南京航空航天大學
Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

De-biasing with Causal Theory

Tang K, Niu Y, Huang J, et al. Unbiased scene graph generation from biased training[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR).

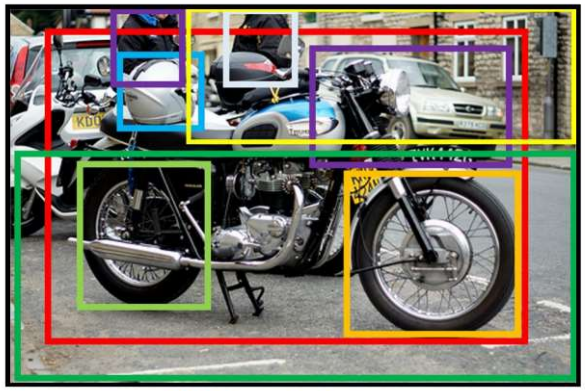
Sun S, Zhi S, Liao Q, et al. Unbiased scene graph generation via two-stage causal modeling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI), 2023.

Certain field

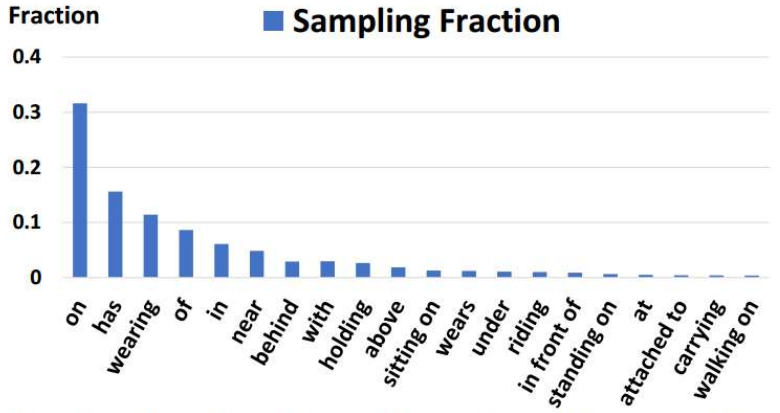
- Existing problems of SGG

  1) Long-Tailed Distribution of Dataset

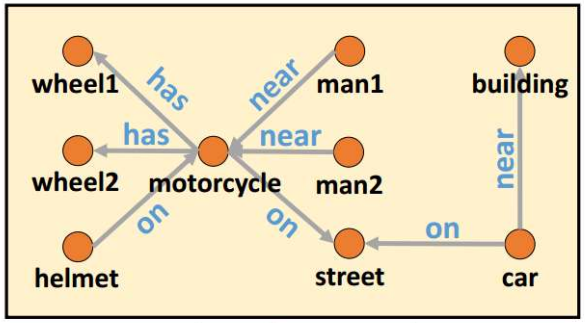
  2) semantic relationship confusion/ coarse-grained predicate



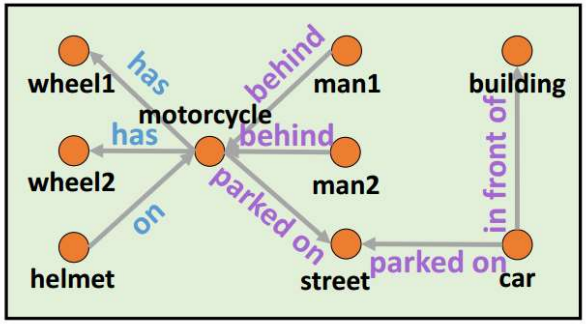
(a) Input Image



(b) Distribution of Predicate Sampling Fraction



Biased Generation



Unbiased Generation

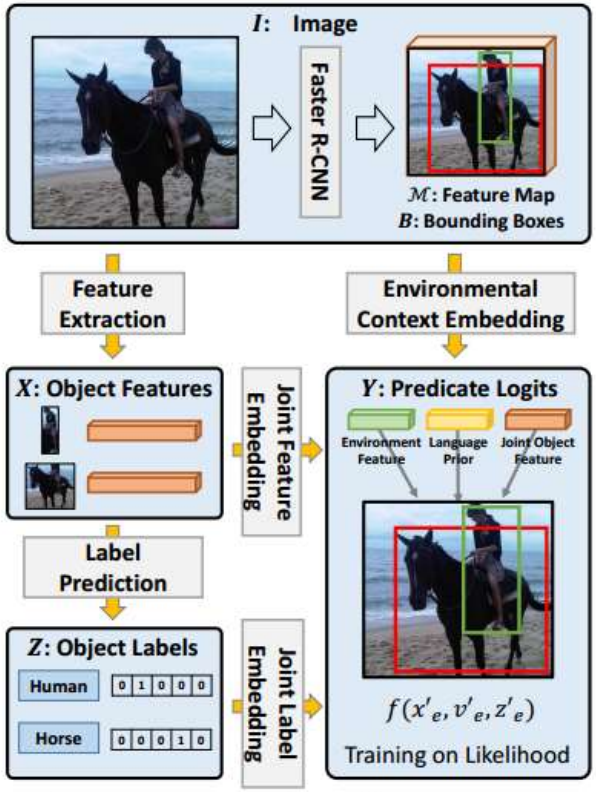
Unbiased Scene Graph Generation from Biased Training

Kaihua Tang[1], Yulei Niu[3], Jianqiang Huang[1,2], Jiaxin Shi[4], Hanwang Zhang[1]
[1]Nanyang Technological University, [2]Damo Academy, Alibaba Group, [3]Renmin University of China, [4]Tsinghua University

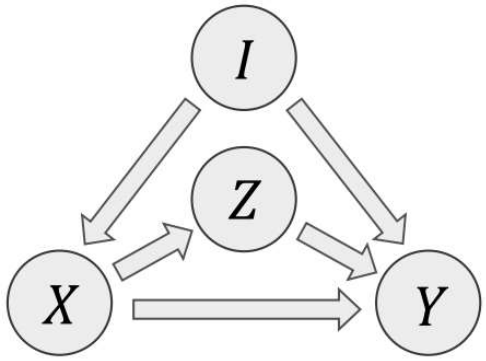
Motivation

Solid points[1]**:**

- Using Total Direct Effect (TDE) to remove bias
- Modeling causal diagrams on Scene Graph Generation(SSG)



The SGG Framework Used for Biased Training



The Causal Graph of SGG

I : Image

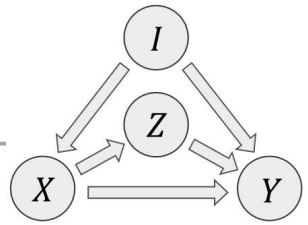
Z : Object Labels

X : Object Features

Y : Predicate Logits

[1] Tang K, Niu Y, Huang J, et al. Unbiased scene graph generation from biased training(CVPR).

Implementation



The Causal Graph of SGG

Variables ： {I、X、Z、Y}

(1) **I** (Input Image & Backbone)：

Faster R-CNN outputs a set of **bounding boxes** B and **feature map** M

(2) **X** (Object Feature)：

It consists of the **pairwise** object feature x_e

(3) **Z** (Object Class)：

It consists of the **pairwise** object label z_e

(4) **Y** (Predicate Classification)：

It generates by combining **three** variables through the **fusion function** such as $y_e = W_r x'_e \cdot \sigma(W_x x'_e + W_v v'_e + z'_e)$

(1) **I→X** (Object Feature Extractor)：

$$Input : \{(r_i, b_i, l_i)\} \xrightarrow{\text{Bi-LSTM}} Output : \{x_i\}$$

(2) **X→Z** (Object Classification)：

$$Input : \{x_i\} \xrightarrow{\text{Bi-LSTM}} Output : \{z_i\}$$

(3) **X→Y** (Object Feature Input for SGG)：

$$Input : \{x_e\} \xrightarrow{\text{merge}} Output : \{x'_e\}$$

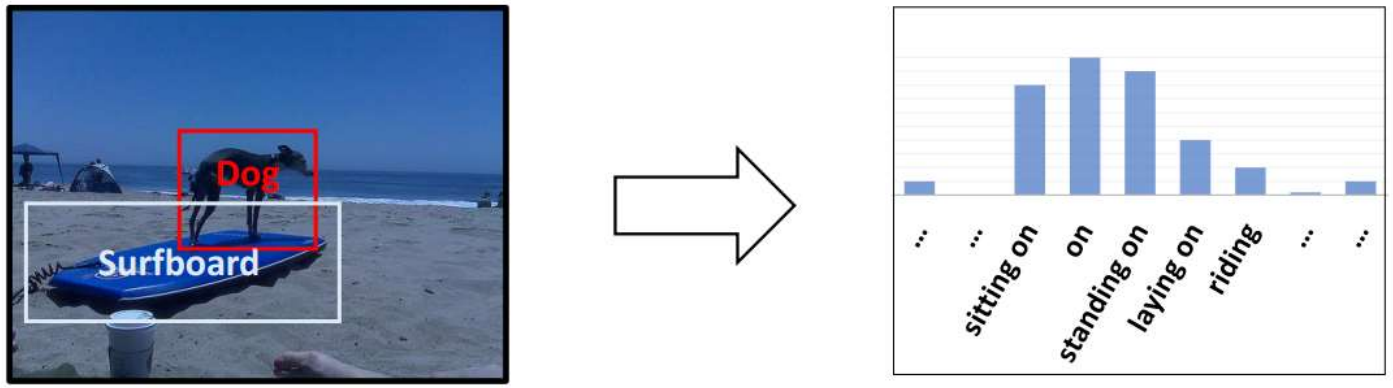
(4) **Z→Y** (Object Class Input for SGG)：

Joint embedding label z'_e

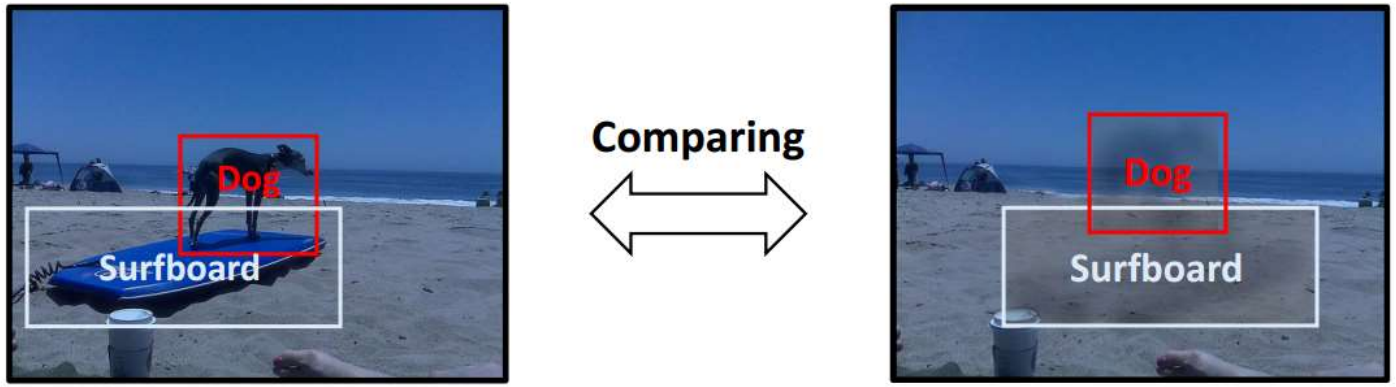
(5) **I→Y** (Visual Context Input for SGG)：

Contextual union region features v'_e

Counterfactual thinking

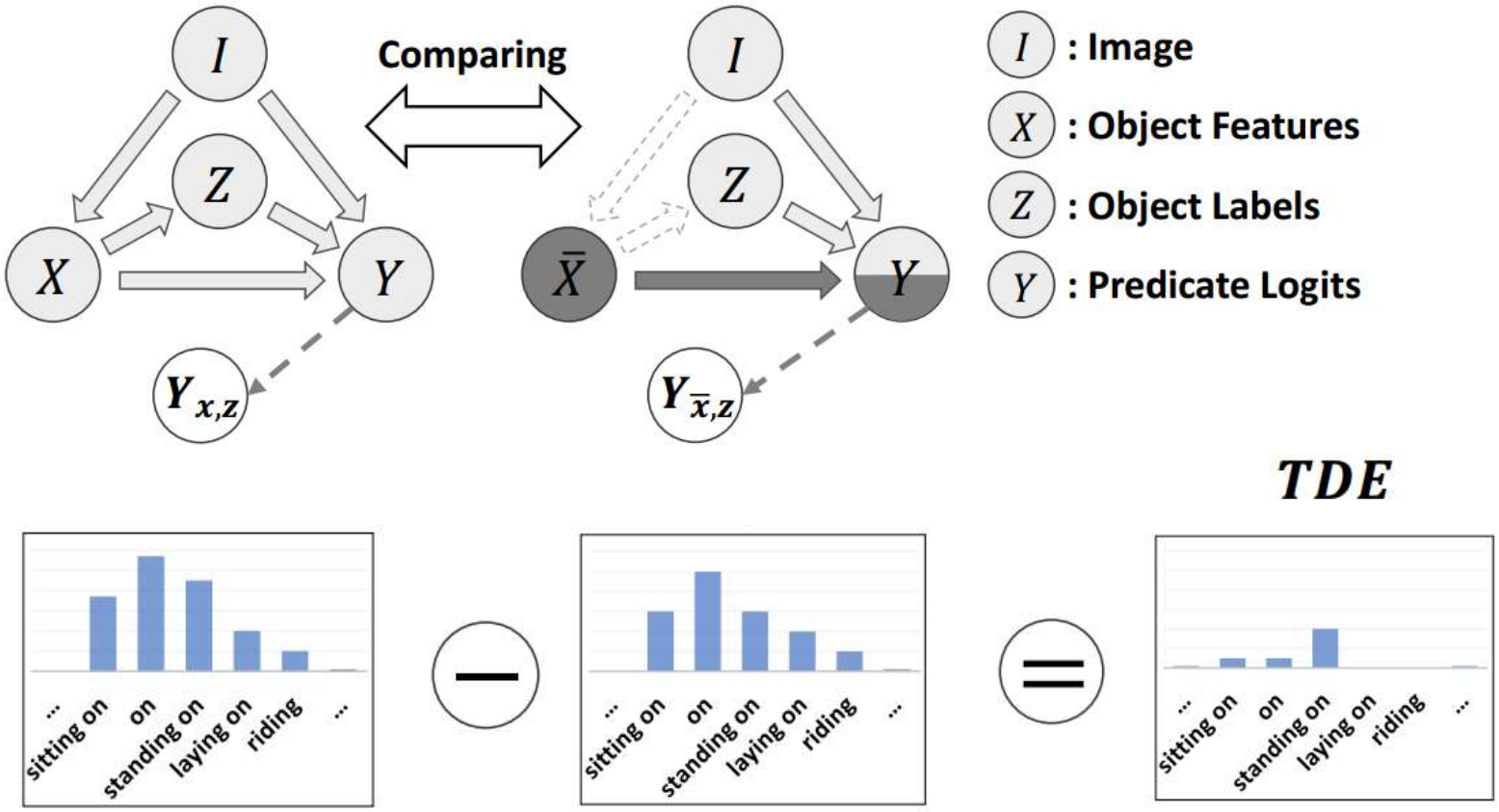


(a) Biased Generation Based on Likelihood



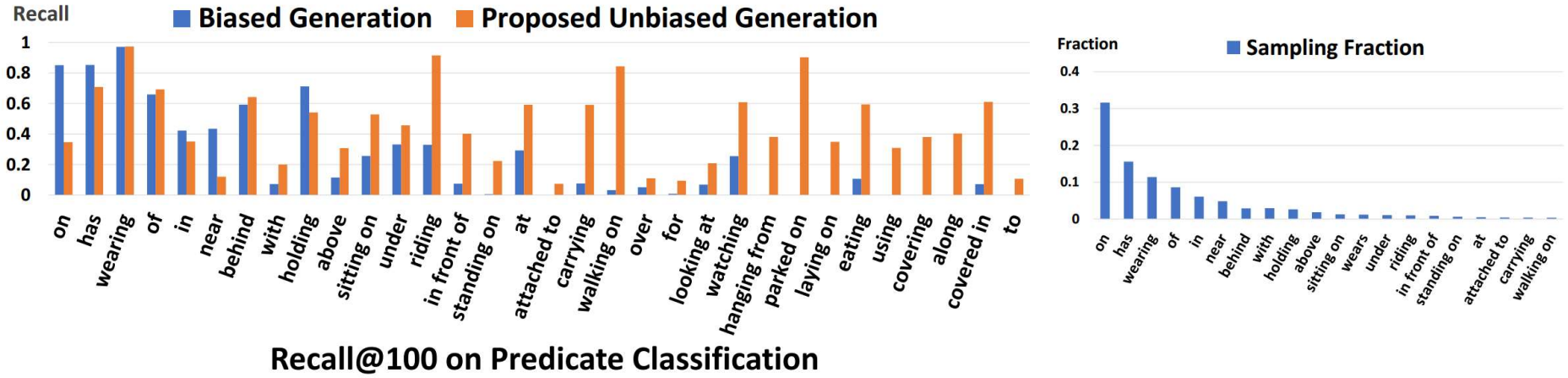
(b) An Intuitive Example of Counterfactual Thinking

Unbiased inference—TDE

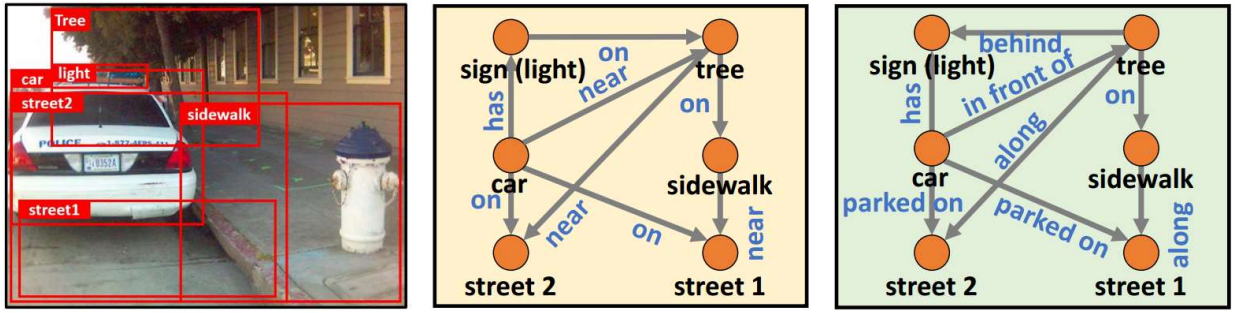


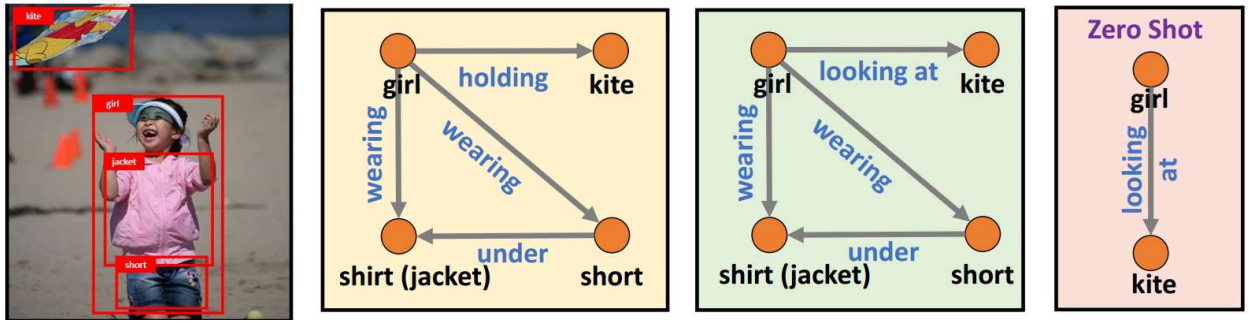
$$TDE = Y_{xz_x} - Y_{\bar{x}z_x}$$

Results

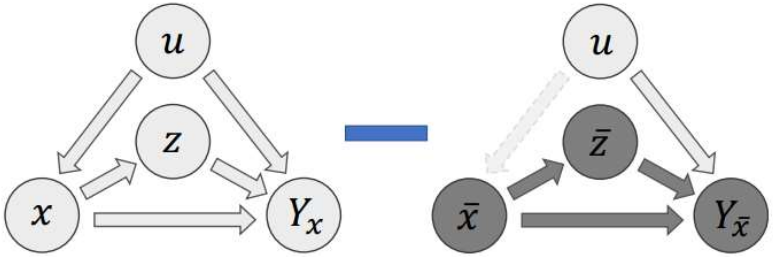


Recall@100 on Predicate Classification





Complementary



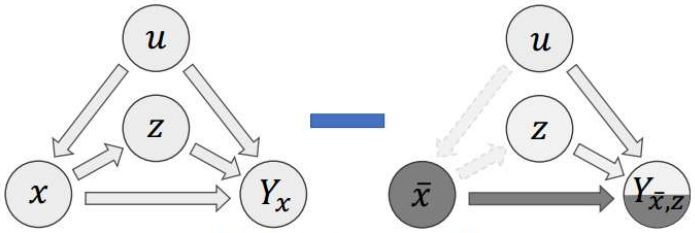
Total Effect

$$TE = Y_x - Y_{x^*} = E[Y|do(X = x)] - E[Y|do(X = x^*)]$$

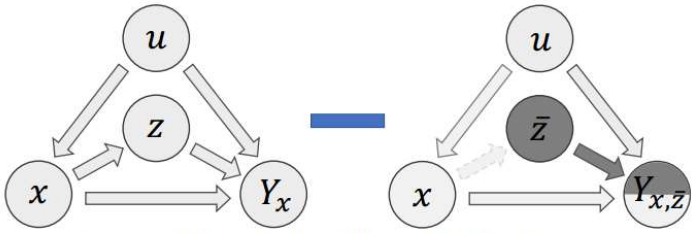
$$ITE = Y(1) - Y(0)$$

$$ATE = E[ITE] = E[Y(1) - Y(0)]$$

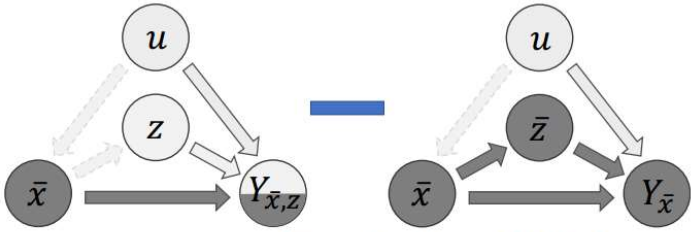
Average Treatment Effect



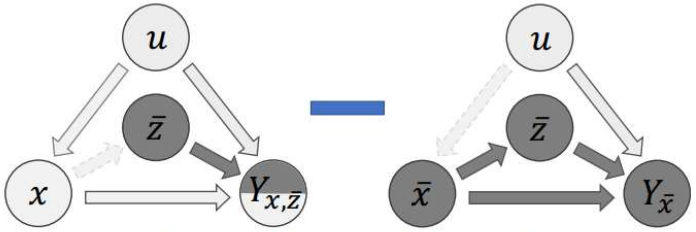
Total Direct Effect



Natural/ Total Indirect Effect



Pure Indirect Effect



Natural/Pure Direct Effect

Pure + mediated interaction effect = Total

Unbiased Scene Graph Generation via Two-stage Causal Modeling

Shuzhou Sun, Shuaifeng Zhi, Qing Liao, Janne Heikkilä, Li Liu

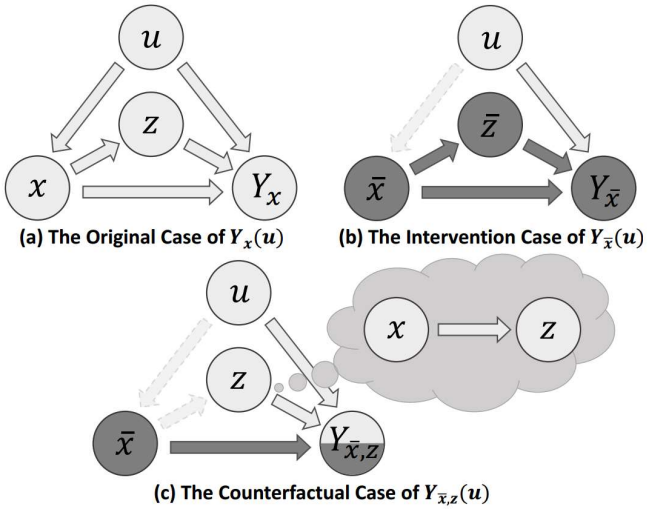
Complementary

the induced submodel $\mathcal{M}_{\tilde{v}}$ in our work is $\langle \mathcal{V}, \mathcal{U}, \mathcal{F}_{\tilde{v}}, P(\mathcal{U}) \rangle$ Where $\mathcal{V} = \{X, Y, S, L\}$, X is input Y is output (relationships), S is the semantic confusion confounder, L is the long-tailed distribution confounder; $\mathcal{U} = \{U_X, U_Y, U_S, U_L\}$; $\mathcal{F}_{\tilde{v}} = \{F_1, F_2, F_3, F_4, F_5\}$; $P(U_X, U_Y, U_S, U_L)$ is the distribution over the exogenous variables.

$$
\begin{aligned}
S &= P(S), \quad (U_S) \\
L &= P(L), \quad (U_L) \\
X &= F_1(L, P(L)) + F_2(S, P(S)), \quad (U_X) \\
Y &= F_3(X, P(X)) + F_4(L, P(L)) + F_5(S, P(S)),
\end{aligned}
$$

Definition 3 (Interventions in SCM). *An intervention* $do(V_i := v')$ *in an SCM* $\mathcal{M}$ *is modeled by replacing the* i*-th structural equation by* $V_i := v'$*, where* v' *is a* V_i*-independent value.*

Definition 4 (Counterfactual in SCM). *A counterfactual in an SCM* $\mathcal{M}$ *is modeled by replacing the* i*-th structural equation by* $V_i := v'$ *and update the* $P(\mathcal{U})$*, where* v' *shares the same meaning as it does in Definition 3. The above counterfactual intervention induces the submodel* $\mathcal{M}^{V_i}$.



Complementary

Independent Causal Mechanisms (ICM) Principle
The causal generative process of a system's variables is composed of autonomous **modules that do not inform or influence each other**. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) **does not inform or influence the other mechanisms**.

Assumption 1 (Causal-insufficient). *The exogenous variable $\mathcal{U}$ in $\mathcal{M}$ satisfies that:* $P(U_1, \ldots, U_n) \neq P(U_1) \times P(U_2) \times \cdots \times P(U_n)$.

Sparse Mechanism Shift (SMS)[2]
Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization, i.e., they should usually not affect all factors simultaneously.

Binary case (X→Y) :

$$P(X, Y) = P(X)P(Y|X)$$

Causal representation:

$$P(V_1, V_2, \ldots, V_n) = \prod_{i=1}^{n} P(V_i \mid \mathrm{pa}(V_i))$$

disentangled

Entangled factorization:

$$P(V_1, V_2, \ldots, V_n) = \prod_{i=1}^{n} P(V_i \mid V_{i+1}, \ldots, V_n)$$

[2]Schölkopf B, Locatello F, Bauer S, et al. Toward causal representation learning[J]. Proceedings of the IEEE, 2021, 109(5): 612-634.

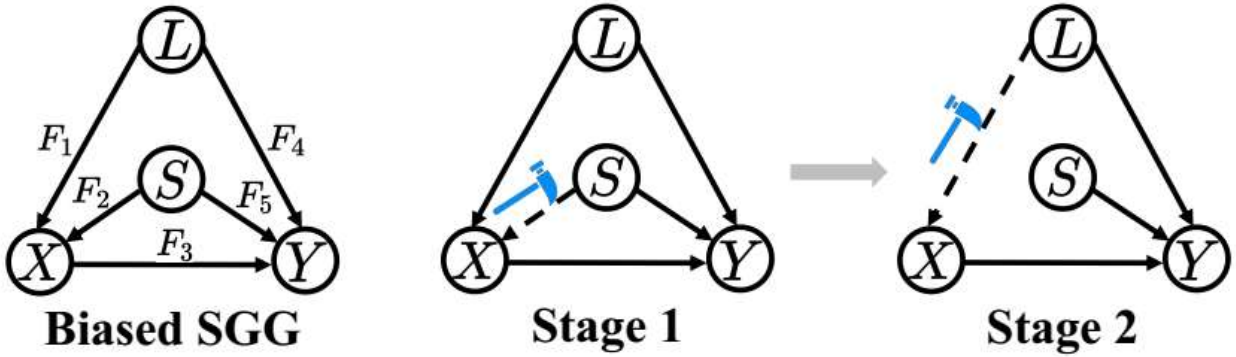
Motivation

Solid points[3]**:**

- The bias from **the long tail** as well as **the semantic bias** and the unknown situation
- Using **sparsity and SMS** for two-stage causal modeling
- Designing **P-loss** to eliminate **semantic bias**, designing **AL-Adjustment** to eliminate **long-tail bias** and calibrating causal representation

[3] Sun S, Zhi S, Liao Q, et al. Unbiased scene graph generation via two-stage causal modeling (TPAMI), 2023.

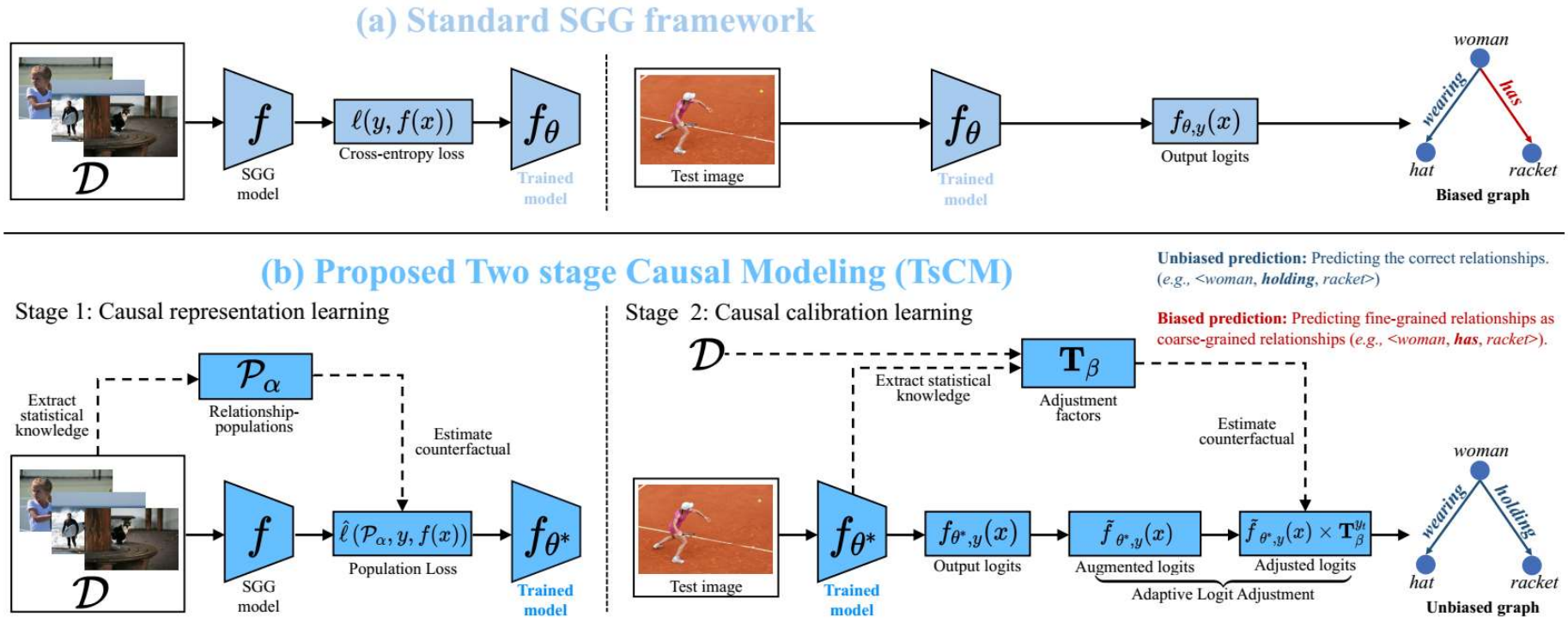
Counterfactual (soft) intervention thinking



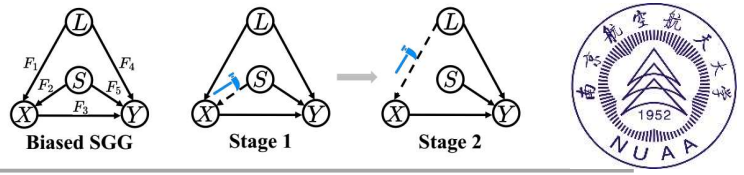
$$\mathbb{E}[Y \mid X, do\,(S := s)\,, do\,(L := l)]$$
$$= \mathbb{E}_X[\underbrace{\mathbb{E}[Y' \mid X, do(S := s)]}_{\text{stage 1}} + \underbrace{\mathbb{E}[Y \mid X, Y', do(L := l)]}_{\text{stage 2}}].$$

where *s*/*l* is a S/L-independent value.

Flow Chart


(a) Standard SGG framework
$\mathcal{D}$
f
SGG model
$\ell(y, f(x))$
Cross-entropy loss
f_θ
Trained model
Test image
f_θ
Trained model
$f_{\theta,y}(x)$
Output logits
woman
wearing
has
hat
racket
Biased graph
(b) Proposed Two stage Causal Modeling (TsCM)
Unbiased prediction: Predicting the correct relationships. (e.g., <woman, holding, racket>)
Biased prediction: Predicting fine-grained relationships as coarse-grained relationships (e.g., <woman, has, racket>).
Stage 1: Causal representation learning
Extract statistical knowledge
$\mathcal{P}_\alpha$
Relationship-populations
Estimate counterfactual
$\mathcal{D}$
f
SGG model
$\hat{\ell}(\mathcal{P}_\alpha, y, f(x))$
Population Loss
f_{θ^*}
Trained model
Stage 2: Causal calibration learning
$\mathcal{D}$
Extract statistical knowledge
$\mathbf{T}_\beta$
Adjustment factors
Estimate counterfactual
Test image
f_{θ^*}
Trained model
$f_{\theta^*,y}(x)$
Output logits
$\tilde{f}_{\theta^*,y}(x)$
Augmented logits
$\tilde{f}_{\theta^*,y}(x) \times \mathbf{T}_\beta^{y_i}$
Adjusted logits
Adaptive Logit Adjustment
woman
wearing
holding
hat
racket
Unbiased graph

Stage 1: Causal representation learning



Would the above error still occur if ***standing on*** and ***walking on*** are no longer similar?

How to exploit the sparsity properties?
How to eliminate semantic bias?

$$P(y|x, do(S := s))$$

ATE

$$= P(y|x, do(S := s_1)) - P(y|x, do(S := s_0))$$

$$= \mathbb{E}_X[\mathbb{E}(Y|X, do(S := s_1)) - \mathbb{E}(Y|X, do(S := s_0))]$$

optimal

$$= P(y|x, \mathcal{P}_\alpha, \pi) = f_{\theta^*}(x)$$

- Relationship similarity set $\mathcal{P}_\alpha = \{\mathbf{P}_\alpha^{y_i}\}_{i=1}^K$

  $\boldsymbol{P}_\alpha^{y_i}$ is a relationship **set** containing the α **most similar** relationships to y_i

- New loss function： P-loss

  more capable of **distinguishing between similar** relationships

P-loss：

$$\hat{\ell}(\mathcal{P}_\alpha, y, f(x)) = \log[1 + \sum_{y' \in \mathbf{P}_\alpha^y} \frac{\pi_{y'}}{\pi_y} \times e^{(f_{y'}(x) - f_y(x))} + \sum_{y' \notin \mathbf{P}_\alpha^y, y' \neq y} e^{(f_{y'}(x) - f_y(x))}],$$

$$\theta^* = \arg\min_\theta \mathbb{E}[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \hat{\ell}(\mathcal{P}_\alpha, y, f(x))],$$

Stage 1: Causal representation learning

- Relationship similarity set $\mathcal{P}_\alpha = \{\mathbf{P}_\alpha^{y_i}\}_{i=1}^K$

two objects, o_i and o_j, whose bounding boxes are $[b_{\bar{x}}^i, b_{\bar{y}}^i, b_h^i, b_w^i]$ and $[b_{\bar{x}}^j, b_{\bar{y}}^j, b_h^j, b_w^j]$, respectively.

the relationship between these two objects as $\psi_{<o_i,o_j>}$:

$$\left[\frac{2(b_{\bar{x}}^i + b_{\bar{x}}^j) - (b_w^i + b_w^j)}{4b_h^i}, \frac{2(b_{\bar{y}}^i + b_{\bar{y}}^j) - (b_h^i + b_h^j)}{4b_h^i}, \frac{2(b_{\bar{x}}^i + b_{\bar{x}}^j) + (b_w^i + b_w^j)}{4b_h^i}, \frac{2(b_{\bar{y}}^i + b_{\bar{y}}^j) + (b_h^i + b_h^j)}{4b_h^i}, \frac{b_h^j}{b_h^i}, \frac{b_w^j}{b_w^i}\right].$$

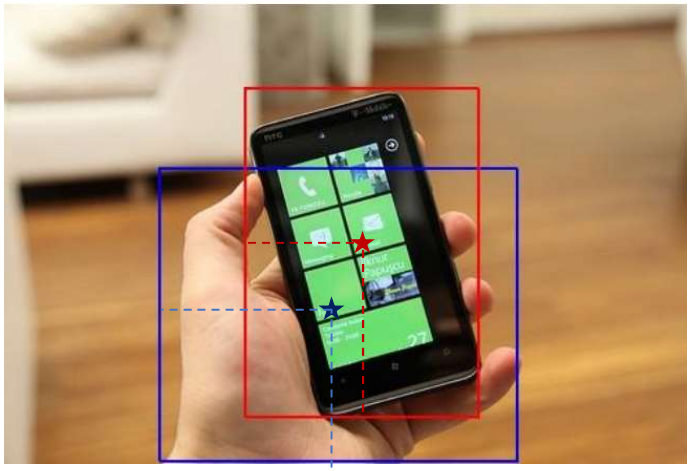
$\boldsymbol{\xi}_y \in \mathbb{R}^4$ $C \times C \times K \times 6$) is a four-dimensional statistic:

$$\boldsymbol{\xi}_y = \begin{bmatrix} \boldsymbol{\xi}_y^{(1,1)} & \boldsymbol{\xi}_y^{(1,2)} & \cdots & \boldsymbol{\xi}_y^{(1,C)} \\ \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{\xi}_y^{(C,1)} & \boldsymbol{\xi}_y^{(C,2)} & \cdots & \boldsymbol{\xi}_y^{(C,C)} \end{bmatrix}$$

where $\boldsymbol{\xi}_y^{(i,j)} = \{\xi_{y_t}^{(i,j)}\}_{t=1}^K$ is the normalized features of relationship $< o_i, y_t, o_j >$, and it can be calculated as:

$$\xi_{y_t}^{(i,j)} = \xi_{y_t}^{<o_i,o_j>} / |\xi_{y_t}^{<o_i,o_j>}|,$$

$\xi_{y_t}^{<o_i,o_j>}$ and $|\xi_{y_t}^{<o_i,o_j>}|$ are the fusion features and numbers of all relationships $< o_i, y_t, o_j >$ in the observed data $\mathcal{D}$, respectively.



$< o_i, y_t, o_j >$

We then calculate the feature of each relationship, for instance, for the t-th relationship ξ_{y_t}:

$$\xi_{y_t} = \sum_{i=1}^{C}\sum_{j=1}^{C} \xi_{y_t}^{(i,j)} / C^{2\S}.$$

For relationship-populations $\mathcal{P}_\alpha$, $\mathcal{P}_\alpha = \{\mathbf{P}_\alpha^{y_i}\}_{i=1}^K$, the population of y_t can be calculated as:

$$\mathbf{P}_\alpha^{y_t} = \underset{t,t' \in \{1,2,\cdots,K\}, t \neq t'}{\arg\operatorname{smal}\alpha} \left\| \xi_{y_t} - \xi_{y_{t'}} \right\|,$$

Stage 2: Causal calibration learning

If one collected the balanced data, or, in particular, *looking at* and near share the same distribution in the observed data D, will the above error still occur?

How to eliminate long-tail bias

$$P(y|x, do(L := l))$$

ATE

$$= P(y|x, do(L := l_1)) - P(y|x, do(L := l_0))$$

$$= \mathbb{E}_X[\mathbb{E}(Y|X, do(L := l_1)) - \mathbb{E}(Y|X, do(L := l_0))]$$

optimal

$$= P(y|x, \mathbf{T}_\beta, \tilde{f}_{\theta^*}) = \tilde{f}_{\theta^*,y}(x) \times \mathbf{T}_\beta \rightarrow \tilde{f}_{\theta^*,y}(x) = e^{f_{\theta^*,y}(x)} \times f^{\text{bg}}_{\theta^*,y}(x)$$

Optimal:

$$y_x = \underset{y \in \{y_1, \cdots, y_K\}}{\arg\max} \{(\tilde{f}_{\theta^*,y}(x) \times \mathbf{T}_\beta)_{y \in \mathbf{T}_\beta} \cap (\tilde{f}_{\theta^*,y}(x))_{y \notin \mathbf{T}_\beta}\}$$

Stage 2: Causal calibration learning

- Adjustment factors $\mathbf{T}_\beta$

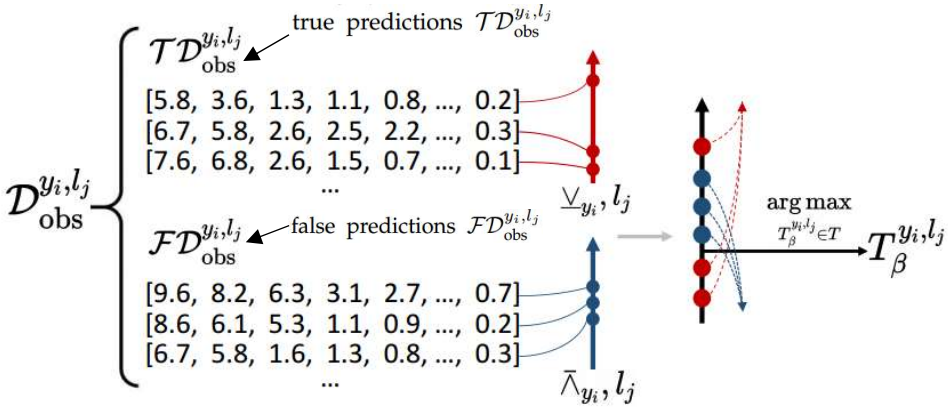
Our proposed adaptive adjustment factors $\mathbf{T}_\beta$ is a two-dimensional $(K \times \beta)$ matrix:

$$\mathbf{T}_\beta = \begin{bmatrix} T_\beta^{y_1,l_1} & T_\beta^{y_1,l_2} & \cdots & T_\beta^{y_1,l_\beta} \\ \cdots & \cdots & \cdots & \cdots \\ T_\beta^{y_K,l_1} & T_\beta^{y_K,l_2} & \cdots & T_\beta^{y_K,l_\beta} \end{bmatrix},$$

where $T_\beta^{y_i,l_j}$ adjusts the j-th largest prediction logit to correspond to the i-th relationship, and it can be calculated as:

$$T_\beta^{y_i,l_j} = \underset{T_\beta^{y_i,l_j} \in T}{\arg\max} \Big(\underbrace{\underset{(x,y)\sim\mathcal{D}^{y_i,l_j}}{\mathbb{TP}} (\tilde{f}_{\theta^*,y}(x) \times T_\beta^{y_i,l_j})}_{\text{true prediction with adjustment}} - \underbrace{\underset{(x,y)\sim\mathcal{D}^{y_i,l_j}}{\mathbb{TP}} (\tilde{f}_{\theta^*,y}(x))}_{\text{true prediction without adjustment}} \Big),$$

where $T \in \mathbb{R}$ and $\underset{(X,Y)}{\mathbb{TP}}(f)$ is a computation kernel that calculates the true prediction numbers (*e.g.*, recall rate (R@K) in SGG task) of model f on dataset (X, Y). $\mathcal{D}^{y_i,l_j}$ is all relationships with the j-th largest prediction logit to correspond to the i-th category reasoned by f_{θ^*}.



$$T_\beta^{y_i,l_j} = \underset{t\in T}{\arg\max} \Big(\sum_{m=1}^{|\underline{\vee}_{y_i,l_j}|} \mathbb{1}(t \geq \underline{\vee}_{y_i,l_j}^m) + \sum_{n=1}^{|\overline{\wedge}_{y_i,l_j}|} \mathbb{1}(t < \overline{\wedge}_{y_i,l_j}^n) \Big),$$

$$T_\beta^{y_i,l_j} = \underset{t\in T}{\arg\max} \Big(\sum_{m=1}^{min(|\underline{\vee}_{y_i,l_j}|,|\overline{\wedge}_{y_i,l_j}|)} \mathbb{1}(t \geq \underline{\vee}'^m_{y_i,l_j}) + \sum_{n=1}^{min(|\underline{\vee}_{y_i,l_j}|,|\overline{\wedge}_{y_i,l_j}|)} \mathbb{1}(t < \overline{\wedge}'^n_{y_i,l_j}) \Big).$$

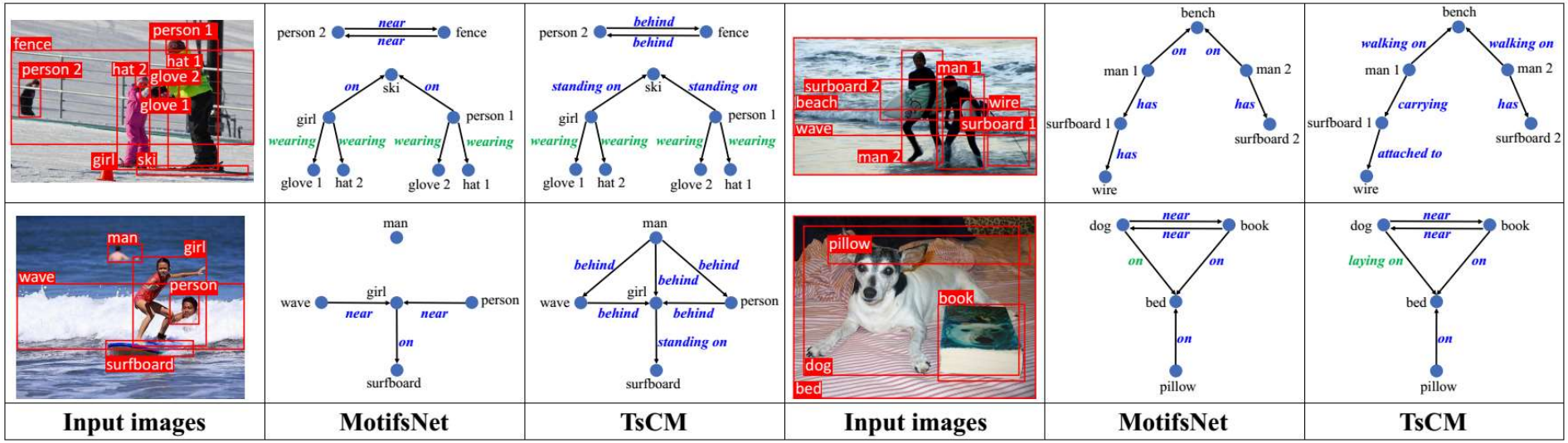
Results

- One of backbone for experiment

	Predicate classification				Scene Graph Classification				Scene Graph Detection			
	PredCls				SGCls				SGDet			
	mR@20	mR@50	mR@100	AVG_{mR}	mR@20	mR@50	mR@100	AVG_{mR}	mR@20	mR@50	mR@100	AVG_{mR}
MotifsNet (backbone) [9]	12.2	15.5	16.8	14.8	7.2	9.0	9.5	8.6	5.2	7.2	8.5	7.0
TDE [14]$^{\diamond\dagger}$ *(CVPR'20)*	18.5	25.5	29.1	24.4	9.8	13.1	14.9	12.6	5.8	8.2	9.8	7.9
SegG [19]$^{\triangle\dagger}$ *(ICCV'21)*	14.5	18.5	20.2	17.7	8.9	11.2	12.1	10.7	6.4	8.3	9.2	8.0
BPL+SA [27]$^{\blacklozenge\diamond\dagger}$ *(ICCV'21)*	24.8	29.7	31.7	28.7	14.0	16.5	17.5	16.0	10.7	13.5	15.6	13.3
CogTree [21]$^{\blacklozenge\dagger}$ *(IJCAI'21)*	20.9	26.4	29.0	25.4	12.1	14.9	16.1	14.4	7.9	10.4	11.8	10.0
DLFE [24]$^{\diamond\dagger}$ *(MM'21)*	22.1	26.9	28.8	25.9	12.8	15.2	15.9	14.6	8.6	11.7	13.8	11.4
EBM-loss [15]$^{\blacklozenge\dagger}$ *(CVPR'21)*	14.2	18.0	19.5	17.2	8.2	10.2	11.0	9.8	5.7	7.7	9.3	7.6
Loss-reweight [44]$^{\blacklozenge\dagger}$ *(ICLR'21)*	26.5	32.9	35.3	31.6	13.8	17.4	19.3	16.8	9.2	12.8	16.5	12.8
Logit-reweight [44]$^{\diamond\dagger}$ *(ICLR'21)*	12.2	15.4	16.7	14.8	6.4	7.6	8.3	7.4	4.5	5.9	7.7	6.0
HML [28]$^{\triangle\blacklozenge\dagger}$ *(ECCV'22)*	30.1	36.3	38.7	35.0	17.1	20.8	22.1	20.0	10.8	14.6	17.3	14.2
FGPL [30]$^{\blacklozenge\dagger}$ *(CVPR'22)*	24.3	33.0	37.5	31.6	17.1	21.3	22.5	20.3	11.1	15.4	18.2	14.9
TransRwt [20]$^{\triangle\dagger}$ *(ECCV'22)*	–	35.8	39.1	–	–	21.5	22.8	–	–	15.8	18.0	–
GCL [16]$^{\blacklozenge\ddagger}$ *(CVPR'22)*	30.5	36.1	38.2	34.9	18.0	20.8	21.8	20.2	12.9	16.8	19.3	16.3
PPDL [17]$^{\blacklozenge\dagger}$ *(CVPR'22)*	–	32.2	33.3	–	–	17.5	18.2	–	–	11.4	13.5	–
RTPB [29]$^{\blacklozenge\diamond\ddagger}$ *(AAAI'22)*	28.8	35.3	37.7	33.9	16.3	19.4	22.6	19.4	9.7	13.1	15.5	12.8
NICE [31]$^{\triangle\blacklozenge\dagger}$ *(CVPR'22)*	–	30.0	32.1	–	–	16.4	17.5	–	–	10.4	12.7	–
PKO [25]$^{\diamond\dagger}$ *(arXiv'22)*	25.0	31.4	34.0	30.1	14.1	17.6	19.1	16.9	9.6	13.4	16.1	13.0
LS-KD(Iter) [32]$^{\blacklozenge\dagger}$ *(arXiv'22)*	–	24.1	27.4	–	–	13.8	15.2	–	–	9.7	11.5	–
CAME [33]$^{\triangle\blacklozenge\ddagger}$ *(arXiv'22)*	18.1	26.2	32.0	25.4	10.5	15.1	18.0	14.5	6.7	9.3	12.1	9.4
TsCM$^{\blacklozenge\diamond\dagger}$	31.8	37.8	40.9	36.8	18.7	22.4	23.8	21.6	13.7	17.4	19.7	16.9

Results

- One of backbone for experiment



Results

- Ablation

Results obtained by different logit augmentation strategies

	Logit augmentation	mR@20	mR@50	mR@100
MotifsNet	No augmentation	28.3	34.1	37.4
	$e^{f_{\theta^*,y}(x)} \times 1$	30.9	36.4	39.7
	$e^{f_{\theta^*,y}(x)} \times 2$	30.7	36.5	40.3
	$e^{f_{\theta^*,y}(x)} \times f^{\text{bg}}_{\theta^*,y}(x)$	31.8	37.8	40.9
VCTree	No augmentation	29.6	35.1	38.2
	$e^{f_{\theta^*,y}(x)} \times 1$	31.6	38.1	40.8
	$e^{f_{\theta^*,y}(x)} \times 2$	31.4	37.3	40.9
	$e^{f_{\theta^*,y}(x)} \times f^{\text{bg}}_{\theta^*,y}(x)$	32.3	38.7	41.5
Transformer	No augmentation	30.4	36.2	38.7
	$e^{f_{\theta^*,y}(x)} \times 1$	31.9	38.4	41.7
	$e^{f_{\theta^*,y}(x)} \times 2$	32.2	38.0	41.5
	$e^{f_{\theta^*,y}(x)} \times f^{\text{bg}}_{\theta^*,y}(x)$	32.8	40.1	42.3

Model performance trained with different loss functions

	Loss	mR/R@20	mR/R@50	mR/R@100
MotifsNet	ℓ	12.2/59.5	15.5/66.0	16.8/67.9
	$\hat{\ell}^{\triangleright}$	12.3/58.8	15.7/65.1	17.2/66.3
	$\hat{\ell}^{\triangle}$	12.3/58.5	15.8/64.2	17.5/65.9
	$\hat{\ell}^{\triangleleft}$	12.5/57.7	16.1/63.3	17.9/65.4
	$\hat{\ell}$	12.9/53.5	16.9/59.4	20.1/61.8
VCTree	ℓ	12.4/59.8	15.4/66.2	16.6/68.1
	$\hat{\ell}^{\triangleright}$	12.3/59.3	15.6/65.4	17.4/67.2
	$\hat{\ell}^{\triangle}$	12.4/58.8	15.6/64.3	17.8/66.4
	$\hat{\ell}^{\triangleleft}$	12.5/58.2	15.9/63.8	18.2/65.9
	$\hat{\ell}$	12.7/54.3	16.4/59.9	19.8/62.3
Transformer	ℓ	12.4/58.5	16.0/65.0	17.5/66.7
	$\hat{\ell}^{\triangleright}$	12.5/58.2	16.2/64.2	17.6/65.1
	$\hat{\ell}^{\triangle}$	12.7/57.9	16.3/63.8	17.6/64.9
	$\hat{\ell}^{\triangleleft}$	12.8/57.4	16.6/63.6	18.1/63.7
	$\hat{\ell}$	13.1/50.8	17.2/58.1	20.3/61.2

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

THANKS