SafetyNet: Detecting and Rejecting Adversarial Examples Robustly Supplementary Materials

Jiajun Lu, Theerasit Issaranon, David Forsyth
University of Illinois at Urbana Champaign
{jlu23, issaran1, daf}@illinois.edu

1. SceneProof Dataset

Our SceneProof dataset is processed from NYU Depth v2 raw captures, Sintel Synthetic RGBD dataset and Middlebury Stereo dataset. The dataset is split into part I and part II. Part I contains NYU natural image & depth pairs, along with manipulated unnatural scenes (swap depth, insert region, predicted depth, scale & shift depth), refer to Figure 1. It is used to train our classifier and work as test data part I. Part II contains unnatural scenes manipulated by other methods (set depth channel to zero, down sample and then up-sample both RGBD channels, aggressively compress the JPG RGBD images), and image & depth pairs from synthetic dataset and stereo dataset, refer to Figure 2. Part II is used as test data part II to test the generalization ability of our SceneProof network, and check the reactions of our detectors to unseen unnatural inputs. A good detector need to tend to reject unfamiliar data type, which does not exist in training data, because it is hard for classifier to do right classifications on unseen data types. In real application scenarios, it needs to be a human computer hybrid system where computer provides suspicious cases and human makes final decisions. Table 1 includes the dataset constitution, and we plan to release the dataset for academia usages.

	Training	Testing I	Testing II
Natural Scene	141780	57542	N/A
Swap Depth	33927	16094	N/A
Insert Region	30426	13741	N/A
Predicted Depth	53904	17026	N/A
Scale&Shift Depth	23523	10681	N/A
zeroD channel	N/A	N/A	1449
down-up sampled	N/A	N/A	1449
low quality JPG	N/A	N/A	1449
Sintel RGBD	N/A	N/A	54
Middlebury RGBD	N/A	N/A	30
Total	283560	115084	4431

Table 1. Number of image & depth pairs for each data type in each dataset split. Natural Scene has true label and the other data types have false labels.

2. Rejecting by Classification Confidence

Our experiments demonstrate that there is a trade-off between classification confidence and detection easiness for adversarial examples. Adversarial examples with high confidence in wrong classification labels tend to have more abnormal activation patterns, so they are easier to be detected by detectors. While adversarial examples with low classification confidence in wrong labels are harder to be detected. For example, attacks like DeepFool add small and just enough perturbations to change the classification label, so these adversarial examples are sometimes hard to detect. However, these adversarial examples could not assign high classification confidence to the wrong label. If they perform more iterations and increase the wrong class classification confidence, our detector could detect them much easier.

Experiments also show that Type II attacks on our quantized SVM detector together with the classifier produce adversarial examples with low confidence. All these experiments mean that we can use classification confidence as a detection criteria, and it could help us increase the detector's detection ability and decrease the potential to be attacked by Type II attacks.

The classification confidence in our experiments is measured by the ratio of the example's second highest classification confidence to the highest classification confidence. For example, if an image has 60% probability to be a dog and 15% probability to be a cat, our classification confidence is 0.25. We reject examples with classification confidence ratio bigger than a threshold, which means the classifier is unsure about the classification.

The classification confidence rejection results for non attack images and various Type II attack adversarial examples are included in Table 2 for Cifar-10 and Table 3 for ImageNet-1000. Both tables show that rejecting by classification confidence rejects few non attack images while hugely increase the rejection of Type II attack adversarial examples. The benefits of rejecting by classification confidence is also demonstrated in the Type II attacks section.

	Statistics	Non Attack	L0 (II)	L2 (II)	Fast (II)	DeepFool (II)
m-SVM Det	Mean-confident	95.45%	73.95%	69.36%	74.73%	73.71%
	Mean-ratio	0.05	0.29	0.36	0.31	0.36
	Rejection-rate	7.22%	43.58%	53.96%	45.46%	63.22%
Subnet Det	Mean-confident	95.45%	95.71%	96.68%	79.21%	73.72%
	Mean-ratio	0.05	0.03	0.04	0.25	0.36
	Rejection-rate	7.22%	3.98%	5.50%	37.73%	63.22%

Table 2. CIFAR-10 classification confidence rejection results on non attack images, and various gradient descent based Type II attack adversarial examples. **Mean-confident** is the mean of classification confidence for the label with highest probability. **Mean-ratio** is the mean of the ratio of the second highest predicted label confidence to the highest predicted label confidence. **Rejection-rate** is the rate that examples are rejected because the ratio is higher than the threshold. The ratio for Cifar-10 is 0.25, which means the first predicted label confidence must be four times higher than the second one. For non attack data, the classification confidence rejection only rejects small amount of examples; for quantized SVM detector, it rejects majority of Type II attack adversarial examples; for detection subnetwork, the rejection is not as efficient as quantized SVM detector, because getting high classification confidence while fooling detection subnetwork is easier (compared to quantized SVM detector).

	Statistics	Non Attack	L0 (II)	L2 (II)	Fast (II)	DeepFool (II)	DeepFool5 (II)
m-SVM Det	Mean-confident	81.55%	76.80%	41.25%	40.64%	43.93%	37.83%
	Mean-ratio	0.15	0.17	0.43	0.49	0.77	0.51
	Rejection-rate	10.98%	14.26%	43.89%	49.55%	95.51%	51.90%
Subnet Det	Mean-confident	81.55%	67.53%	67.13%	36.65%	43.93%	37.82%
	Mean-ratio	0.15	0.28	0.30	0.51	0.77	0.51
	Rejection-rate	10.98%	25.21%	28.55%	51.80%	95.51%	51.84%

Table 3. ImageNet-1000 classification confidence rejection results on non attack images, and various gradient descent based Type II attack adversarial examples. The table arrangement is same to Table 2, and DeepFool5 is top-5 DeepFool. The rejection ratio threshold is 0.5. For non attack data, the classification confidence rejection only rejects small amount of examples; for quantized SVM detector and detection subnetwork, they reject majority of Type II attack adversarial examples.

3. Type II Attacks on Cifar-10 and ImageNet-1000

Our main paper has results of detections' reactions on non attack data along with Type I attacks for Cifar-10, ImageNet-1000 and SceneProof dataset, and various Type II attacks for SceneProof dataset. In this section, we include the gradient descent based Type II attacks for Cifar-10 and ImageNet-1000 with SVM detector, and compare to detection subnetwork [1]. Metzen et al. [1] only investigated type I attacks and has not investigated type II attacks on the detection subnetwork. Because the gradients of detection subnetwork are better formed, it should be easier to attack with Type II gradient descent attacks.

In our experiments for Cifar-10 and ImageNet-1000, we use different gradient descent based Type II attacks (L0, L2, Fast, DeepFool and top-5 DeepFool) to attack the detector and classifier at the same time. In the main paper, gradient descent based Type II attacks on SceneProof dataset use L2 LBFGS method.

The summary for Type II attacks on Cifar-10 could be found in Table 4. The numbers reported in the table are the percentages of adversarial examples that are both misclassified and undetected (lower is better). Without classification confidence rejection, quantized SVM detector and detection subnetwork perform similar under Type II attacks for L0, L2 and Fast methods, and quantized SVM detector performs significantly better under DeepFool Type II attacks. With classification confidence rejection, quantized SVM detector is very hard to attack and performs better than detection subnetwork on almost all attacking methods. The classification confidence rejection increases at maximum 7% false rejection on non attack images. The detailed percentages of Type II attacks on Cifar-10 could be found in Table 6.

The summary for Type II attacks on ImageNet-1000 could be found in Table 5.The table arrangement is same to Table 4, and DeepFool5 is top-5 DeepFool attack. Quantized SVM detector consistently performs better than detection subnetwork for various attacking methods and for both with classification confidence rejection and without. It's very difficult to perform Type II attacks on quantized SVM detector with rejection. The classification confidence rejection increases at maximum 10% false rejection on non attack images. The detailed percentages of Type II attacks on ImageNet-1000 could be found in Table 7.

References

[1] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017. 2

Method	L0 (II)	L2 (II)	Fast (II)	DeepFool (II)
m-SVM Det	19.73	18.70	6.86	22.01
m-SVM Det - R	**9.86**	**7.32**	3.41	**8.32**
Subnet Det	20.73	12.30	1.89	96.24
Subnet Det - R	19.69	11.57	**1.19**	35.39

Table 4. Percentages of CIFAR-10 Type II attack adversarial examples that are both misclassified and undetected, lower is better. - R means classification confidence rejection is used (rejection ratio is 0.25), otherwise only the detector is on duty. Without classification confidence rejection, quantized SVM detector and detection subnetwork perform similar under Type II attacks for L0, L2 and Fast methods, and quantized SVM detector performs significantly better under DeepFool Type II attacks. With classification confidence rejection, quantized SVM detector is hard to attack and performs better than detection subnetwork on almost all attacking methods.

Method	L0 (II)	L2 (II)	Fast (II)	DeepFool (II)	DeepFool5 (II)
m-SVM Det	25.15	26.40	12.97	45.26	30.08
m-SVM Det - R	**23.19**	**15.05**	**8.26**	**2.32**	**15.52**
Subnet Det	70.52	36.43	21.25	100.00	42.24
Subnet Det - R	52.56	26.66	12.16	4.49	21.99

Table 5. Percentages of IMAGENET-1000 Type II attack adversarial examples that are misclassified and undetected, lower is better. - R means classification confidence rejection is used (rejection ratio is 0.5), otherwise only the detector is on duty. Quantized SVM detector consistently performs better than detection subnetwork for various attacking methods and for both with classification confidence rejection and without. It is difficult to perform Type II attacks on quantized SVM detector with rejection.

		L0 (II)		L2 (II)		Fast (II)		DeepFool (II)	
Cifar-10		undet	det	undet	det	undet	det	undet	det
m-SVM Det	=	37.95	22.58	51.16	19.23	33.45	41.75	1.03	2.71
	$\neq$	19.73	19.75	18.70	10.90	6.86	17.95	22.01	74.23
m-SVM Det - R	=	31.79	28.74	42.23	28.17	30.87	44.32	0.41	3.34
	$\neq$	9.86	29.62	7.32	22.29	3.41	21.39	8.32	87.94
Subnet Det	=	16.91	21.64	28.57	32.01	8.06	66.57	3.76	0.00
	$\neq$	20.73	40.72	12.30	27.13	1.89	23.48	96.24	0.00
Subnet Det - R	=	16.25	22.30	28.02	32.56	7.53	67.10	1.15	2.61
	$\neq$	19.69	41.76	11.57	27.85	1.19	24.18	35.39	60.85

Table 6. Percentage details of Table 4 with correct classification (=) and undetected as adversarials (undet), correct classification and detected as adversarials (det), misclassification ($\neq$) and undetected as adversarials, misclassification and detected as adversarials. Table 4 comes from misclassification and undetected as adversarials (left down corner). *For all Type II attacks, correct classification and detected as adversarials percentage does not matter, because attacks tend to distort activation patterns even when the labels have not been changed.*

		L0 (II)		L2 (II)		Fast (II)		DeepFool (II)		DeepFool5 (II)	
ImageNet-1000		undet	det	undet	det	undet	det	undet	det	undet	det
m-SVM Det	=	0.00	0.00	3.12	1.58	55.21	7.04	0.00	0.00	0.00	0.00
	$\neq$	25.15	74.84	26.40	68.90	12.97	24.78	45.26	54.74	30.08	69.92
m-SVM Det - R	=	0.00	0.00	2.43	2.27	53.06	9.19	0.00	0.00	0.00	0.00
	$\neq$	23.19	76.80	15.05	80.24	8.26	29.48	2.32	97.67	15.52	84.48
m-SVM Det	=	17.67	4.13	33.13	20.69	21.96	13.28	0.00	0.00	0.00	0.00
	$\neq$	70.52	7.68	36.43	9.74	21.25	43.52	100.00	0.00	42.24	57.76
m-SVM Det - R	=	16.03	5.77	30.86	22.97	19.63	15.61	0.00	0.00	0.00	0.00
	$\neq$	52.56	25.64	26.66	19.51	12.16	52.61	4.49	95.51	21.99	78.01

Table 7. Percentage details of Table 5 with correct classification (=) and undetected as adversarials (undet), correct classification and detected as adversarials (det), misclassification ($\neq$) and undetected as adversarials, misclassification and detected as adversarials. Table 5 comes from misclassification and undetected as adversarials (left down corner). *For all Type II attacks, correct classification and detected as adversarials percentage does not matter, because attacks tend to distort activation patterns even when the labels have not been changed.*

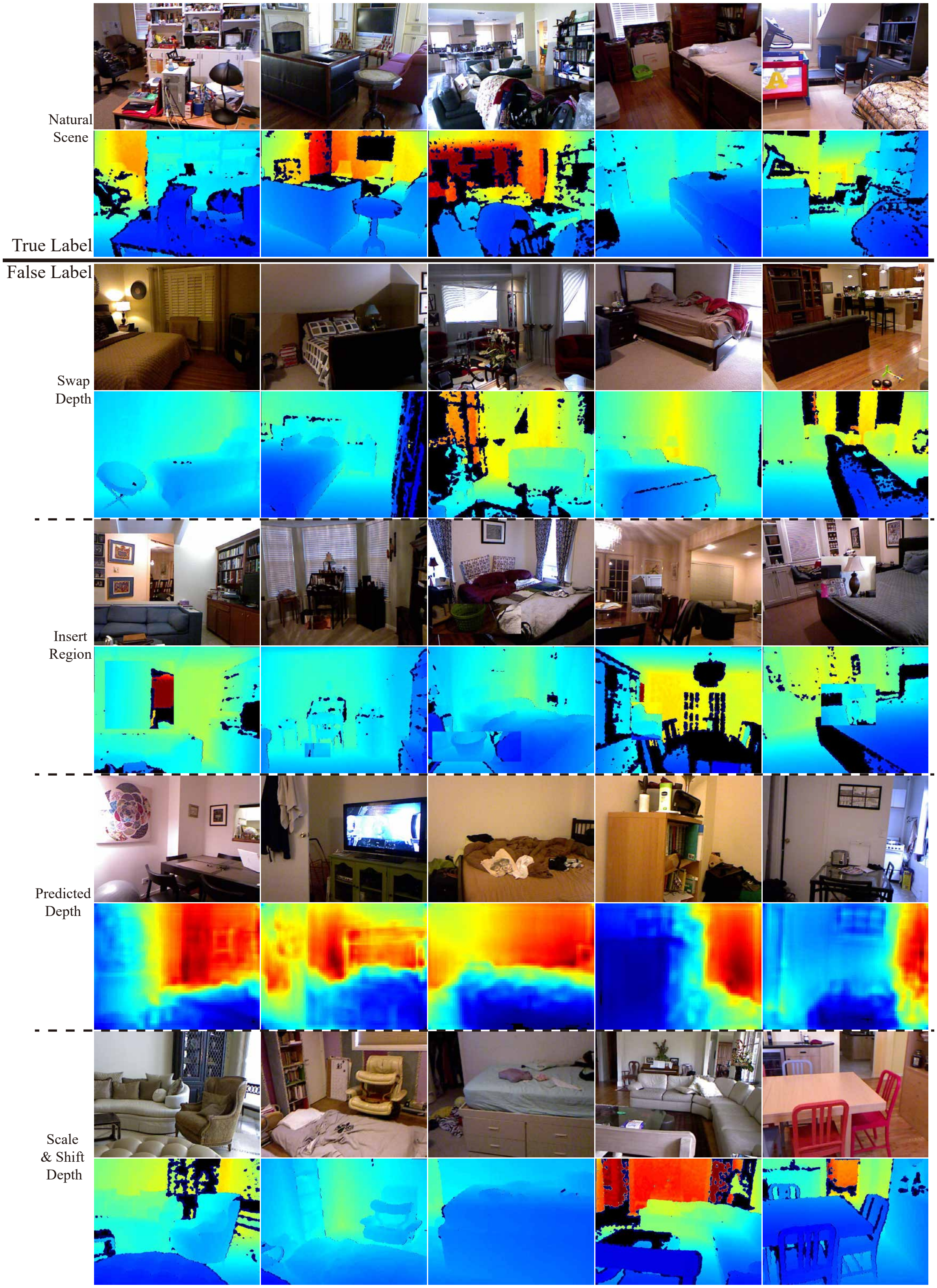

Figure 1. SceneProof dataset part I. Natural Scene has true label, and others have false labels.

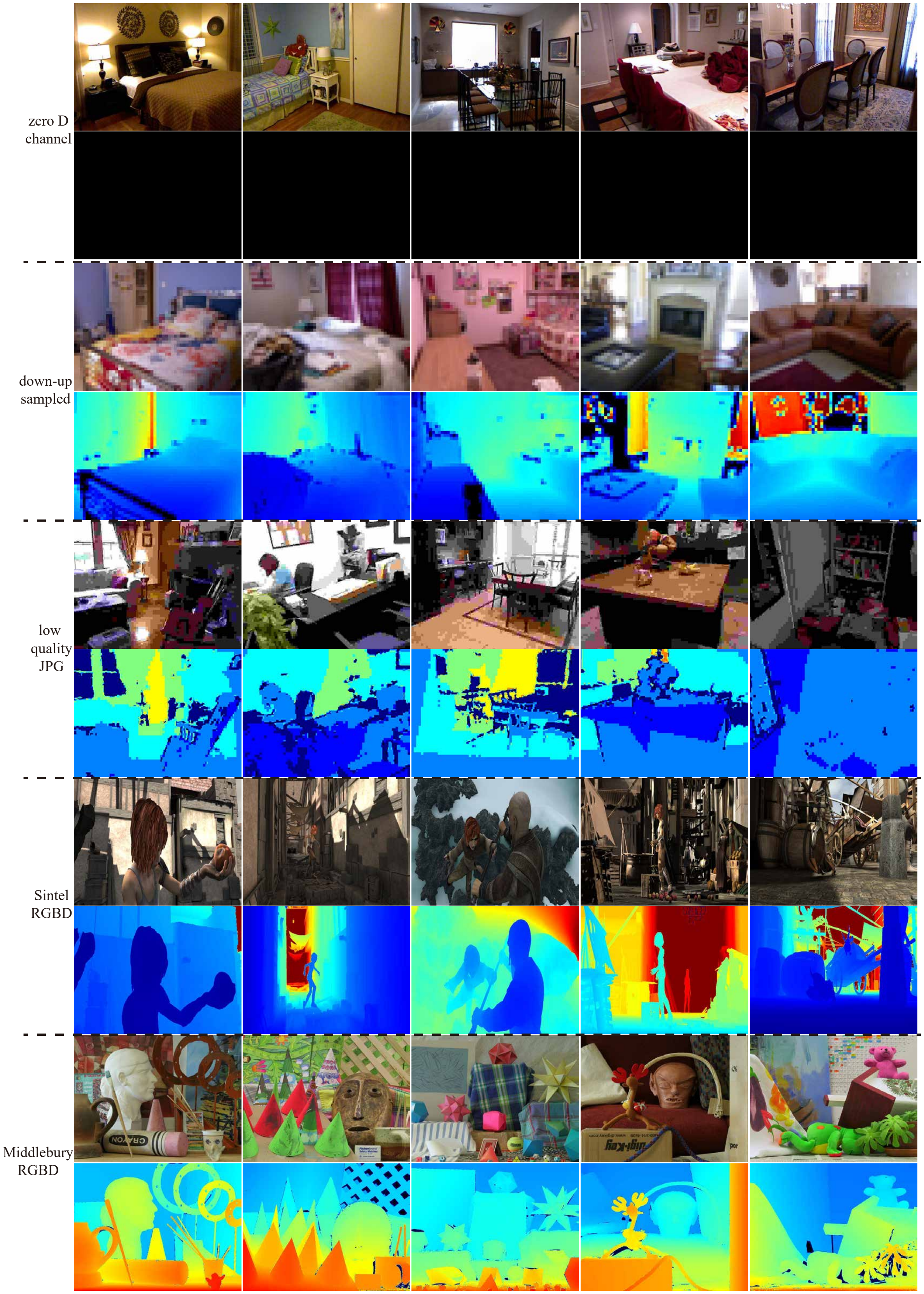

Figure 2. SceneProof dataset part II. All have false labels.