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.

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 predictions with high confidence, which 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.

4) Results for adversarial distortions

- **Improved robustness to attacks without adversarial training**

5) Results for natural distortions

- **Notable improvements especially in fine-grained regions**

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**.