Poster Session:
THU-PM-367

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

# Vector Quantization with Self-Attention for Quality-Independent Representation Learning

Zhou Yang[1] Weisheng Dong[1*] Xin Li[2] Mengluan Huang[1]
Yulin Sun[1] Guangming Shi[1]

[1]School of Artificial Intelligence
Xidian University

西安電子科技大學
XIDIAN UNIVERSITY

2Lane Department. of CSEE
West Virginia University

West Virginia University

# Overview

- Problems:
  The robustness of deep models on the potential data distribution shift

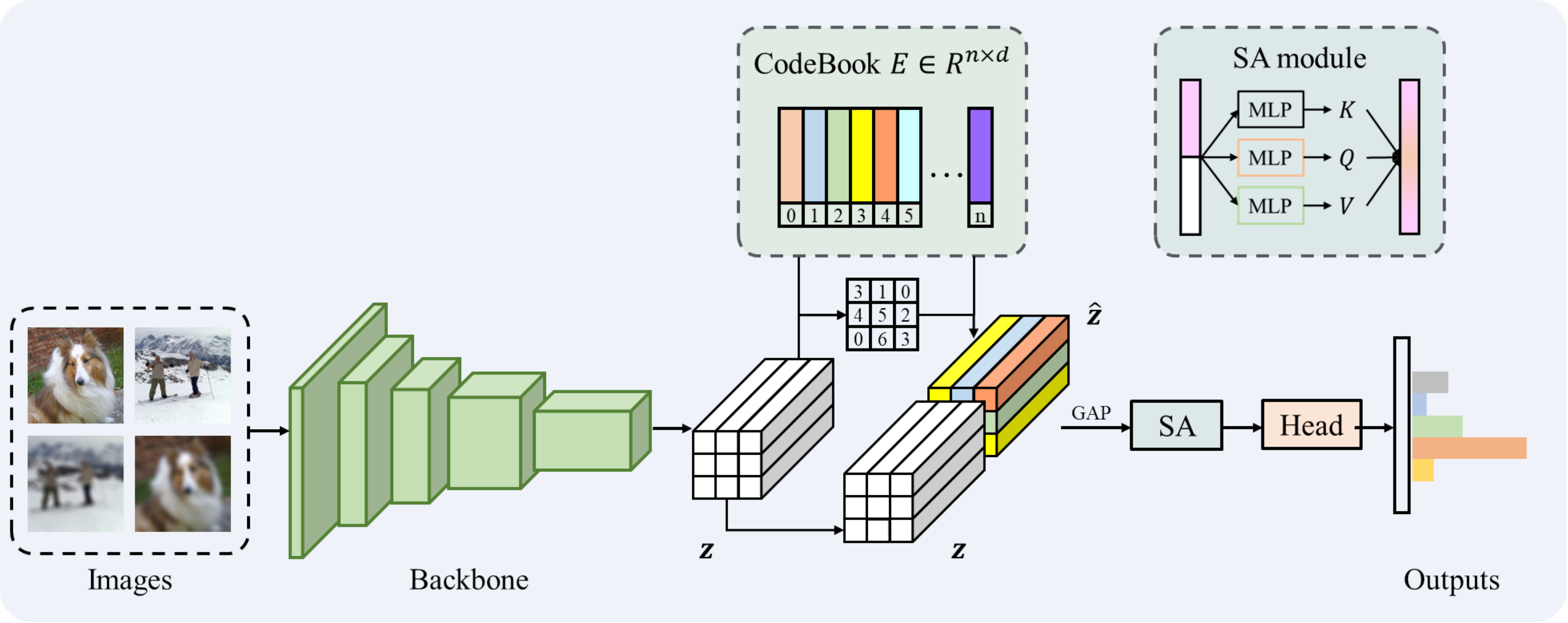
- Motivation:
  Sparse representation can remove redundancy in signals and works well in image restoration

- Method:
  Vector quantization for quality-independent feature representation learning.

- Results:
  Better recognition performance on several benchmark datasets.

# Overview

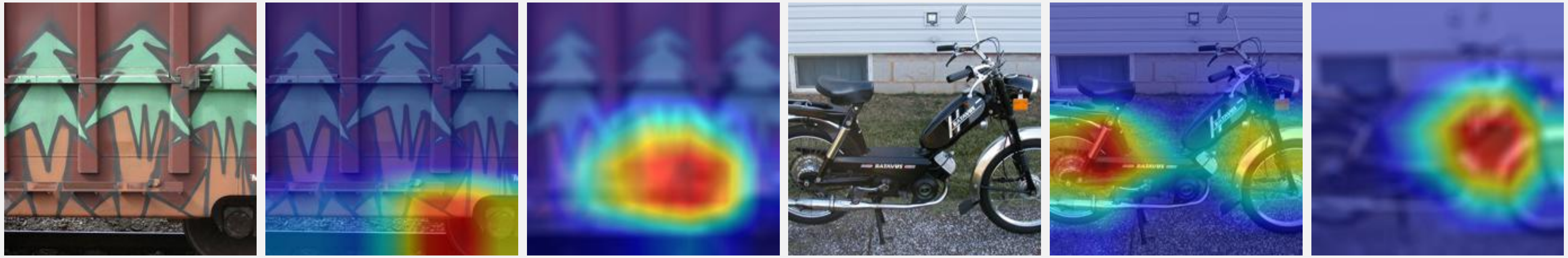
➢ The overall architecture of our proposed method:



# background

clean fog blur noise

Deep models are typically trained on high-quality images but have poor recognition accuracy for low-quality ones



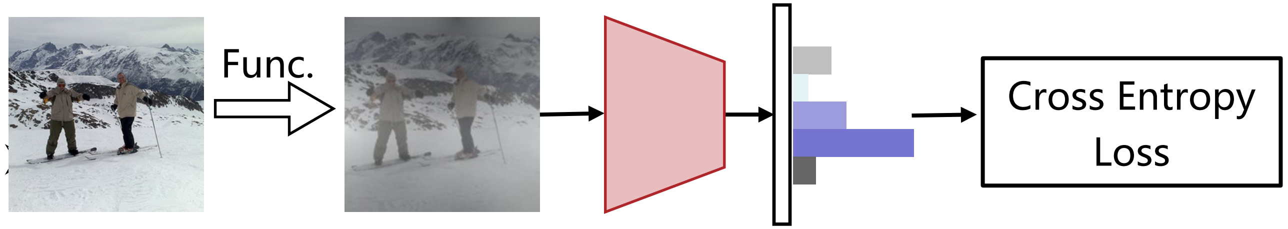
Visualization of CAM Class Activation Maps for Deep Models on Clean and Defocus-Blur Images

Deep features extracted from low-quality images are interfered, affecting the recognition

# Related Work

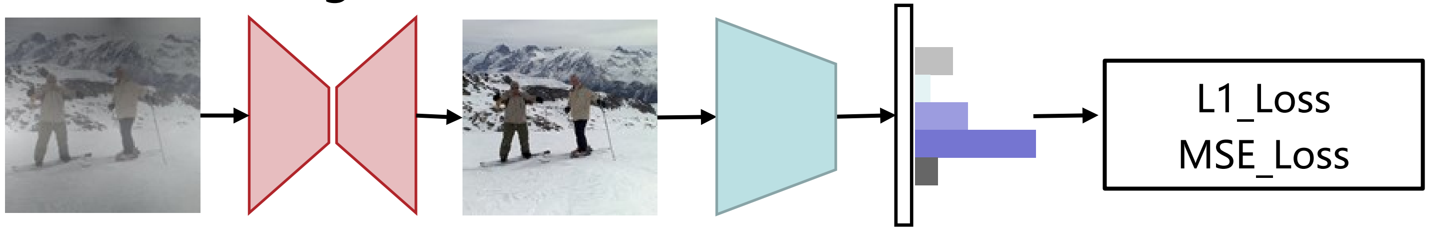


- Existing recognition methods for degraded images
  - Fine-tuning the models by mimicking real-world corruption



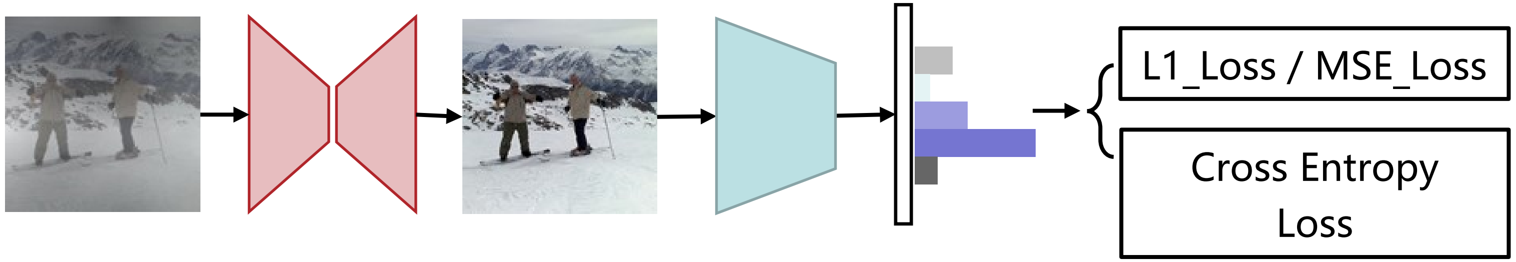

Average Distribution

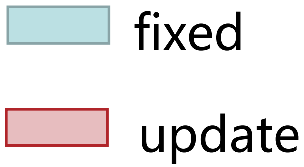
  - Based on image restoration：
    - a): recognition after restoration, Haze removal (ECCV'18)




Two Stage

    - b): Recognition friendly: Denoise&Recog (IJCAI'19), URIE (ECCV'20)

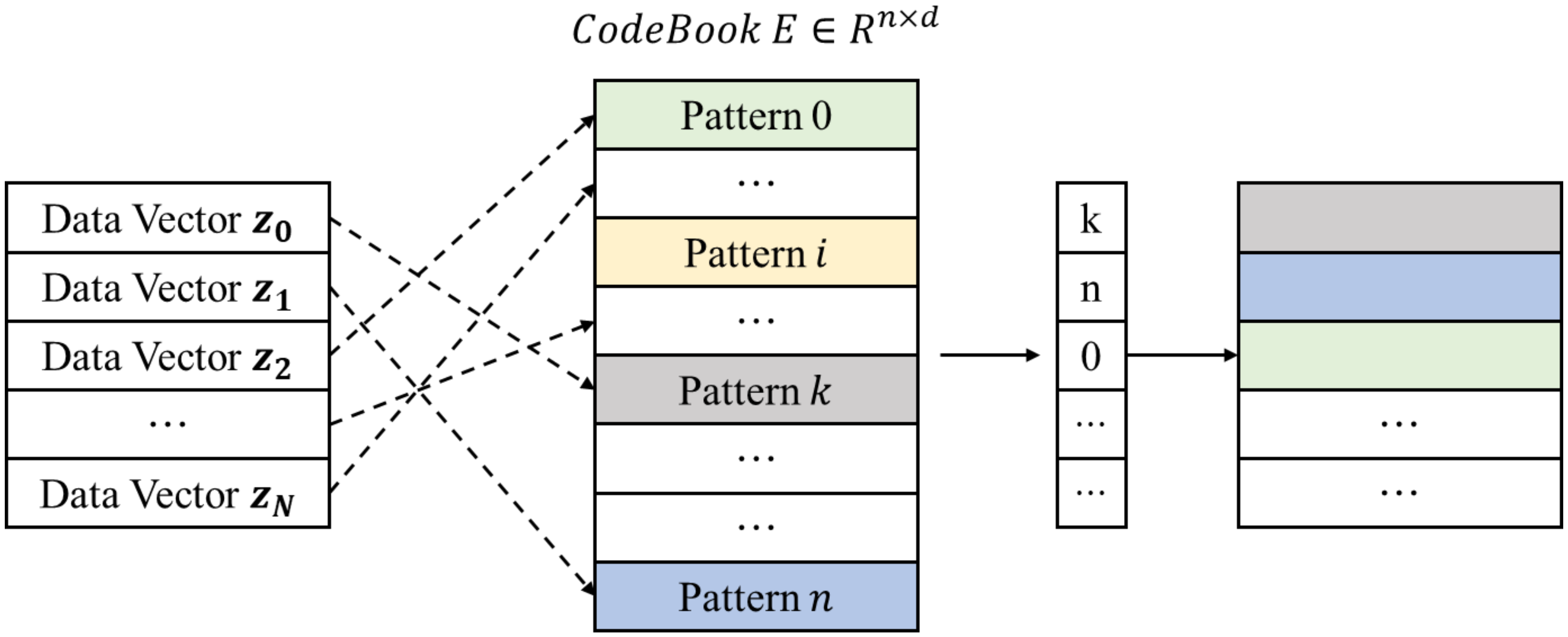


# Related Work



- Vector quantization:




$$\hat{\boldsymbol{z}} = E(\boldsymbol{z}) = \boldsymbol{e_k}, \qquad where\ k = arg \min_i \left|\left|\boldsymbol{z} - \boldsymbol{e_i}\right|\right|_2^2$$

# Method

- Quality-independent feature representation learning:
  - Assuming that the quality-independent feature vector $\hat{\boldsymbol{z}}$ of an image is a linear combination of a series of features (atoms):

$$\hat{z} = \sum_i \alpha_i * \boldsymbol{e_i} = \alpha_0 * \boldsymbol{e_0} + \alpha_1 * e_1 + \cdots + \alpha_n * e_n,\ \boldsymbol{e_i} \in E,\ \mathrm{E} \in R^{n \times d} \quad (1)$$

  - We have this sparse representation, which need to optimize $\boldsymbol{\alpha}$ and $E$ alternately:

$$\hat{\boldsymbol{z}} = E * \hat{\alpha},\ \ \hat{\alpha} = arg\min_{\alpha} \left|\left|\boldsymbol{z} - E * \boldsymbol{\alpha}\right|\right|_2^2 + \lambda * \left|\left|\boldsymbol{\alpha}\right|\right|_0 \quad (2)$$

  - Simplify $\boldsymbol{\alpha}$ as an one-hot vector, we have:

$$\hat{\boldsymbol{z}} = e_k \quad (3)$$

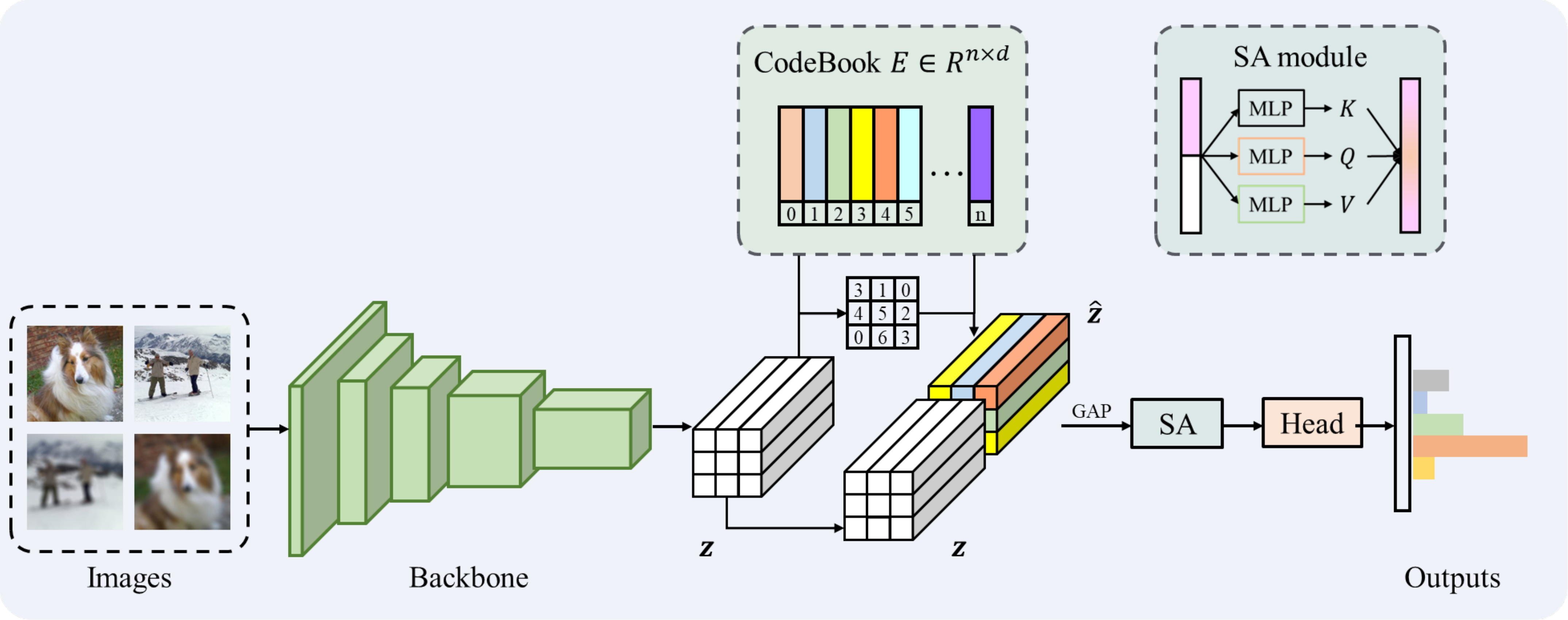
  - Vector quantization as VQ-VAE:

$$L_{vq} = \left|\left|sg(\boldsymbol{z}) - \hat{\boldsymbol{z}}\right|\right|_2^2 \quad (4)$$

$$L_{cmt} = \left|\left|\boldsymbol{z} - sg(\hat{\boldsymbol{z}})\right|\right|_2^2 \quad (5)$$

# Method

➢ Overall architecture of our proposed approach:



# Method

- Training Loss:

$$L_{total} = L_{ce} + \lambda * (L_{vq} + \beta * L_{cmt}) \tag{6}$$

- Concatenate & self-attention:

$$f = Cat(\boldsymbol{z}, \hat{\boldsymbol{z}}) \tag{7}$$

$$f_{sa} = softmax\left(K * \frac{Q^T}{\sqrt{d^n}}\right) * V \tag{8}$$

- Experiments have shown that these skills can further improve performance.

# Ablation Study

- The impact of codebook size n:

| Size | # Params | clean ↑ | mCE ↓ |
|---|---|---|---|
| $n = 1k$ | $4.0 \times 10^7$ | 76.1 | 45.7 |
| $n = 10k$ | $5.8 \times 10^7$ | 76.6 | 43.1 |
| $n = 100k$ | $2.4 \times 10^8$ | 76.6 | 42.9 |

- The choice of fusion mode:

| CodeBook | Fusion mode | SA | clean ↑ | mCE ↓ |
|---|---|---|---|---|
| - | - | - | 73.1 | 53.7 |
| ✓ | replace | - | 74.3 | 50.1 |
| ✓ | add | - | 74.7 | 48.9 |
| ✓ | concat | - | 76.2 | 45.7 |
| ✓ | concat | ✓ | **76.6** | **43.1** |

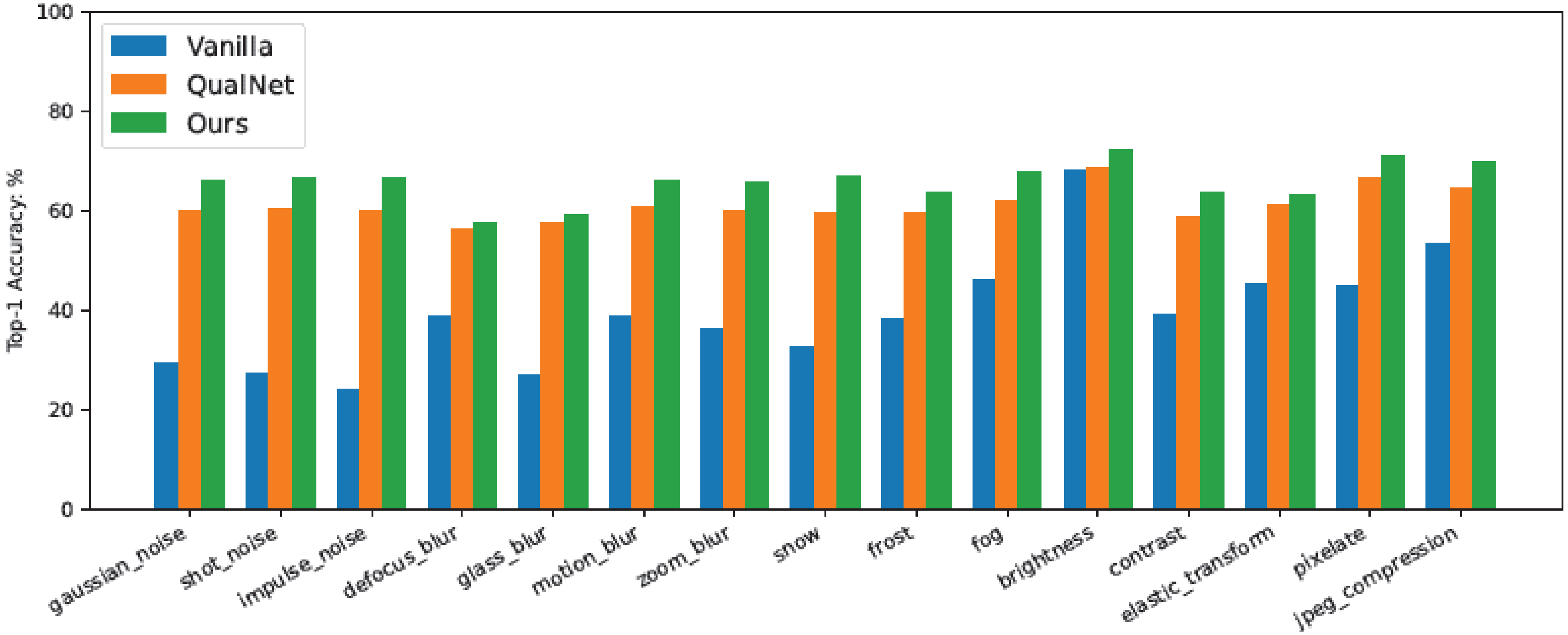
# Results

| Method | Backbone | Clean | Known | UnKnown | mCE ↓ |
|---|---|---|---|---|---|
| Vanilla [20] | ResNet50 | 76.1 | 39.1 | 46.7 | 76.7 |
| DDP [55] | | 72.1 | 48.2 | 50.7 | 62.78 |
| URIE [47] | | 73.8 | 55.1 | 56.5 | 55.7 |
| QualNet [29] | | 75.4 | 61.1 | 58.1 | 50.3 |
| Ours | | **76.6** | **65.6** | **60.2** | **43.1** |
| Vanilla [20] | ResNeXt101 | 79.6 | 47.1 | 55.5 | 69.7 |
| QualNet [29] | | 77.8 | 65.5 | 63.3 | 42.6 |
| Ours | | **80.3** | **68.6** | **64.5** | **37.9** |

| Method | Clean | ImageNet-C ↓ | ImageNet-A | ImageNet-R |
|---|---|---|---|---|
| Vanilla [20] | 76.1 | 76.7 | 0.0 | 36.2 |
| + Ours | **76.6** | **71.1** | **3.7** | **38.6** |
| DeepAugment [21] | 76.6 | 60.4 | 3.5 | 42.2 |
| + AugMix [23] | 75.8 | 53.5 | 3.9 | 46.8 |
| + DAT [35] | 77.1 | 50.8 | **6.8** | 47.8 |
| DAu+AM+Ours | **77.4** | **48.7** | 5.9 | **49.3** |

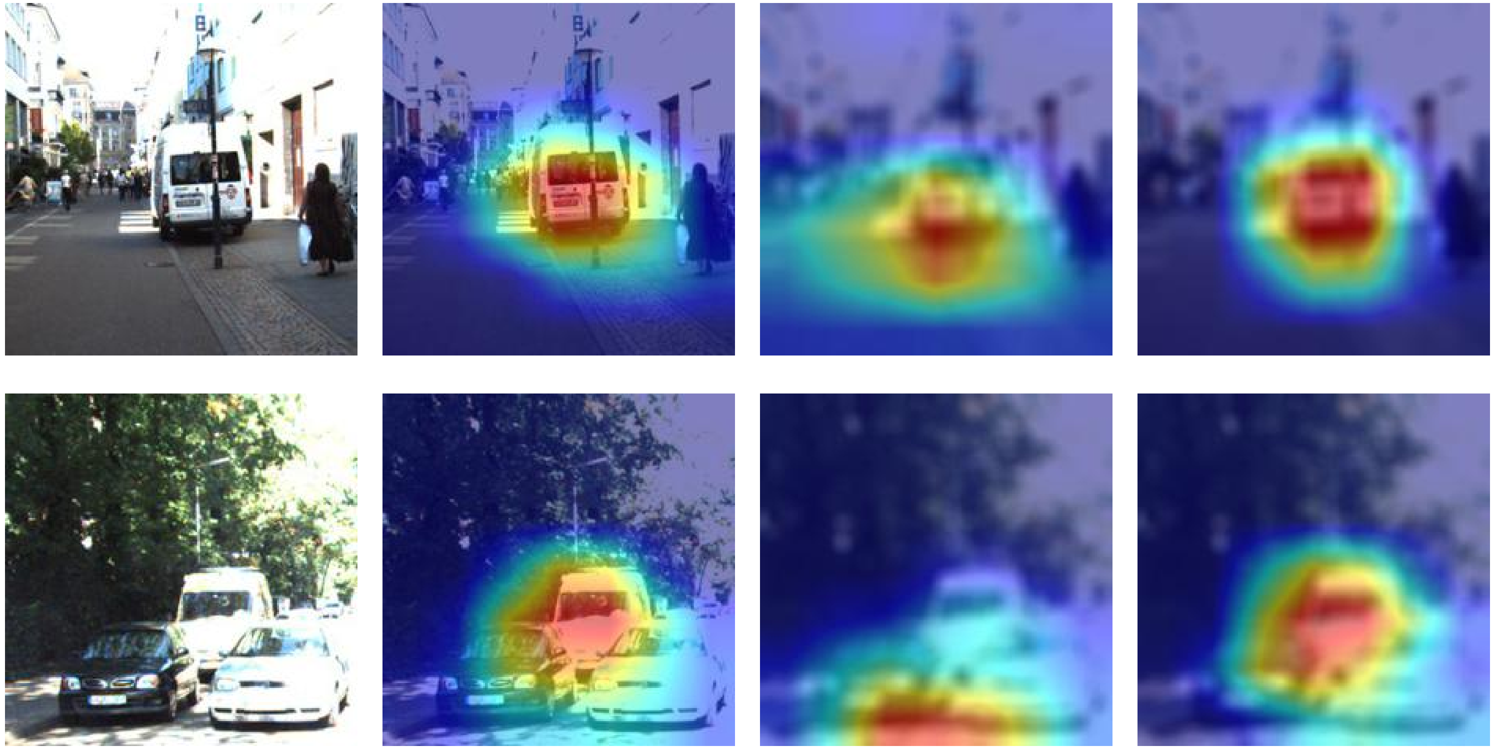
The top-1 accuracy of each method on several benchmark datasets.

# Results




The detailed top-1 accuracy results of the different methods for each corruption type in the benchmark dataset ImageNet-C.

# Results



Clean image | CAM on clean | CAM on blur | After our method

# Summary

- We propose to introduce vector quantization into the recognition model and improve the models' robustness on common corruptions.
- We concatenate the quantized feature vector with the original one and use the self-attention module to enhance the quality-independent feature representation instead of direct replacement in the standard vector quantization method.
- Extensive experimental results show that our method has achieved higher accuracy on benchmark low-quality datasets than several current sota methods.

# Thank you for watching!

Code is available at: https://github.com/yangzhou321/VQSA