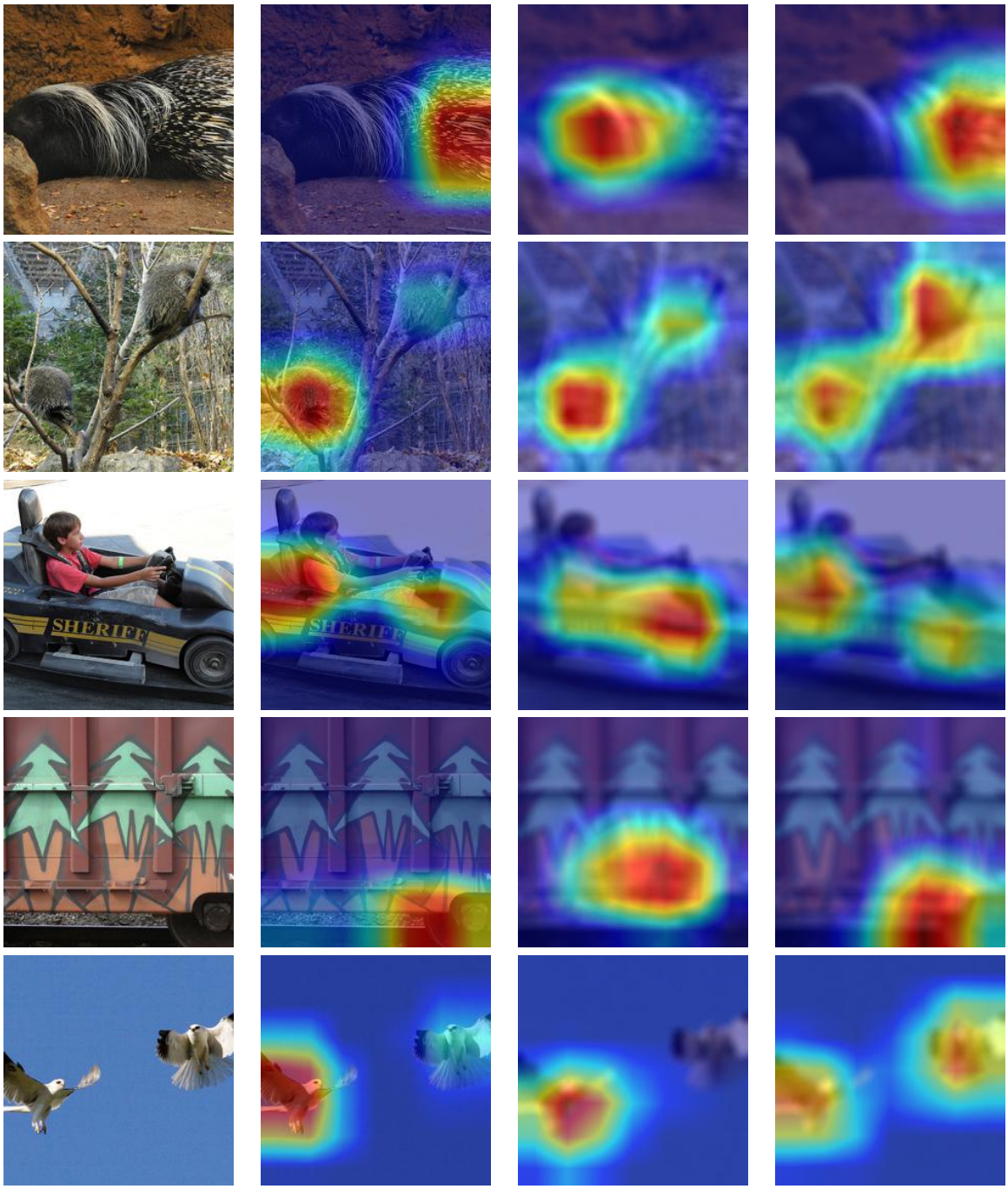
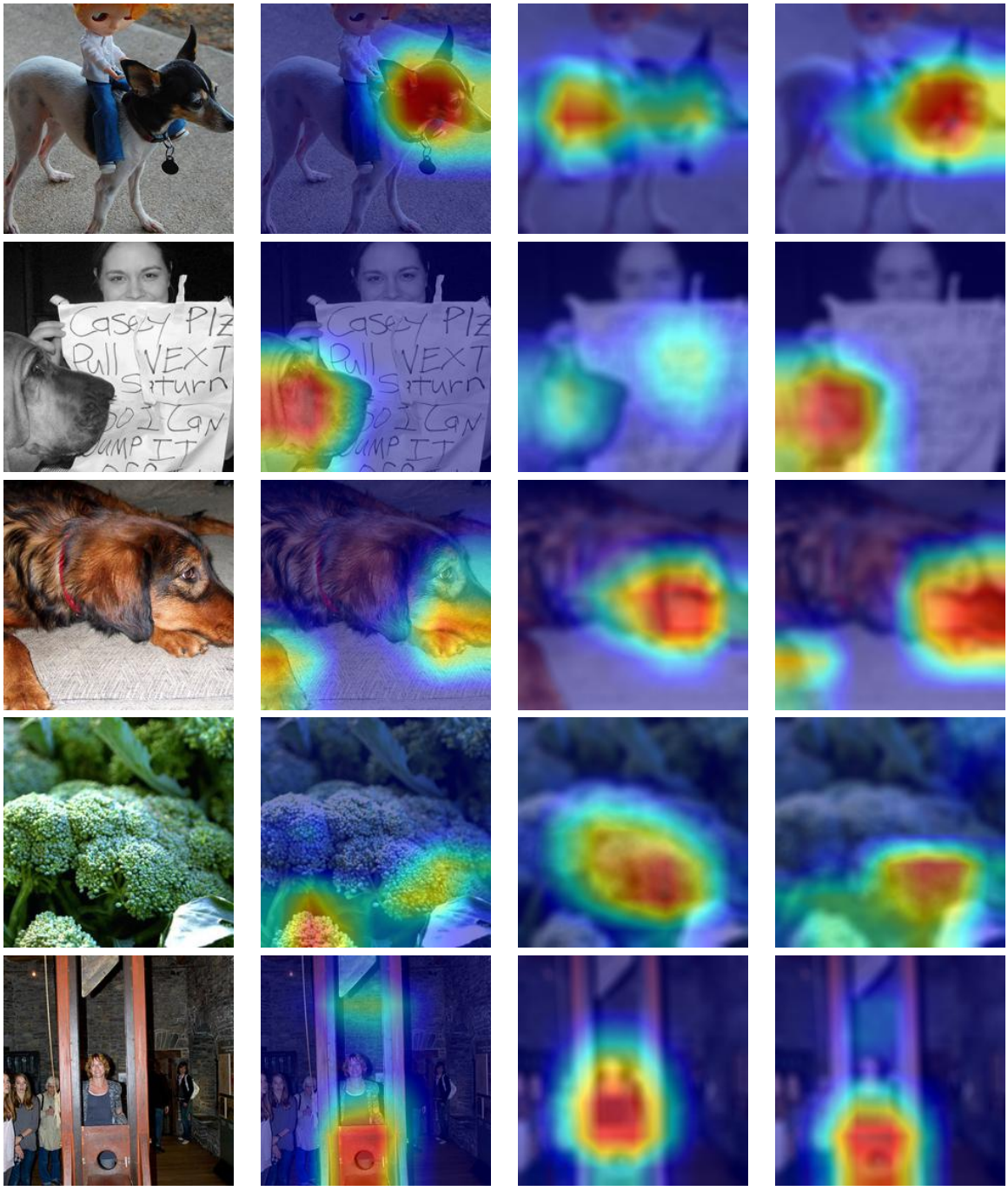
Supplementary Material for ”Vector Quantization with Self-Attention for Quality-Independent Representation Learning”

A. More Visualization Results

In this section, we show more class activation map results based on the Grad-CAM [4] technology to verify the superiority of our method. The results are shown in Tab. 1.





(a) Image (b) Clean (c) QualNet [2] (d) Ours

Table 1. The Grad-CAM [3] maps of different models on defocus blur images. (a) The original clean images. (b) The maps of vanilla ResNet50 [1] model on clean images. (c) and (d) show the maps of QualNet50 [2] and our proposed method on defocus blur images with severity level 3. The results show that our method still can focus on the salient object area without being seriously affected by corruption. **Best viewed in color.**

From the figure, we can draw a conclusion that our method can still focus on the regions more relevant to the image labels of degraded images. This proves that the features extracted by our method are independent of the image quality, that is, no matter what the image quality is, we can learn effective feature representation for recognition.

B. Additional Ablation Study

In this section, we investigate the impact of the size of the codebook $E \in R^{n \times d}$ on the parameters and model performance. The size is the number of features contained in the codebook, that is, the value of n. We tested codebook in three sizes, i.e., $n = 1000$, $n = 10000$, $n = 100000$, specifically. The results are shown in Tab. 2. Noted that codebook under the **Pytorch** framework is implemented using the *torch.nn.embedding* method, and this operator temporarily does not support the calculation of *params* in the **torchstat** library. Therefore, we have manually recalculated it and this is the correct version.

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

Table 2. The influence of codebook size n on parameters and model performance.

From the table, we can see that when the size n is increased to 100000, the improvement of performance is limited, and the number of parameters increases significantly. Therefore, we set $n = 10000$ in our experiments.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[2] Insoo Kim, Seungju Han, Ji-won Baek, Seong-Jin Park, Jae-Joon Han, and Jinwoo Shin. Quality-agnostic image recognition via invertible decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12257–12266, 2021. 2

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[4] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1