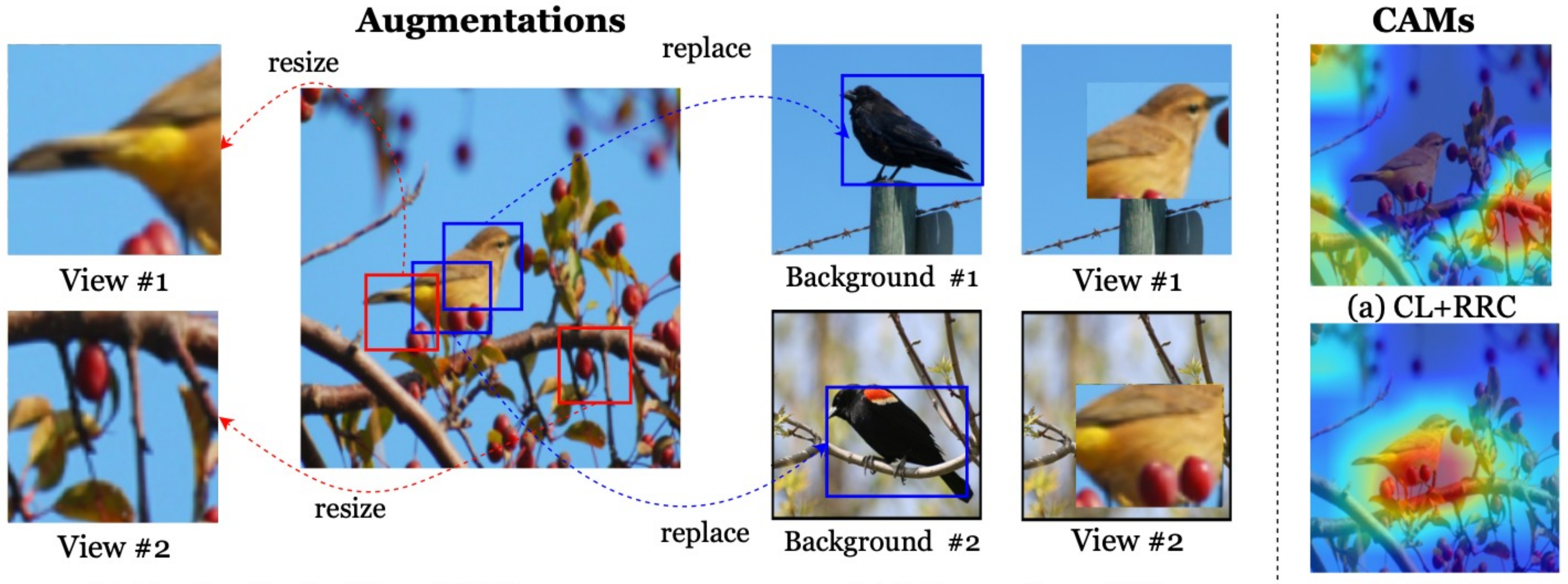
Exploring Localization for Self-supervised Fine-grained Contrastive Learning

Di Wu[1,2], Siyuan Li[1,2], Zelin Zang[1,2], and Stan Z. Li[1]
[1] Westlake University, [2] Zhejiang University

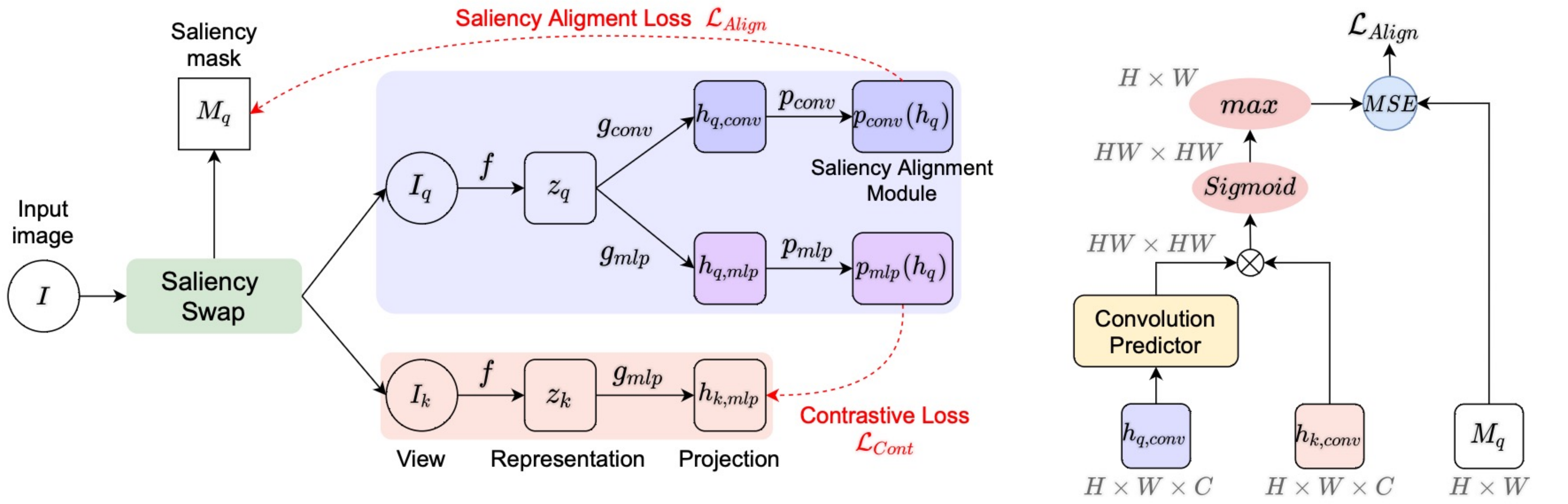
Introduction



(a) RandomResizedCrop (RRC) (b) Saliency Swap (SS) (b) CL+SS

Comparison of RandomResizedCrop (RRC) and the proposed Saliency Swap (SS). (a) shows the commonly-used RRC in contrastive learning (CL) methods. (b) shows our proposed SS, which crops from regions of interest of the reference image and replaces the saliency regions of two randomly selected images to guarantee semantic consistency.

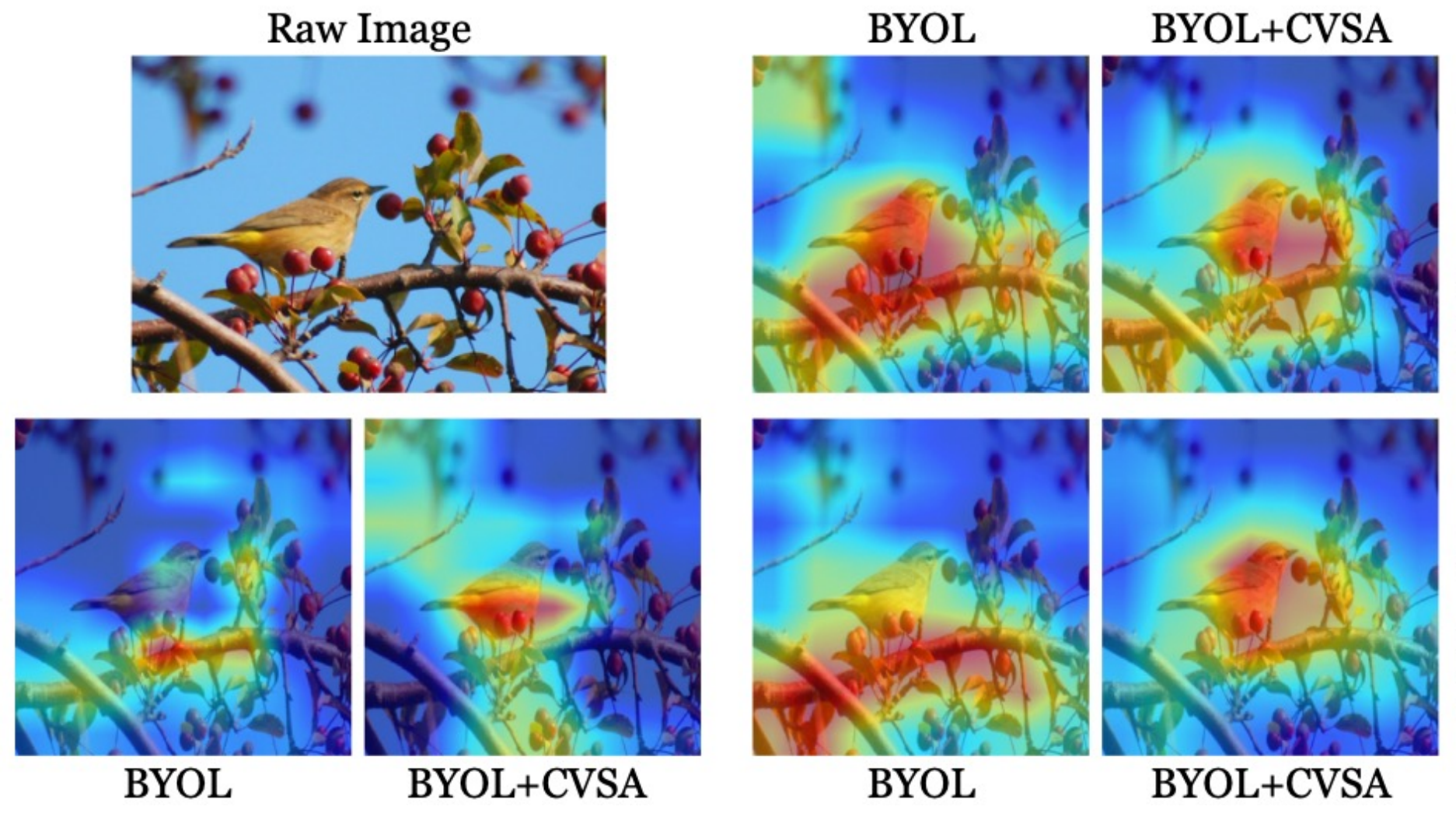
Method

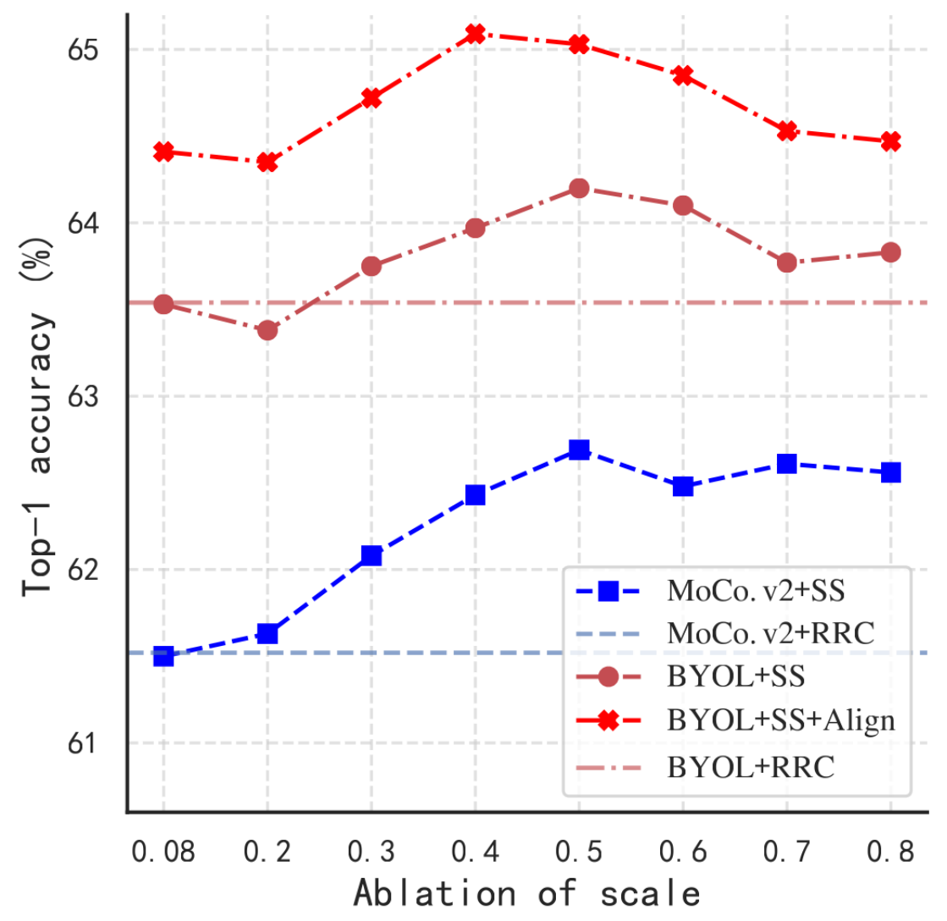


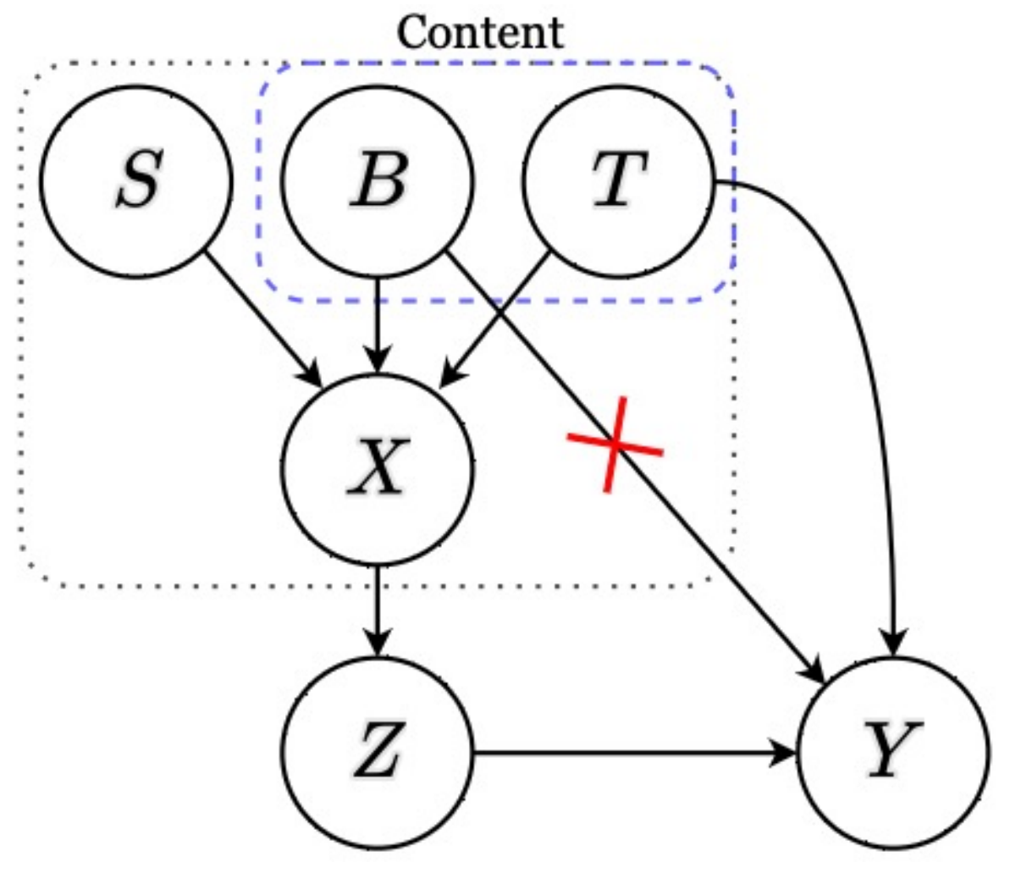
(a) CVSA Architecture (b) Saliency Alignment Module

Learning paradigm of our proposed Cross-view Saliency Alignment (CVSA). (a) The network parameters in red are an exponential moving average of the purple part, and the contrastive loss L_{Cont} is calculated between $p_{q,mlp}$ and $h_{k,mlp}$ (stop-gradient) as BYOL. (b) Saliency alignment module calculates L_{Align} between predicted attention maps and input saliency masks M. The final learning objective is$L_{CVSA} = L_{Cont} + L_{Align}$.

Analysis







We analyze the Grad-CAM of learned representation on CUB-200 (left), the scale hyper-parameter of SS compared to RRC (right 1), and the causal interpretation graph of CVSA which weakens the causality between the background B and the semantic labels Y for better localization abilities.

Conclusion: CVSA learns discriminative fine-grained representation with better localization abilities.

Experiments

Methods	CUB	NAbirds	Aircrafts	Cars
Random	58.51 (-6.34)	65.78 (-6.76)	70.58 (-2.02)	70.51 (-5.36)
Rel-Loc	65.89 (+1.04)	72.60 (+0.06)	72.24 (-0.36)	75.27 (-0.60)
Rot-Pred	66.67 (+1.82)	73.01 (+0.47)	72.67 (+0.07)	75.48 (-0.39)
SimCLR	63.43 (-1.42)	72.05 (-0.49)	72.42 (-0.18)	75.12 (-0.75)
MoCo.v2	63.21 (-1.64)	71.36 (-1.18)	71.93 (-0.67)	74.89 (-0.98)
LooC*	66.42 (+1.57)	72.84 (+0.30)	72.49 (-0.11)	75.69 (-0.18)
InsLoc	64.87 (+0.02)	72.80 (+0.26)	73.43 (+0.83)	76.61 (+0.64)
BYOL	64.85 (+0.00)	72.54 (+0.00)	72.60 (+0.00)	75.87 (+0.00)
BYOL+DiLo	66.16 (+1.31)	73.12 (+0.58)	73.52 (+0.92)	76.36 (+0.49)
BYOL+CVSA	**66.88** (+2.03)	**73.75** (+1.21)	**74.55** (+1.95)	**77.45** (+1.58)

Methods	Stage 2	CUB	NAbirds	Aircrafts	Cars
Rel-Loc	✓	67.33 (-1.22)	73.82 (-2.47)	81.20 (-0.53)	85.80 (-0.56)
Rot-Pred	✓	67.75 (-0.80)	74.26 (-2.03)	81.58 (-0.15)	85.74 (-0.62)
SimCLR	×	68.30 (-0.30)	73.51 (-2.78)	81.18 (-0.55)	85.93 (-0.43)
MoCo.v2	×	68.47 (-0.08)	73.73 (-2.56)	81.78 (+0.05)	86.08 (-0.28)
MoCo.v2	✓	67.60 (-1.05)	73.18 (-3.11)	81.04 (-0.69)	85.71 (-0.65)
LooC*	✓	68.71 (+0.16)	74.45 (-1.84)	81.75 (+0.02)	85.90 (-0.46)
InsLoc	✓	67.94 (-0.20)	76.36 (+0.07)	81.54 (-0.23)	86.38 (+0.02)
BYOL	×	68.55 (+0.00)	76.29 (+0.00)	81.73 (+0.00)	86.36 (+0.00)
BYOL	✓	68.01 (-0.54)	75.82 (-0.47)	80.53 (-1.20)	85.49 (-0.87)
BYOL+DiLo	✓	68.70 (+0.15)	76.94 (+0.65)	82.04 (+0.31)	86.46 (+0.10)
BYOL+CVSA	✓	**69.14** (+0.59)	**77.57** (+1.28)	**82.77** (+1.04)	**87.13** (+0.77)

We evaluate CVSA in three aspects: (1) *Second-stage only* pre-training on fine-grained benchmarks based on ResNet-50, (2) *Dual-stage* pre-training with the first-stage using COCO dataset, and (3) *Dual-stage* pre-training on fine-grained benchmarks with the first-stage using ImageNet-1K dataset. Top-1 accuracy of fine-tuning evaluation is reported. CVSA yields significant improvements for *second-stage only* pre-training, which learns better localization abilities than BYOL baseline, while consistently outperforming existing contrastive learning methods with *dual-stage* pre-training for more practical usage.

Observation: CVSA brings consistent improvements of fine-grained representation learning.

Conclusions

We design a dual-stage pre-training pipeline with the first-stage to train feature extraction and the second-stage to train localization. To empower the model with localization abilities in the second-stage, we propose cross-view saliency alignment (CVSA), a new self-supervised contrastive learning framework. Extensive experiments on fine-grained benchmarks demonstrate the effectiveness of our contributions in learning better fine-grained representations.