In Pursuit of Causal Label Correlations for Multi-label Image Recognition

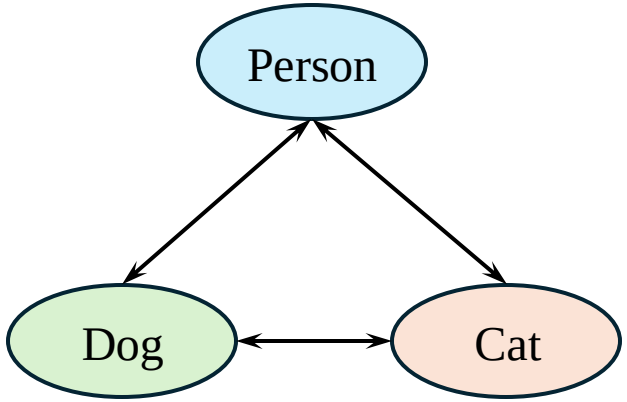
Zhao-Min Chen[1] Xin Jin[2] Yisu Ge[1,*] Sixian Chan[3]

[1]Wenzhou University [2]Samsung Electronic [3]Zhejiang University of Technology

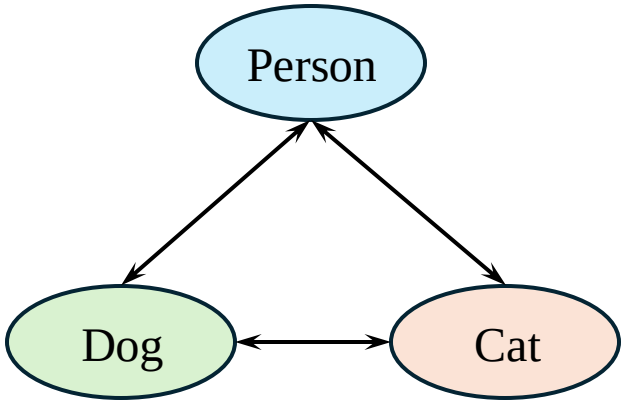
Background

- Traditional multi-label recognition methods assumed that the training and test sets follow independent and identically distributions (i.i.d.), and the label correlations are consistent.

Training



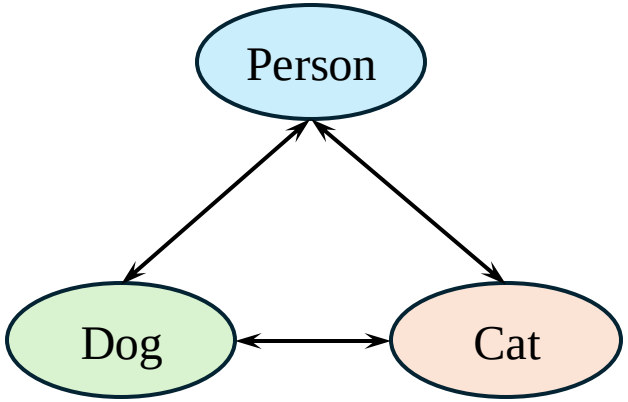
Testing



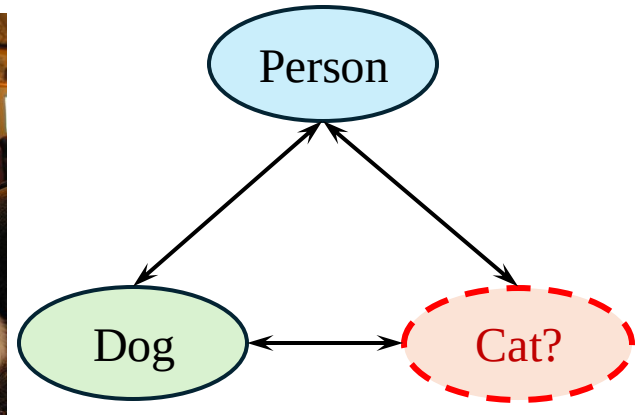
Background

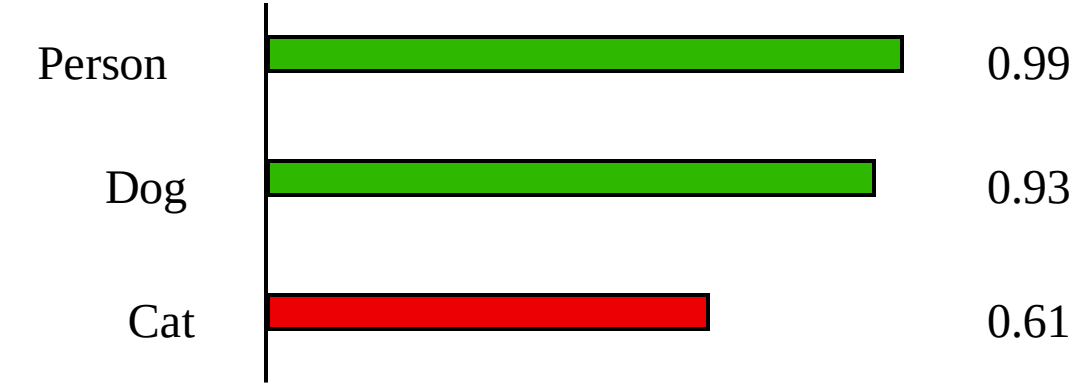
- Previous multi-label recognition methods may fall short when there exists contextual bias in the training set.

Training



Testing



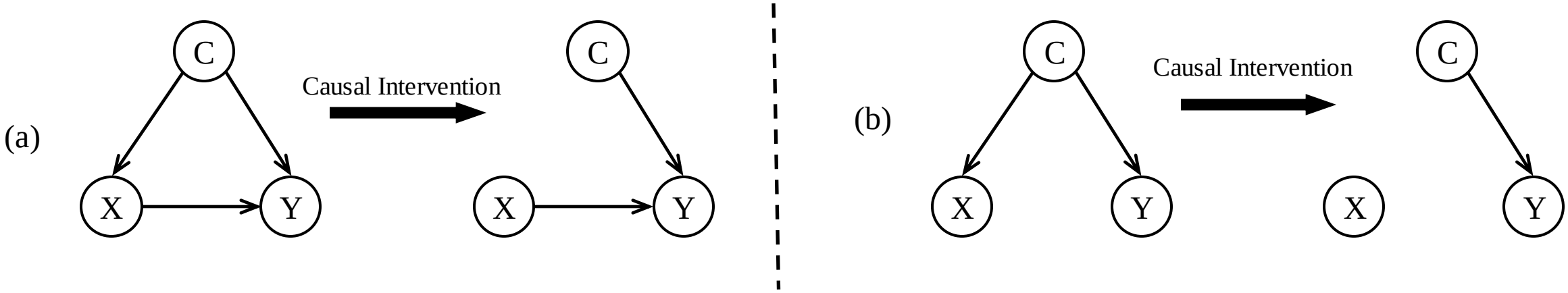


Motivation

- The causal label correlations are stable
- Pursuing causal correlations
- Suppressing spurious correlations

Motivation

- Causal intervention theory can remove the effects of confounders.
- After applying causal intervention, causal correlations will be preserved, spurious correlations will be removed.



Method

- Statement:
  - If P(Y | do(X)) > P(Y), then a causal correlation exists from X to Y in a probability-raising sense, where X and Y represent two labels.

- Backdoor adjustment (C denote the confounders):

$$P(Y|do(X)) = \sum_{c} P(Y|X, C = c)P(C = c)$$

Method

- Estimate the probability of each category:

$$\hat{y}^j_{causal} = f_{merge}([P(Y_j|do(X_1), ..., P(Y_j|do(X_N)])$$

- Normalized weighted geometric mean approximation:

$$\begin{aligned} P(Y_j|do(X_i)) &= \mathbb{E}_c[\sigma(f_{y_j}(x_i, c))] \\ &\approx \sigma(\mathbb{E}_c[f_{y_j}(x_i, c)]) \\ &= \sigma(\sum_c f_{y_j}(x_i, c) \cdot P(c)) \end{aligned}$$

Method

- Effective modeling by cross-attention:

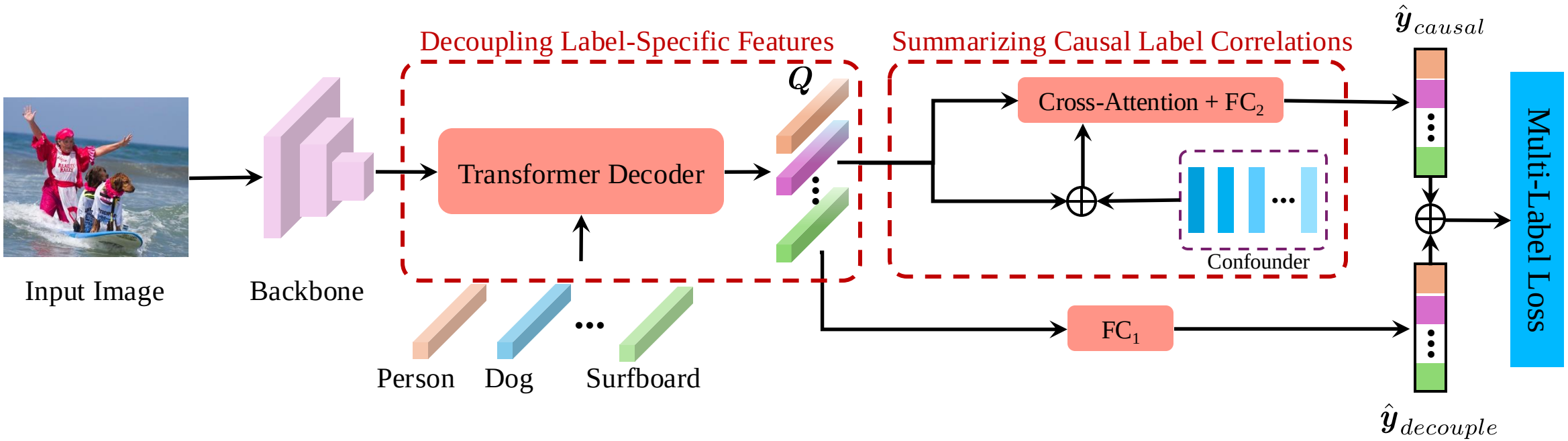
$$\mathbf{Z}_c = \mathbf{X} + c,$$

$$\begin{aligned}
\hat{y}^j_{causal} &= f_{merge}([P(Y_j|do(X_1), ..., P(Y_j|do(X_N)]) \\
&= f_{merge}([\sigma(\sum_c f_{y_j}(x_1, c) \cdot P(c)), ..., \sigma(\sum_c f_{y_j}(x_N, c) \cdot P(c))]) \\
&\approx \sigma(\sum_c f_{y_j}(\mathbf{X}, c) \cdot P(c)) \\
&= \sigma(\sum_c f_{fc2}(f_{cross_atten}(y_j, \mathbf{Z}_c, \mathbf{Z}_c)) \cdot P(c))
\end{aligned}$$

Where **X** denote the label-specific features. We model the confounders by clustering spatial features with K-means algorithm

Method

- Framework Overview



Experiments

Common Setting

Method	COCO-Stuff (mAP)			Deepfashion (top-3 recall)		
	Exclusive	Co-occur	All	Exclusive	Co-occur	All
Q2L [19]	23.5	67.1	57.2	12.8	26.3	26.1
ADD-GCN [12]	20.6	64.8	55.2	8.2	22.6	23.5
ML-GCN [4]	18.6	67.1	55.1	10.3	23.7	24.0
SSGRL [28]	18.1	66.6	54.9	7.9	22.8	23.1
C-Tran [15]	22.4	65.1	55.4	11.4	24.6	24.8
CCD [17]	23.8	65.3	55.9	11.5	24.2	24.6
TDRG [38]	20.0	64.8	56.2	8.1	22.9	23.6
IDA [18]	25.2	64.9	57.0	11.3	25.1	25.4
CAM-Based [14]	26.4	64.9	–	–	–	–
feature-split [14]	28.8	66.0	–	9.2	20.1	–
Baseline (R50)	21.9	65.5	55.0	11.5	24.1	24.1
Ours	**29.7**	**69.6**	**60.6**	**14.6**	**27.4**	**28.8**

Experiments

Cross-dataset Setting

Method	MS-COCO → NUS-WIDE	NUS-WIDE → MS-COCO
ADD-GCN [12]	81.8	77.2
ML-GCN [4]	81.4	77.2
SSGRL [28]	80.2	76.1
C-Tran [15]	80.9	76.9
CCD [17]	81.9	78.3
Q2L [19]	82.1	78.6
IDA [18]	82.3	78.9
CAM-Based [14]	81.0	77.8
feature-split [14]	81.9	78.3
Baseline (R101)	81.1	77.1
Ours	**83.2**	**80.2**

Experiments

Ablation Studies

Table 3: The impacts of different modules.

Decouple	Causal	Exclusive	Co-occur	All
		21.9	65.5	55.0
√		22.1	67.0	56.8
√	√	**29.7**	**69.6**	**60.6**

Table 4: The impacts of clustering center number.

Number	Exclusive	Co-occur	All
20	26.9	68.9	59.6
40	26.8	69.3	60.1
60	29.3	**69.8**	60.2
80	**29.7**	69.6	**60.6**
100	29.5	69.4	60.5

Table 5: The impact of backbones for clustering.

Confounder Backbone	Exclusive	Co-occur	All
ResNet-50	**29.7**	69.6	**60.6**
ResNet-101	29.6	**69.9**	60.5
BEIT3-Large	29.4	69.7	60.5

Table 6: The impact of different modeling approaches for confounders.

Method	Exclusive	Co-occur	All
Random	22.1	66.3	56.1
Early	27.8	69.1	60.1
Label	28.0	69.3	60.3
K-means	**29.7**	**69.6**	**60.6**

Thank You for Your Attention !