Self-Distribution Distillation: Efficient Uncertainty Estimation (Supplementary material)

Yassir Fathullah[1] **Mark J. F. Gales**[1]

[1]Engineering Department, University of Cambridge, UK

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

A EXPERIMENTAL CONFIGURATION

Table 1: Description of datasets used in training and evaluating models.

Dataset	Train	Test	Classes
CIFAR-100	50000	10000	100
LSUN	-	10000	10
SVHN	-	26032	10
Tiny ImageNet	-	10000	200

All models were trained on the CIFAR-100 dataset, with and without data augmentation. The augmentation scheme involves randomly mirroring and shifting images following He et al. [2016], Huang et al. [2016]. Remaining datasets were used as out-of-distribution samples in the detection task.

All individual models, and ensemble members were based of off the DenseNet-BC ($k = 12$, 100 layers) architecture and trained according to Huang et al. [2017]. SWAG-Diag was obtained by checkpointing the weights of the last 20 epochs with a reduced learning rate of $\eta = 1.0 \times 10^{-4}$. MIMO with two output heads was trained using the same setup as for the standard model. To keep training costs comparable to (S2D) individual models, no batch or input repetition was used [Havasi et al., 2021]. Similarly all self-distribution distilled equivalents were trained with identical training recipes with the addition of a student loss ($\mu = 1.28 \times 10^{-4}$).

Regarding distilled based models, the EnD baseline was trained using negative log-likelihood using the average temperature scaled prediction of the teacher ensemble, with $T \in \{1.0, 2.0, 3.0, 4.0, 5.0\}$. For the hierarchical distribution distillation approaches the students were first initialised with the weights of an S2D model trained for 150 epochs, for increased stability. Thereafter, each student was trained using the appropriate H2D criteria with a significantly reduced learning rate. H2D-Dir was trained using $\eta = 5.0 \times 10^{-5}$ for an additional 150 epochs. H2D-Gauss required an initial learning rate of $\eta = 5.0 \times 10^{-3}$ which was reduced by a factor of 2 after 75 and 150 epochs. It was trained for 170 epochs. Additionally, uncertainties were computed by generating 50 samples from each Gaussian prediction, since this modelling choice does not result in closed form expressions.

A.1 PROXY TARGET TRAINING

Since the use of negative log-likelihood can be unstable in training S2D and distilling H2D models we utilise proxy targets and KL-divergence. It has already been mentioned that the proxy target in S2D follows:

$$\tilde{\boldsymbol{\alpha}} = \arg\max_{\hat{\boldsymbol{\alpha}}} \sum_m \ln \texttt{Dir}(\boldsymbol{\pi}^{(m)}; \hat{\boldsymbol{\alpha}}), \quad \boldsymbol{\pi}^{(m)} = \texttt{Softmax}(\boldsymbol{z}^{(m)}, T) \tag{1}$$

Each categorical prediction will be temperature scaled, with $T = 1.5$, to mitigate overconfident predictions. While H2D-Dir does not require any proxy targets, the Gaussian equivalent does. The proxy diagonal Gaussian, estimated according to maximum log-likelihood, has a closed-form expression:

$$\tilde{\boldsymbol{\mu}} = \frac{1}{M}\sum_{m=1}^{M} \ln \boldsymbol{\alpha}^{(m)}, \quad \tilde{\boldsymbol{\sigma}}^2 = \frac{1}{M}\sum_{m=1}^{M} (\ln \boldsymbol{\alpha}^{(m)} - \tilde{\boldsymbol{\mu}})^2 \tag{2}$$

where $\boldsymbol{v}^2 = \boldsymbol{v} \odot \boldsymbol{v}$ represents an element-wise multiplication. This is then used in a KL-divergence based loss, training the student with prediction $\boldsymbol{\mu}, \boldsymbol{\sigma}$ according to:

$$\texttt{KL}\Big(\mathcal{N}(\boldsymbol{z}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}^2) \,\Big\|\, \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2)\Big) = \sum_{c=1}^{K} \ln\left(\frac{\sigma_c}{\tilde{\sigma}_c}\right) + \frac{\tilde{\sigma}_c^2 + (\mu_c - \tilde{\mu}_c)^2}{2\sigma_c^2} - \frac{1}{2} \tag{3}$$

Note however, that the proxy targets are detached from any back gradient propagation calculations. This is to simulate typical teacher-student knowledge transfer where teacher weights are kept fixed during student training.

B OUT-OF-DISTRIBUTION DETECTION

This section covers remaining out-of-distribution detection experiments. First, we cover the LSUN and Tiny ImageNet detection problem for all models considered in section 5.2. Thereafter, additional experiments will be run on ensembles of various sizes. This is to investigate if the low quality of knowledge uncertainty estimates is caused by a limited number of ensemble members.

B.1 TINY IMAGENET EXPERIMENTS

Similar to the results section 5.2 the S2D Deep ensemble and H2D-Gauss outperformed all other models, see Table 3 and 4. The only exception is the use of confidence on resized TIM with the AUROC metric where the Deep ensemble marginally outperforms the S2D equivalent. However, unlike previous results, knowledge uncertainty seems to perform on par with or outperform confidence. The one exception is the MC ensemble.

Table 2: OOD detection results (LSUN random crop) trained on C100. **Best** in column and **best** overall.

Model	OOD %AUROC				OOD %AUPR			
	Conf.	TU	DU	KU	Conf.	TU	DU	KU
Individual	83.2 ±2.1	85.7 ±4.5			79.4 ±5.6	83.0 ±5.9		
S2D Individual	85.4 ±4.5	88.9 ±4.1	90.3 ±4.0	84.1 ±4.8	81.9 ±6.2	86.6 ±5.7	90.3 ±5.0	76.0 ±5.1
MIMO	83.3 ±3.9	86.2 ±4.2	86.3 ±4.3	80.9 ±1.6	79.6 ±6.6	83.8 ±6.8	83.8 ±6.9	72.4 ±3.7
S2D MIMO	85.8 ±2.5	89.5 ±2.8	90.7 ±2.8	85.5 ±2.8	78.0 ±3.4	84.8 ±3.5	89.4 ±3.4	75.2 ±3.3
SWAG-Diag	84.3 ±2.8	87.1 ±3.1	87.1 ±3.1	80.8 ±7.2	80.8 ±4.0	84.5 ±3.8	84.6 ±3.8	73.4 ±14.2
S2D SWAG-Diag	85.6 ±2.7	89.1 ±2.5	90.4 ±2.5	85.3 ±3.0	81.8 ±4.0	86.5 ±3.6	90.2 ±3.4	76.4 ±3.6
MC ensemble	81.0 ±3.5	84.4 ±4.0	86.4 ±3.8	63.0 ±4.0	77.0 ±3.6	81.7 ±4.0	84.9 ±4.0	53.1 ±3.1
S2D MC ensemble	83.3 ±2.3	86.9 ±3.2	90.0 ±2.5	77.7 ±5.3	79.3 ±3.2	83.8 ±4.3	90.1 ±3.2	69.8 ±4.8
Deep ensemble	85.9	89.1	90.9	80.4	82.0	86.3	89.1	72.5
S2D Deep ensemble	86.8	90.5	93.7	81.5	**83.0**	87.9	93.9	73.4
EnD	84.7	87.4			81.1	84.9		
H2D-Dir	85.3	88.9	88.8	**91.7**	82.5	87.4	87.6	**87.1**
H2D-Gauss	**86.9**	**90.6**	**95.1**	76.0	82.9	**88.0**	**95.7**	67.0

Table 3: OOD detection results (TIM resize) trained on C100. **Best** in column and **best** overall.

Individual	77.6 ±0.7	79.5 ±0.7			74.2 ±0.7	77.1 ±0.9		
S2D Individual	78.0 ±0.8	80.1 ±0.7	79.6 ±0.8	78.1 ±0.4	75.3 ±0.9	77.7 ±0.9	76.6 ±1.2	76.3 ±0.5
MIMO	78.1 ±0.4	79.9 ±0.7	79.9 ±0.8	76.3 ±1.5	74.6 ±1.0	77.3 ±1.3	77.4 ±1.3	69.6 ±2.0
S2D MIMO	80.1 ±1.2	80.7 ±1.2	80.7 ±1.2	80.4 ±1.2	77.3 ±1.6	77.8 ±1.6	77.7 ±1.5	77.5 ±1.6
SWAG-Diag	77.7 ±0.7	79.6 ±0.6	79.6 ±0.6	76.4 ±0.7	74.2 ±0.8	77.0 ±0.8	77.1 ±0.8	70.0 ±0.7
S2D SWAG-Diag	78.6 ±0.7	80.5 ±0.6	80.1 ±0.7	79.2 ±0.5	75.6 ±0.9	78.1 ±1.1	77.1 ±1.0	76.5 ±0.9
MC ensemble	78.5 ±0.5	80.6 ±0.3	80.8 ±0.4	76.6 ±0.6	75.2 ±0.5	78.1 ±0.6	78.4 ±0.5	70.9 ±1.1
S2D MC ensemble	79.3 ±0.5	81.1 ±0.5	81.1 ±0.5	80.4 ±0.6	76.4 ±0.7	78.5 ±0.8	78.1 ±1.0	77.1 ±0.7
Deep ensemble	**81.7**	83.6	83.5	81.0	78.9	81.6	81.5	76.6
S2D Deep Ensemble	81.5	**84.2**	82.8	82.8	**79.1**	**82.0**	79.9	80.0
EnD	78.7	80.4			75.4	78.0		
H2D-Dir	77.3	79.8	79.6	81.6	74.5	77.9	77.7	79.2
H2D-Gauss	80.5	82.6	**83.7**	**82.8**	78.8	81.4	**82.5**	**80.1**

Table 4: OOD detection results (TIM random crop) trained on C100. **Best** in column and **best** overall.

Individual	76.7 ±4.1	79.2 ±4.2			74.7 ±3.6	78.5 ±3.8		
S2D Individual	80.2 ±5.9	85.4 ±6.2	84.5 ±5.9	86.4 ±6.3	79.3 ±6.3	83.3 ±6.7	81.9 ±6.6	83.1 ±6.7
MIMO	79.4 ±4.8	81.9 ±5.3	81.9 ±5.3	79.8 ±4.6	77.1 ±4.8	80.9 ±5.2	80.8 ±5.3	74.9 ±8.1
S2D MIMO	80.3 ±8.6	86.5 ±8.5	86.5 ±8.5	86.9 ±8.6	80.0 ±6.5	82.9 ±6.4	83.0 ±6.4	84.9 ±6.5
SWAG-Diag	78.4 ±3.5	80.9 ±3.7	80.9 ±4.0	78.6 ±2.0	76.0 ±3.3	79.8 ±3.4	79.7 ±3.7	73.7 ±3.5
S2D SWAG-Diag	80.5 ±6.0	84.8 ±6.5	83.8 ±6.3	86.6 ±6.6	79.4 ±5.5	83.4 ±6.1	81.8 ±6.2	83.4 ±6.0
MC ensemble	75.8 ±4.5	78.8 ±4.8	79.7 ±4.9	69.3 ±3.7	74.3 ±4.0	78.5 ±4.3	80.0 ±4.3	60.8 ±3.7
S2D MC ensemble	78.8 ±6.3	82.1 ±6.4	82.6 ±6.5	82.0 ±6.1	77.1 ±5.2	81.1 ±5.1	81.8 ±5.1	79.8 ±4.9
Deep ensemble	80.9	84.2	83.5	82.3	79.3	83.9	83.2	79.8
S2D Deep ensemble	**84.8**	**88.5**	86.4	**89.7**	**82.8**	**87.3**	84.4	**87.7**
EnD	72.7	74.8			71.4	75.0		
H2D-Dir	74.7	78.2	77.9	84.2	73.2	77.7	77.5	81.7
H2D-Gauss	83.2	88.0	**88.0**	88.5	81.0	86.0	**87.2**	84.1

B.2 ENSEMBLE SIZE EXPERIMENTS

Knowledge uncertainty was found to have underwhelming performance (especially for MC and Deep ensembles) and did not show similar trends to prior work [Malinin and Gales, 2018, 2021, Malinin et al., 2020]. To possibly mitigate this, the ensemble size was increased as a smaller number of models could lead to inaccurate measures of diversity and knowledge uncertainty. Results are compiled in Tables 5-10.

Performance on the CIFAR-100 test set is shown in Table 5. Increasing the ensemble size leads to improved accuracy and lower negative log-likelihoods as would be expected. The MC ensemble also becomes better calibrated. The Deep ensemble on the other hand has increasing calibration error with the number of members. This is due to the ensemble prediction becoming under-confident when averaging over a large number of members.

Out-of-distribution detection performance on LSUN, SVHN and TIM are compiled in Tables 6-10. Although the MC ensemble enjoys improved accuracy when increased in size, it seems to remain relatively unaffected in terms of OOD detection using any uncertainty metric. In detecting LSUN using random crops, the performance of KU interestingly deteriorates notably. Overall this points to MC ensembles' lacking ability in utilising new information from additional ensemble member draws/samples for better uncertainty estimation. Regarding the Deep ensemble, it generally improves with increasing size with any metric, however with diminishing returns. In this case all uncertainties improve with ensemble size, not only knowledge uncertainty. Therefore it seems that the cause for confidence, total and data outperforming knowledge uncertainty is not due to the ensemble size being limited to five members.

Table 5: Test performance of various ensembles and sizes (± 2 std). All models are trained on C100.

Ensemble Type	Ensemble Size (M)	Acc.	NLL	%ECE
MC	5	75.6 ±0.9	0.94 ±0.04	6.67 ±1.18
	10	75.8 ±0.9	0.92 ±0.04	6.11 ±1.11
	20	76.0 ±1.0	0.91 ±0.04	5.81 ±1.12
Deep	5	79.3	0.76	1.44
	10	80.1	0.71	1.91
	20	80.3	0.68	2.19

Table 6: OOD detection results (LSUN resize) trained on C100.

Type	M	OOD %AUROC Conf.	TU	DU	KU	OOD %AUPR Conf.	TU	DU	KU
MC	5	76.6 $\pm$0.8	78.3 $\pm$0.8	78.9 $\pm$0.8	72.4 $\pm$1.2	72.2 $\pm$1.0	74.6 $\pm$1.6	75.6 $\pm$1.7	64.2 $\pm$2.0
	10	76.7 $\pm$0.6	78.3 $\pm$0.8	79.1 $\pm$0.9	72.6 $\pm$1.2	72.3 $\pm$1.1	74.6 $\pm$1.6	75.9 $\pm$1.7	64.3 $\pm$2.0
	20	76.8 $\pm$0.7	78.4 $\pm$0.8	79.2 $\pm$0.8	72.7 $\pm$1.3	72.4 $\pm$1.2	74.6 $\pm$1.6	76.0 $\pm$1.7	64.3 $\pm$2.3
Deep	5	81.1	82.9	83.4	79.2	77.7	80.4	81.2	73.6
	10	82.0	83.9	84.8	80.3	79.1	81.8	83.4	74.9
	20	82.2	84.0	85.1	80.9	79.4	81.8	83.6	75.7

Table 7: OOD detection results (LSUN random crop) trained on C100.

Type	M	Conf.	TU	DU	KU	Conf.	TU	DU	KU
MC	5	81.0 $\pm$3.5	84.4 $\pm$4.0	86.4 $\pm$3.8	63.0 $\pm$4.0	77.0 $\pm$3.6	81.7 $\pm$4.0	84.9 $\pm$4.0	53.1 $\pm$3.1
	10	81.0 $\pm$3.5	84.4 $\pm$3.9	86.7 $\pm$3.7	61.6 $\pm$3.9	77.0 $\pm$3.7	81.8 $\pm$4.0	85.4 $\pm$4.0	52.2 $\pm$3.0
	20	80.8 $\pm$3.7	84.1 $\pm$4.1	86.6 $\pm$3.9	60.9 $\pm$4.0	76.7 $\pm$3.9	81.3 $\pm$4.2	85.3 $\pm$4.2	51.7 $\pm$3.0
Deep	5	85.9	89.1	90.9	80.4	82.0	86.3	89.1	72.5
	10	85.7	89.3	91.3	81.3	81.8	86.4	89.9	73.1
	20	86.2	89.8	92.2	82.0	82.1	86.8	91.0	73.1

Table 8: OOD detection results (SVHN) trained on C100.

Type	M	Conf.	TU	DU	KU	Conf.	TU	DU	KU
MC	5	79.0 $\pm$4.3	81.6 $\pm$4.7	83.1 $\pm$4.6	68.3 $\pm$3.0	88.1 $\pm$2.8	89.3 $\pm$3.3	90.7 $\pm$3.1	77.4 $\pm$1.8
	10	78.9 $\pm$4.4	81.5 $\pm$4.7	83.3 $\pm$4.7	67.5 $\pm$3.1	88.0 $\pm$2.7	89.3 $\pm$3.3	90.9 $\pm$3.1	76.6 $\pm$2.0
	20	78.9 $\pm$4.4	81.5 $\pm$4.7	83.3 $\pm$4.7	67.1 $\pm$3.3	88.1 $\pm$2.7	89.2 $\pm$3.3	90.9 $\pm$3.1	76.3 $\pm$2.0
Deep	5	84.5	87.2	86.8	85.0	91.3	92.5	92.2	91.5
	10	84.1	87.0	87.5	83.9	91.2	92.4	93.1	90.3
	20	83.7	86.6	87.2	84.1	91.0	92.2	92.9	90.6

Table 9: OOD detection results (TIM resize) trained on C100.

Type	M	Conf.	TU	DU	KU	Conf.	TU	DU	KU
MC	5	78.5 $\pm$0.5	80.6 $\pm$0.3	80.8 $\pm$0.4	76.6 $\pm$0.6	75.2 $\pm$0.5	78.1 $\pm$0.6	78.4 $\pm$0.5	70.9 $\pm$1.1
	10	78.7 $\pm$0.6	80.8 $\pm$0.4	81.0 $\pm$0.5	77.4 $\pm$0.7	75.4 $\pm$0.6	78.4 $\pm$0.6	78.7 $\pm$0.5	72.2 $\pm$1.1
	20	78.8 $\pm$0.5	80.9 $\pm$0.4	81.2 $\pm$0.4	77.9 $\pm$0.7	75.6 $\pm$0.5	78.4 $\pm$0.4	78.8 $\pm$0.4	72.9 $\pm$1.4
Deep	5	81.7	83.6	83.5	81.0	78.9	81.6	81.5	76.6
	10	82.3	84.1	84.2	82.4	79.8	82.2	82.4	78.7
	20	82.6	84.4	84.5	83.0	80.1	82.4	82.8	79.6

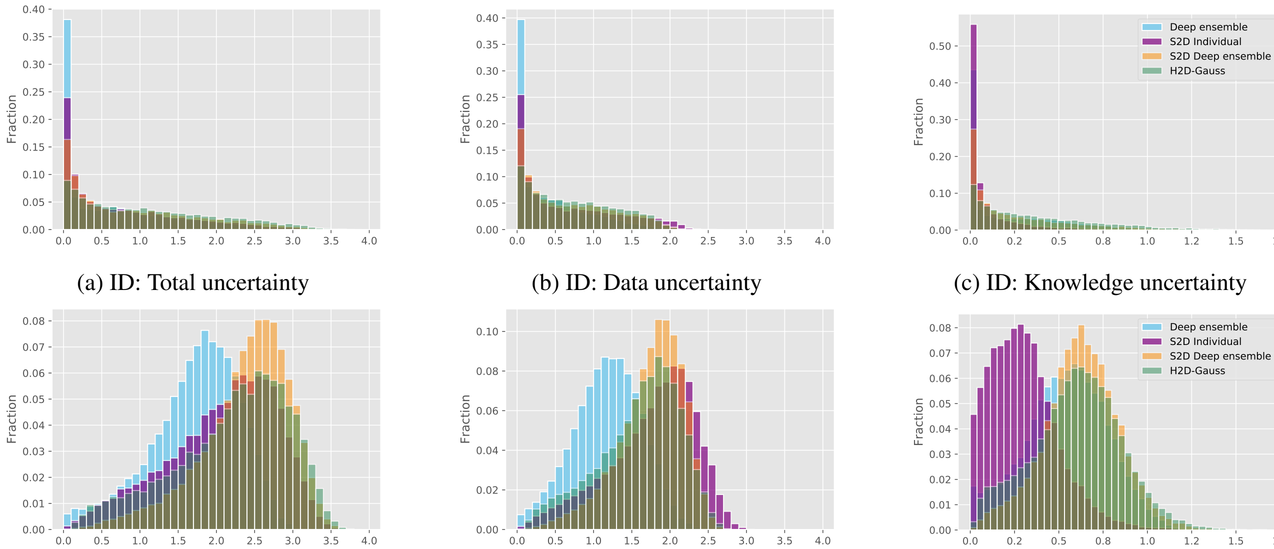
Table 10: OOD detection results (TIM random crop) trained on C100.

Type	M	Conf.	TU	DU	KU	Conf.	TU	DU	KU
MC	5	75.8 $\pm$4.5	78.8 $\pm$4.8	79.7 $\pm$4.9	69.3 $\pm$3.7	74.3 $\pm$4.0	78.5 $\pm$4.5	80.0 $\pm$4.3	60.8 $\pm$3.7
	10	75.7 $\pm$4.8	78.7 $\pm$5.1	79.7 $\pm$5.2	69.1 $\pm$3.9	74.2 $\pm$4.2	78.5 $\pm$4.5	80.2 $\pm$4.5	60.7 $\pm$3.8
	20	75.7 $\pm$4.7	78.6 $\pm$5.0	79.7 $\pm$5.2	69.0 $\pm$4.1	74.3 $\pm$4.1	78.4 $\pm$4.4	80.3 $\pm$4.3	60.6 $\pm$4.4
Deep	5	80.9	84.2	83.5	82.3	79.3	83.9	83.2	79.8
	10	82.8	86.5	85.7	85.5	81.0	85.8	85.0	83.7
	20	83.4	87.1	86.1	86.8	81.6	86.4	85.4	85.4

C BEHAVIOUR OF UNCERTAINTIES

This section investigates how the uncertainties produced from a vanilla Deep ensemble differ from self-distribution distilled derived systems, and how well hierarchical distribution distillation captures the behaviour of its teacher. The comparison will be made between the in-domain CIFAR-100 and, out of simplicity, only the out-of-domain SVHN test set.

Figure 1 shows the contrast of various uncertainties between an CIFAR-100 (ID) and SVHN (OOD) test sets. Clearly, the S2D



(a) ID: Total uncertainty (b) ID: Data uncertainty (c) ID: Knowledge uncertainty

(d) OOD: Total uncertainty (e) OOD: Data uncertainty (f) OOD: Knowledge uncertainty

Figure 1: Histograms of various uncertainties produced by Deep ensemble, S2D, S2D Deep ensemble and H2D-Gauss systems. Out-of-distribution data was generated from the SVHN test set.

systems output ID uncertainties in a consistent manner, even matching the conceptually different Deep ensemble. Observe that S2D integrates temperature scaling (smoothing predictions) into the training of models; total and data uncertainties[1] estimated by these models will naturally have larger entropy than Deep ensembles. While it is expected that the Deep ensemble would have different behaviour on the SVHN OOD set, it is surprising to observe how well H2D-Gauss aligns with its S2D Deep ensemble teacher. An individual S2D model was also able to generate closely related total and data uncertainty estimates, but suffers significantly in producing consistent knowledge uncertainties. These results raise the question if a Gaussian student could capture the diversity in a vanilla Deep ensemble by modelling the logits, in a similar fashion to how H2D-Gauss models its teacher—a possible avenue for future work.

D ADDITIONAL EXPERIMENTS: WIDERESNET

Following the DenseNet-BC experiments in section 5 we repeated them with a different architecture. In this section we focus on a significantly larger WideResNet [Zagoruyko and Komodakis, 2016] model with a depth of 28 and a widening factor of 10. The standard and S2D models were both trained as described in Zagoruyko and Komodakis [2016], with the S2D specific parameters being the same as previously described. The only difference is that teacher predictions were generated using multiplicative Gaussian noise with a fixed standard deviation of 0.10.

The H2D-Gauss model was also trained in a different manner. First, it was initialised from an S2D model trained for 150 epochs. Thereafter it was trained for an additional 80 epochs with a starting learning rate of $\eta = 2 \times 10^{-3}$ which was reduced by a factor of 4 after 60 epochs. For this section, EnD and H2D-Dir were not investigated.

Table 11 shows test set performance. Unlike previous experiments, S2D was not able to outperform an individual model by more than two standard deviations, in this case achieving around one standard deviation improvement in accuracy. Interestingly, the MC approach has worse accuracy for both the standard and S2D case, however this could be due to the small number of drawn samples ($M = 5$). Furthermore, both Deep ensembles significantly outperform their individual equivalents with the S2D version being slightly better in all measured performance metrics. The notable result in this table is the high performance of H2D-Gauss, able to outperform the Deep ensemble in C100 and achieve near ensemble performance in C100+.

In the OOD detection task we observe that both versions of the MC ensemble struggle to outperform their individual counterparts. There also seems to be a disparity in performance when comparing resize and random cropped LSUN and

[1]Knowledge uncertainty does not necessarily increase with temperature.

Table 11: Test performance (± 2 std).

Dataset	C100			C100+		
Model	Acc.	NLL	%ECE	Acc.	NLL	%ECE
Individual	73.9 ±0.5	1.05 ±0.02	5.26 ±0.78	81.1 ±0.3	0.76 ±0.01	5.21 ±0.44
S2D Individual	74.2 ±0.5	1.06 ±0.05	5.48 ±2.25	81.3 ±0.3	0.74 ±0.01	4.24 ±0.74
MC ensemble	73.6 ±0.5	1.05 ±0.03	4.70 ±0.88	81.0 ±0.5	0.74 ±0.01	3.29 ±0.36
S2D MC ensemble	73.8 ±0.4	1.03 ±0.04	2.95 ±1.01	81.0 ±0.3	0.73 ±0.01	1.99 ±0.35
Deep ensemble	77.1	0.88	5.08	83.4	0.63	2.27
S2D Deep ensemble	77.9	0.86	4.52	83.6	0.63	1.84
H2D-Gauss	77.4	0.95	5.19	82.8	0.71	2.45

TIM. With random crops, all S2D systems notably outperform their standard counterparts. In this case both S2D Individual and H2D-Gauss were able to outperform the Deep ensemble using any uncertainty metric. In the other case of resizing LSUN and TIM images and in SVHN the detection performance difference is smaller but the S2D Deep ensemble still remains the best model with both H2D-Gauss and Deep ensemble performing similarly.

Table 12: LSUN (resize) OOD detection results. **Best** in column and **best** overall.

Model	OOD %AUROC				OOD %AUPR			
	Conf.	TU	DU	KU	Conf.	TU	DU	KU
Individual	76.3 ±0.5	76.7 ±0.6			70.7 ±0.8	71.1 ±0.9		
S2D Individual	76.0 ±1.1	76.5 ±1.5	76.7 ±1.4	75.7 ±1.6	71.4 ±1.8	72.0 ±2.7	72.8 ±3.7	69.7 ±2.0
MC ensemble	75.8 ±0.6	76.2 ±0.7	76.4 ±0.7	65.2 ±1.7	70.3 ±1.0	70.5 ±1.1	70.8 ±1.2	56.2 ±1.5
S2D MC ensemble	75.7 ±1.0	76.4 ±1.7	77.0 ±1.6	75.2 ±2.1	71.0 ±1.6	71.6 ±2.7	73.1 ±3.8	69.6 ±2.6
Deep ensemble	77.6	78.0	78.4	68.0	72.3	72.6	73.1	58.8
S2D Deep ensemble	**77.7**	**78.5**	**79.3**	76.8	**73.2**	**74.1**	**75.9**	71.3
H2D-Gauss	77.1	77.2	77.8	**77.5**	72.0	71.8	71.9	**72.3**

Table 13: LSUN (random crop) OOD detection results. **Best** in column and **best** overall.

Individual	72.4 ±5.0	73.9 ±5.4			67.0 ±2.9	68.7 ±3.1		
S2D Individual	75.8 ±3.4	77.6 ±4.3	77.9 ±4.7	**76.5** ±4.6	70.5 ±3.9	72.6 ±4.9	74.4 ±4.7	**71.4** ±5.5
MC ensemble	68.9 ±5.6	70.3 ±6.0	70.9 ±6.2	50.8 ±3.7	64.0 ±3.0	65.2 ±3.5	66.1 ±3.6	45.7 ±1.5
S2D MC ensemble	72.7 ±3.2	74.5 ±4.1	75.9 ±4.3	72.0 ±4.4	67.7 ±3.3	69.7 ±4.6	73.4 ±4.4	65.7 ±5.0
Deep ensemble	72.1	74.2	75.2	60.6	67.2	69.2	70.5	51.6
S2D Deep ensemble	75.5	**78.4**	**80.0**	75.4	**70.7**	**73.9**	**77.2**	69.0
H2D-Gauss	**76.0**	77.6	77.8	76.4	69.6	71.5	74.1	70.9

Table 14: SVHN OOD detection results. **Best** in column and **best** overall.

Individual	80.1 ±4.6	81.6 ±4.4			88.3 ±2.4	89.0 ±2.3		
S2D Individual	80.1 ±4.4	81.6 ±4.4	81.9 ±4.8	81.4 ±5.4	88.6 ±2.3	89.2 ±2.5	90.1 ±2.5	87.8 ±4.1
MC ensemble	77.6 ±4.9	79.1 ±4.5	79.7 ±4.5	56.6 ±2.5	86.9 ±2.3	87.5 ±2.2	88.0 ±2.2	70.2 ±1.2
S2D MC ensemble	77.3 ±4.7	79.0 ±4.8	80.1 ±4.6	77.3 ±5.6	87.1 ±2.5	87.7 ±2.7	89.6 ±2.5	85.7 ±3.9
Deep ensemble	**81.5**	83.4	84.0	68.3	89.2	89.9	90.4	77.9
S2D Deep ensemble	81.5	**83.7**	**84.6**	**81.8**	**89.6**	**90.5**	**92.0**	**88.1**
H2D-Gauss	81.5	82.1	83.2	80.6	88.6	88.4	90.5	87.1

Table 15: TIM (resize) OOD detection results. **Best** in column and **best** overall.

Individual	79.7 ±0.4	80.5 ±0.4			75.9 ±0.5	76.9 ±0.5		
S2D Individual	79.2 ±0.6	80.0 ±0.5	80.2 ±0.3	80.2 ±0.4	76.0 ±1.0	77.1 ±1.0	77.1 ±0.7	76.7 ±0.7
MC ensemble	79.8 ±0.4	80.6 ±0.3	80.7 ±0.4	68.3 ±1.7	76.1 ±0.7	77.0 ±0.6	77.1 ±0.6	59.5 ±1.6
S2D MC ensemble	79.4 ±0.6	80.3 ±0.7	80.2 ±1.0	80.1 ±0.7	75.9 ±0.9	77.1 ±1.0	77.2 ±1.1	76.8 ±0.6
Deep ensemble	81.8	82.7	82.7	72.5	78.4	79.3	79.2	64.1
S2D Deep ensemble	**81.9**	**82.9**	**82.9**	**82.5**	**79.0**	**80.2**	**80.2**	**79.6**
H2D-Gauss	80.9	81.4	81.4	81.5	77.4	79.0	78.9	78.0

Table 16: TIM (random crop) OOD detection results. **Best** in column and **best** overall.

Individual	71.2 ±3.8	72.8 ±4.0			68.9 ±3.5	70.9 ±4.0		
S2D Individual	73.1 ±3.0	74.9 ±3.6	76.3 ±3.9	75.9 ±3.4	71.4 ±1.7	73.7 ±2.2	74.5 ±2.4	73.4 ±2.4
MC ensemble	70.1 ±3.5	71.8 ±3.7	72.1 ±3.7	57.1 ±1.0	68.1 ±3.6	70.2 ±3.9	70.6 ±3.9	50.4 ±1.1
S2D MC ensemble	71.7 ±2.7	73.8 ±3.2	74.2 ±3.3	73.7 ±3.1	70.0 ±1.5	72.6 ±1.7	73.3 ±1.8	71.9 ±1.6
Deep ensemble	72.2	74.5	74.7	65.2	70.3	72.9	73.0	58.1
S2D Deep ensemble	74.3	**77.0**	77.3	**77.1**	**72.6**	**75.9**	**76.2**	**75.5**
H2D-Gauss	**75.2**	76.9	**77.3**	76.4	72.0	74.0	74.5	73.5