

How Do Vision Transformers Work?

Namuk Park^{1,2}, Songkuk Kim¹

¹Yonsei University, ²NAVER AI Lab

Paper: <https://arxiv.org/abs/2202.06709>

Code: <https://github.com/xxxnell/how-do-vits-work>

What Properties of Self-Attentions Do We Need?

MSAs (multi-head self-attention) have flat but non-convex losses. In contrast, **Convs** have convex but sharp losses.

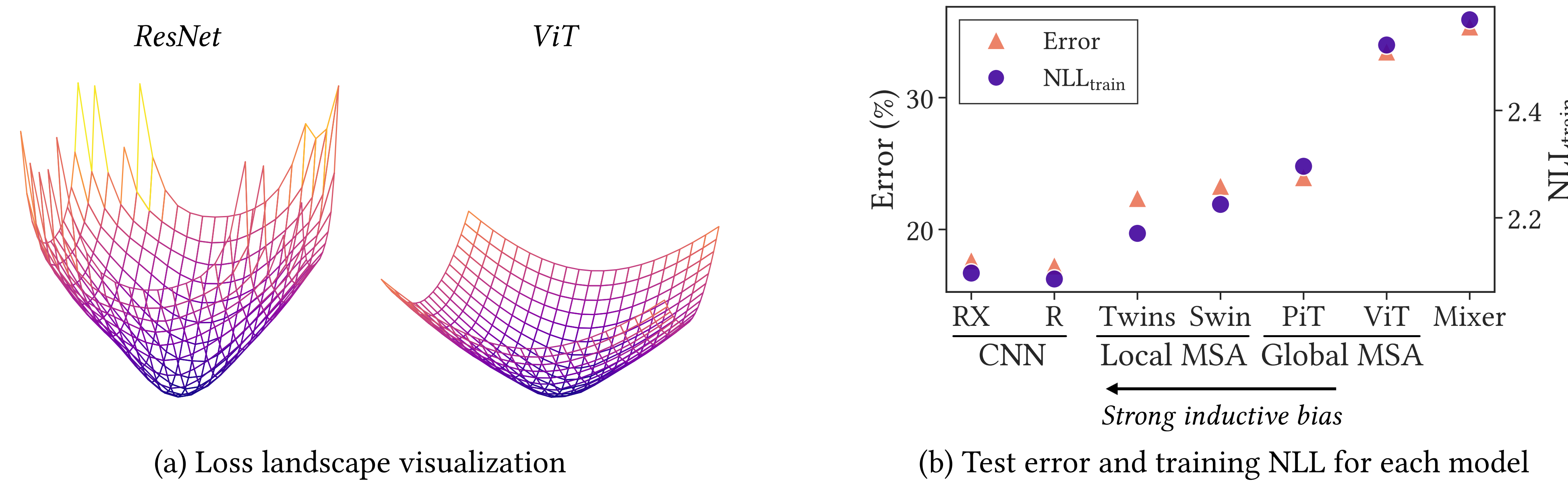


Figure 1: **Left: Loss landscape visualization (Left)** show that **ViT** has a flatter loss than **ResNet**. **Right:** Weak inductive bias (e.g. long-range dependency) disturbs NN optimization.

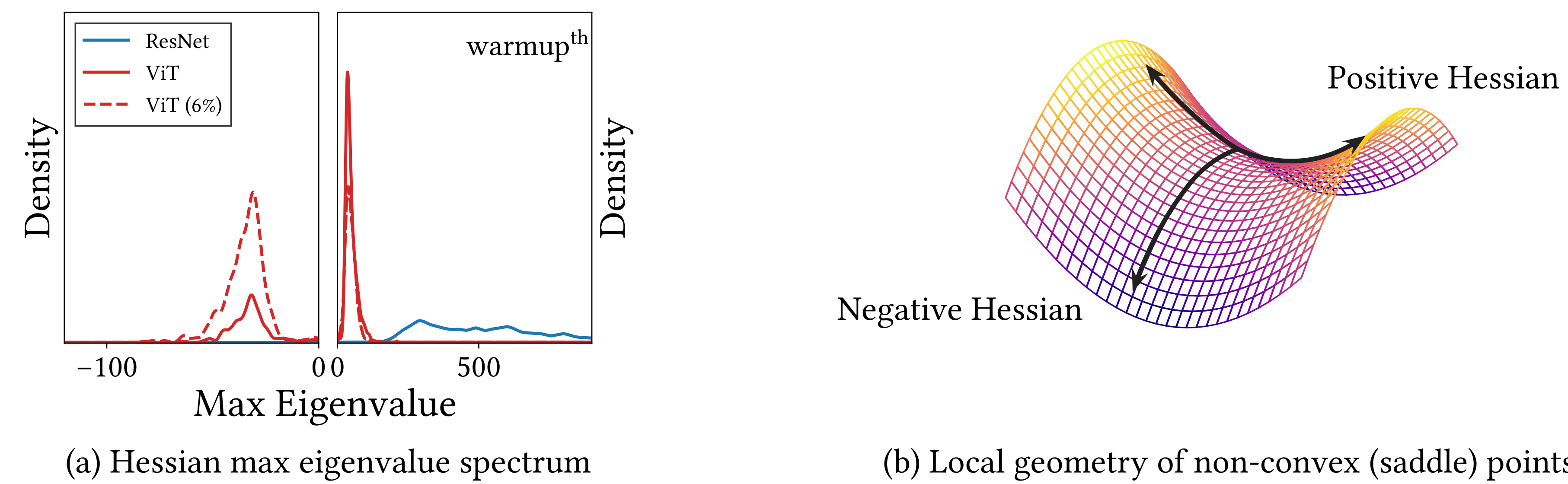


Figure 2: **Hessian max eigenvalue spectra** show that MSAs have their pros and cons. **ViT** has a number of negative Hessian eigenvalues, while **ResNet** only has a few. The magnitude of **ViT**'s positive Hessian eigenvalues is small.

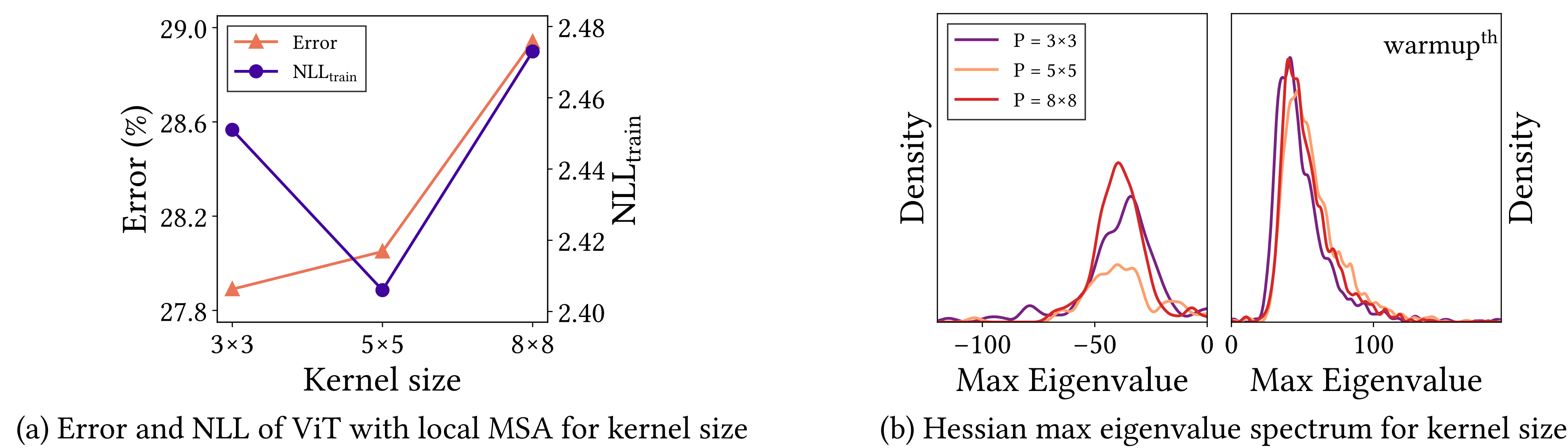


Figure 3: The key feature of **MSA** is data specificity, not long-range dependency. **Left:** Convolutional ViT demonstrates that locality constraint improves ViT. **Right:** Locality inductive bias suppresses the negative Hessian eigenvalues.

Do Self-Attentions Act Like Convs?

MSAs are low-pass filter, but **Convs** are high-pass filter. It suggests that MSAs are shape-biased, whereas Convs are texture-biased.

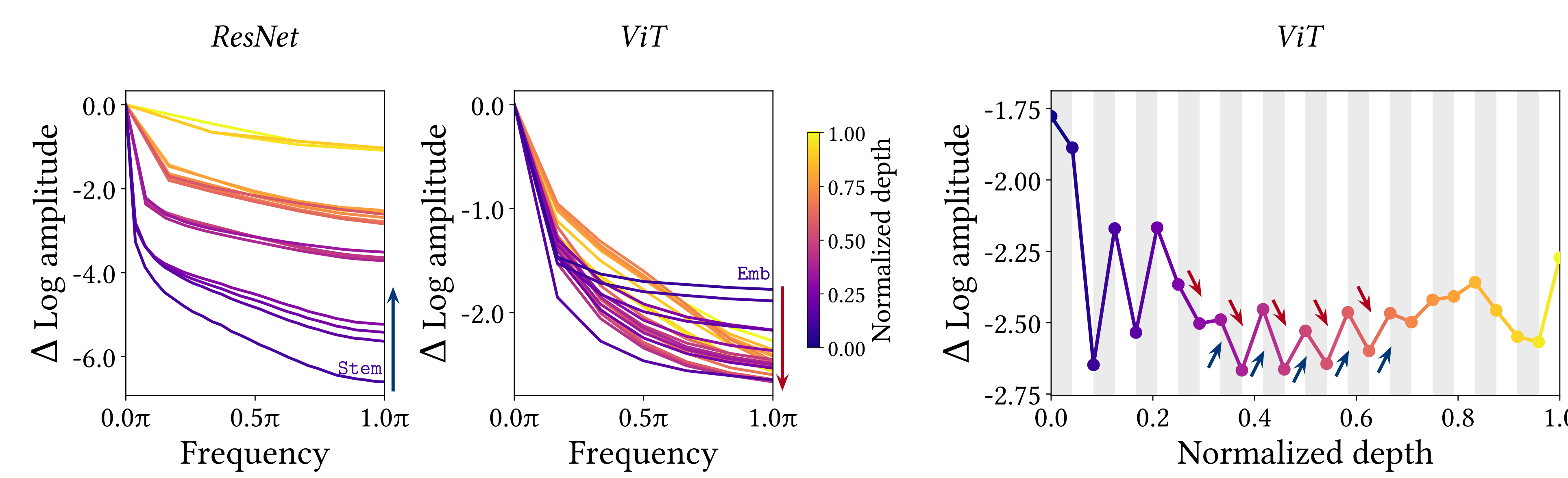


Figure 4: Relative log amplitudes of **Fourier transformed feature map** show that **ViT** tend to reduces high-frequency signals, while **ResNet** amplify them. **Left:** In ViT, **MSAs** (gray area) generally reduce the high-frequency (1.0π) component of feature map, and **Conv/MLPs** (white area) amplify it.

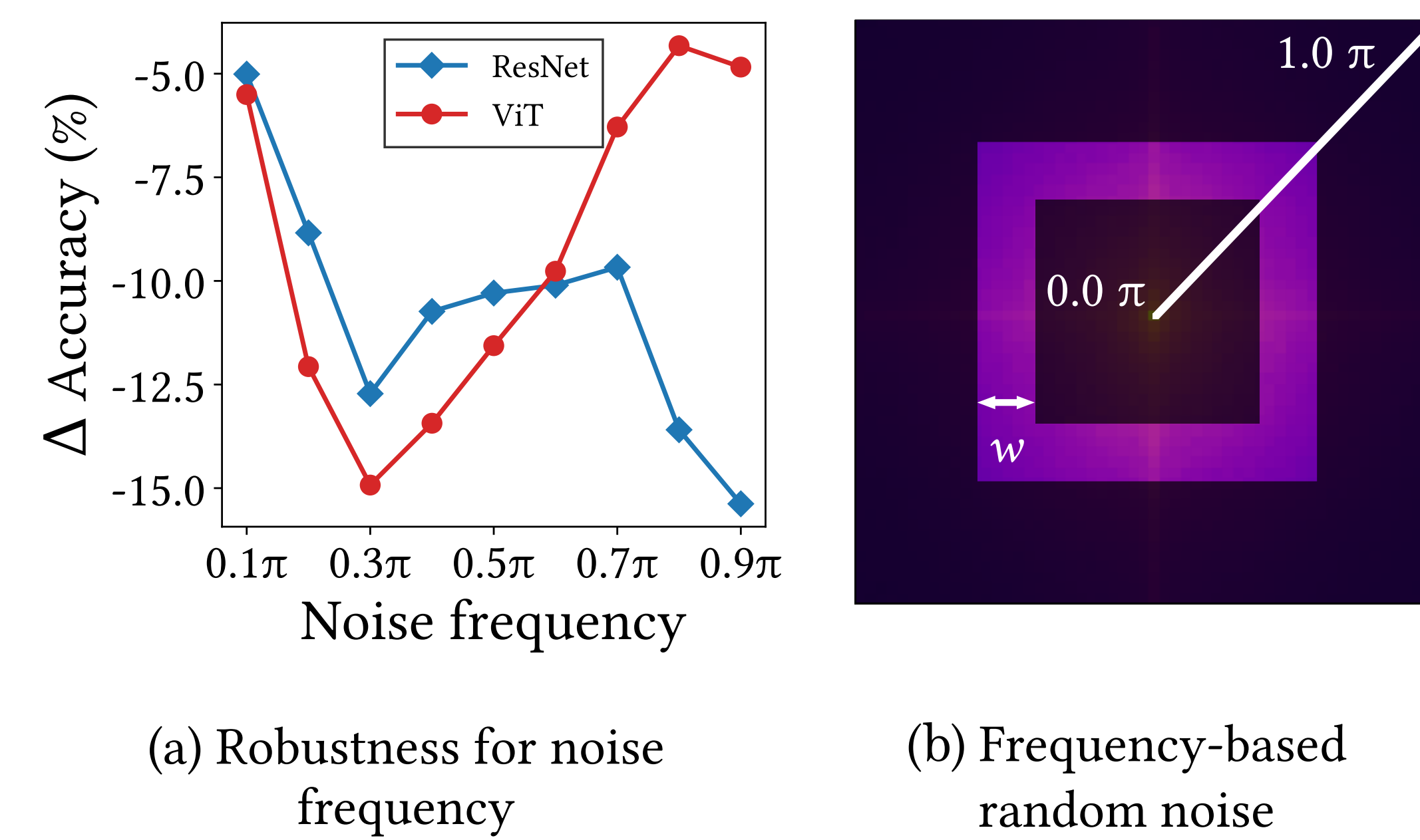


Figure 5: We measure the **decrease in accuracy against frequency-based random noise**. **ViT** is robust against high-frequency noise, while **ResNet** is vulnerable to them.

It suggests that low-frequency signals and high-frequency signals are informative to MSAs and Convs.

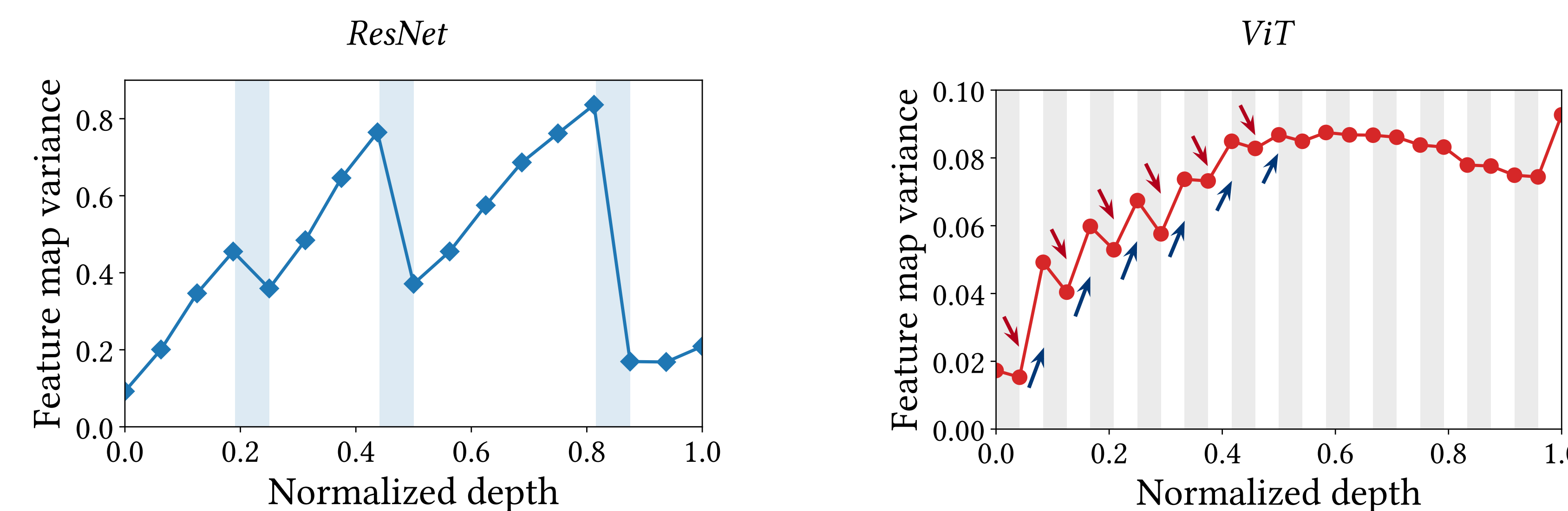


Figure 6: **MSAs** (gray area) reduce **the variance of feature map points**, but **Convs/MLPs** (white area) increase the variance. The blue area is subsampling layer. The results implies that **MSAs** aggregate feature maps, and **Convs** convert them.

How Can We Harmonize Self-Attentions with Convs?

MSAs closer to the end of a stage (not a model) and **Convs** at the beginning of a stage significantly improve the performance.

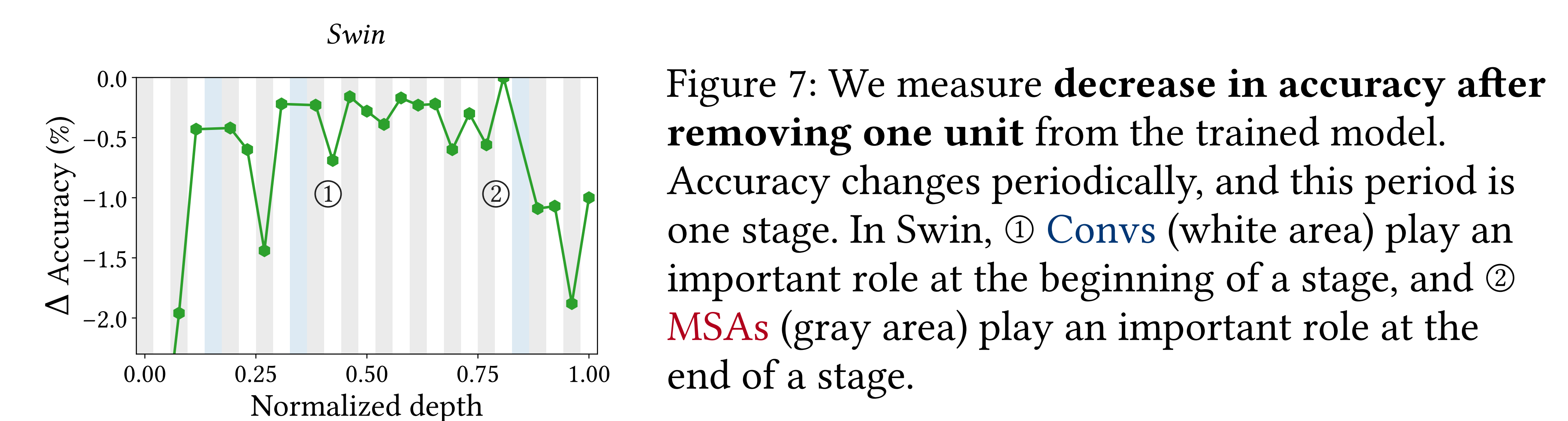


Figure 7: We measure **decrease in accuracy after removing one unit** from the trained model. Accuracy changes periodically, and this period is one stage. In Swin, ① **Convs** (white area) play an important role at the beginning of a stage, and ② **MSAs** (gray area) play an important role at the end of a stage.

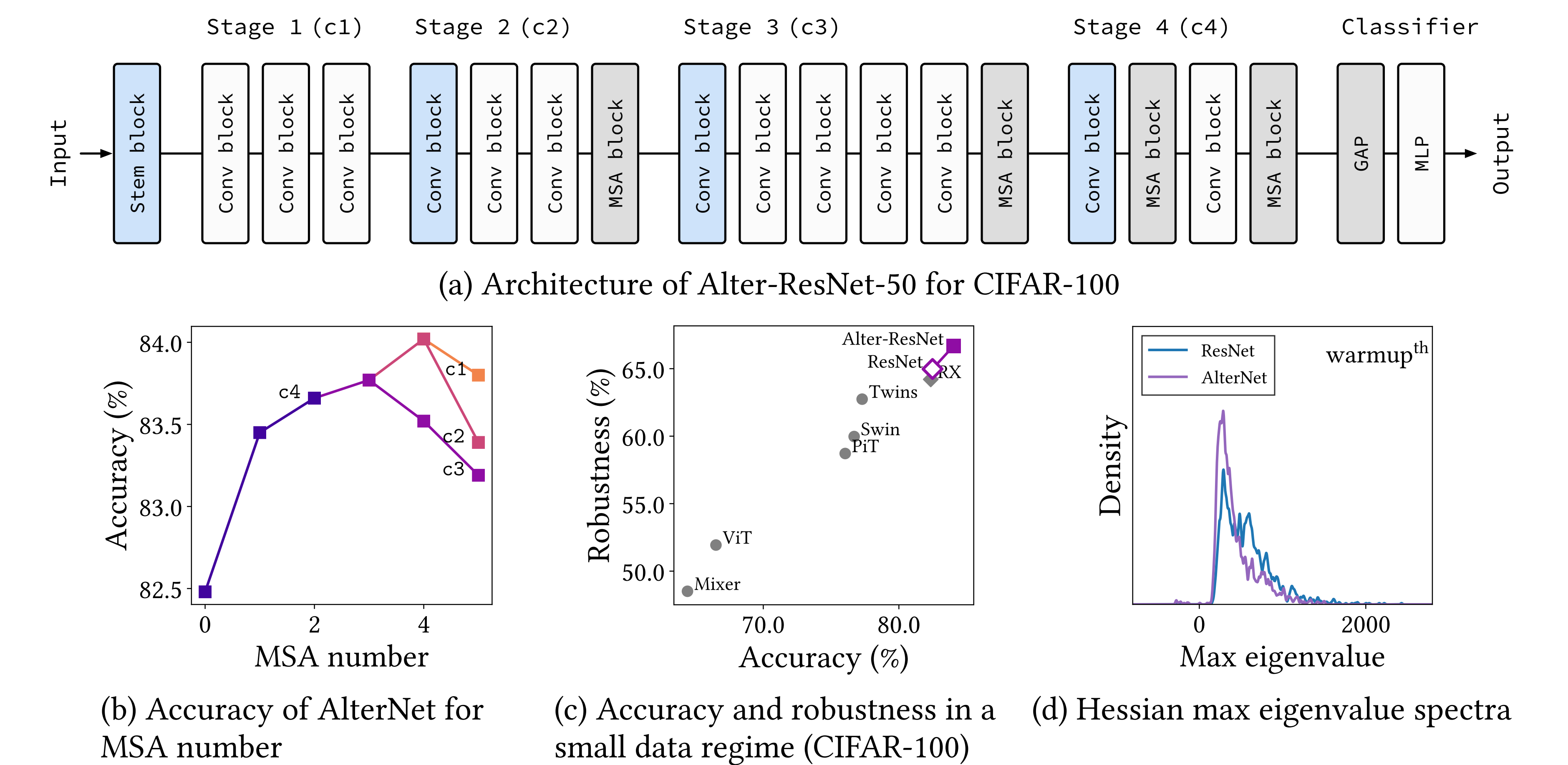


Figure 8: We propose **AlterNet**, a model in which **Conv** blocks at the end of a stage are replaced with **MSA** blocks. AlterNet outperforms CNNs even in small data regimes.

In summary, appropriate inductive biases improves NN optimization, and self-attentions have a spatial smoothing inductive bias.

	Self-Attention	Convolution
Loss Landscape	Flat but non-convex	Convex but sharp
Fourier Analysis	Low-pass filter (shape-biased)	High-pass filter (texture-biased)
Best Practice	The end of a stage	The beginning of a stage