ATS: Adaptive Temperature Scaling for Enhancing Out-of-Distribution Detection Methods - Supplementary Material

Gerhard Krumpl [†1,2] Henning Avenhaus[2] Horst Possegger[1] Horst Bischof[1]

[1]Institute of Computer Graphics and Vision, Graz University of Technology, Austria
[2]KESTRELEYE GmbH, Austria

In the following, we provide detailed results accompanying the evaluations reported in the main manuscript, as well as a breakdown of our runtime analysis.

1. Detailed Results

CIFAR. Table 1, Table 2, and Table 3 provide a comprehensive breakdown of the CIFAR [9] results for each of the eight out-of-distribution (OOD) datasets (SVHN [12], Textures [2], iSUN [18], LSUN (resized and cropped) [19], Places365 [20], MNIST [10], and Fashion-MNIST [16]), supplementing the average results presented in Table 2 of the main manuscript. Our Adaptive Temperature Scaling method consistently enhances the performance for CIFAR [9] against various OOD datasets.

However, an exception is observed with Places365 [20], where performance is slightly diminished. The observed discrepancy can be attributed to the inherent similarities between samples in Places365 [20] and those in the CIFAR [9] datasets. We manually curated exemplary Places365 [20] classes that exhibit a striking resemblance to CIFAR's in-distribution classes in Fig. 1. Given the semantic similarity or outright overlap of multiple samples with ID classes, the validity of using the whole Places365 [20] dataset as an OOD test set is questioned. We further analyzed Places365 [20] as an OOD test set for ResNet18 trained on CIFAR-100 [9] in order to investigate the observed drop in FPR95 performance for this setting (decrease by 4.7% using MSP with ATS, see Table 1 in the main manuscript). In Fig. 2, we present samples from Places365 [20] that were identified as OOD under MSP [7] but were reclassified as ID upon applying ATS. A closer examination reveals that these samples have high similarities or overlap with CIFAR-100's ID classes. We argue that the observed performance dip can be ascribed predominantly to these overlapping samples, further emphasizing the efficacy of ATS.

Across all CIFAR benchmarks, the per-sample temperature used as a standalone method demonstrates promising results. However, the performance is notably enhanced when we use the temperature to scale the logits, increasing the gap between ID and OOD samples and thereby improving the model's overall performance in out-of-distribution detection.

ImageNet. Table 4 shows the per-dataset (iNaturalist [15], SUN [17], Places [20], Textures [2], NINCO [1], and Fashion-MNIST [16]) results for ImageNet [3] (average results reported in Table 3 in the main manuscript). Our approach significantly improves the performance of methods such as MSP [7], ODIN [11], and MLS [6], as they have a comparatively low baseline. Remarkably, ATS still improves the overall result on strong baseline methods, *i.e.*, ReAct [13], DICE [14] and ASH [4], but the improvement is less pronounced. The limited improvement observed compared to CIFAR, especially on strong baselines, can be attributed to the OOD datasets' specific characteristics and the intermediate layer selection process (see Section 4.3 in the main manuscript). The per-sample temperature alone shows weak performance on the ImageNet dataset, but the incorporation of ATS enhances the results. This improvement is particularly prominent when detecting far-OOD data such as Fashion-MNIST [16], emphasizing the effectiveness of ATS in boosting the robustness of OOD detection methods.

We also note that ATS consistently improves the performance on NINCO [1]. This is relevant because NINCO [1] has been cleared from semantically similar samples, where the OOD-ness is unclear, while OOD datasets such as Places or iNaturalist [15] contain overlap with ImageNet [3], calling their value as OOD test sets into question [1].

Summary. ATS provides a favorable trade-off between performance on near-OOD and far-OOD datasets. While ATS generally enhances the robustness of OOD detection, especially against far-OOD samples, there are combinations of ID and OOD datasets where its inclusion results in slightly degraded performance. This trade-off highlights the

[†]Correspondence: gerhard.krumpl@icg.tugraz.at

need for further research on per-sample optimized information extraction from the intermediate layers of the network.

Our analysis further shows that using Places365 [20] as an OOD test set for the CIFAR [9] datasets, and further supported by Bitterwolf *et al.* [1] regarding overlaps with ImageNet [3], not all datasets are suited as impeccable OOD test sets. Nevertheless, we opted to include all relevant datasets in our evaluation for a comprehensive evaluation and to maintain consistency with state-of-the-art benchmarks. This leads to an important realization: there is a pressing need for more rigorous OOD test set design and consequent research, ensuring precise OOD evaluation.

2. Computational Details

In this section, we perform a runtime analysis of our approach, detailed in Algorithm 1 and Algorithm 2, to understand the practical implications. These algorithms provide a detailed procedure for the calibration and inference phases of our Adaptive Temperature Scaling approach.

It should be emphasized that the calibration phase (Algorithm 1) involves pre-computing the eCDF for each chosen intermediate layer based on the in-distribution training set and is conducted offline, thereby not contributing to the computational overhead during test-time. During inference (Algorithm 2), the overhead primarily arises from computing the mean layer activation, which is the "bottleneck" of our approach and scales linearly with the number of activations. The subsequent steps, including p-value retrieval, their aggregation, and temperature scaling, are relatively inconsequential in terms of computational demand.

We conducted an experiment to measure the inference time with and without ATS for three models. The experiment measures the inference time, including temperature calculation and logit scaling, but excludes the logit-based OOD scoring function, comparing the overhead against a normal forward pass. For a precise evaluation, we measure inference times across 10k iterations, preceded by 1k warmup runs using a synthetic input.

Figure 3 depicts the runtime overhead of ATS for three models: ResNet18, ResNet50, and DenseNet100, based on the choice of intermediate layers. In the analysis, we focus on the initial l layers (left plot of Fig. 3) and the concluding l layers (right plot of Fig. 3) for ATS computations. When layers are selected uniformly across the model's depth, the computational overhead remains moderate: 6.18% for ResNet18, 11.94% for ResNet50, and 4.71% for DenseNet100. Furthermore, Fig. 3 underlines that while this overhead is acceptable considering the performance gains, shallow layers impose a greater computational burden due to their large feature dimensions.

Optimizing the runtime overhead is possible by refining the number of intermediate layers, considering the preservation of robustness. Building on insights from Section 4.3 of our main manuscript, it is evident that optimal intermediate layers for ATS vary depending on the ID vs. OOD datasets. Harnessing this understanding, a prudential strategy would dynamically select the best N layers, harmonizing computational speed with OOD detection performance. This synergy accentuates the exciting prospects in fine-tuning layer choices, signaling the need for further research on adaptive layer selection strategies.

The experiment is conducted on a server equipped with an Intel(R) Core(TM) i9-9900X CPU @ 3.50GHz, paired with an NVIDIA GeForce RTX 3080 GPU. The computational setup operates on Ubuntu 22.04, incorporating PyTorch 1.13, and leveraged CUDA 11.6 and cuDNN 8.3.2. The standard neural network forward pass and the adaptive temperature scaling (ATS) operate independently without contending for computational resources.

We reckon that opportunities exist for even more computationally efficient implementations of ATS. Thus, the current evaluation serves as an initial benchmark for the computational overhead.

References

[1] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or Out? Fixing ImageNet Out-of-Distribution Detection Evaluation. In *Proc. ICLR Workshops*, 2023.

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *Proc. CVPR*, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

[4] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely Simple Activation Shaping for Out-of-Distribution Detection. In *Proc. ICLR*, 2023.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. CVPR*, 2016.

[6] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling Out-of-Distribution Detection for Real-World Settings. In *Proc. ICML*, 2022.

[7] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proc. ICLR*, 2017.

[8] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *Proc. CVPR*, 2017.

[9] Alex Krizhevsky and Geoffrey E. Hinton. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[10] Yann Lecun, Lé'on Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based Learning Applied to Document Recognition. *IEEE*, 86(11):2278–2324, 1998.

[11] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *Proc. ICLR*, 2018.

Method	OOD-Dataset																	
	SVHN		Textures		iSUN		LSUN		LSUN-Crop		Places365		MNIST		fMNIST		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
AT-only	96.06	17.02	88.46	43.04	92.42	32.89	93.67	28.79	97.49	15.11	59.62	89.57	97.77	12.00	87.75	55.45	89.16	36.73
MSP	90.23	65.17	89.64	62.79	92.00	55.45	92.49	53.64	93.83	44.71	88.73	62.52	93.08	51.57	93.49	48.10	91.69	55.49
MSP+	**98.33**	8.62	**95.26**	**25.92**	**98.42**	**7.75**	**98.73**	**5.73**	99.41	2.61	87.55	52.23	99.46	1.69	98.26	9.14	**96.93**	**14.21**
ODIN	89.96	48.14	87.50	51.32	96.89	16.65	97.43	13.62	98.18	9.58	89.19	46.73	99.42	2.06	**98.57**	**6.66**	94.64	24.35
ODIN+	97.96	10.53	94.03	28.98	98.26	8.88	98.68	6.40	**99.45**	**2.28**	82.72	66.09	**99.80**	**0.24**	98.23	9.42	96.14	16.60
MLS	91.36	51.01	89.08	55.46	94.26	36.40	94.97	32.10	97.53	14.11	90.62	44.93	97.13	17.08	97.13	16.95	94.01	33.50
MLS+	98.33	**8.61**	95.23	25.98	98.42	7.77	98.73	5.74	99.42	2.58	87.44	52.45	99.46	1.67	98.26	9.14	96.91	14.24
Energy	91.37	50.65	89.03	55.64	94.30	35.89	95.02	31.52	97.64	13.41	**90.67**	**44.44**	97.26	16.05	97.24	16.09	94.07	32.96
Energy+	97.22	14.16	91.30	39.40	96.56	18.66	97.23	14.28	99.17	4.20	78.85	71.47	99.41	2.02	96.60	19.81	94.54	23.00
ReAct	74.76	76.20	67.58	80.12	65.41	82.47	65.03	82.39	74.33	70.42	65.61	81.48	79.67	65.63	72.51	78.30	70.61	77.12
ReAct+	93.98	25.24	84.58	51.61	88.56	47.62	89.55	47.26	93.38	30.01	65.33	82.44	93.06	27.62	84.31	54.91	86.59	45.84
DICE	72.08	75.29	62.20	79.64	63.43	87.55	66.11	85.37	92.21	27.93	60.33	87.08	97.39	12.96	90.30	41.93	75.51	62.22
DICE+	95.19	25.29	84.30	53.58	89.70	41.68	90.50	41.16	95.91	20.85	67.86	80.74	98.58	6.74	90.43	41.14	89.06	38.90
ASH-B	67.08	76.65	62.32	81.76	64.37	80.72	64.46	80.22	74.73	63.44	59.64	85.87	79.71	60.77	72.63	71.71	68.12	75.14
ASH-B+	87.92	44.58	81.05	57.32	85.65	49.64	86.69	49.35	94.09	26.38	60.96	86.75	95.57	24.18	86.76	51.18	84.84	48.67

Table 1. Detailed performance of OOD detection methods with and without ATS for ResNet18 [5] trained on CIFAR-10 [9]. Method AT-only denotes the performance when the per-sample temperature is used directly as the OOD detection score. *Method*+ denotes that our ATS is applied on top of the method (*i.e*., rows with gray background). ↑/↓ indicates that larger/smaller values are better. The **best** and second-best results for each OOD dataset (*i.e*., each column) are shown in bold or underlined, respectively. All values are reported as percentages.

Method	OOD-Dataset																	
	SVHN		Textures		iSUN		LSUN		LSUN-Crop		Places365		MNIST		fMNIST		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
AT-only	97.27	12.96	91.25	33.37	95.29	24.25	95.76	22.05	97.40	14.56	57.85	90.90	98.99	1.72	94.21	34.85	91.00	29.33
MSP	88.21	70.92	89.65	62.20	94.43	43.10	94.97	39.36	95.92	30.82	88.69	63.88	93.99	49.37	94.88	38.00	92.59	49.71
MSP+	98.81	6.00	96.51	19.84	99.49	2.13	99.57	1.59	99.77	0.69	90.83	44.71	99.88	0.02	99.45	1.94	98.04	9.62
ODIN	92.41	45.53	89.02	49.43	99.05	3.94	99.27	2.21	99.34	2.45	91.79	39.74	99.53	0.99	98.93	4.61	96.17	18.61
ODIN+	98.78	6.37	96.31	19.91	**99.59**	**1.62**	**99.67**	**1.16**	99.77	0.56	89.81	49.24	99.94	**0.00**	99.46	1.77	97.92	10.08
MLS	92.15	47.55	88.53	53.40	97.95	10.13	98.25	7.70	99.21	2.83	91.78	40.46	98.79	4.20	98.65	6.03	95.66	21.54
MLS+	98.81	6.00	96.51	19.86	99.49	2.13	99.57	1.60	99.77	0.69	90.81	44.77	99.89	0.02	99.45	1.96	98.04	9.63
Energy	92.24	46.32	88.44	53.03	98.01	9.45	98.31	7.16	99.29	2.57	91.83	39.59	98.92	3.53	98.74	5.42	95.72	20.88
Energy+	98.75	6.56	94.39	26.45	99.29	3.00	99.36	2.42	99.75	0.77	85.55	58.85	99.92	0.01	99.22	3.12	97.03	12.65
ReAct	93.23	44.38	92.12	45.78	98.15	8.56	98.45	6.17	99.30	2.53	**92.09**	**39.47**	98.84	4.44	98.68	5.81	96.36	19.64
ReAct+	98.96	4.90	97.08	16.97	99.47	2.39	99.55	1.74	99.76	0.78	90.70	45.55	99.87	0.02	99.43	2.11	**98.10**	**9.31**
DICE	95.58	25.66	89.16	43.69	99.27	3.03	99.42	2.08	**99.93**	**0.22**	91.00	43.21	**99.97**	**0.00**	**99.62**	**1.51**	96.74	14.92
DICE+	98.29	9.70	96.41	20.82	99.42	2.41	99.50	1.72	99.75	0.72	90.77	43.74	99.87	0.02	99.41	2.18	97.93	10.16
ASH-B	97.34	14.89	94.78	28.14	98.87	5.28	99.06	3.97	99.68	0.97	90.16	45.32	99.69	0.32	99.11	3.96	97.34	12.86
ASH-B+	**99.23**	**3.30**	**97.16**	**15.87**	99.35	2.82	99.44	2.27	99.77	0.61	87.18	55.45	99.91	**0.00**	99.34	2.45	97.67	10.35

Table 2. Detailed performance of OOD detection methods with and without ATS for DenseNet [8] trained on CIFAR-10 [9].

[12] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. *NeurIPS Workshops*, 2011.

[13] Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution Detection With Rectified Activations. In *NeurIPS*, 2021.

[14] Yiyou Sun and Yixuan Li. DICE: Leveraging Sparsification for Out-of-Distribution Detection. In *Proc. ECCV*, 2022.

[15] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The INaturalist Species Classification and Detection Dataset. In *Proc. CVPR*, 2018.

[16] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[17] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proc. CVPR*, 2010.

Method	OOD-Dataset																Average	
	SVHN		Textures		iSUN		LSUN		LSUN-Crop		Places365		MNIST		fMNIST			
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
AT-only	95.55	20.39	86.38	48.40	89.69	44.97	91.12	43.65	93.61	31.66	50.65	94.87	99.02	0.35	92.39	46.11	87.30	41.30
MSP	75.24	85.16	71.62	86.19	77.05	81.15	77.93	79.54	85.91	62.66	74.57	83.41	71.67	89.37	87.26	59.47	77.66	78.37
MSP+	96.86	15.42	90.96	45.76	95.30	26.38	95.95	23.31	99.24	3.42	75.30	82.87	99.07	4.08	99.00	4.67	93.96	25.74
ODIN	71.30	94.83	72.14	85.20	86.59	62.12	87.37	60.18	97.08	17.52	**78.24**	**78.66**	89.44	56.90	97.49	14.23	84.96	58.70
ODIN+	96.04	19.71	91.23	42.77	**96.26**	20.54	96.93	17.09	99.25	3.24	74.79	83.55	99.46	1.52	98.97	4.91	**94.12**	**24.17**
MLS	79.13	86.00	70.98	85.51	81.68	78.17	82.70	76.32	96.74	19.43	77.62	79.85	84.95	73.72	97.16	16.75	83.87	64.47
MLS+	96.86	15.41	90.96	45.66	95.31	26.32	95.95	23.24	99.24	3.42	75.28	82.90	99.07	4.01	99.00	4.66	93.96	25.70
Energy	79.00	86.86	70.79	86.15	81.59	78.82	82.60	77.45	97.10	17.22	77.52	80.23	85.39	71.94	97.49	14.29	83.93	64.12
Energy+	94.47	25.08	86.38	55.83	92.92	38.60	93.68	36.88	98.96	4.85	62.90	92.75	99.46	0.34	98.42	7.64	90.90	32.75
ReAct	76.77	89.66	75.74	84.17	88.73	56.55	89.90	54.55	85.67	57.61	70.20	85.02	68.19	90.55	91.42	43.30	80.83	70.18
ReAct+	**97.85**	**10.21**	92.01	35.12	95.84	**20.37**	**96.95**	**15.13**	97.48	13.43	66.84	88.00	99.18	1.93	97.54	13.03	92.96	24.65
DICE	83.56	70.06	74.46	68.26	73.11	85.90	72.68	86.98	**99.69**	**1.12**	76.38	81.82	96.92	16.39	**99.02**	**4.50**	84.48	51.88
DICE+	96.50	16.02	89.45	47.22	94.62	29.74	95.16	27.63	99.19	3.46	75.71	80.83	99.07	3.93	98.75	5.99	93.56	26.85
ASH-B	89.78	50.45	85.35	55.89	88.26	52.53	88.37	51.73	98.80	6.43	72.78	85.06	97.69	13.74	98.18	9.97	89.90	40.72
ASH-B+	97.12	13.99	**92.32**	**35.00**	94.53	28.49	95.23	25.52	98.84	5.41	68.21	90.70	**99.65**	**0.08**	98.41	7.63	93.04	25.85

Table 3. Detailed performance of OOD detection methods with and without ATS for DenseNet [8] trained on CIFAR-100 [9].

Method	OOD-Dataset												Average	
	iNaturalist		SUN		Places		Textures		NINCO		fMNIST			
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
AT-only	80.85	65.40	77.62	71.87	70.75	80.27	89.60	42.13	65.56	81.17	94.93	42.70	79.89	63.92
MSP	88.42	52.71	81.75	68.56	80.63	71.60	80.46	66.15	79.97	75.96	86.46	64.15	82.95	66.52
MSP+	92.66	35.99	88.87	45.53	84.85	57.46	94.00	25.73	80.45	67.11	96.56	24.58	89.57	42.73
ODIN	91.38	41.56	86.89	54.02	84.44	62.15	87.57	45.55	77.70	75.13	96.76	19.51	87.46	49.65
ODIN+	91.78	36.63	88.30	48.24	84.01	59.96	93.42	26.42	77.11	69.16	**98.77**	**4.22**	88.90	40.77
MLS	91.13	50.87	86.59	59.87	84.18	65.64	86.40	54.29	80.40	76.62	86.47	69.84	85.86	62.86
MLS+	92.65	36.01	88.87	45.54	84.85	57.47	94.00	25.73	80.44	67.17	96.56	24.60	89.56	42.75
Energy	90.59	53.96	86.73	58.25	84.12	65.40	86.73	52.30	79.69	77.63	85.32	73.57	85.53	63.52
Energy+	79.49	65.77	81.92	62.36	75.29	73.29	90.71	35.53	65.87	79.72	86.39	64.47	79.94	63.52
ReAct	96.39	19.55	94.41	24.01	91.93	**33.45**	90.45	45.85	80.13	71.50	88.70	68.16	90.33	43.75
ReAct+	93.83	32.14	90.60	40.89	86.62	52.99	94.91	24.77	79.91	68.73	97.84	9.25	90.62	38.13
DICE	94.49	26.64	90.99	36.08	87.73	47.65	90.46	32.46	77.49	74.12	84.17	62.17	87.55	46.52
DICE+	91.97	34.71	88.39	43.25	83.43	56.01	94.25	24.89	75.85	68.82	96.42	28.49	88.38	42.69
ReAct+DICE	96.04	20.17	93.77	26.61	90.53	38.53	92.42	29.96	73.84	74.69	85.91	64.62	88.75	42.43
ReAct+DICE+	92.06	34.23	88.88	42.20	83.86	54.38	94.13	26.76	74.12	70.86	96.60	23.33	88.28	41.96
ASH-B	**97.32**	**14.21**	**95.10**	**22.11**	**92.31**	**33.45**	95.50	21.13	**82.32**	69.39	93.56	43.99	**92.69**	34.05
ASH-B+	95.19	24.07	92.37	32.70	88.33	45.63	**96.29**	**18.71**	79.93	**66.14**	97.51	13.33	91.61	**33.43**

Table 4. Detailed performance of OOD detection methods with and without ATS for ResNet50 [5] trained on ImageNet [3].

[18] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R. Kulkarni, and Jianxiong Xiao. TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[19] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *arXiv preprint arXiv:1506.03365*, 2016.

[20] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million Image Database for Scene Recognition. *TPAMI*, 2017.

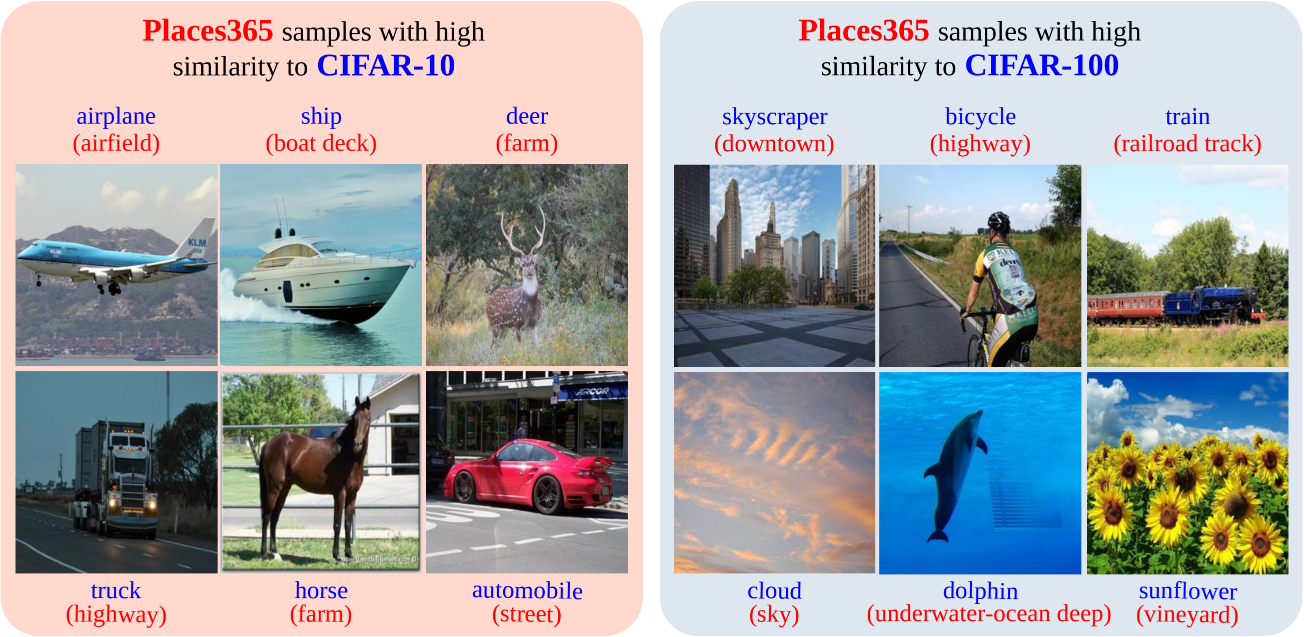

Figure 1. Examples showing potential contamination when using Places365 [20] for OOD detection with a model trained on CIFAR [9] as ID data. Annotations in blue indicate the corresponding CIFAR [9] classes present in the image, while those in red denote the original class from the Places365 [20] dataset. **Left figure:** Samples from classes of the Places365 [20] dataset that overlap with CIFAR-10 [9] classes. **Right figure:** Samples from classes of the Places365 [20] dataset that overlap with CIFAR-100 [9] classes.

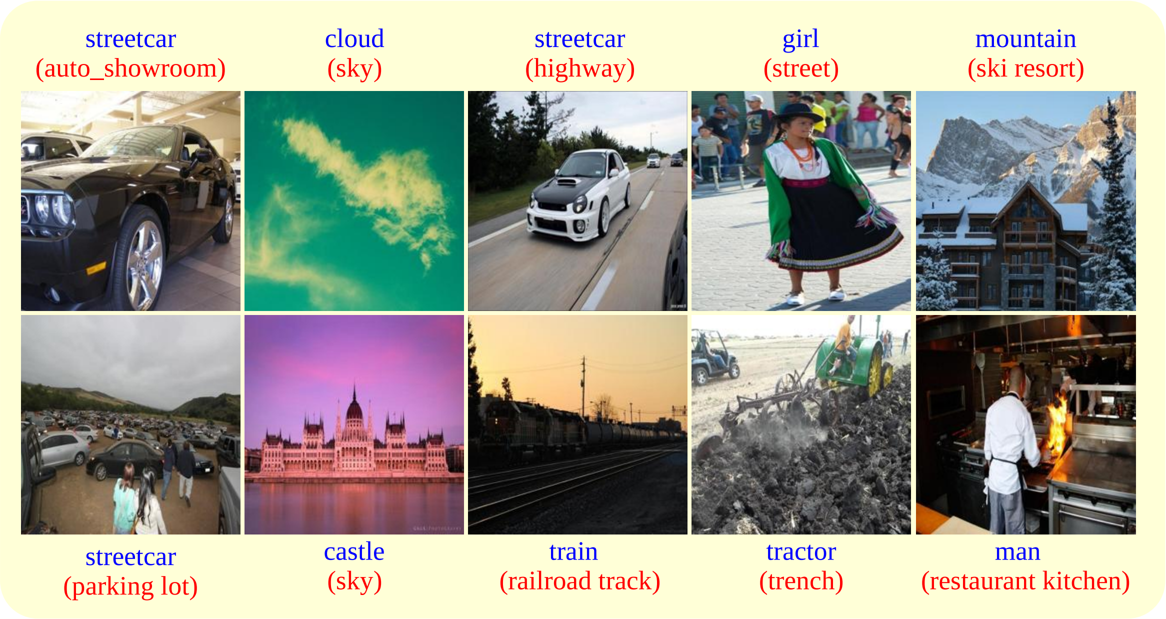

Figure 2. Analysis of Places365 [20] OOD detection using Maximum Softmax Probability (MSP) [7] with and without Adaptive Temperature Scaling (ATS) on ResNet18 [5] trained with CIFAR-100 [9]. These exemplary Places365 images were initially categorized as OOD by MSP, but then reclassified as ID after applying ATS. We set as threshold the score value with 95% recall. Annotations in blue denote the corresponding CIFAR-100 [9] classes present in the image, while those in red denote the original class from the Places365 [20] dataset.

Algorithm 1 Calibration: Offline eCDF computation

```
Input: $\mathcal{D}_{\text{in}} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$    ▷ In-distribution training dataset
Input: $\mathcal{L}$    ▷ Selected intermediate layers for adjustment
Input: $f$    ▷ Trained neural network
Output: $\hat{F}_{l,\mathcal{D}_{\text{in}}}, \forall l \in \mathcal{L}$    ▷ eCDF per intermediate layer

// Compute mean layer activation for all train samples
for $\mathbf{x}_i \in \mathcal{D}_{\text{in}}$ do
    for $l \in \mathcal{L}$ do
        // Feature map of the $l$-th intermediate layer
        $\mathbf{z}_{l,i} \leftarrow f(\mathbf{x}_i)$

        // Mean layer activation
        $\mu_{l,i}(\mathbf{x}_i) \leftarrow \frac{1}{C_l H_l W_l} \sum_c^{C_l} \sum_h^{H_l} \sum_w^{W_l} \max(\mathbf{z}_{l,i}(c,h,w),0),$
    end for
end for

// Pre-Compute eCDF for intermediate layers
for $l \in \mathcal{L}$ do
    $\hat{F}_{l,\mathcal{D}_{\text{in}}} \leftarrow$ Pre-compute eCDF from $\mu_{l,i}, i \in [1, N]$
end for

return $\hat{F}_{l,\mathcal{D}_{\text{in}}}, \forall l \in \mathcal{L}$
```

Algorithm 2 Inference: Test-time OOD score computation

```
Input: $\mathbf{x}$    ▷ Test sample
Input: $f$, $G$    ▷ Trained NN, and logit-based OOD scoring function
Input: $\hat{F}_{l,\mathcal{D}_{\text{in}}}, \forall l \in \mathcal{L}$    ▷ eCDF per selected intermediate layer
Output: $s(\mathbf{x})$    ▷ Sample specific OOD score

// Calculate layer-specific $p$-values
for $l \in \mathcal{L}$ do
    // Feature map of the $l$-th intermediate layer
    $\mathbf{z}_l \leftarrow f(\mathbf{x})$

    // Mean layer activation
    $\mu_l(\mathbf{x}) \leftarrow \frac{1}{C_l H_l W_l} \sum_c^{C_l} \sum_h^{H_l} \sum_w^{W_l} \max(\mathbf{z}_l(c,h,w),0),$

    // $p$-value per-intermediate layer using two-sided test
    $p_l(\mathbf{x}) \leftarrow 2 \min(\hat{F}_{l,\mathcal{D}_{\text{in}}}(\mu_l(\mathbf{x})), 1 - \hat{F}_{l,\mathcal{D}_{\text{in}}}(\mu_l(\mathbf{x})))$
end for

// Derive sample specific temperature by aggregating the layer-specific
    $p$-values via Fisher's method
$\hat{T}(\mathbf{x}) \leftarrow -2 \sum_{l \in \mathcal{L}} \log(p_l(\mathbf{x}))$

// Scale logits with sample-specific temperature value and derive OOD
    score with given scoring function
$s(\mathbf{x}) = G(f(\mathbf{x}) \,/\, \hat{T}(\mathbf{x}))$

return $s(\mathbf{x})$
```

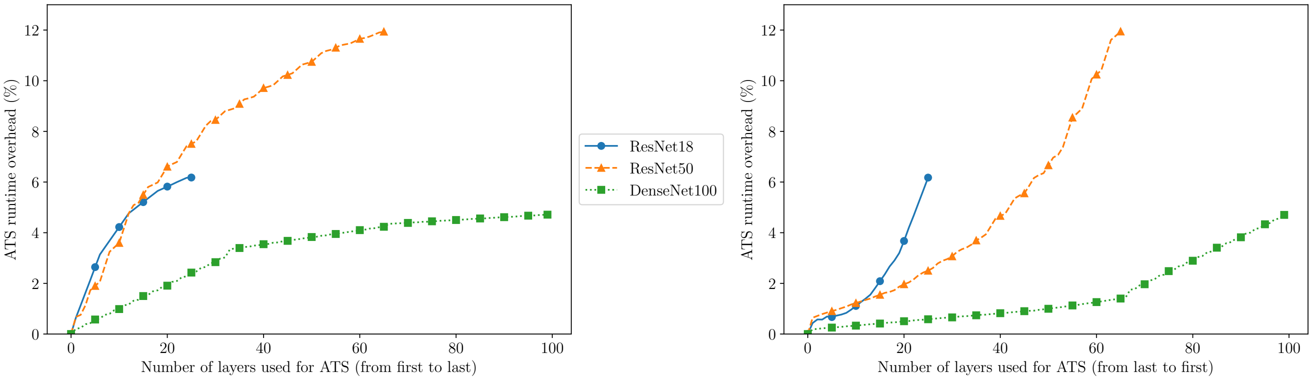

Figure 3. Evaluation of the runtime overhead across three model architectures (ResNet18, Renset50, and DenseNet100) illustrating the impact of incorporating varying numbers of intermediate layers. **Left figure:** Overhead when the first l layers are employed for adaptive temperature scaling. **Right figure:** Overhead when the last l layers are employed for adaptive temperature scaling.