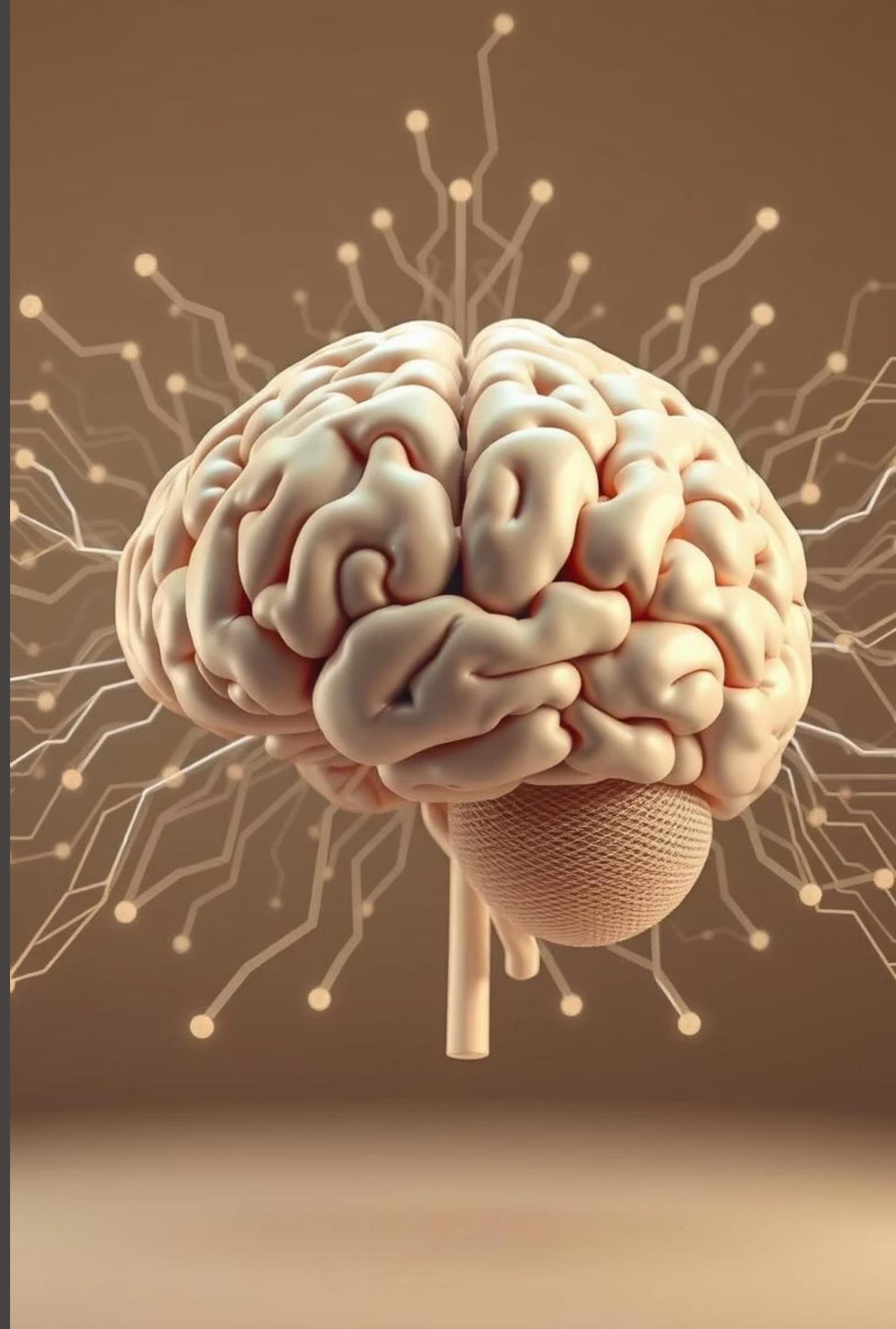
CNN for Object Recognition

A Seminar Presentation for Visual Processing Course (COMP-8510)

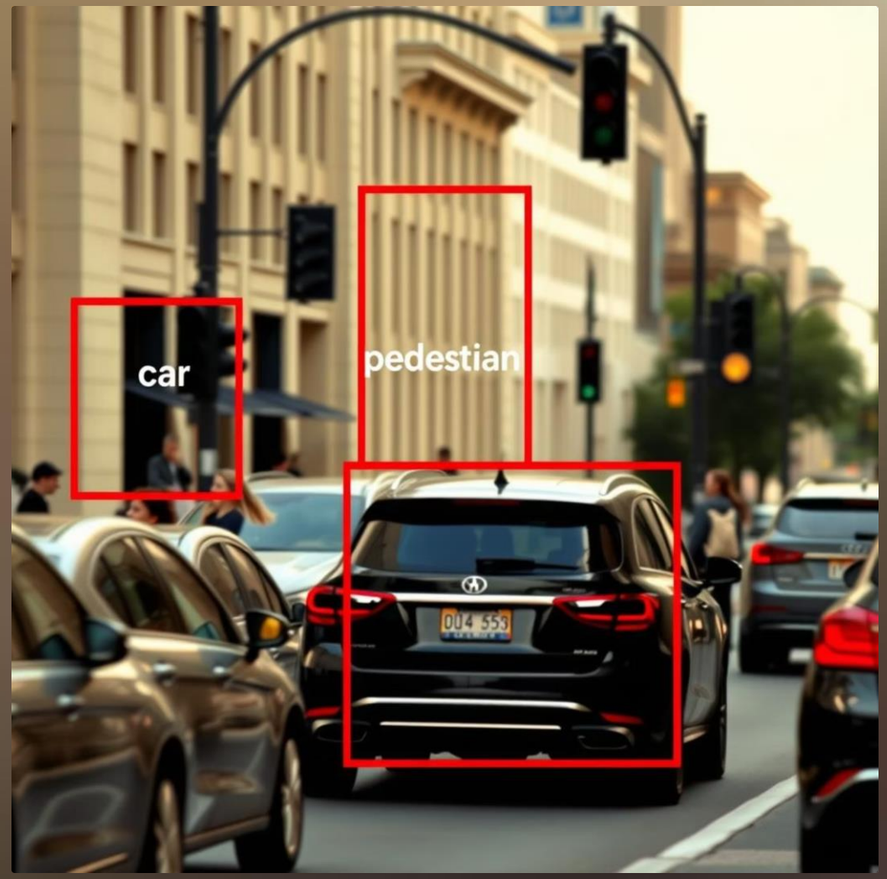
Presented by: Shaurya Parshad (SID: 110191553)



What is Object Recognition?

Object recognition is a fundamental computer vision task that involves identifying and localizing objects within an image or video. It's a two-fold process:

- **Classification:** Determining what an object is (e.g., "cat", "car").
- **Localization:** Pinpointing where the object is in the image, often with a bounding box.



Importance of Object Recognition

Autonomous Systems

Enables self-driving cars and robots to "see" and interact safely with their environment.

Enhanced Search

Powers image and video search engines, allowing users to find content based on visual cues.

Security & Surveillance

Automates threat detection, anomaly identification, and monitoring in real-time video feeds.

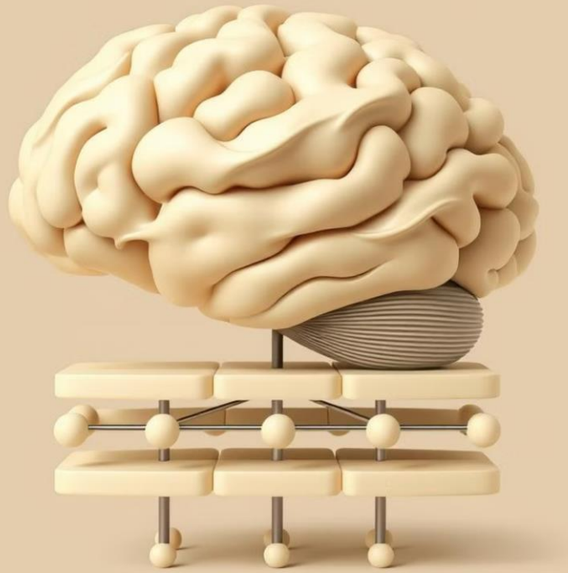
Bridging Perception

Connects machine intelligence with human-like visual perception capabilities.

Introduction to Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a specialized class of deep neural networks particularly adept at processing visual data. They draw inspiration from the biological visual cortex.

- **Bio-Inspired:** Mimic the hierarchical processing of the human visual system.
- **Feature Extraction:** Excel at automatically learning and extracting features like edges, textures, and shapes.
- **End-to-End Learning:** Can learn directly from raw pixel data, eliminating the need for manual feature engineering.



How CNNs Work: Key Layers

CNNs process images through a series of specialized layers, each performing a distinct function to transform raw pixel data into meaningful features.

1

Convolutional Layers

Apply filters to input images, detecting local patterns such as edges, corners, and textures. This generates feature maps.

2

Pooling Layers

Reduce the dimensionality of feature maps, maintaining important information while reducing computational load and increasing invariance to minor shifts.

3

Fully Connected Layers

Take the high-level features extracted by previous layers and use them to make final classifications or predictions, connecting all neurons from one layer to the next.

CNNs in Object Detection Architectures

CNNs form the backbone of many state-of-the-art object detection models, known for their ability to precisely localize objects.

- **Classic Examples:** Faster R-CNN, YOLO (You Only Look Once), SSD (Single Shot Detector) are prominent CNN-based models.
- **Methodology:** They typically involve generating region proposals, extracting features from these regions, and then classifying them while predicting bounding box coordinates.
- **Strength:** Exceptional at extracting fine-grained local details and achieving accurate localization of objects within an image.

Transformers in Vision

Originally designed for natural language processing, Transformer architectures have recently been adapted for computer vision tasks, offering a new paradigm for image understanding.

Vision Transformers (ViT)

Adapt the self-attention mechanism to process image patches like words.

Global Context

Excellent at capturing long-range dependencies and relationships across the entire image.

Spatial Relationships

Effective at understanding the overall context and spatial arrangement of objects.

Local Detail Limitation

Often less proficient than CNNs in capturing fine-grained local textures and precise details.

CNNs vs. Transformers: A Comparison

Both CNNs and Transformers offer distinct advantages in visual processing, making them suitable for different aspects of image understanding.

Strengths	Sharp on tiny details (edges, textures) and see small patches at one time	Global context, long-range dependencies, spatial relationships
Best For	Precise localization, speed and details	Understanding overall scene and broad context and relationships.
Examples	Used in VGGNet , AexNet	Used in Vision Transformer, ChatGPT, BERT

Primary Article: "A Dynamic Dual-Processing Object Detection Framework"

Article Title: "A Dynamic Dual-Processing Object Detection Framework Inspired by the Brain's Recognition Mechanism."

Authors: Minying Zhang, Tianpeng Bu, Lulu Hu (Alibaba Group)

Key Idea: Proposes a novel framework that integrates the strengths of CNNs for local detail extraction and Transformers for global context understanding, mimicking human visual processing.

Motivation and Biological Inspiration

The article's core motivation stems from neuroscience, which suggests that the human brain employs a dual-processing mechanism for visual recognition.

1

Familiarity Mechanism

Quick Recognition: Relies on rapid, local cue processing, enabling instant identification of familiar objects based on specific features.

CNN Analogy: Similar to how CNNs excel at extracting and using hierarchical local details.

2

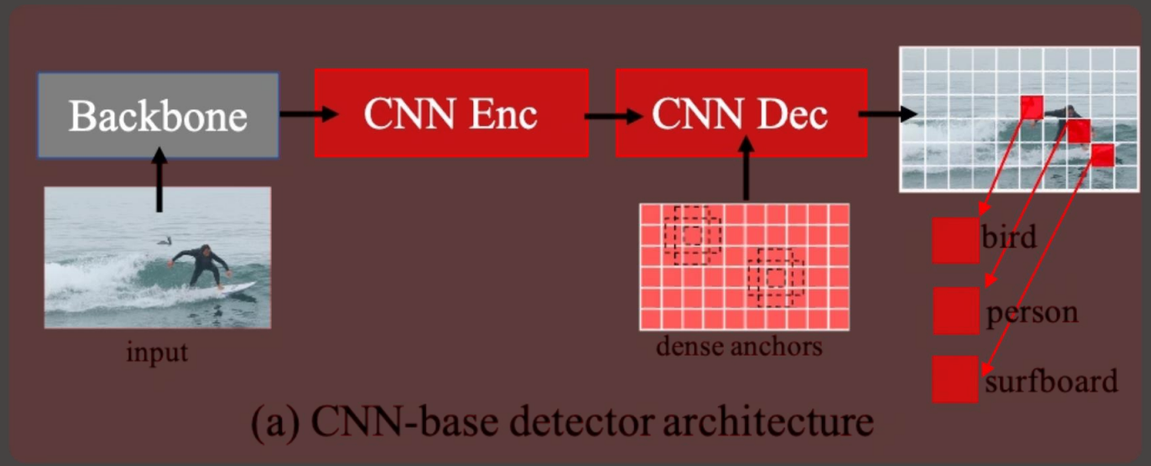
Recollection Mechanism

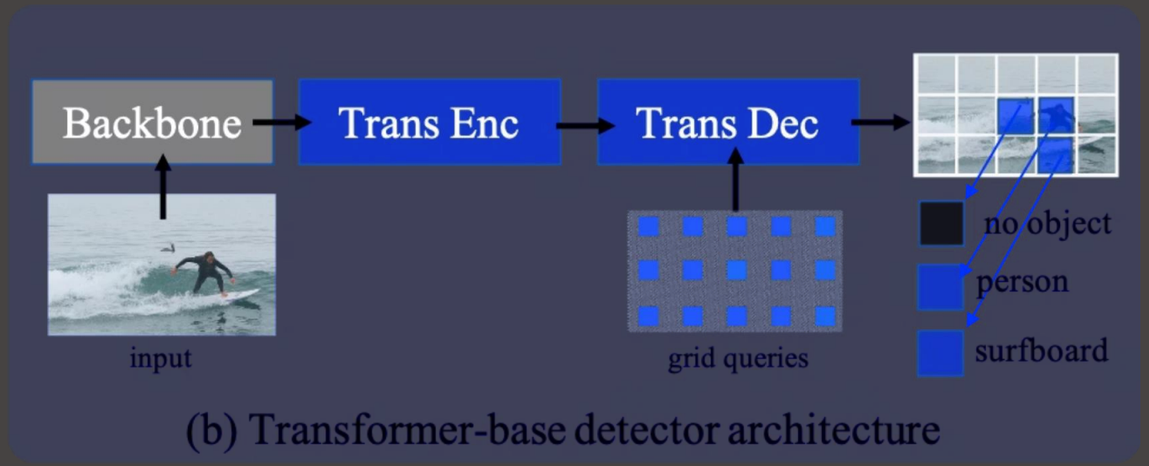
Contextual Recognition: Involves slower, more deliberate processing that leverages global context and relationships between elements.

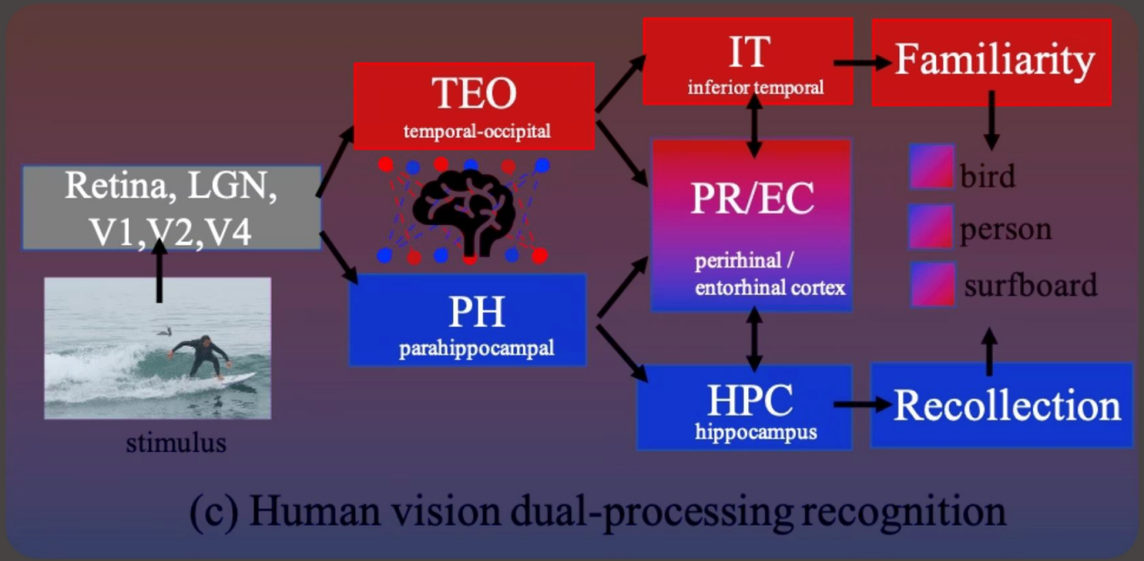
Transformer Analogy: Mirrors the Transformer's ability to grasp long-range dependencies and overall scene understanding.

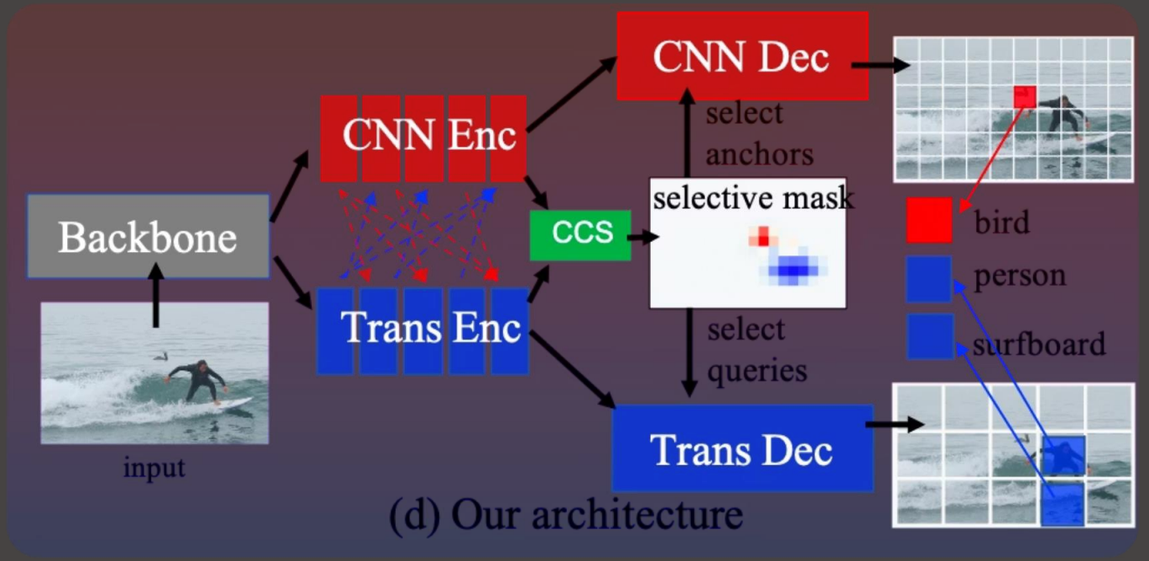
The research posits that by combining these complementary approaches, a hybrid model can overcome the limitations of single-paradigm systems, leading to more robust and brain-like object detection.

Various Architectures



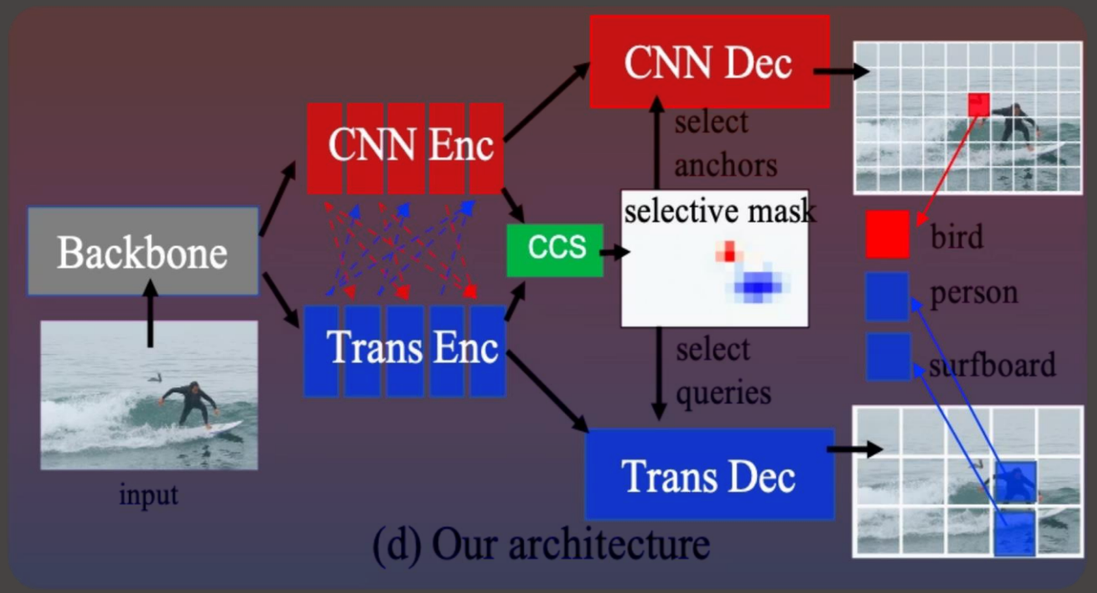






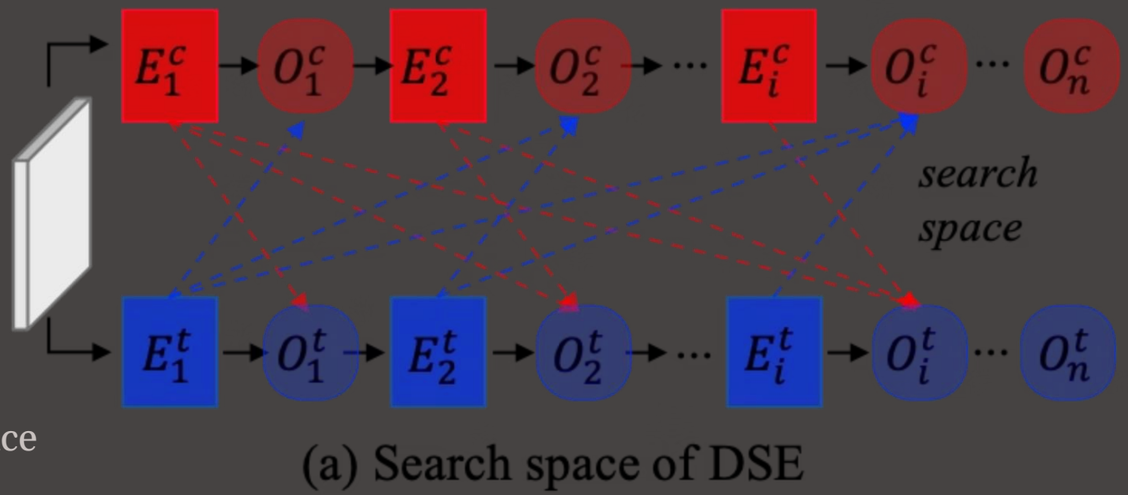
Proposed Framework Overview

- Dynamic Dual-Processing (DDP) architecture consists of:
  - Shared Backbone (common feature extraction).
  - Dual-stream Encoder (CNN + Transformer for parallel processing).
  - Dynamic Dual-Decoder (adaptive choice between CNN and Transformer decoding per image region).



Dual-Stream Encoder

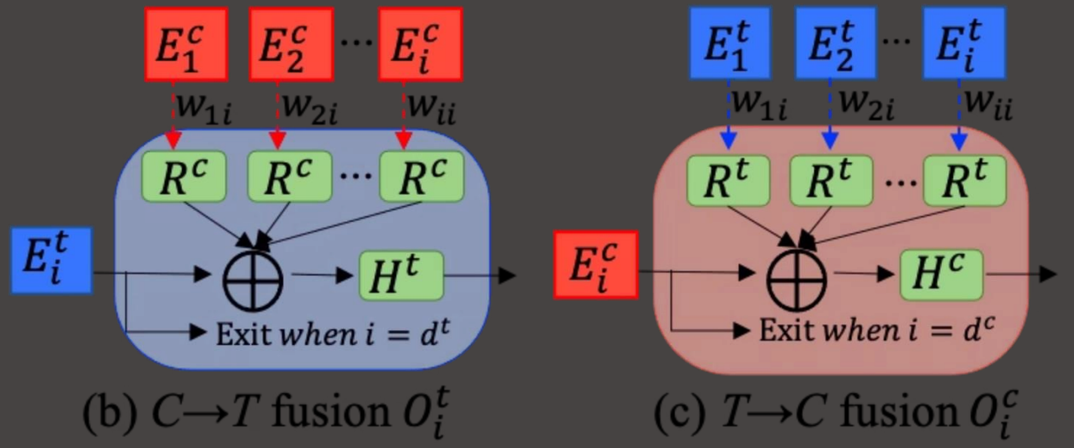
- Two parallel processing streams:
  - CNN pathway (extracts detailed local features).
  - Transformer pathway (captures global context information).
- Intermediate interactions and feature fusion between streams enhance representations.



(a) Search space of DSE

Neural Architecture Search for Feature Fusion

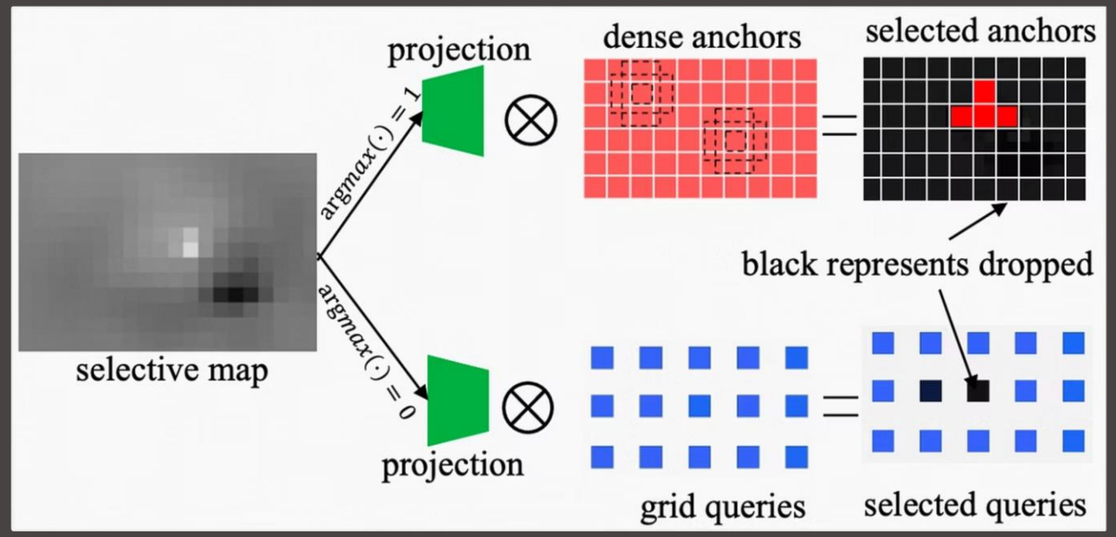
- Automated search (NAS) finds best method for combining CNN and Transformer features.
- Optimizes performance by intelligently merging local details and global context.
- Ensures computational efficiency and effectiveness.



(b) $C \rightarrow T$ fusion O_i^t (c) $T \rightarrow C$ fusion O_i^c

Dynamic Dual-Decoder

- Adaptive decoder comprising:
  - CNN decoder (optimal for clear textures and local details).
  - Transformer decoder (optimal for ambiguous, context-dependent regions).
- Uses selective mask to dynamically assign each object or region to the most suitable decoder.



Selective Mask Mechanism

- Binary selective mask determines whether CNN or Transformer decoder is activated at each location.
- Mask generation involves Gumbel-Softmax trick for differentiable training.
- Dynamically routes detection tasks, significantly boosting detection accuracy without extra computation overhead.

Visualization of Selective Mask Decision-Making

- Visual examples of mask decisions:
  - Blue regions handled by Transformer.
  - Red regions handled by CNN.
- Shows adaptive allocation based on object features (local vs. global).

Multi-Stage Training Strategy

Training involves a careful multi-stage process:

Stand-alone Pre-training

Initializes CNN and Transformer streams independently.

Encoder NAS Search

Finds the optimal encoder feature fusion structure through automated search.

Selective Mask Learning

Trains the selective mask to dynamically decide between CNN and Transformer decoders.

Joint Fine-tuning

Refines the entire model for cohesive performance and overall optimization.

Experimental Setup

- Evaluated on widely-used MS COCO object detection benchmark.
- Measures performance using mean Average Precision (mAP).
- Comparison against state-of-the-art single and hybrid models.

Key Insights:

- **DDP (94G)** matches the top-performing **DN-DETR-R101** in mAP (**45.2**) but uses **46% fewer FLOPs** and has **~60% higher FPS**.

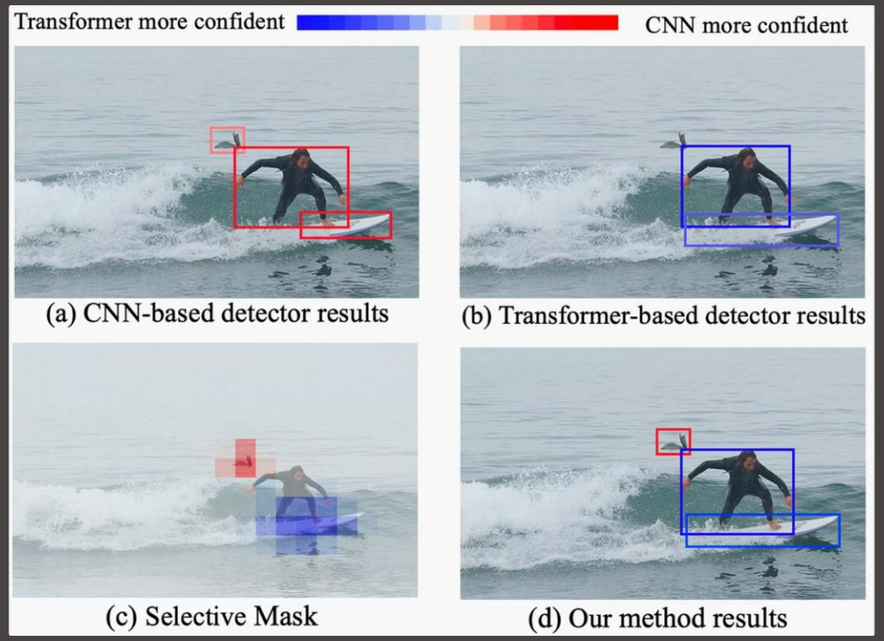
- **DDP (80G)** variant achieves **real-time performance** (41 FPS) with minimal accuracy trade-off, outperforming YOLOF-R50 by **+5.7 mAP** despite similar FLOPs.

- The model is **efficient**, **lightweight**, and **high-performing**, validating the effectiveness of combining CNN and Transformer decoding dynamically.

Model	FLOPs	Params	FPS	mAP
DAB-DETR-R50	94G	44M	21	42.2
YOLOF-R50	86G	44M	39	37.7
DN-DETR-R101	174G	63M	17	45.2
SparseRCNN-R101	206G	125M	19	44.1
DDP(94G)	94G	51M	27	**45.2**
DDP(80G)	**80G**	**42M**	**41**	43.4

Qualitative Examples

- Illustration of dynamic mask behavior:
  - CNN decoder localizes bird effectively due to clear local textures.
  - Transformer decoder identifies surfboard using surrounding context clues.



Advantages of Dynamic Dual-Processing

- Successfully integrates local CNN strengths and global Transformer advantages.
- Selective masking efficiently manages computational resources.
- Significantly improves real-world object detection performance.

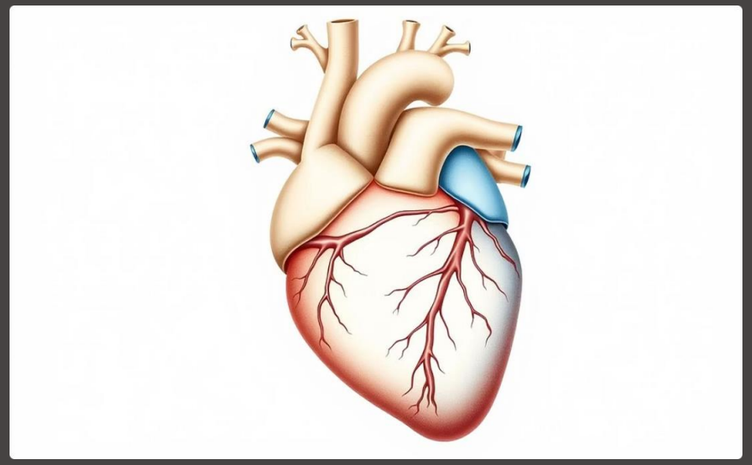
Limitations

- Increased complexity in architecture design (two encoders, two decoders).
- Requires detailed multi-stage training strategy.
- Slight increase in model parameters.

Real-world Impact and Applications

Autonomous Vehicles

Enhances object detection capabilities for safer and more reliable navigation in self-driving cars.



Medical Imaging

Improves diagnostic accuracy and efficiency by precisely identifying anomalies in complex medical scans.

Surveillance Systems

Offers more reliable visual perception under varied and challenging conditions for enhanced security and monitoring.

Future Research Directions

- Develop more streamlined and computationally efficient hybrid models to broaden their applicability.
- Further investigate deeper biological analogies, such as memory and attention mechanisms, to inspire more sophisticated AI architectures.
- Conduct more extensive evaluations across broader and more diverse datasets to ensure robustness and generalization capabilities.

Conclusion

- CNNs are important but have limits; Transformers add useful benefits.
- The Dynamic Dual-Processing (DDP) framework shows how models inspired by biology can work better.
- This is a big step towards smarter computer vision systems that can adapt.

References

- Berrios, W., & Deza, A. (2022). *Joint Rotational Invariance and Adversarial Training of a Dual-Stream Transformer Yields State of the Art Brain-Score for Area V4*.

- Zhang, M., Bu, T., & Hu, L. (2023). *A Dynamic Dual-Processing Object Detection Framework Inspired by the Brain's Recognition Mechanism*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023)*.

Thank You!