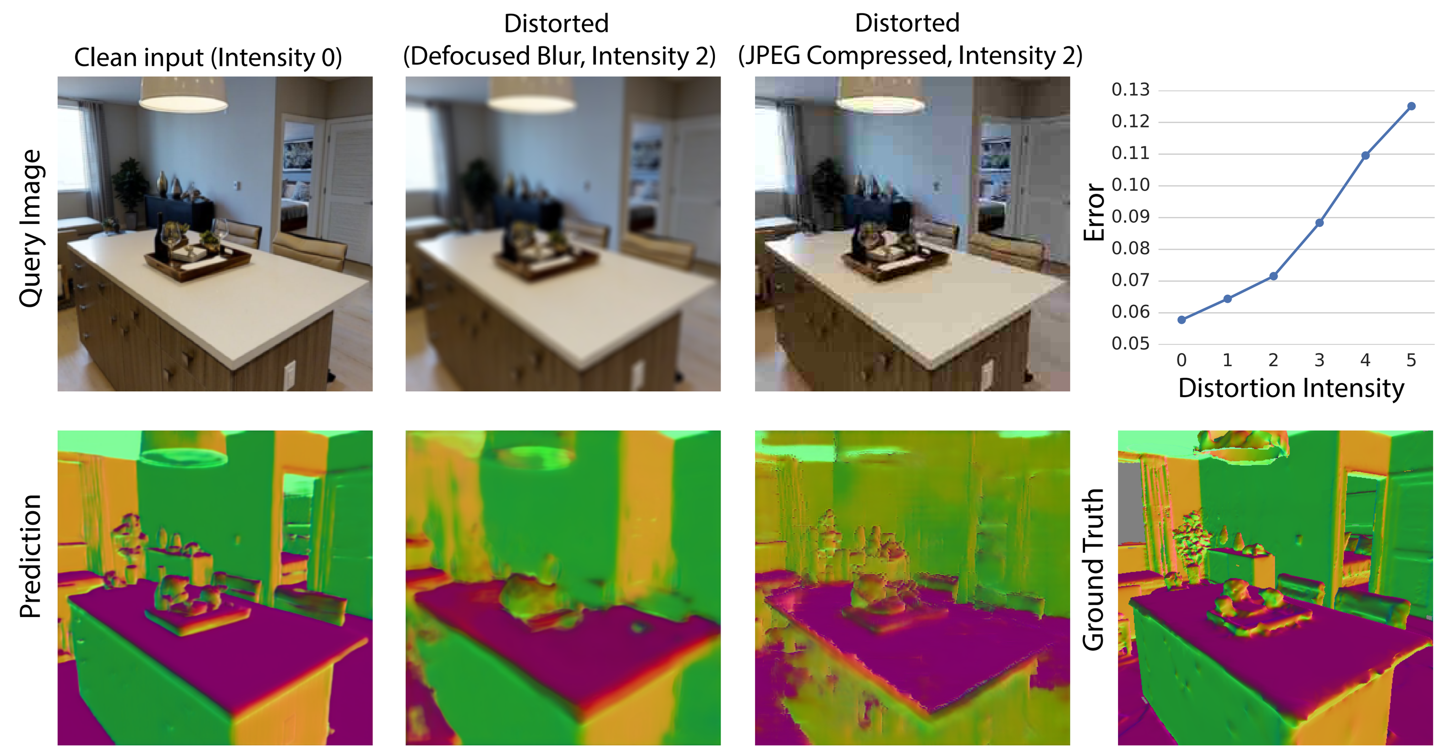
Robustness via Cross-Domain Ensembles

Teresa Yeo*, Oğuzhan Fatih Kar*, Amir Zamir
crossdomain-ensembles.epfl.ch

1) Problem

- Neural networks (NNs) deployed in real world will encounter data with naturally occurring distortions.
- Their predictions under such shifts from the training data are unreliable, e.g. see surface normals results below.



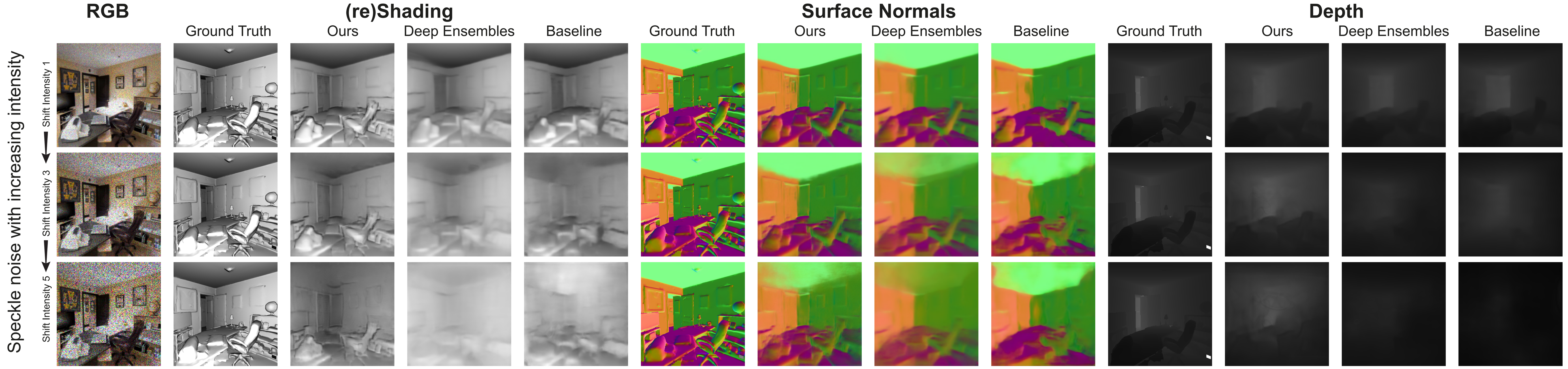
4) Results for adversarial distortions

★ Improved robustness to attacks *without* adversarial training

Method \ ϵ	Normal 2	Normal 4	Normal 8	Normal 16	Reshade 2	Reshade 4	Reshade 8	Reshade 16	Depth 2	Depth 4	Depth 8	Depth 16
Baseline UNet	8.23	11.53	13.03	14.37	17.92	22.78	27.26	34.40	5.50	6.76	8.36	9.80
Deep ensembles	**7.49**	11.13	13.36	15.65	15.66	21.95	27.75	34.98	5.45	6.68	8.27	10.52
Inv. var. merging	7.60	**8.89**	**10.40**	**12.77**	**15.56**	**16.55**	**18.93**	**22.01**	**4.94**	**4.99**	**5.93**	**6.75**
Adv. T. (lower bound error)	5.78	5.74	5.45	5.53	9.39	8.98	8.07	8.20	2.23	2.27	2.39	2.74

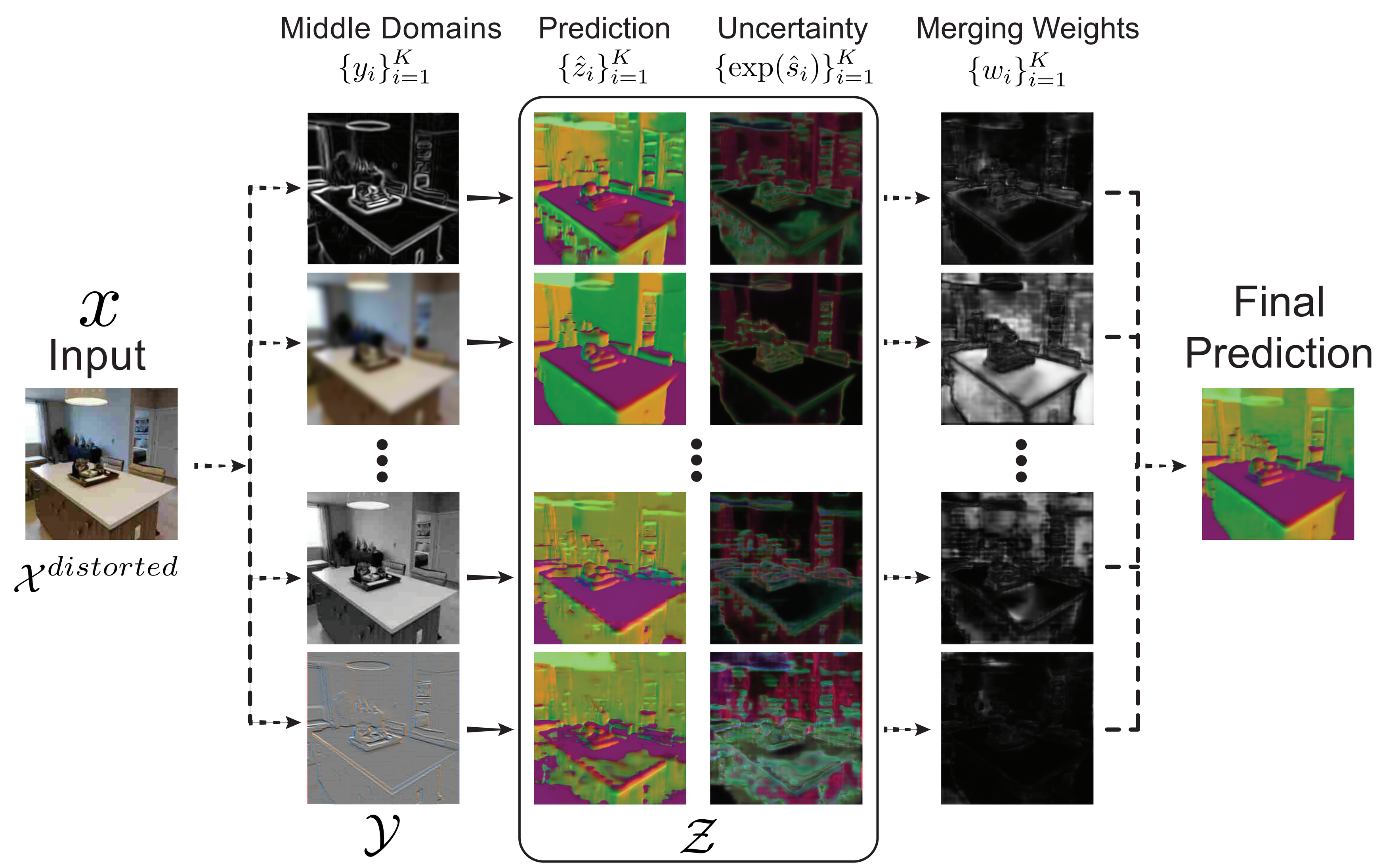
5) Results for natural distortions

★ Notable improvements especially in fine-grained regions



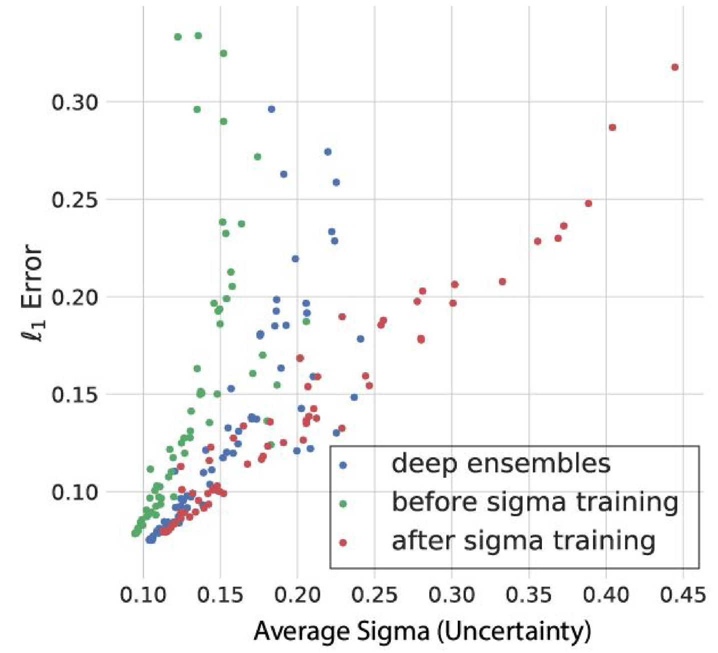
2) How do we obtain robust predictions?

- **Middle domains:** Consider a set of transformations (examples below) that are each invariant to a particular change in the input image (e.g. brightness).
- **Learn the mappings:** Train a model from each middle domain to target domain. Also estimate the uncertainty by a simple parameterization of the output, using a likelihood loss.
- **Uncertainty-guided merging:** Each model will contribute to the final prediction based on its confidence.



3) Avoiding overconfident predictions

- **Uncertainty estimates under distribution shifts are poorly calibrated:** Models output poor predictions with high confidence. This reduces the quality of uncertainties as merging weights.
- **Calibration:** We propose *sigma training* as a calibration stage to alleviate this. It encourages the model to output high uncertainties while keeping predictions fixed.



★ The resulting uncertainties have stronger correlation with errors for *unseen* corruptions.

6) Key takeaways

★ Middle domains promote ensemble diversity and reduce NN tendency to learn from superficial cues. Manual handpicking is not required. They add negligible computation overhead.
★ Using uncertainties as weights significantly outperforms uniformly averaging the ensemble predictions.
★ The proposed method improves robustness for several tasks and datasets under unseen adversarial & non-adversarial shifts.
★ Performance on clean data is not sacrificed.