LocNet: Does Depth Density Estimation help learn Shape Bias in CNNs?

Gaurav Gaonkar, Leyang Hu, Shangyang Min, Zhuoyue Jin
Department of Computer Science
Brown University
Providence, RI 02912, USA

1 Introduction

There is a shred of strong evidence in the field of cognitive science (Haushofer et al., 2008; Allen et al., 2009) showing human biological vision is both texture and shape bias. However, state-of-the-art image classification models predominantly rely on texture features, in comparison to the shape information of the object (Geirhos et al., 2021). This study aims to bridge this gap between models and humans by introducing a new training objective. Here, we perform a joint optimization of image classification and depth estimation tasks. The hypothesis posits that training a model jointly for depth estimation tasks will enhance its awareness of object structure properties, improving the model's ability to learn texture and shape information. We assess the classification performance of the model on both in-distribution and out-of-distribution datasets to comprehensively evaluate its robustness and effectiveness under diverse texture and shape perturbations. Our code is available at https://github.com/gaga1313/LOCNet.

2 Methodology

Here, we investigate whether optimizing weights simultaneously for depth estimation and classification tasks makes the model more biased toward the shape and texture of the object. We developed an encoder-decoder architecture (Rumelhart et al., 1986) utilizing ResNet50 (He et al., 2016) backbone, designed for simultaneous classification and depth density prediction. This architecture choice is predicated on our hypothesis that by learning depth information, the encoder exploits specific features crucial for discerning the 2D or 3D aspects of the objects. We evaluate the shape bias of the model on an out-of-distribution (OOD) dataset (Geirhos et al., 2018). Evaluating the jointly trained model on the OOD benchmark dataset helps determine whether the combined processing of textural and shape-related information inculcates shape and texture bias in the model, making it more robust and accurate than the classification-only trained model.

2.1 Data

2.1.1 Training and Validation Dataset

Here we develop a novel dataset by creating (image, depth map) pairs for a subset of the Imagenet 1k dataset (Deng et al., 2009). Utilizing sixteen animate and inanimate classes from Imagenet 1k, we generate the pseudo-depth density maps using the depth-anything model (Yang et al., 2024). The sixteen inanimate classes include *knife, keyboard, elephant, bicycle, airplane, clock, oven, chair, bear, boat, cat, bottle, truck, car, bird, and dog*. We specifically select the above 16 classes to maintain the overlap with the classes in the out-of-distribution (OOD) dataset (Geirhos et al., 2018).

2.1.2 Evaluation Dataset

Geirhos et al. (2018) created OOD dataset to measure and compare the robustness of humans and the SOTA computer vision networks on the image classification task. It includes twelve parametric texture-based image degradations and five shape-cue-based degradations shown in Figure 1 and Figure 2. It further evaluates the performance of humans and CNNs and shows that humans are shape-biased and robust towards perturbations. We utilize this dataset as a benchmark for our jointly trained model.

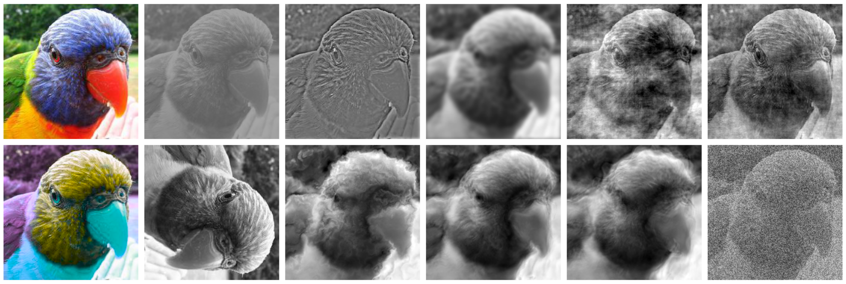

Figure 1: Twelve parametric filters in the OOD dataset.

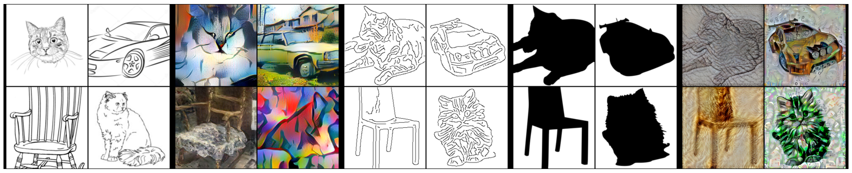

Figure 2: Five shape-cue filters in the OOD dataset.

2.2 Metrics

Here we optimize a combination of Categorical-Cross-Entropy (CCE) for classification, and Mean Squared Error (MSE) for depth density reconstruction. The combined loss functions are described in Equation 1.

$$L_{joint} = \alpha \times \lambda \times \underbrace{-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(p_{ij})}_{\text{Cross-Entropy (CCE) Loss}} + \beta \times \underbrace{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}_{\text{Mean Squared Error (MSE) Loss}} \tag{1}$$

where, $\boldsymbol{\alpha}$ is CCE scaling factor, $\boldsymbol{\beta}$ is MSE scaling factor, and $\boldsymbol{\lambda}$ is CCE loss annealing.

2.3 Model Structure

Following our hypothesis, we create an encoder-decoder architecture for joint optimization of the loss function given in Equation 1. We utilize an Autoencoder (Rumelhart et al., 1986) architecture with a ResNet-50 encoder and transpose a ResNet-50 decoder as shown in Figure 3. The encoder is followed by another linear layer to do the classification task. Such architecture design compels the encoder to preserve the object's structural information until the bottleneck layer, allowing structural and textural information to be further used by the decoder for depth density estimation and linear classifier for classification respectively.

On the contrary UNET-like architecture (Ronneberger et al., 2015) with skip connections between encoder and decoder convolution blocks are prone to shortcuts and important structure information might not propagate until the bottleneck of the encoder.

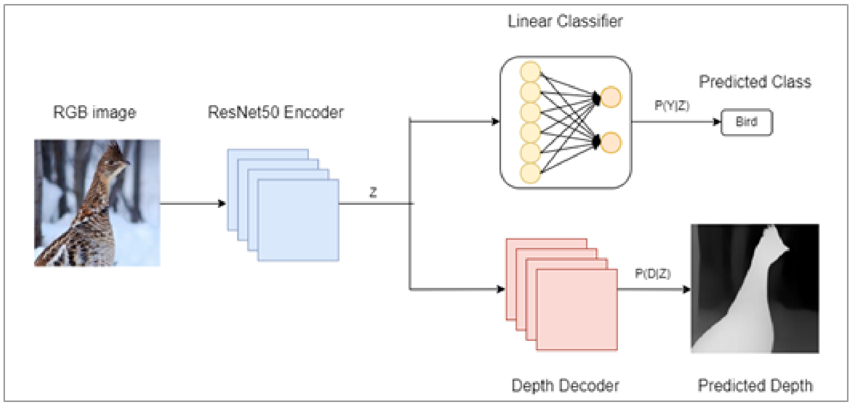

Figure 3: Architecture of the proposed model. An RGB image is encoded into the latent representation using ResNet50 as the encoder. Then the representation is simultaneously passed to a linear classifier and a depth decoder to obtain a classification result and a depth prediction map.

3 Challenges

3.1 UNet Architecture Learning Shortcuts

During initial training experiments, we utilized UNET (Ronneberger et al., 2015) architecture for joining training on classification and depth estimation tasks. Using UNET architecture led to an excellent reconstruction of depth maps shown in Figure 4. However, the performance on the OOD dataset was much worse than expected. Later, we found out that the skip connections in the convolution blocks led to shortcut paths for passing structural information of the object to the decoder leading to good depth density estimation maps but poor performance on OOD classification. This inspired us to use a bottleneck architecture like Autoencoder to avoid shortcuts through skip connections.

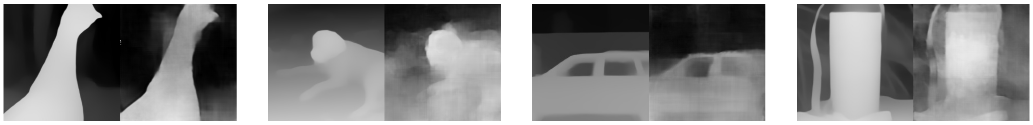

Figure 4: Ground Truth depth maps (left) and Reconstructed depth maps by UNET (right).

3.2 Joint Optimization of Loss Fucntion

One of the most important issues we dealt with was jointly training the model. As the approach requires the simultaneous optimization of both classification accuracy and depth map reconstruction, challenges come from the objective of our model and the nature of the two loss functions. CCE loss starts at a very high scale and converges quickly to a low scale, however, MSE starts from a very low scale and converges slowly. Balancing these two loss functions to train a unified model framework required careful tuning of hyperparameter α, warm-up epochs, and cosine learning rate schedule shown in Figure 5.

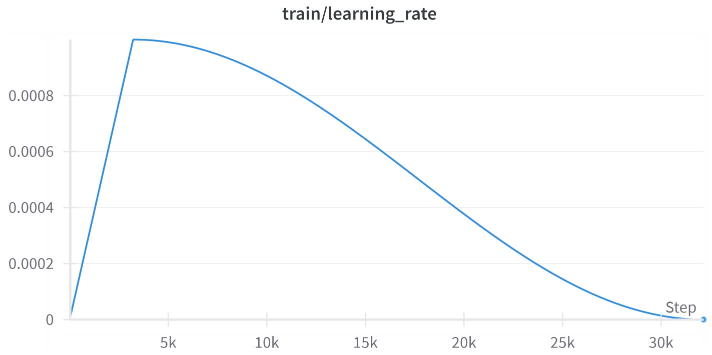

Figure 5: Learning rate scheduler applied in the training.

3.3 MSE CONVERGENCE

Due to the difference in convergence of the loss function, the model trained much faster for classification than depth density estimation. To boost performance on the depth density estimation task we employ tricks like warmup training with only MSE optimization for the initial 5 epochs, and CCE loss annealing shown in Figure 6. The depth density maps of our best-converged Autoencoder model are shown in Figure 7.

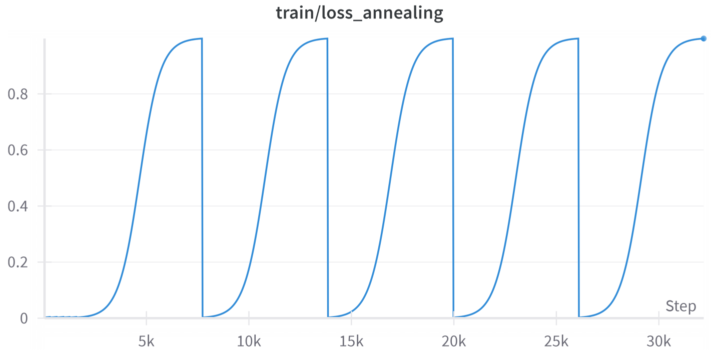

Figure 6: CCE loss annealing.

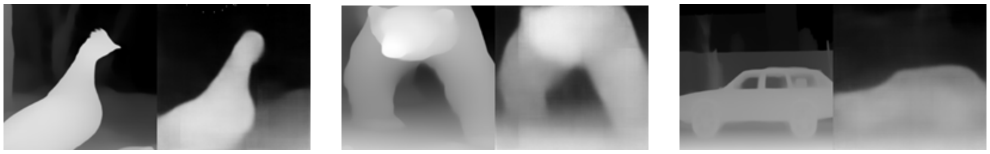

Figure 7: Ground Truth depth maps (left) and Reconstructed depth maps by Autoencoder (right).

4 RESULTS

4.1 BASELINE

To compare the performance of our Jointly trained model, we create a baseline ResNet50 model trained only for classification on sixteen class classification tasks. We ensure that both the models converge for the classification tasks and additionally the convergence of the jointly trained model on the depth estimation task. The convergence behavior of both models is shown in Figure 8, Figure 9 and Figure 10.

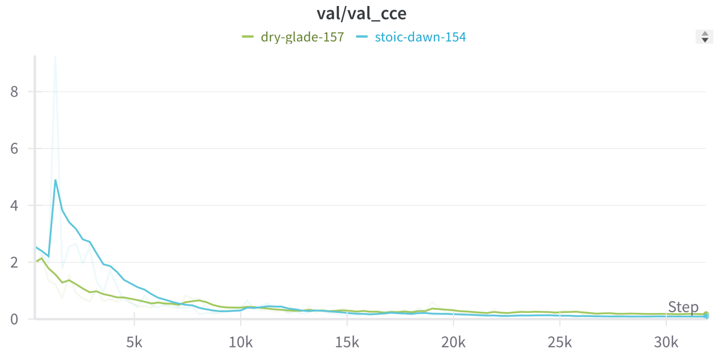

Figure 8: CCE loss on the validation set.

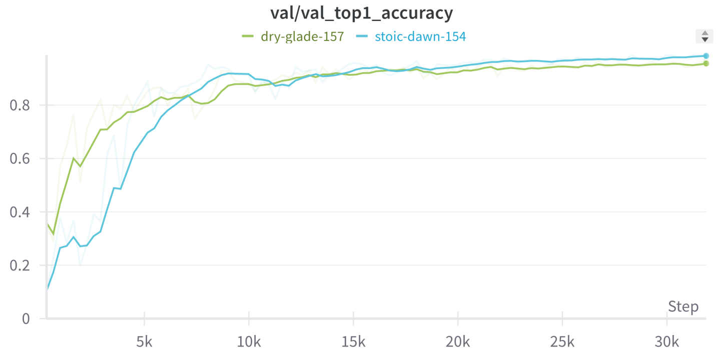

Figure 9: Accuracy on the validation set.

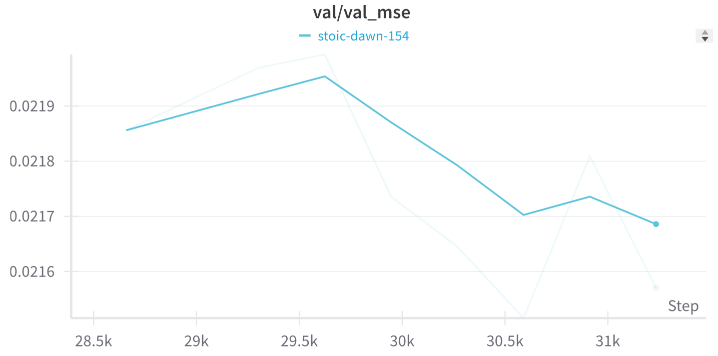

Figure 10: MSE on the validation set.

4.2 OOD Evaluation

The jointly trained model outperforms the classification-only model on all the benchmark filters on top-5 accuracy metrics as shown in Figure 12 and Table 1. Evaluated on top-1 accuracy the jointly trained model outperforms the classification-only model sketch, stylized image, and all parametric benchmark filters as shown in Figure 11 and Table 1.

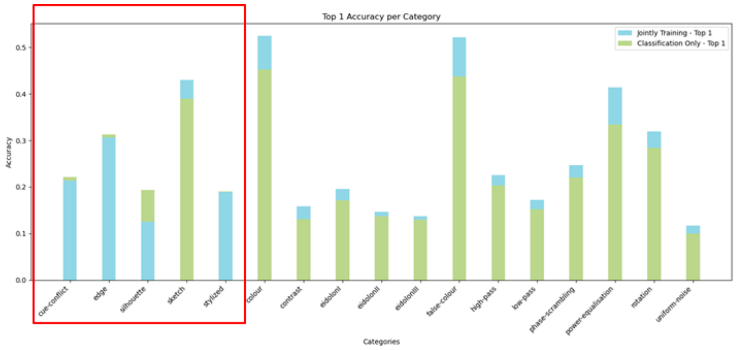

Figure 11: Top-1 accuracy on the OOD dataset. The red rectangle highlights categories specifically used to test the shape bias of the model. The jointly trained model has higher top-1 accuracy on the sketch filter and all texture-based filters.

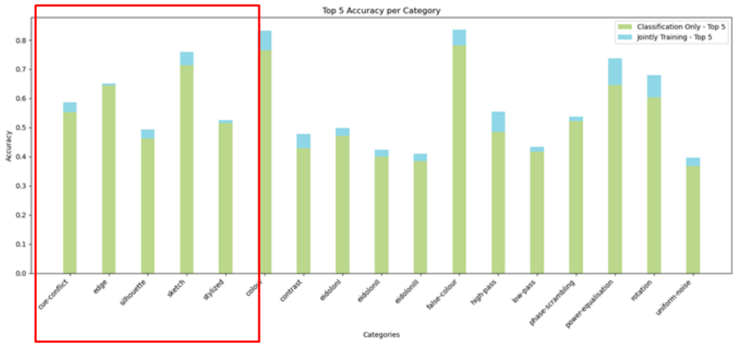

Figure 12: Top-5 accuracy on the OOD dataset. The red rectangle highlights categories specifically used to test the shape bias of the model. The jointly trained model has higher top-5 accuracy on all filters.

4.3 Sketch Filter - Gradient Weighted Class Activation Maps

We further investigated the reason for the better top-1 accuracy on the sketch filter using Grad-CAM to identify the regions important to the model while performing classification. We found out that the model trained with joint optimization has more spread out Grad-CAM maps, in shallow layers, indicating the model focuses on the entire object in the shallow layers and later forms a sub-network, which we discuss in detail in the Reflection and Discussion section.

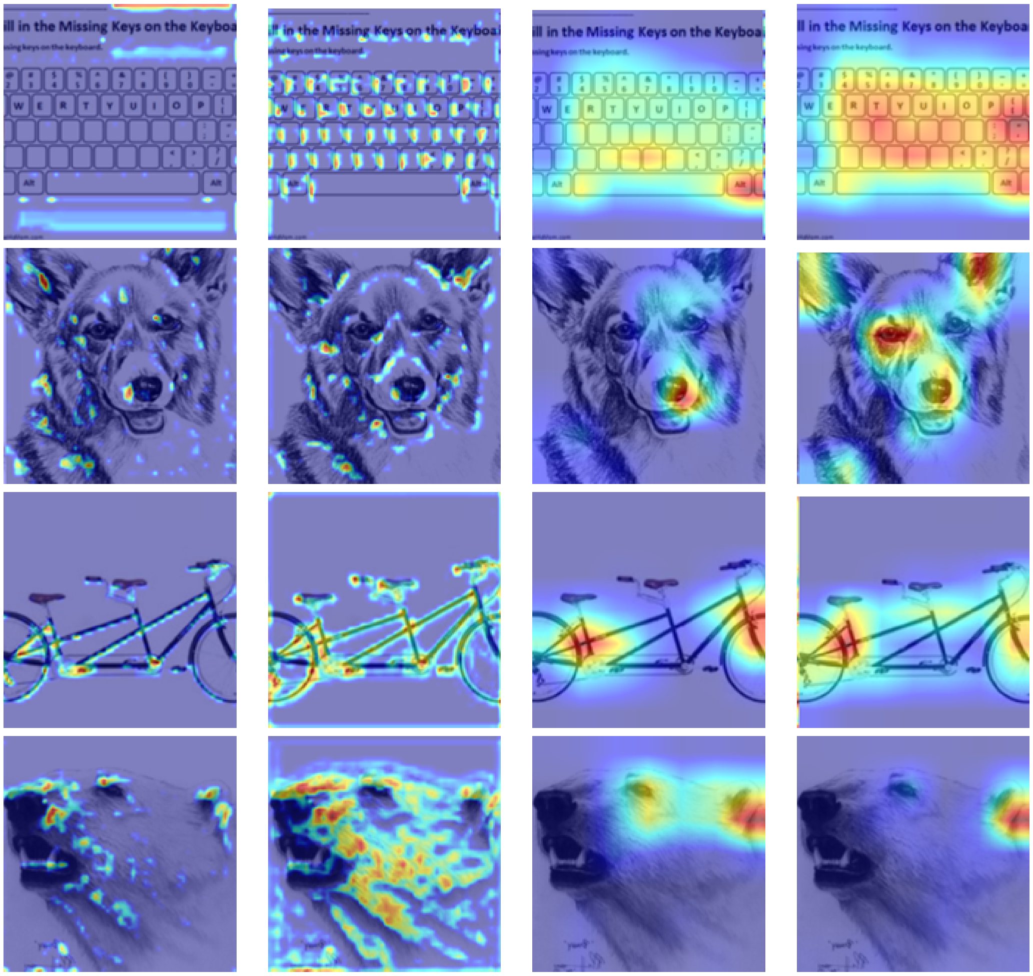
ill in the Missing Keys on the Keyboa
il in the Missing Keys on the Keyboa
ill in the Missing Keys on the Keyboa
ll in the Missing Keys on the Keyboa

Figure 13: Grad-CAM Visualizations Across Different Model Layers. This figure displays Grad-CAM visualizations related to the activation maps of our models. The first and third images from the left depict the shallow layers of the base model and our modified model, respectively. The second and fourth images illustrate the deep layers of the base model and our modified model, respectively. These visualizations highlight differences in focus areas between shallow and deep network layers under different model architectures.

Table 1: Top-1 and top-5 accuracy of the jointly trained model and baseline model on the OOD dataset across various categories.

Category	**Top-5 Acc (Joint)**	**Top-5 Acc (Baseline)**	**Top-1 Acc (Joint)**	**Top-1 Acc (Baseline)**
Cue-conflict	**58.67**%	55.16%	21.33%	**22.11**%
Edge	**65.00**%	64.38%	30.63%	**31.25**%
Silhouette	**49.38**%	46.25%	12.50%	**19.37**%
Sketch	**76.00**%	71.38%	**43.00**%	39.00%
Stylized	**52.50**%	51.50%	18.87%	**19.00**%
Colour	**83.28**%	76.41%	**52.50**%	45.23%
Contrast	**47.81**%	42.81%	**15.86**%	13.05%
Eidolon I	**49.77**%	47.03%	**19.61**%	17.11%
Eidolon II	**42.27**%	39.92%	**14.69**%	13.75%
Eidolon III	**40.94**%	38.36%	**13.67**%	12.97%
False-colour	**83.57**%	78.13%	**52.14**%	43.75%
High-pass	**55.39**%	48.44%	**22.58**%	20.31%
Low-pass	**43.36**%	41.64%	**17.19**%	15.16%
Phase-scrambling	**53.75**%	52.14%	**24.73**%	21.96%
Power-equalisation	**73.75**%	64.55%	**41.43**%	33.39%
Rotation	**67.97**%	60.39%	**31.87**%	28.44%
Uniform-noise	**39.61**%	36.72%	**11.72**%	10.00%

5 Reflection and Discussion

Our results show that the jointly trained model outperforms the classification-only model on all OOD filters for top-5 accuracy. However, it only outperforms two out of five shape-cue filters on top-1 accuracy, indicating three possible speculations: 1) There is a formation of a sub-network doing both tasks simultaneously. 2) The model is finding a shortcut to do the depth estimation task, possible shortcut cues would include the blurriness or out-of-focus of the image on the background. Such queues can help in doing the depth estimation task without learning the shape of the object. 3) Both models are over-fitting on the task and require more number classes to make the task more difficult.

In our future investigations, we intend to explore the formation of a sub-network using the methodology outlined in Zhang et al. (2023). Should such a sub-network exist, we propose the implementation of drop-outs across the entire encoder-network architecture to mitigate the emergence of such sub-networks. Additionally, we aim to substitute the depth density estimation task with the image reconstruction task, thereby preempting any shortcuts exploited by the model in executing auxiliary reconstruction. Furthermore, we plan to experiment with more intricate architectures such as the Transformer Vaswani et al. (2017), and HGRU Linsley et al. (2018) to harness the advantages of learning long-range dependencies across both tasks.

References

Harriet A Allen, Glyn W Humphreys, Jessica Colin, and Heiko Neumann. Ventral extra-striate cortical areas are required for human visual texture segmentation. *Journal of vision*, 9(9):2–2, 2009.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

Johannes Haushofer, Margaret S Livingstone, and Nancy Kanwisher. Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. *PLoS biology*, 6(7):e187, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Drew Linsley, Junkyung Kim, Vijay Veerabadran, Charles Windolf, and Thomas Serre. Learning long-range spatial dependencies with horizontal gated recurrent units. *Advances in neural information processing systems*, 31, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

Enyan Zhang, Michael A Lepori, and Ellie Pavlick. Instilling inductive biases with subnetworks. *arXiv preprint arXiv:2310.10899*, 2023.