



# BÁO CÁO KHÓA LUẬN TỐT NGHIỆP








# PHÂN LỚP ĐA ĐỐI TƯỢNG DỰA TRÊN MÔ HÌNH HỌC SÂU



Giảng viên hướng dẫn:  
TS. Bùi Tiến Lên

Sinh viên thực hiện:  
Hồ Đăng Cao  
Đỗ Đức Duy

## NỘI DUNG

-  Giới Thiệu Bài Toán
-  Phương Pháp Giải Quyết
  -  Công trình đề xuất
  -  Cải tiến đề xuất
  -  Thực nghiệm và Đánh giá
-  Kết Luận Chung
-  Hướng Phát Triển



# GIỚI THIỆU BÀI TOÁN



**Mô hình  
phân lớp  
đa đối tượng  
từ hình ảnh**

Salmon

Bean

Mushroom

⋮

Rice



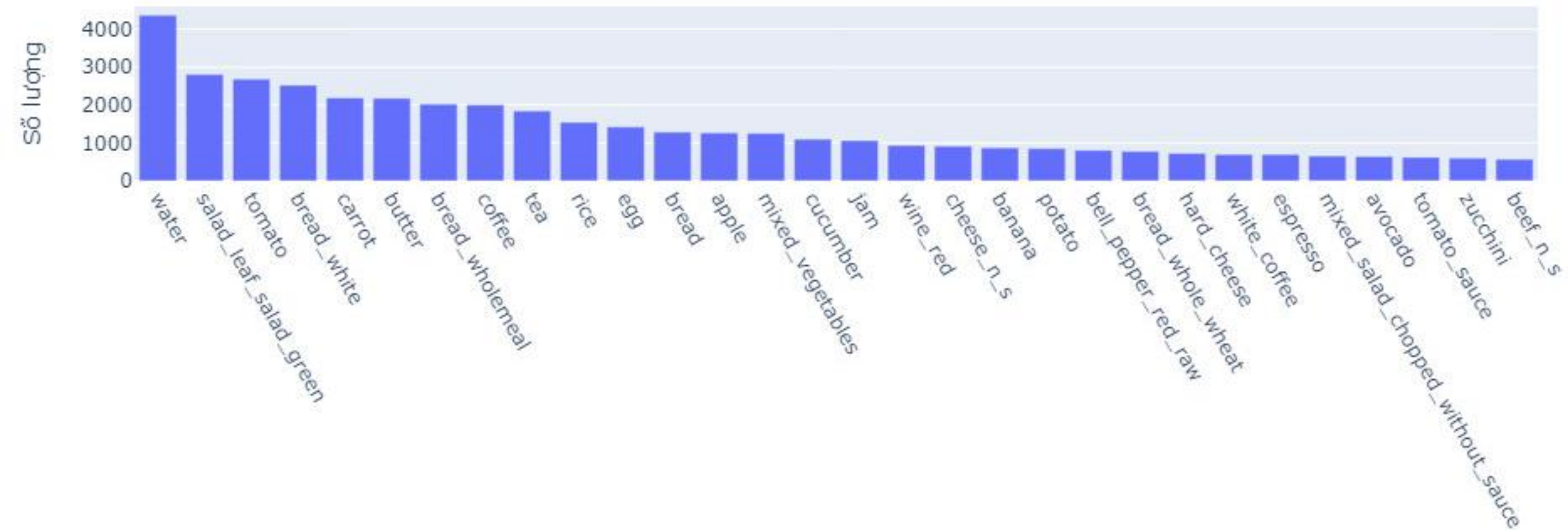
Vấn đề 1: Ảnh xuất hiện các **đối tượng thừa**, ngoài tập nhãn.





## Giới Thiệu Vấn Đề

Vấn đề 2: Số lượng ảnh ở mỗi nhãn có sự chênh lệch lớn.





Vấn đề 3: Các **nhãn** có nghĩa thuộc **cùng trường từ vựng**.

coffee – espresso

bread – bread\_white



coffee



espresso



- **Nguồn gốc:** từ cuộc thi Food Recognition Benchmark 2022.
- **Mô tả:** các bức ảnh về các món ăn trong bữa ăn hằng ngày.
- **Số lượng ảnh:**
  - Tập huấn luyện: 54392.
  - Tập đánh giá: 946.
- **Số lượng nhãn:** 323.





## Tóm Tắt Các Đóng Góp

- Đề xuất sử dụng hai công trình **Đơn nhãn dương** và **C-Tran**.
- Thực hiện **tiền xử lý**, loại bỏ **đối tượng** ngoài **tập nhãn** trên ảnh.
- **Cải tiến, thay đổi** kiến trúc của **Đơn nhãn dương** và **C-Tran**.  
Đồng thời **kết hợp** 2 mô hình.



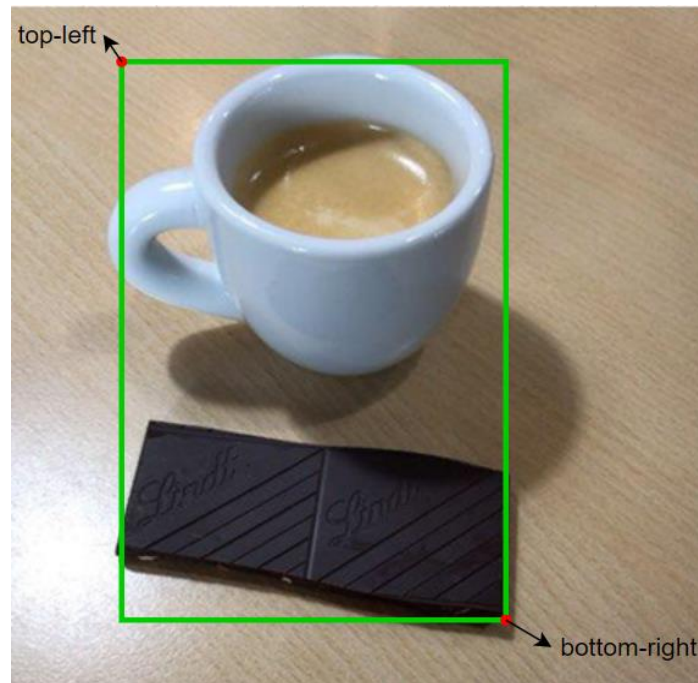
# PHƯƠNG PHÁP GIẢI QUYẾT



## Loại bỏ đối tượng nằm ngoài tập nhãn



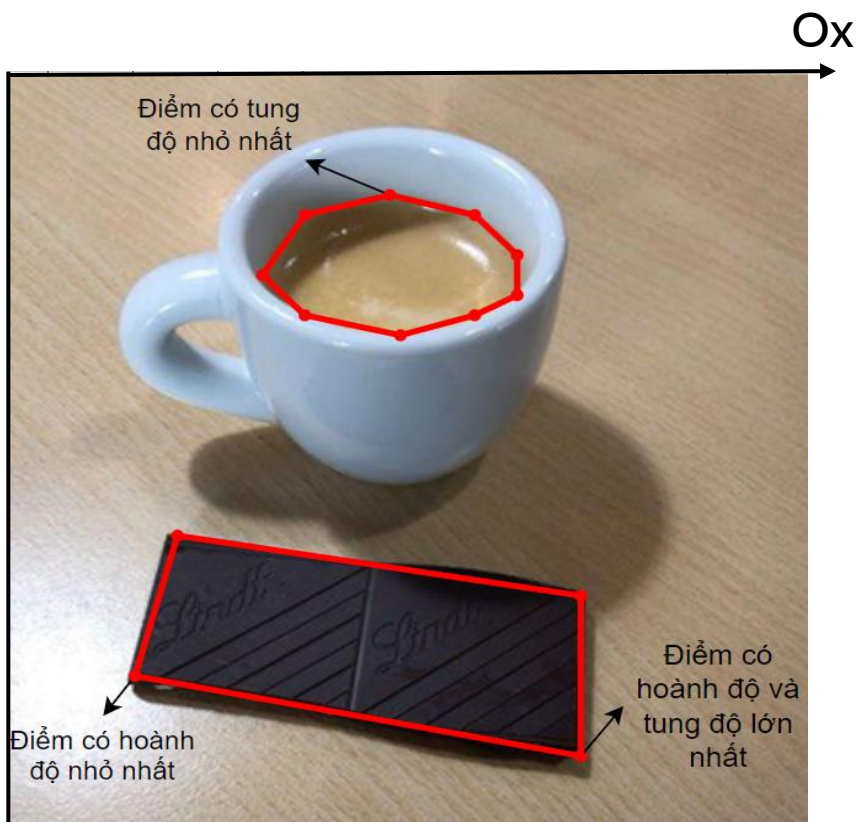
Ảnh gốc



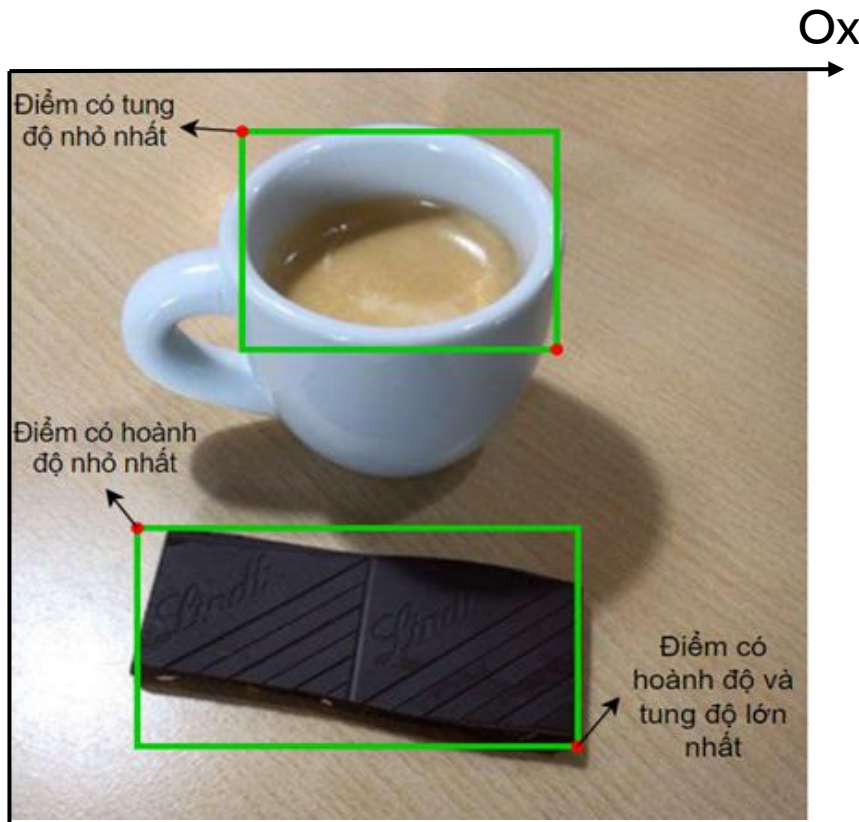
Xác định vùng cắt



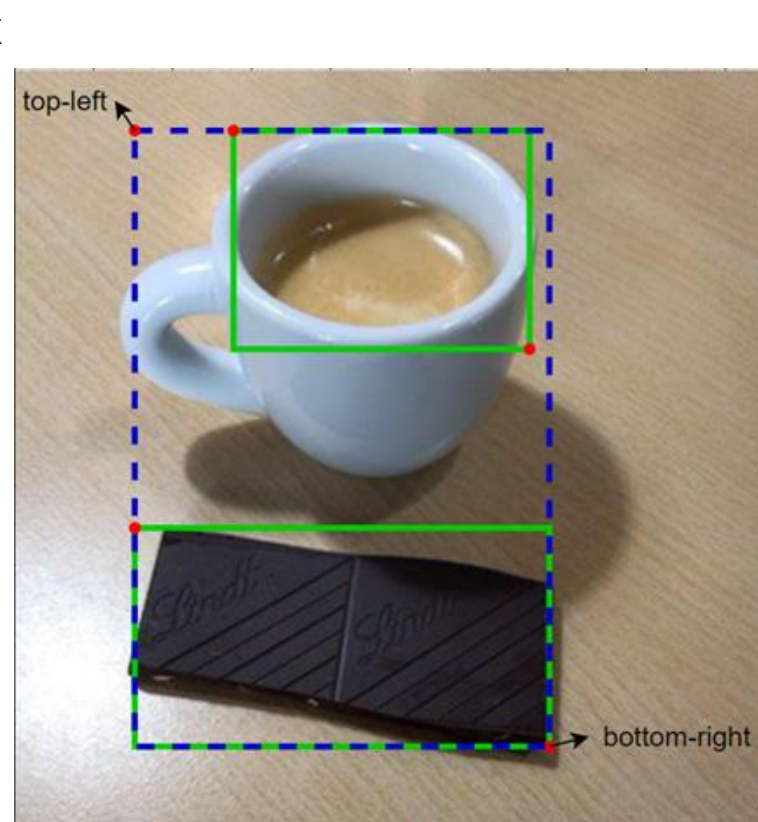
Ảnh sau khi cắt



Xác định từ các **phân vùng**.



Xác định từ các **khung chứa**.



Vùng cắt được xác định

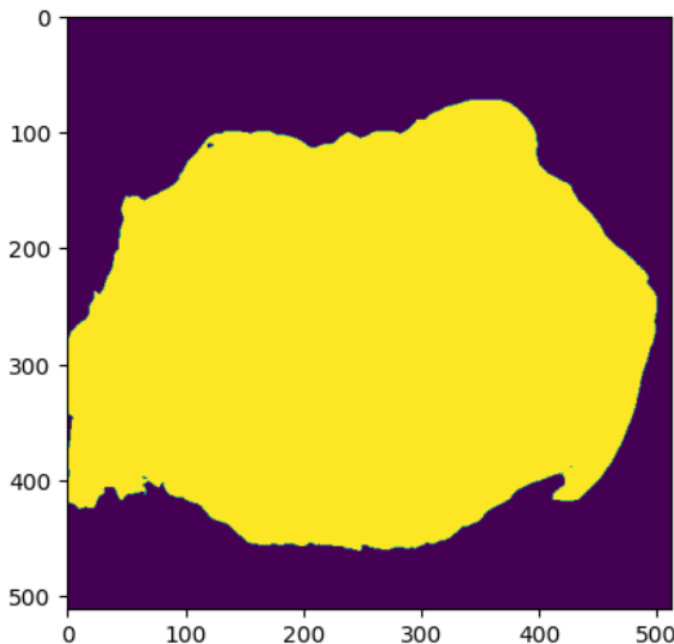




Dùng mô hình được huấn luyện sẵn



Ảnh gốc



Ảnh sau khi qua mạng **U-Net**



Ảnh sau khi cắt



# Công Trình Đề Xuất Đơn nhận dương



## Mục tiêu của công trình

Chỉ cần ghi nhận 1 đối tượng có trong mỗi ảnh cho quá trình huấn luyện.

- vector nhãn của ảnh thứ  $n$  **chỉ có 1 nhãn dương**.

1	∅	∅	...	∅
---	---	---	-----	---

- Hàm mất mát  $\mathcal{L}_{BCE}^+$  đo độ lỗi giữa **vector nhãn** và vector **dự đoán**.

**Hạn chế:** Mô hình sẽ dự đoán tất cả các lớp là **dương**.

**Giải pháp:**

- Bổ sung nhãn âm.
- Phạt dự đoán nhiều nhãn dương.



## Các hàm mất mát - Bổ sung nhãn âm.

Giả sử các nhãn không được quan sát là âm:  $\mathcal{L}_{AN}$  → **hiều nhãn** làm giảm độ chính xác.

- Thêm **trọng số**:  $\mathcal{L}_{WAN}$
- Kết hợp **làm mịn nhãn** ( $LS$ ) cho mỗi lớp :  $\mathcal{L}_{AN-LS}$





## Các hàm mất mát - Phạt dự đoán nhiều nhãn dương.

Điều chuẩn dương kì vọng:  $\mathcal{L}_{EPR}$

$\mathcal{L}_{BCE}^+$

$R_k$  →  $\neq$

$k$  số nhãn dương kỳ vọng

$\hat{k}$  số nhãn dương trung bình dự đoán

vector nhãn ước lượng

Bộ Phận Ước Lượng Nhãn

0.8

...

0.12

Ước lượng nhãn trực tuyến điều chuẩn:  $\mathcal{L}_{ROLE}$

$\mathcal{L}_{BCE}$

$\mathcal{L}_{EPR}$

vector nhãn ước lượng

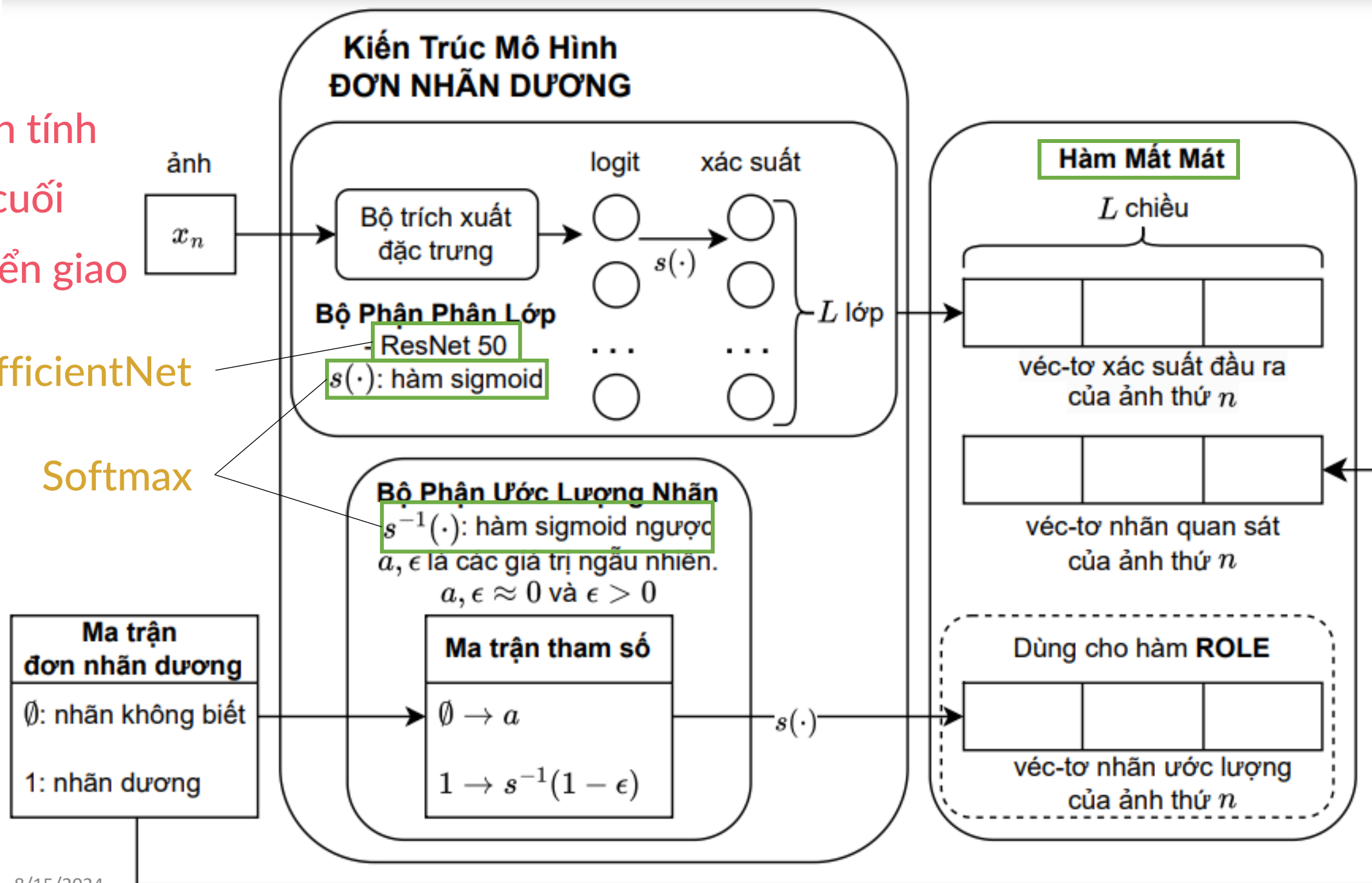
$\neq$

Vector dự đoán

Tuyển tính  
Đầu cuối  
Chuyển giao

# EfficientNet

# Softmax





## Cải Tiến Đề Xuất Đơn nhận dương



Chuyển vector nhần  $z$  sang các giá trị xác suất.  
Nhần không biết xem như nhần âm.

1	0	0	...	0
---	---	---	-----	---

**Mục tiêu:**  $x_i = \sigma^{-1}(z_i)$

Áp dụng kĩ thuật làm mịn nhần.

$1 - \epsilon$	$\frac{\epsilon}{L - 1}$	$\frac{\epsilon}{L - 1}$	...	$\frac{\epsilon}{L - 1}$
----------------	--------------------------	--------------------------	-----	--------------------------

$\parallel$   $\parallel$   
 $\sigma(x^+)$   $\sigma(x^-) \rightarrow x^+$  phụ thuộc  $x^-$ .



Huber:  $\mathcal{L}_{HU}$   $\begin{cases} \text{MSE} \\ \text{MAE} \end{cases}$

Focal:  $\mathcal{L}_{FO} = \mathcal{L}_{AN}$   $\begin{cases} \alpha_i: \text{nghịch đảo tần suất xuất hiện của nhãn thứ } i \\ (1 - f_{ni})^\gamma: \text{dự đoán càng sai thì phạt càng nặng.} \end{cases}$



## Thực Nghiệm và Đánh Giá Đơn nhân dương



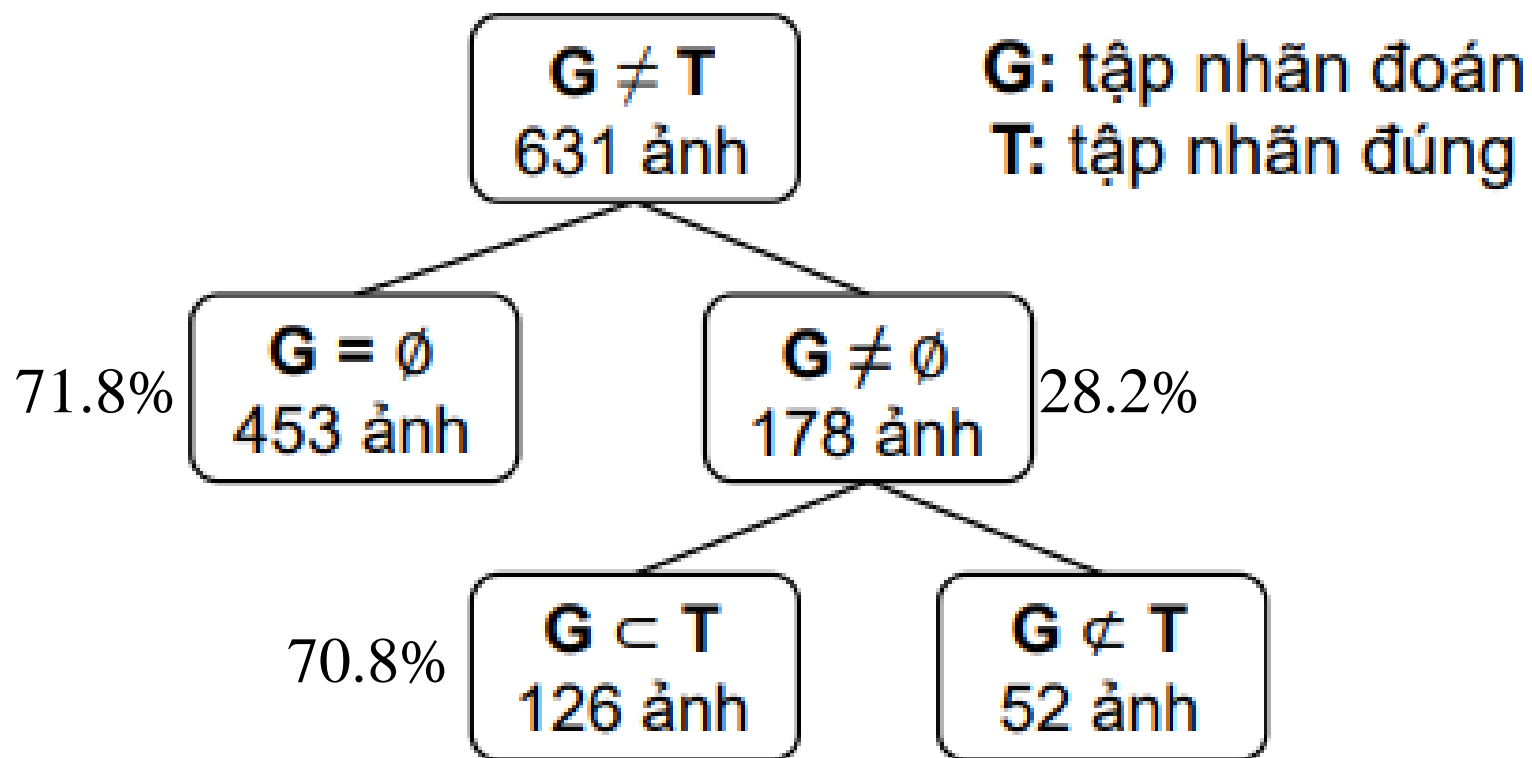
Các kết quả tốt nhất với **ResNet 50** giữa:

- Các tốc độ học ( $10^{-3}$ ,  $10^{-4}$ ,  $10^{-5}$ ).
- Các kích thước lô (8, 16, 32).
- Trên tập dữ liệu gốc (chưa cắt).

Hàm mất mát	Chế độ huấn luyện	Hàm kích hoạt	mAP tập kiểm tra
FO	đầu cuối	sigmoid	0.7938
ROLE	đầu cuối	sigmoid	26.5441
HU	đầu cuối	sigmoid	32.6947
AN-LS	đầu cuối	sigmoid	34.6173
AN-LS	chuyển giao	sigmoid	34.8327
HU	chuyển giao	softmax	1.5206



Ở ngưỡng phân lớp 0.5







Số lượng	Các cặp (nhãn đúng, nhãn đoán)
4	(water, soft_drink_with_a_taste), (espresso, coffee)
3	(water, glucose_drink_50g), (espresso, ristretto_with_caffeine)
2	(water, water_with_lemon_juice), (bread_wholemeal, bread_whole_wheat), (mixed_salad_chopped_without_sauce, salad_leaf_salad_green), (coffee, white_coffee), (coffee, ristretto_with_caffeine)

Các **nhóm** dễ nhầm lẫn gồm **nước**, **cà phê**, **bánh mì** và **salad**.

Các nhãn **tương đồng** về **nghĩa**. → **Ảnh hưởng** đến **độ tin cậy** về **kết quả** của mô hình.



water



soft\_drink\_with\_a\_taste



glucose\_drink\_50g



water\_with\_lemon\_juice

→ Ở mức con người cũng khó có thể phân biệt được các bức ảnh cùng nhóm.



**Ưu điểm:** làm tốt trong vấn đề trích xuất và nhận diện các đặc trưng ảnh với số lượng nhãn cần đánh thấp.

**Hạn chế:** các vấn đề về ý nghĩa nhãn làm ảnh hưởng lớn đến kết quả phân lớp.

**Giải pháp:** cần một mô hình có thể học được mối liên hệ giữa các nhãn.

→ C-Tran.



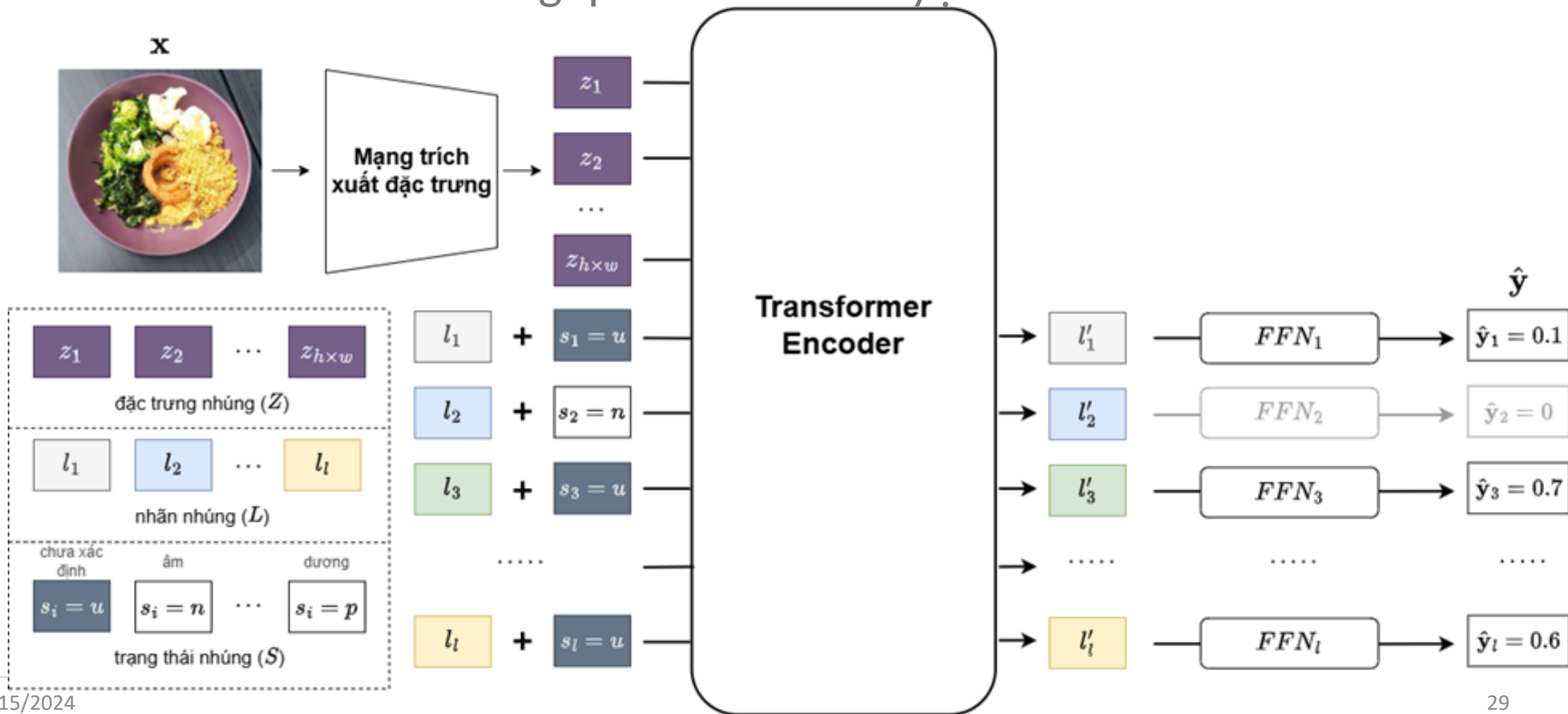
## Công Trình Liên Quan

# C-Tran



# Tổng quan

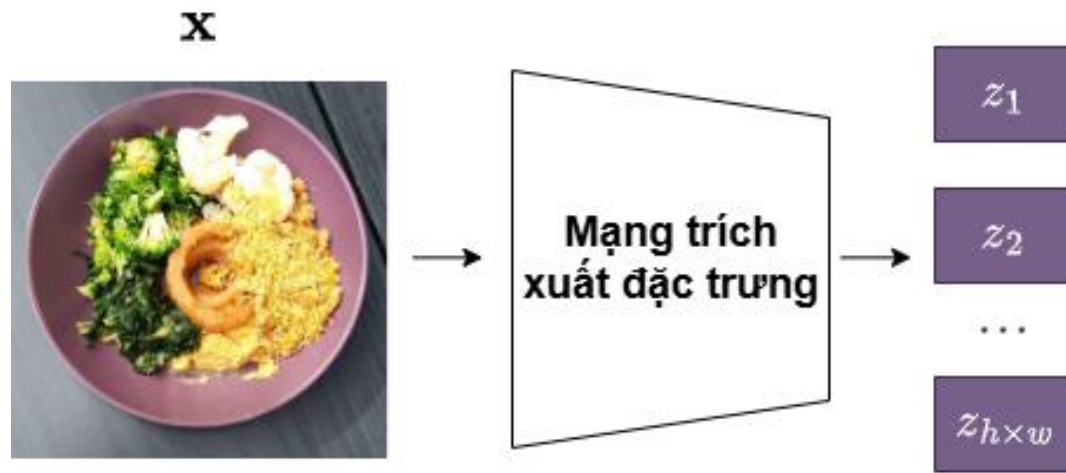
- Khai thác sự phụ thuộc giữa các đặc trưng và nhãn trong hình ảnh.
- Che nhãn hình ảnh trong quá trình huấn luyện.





## Các thành phần chính

### 1. Nhúng đặc trưng $Z$

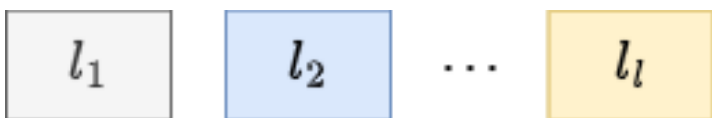


Các vector  $z_i \in Z = \mathbb{R}^d$ , đại diện cho các vùng được ánh xạ từ các mảng không gian gốc của ảnh thông qua mạng trích xuất đặc trưng.



## Các thành phần chính

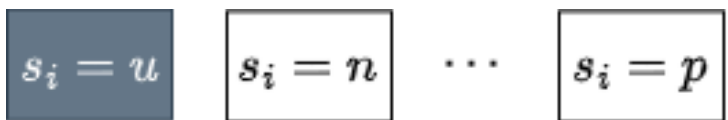
### 2. Nhúng nhãn L



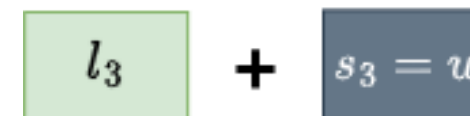
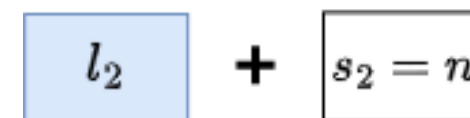
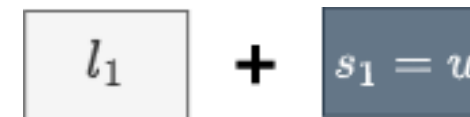
$L = \{l_1, l_2, \dots, l_l\}, l_i \in \mathbb{R}^d$ , đại diện cho các nhãn  $l$  có thể có trong tập dữ liệu.

### 3. Thêm thông tin về nhãn thông qua nhúng trạng thái S

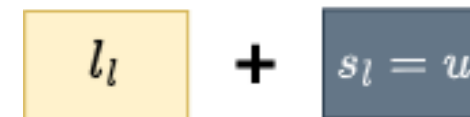
$$\tilde{l}_i = l_i + s_i$$



$s_i \in \{u, n, p\}$ : không xác định ( $u$ ), âm ( $n$ ), dương ( $p$ ).



.....

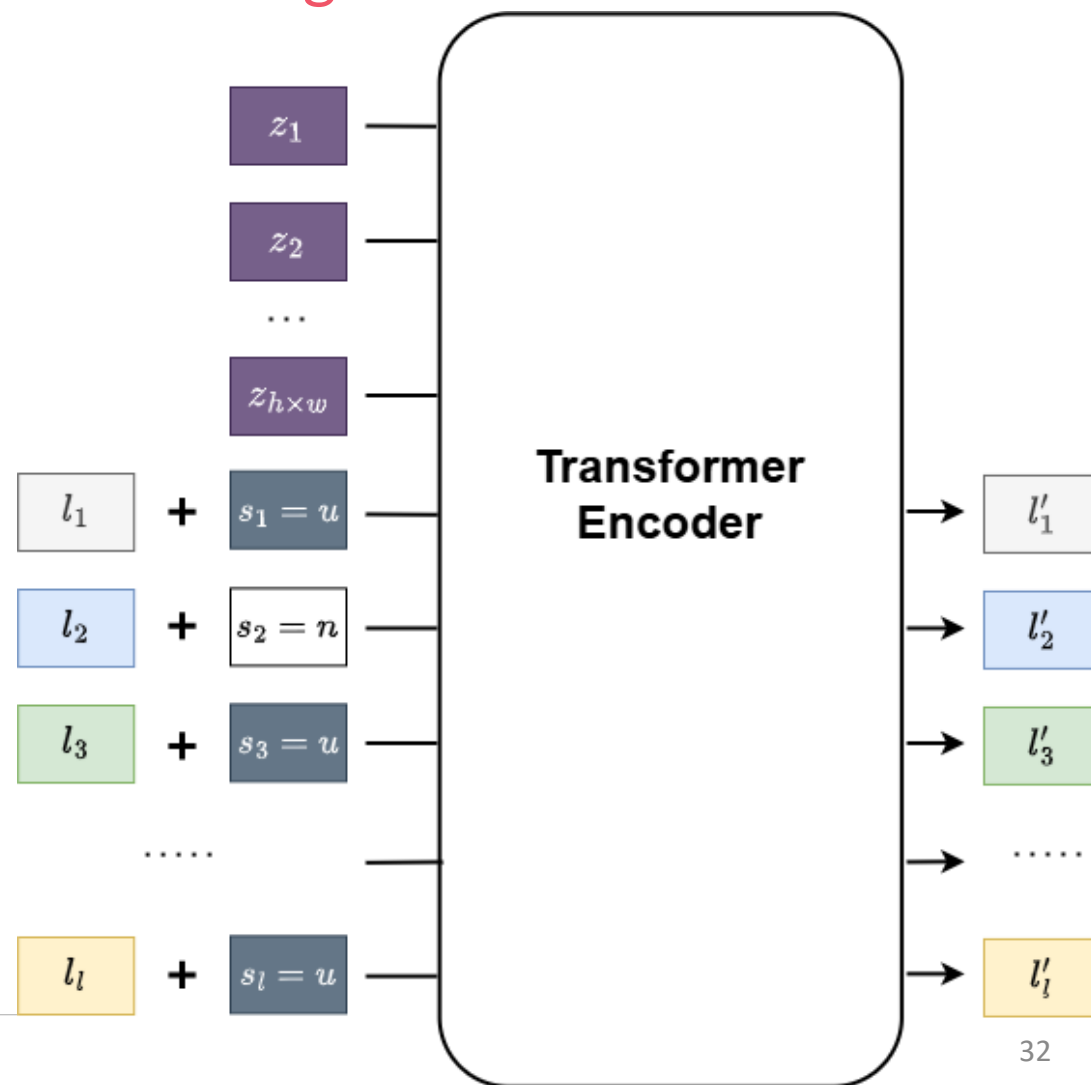




## Các thành phần chính

### 4. Mô hình hóa sự tương tác giữa đặc trưng và nhãn bằng Transformer Encoder

- $H = \{z_1, \dots, z_{h \times w}, \tilde{l}_1, \dots, \tilde{l}_l\}$  là đầu vào của Transformer Encoder.
- Đầu ra là  $H' = \{l'_1, \dots, l'_l\}$

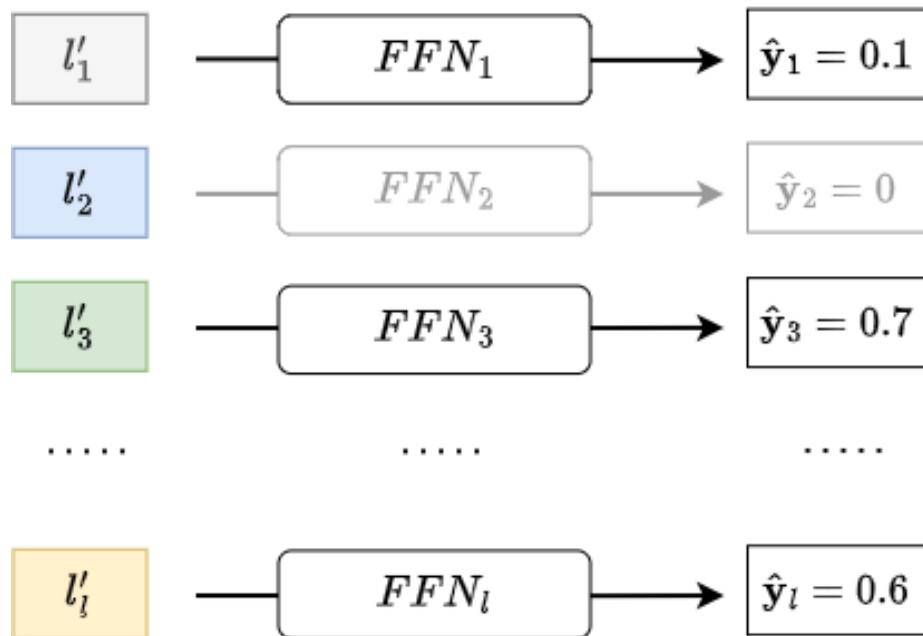






## Các thành phần chính

### 5. Quá trình suy luận để phân loại nhãn



### 6. Hàm mất mát

$L_{BCE}$ : Binary Cross Entropy

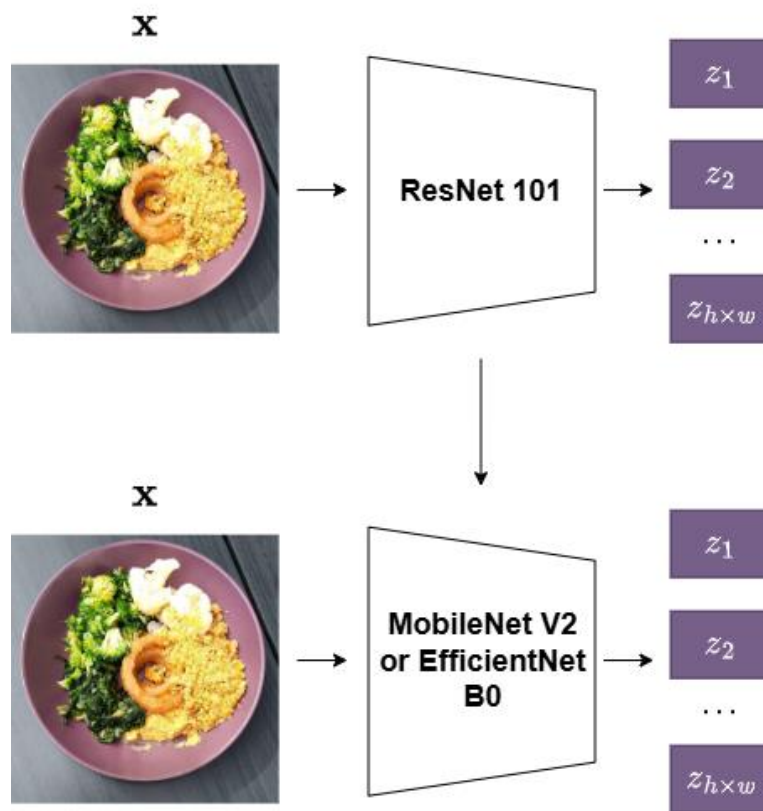
Mạng chuyển tiếp độc lập  $FFN_i$  cho  $l'_i$  gồm 1 lớp tuyến tính.  
Sau đó, dùng hàm **sigmoid** để tính giá trị xác suất cho các nhãn  $l'_i$ .



## Cải Tiến Đề Xuất **C-Tran**



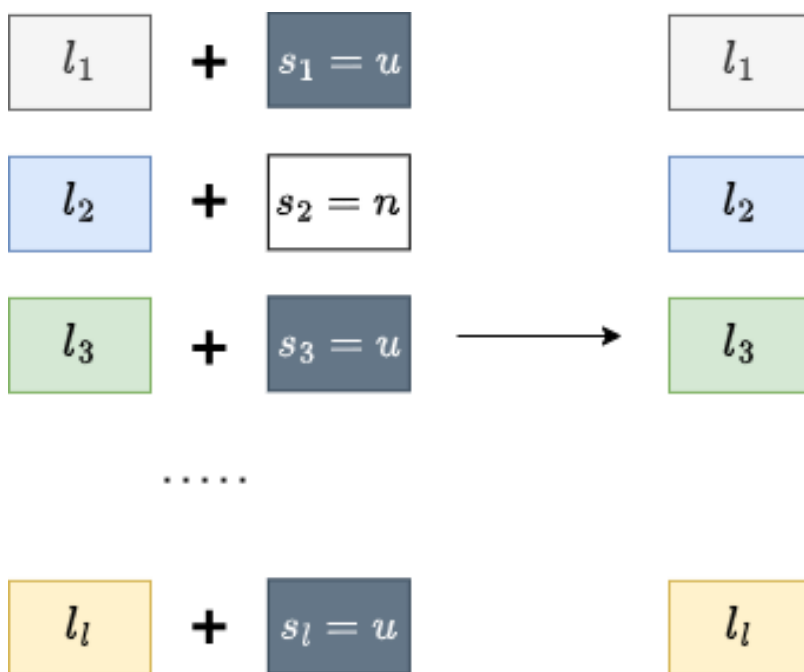
# Thay đổi mạng trích xuất đặc trưng



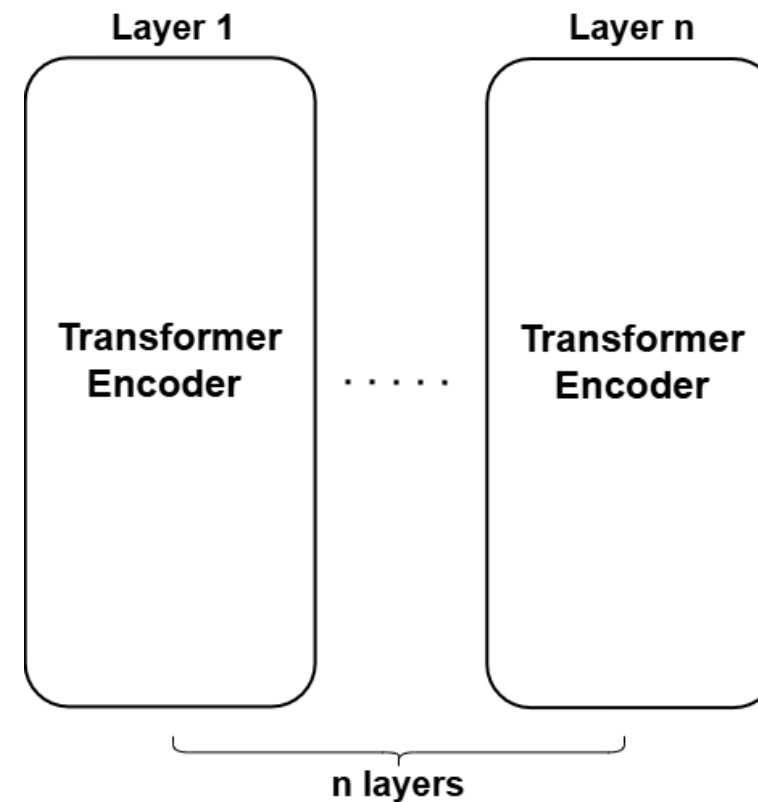
Các mạng mới tối ưu về mặt **số lượng tham số**.



Loại bỏ thông tin trạng thái nhần.



Thay đổi số lượng lớp Encoder.





## Thực Nghiệm và Đánh Giá **C-Tran**



## Bảng kết quả



Mạng trích xuất đặc trưng	Hàm kích hoạt	Số lớp En-coder	Che nhân huấn luyện	Lượng nhân biết trước	Trạng thái nhân	Kết quả kiểm tra
ResNet 101	sigmoid	3	có	0	tổng	91.3
ResNet 101	softmax	3	có	0	tổng	90.1
ResNet 101	sigmoid	3	có	243	tổng	91.3
EfficientNet B0	sigmoid	3	có	0	tổng	91.3
MobileNet V2	sigmoid	3	có	0	tổng	91.3
MobileNet V2	sigmoid	3	có	243	tổng	91.3
MobileNet V2	sigmoid	3	có	0	tích	90.6
MobileNet V2	softmax	3	có	0	tổng	89.8
MobileNet V2	sigmoid	4	có	0	tổng	91.3
MobileNet V2	sigmoid	2	không	0	tổng	91.3
MobileNet V2	sigmoid	2	có	0	tổng	91.3

Mô hình gốc  
(~42.5M)

Mô hình đề xuất  
(~2.5M-4.3M)

Kết quả kiểm tra là độ chính xác **mAP**.



# KẾT LUẬN CHUNG



## Kết Luận Chung

Cắt vùng dư thừa trong ảnh,  
nâng cao chất lượng phân loại.

Cải tiến các mô hình và  
thu được kết quả tốt với  
kích thước nhỏ.



Đã giải  
quyết





## Kết Luận Chung

Tối ưu vấn đề ngữ nghĩa  
nhân ở Đơn nhân dương.

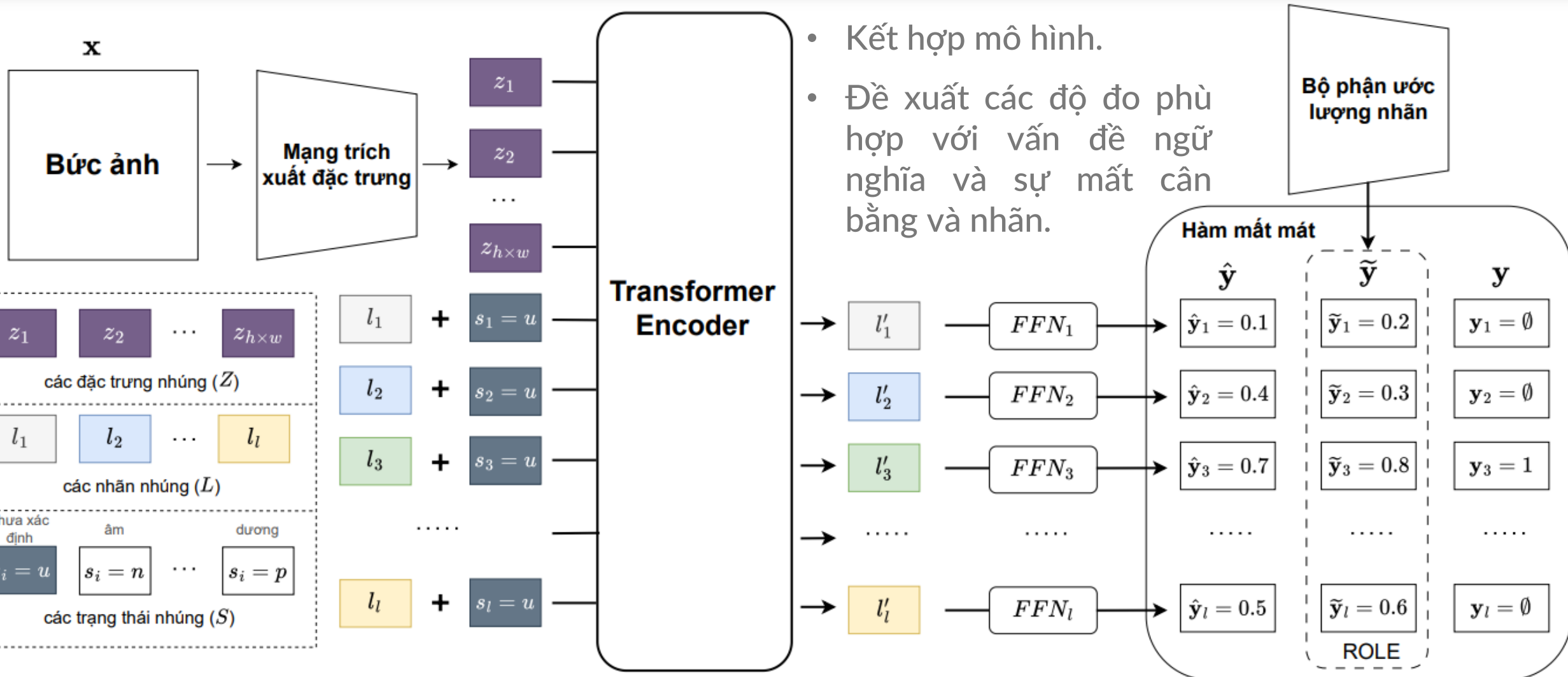
Thử nghiệm mô hình kết hợp.



Chưa giải  
quyết



# HƯỚNG PHÁT TRIỂN



XIN CẢM ƠN  
QUÝ THẦY CÔ  
ĐÃ LẮNG NGHE