Supplementary Material for Improving Equivariance in State-of-the-Art Supervised Depth and Normal Predictors

Yuanyi Zhong, Anand Bhattad, Yu-Xiong Wang, David Forsyth
University of Illinois Urbana-Champaign
{yuanyiz2, bhattad2, yxw, daf}@illinois.edu

A. Quantitative study of equivariant error

In Figure 1, we qualitatively show that the state-of-the-art depth predictor, MiDaS-v3.0 DPT-Large [6], possess insufficient equivariance to cropping transform. Here, we illustrate the same problem in a quantitative manner. From the input image in Figure 1, we generate 5,000 random pairs of crops with scale variation 0.85-1 and aspect ratio variation 3/4-4/3. Note the scale variation is deliberately chosen not to be drastic. We resize them and pass all of them to the pre-trained network to get 5,000 depth map predictions. After that, we compute the AbsRel (absolute relative error of depth, averaged over pixels) between the overlapped region of the pairs of predictions (using one as the target), and call this number $\text{eqerr}_{\text{depth}}(f, t_1, t_2)$. See the following equation. Here, f is the depth predictor, t_1, t_2 represent a pair of randomly sampled crop transforms.

$$\text{eqerr}_{\text{depth}}(f, t_1, t_2) = \text{AbsRel}(t_1^{-1} \circ f \circ t_1(x), t_2^{-1} \circ f \circ t_2(x)).$$

This number essentially measures the degree of variation caused by random cropping. A perfectly equivariant predictor will have $\text{eqerr}_{\text{depth}} = 0$ for any crop transforms t_1, t_2. In Figure A.1, we draw the distribution of the 5,000 $\text{eqerr}_{\text{depth}}$'s in a box plot, and compare to the AbsRel between the prediction and the ground truth (red line), for both the model before and after our equivariant fine-tuning.

From the left part of Figure A.1, we observe that the variation caused by random cropping (the box plot) is very large compared to the AbsRel to ground truth. The mean of variation is almost 6%. The largest error can go beyond 10%, whereas the error against ground truth is just 13.6%. The right part shows the same quantities after our equivariant fine-tuning. We observe that the the variation caused by cropping is much smaller now, while the accuracy with respect to ground truth also improves.

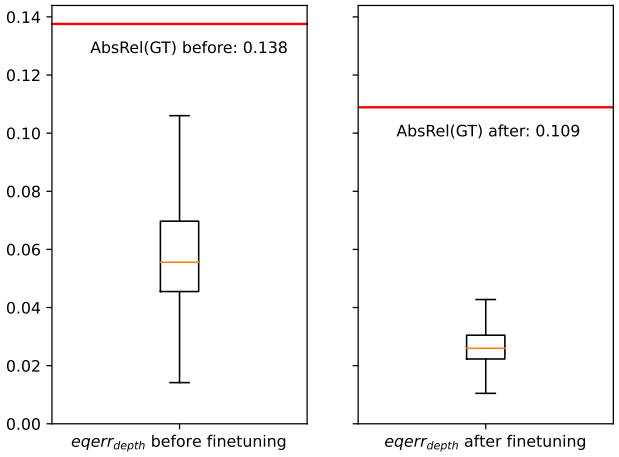

Figure A.1. Box plots of variations caused by random cropping, before and after our unsupervised equivariant finetuning, for the example picture in Figure 1 of the main paper. Red lines show the error against ground truth. The model is pre-trained MiDaS-v3.0 DPT-Large [6]. Equivariant fine-tuning shrinks the variation by random cropping considerably.

B. Additional results

Using ImageNet supervised and dense contrastive learning pre-trained ResNets as initialization. The main paper Table 1 shows supervised Taskonomy [11] results with the UNet architecture defined in [10]. Here, we consider ResNet-50 [4] based encoder-decoder architecture defined in [11] for depth prediction. This architecture uses the ResNet-50 backbone as encoder to obtain a 2048x8x8 feature map for 3x256x256 RGB input, then uses a 10-layer convolutional decoder (with transposed convolutions in last 5 layers) to decode a 256x256 full-sized output. We put our equivariance loss (EqLoss) with K=3 on the second to last layer.

With ResNet as encoder, we have the possibility to initialize our training with pre-trained models. We examine ImageNet supervised classification pretrained from torchvision [5] and PixelPro [9] ImageNet-pretrained model. The reason to study pixel-wise dense contrastive learning pre-trained models such as DenseCL [8] and PixelPro (PixPro) [9], is that, (1) they are pre-trained with dense correspondence between two random crops – the idea is sim-

Table B.1. ResNet-50 results on Taskonomy Depth-ZBuffer [11] with random, ImageNet sup., and PixelPro [9] initializations. '+EqLoss' rows are adding our equivariant loss. $\delta > 1.25$ and AbsRel are validation error metrics (lower the more accurate). 'EqLoss' column is the validation equivariant loss (lower the more equivariant). Other experiment settings are the same as Table 1 in the main paper.

Pretrain	+EqLoss	δ>1.25 (%)↓	AbsRel (%)↓	EqLoss↓
Random init.		31.6	21.0	0.485
Random init.	✓	30.3	20.3	0.349
ImageNet sup.		24.7	17.6	0.403
ImageNet sup.	✓	24.6	17.5	0.358
PixelPro [9]		22.8	17.0	0.514
PixelPro	✓	22.7	16.7	0.463

Table B.2. Compare the K=2 pixel-wise dense contrastive loss variant and the K=3 variant in our main results.

	Dense CL (K=2)	Our EqLoss (K=3)
Depth: δ>1.25 (%)↓	25.3	**25.0**
Normal: Ang error °↓	6.53	**6.47**

ilar to ours, therefore may achieve higher equivariance to cropping; (2) better suited for downstream dense prediction tasks than image-level pretrained models. We are curious if initializing from these feature stacks would alleviate the equivariance problem in depth predictors.

Table B.1 lists the results. We can make several observations: (1) Initializing from PixelPro is better than from ImageNet sup., which is in turn better than random init. (2) PixelPro initialization does not completely resolve the non-equivariance issue, as the EqLoss of the final depth predictor is still high, which means there is still inconsistency between different crops of the same images. (3) Regardless of initialization, adding our EqLoss technique improves the final accuracy and equivariance. The improvement is larger for random init than others. The equivariance for PixelPro is improved as well, measured by lower validation EqLoss. (4) Our method generalizes to ResNet architecture, in addition to the UNet (Tables 1,2) and Dense Prediction Transformer (Table 3, [6]) architectures in the main paper.

Using dense CL loss during depth/normal network training. When K=2, our equivariant loss reduces to a type of pixel-wise dense contrastive loss, while our final results use K=3. Pixel-wise dense contrastive methods such as DenseCL [8] and PixelPro [9] appear in self-supervised learning literature. They are relevant to our paper because the idea is also to learn equivariant rather than invariant representations in contrastive learning. Pixel-level contrastive learning have shown to improve upon image-level contrastive learning [1,3] when transferring to detection and segmentation downstream tasks. However, they have not been applied to state-of-the-art depth and normal predictors to the best of our knowledge. Figure 4 includes this comparison for depth prediction. In Table B.2, we show that result again and supplement the surface normal result. In both tasks, the K=3 variant is better than the pixel-wise dense contrastive loss variant, under the same wall-clock time budget. Regardless of the variants, our main point is that equivariance is missing from current depth and normal predictors and the equivariant regularization technique improves the performance of them by increasing equivariance.

Table B.3. Compare imposing equivariant loss in label space and feature space in supervised Taskonomy [11] settings.

	L: Label Space	L-1: Feature Space
Depth: δ>1.25 (%)↓	25.7	**25.0**
Normal: Ang error °↓	6.54	**6.47**

Comparison of label and feature space equivariant regularization. We show additional surface normal prediction result for the comparison of loss layer in Table B.3 to supplement Table 6. We find applying our EqLoss on feature space is also better than label space for surface normal.

C. Additional details

Cosine weighting window. In Section 5.1, we mention the use of a weighting window with smooth edges when computing the equivariant average to suppress the boundary artifacts. The motivation is that the boundary predictions (of depths, for example) may not be accurate because the input may not contain enough context for those pixels. It is beneficial to down-weight them in averaging. Figure C.1 shows the illustration of the actual weighting window used in our experiments. The smooth edges are generated from a smooth-changing cosine function. The average operation in Eq. 1 will become a weighted average.

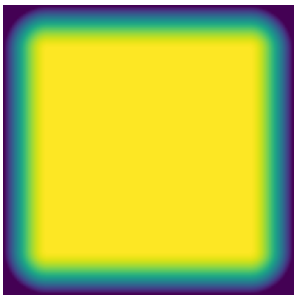

Figure C.1. Cosine window weighting map when computing the output average by Eq. 1. The brighter pixels mean weights close to 1, and the darker pixels mean weights close to 0. The edge transition follows a cosine function. The output map of each crop is weighted by this map. The boundary outputs will contribute less to the average to avoid artifacts.

Linear predictor and stop gradient. Inspired by contrastive learning with predictor [2], we compared equivariant loss with or without a predictor function between the individual crop outputs and the average output. The intuition is that the predictor and stop gradient technique prevents the network from learning a collapsed constant representation. We found that training with a predictor is usually more stable and better-performing. In one experiment, the validation L1 loss improved from 5.61e-2 to 5.49e-2. Our predictor is a linear layer without bias terms, initialized to be the identity function, predicting from each crop output to the average output. The stop gradient is on the average output.

Applying equivariant loss to more than one layer. Table 6 of the main paper studies the location of our equivariant loss. It is natural to consider applying the loss on more than one layer. We attempted applying it on both L-1 and up1. However, there seems to be no additional gain from this: the $\delta > 1.25$ becomes 25.3%. While we believe there might be more potential in general, we feel applying to multiple layers requires more effort on hyper-parameter tuning and causes unnecessary complexity. We thus stick to applying on only one layer for simplicity.

Special considerations for depth predictors. The pre-trained model from [6, 7] are trained with a combination of loss functions in the disparity space (inverse depth), and the prediction only satisfies $p \approx \alpha + \beta \frac{1}{d}$ where p is the predicted disparity and d is the actual depth. Following their practice, we train with the L1 loss on the disparities (inverse depths) instead of depths. Due to the same reason, before computing the evaluation metrics, we also follow their practice to use least square regression, i.e., compute the best α and β, to align the values of predicted disparity map to the ground truth disparity map.

D. Is it just depth or normal?

Our paper focuses on state-of-the-art depth and normal prediction models, and finds them not very equivariant to crop transform. Does the problem only exist for depth and normal predictors? We believe the problem is actually quite prevalent and was previously under-explored. In many image-to-image translation tasks where equivariance is desired, the network is not explicitly trained to be equivariant –It does not have strong preference that the output of cropping should not change.

Here, we use the CycleGAN horse-to-zebra translation [12] as yet another failure example of equivariance in dense prediction models. Figure D.1 shows that the resulting stripes on the zebra are sensitive to the crop locations, while ideally the translated image should be a deterministic mapping of the input image contents. This problem is even more salient if the method is used to translate a video, as in the official gif example of CycleGAN[1], where we can visually see the unstable predictions of stripes. Therefore, we believe our approach is broadly relevant to the field. Our equivariant regularization loss can be employed as an additional loss during training to promote equivariance.

E. More visualization

As a supplementary visualization to Figure 3 (depth and normal) of the main paper, we provide edge detection results on Taskonomy [11] in Figure D.2. Our model with equivariant loss is more robust to cropping.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[2] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 3

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[5] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 1

[6] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 1, 2, 3

[7] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2020. 3

[8] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021. 1, 2

[9] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, 2021. 1, 2

[10] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *CVPR*, 2020. 1

[11] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 1, 2, 3

[1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

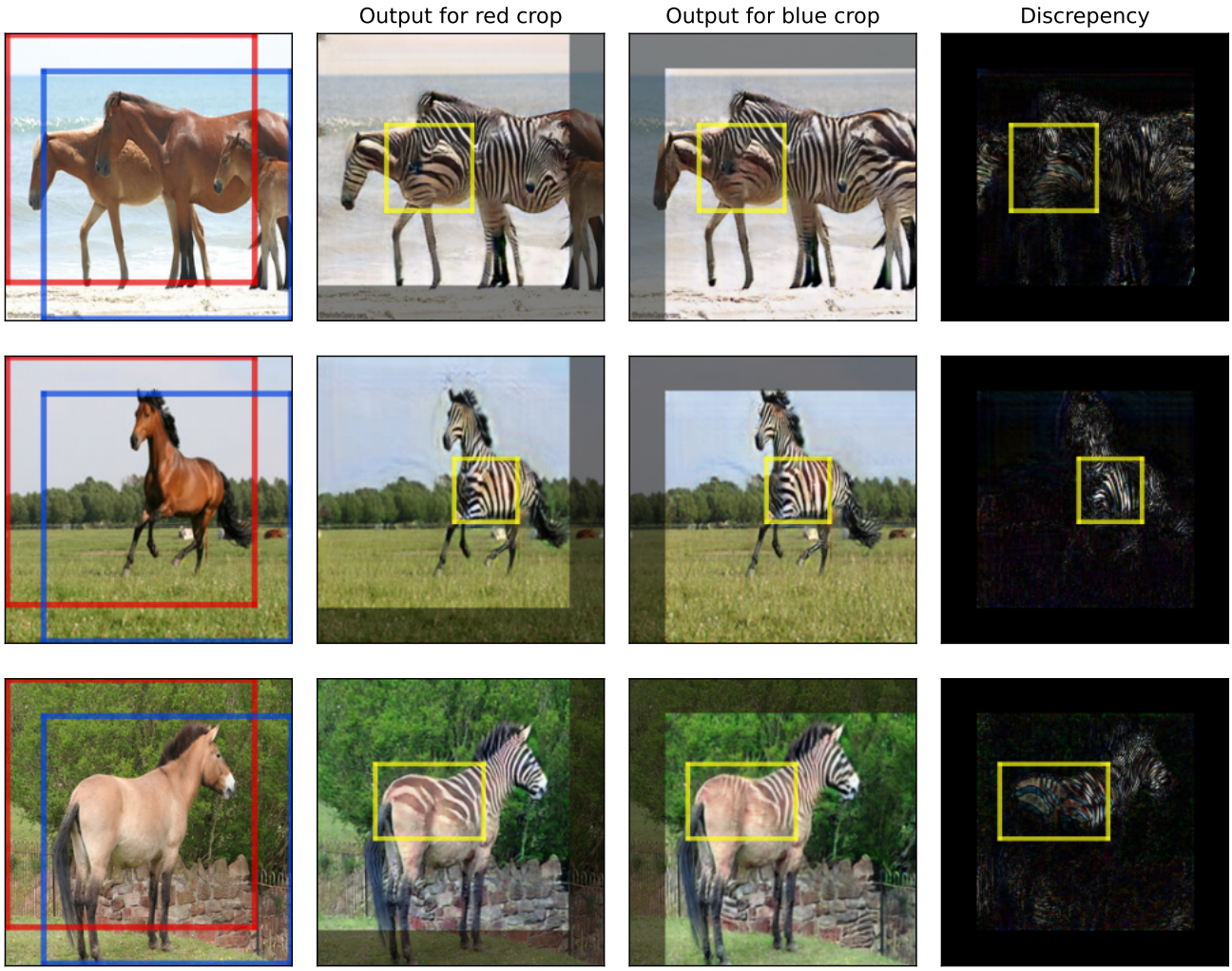

Figure D.1. CycleGAN fails to be equivariant in the horse-to-zebra examples, in addition to the depth and surface normal cases of the main paper. We believe the issue of non-equivariance is quite common in image dense prediction models and has been overlooked by prior work (besides in semantic segmentation). Our equivariant regularization approach has potential uses in these other domains.

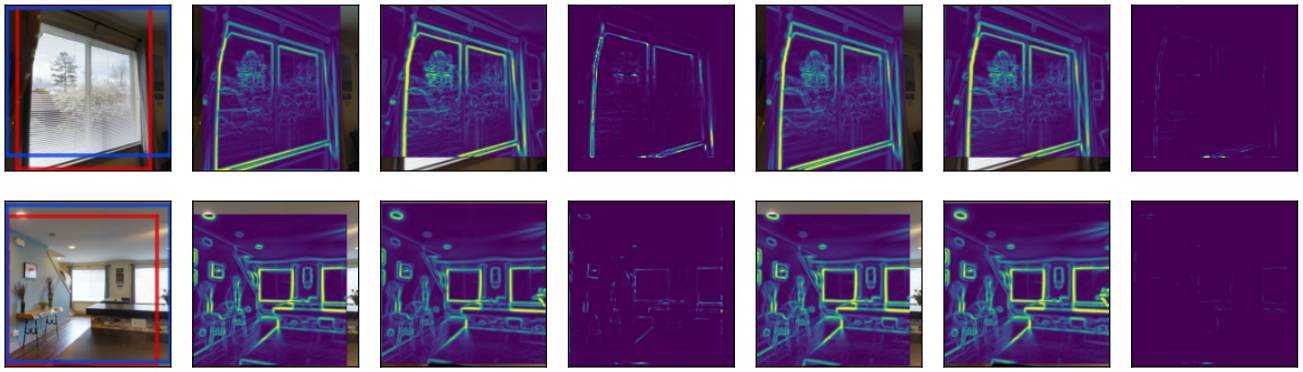

Figure D.2. Visualization of Table 1 edge detection results of Taskonomy validation images. From left to right, the columns are images (red, blue crops), predictions of the baseline model (without equivariant loss) and their discrepancy, predictions of our model (with equivariant loss) and their discrepancy.

[12] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3