We sincerely appreciate all reviewers' efforts and valuable comments. Please find our point-by-point rebuttal below:

Reviewer 1: Thanks for detailed comments and constructive advice. **Q1:** Data augmentation and latent dimension for AE. **Reply:** First, we must point out that using data from pseudo classes, as SSD of *E^3Outlier* did, cannot make AE perform better. Since AE cannot exploit the discriminative label information of pseudo classes, in original paper we did not use their data to train AE. As suggested, we train AE with the same augmented data, but the performance typically becomes worse (e.g. 55.5%/63.9%/54.2%/50.0%/53.8% AUROC on MNIST/F-MNIST/CIFAR10/SVHN/CIFAR100 when $\rho = 10\%$). Second, to fairly compare the quality of learned representation, we must ensure its dimension to be equal for AE and SSD. Note that SSD's penultimate layer, rather than its final K-node classification layer, is used to yield SSD's learned representation (explained in line 136-138). Thus, AE's hidden layer shares SSD's penultimate layer dimension, which is fixed to 256 by Wide-ResNet architecture. It is already smaller than input dimension (3072 or 1024) here. We also test even smaller AE latent dimensions (16, 32, 64, 128): The results show that even for optimal latent dimension (64) that performs best on most benchmarks, it brings minimal gain to AE performance on difficult datasets CIFAR10/CIFAR100 (e.g. 56.3%/56.1% AUROC when $\rho = 10\%$), and on simpler datasets MNIST/F-MNIST/SVHN AE's performance (71.9%/75.6%/53.4%, $\rho = 10\%$) is still far behind *E^3Outlier* (94.1%/93.3%/86.0%) despite limited improvement. In fact, choosing AE's latent dimension priorly is difficult in itself. Our test shows that neither way above helps AE-based methods perform comparably to *E^3Outlier*, and we will add detailed comparisons to paper as suggested. **Q2:** The point of derivation in line 155-180. **Reply:** Although inlier priority seems intuitive, the derivation not only justifies this intuition theoretically, but also provides quantitative measure on "how much" priority inliers will gain in SSD training, i.e. it reveals the quantitative correlation between inliers/outliers' gradient magnitude and the inlier/outlier ratio. We believe that a rigorous conclusion that matches intuition does NOT make it "trivial". Meanwhile, the method to analyze the simplified case can serve as a foundation to inspire further theoretical analysis of complex cases. As to the gradient evolution in training, we did observe that the inliers' gradient magnitude decreases as the training continues, and the performance will drop moderately if too many training epochs are used (see Fig. 4(i)), which implies a better fitting of outliers at this stage. However, the network is still observed to classify inliers better and achieve satisfactory UOD performance. **Q3:** The analysis in line 181-197. **Reply:** We discuss the network updating direction here to provide a more holistic empirical justification of inlier priority, as magnitude and direction are two key factors for the back-propagated gradient vector. **Q4:** The choice of operations. **Reply:** We have conducted an ablation study in Sec. 4.2 based on the combination of different operation sets (see Fig. 4(f)) instead of each individual operation, as it is very time-consuming. In fact, the evaluation and selection of operations is exactly what we are interested in for our next-step research, and our solution will be training a network to examine the geometric property (e.g. symmetry, straightness) in an image to guide the selection of operations. **Q5:** The image artifact of shifting operation. **Reply:** In fact, our experiments show that using operations that create image artifacts (e.g. shifting) alone as surrogate supervision indeed leads to poor performance just as the reviewer pointed out. However, when they are combined with those operations that do not create artifacts (e.g. rotation by $90°$), they can improve the performance of surrogate supervision. We simply fill the artifact with 0, as other padding methods (e.g. nearest neighbor) produce very similar performance. **Improvements:** **(1)** Please see reply to Q1. **(2)** Please see reply to Q2 and Q3, and it should be clarified that the theoretical analysis was NOT intended to illustrate the advantages of *E^3Outlier* over AE-based methods (in fact the discussion for this purpose is given in Sec. 3.1). More importantly, we must point out that inlier priority does NOT apply to AE-based methods: First, AE uses the raw image pixels as learning targets, but the intra-class difference of inlier images can be very large, which means AE does not have an unified learning target. Second, AE is ineffective in learning high-level representations (discussed in Sec. 3.1), which makes it difficult to capture common high-level semantics of inlier images. Both factors above disable inliers from being a joint force to dominate the training of AE and produce inlier priority like SSD, which is also demonstrated by AE's poor UOD performance in empirical evaluations. **(3)** Please see reply to Q4.

Reviewer 2: Thanks for the comments and beneficial suggestions. **Q1:** More introduction to self-supervision. **Reply:** Although both are used in the literature, we prefer surrogate supervision to self-supervision here because it can better reflect the fundamental difference between our method and the commonly-used autoencoder in UOD (autoencoder is also viewed as "self-supervised" in some literature). In the camera-ready version where an additional page is granted, we will provide a more detailed review on this topic as suggested. **Improvements:** **(1)** Please see reply to Q1.

Reviewer 3: Thanks for the comments and useful feedback. **Q1:** Theoretical analysis of AE/CAE. **Reply:** We have made an attempt to analyze AE/CAE theoretically just like *E^3Outlier*. However, since AE/CAE is trained to reconstruct the images from the latent representation, its learning target is different from one image to another, which prevents us from yielding the expectation of its gradient magnitude in theory. In fact, the reason why our method outperforms AE-based methods is mainly discussed in Sec. 3.1. **Q2:** Smaller outlier ratio. **Reply:** In fact, experiments show that our method will perform even better when the outlier ratio ρ is set to smaller values, since the inlier priority will be even larger in such case. For example, when $\rho = 0.5\%$, *E^3Outlier* achieves 96.0%/93.6%/87.4%/91.0%/80.7% AUROC for MNIST/F-MNIST/CIFAR10/SVHN/CIFAR100, and results for $\rho = 1\%$ show a similar trend. These results indicate that our method can perform satisfactorily in a wide range of outlier ratio. **Improvements:** **(1)** Please see reply to Q1.