Masking Improves Contrastive Self-Supervised Learning for ConvNets, and Saliency Tells You Where

Zhi-Yi Chin*[1], Chieh-Ming Jiang*[1], Ching-Chun Huang[1], Pin-Yu Chen[2], Wei-Chen Chiu[1]

[1]National Yang Ming Chiao Tung University, Taiwan [2]IBM Research

(* denotes equal contribution)

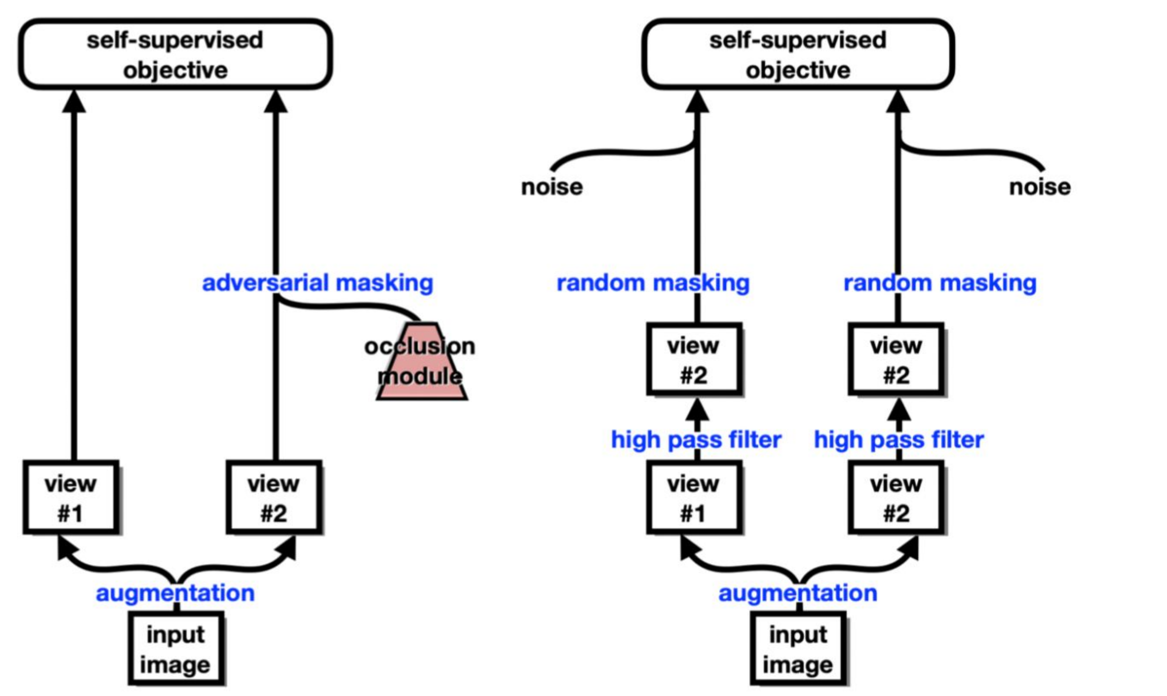
JAN 4-8 WACV 2024 WAIKOLOA HAWAII

Introduction

- **Self-Supervised Learning (SSL)**
  - Training a model (feature extractor) via leveraging the data itself to define a task for providing supervisory signals
- **Motivation**
  - Most existing SSL methods that incorporate masking as augmentation are based on Vision Transformers (ViT) and utilize reconstruction as the training objective.
  - Most of existing contrastive SSL methods for CNNs seldom apply masking as an augmentation technique due to the negative effects (e.g. unwanted edges) which the masked patches could introduce to the convolution layer.
  - Our question: Are we able to include masking as an extra augmentation method into contrastive SSL framework with CNN as its backbone?

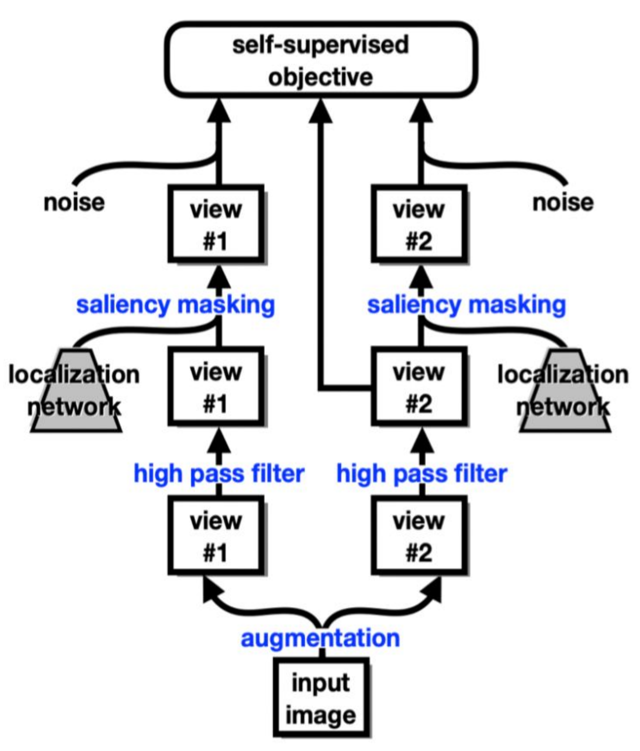
Masking in CNN-based SSL

- **MSCN** (Jing et al., 2022) tackles the issue **"How to mask"** by incorporating high pass filter to alleviate unwanted edges problem.
- **ADIOS** (Shi et al., 2022) tackles the issue **"Where to mask"** instead of random masking, it particularly adopts an occlusion module which learns adversarially along the feature extractor to produce semantically meaningful masks
- Previous studies have delved into the singular aspects of either how or where to apply masking, instead, our focus is to jointly take both where and how to implement masking into consideration.



ADIOS MSCN

We extend the "How to mask" idea in MSCN by providing more solutions to mitigate the unwanted edges problem as well as tackling "Where to mask" by producing semantic meaningful mask with far less computation than ADIOS.



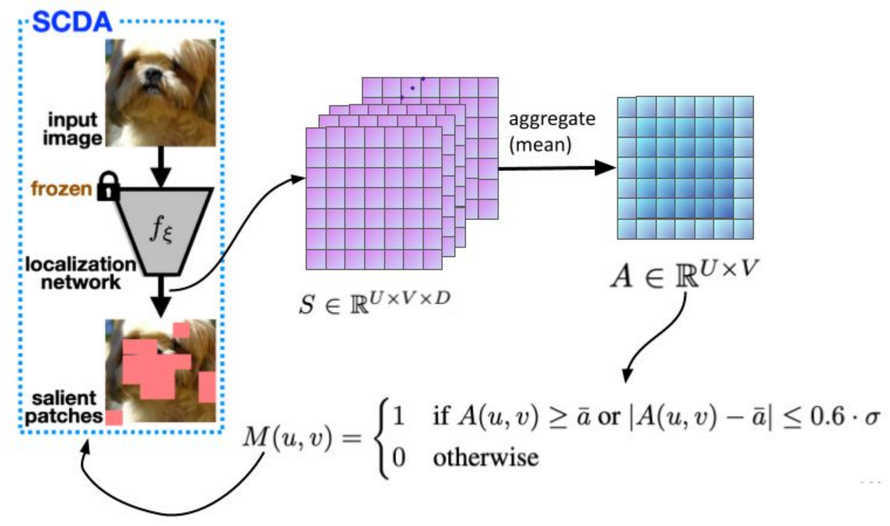
Ours

Network Architecture

Our proposed method utilizes saliency-guided masking for contrastive SSL with CNNs, leveraging saliency information before applying random masking.

- To mitigate parasitic edges and improve performance, three masking strategies are introduced.
- Extra hard negative samples are introduced by masking more salient patches.

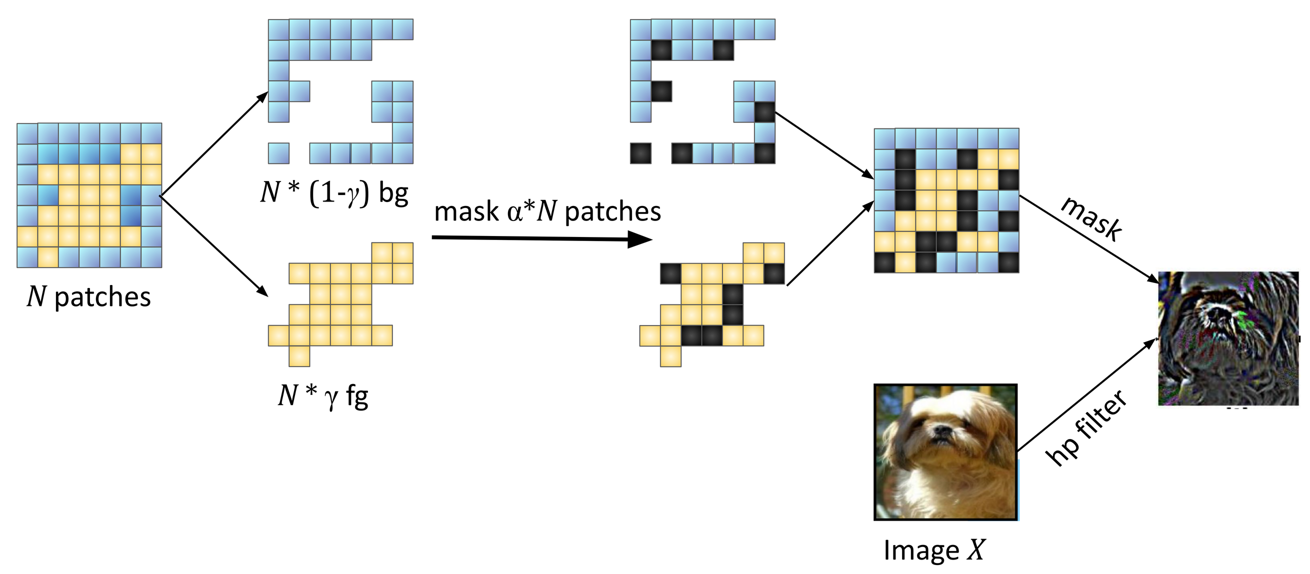
Retrieve Saliency Map with Relaxed Constrained SCDA (Wei et al. 2017)



$S \in \mathbb{R}^{U \times V \times D}$ $A \in \mathbb{R}^{U \times V}$

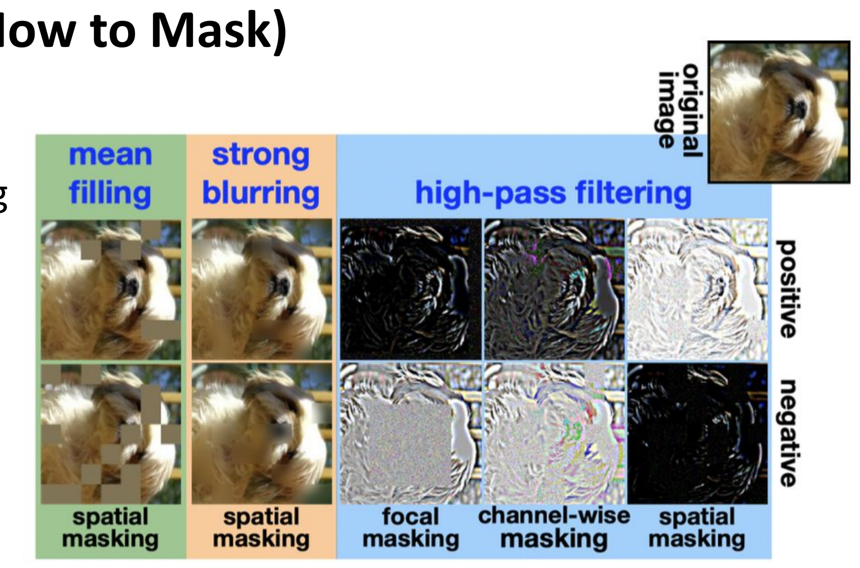
$$M(u,v) = \begin{cases} 1 & \text{if } A(u,v) \geq \bar{a} \text{ or } |A(u,v) - \bar{a}| \leq 0.6 \cdot \sigma \\ 0 & \text{otherwise} \end{cases}$$

Saliency-Guided Masking

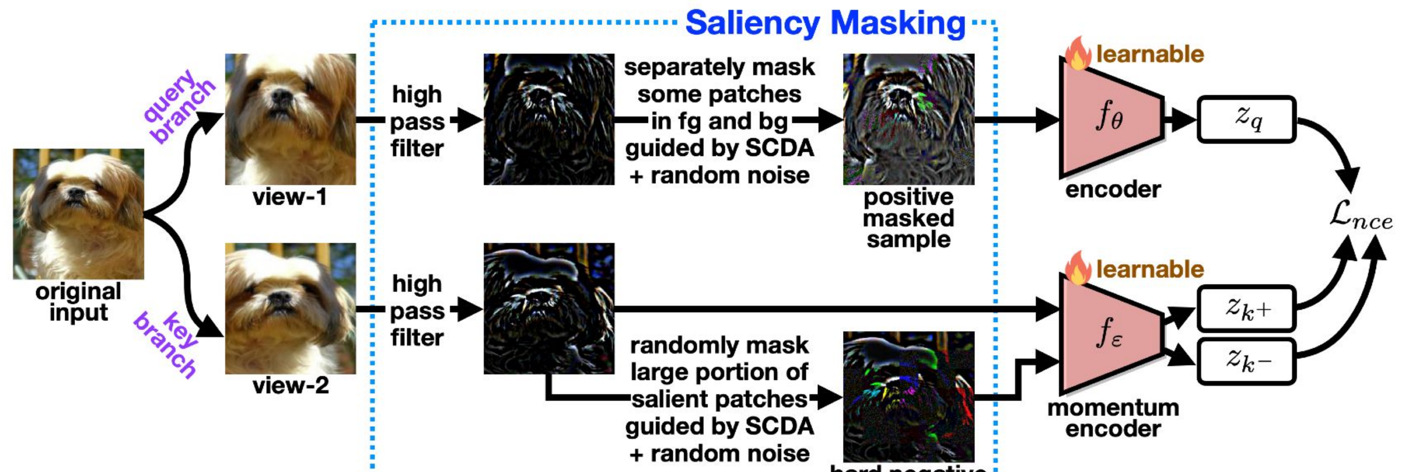


Three Masking Strat. (How to Mask)

- High passs filtering: apply high-pass filter before masking and add Gaussian noise to the mask area.
- Strong blurring: Blur the masked area.
- Mean filling: Fill the masked area with the mean pixel value.



Overall Training Pipeline



$$\mathcal{L}_{nce} = -\log \frac{\exp(z_q^\top z_{k+}/\tau)}{\sum_{X_\neg} \exp(z_q^\top z_\neg/\tau) + \exp(\rho z_q^\top z_{k-}/\tau)}$$

- z_q: query branch, perform saliency masking
- z_{k+}: key branch, **does not perform masking**
- z-: other images in the queue
- z_{k-}: hard negative sample, perform saliency making with higher masking ratio on the key view

Experimental Results

Transfer Classification Results

Method	ImageNet-100	Caltech-101	Flowers-102
Supervised	82.72	21.99	20.29
MoCov2	68.22	81.87	88.39
+ MSCN [10]	70.28	**84.13**	90.10
+ ADIOS [18]	62.76	79.83	88.39
+ OURS (High-pass filtering)	**73.80**	**84.91**	**90.95**
+ OURS (Strong blurring)	**72.50**	83.95	90.59
+ OURS (Mean filling)	70.84	82.68	**90.83**
SimCLR	69.77	78.20	85.21
+ MSCN [10]	77.18	**86.99**	**91.08**
+ ADIOS [18]	71.12	81.96	87.53
+ OURS (High-pass filtering)	**77.90**	**87.04**	90.71
+ OURS (Strong blurring)	**77.78**	83.41	**91.93**
+ OURS (Mean filling)	77.36	83.55	90.83

Transfer Detection/ Instance Segmentation Results

Method	VOC07+12 detection			COCO detection			COCO instance segmentation		
	AP_{all}	AP_{50}	AP_{75}	AP^{bb}_{all}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}_{all}	AP^{mk}_{50}	AP^{mk}_{75}
Supervised	44.30	73.47	46.50	37.84	57.09	40.67	33.14	53.95	35.31
MoCov2	50.27	76.68	54.76	38.52	57.62	41.67	33.75	54.70	35.86
+ MSCN	50.27	76.99	54.70	38.80	58.09	**42.20**	33.89	54.78	**36.36**
+ ADIOS	45.85	73.44	48.45	38.12	57.38	41.29	33.38	54.25	35.63
+ OURS (High-pass filtering)	**50.89**	**77.66**	**55.44**	**39.16**	**58.62**	**42.45**	**34.22**	**55.28**	36.30
+ OURS (Strong blurring)	**50.76**	**77.29**	54.75	38.90	**58.13**	42.11	**33.93**	54.77	**36.53**
+ OURS (Mean filling)	50.59	76.97	**55.30**	**38.93**	58.08	42.17	33.92	**54.86**	36.27
SimCLR	40.34	69.86	40.96	36.30	55.55	38.80	31.99	52.28	33.80
+ MSCN	43.50	73.18	**45.04**	37.88	57.44	40.68	33.36	54.15	35.57
+ ADIOS	**43.83**	**73.42**	**45.01**	**38.76**	**58.35**	**41.96**	**33.94**	**54.96**	**36.23**
+ OURS (High-pass filtering)	**43.76**	**73.43**	44.90	**38.45**	**57.79**	**41.58**	**33.90**	**54.70**	**35.93**
+ OURS (Strong blurring)	43.20	73.15	44.27	37.44	56.80	39.96	32.92	53.73	35.00
+ OURS (Mean filling)	43.20	72.54	44.79	37.27	56.46	40.10	32.68	53.35	34.54

Manipulating Variance

Setting	Mask branch	Top1
Baseline MoCov2	✗	56.00
High-pass filtering	key	52.25
	both	56.29
	query	**58.19**
Strong blurring	key	51.06
	both	56.83
	query	**58.28**
Mean filling	key	47.53
	both	56.86
	query	**58.34**

We observe that exclusively masking in the query branch leverages variance manipulation in the two branches of the siamese network, resulting in improved training benefits.

Conclusion

- We introduce a salient masking augmentation method for contrastive self-supervised learning using a ConvNet backbone.
- In comparison to randomly masking patches of the input image, our salient masking approach generates masks with higher semantic relevance.
- Alongside masked positive samples, we additionally present a straightforward technique to generate hard negative samples based on three distinct masking strategies, which further enhances the capacity for training the feature encoder.
- The extensive outcomes of our experiments distinctly validate the effectiveness and superiority of our proposed method.

Acknowledgement

This work is supported by NSTC (National Science and Technology Council) 111-2628-E-A49-018-MY4 and 112-2221-E-A49-087-MY3, Higher Education Sprout Project of the National Yang Ming Chiao Tung University, as well as Ministry of Education, Taiwan.