A Appendix

In Section A.1, we compare the performance of GradCon with other benchmarking and state-of-the-art algorithms on fMNIST. In Section A.2, we perform statistical analysis and highlight the separation between inliers and outliers achieved by using the gradient-based representations in CIFAR-10. In Section A.3, we analyze different parameter settings for GradCon. Finally, we provide additional details on CURE-TSR dataset in Section A.4.

A.1 Additional Results on fMNIST

We compared the performance of GradCon with other benchmarking and state-of-the-art algorithms using CIFAR-10 and MNIST in Table 3 and 4. In Table 5 of the paper, we mainly focused on rigorous comparison between GradCon and GPND which shows the second best performacne in terms of the average AUROC on fMNIST. In this section, we report the average AUROC performance of GradCon in comparison with that of additional benchmarking and state-of-the-art algorithms using fMNIST in Table 7. The same experimental setup for fMNIST described in Section 5.1 is utilized and the test set contains the same number of inliers and outliers. GradCon outperforms all the compared algorithms including GPND. Given that ALOCC, OCGAN, and GPND are all based on adversarial training to further constrain the activation-based representations, GradCon achieves the best performance in fMNIST only based on a CAE and requires significantly less computations.

Method	ALOCC DR [29]	ALOCC D [29]	DCAE [30]	OCGAN [22]	GPND [24]	**GradCon**
AUROC	0.753	0.601	0.908	0.924	0.933	**0.934**

Table 7: Average AUROC result of GradCon compared with benchmarking and state-of-the-art anomaly detection algorithms on fMNIST.

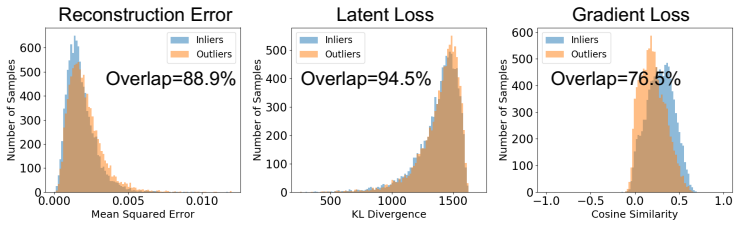

Figure 6. Histogram analysis on activation losses and gradient loss in CIFAR-10. For each class, we calculate the activation losses and the gradient loss from inliers and outliers. The losses from all 10 classes are visualized using histograms. The percentage of overlap is calculated by dividing the number of samples in the overlapped region of the histograms by the total number of samples.

A.2 Histogram Analysis on CIFAR-10

We presented histogram analysis using gray scale digit images in MNIST to explain the state-of-the-art performance achieved by GradCon in Fig. 5. In this section, we perform the same histogram analysis using color images of general objects in CIFAR-10 to further highlight the separation between inliers and outliers achieved by the gradient-based representations. We obtain histograms for CIFAR-10 through the same procedures that are used to generate histograms for MNIST visualized in Fig. 5. In Fig. 6, we visualize the histograms of the reconstruction error, the latent loss, and the gradient loss in CIFAR-10. Also, we provide the percentage of overlap between histograms from inliers and outliers. The measured error on each representation is expected to differentiate inliers from outliers and achieve as small as possible overlap between histograms. The gradient loss shows the smallest overlap compared to other two losses defined in activation-based representations. This statistical analysis also supports the superior performance of GradCon compared to other reconstruction error or latent loss-based algorithms reported in Table 3.

Comparison between histograms from MNIST visualized in Fig. 5 and those from CIFAR-10 shows that the gradient loss is more effective when data becomes complicated and challenging for anomaly detection. In MNIST, simple low-level features such as curved edges or straight edges can be class discriminant features for anomaly detection. On the other hand, CIFAR-10 contains images with richer structure and features than MNIST. Therefore, normal and abnormal data are not easily separable and the overlap between histograms is significantly larger in CIFAR-10 than MNIST. In CIFAR-10, the overlap of the gradient loss is smaller than the second smallest overlap of the reconstruction error by 12.4%. In MNIST, the overlap of the gradient loss is smaller than the second smallest overlap by 5.7%. GradCon also outperforms other state-of-the-art methods by a larger margin of AUROC in CIFAR-10 compared to MNIST. The overlap and performance differences show that the contribution of the gradient loss becomes more significant when data is complicated and challenging for anomaly detection.

A.3 Parameter Setting for the Gradient Loss

We analyze the impact of different parameter settings on the performance of GradCon. The final anomaly score of GradCon is given as $\mathcal{L} + \beta\mathcal{L}_{grad}$, where $\mathcal{L}$ is the reconstruction error and $\mathcal{L}_{grad}$ is the gradient loss. While we use α parameter to weight the gradient loss and constrain the gradients during training, we observe that the gradient loss generally shows better performance as an anomaly score than the reconstruction error. Hence, we use $\beta = n\alpha$, where n is constant, to weight the gradient loss more for the anomaly score. We evaluate the average AUROC performance of GradCon with different β parameters using CIFAR-10 in Fig. 7. In particular, we change the scaling constant, n, to change β in the x-axis of the plot. The performance of GradCon improves as we increase β in the range of $\beta = [0, 2\alpha]$. Also, GradCon consistently achieves state-of-the-art performance across a wide range of β parameter settings when $\beta \geq 1.67\alpha$. To

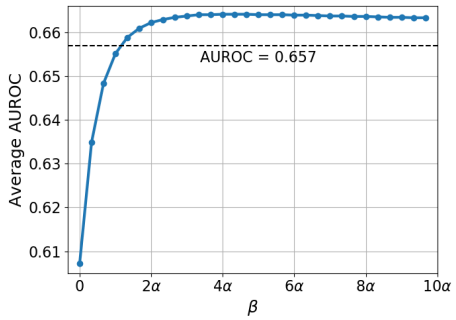

Figure 7. Average AUROC results with different β parameters in CIFAR-10. $\alpha = 0.03$ is utilized to train the CAE. The dotted line (average AUROC = 0.657) indicates the performance of OCGAN which achieves the second best performance in CIFAR-10.

be specific, GradCon always outperforms OCGAN which achieves the second best average AUROC performance of 0.657 in CIFAR-10 when $\beta \geq 1.67\alpha$. This analysis shows that GradCon achieves the best performance in CIFAR-10 across a wide range of β.

A.4 Additional Details on CURE-TSR Dataset

We visualize traffic sign images with 8 different challenge types and 5 different levels in Fig. 8. Level 5 images contain the most severe challenge effect and level 1 images are least affected by the challenging conditions. Since level 1 images are perceptually most similar to the challenge-free image, it is more challenging for anomaly detection algorithms to classify level 1 images as outliers. The gradient loss from CAE + Grad outperforms the reconstruction error from CAE in all level 1 challenge types. This result shows that the gradient loss consistently outperforms the reconstruction error even when inliers and outliers become relatively similar under mild challenging conditions.

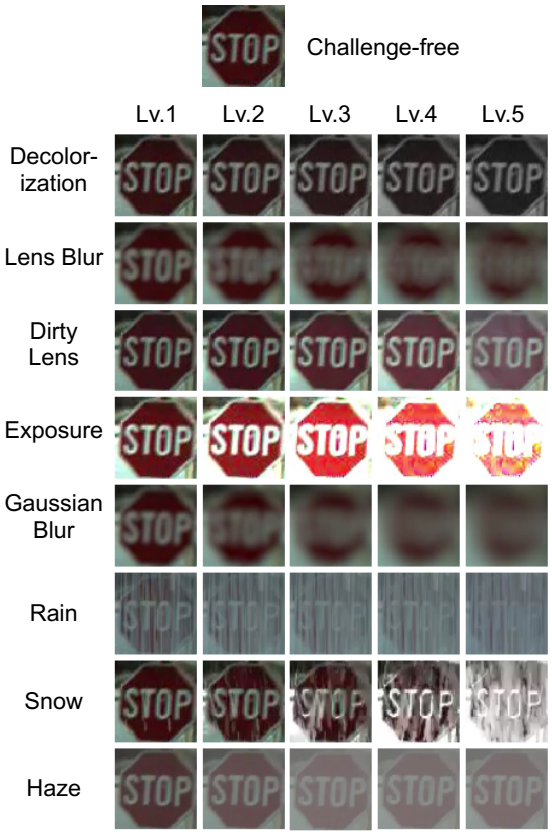

Figure 8. A challenge-free stop sign and stop signs with 8 different challenge types and 5 different challenge levels. Challenging conditions become more severe as the level becomes higher.