Semantic Complexity vs Spatial Frequency, a Perceived Tension in Objectives

James Chen

May 2025

There is a subtle interplay along these two axes. Humans naturally ignore higher frequency patterns, and in practice generally more semantically meaningful regularities have lower frequency. However I'd argue that this is a coincidence spawning from the human viewpoint, rather than some direct causal relationship. In the extreme case, biological cell composition of a blade of grass is very meaningful, it's just that most of the time, its state has less meaning to a human than a field of grass. A hard bias that discards all high-frequency content is undesirable. Instead I believe that we should capture *all* statistically-regular patterns. High frequency patterns can have semantics that we want to capture. In some sense, a model which captures all regularities is a model which is thoroughly descriptive, and needed for an ideal model.

Thesis. As far as I'm aware, pixel-space methods like denoising/reconstruction are the only ones we know which preserve high-frequency patterns well. In my opinion, the deterioration in representation quality seen in pixel-reconstruction pre-training is not due to focusing on high-frequency detail itself, but to the indiscriminate L2 loss that cannot tell structured signal from unstructured noise when attempting to learn regularities at higher-frequencies, and thus creates a soft-ceiling on learned representations.

Pixel-space reconstructive objectives. The most straightforward way to optimize for all regularities are pixel-space reconstructive objectives [He et al., 2021], which aim to capture all patterns (and thus regularities), however these produce brittle downstream representations [Assran et al., 2023]. Pixel-space objectives treat every pixel as equally reliable; in practice this forces them to chase artefacts and sensor noise, yielding brittle features.

Note that in objectives which optimize directly on pixel space (e.g., denoising, reconstruction), networks exhibit both spectral bias (preferring low spatial frequencies) [Rahaman et al., 2018] and simplicity bias (preferring semantically simpler features) [Shah et al., 2020]. This is actually quite convenient, as noise mostly lives at higher frequencies, so during the earlier bits of training, the learning of lower frequencies and lower semantic complexity regularities do not greatly conflict, as lower frequencies likely just have higher signal-to-noise in their error signal for pixel space reconstruction, and most semantics are low frequency too.

However, as the model starts learning harder-to-fit patterns, it likely has to learn higher semantic complexities as well as higher spatial frequency reconstruction at the same time with an error signal that likely has poor signal-to-noise (as most noise lives at higher spatial frequencies). At this stage the residual loss is driven by three distinct sources together:

1. High-frequency structure that is still predictable,

2. high-frequency idiosyncrasies that are effectively noise,
3. semantically complex long-range interactions of all frequencies.

Because the loss treats them identically, the network is forced to spend capacity on all three at once.

We cannot learn to ignore noise, as we have to reconstruct it, and so we have to spend arbitrary amounts of model capacity trying to fit it. This empirically degrades the overall representation quality of the model a lot [Assran et al., 2023].

A way to think about this is that noise often sets a **ceiling in effective capacity usage** for pixel-space reconstructive objectives, after which every extra bit of regularity we attempt to learn takes a significant hit from the noise. The optimal predictor for unpredictable noise is its expected value, so after a while the loss term is dominated by variance we can never remove; this keeps gradient signal alive on irrelevant pixels and makes it hard to learn more useful information.

For illustrative purposes, we would like the model to learn progressively higher-level semantic regularities (irrespective of frequency) first and only later allocate capacity to idiosyncratic, hard-to-predict patterns — the kind it still needs to reproduce faithfully when the downstream task demands pixel-accurate recall (e.g. reference-conditioned image generation). Because the reconstruction loss cannot distinguish those idiosyncrasies from pure sensor noise, its gradient begins to corrupt learning far earlier than we would like.

This phenomenon is often interpreted as "caring about high frequency details too much", i.e., we drop semantics in favor of high frequency details. However, I'd argue this may not be the entire story. Rather, while attempting to capture high complexity/frequency regularities with pixel-space reconstruction, we just so happen to degrade our representations due to said objectives also incentivizing the capturing of noise which starts degrading representations after we learn low frequency regularities and most semantic regularities.

Objectives which filter out high frequencies. This often leads to objectives for visual understanding which are often effective because they are biased to filter out high frequency components to varying degrees, while still being able to capture a lot of useful semantic information (as semantic information is often low frequency.) In some sense, this is a form of early stopping with respect to the frequency of the regularities we capture.

CLIP [Radford et al., 2021] retains information only useful for captioning (which is often low frequency,) and is more susceptible to lower frequency noise (vs higher frequency noise) De Rosa et al. [2024] (i.e., it likely filters out a lot of higher frequency patterns.)

Though DINOv2 [Oquab et al., 2023] can pick up fine-grained or high-frequency cues when they dominate an entire 14×14 patch, its patch token still collapses all intra-patch positions into a single vector. The features are therefore semantically rich but spatially coarse, largely discarding where those local details sit inside each patch (so it's bad at intra-patch spatial reasoning.)

In both of these cases, the objective is free to filter out a lot of high frequency patterns (a lot of which is noise) to learn more complex semantics. This raises the ceiling in effective capacity usage.

I don't believe that this approach to objective design is entirely correct. There actually isn't an issue with capturing all meaningful high-frequency regularities, it's just that pixel-space reconstruction (which is really the only way we know how to do so) enjoys capturing noise when it tries to learn higher-frequency and more complex regularities.

Better representations for understanding. Vision encoders shouldn't inherit human bias toward low-frequency patterns. While pixel-space reconstruction objectives (like MAE) fail by forcing models to reconstruct noise alongside meaningful patterns, the common fix of designing objectives with the bias of filtering out high-frequency components is misguided. The real problem isn't paying too much attention to high frequencies—it's that methods that allow us to learn high frequency regularities (so far to my knowledge only pixel-space reconstruction methods) are often sensitive to unpredictable noise. Thus the goal should be designing objectives which are resistant to noise, but can still capture all regular patterns regardless of frequency.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023.

Vincenzo De Rosa, Fabrizio Guillaro, Giovanni Poggi, Davide Cozzolino, and Luisa Verdoliva. Exploring the adversarial robustness of clip for ai-generated image detection, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

Maximilien Oquab et al. DINOv2: Learning robust visual features without supervision, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks, 2018.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks, 2020.