Inception
- ...

2014-2015

- CycleMLP
- ResMLP
- ...

2020-2022

2012

CNN

2017

- GCN
- Transformer

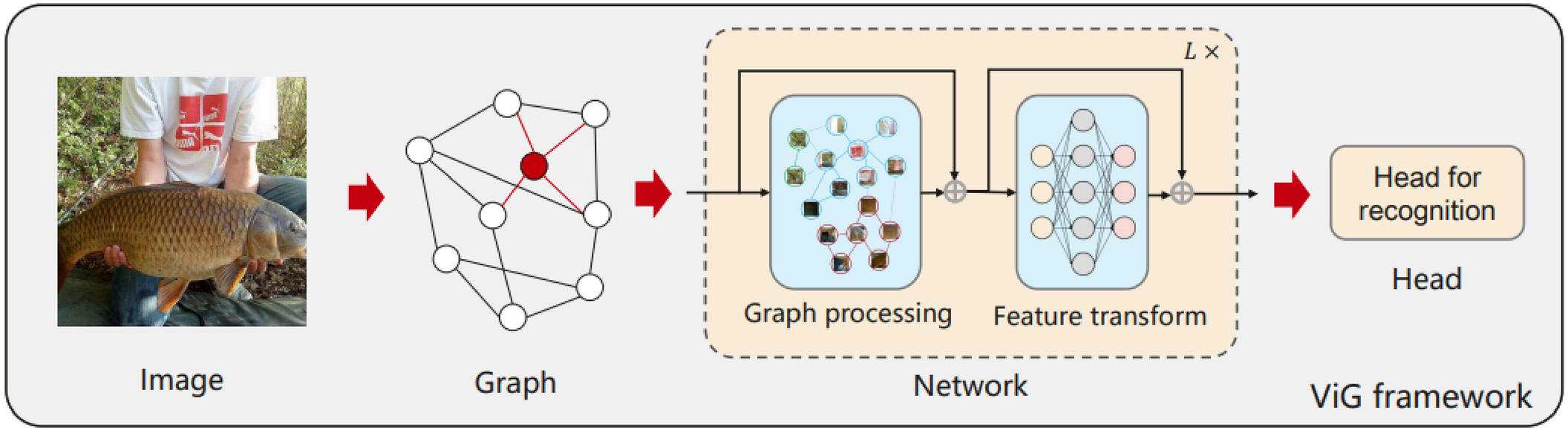
2020

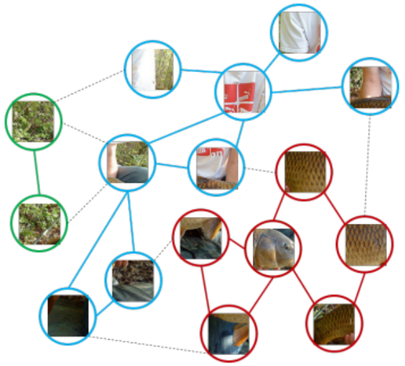
ViT

2022

ViG

Vision GNN (ViG)





03

Các kiến thức cơ bản cần biết

01 Thị giác máy tính (Computer Vision)

02 Mạng nơ-ron đồ thị (Graph Neural Network - GNN)

03 Multilayer perceptron

04 Plan

CÔNG VIỆC	NGHIÊN CỨU VỀ PHƯƠNG PHÁP VÀ KẾT QUẢ THỰC NGHIỆM	TÌM HIỂU SOURCE CODE	CHẠY THỰC NGHIỆM	PHÂN TÍCH SO SÁNH KẾT QUẢ VÀ BÁO CÁO	ABLATION STUDY (NẾU CÒN THỜI GIAN)
MÔ TẢ	• Đọc và hiểu rõ các thành phần của Vision GNN và kết quả thực nghiệm • Tìm hiểu và chuẩn bị các bộ dữ liệu cần thiết: ImageNet và COCO.	• Tìm đọc và hiểu rõ source code có liên quan. • Chuẩn bị các môi trường và công cụ cần thiết.	Chạy các thí nghiệm để so sánh ViG với các nhóm kiến trúc như CNN, MLP, Transformer	• Tổng hợp kết quả thực nghiệm và viết báo cáo. • Chuẩn bị slide và thuyết trình	Thực hiện các nghiên cứu ablation để phân tích sâu hơn về các yếu tố ảnh hưởng đến hiệu suất của Vision GNN.
THỜI GIAN (ƯỚC TÍNH)	• 1 tuần	• 1 tuần	• 1 tuần	• 1 tuần	• 1 tuần

Content

01 Overview

02 Approach

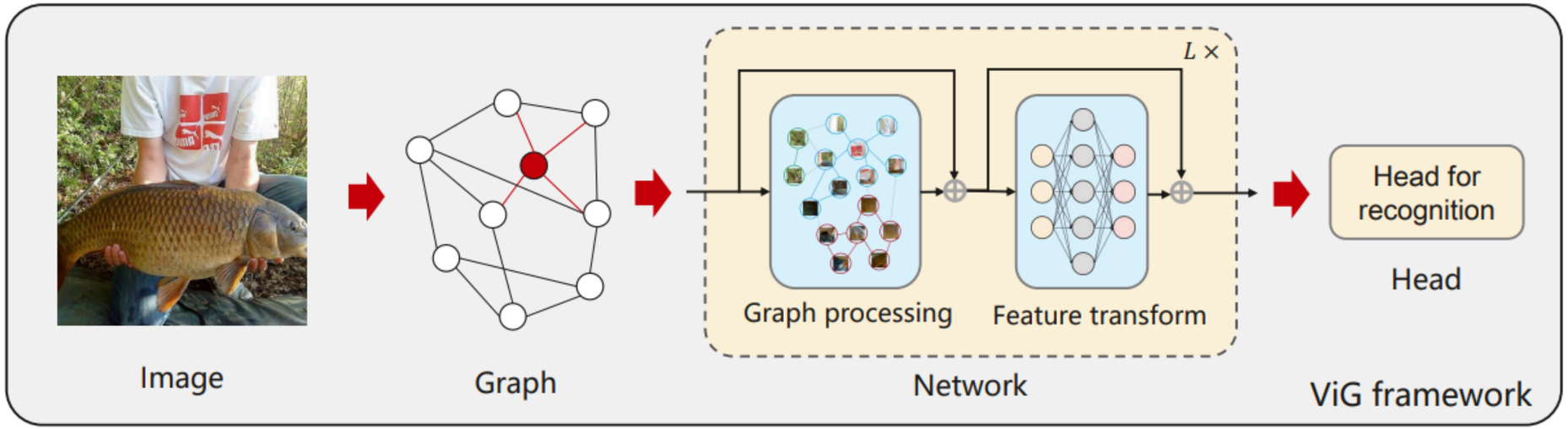
03 Experiments

04 Conclusion

01

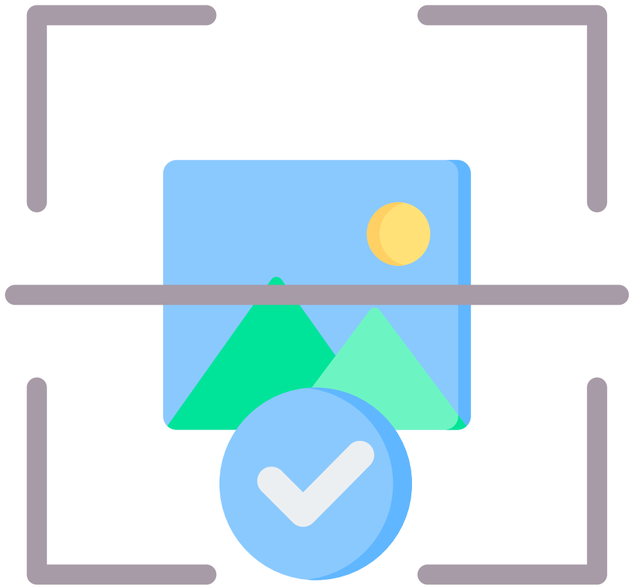
Tổng quan

- ViG: Kiến trúc mạng mới, biểu diễn hình ảnh bằng đồ thị.
- Mô hình hóa đối tượng phức tạp, nắm bắt mối quan hệ hiệu quả.
- Khắc phục hạn chế của CNN và Transformer.
- Nghiên cứu tiên phong áp dụng GNN trực tiếp cho thị giác máy tính



01 Khoảng trống nghiên cứu

- **Gap:** Chưa có GNN xử lý trực tiếp dữ liệu hình ảnh. Nghiên cứu này là tiên phong trong việc áp dụng GNN cho các tác vụ thị giác tổng quát.



01

Công trình dựa trên

Nguồn cảm hứng:

- Tác giả của nghiên cứu này đã dựa trên sự phát triển và thành tựu của các mô hình CNN, Transformer, MLP, và GNN để phát triển mô hình ViG. Mặc dù không hoàn toàn đi theo hướng mới.

Approach

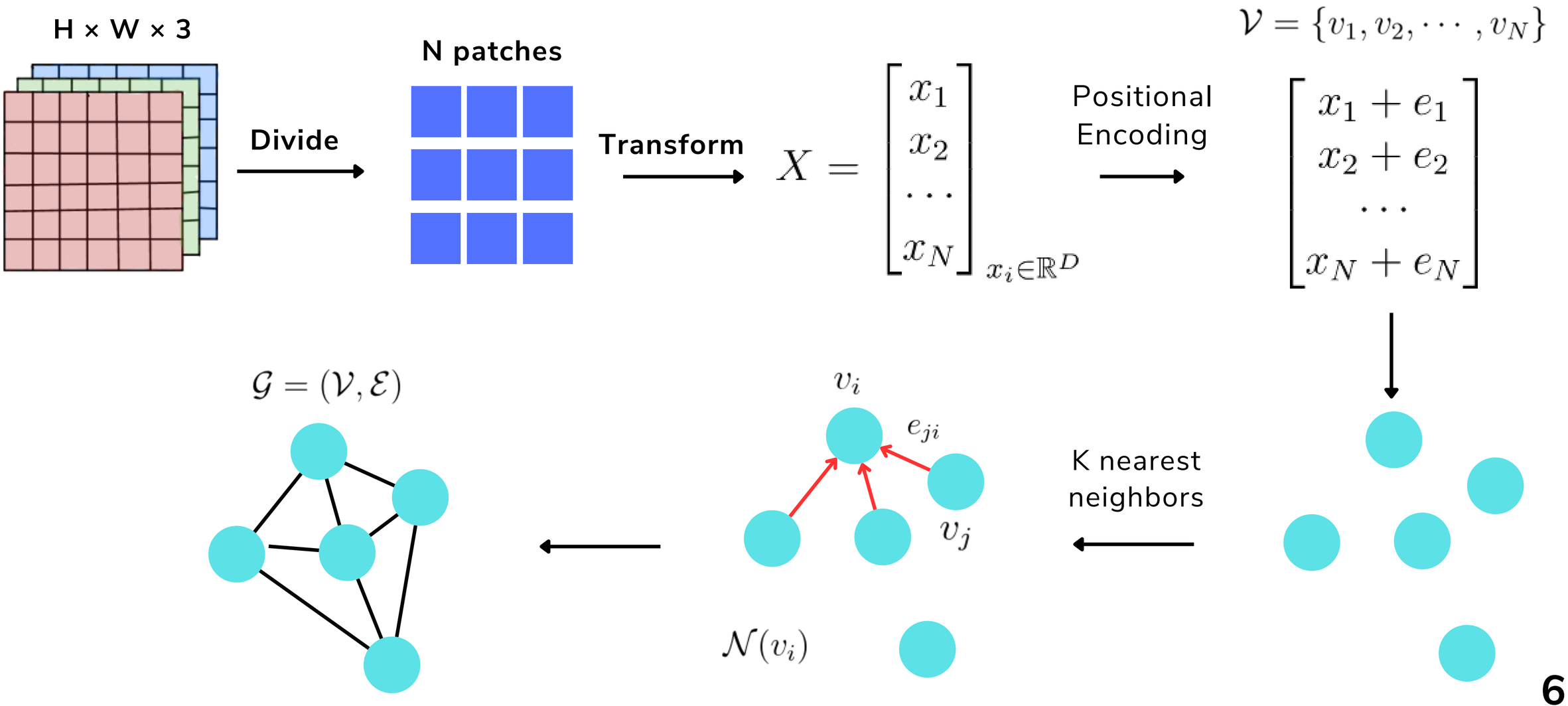
01 ViG Block

02 ViG Network

02

Graph Structure of Image

$$\mathcal{G} = G(X)$$

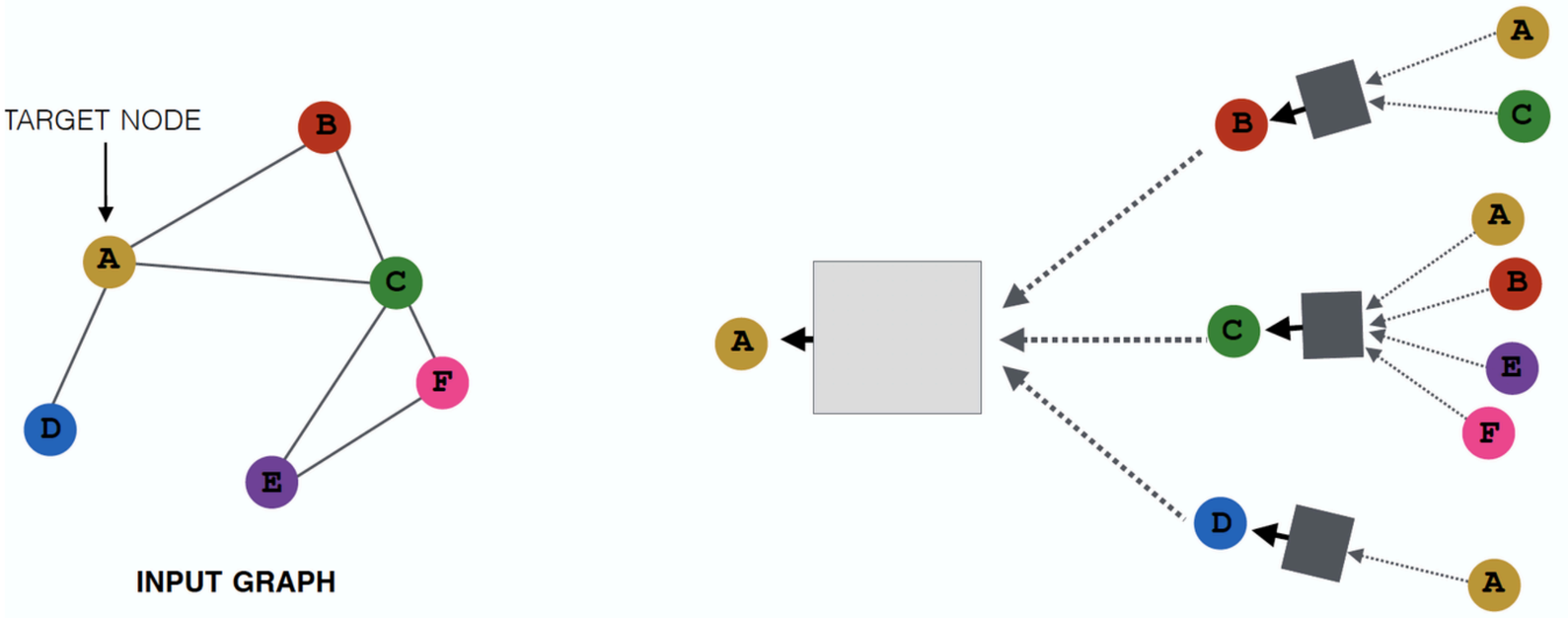


02

Graph-level Processing

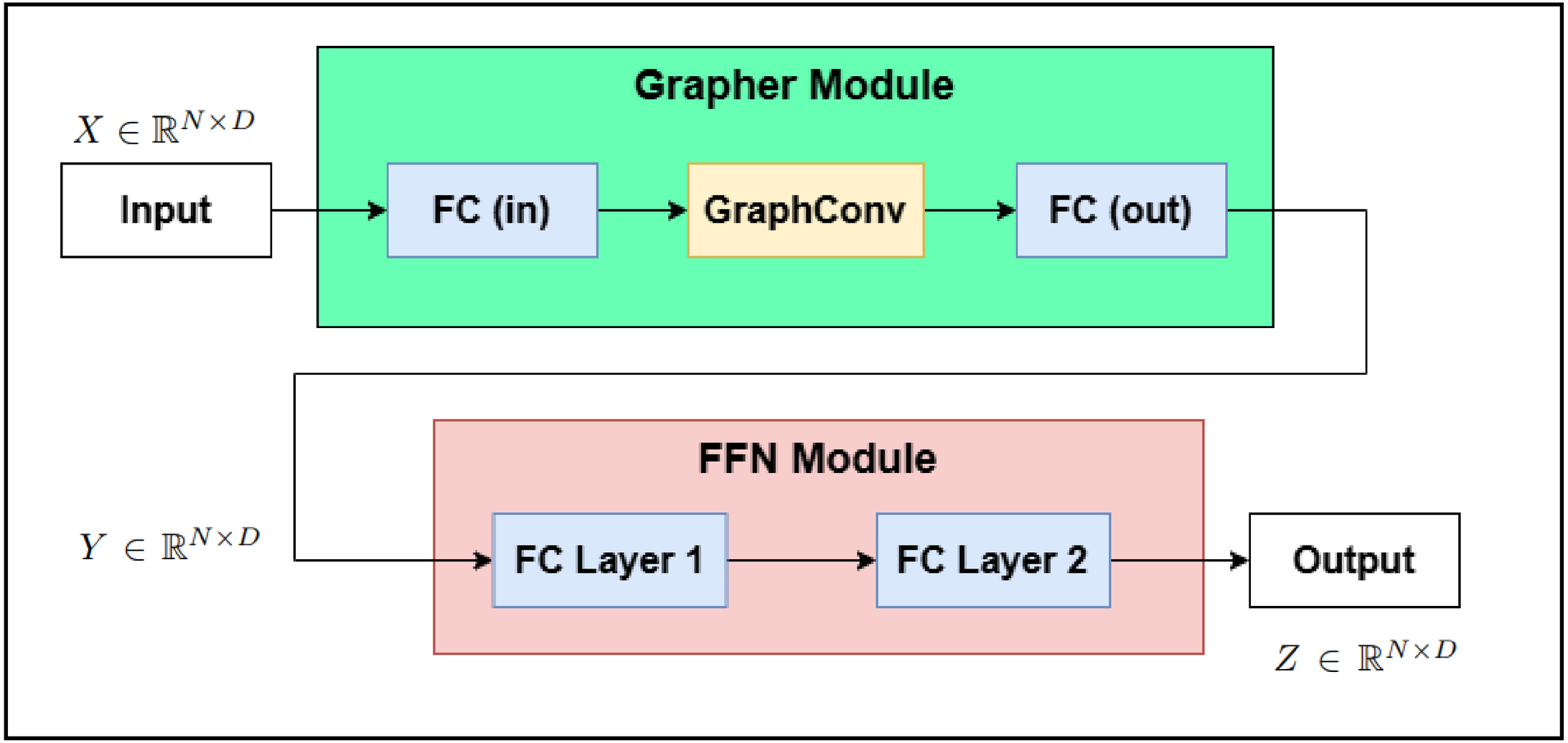
$$\mathcal{G}' = F(\mathcal{G}, \mathcal{W})$$
$$= Update(Aggregate(\mathcal{G}, W_{agg}), W_{update})$$

$$\mathbf{x}_i' = h(\mathbf{x}_i, g(\mathbf{x}_i, \mathcal{N}(\mathbf{x}_i), W_{agg}), W_{update})$$



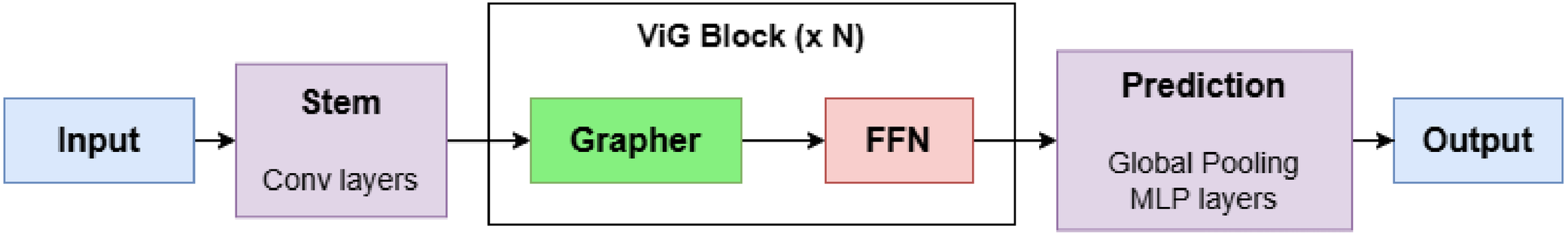
$$\mathbf{x}_i' = [head^1 W^1_{update}, head^2 W^2_{update}, \cdots, head^h W^h_{update}].$$

ViG Block


Grapher Module
$X \in \mathbb{R}^{N \times D}$
Input
FC (in)
GraphConv
FC (out)
FFN Module
$Y \in \mathbb{R}^{N \times D}$
FC Layer 1
FC Layer 2
Output
$Z \in \mathbb{R}^{N \times D}$

02

ViG Network



Isotropic architecture

Table 1: Variants of our isotropic ViG architecture. The FLOPs are calculated for the image with 224×224 resolution. 'Ti' denotes tiny, 'S' denotes small, and 'B' denotes base.

Model	Depth	Dimension D	Params (M)	FLOPs (B)
ViG-Ti	12	192	7.1	1.3
ViG-S	16	320	22.7	4.5
ViG-B	16	640	86.8	17.7

02

Pyramid Architecture

Stage	Output size	PyramidViG-Ti	PyramidViG-S	PyramidViG-M	PyramidViG-B
Stem	$\frac{H}{4} \times \frac{W}{4}$	Conv×3	Conv×3	Conv×3	Conv×3
Stage 1	$\frac{H}{4} \times \frac{W}{4}$	$\begin{bmatrix} D=48 \\ E=4 \\ K=9 \end{bmatrix} \times 2$	$\begin{bmatrix} D=80 \\ E=4 \\ K=9 \end{bmatrix} \times 2$	$\begin{bmatrix} D=96 \\ E=4 \\ K=9 \end{bmatrix} \times 2$	$\begin{bmatrix} D=128 \\ E=4 \\ K=9 \end{bmatrix} \times 2$
Downsample	$\frac{H}{8} \times \frac{W}{8}$	Conv	Conv	Conv	Conv
Stage 2	$\frac{H}{8} \times \frac{W}{8}$	$\begin{bmatrix} D=96 \\ E=4 \\ K=9 \end{bmatrix} \times 2$	$\begin{bmatrix} D=160 \\ E=4 \\ K=9 \end{bmatrix} \times 2$	$\begin{bmatrix} D=192 \\ E=4 \\ K=9 \end{bmatrix} \times 2$	$\begin{bmatrix} D=256 \\ E=4 \\ K=9 \end{bmatrix} \times 2$
Downsample	$\frac{H}{16} \times \frac{W}{16}$	Conv	Conv	Conv	Conv
Stage 3	$\frac{H}{16} \times \frac{W}{16}$	$\begin{bmatrix} D=240 \\ E=4 \\ K=9 \end{bmatrix} \times 6$	$\begin{bmatrix} D=400 \\ E=4 \\ K=9 \end{bmatrix} \times 6$	$\begin{bmatrix} D=384 \\ E=4 \\ K=9 \end{bmatrix} \times 16$	$\begin{bmatrix} D=512 \\ E=4 \\ K=9 \end{bmatrix} \times 18$
Downsample	$\frac{H}{32} \times \frac{W}{32}$	Conv	Conv	Conv	Conv
Stage 4	$\frac{H}{32} \times \frac{W}{32}$	$\begin{bmatrix} D=384 \\ E=4 \\ K=9 \end{bmatrix} \times 2$	$\begin{bmatrix} D=640 \\ E=4 \\ K=9 \end{bmatrix} \times 2$	$\begin{bmatrix} D=768 \\ E=4 \\ K=9 \end{bmatrix} \times 2$	$\begin{bmatrix} D=1024 \\ E=4 \\ K=9 \end{bmatrix} \times 2$
Head	1×1	Pooling & MLP	Pooling & MLP	Pooling & MLP	Pooling & MLP
Parameters (M)		10.7	27.3	51.7	92.6
FLOPs (B)		1.7	4.6	8.9	16.8

Experiments

01 Experimental Settings

02 Main Results on ImageNet

03 Object Detection

04 Ablation Study

03 Datasets & Experimental Settings

- **Datasets:**
  - ImageNet: 1.2M ảnh training, 50K ảnh validation, 1,000 loại
  - COCO 2017: 118K ảnh training, 5K ảnh validation images, 80 loại
- **Experimental Settings:**
  - **ViG Models**: Dilated aggregation, GELU activation.
  - **ImageNet**: Chiến lược huấn luyện DeiT, các phương pháp tăng cường dữ liệu (RandAugment, Mixup, Cutmix ...)
  - **COCO Detection**: Sử dụng RetinaNet và Mask R-CNN với Pyramid ViG làm backbone
  - **Implementation:** Huấn luyện trên 8 NVIDIA V100 GPUs dùng PyTorch và MindSpore.

Table 3: Training hyper-parameters for ImageNet.

(Pyramid) ViG	Ti	S	M	B
Epochs	300			
Optimizer	AdamW [37]			
Batch size	1024			
Start learning rate (LR)	2e-3			
Learning rate schedule	Cosine			
Warmup epochs	20			
Weight decay	0.05			
Label smoothing [47]	0.1			
Stochastic path [22]	0.1	0.1	0.1	0.3
Repeated augment [20]	✓			
RandAugment [6]	✓			
Mixup prob. [73]	0.8			
Cutmix prob. [72]	1.0			
Random erasing prob. [74]	0.25			
Exponential moving average	0.99996			

03

Evaluation Metrics

- **Params (M)**: Số lượng tham số của mô hình tính bằng triệu (Million).
- **FLOPs (B)**: Số phép toán dấu chấm động cần thiết để xử lý một hình ảnh, tính bằng tỉ (Billion).
- **Top-1 Accuracy**: Tỷ lệ phần trăm số lần dự đoán hàng đầu của mô hình là chính xác
- **Top-5 Accuracy**: Tỷ lệ phần trăm số lần lớp đúng nằm trong top 5 dự đoán của mô hình.

03

Main Result

- **Kiến trúc Isotropic:** Duy trì kích thước đặc trưng trong các lớp chính, giúp dễ dàng mở rộng và thân thiện với phần cứng.
- So sánh với CNN, Transformer và MLP isotropic.
- **Kết quả**: ViG vượt trội hơn các mô hình khác. ViG-Ti đạt độ chính xác top-1 là 73.9%, cao hơn 1.7% so với mô hình DeiT-Ti với chi phí tính toán tương đương.

Table 4: Results of ViG and other isotropic networks on ImageNet. ♠ CNN, ■ MLP, ♦ Transformer, ★ GNN.

Model	Resolution	Params (M)	FLOPs (B)	Top-1	Top-5
♠ ResMLP-S12 conv3x3 [50]	224×224	16.7	3.2	77.0	-
♠ ConvMixer-768/32 [52]	224×224	21.1	20.9	80.2	-
♠ ConvMixer-1536/20 [52]	224×224	51.6	51.4	81.4	-
♦ ViT-B/16 [9]	384×384	86.4	55.5	77.9	-
♦ DeiT-Ti [51]	224×224	5.7	1.3	72.2	91.1
♦ DeiT-S [51]	224×224	22.1	4.6	79.8	95.0
♦ DeiT-B [51]	224×224	86.4	17.6	81.8	95.7
■ ResMLP-S24 [50]	224×224	30	6.0	79.4	94.5
■ ResMLP-B24 [50]	224×224	116	23.0	81.0	95.0
■ Mixer-B/16 [49]	224×224	59	11.7	76.4	-
★ ViG-Ti (ours)	224×224	7.1	1.3	**73.9**	**92.0**
★ ViG-S (ours)	224×224	22.7	4.5	**80.4**	**95.2**
★ ViG-B (ours)	224×224	86.8	17.7	**82.3**	**95.9**

03

Main Result

- **Kiến trúc Pyramid:** Giảm dần kích thước feature map, giúp thu nhận các đặc trưng nhiều tỷ lệ và tận dụng tính không thay đổi kích thước của hình ảnh.
- So sánh với CNN, MLP, Transformer
- **Kết quả**: Đạt hiệu suất tương đương hoặc vượt trội so với các mạng hàng đầu, cho thấy rằng mạng nơ-ron đồ thị có thể hoạt động xuất sắc trong các tác vụ thị giác và có tiềm năng trở thành thành phần cơ bản trong hệ thống thị giác máy tính.

Table 5: Results of Pyramid ViG and other pyramid networks on ImageNet. ♠ CNN, ■ MLP, ♦ Transformer, ★ GNN.

Model	Resolution	Params (M)	FLOPs (B)	Top-1	Top-5
♠ ResNet-18 [17, 59]	224×224	12	1.8	70.6	89.7
♠ ResNet-50 [17, 59]	224×224	25.6	4.1	79.8	95.0
♠ ResNet-152 [17, 59]	224×224	60.2	11.5	81.8	95.9
♠ BoTNet-T3 [46]	224×224	33.5	7.3	81.7	-
♠ BoTNet-T3 [46]	224×224	54.7	10.9	82.8	-
♠ BoTNet-T3 [46]	256×256	75.1	19.3	83.5	-
♦ PVT-Tiny [57]	224×224	13.2	1.9	75.1	-
♦ PVT-Small [57]	224×224	24.5	3.8	79.8	-
♦ PVT-Medium [57]	224×224	44.2	6.7	81.2	-
♦ PVT-Large [57]	224×224	61.4	9.8	81.7	-
♦ CvT-13 [60]	224×224	20	4.5	81.6	-
♦ CvT-21 [60]	224×224	32	7.1	82.5	-
♦ CvT-21 [60]	384×384	32	24.9	83.3	-
♦ Swin-T [35]	224×224	29	4.5	81.3	95.5
♦ Swin-S [35]	224×224	50	8.7	83.0	96.2
♦ Swin-B [35]	224×224	88	15.4	83.5	96.5
■ CycleMLP-B2 [5]	224×224	27	3.9	81.6	-
■ CycleMLP-B3 [5]	224×224	38	6.9	82.4	-
■ CycleMLP-B4 [5]	224×224	52	10.1	83.0	-
■ Poolformer-S12 [71]	224×224	12	2.0	77.2	93.5
■ Poolformer-S36 [71]	224×224	31	5.2	81.4	95.5
■ Poolformer-M48 [71]	224×224	73	11.9	82.5	96.0
★ Pyramid ViG-Ti (ours)	224×224	10.7	1.7	**78.2**	**94.2**
★ Pyramid ViG-S (ours)	224×224	27.3	4.6	**82.1**	**96.0**
★ Pyramid ViG-M (ours)	224×224	51.7	8.9	**83.1**	**96.4**
★ Pyramid ViG-B (ours)	224×224	92.6	16.8	**83.7**	**96.5**

03

Object Detection

- **Evaluation**: Pyramid ViG-S pretrained trên ImageNet cho object detection sử dụng các framework RetinaNet và Mask R-CNN. Các mô hình được huấn luyện theo lịch trình "1x" và FLOPs được tính với kích thước đầu vào 1280×800.
- **Kết quả**: Pyramid ViG-S vượt trội hơn các mô hình backbone khác như ResNet, CycleMLP và Swin Transformer trên cả RetinaNet và Mask R-CNN. Điều này chứng tỏ khả năng tổng quát mạnh mẽ của kiến trúc ViG.

Table 10: Object detection and instance segmentation results on COCO val2017. Our Pyramid ViG is compared with other backbones on RetinaNet and Mask R-CNN frameworks.

Backbone	RetinaNet 1×							
	Param	FLOPs	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet50 [17]	37.7M	239.3B	36.3	55.3	38.6	19.3	40.0	48.8
ResNeXt-101-32x4d [62]	56.4M	319B	39.9	59.6	42.7	22.3	44.2	52.5
PVT-Small [57]	34.2M	226.5B	40.4	61.3	44.2	25.0	42.9	55.7
CycleMLP-B2 [5]	36.5M	230.9B	40.6	61.4	43.2	22.9	44.4	54.5
Swin-T [35]	38.5M	244.8B	41.5	62.1	44.2	25.1	44.9	**55.5**
Pyramid ViG-S (ours)	36.2M	240.0B	**41.8**	**63.1**	**44.7**	**28.5**	**45.4**	53.4
Backbone	**Mask R-CNN 1×**							
	Param	FLOPs	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet50 [17]	44.2M	260.1B	38.0	58.6	41.4	34.4	55.1	36.7
PVT-Small [57]	44.1M	245.1B	40.4	62.9	43.8	37.8	60.1	40.3
CycleMLP-B2 [5]	46.5M	249.5B	42.1	64.0	45.7	38.9	61.2	41.8
PoolFormer-S24 [71]	41.0M	-	40.1	62.2	43.4	37.0	59.1	39.6
Swin-T [35]	47.8M	264.0B	42.2	64.6	**46.2**	39.1	61.6	**42.0**
Pyramid ViG-S (ours)	45.8M	258.8B	**42.6**	**65.2**	46.0	**39.4**	**62.4**	41.6

Ablation Study

Ablation Study

- So sánh các loại graph convolution khác nhau
- EdgeConv, GIN, GraphSAGE và Max-Relative đều tốt hơn baseline
- Max-Relative cho kết quả tốt nhất và cân bằng giữa độ chính xác và hiệu suất tính toán

Table 6: ImageNet results of different types of graph convolution. The basic architecture is ViG-Ti.

GraphConv	Params (M)	FLOPs (B)	Top-1
EdgeConv [58]	7.2	2.4	74.3
GIN [64]	7.0	1.3	72.8
GraphSAGE [13]	7.3	1.6	74.0
Max-Relative GraphConv [30]	7.1	1.3	73.9

03

Ablation Study

- So sánh hiệu quả khi thêm các module vào mô hình
- Chỉ dùng GraphConv cho kết quả thấp nhất
- Thêm FC layer và FFN cải thiện đáng kể độ chính xác
- Kết hợp đầy đủ GraphConv, FC và FFN cho kết quả tốt nhất

Table 7: The effects of modules in ViG on ImageNet.

GraphConv	FC in Grapher module	FFN module	Params (M)	FLOPs (B)	Top-1
✓	✗	✗	5.8	1.4	67.0
✓	✓	✗	4.4	1.4	73.4
✓	✗	✓	7.7	1.3	73.6
✓	✓	✓	7.1	1.3	73.9

Ablation Study

- Khảo sát ảnh hưởng của số lượng node láng giềng K
- K quá nhỏ (3) hoặc quá lớn (20) đều không tốt
- K trong khoảng 9-15 cho kết quả tốt nhất

Table 8: Top-1 accuracy *vs.* K on ImageNet.

K	3	6	9	12	15	20	9 to 18
Top-1	72.2	73.4	73.6	73.6	73.5	73.3	73.9

Ablation Study

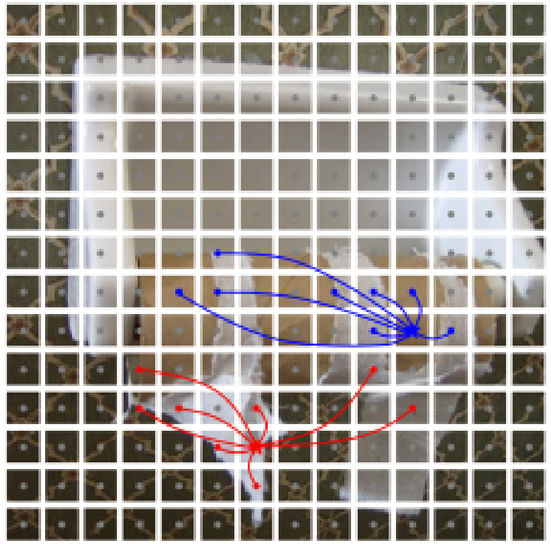
- So sánh hiệu quả với số lượng heads khác nhau
- Số lượng heads ảnh hưởng không đáng kể đến độ chính xác
- Tăng số heads làm giảm nhẹ FLOP
- Chọn 4 heads là cân bằng tốt nhất giữa FLOPs và accuracy

Table 9: Top-1 accuracy *vs.* h on ImageNet.

h	1	2	4	6	8
FLOPs / Top-1	1.6B / 74.2	1.4B / 74.0	1.3B / 73.9	1.2B / 73.7	1.2B / 73.7

03

Visualization



(a) Input image.

(b) Graph connection in the 1st block.

(c) Graph connection in the 12th block.

03

Verifying ViG Result

- Dataset: Tiny ImageNet 200
- Kiến trúc Isotropic
- Inference

Model	Param (M)	FLOPs (B)	Loss	Top-1	Top-5
ViG-Ti	7.1	1.3	7.957	0.07	0.4
ViG-S	22.7	4.5	8.305	0.06	0.51
ViG-B	86.8	17.7	7.768	0.06	0.63

Conclusion

04

Conclusion

- **Đổi mới**: Nghiên cứu giới thiệu việc đại diện hình ảnh dưới dạng dữ liệu đồ thị và sử dụng mạng nơ-ron đồ thị (GNN) cho các nhiệm vụ thị giác.
- **Thách thức**: Convolution đồ thị trực tiếp gặp phải vấn đề over-smoothing và hiệu suất kém.
- **Giải pháp**: Nghiên cứu cải thiện tính đa dạng thông tin bằng cách giới thiệu các biến đổi đặc trưng bổ sung trong các nút.
- **Kết quả**: Các mạng ViG, với cả kiến trúc isotropic và pyramid, cho thấy hiệu suất vượt trội trong nhận diện hình ảnh và phát hiện đối tượng.

Thank You