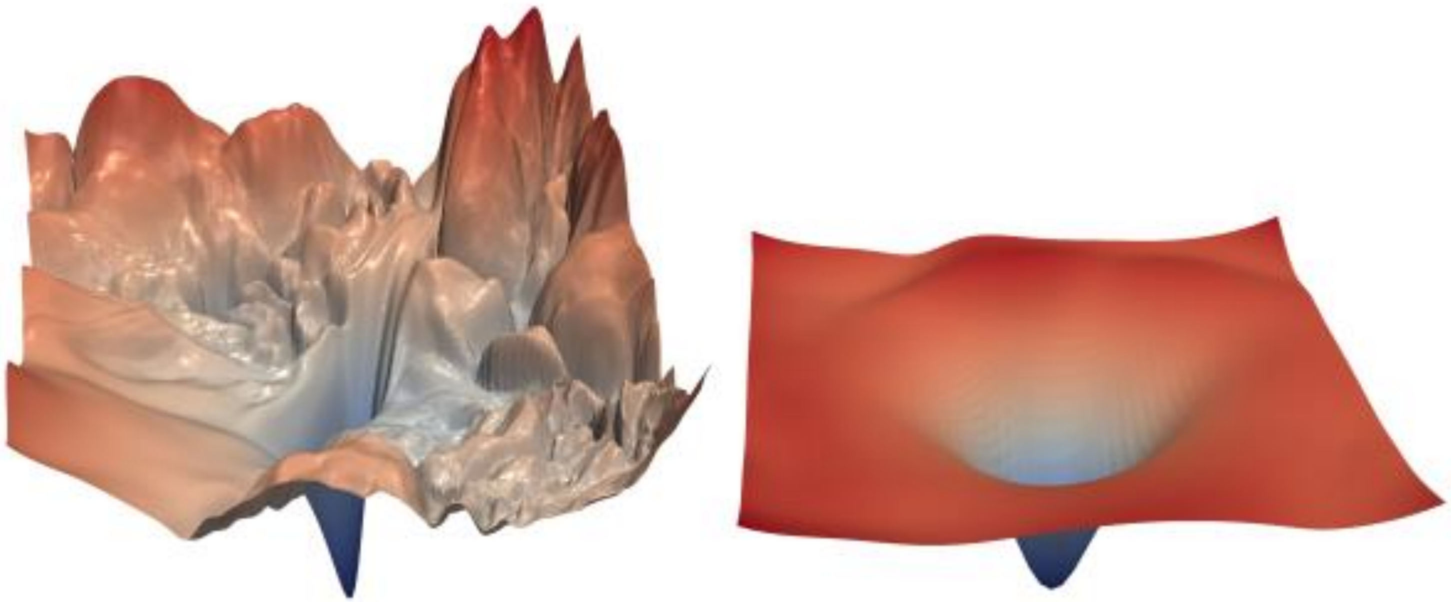
Depth
Trainability
Generalization



(a) without skip connections

(b) with skip connections

Li, Xu, Taylor, Studer, Goldstein, Visualizing the Loss Landscape of Neural Nets, NeurIPS, 2018

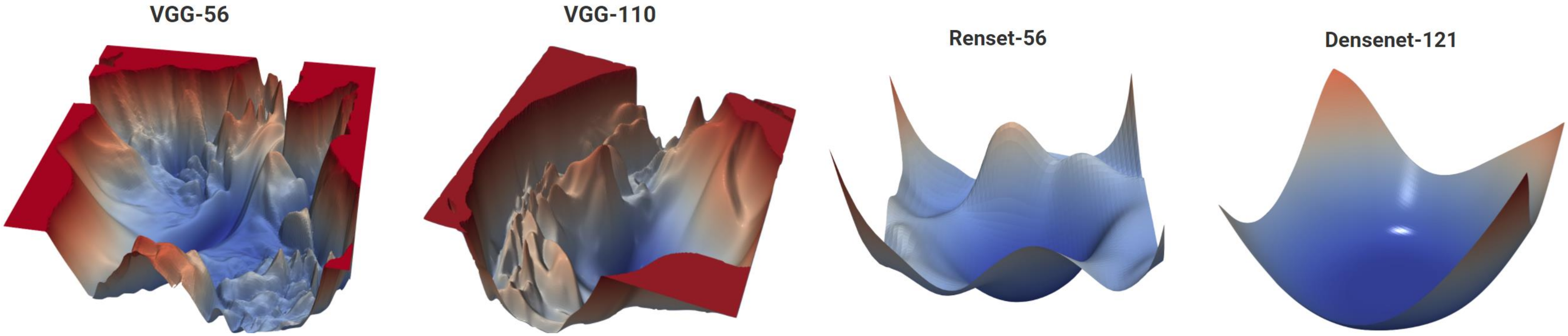
Trainability depends on model choices

- Neural network architecture
- Optimizer
- Initialization
- Hyperparameter choices

- Why residual connections make networks more trainable?

Smoothening the loss surface

- Adding skip connections makes the loss surface less rough
- Gradients more representative of the direction to good local minima
- Use visualizations with a grain of salt: dramatic dimensionality reduction!



The effect of depth

- Deeper architectures have more uneven, chaotic surfaces and many minima
- Removing skip connections fragments and elongates the loss surface
- Fragmentation requires good initialization
- Flatter minima accompanied by lower test errors

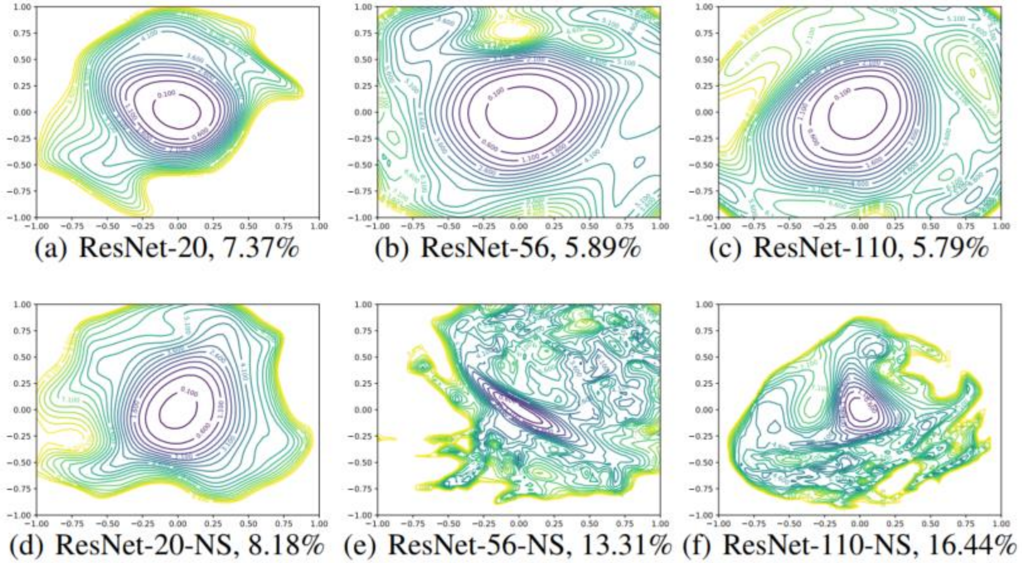


Figure 5: 2D visualization of the loss surface of ResNet and ResNet-noshort with different depth.

The effect of depth in wider architectures

- Similar conclusions when increasing width
- Width makes the loss surface even smoother and flatter

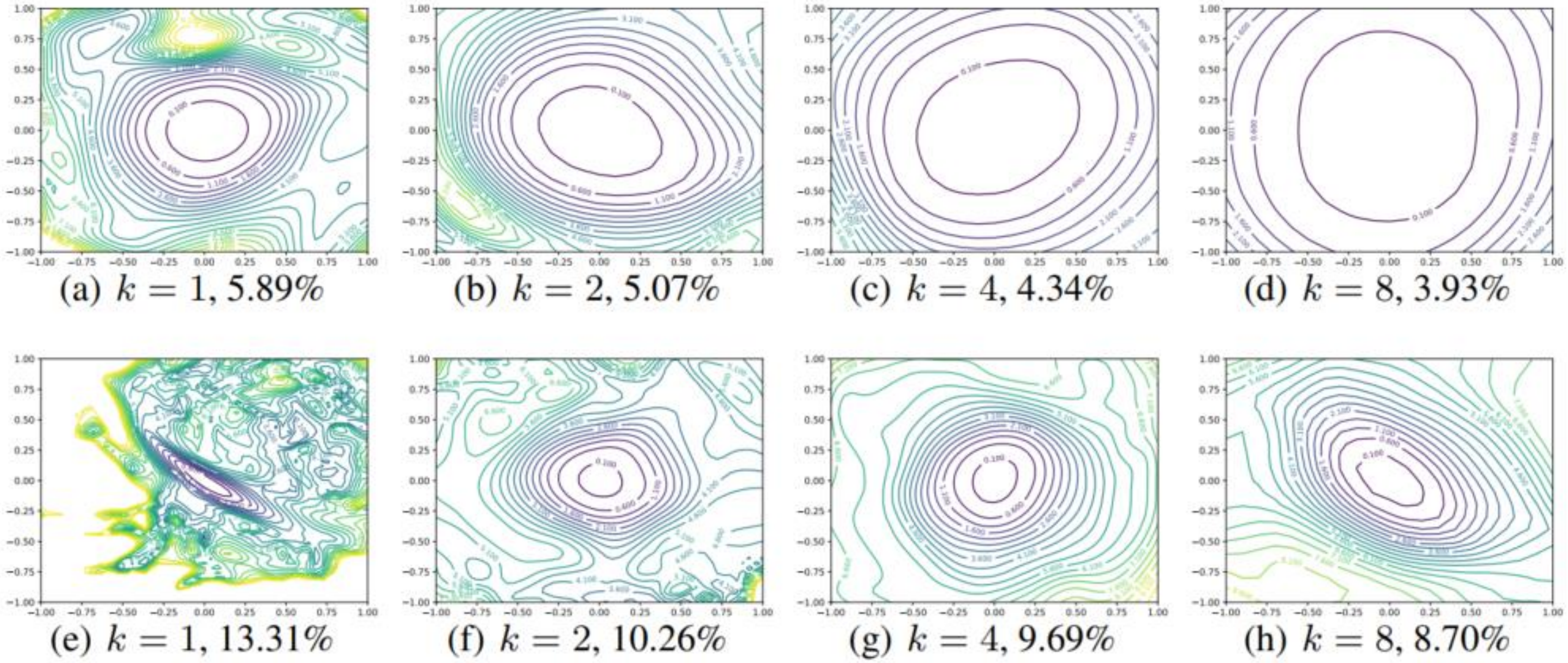


Figure 6: Wide-ResNet-56 on CIFAR-10 both with shortcut connections (top) and without (bottom). The label $k = 2$ means twice as many filters per layer. Test error is reported below each figure.

The effect of the optimizer

- Weight decay encourages optimization trajectory perpendicular to isocurves
- Turning off weight decay, the optimizer often goes in parallel with isocurves

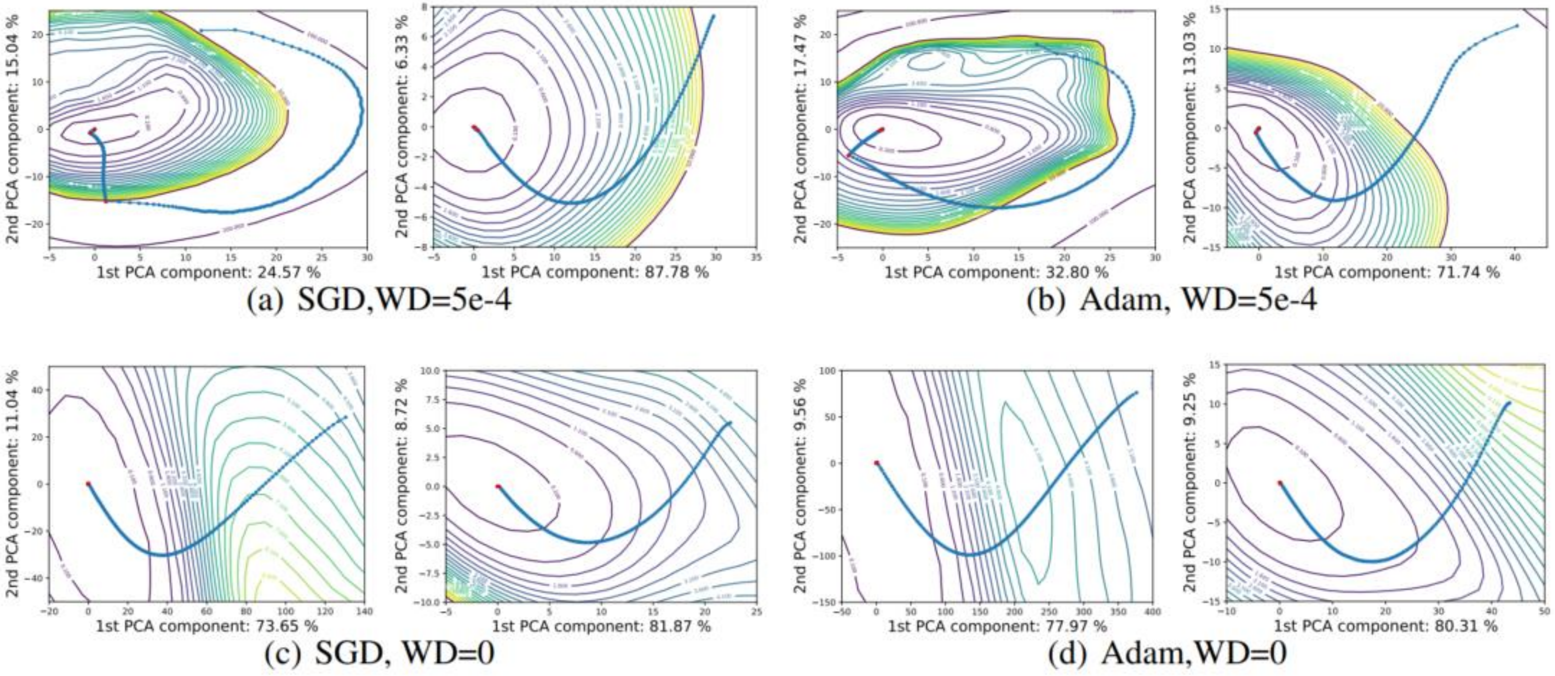


Figure 9: Projected learning trajectories use normalized PCA directions for VGG-9. The left plot in each subfigure uses batch size 128, and the right one uses batch size 8192.

Residual connections "stabilize" gradients

- The gradient with skip connection becomes

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{h}} \cdot \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{h}} \cdot \left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}} \right) = \frac{\partial \mathcal{L}}{\partial \boldsymbol{h}} \cdot \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{x}} + \frac{\partial \mathcal{L}}{\partial \boldsymbol{h}}$$

- The previous layer gradient is **carried to** the next module untouched
- Seen otherwise, the loss surface corresponds to stronger gradients, *i.e.*, smoother