Enhancing Intrinsic Features for Debiasing via Investigating Class-Discerning Common Attributes in Bias-Contrastive Pair

Jeonghoon Park*[1], Chaeyeon Chung*[1], Juyoung Lee[2], Jaegul Choo[1]
[1]Korea Advanced Institute of Science and Technology, South Korea, [2]Kakao Brain, South Korea.
[1]{jeonghoon_park, cy_chung, jchoo}@kaist.ac.kr, [2]michael.ljy@kakaobrain.com

Code: None

—— CVPR 2024

2024.5.16 • ChongQing

Reported by Renhui Luo

1.Introduction

2.Overview

3.Methods

4.Experiments

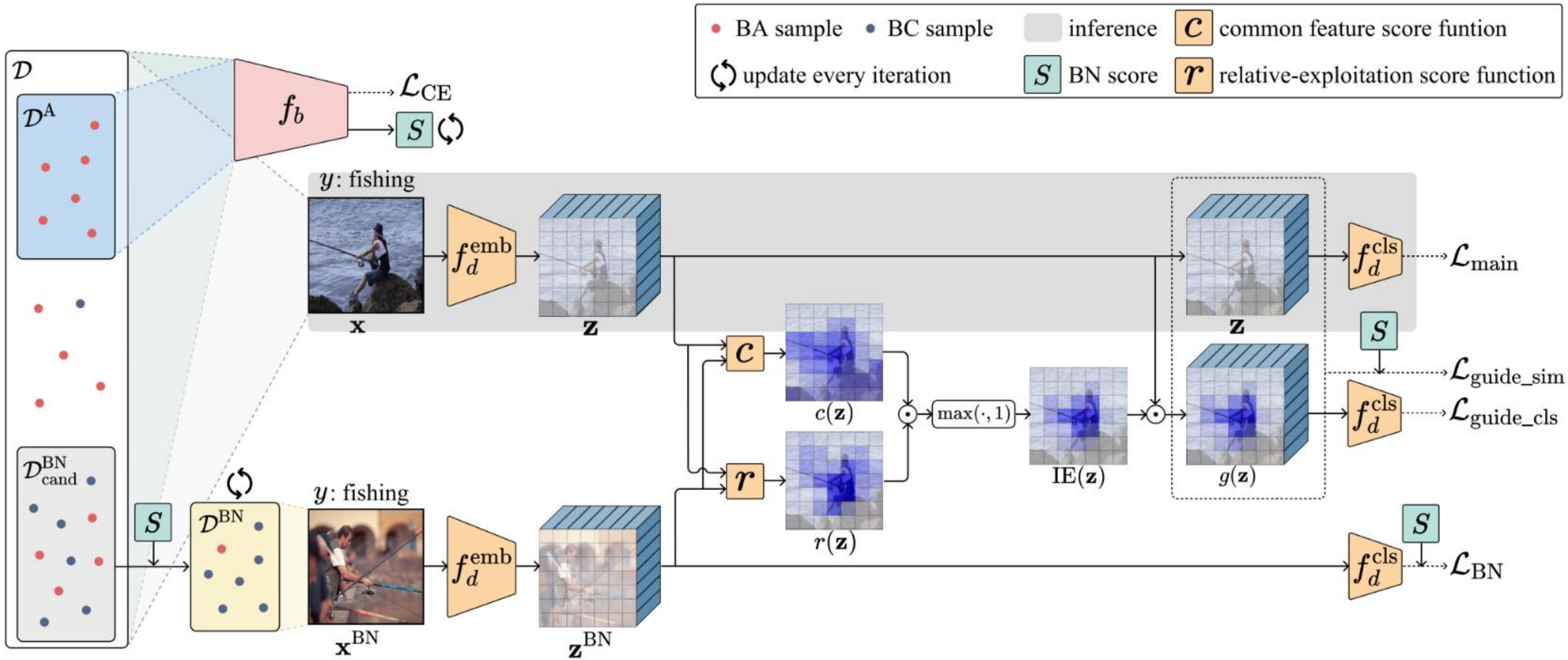
Introduction

Bias-Conflict

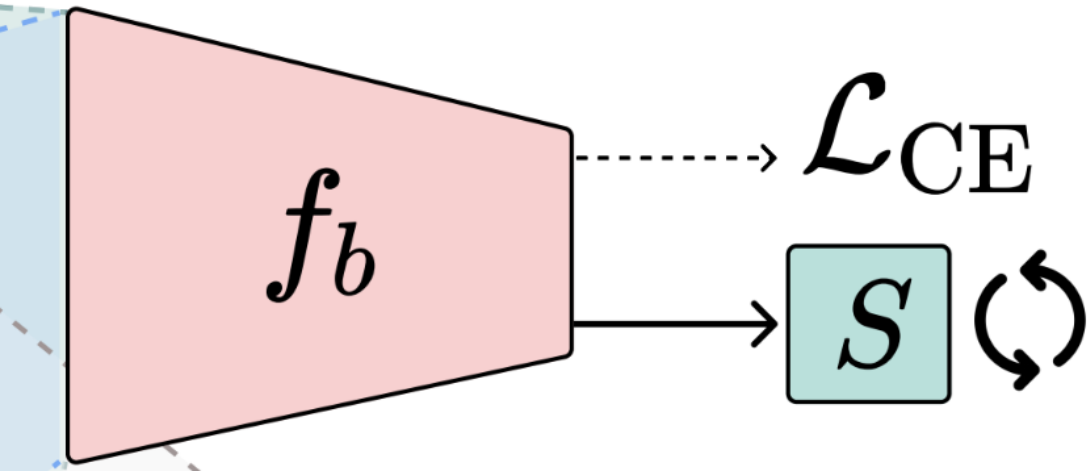
Bias-Align

Model

Overview


BA sample
BC sample
inference
common feature score funtion
update every iteration
BN score
relative-exploitation score function
$\mathcal{D}$
$\mathcal{D}^{\mathrm{A}}$
$\mathcal{D}^{\mathrm{BN}}_{\mathrm{cand}}$
f_b
$\mathcal{L}_{\mathrm{CE}}$
S
y: fishing
$\mathbf{x}$
f_d^{emb}
$\mathbf{z}$
$\boldsymbol{c}$
$\boldsymbol{r}$
$c(\mathbf{z})$
$r(\mathbf{z})$
$\max(\cdot, 1)$
$\mathrm{IE}(\mathbf{z})$
$g(\mathbf{z})$
f_d^{cls}
$\mathcal{L}_{\mathrm{main}}$
$\mathcal{L}_{\mathrm{guide\_sim}}$
$\mathcal{L}_{\mathrm{guide\_cls}}$
$\mathcal{L}_{\mathrm{BN}}$
$\mathcal{D}^{\mathrm{BN}}$
$\mathbf{x}^{\mathrm{BN}}$
$\mathbf{z}^{\mathrm{BN}}$

Method

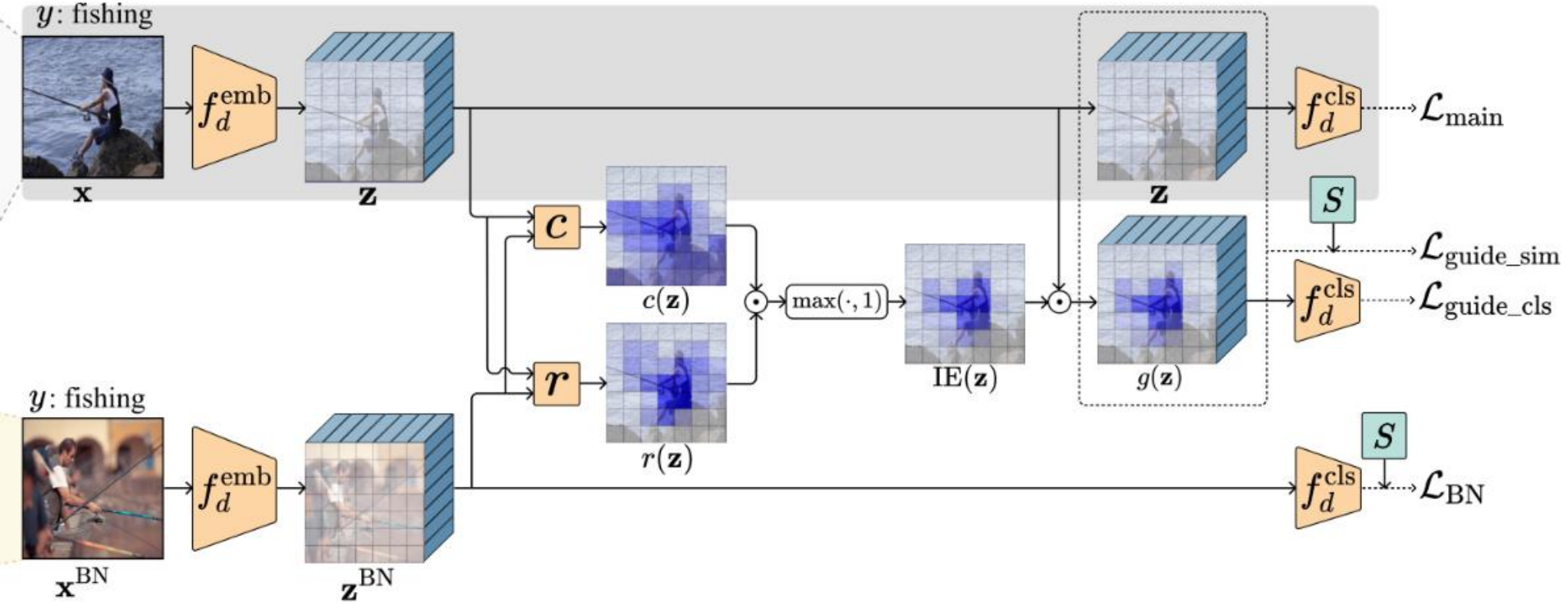


$$l_t(\mathbf{x}) = \alpha_l \cdot \mathcal{L}_{\text{CE}}(f_b(\mathbf{x}), y) + (1 - \alpha_l) \cdot l_{t-1}(\mathbf{x}), \quad (1)$$

$$s_t(\mathbf{x}) = \alpha_s \cdot (l_t(\mathbf{x}) - l_{\text{ref}}(\mathbf{x})) + (1 - \alpha_s) \cdot s_{t-1}(\mathbf{x}), \quad (2)$$

$$\mathcal{D}_t^{\text{BN}} = \{\mathbf{x} \mid s_t(\mathbf{x}) > 0, \mathbf{x} \sim \mathcal{D}_{\text{cand}}^{\text{BN}}\}, \quad (3)$$

Method



$$f_d(\mathbf{x}) = f_d^{\text{cls}}\left(f_d^{\text{emb}}(\mathbf{x})\right).$$

$$\mathbf{z} = f_d^{\text{emb}}(\mathbf{x})$$

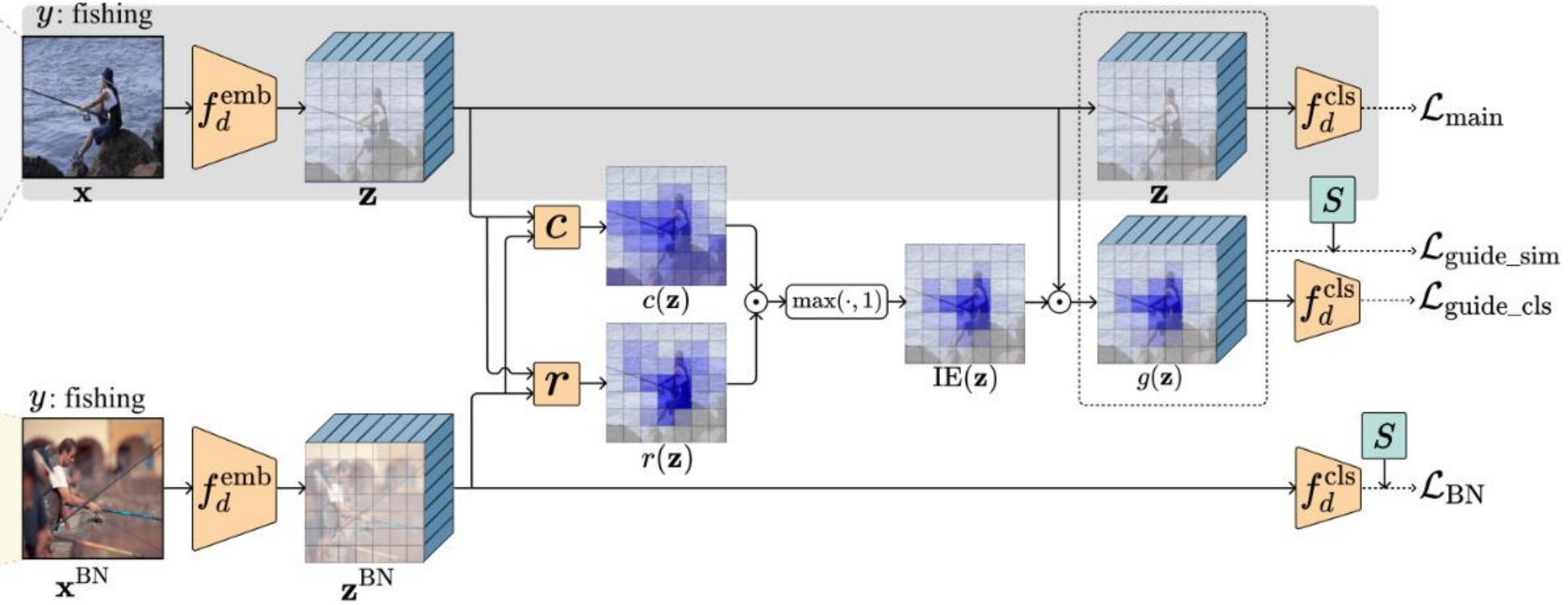
$$\mathbf{z}^{\text{BN}} = f_d^{\text{emb}}(\mathbf{x}^{\text{BN}})$$

$$c(\mathbf{z})_n = \frac{\mathbf{z}_{i^*}^{\text{BN}} \cdot \mathbf{z}_n}{\max_{i,j}(\mathbf{z}_i^{\text{BN}} \cdot \mathbf{z}_j)}, \quad (4)$$

$$i^* = \arg\max_i \left(\mathbf{z}_i^{\text{BN}} \cdot \mathbf{z}_n\right)$$

$$r(\mathbf{z})_n = \left(\frac{2\text{E}(\mathbf{z}^{\text{BN}})_{i^*}}{\text{E}(\mathbf{z}^{\text{BN}})_{i^*} + \text{E}(\mathbf{z})_n}\right)^{\tau}, \quad (5)$$

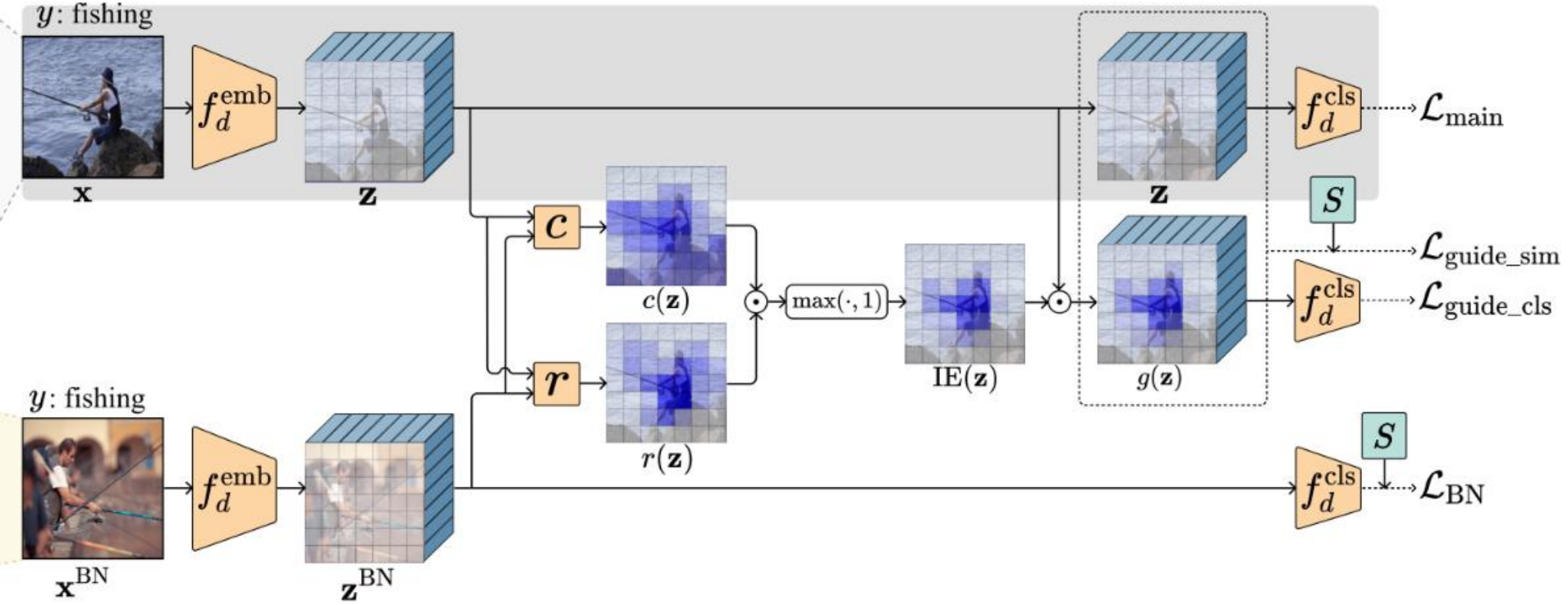
Method



$$\mathrm{IE}(\mathbf{z})_n = \max(c(\mathbf{z})_n \odot r(\mathbf{z})_n, 1), \qquad (6)$$

$$g(\mathbf{z}) = \mathbf{z} \odot \mathrm{IE}(\mathbf{z}). \qquad (7)$$

Method



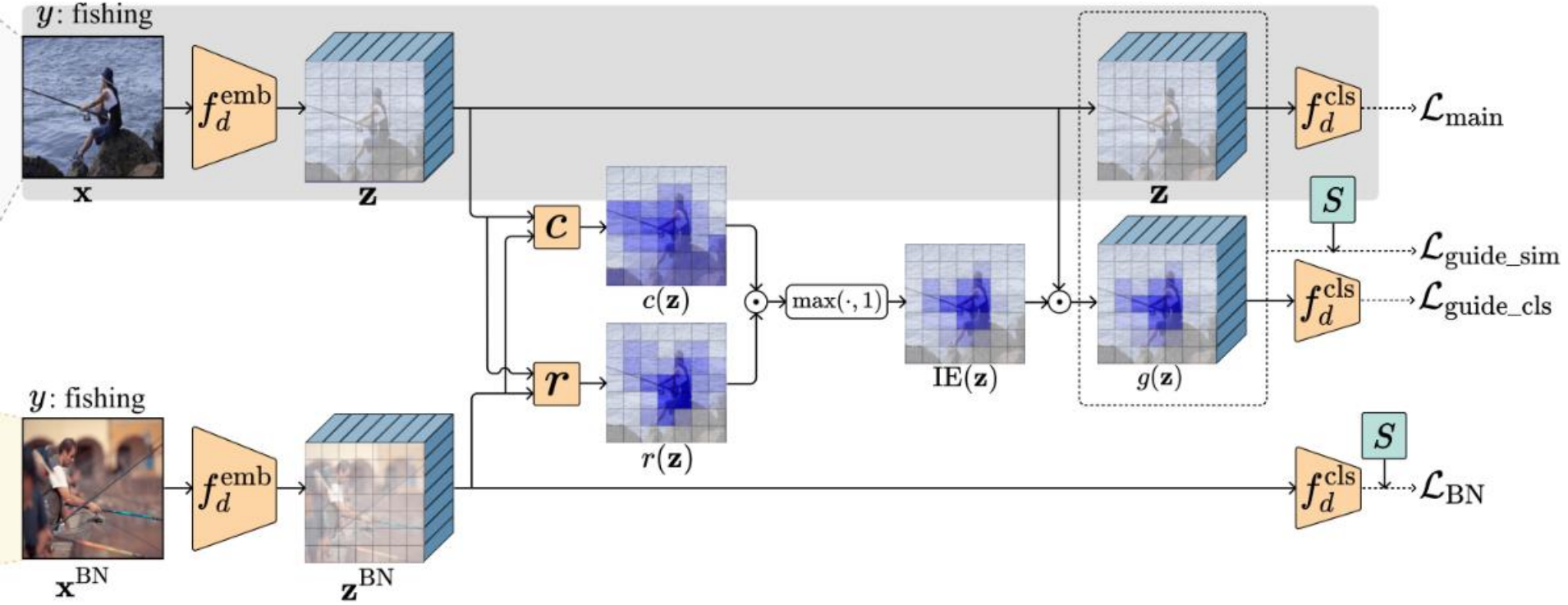
$$\mathcal{L}_{main} = w(\mathbf{x})\mathcal{L}_{CE}(f_d(\mathbf{x}), y) \quad (8)$$

$$\mathcal{L}_{guide\_sim} = s(\mathbf{x}^{BN})\|\mathrm{GAP}(\mathbf{z}) - \mathrm{GAP}(g(\mathbf{z}))\|_1, \quad (9)$$

$$w(\mathbf{x}) = \frac{\mathcal{L}_{CE}(f_b(\mathbf{x}), y)}{\mathcal{L}_{CE}(f_b(\mathbf{x}), y) + \mathcal{L}_{CE}(f_d(\mathbf{x}), y)}. \quad (14)$$

$$\mathcal{L}_{guide\_cls} = w(\mathbf{x})\mathcal{L}_{CE}\left(f_d^{cls}(g(\mathbf{z})), y\right), \quad (10)$$

Method



$$\mathcal{L}_{\text{main}} = w(\mathbf{x})\mathcal{L}_{\text{CE}}(f_d(\mathbf{x}), y) \tag{8}$$

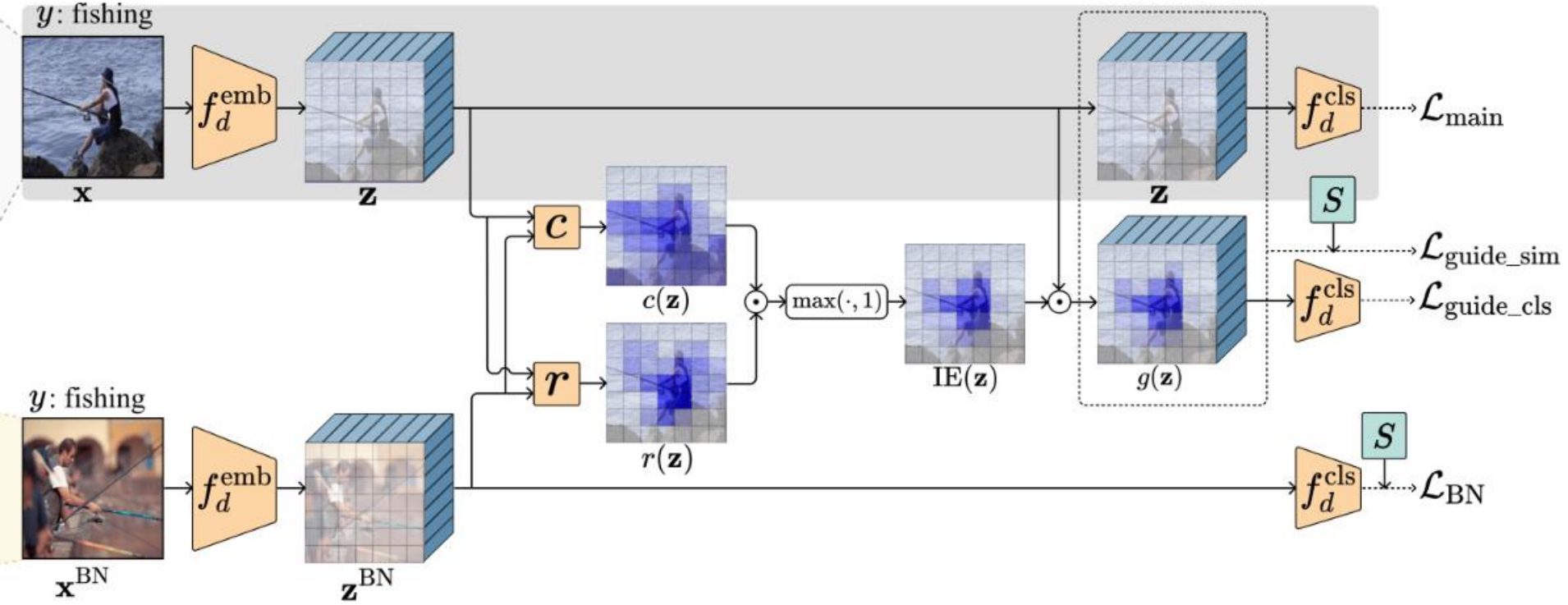
$$w(\mathbf{x}) = \frac{\mathcal{L}_{\text{CE}}(f_b(\mathbf{x}), y)}{\mathcal{L}_{\text{CE}}(f_b(\mathbf{x}), y) + \mathcal{L}_{\text{CE}}(f_d(\mathbf{x}), y)}. \tag{14}$$

$$\mathcal{L}_{\text{guide\_sim}} = s(\mathbf{x}^{\text{BN}})\|\text{GAP}(\mathbf{z}) - \text{GAP}\left(g(\mathbf{z})\right)\|_1, \tag{9}$$

$$\mathcal{L}_{\text{guide\_cls}} = w(\mathbf{x})\mathcal{L}_{\text{CE}}\left(f_d^{\text{cls}}(g(\mathbf{z})), y\right), \tag{10}$$

$$\mathcal{L}_{\text{guide}} = \lambda_{\text{sim}}\mathcal{L}_{\text{guide\_sim}} + \mathcal{L}_{\text{guide\_cls}}, \tag{11}$$

Method



$$\mathcal{L}_{\text{BN}} = s(\mathbf{x}^{\text{BN}})\mathcal{L}_{\text{CE}}(f_d(\mathbf{x}^{\text{BN}}), y). \qquad (12)$$

$$\mathcal{L}_{\text{total}} = \lambda_{\text{main}}\mathcal{L}_{\text{main}} + \mathcal{L}_{\text{guide}} + \mathcal{L}_{\text{BN}}, \qquad (13)$$

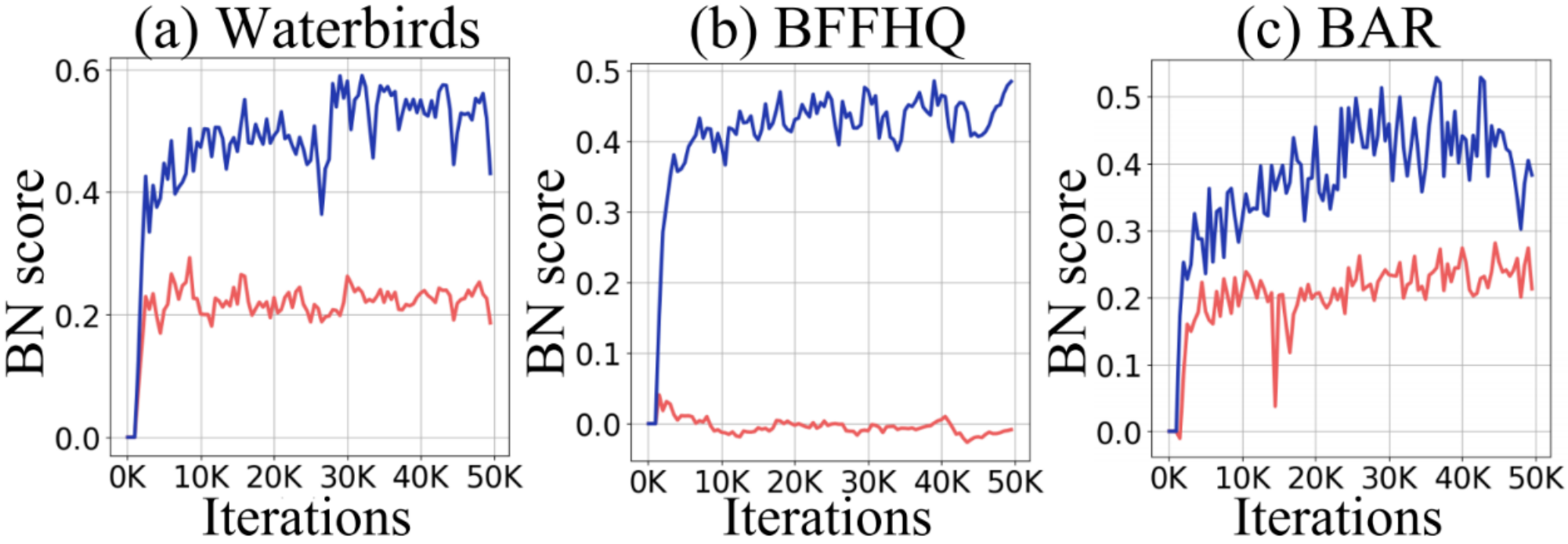
Experiments

Method	Waterbirds				BFFHQ				BAR	
	0.5	1.0	2.0	5.0	0.5	1.0	2.0	5.0	1.0	5.0
Vanilla [5]	57.41	58.07	61.04	64.13	55.64	60.96	69.00	82.88	70.55	82.53
HEX [25]	57.88	58.28	61.02	64.32	56.96	62.32	70.72	83.40	70.48	81.20
LNL [9]	58.49	59.68	62.27	66.07	56.88	62.64	69.80	83.08	-	-
EnD [22]	58.47	57.81	61.26	64.11	55.96	60.88	69.72	82.88	-	-
ReBias [2]	55.44	55.93	58.53	62.14	55.76	60.68	69.60	82.64	73.04	83.90
LfF [15]	60.66	61.78	58.92	61.43	65.19	69.24	73.08	79.80	70.16	82.95
DisEnt [12]	59.59	60.05	59.76	64.01	62.08	66.00	69.92	80.68	70.33	83.13
LfF+BE [13]	61.22	62.58	63.00	63.48	67.36	75.08	80.32	85.48	73.36	83.87
DisEnt+BE [13]	51.65	54.10	53.43	54.21	67.56	73.48	79.48	84.84	73.29	84.96
Ours	**63.64**	**65.22**	**65.23**	**66.33**	**71.68**	**77.56**	**83.08**	**87.60**	**75.14**	**85.03**

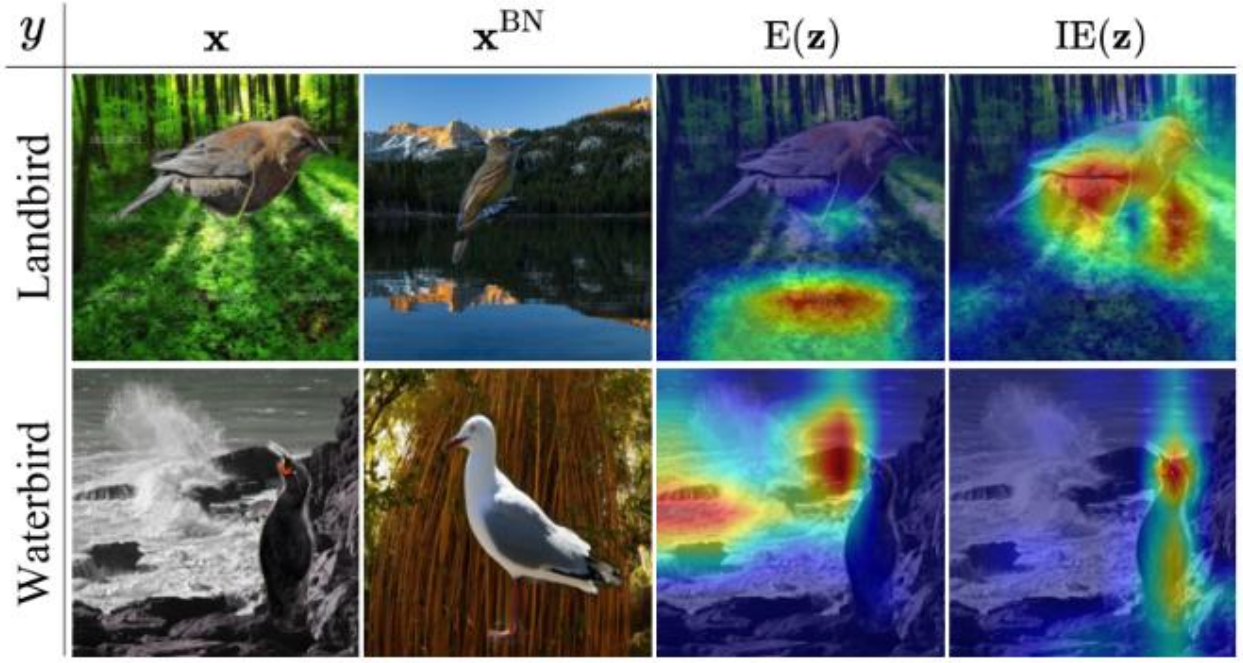
Experiments

Dataset	$\mathcal{D}^{BN}_{cand}$ - $\mathcal{D}^{BN}$		$\mathcal{D}^{BN}/\mathcal{D}$ (%)	
	BA	BC	BA	BC
Waterbirds	26.50 ±5.32	0.75 ±0.83	2.75 ±0.31	79.69 ±3.72
BFFHQ	199.80 ±40.14	8.00 ±2.76	0.46 ±0.09	50.00 ±1.04
BAR	30.60 ±3.83	3.20 ±1.60	3.58 ±0.14	47.14 ±5.71

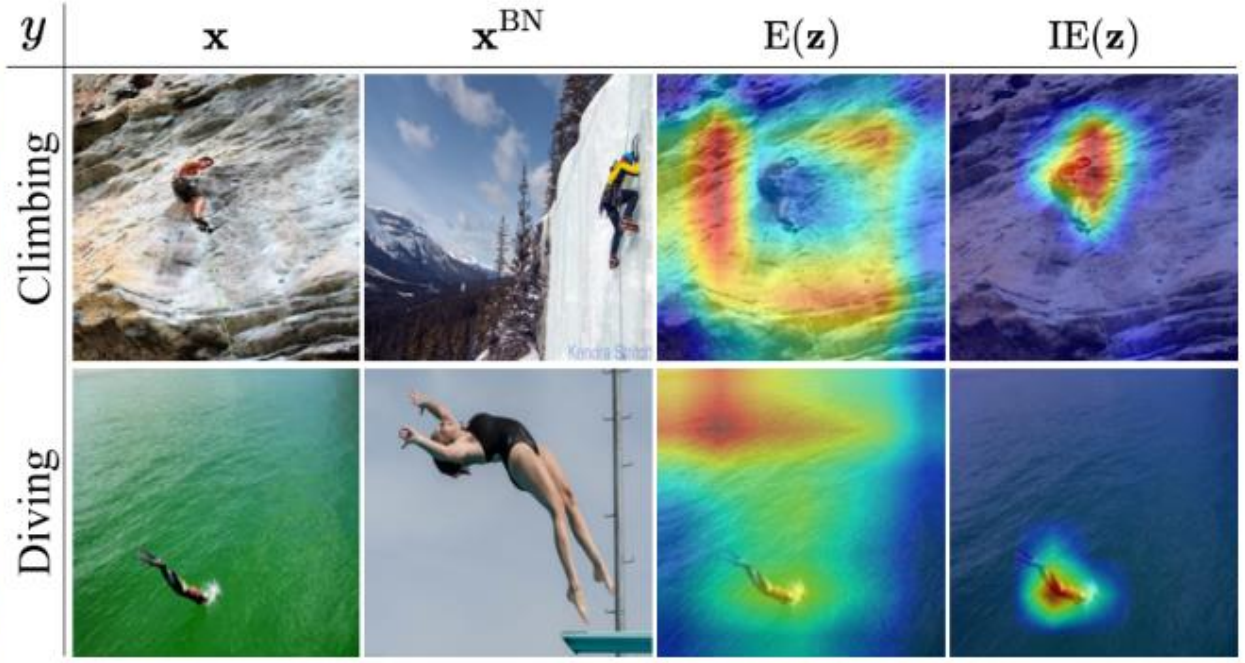
Experiments



Experiments

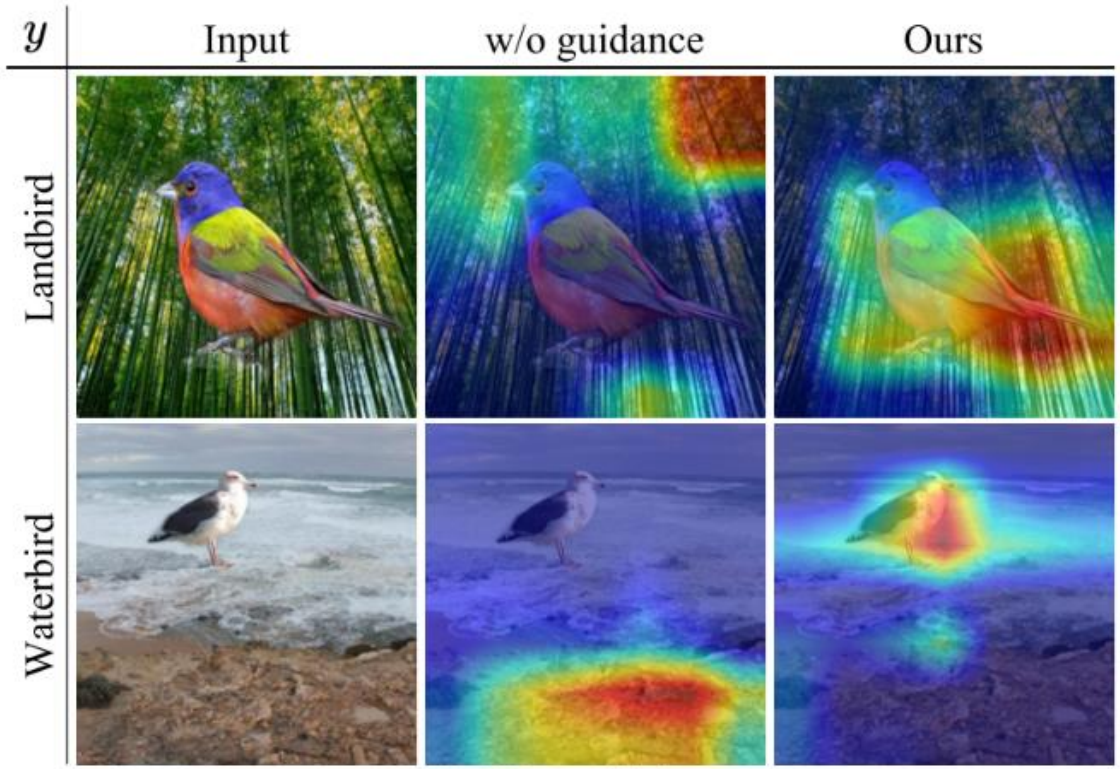


(a) Waterbirds

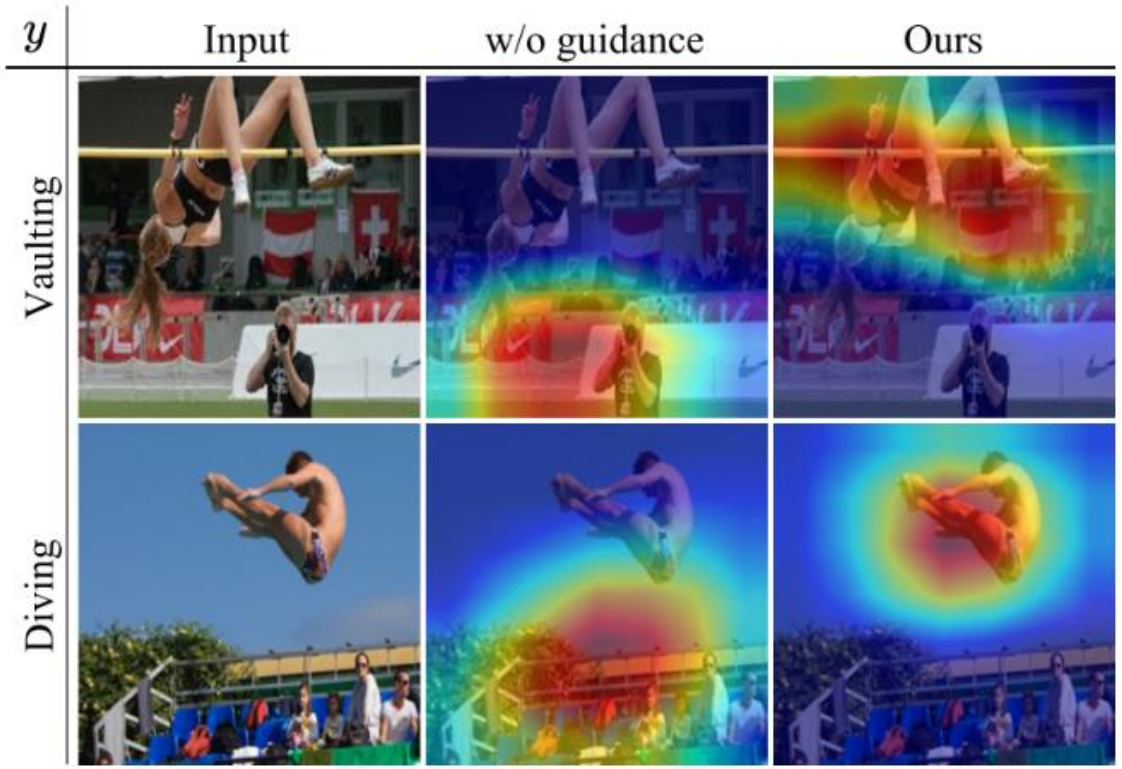


(b) BAR

Experiments



(a) Waterbirds

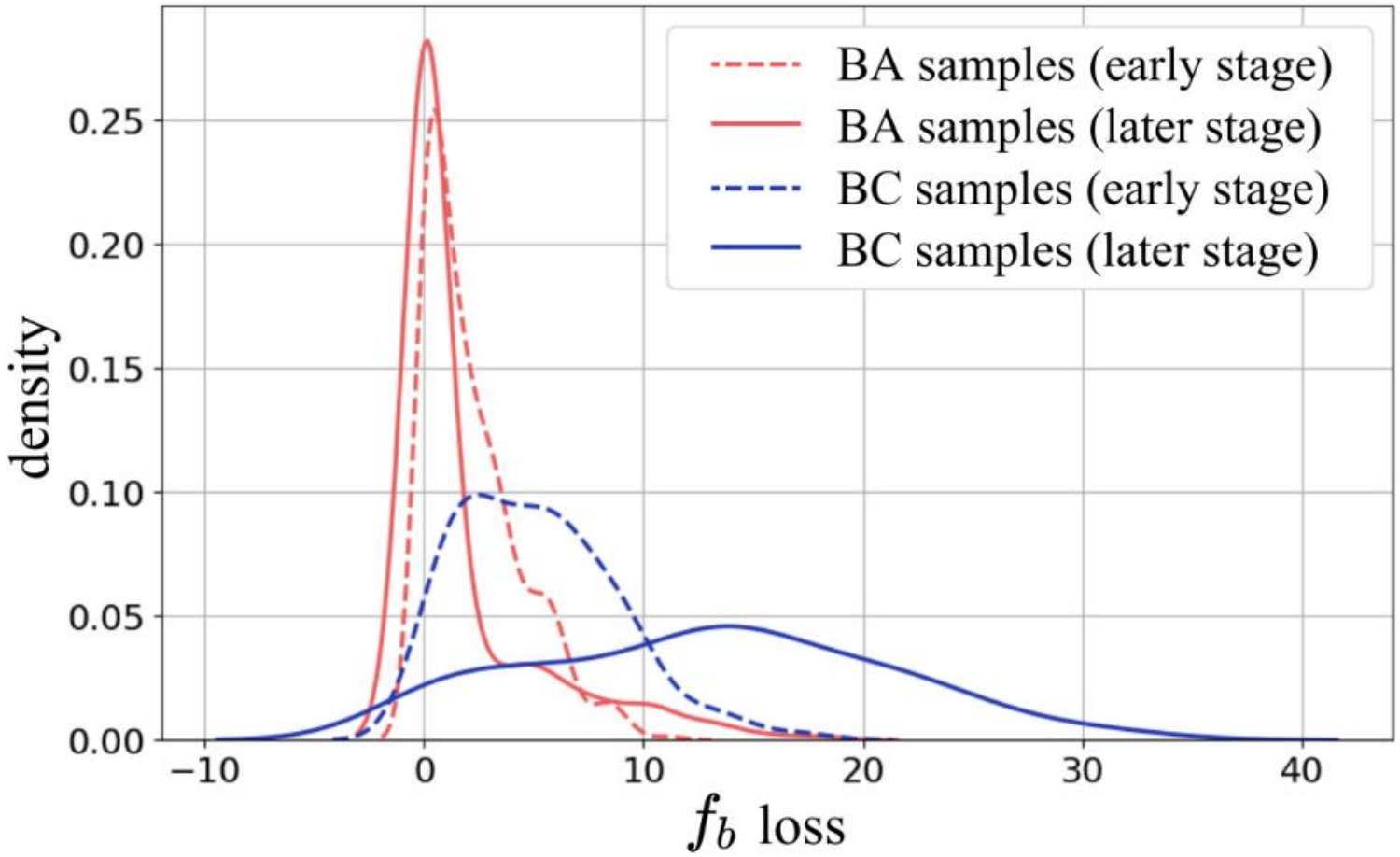


(b) BAR

Experiments

$\mathcal{L}_{\text{guide}}$	$\mathcal{L}_{\text{BN}}$	$\mathbf{x}^{\text{BN}}$	$s(\mathbf{x}^{\text{BN}})$ as loss weight	Waterbirds	BFFHQ	BAR
✓	✓	$\mathcal{D}$	✗	62.79 ±1.21	71.04 ±2.55	73.36 ±1.40
✓	✓	$\mathcal{D}^{\text{BN}}_{\text{cand}}$	✗	64.65 ±1.23	75.64 ±1.87	74.27 ±0.66
✓	✓	$\mathcal{D}^{\text{BN}}$	✗	65.10 ±0.87	77.08 ±2.05	74.62 ±1.07
✗	✓	$\mathcal{D}^{\text{BN}}$	✓	63.81 ±1.24	76.92 ±1.03	74.03 ±1.13
✓	✗	$\mathcal{D}^{\text{BN}}$	✓	62.10 ±3.35	74.84 ±2.00	74.87 ±1.51
✓	✓	$\mathcal{D}^{\text{BN}}$	✓	**65.22** ±0.95	**77.56** ±1.24	**75.14** ±0.82

Experiments



Experiments

Experiments

#BC in $\mathcal{D}^{\mathrm{BN}}$/#BC in $\mathcal{D}$	0.1	0.5	1.0	1.0	1.0	1.0	1.0
#BA in $\mathcal{D}^{\mathrm{BN}}$/#BC in $\mathcal{D}^{\mathrm{BN}}$	0.0	0.0	0.0	0.1	1.0	2.0	10.0
Accuracy	75.84	78.12	81.40	80.24	77.48	75.48	70.90

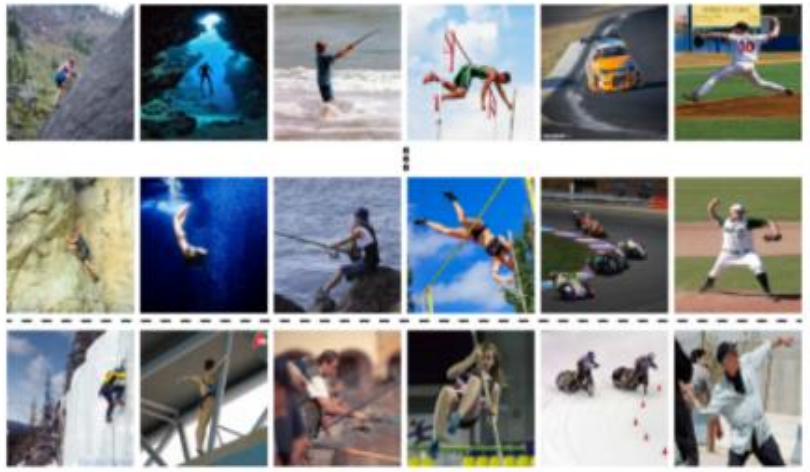
Experiments

BS	Vanilla [5]	HEX [25]	LNL [9]	EnD [22]	ReBias [2]	LfF [15]	DisEnt [12]	LfF+BE [13]	DisEnt+BE [13]	Ours
0.5	24.08 ±1.56	28.20 ±3.07	26.08 ±1.64	28.29 ±3.53	27.00 ±1.10	56.22 ±6.07	38.07 ±11.01	55.15 ±2.78	36.60 ±10.88	59.12 ±3.67
1.0	24.78 ±2.45	26.32 ±2.90	29.72 ±3.45	25.69 ±2.41	27.95 ±1.56	59.07 ±3.40	47.02 ±7.26	55.53 ±1.60	28.35 ±4.17	63.05 ±1.97
2.0	34.39 ±2.24	32.12 ±2.89	33.92 ±1.94	32.94 ±1.48	32.16 ±0.76	53.07 ±6.74	44.93 ±8.54	52.91 ±2.62	31.08 ±6.01	61.71 ±4.94
5.0	38.34 ±1.05	39.08 ±0.92	43.22 ±1.94	40.91 ±1.11	39.72 ±1.11	58.05 ±2.37	52.96 ±6.33	48.48 ±3.72	37.92 ±6.47	58.60 ±3.32

Experiments

(a) Waterbirds

(b) BFFHQ



(c) BAR

Thanks!