

Perception-Guided Accuracy Estimation: A Universal Framework for Robust Model Evaluation

Hao Sun, Zhongyi Han, Yilong Yin

Frontiers of Computer Science, DOI: [10.1007/s11704-026-51697-6](https://doi.org/10.1007/s11704-026-51697-6)

Problems & Ideas

- Problems of conventional stereo matching approaches:
 - Conventional accuracy estimation methods rely on strong assumptions that may not always hold in real-world scenarios.
 - Existing approaches fail to provide stable dataset-level accuracy estimates without target labels.
- Ideas:
 - Reformulate accuracy estimation as a perception-driven regression problem based on feature visualization.
 - Human-like perception of class structure provides robust cues for model reliability under distribution shift.

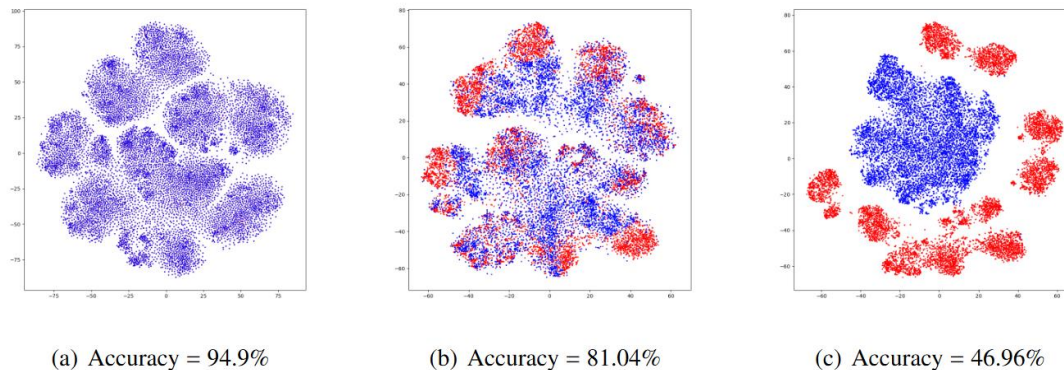


Fig. 1 Feature representation visualization (t-SNE) corresponding to different model performance levels. From left to right, the classification accuracies of the evaluated model on the test set (blue) decrease, with values of 94.9%, 81.04%, and 46.96%, respectively. This visualization illustrates how diminishing feature quality (as visually represented) correlates with a substantial degradation in model accuracy.

Main Contributions

- Contributions:
 - A perception-guided framework for estimating model accuracy under distribution shift without strong assumptions.
 - A representation and mentoring strategy that stabilizes regression from visualized feature distributions.;
 - Extensive validation across multiple benchmarks demonstrates robustness and generality.

Table 1 Mean Absolute Error (MAE) (%) of predicted accuracy on the test set with distribution shift. HPT outperforms other methods on most test datasets. For CIFAR10C and CIFAR100C datasets, we choose the most severe corruption to report the method's ability to deal with severe domain shift. *CIFAR10C-Average represents the average result over all CIFAR10C datasets.

Training set	Test set	AC	DoC	Rotation	AgreeScore	ATC	ProjNorm	MANDOLINE	HPT (ours)
CIFAR10	CIFAR-10.1	5.93±0.92	3.86±0.64	5.40±0.81	1.00±0.18	0.90±0.17	3.74±0.58	2.41±0.41	0.14±0.05
CIFAR10	CIFAR10C-brightness	4.29±0.72	2.22±0.38	0.43±0.07	1.72±0.31	0.61±0.12	0.01±0.01	0.62±0.14	1.31±0.29
	CIFAR10C-contrast	28.58±4.85	26.51±4.30	3.49±0.52	19.88±3.15	7.00±1.26	13.65±2.40	7.41±1.35	4.30±0.86
	CIFAR10C-defocus-blur	9.91±1.65	7.83±1.32	2.27±0.39	3.25±0.57	0.90±0.17	8.35±1.41	5.21±0.96	4.58±0.92
	CIFAR10C-elastic-transform	14.75±2.35	12.68±2.03	6.68±1.07	3.96±0.69	2.84±0.51	9.55±1.62	3.91±0.73	1.80±0.41
	CIFAR10C-fog	10.41±1.78	8.34±1.42	9.31±1.59	0.03±0.02	1.49±0.27	6.25±1.08	3.02±0.58	0.69±0.18
	CIFAR10C-frost	15.37±2.57	13.50±2.19	1.44±0.24	3.86±0.66	5.34±0.93	1.61±0.29	3.92±0.71	5.15±1.01
	CIFAR10C-gaussian-blur	12.47±2.01	10.40±1.70	5.15±0.88	1.39±0.25	0.37±0.08	8.95±1.49	6.11±1.10	8.52±1.62
	CIFAR10C-gaussian-noise	20.56±3.28	18.49±2.95	17.60±2.81	15.93±2.54	15.65±2.49	7.53±1.20	3.90±0.70	1.81±0.42
	CIFAR10C-glass-blur	32.81±5.26	30.74±4.87	7.91±1.28	18.55±2.93	16.04±2.61	8.54±1.37	6.10±1.12	5.12±1.03
	CIFAR10C-impulse-noise	19.58±3.14	17.51±2.83	9.68±1.57	6.49±1.12	20.60±3.22	22.70±3.60	10.82±1.95	2.79±0.73
	CIFAR10C-jpeg-compression	19.56±3.11	17.49±2.80	4.77±0.83	4.25±0.75	2.91±0.53	2.91±0.53	2.21±0.42	2.95±0.61
	CIFAR10C-motion-blur	12.77±2.10	10.70±1.79	12.23±2.06	4.99±0.87	1.49±0.27	6.23±1.08	4.10±0.77	3.47±0.66
	CIFAR10C-pixelate	42.56±6.81	40.49±6.32	21.73±3.28	27.58±4.53	5.80±0.99	4.81±0.89	12.72±2.43	
	CIFAR10C-saturate	4.72±0.85	2.65±0.48	2.08±0.39	0.65±0.13	0.20±0.05	5.04±0.91	2.40±0.47	0.73±0.18
	CIFAR10C-shot-noise	16.80±2.85	14.73±2.50	14.18±2.41	12.46±2.09	16.75±2.82	4.29±0.86	2.10±0.41	0.24±0.08
	CIFAR10C-snow	9.18±1.60	7.11±1.27	5.92±1.06	0.51±0.12	1.20±0.24	1.69±0.33	1.12±0.24	0.99±0.25
	CIFAR10C-spatter	7.88±1.36	5.81±1.01	3.21±0.58	1.41±0.27	0.94±0.21	0.99±0.22	0.52±0.13	0.08±0.03
	CIFAR10C-speckle-noise	14.56±2.42	12.49±2.06	10.32±1.76	10.93±1.86	16.81±2.86	10.22±1.75	6.01±1.09	4.82±0.97
	CIFAR10C-zoom-blur	8.99±1.53	6.91±1.19	0.14±0.04	3.24±0.58	0.04±0.02	4.39±0.79	1.94±0.38	0.43±0.11
	CIFAR10C-Average*	9.94±1.72	7.87±1.36	5.19±0.91	4.15±0.80	3.94±0.78	4.39±0.86	3.10±0.62	2.49±0.55
MNIST	SVHN	23.42±3.98	23.42±3.98	1.78±0.36	52.89±8.45	12.07±2.02	54.65±8.90	6.52±1.21	0.58±0.19
	USPS	34.09±5.12	34.08±5.11	12.42±2.05	10.91±1.80	3.19±0.62	23.71±3.98	4.88±0.93	0.05±0.02

Table 2 Mean Absolute Error (MAE) (%) of predicted accuracy on the test set with distribution shift. *CIFAR100C-Average represents the average result over all CIFAR100C datasets.

Training set	Test set	AC	DoC	Rotation	AgreeScore	ATC	ProjNorm	MANDOLINE	HPT (ours)
CIFAR100	CIFAR100C-brightness	9.65±1.60	7.58±1.25	5.56±0.96	0.99±0.24	0.32±0.09	8.30±1.38	4.91±0.92	0.51±0.15
	CIFAR100C-contrast	39.95±6.90	37.88±6.40	10.85±2.00	20.07±3.80	23.27±4.20	5.72±1.10	4.11±0.83	3.81±0.92
	CIFAR100C-defocus-blur	25.00±4.10	22.93±3.70	4.28±0.78	9.53±1.70	9.13±1.60	4.78±0.86	3.10±0.62	0.74±0.21
	CIFAR100C-elastic-transform	20.06±3.50	17.99±3.10	1.29±0.32	4.99±0.98	2.42±0.55	0.17±0.06	0.64±0.19	3.36±0.82
	CIFAR100C-fog	24.98±4.00	22.91±3.60	3.70±0.74	3.48±0.70	13.10±2.30	6.79±1.25	4.20±0.84	7.94±1.90
	CIFAR100C-frost	21.04±3.50	18.97±3.10	5.17±0.95	3.14±0.63	7.83±1.45	7.07±1.32	4.81±0.96	3.05±0.86
	CIFAR100C-gaussian-blur	28.97±4.90	26.90±4.40	4.93±0.91	13.03±2.35	11.75±2.10	3.09±0.61	2.10±0.47	0.06±0.03
	CIFAR100C-gaussian-noise	36.27±6.10	34.20±5.70	3.15±0.64	3.27±0.65	7.31±1.36	3.25±0.71	2.51±0.52	0.36±0.11
	CIFAR100C-glass-blur	36.83±6.20	34.76±5.80	4.46±0.86	18.08±3.25	18.65±3.32	0.18±0.06	0.91±0.23	0.61±0.19
	CIFAR100C-impulse-noise	42.39±7.30	40.32±6.90	9.39±1.76	18.82±3.40	13.01±2.35	10.76±2.01	6.52±1.28	1.96±0.58
	CIFAR100C-jpeg-compression	21.09±3.40	19.02±3.00	6.38±1.16	6.79±1.23	3.21±0.64	5.10±0.96	3.21±0.64	1.81±0.55
	CIFAR100C-motion-blur	20.24±3.30	18.17±2.90	5.10±0.93	5.07±0.92	5.10±0.96	1.15±0.41	1.41±0.36	1.76±0.52
	CIFAR100C-pixelate	38.96±6.70	36.89±6.20	16.35±3.05	16.66±3.09	22.92±4.15	9.62±1.76	6.10±1.22	4.69±1.32
	CIFAR100C-saturate	16.63±2.80	14.55±2.40	2.27±0.43	2.06±0.40	6.17±1.19	8.56±1.62	5.12±1.02	1.51±0.45
	CIFAR100C-shot-noise	35.44±5.90	33.36±5.40	1.06±0.24	6.55±1.21	6.80±1.29	6.21±1.16	3.95±0.79	2.64±0.79
	CIFAR100C-snow	13.83±2.30	11.76±2.00	2.03±0.39	1.19±0.24	0.72±0.16	1.39±0.29	0.81±0.21	0.27±0.08
	CIFAR100C-spatter	11.05±1.90	8.98±1.60	6.71±1.25	1.09±0.23	0.57±0.12	9.10±1.72	5.41±1.05	3.51±1.01
	CIFAR100C-speckle-noise	35.19±5.80	33.12±5.30	0.46±0.14	7.75±1.43	9.53±1.81	5.01±0.94	2.92±0.58	1.12±0.34
	CIFAR100C-zoom-blur	20.07±3.30	18.00±2.90	1.70±0.34	7.33±1.35	5.63±1.02	4.77±0.86	3.14±0.63	1.68±0.50
	CIFAR100C-Average*	17.21±2.90	15.14±2.50	5.43±1.01	5.18±0.98	4.71±0.91	8.51±1.62	5.32±1.06	2.07±0.61