Abstract

Image generation has been a long sought-after but challenging task, and performing the generation task in an efficient manner is similarly difficult. Often researchers attempt to create a "one size fits all" generator, where there are few differences in the parameter space for drastically different datasets. Herein, we present a new transformer-based framework, dubbed StyleNAT, targeting high-quality image generation with superior efficiency and flexibility. At the core of our model, is a carefully designed framework that partitions attention heads to capture local and global information, which is achieved through using Neighborhood Attention (NA). With different heads able to pay attention to varying receptive fields, the model is able to better combine this information, and adapt, in a highly flexible manner, to the data at hand. StyleNAT attains a new SOTA FID score on FFHQ-256 with 2.046, beating prior arts with convolutional models such as StyleGAN-XL and transformers such as HIT and StyleSwin, and a new transformer SOTA on FFHQ-1024 with an FID score of 4.174. These results show a 6.4% improvement on FFHQ-256 scores when compared to StyleGAN-XL with a 28% reduction in the number of parameters and 56% improvement in sampling throughput. In addition, we demonstrate a visualization technique for viewing the attention maps for local attentions such as NA and Swin. Code and models will be open-sourced at: https://github.com/SHI-Labs/StyleNAT.

1. Introduction

Adversarially trained generative image modeling has been long dominated by Convolutional Neural Networks (CNNs). More recently, however, there have been strides in shifting towards new architectures, such as Transformer-based GANs [58, 59] and Diffusion-based GANs [54]. In addition to this, other generative modeling methods, such as pure Diffusion [7, 48] and Variational Autoencoders [5, 15], are catching up to the standards set by GANs. Diffusion and VAEs have some advantages in that they are able to approximate the densities of the training distribution, and thus can incorporate more features and can have a higher recall. Both these methods have major disadvantages in that they are more computationally expensive when compared to GANs, with diffusion models being known for slower inference even after training.

While convolutional models have the advantage of utilizing local receptive fields, transformer-based models, with self attention mechanisms, have the advantage of utilizing a global receptive field. Transformers hope to resolve issues that convolutional based models have such as heterochromia, inconsistent pupil size, and other effects that aren’t well captured by metrics such as FID and Inception Score. Unfortunately, self attention bears with it a quadratically growing memory and FLOPs, for transformers means that it may simply be computationally infeasible to apply a standard transformer to the entirety of the latent structure. These are some of the reasons early transformer-based GANs had difficulties competing with convolutional based ones [23, 20, 19].
Quite recently, there has been a push for restricted attention mechanisms, which typically do not pay attention to the entire feature space. By doing this, these transformers have a reduced computational burden and also put pressure on fixed-size receptive fields, which can help with convergence. The most popular of these, Swin Transformer [36, 35], uses a shifted window self attention (WSA) mechanism. This shifted window breaks up the global attention into smaller non-overlapping patches, which results in a linear time and space complexity. Swin also follows every WSA with a shifted variant, which shifts the dividing lines to allow for out-of-patch interactions. Unfortunately Swin has issues with translation and rotation [12]. StyleSwin [58] was one of the earliest transformer-based GANs that caught up to its convolutional counterparts. Its style-based generator [29] was built upon Swin’s WSA attention mechanism. Despite its success, it failed to excel beyond existing state-of-the-art GANs at the time.

Localized attention mechanisms come with other costs though: the inability to attend globally and therefore capture long-range dependencies. This led to works such as Dilated Neighborhood Attention (DiNA) [11], which was proposed to tackle this issue in models based on localized self attention. This work further extends Neighborhood Attention (NA) [12], which introduced an efficient sliding-window attention mechanism that directly localizes self attention for any point to its nearest neighbors. NA also resolves many of the shift and rotation issues that Swin has. Due to the flexibility which sliding windows provide, NA/DiNA can easily be dilated to span the entire feature map and capture more global context. Within this work we seek to extend the power of the attention heads of NA/DiNA by partitioning them and allowing them to incorporate different vantage points across the latent structure for high-resolution image generation. While this method, called Hydra-NA, we greatly extend the capabilities of previous local attention mechanisms with almost no additional computational cost on top of NA/DiNA. Through these partitions and the use of various kernels and dilations, these heads are able to incorporate a much more feature rich landscape into their structure, thus increasing the information gain that they can obtain on a given latent structure. Our Hydra-NA module increases the flexibility of neighborhood attention modules, allowing them to be better adapted to the data they are trying to learn on various generation tasks.

Our paper’s contributions are:

- We introduce Hydra-NA, which extends NA/DiNA and provides a flexible design to combine local and long-range receptive fields via arbitrary partitioning of the attention heads.
- We propose StyleNAT, an efficient and flexible image generation framework that is not only memory and compute efficient but also adaptive to different datasets and tasks via neighborhood attention and its dilated variants with our Hydra-NA design.

- We achieve new state of the arts on FFHQ-256 (among all generative models) and FFHQ-1024 (among all transformer-based GANs).
- We develop a method to visualize the attention maps of these localized attention windows, for both Swin and NA.

We are able to accomplish all this with significantly fewer computational resources than many of our competitors as well as minimal tuning of our hyper-parameters.

2. Related Works

In this section, we will briefly introduce attention modules introduced in Swin Transformer [36], NAT [12], and DiNAT [11]. We will then move on to style-based generative adversarial models, particularly StyleGAN [29] and StyleSwin [58].

2.1. Attention-based Models

The Transformer [53] is arguably one of the most prevalent architectures in language processing. The architecture simply consists of linear projections and dot product attention, along with normalization layers and skip connections. Other than large language models directly using this architecture [42, 6], the work inspired research into models built with dot product self attention [41, 43]. Later in 2020, Vision Transformer [8] applied a plain transformer encoder to image classification, which was outperformed existing CNNs at large scale classification. This inspired many researchers to study Transformers as direct competitors to CNNs in different settings [50, 51, 14] and across different vision tasks [33, 16].

2.1.1 Swin Transformer

Liu et al. [36] proposed Window Self Attention (WSA) and Shifted Window Self Attention (SWSA), both of which partition feature maps into windows of fixed size, and apply self attention to each window separately. The difference between the regular and shifted variants is that the latter shifts the partition boundaries by shifting pixels, therefore allowing out-of-window interactions and receptive field growth. Because of the fixed window size, self attention’s quadratic complexity drops to a linear complexity. Through these mechanisms, they propose a hierarchical transformer model, which is a stack of 4 transformer encoders with self attention replaced with WSA and SWSA (every other layer...
Figure 2: Architecture of StyleNAT. We follow the standard StyleGAN architecture. The Mapping Network takes in a random normal latent variable of dimension 512, followed by affine functions, that feed into a Synthesis Network, which has a latent learnable parameter that is initialized with normal noise of shape $[512, 4, 4]$. Differing from StyleGAN, we replace the convolutions with transformers, using Neighborhood Attention. The first step, size $4 \times 4$, we use a Multi-Headed Self Attention, as this is approximately equivalent and computationally more efficient. Our Hydra-NA architecture, to the right, gives us the unique ability to have arbitrary kernel sizes, dilations, and strides unique to each attention head. This gives our model additional flexibility over the receptive field while maintaining computational efficiency. The full code is provided in Appendix B.

2.1.2 Neighborhood Attention Transformer

Neighborhood Attention (NA) [12] was proposed as a direct restriction of self attention to local windows. The key difference between NA and WSA is that the restriction is pixel-wise, leading to each pixel attending to only its nearest-neighboring pixels. The resulting attention spans would be in theory similar to how convolutions apply weights, with the exception of how cornering pixels are handled. Compared to WSA, in addition to introducing locality [36], NA also maintains translational equivariance. NA also approaches self attention itself as its window size grows, and unlike WSA, would not need pixel shifts as it is a dynamic operation. Similar operations, in which self attention is restricted in a token-wise manner, had been investigated prior to this work [43], but were less actively studied due to implementation difficulties [43, 52, 36]. To that end, Neighborhood Attention Extension (NATTEN) [12] was created as an extension to PyTorch, with efficient CUDA kernels, which allow NA to run even faster than WSA, while using less memory. The model built with NA, Neighborhood Attention Transformer (NAT) [12], was shown to have superior classification performance compared to Swin, and competitive object detection, instance segmentation, and semantic segmentation performance.

2.1.3 Dilated Neighborhood Attention Transformer

Dilated Neighborhood Attention Transformer (DiNAT) [11] followed NAT [12] by extending NA to Dilated NA (DiNA), which allows models to use extended receptive fields and capture more global context, all with no additional computational cost. DiNA was added to the existing NATTEN package and the underlying CUDA kernels, which allowed the easy utilization of this attention pattern. By simply stacking NA and DiNA layers in the same manner that Swin stacks WSA and SWSA, DiNAT models significantly outperformed Swin Transformer [36], as well as NAT [12] across multiple vision tasks, especially downstream tasks. DiNAT also outperformed Swin’s convolutional competitor, ConvNeXt [37], in object detection, instance segmentation, and semantic segmentation.

2.2. Style-based GANs

Style based generators have long been the de facto architecture choice for generative modeling. In this section we...
discuss different relevant generators and their relationship to our network.

2.2.1 StyleGAN

Karras et al. first introduced the idea of using progressive generating using convolutions where the one would begin with a small latent representation, process each layer, and then scale up and repeat the process until the desired image size is achieved [25]. This architecture has major advantages in that it reduces the size of the latent representation, and thus the complexity of data that the model needs to learn. This comes with the obvious drawbacks that we make a trade-off of diversity for fidelity and computational simplicity. Karras et al. later produced several papers that built on this idea [29, 30, 26, 28], improving the progressive network, but also introducing a sub-network, called the “style network”, that allows one to control the latent representations at each of these levels. This innovation dramatically increased the usability of this type of network architecture because it allows for simpler feature interpolation, which is often difficult in implicit density models. Because of the high fidelity, flexibility, and the ability to perform feature interpolation, this style of network has become the de facto choice for many researchers [1, 21, 22, 24].

2.2.2 StyleSwin

StyleSwin [58] introduced Swin based transformers into the StyleGAN framework, beating other transformer based models, but still not achieving any state of the art results. This work did show the power of Swin and how localized attention mechanisms are able to out perform many CNN based architectures and did extend the state of the art for style based networks. StyleSwin achieved this not only by replacing the CNN layers with Swin layers, but also partitioned the transformer heads into two, so that each half could operate with a different kernel. This was key to improving their generative quality, allowing the framework to increase flexibility and pay attention to distinct parts of the scene at the same time. StyleSwin also included a wavelet based discriminator, as was introduced in SWAGAN [9] and also used by StyleGAN3 [28], which all show improvements for addressing high-frequency content that causes artifacts. StyleSwin uses a constant window size of 8 for all latents larger than the initial $4 \times 4$, wherein they use a window size of 4. We follow a similar structure except replace the initial layer with MHSA for efficiency.

3. Methodology

Our StyleNAT network architecture closely follows that of StyleSwin [58], which closely follows the architecture of StyleGAN2. Herein we described the differences in architecture and the roles that they play in generating our images. We describe how our network can be adapted to many different tasks and how to choose the suitable configurations. The simplicity of our design is its innovation, demonstrating the powerful role of Neighborhood Attention and the partitioning of attention heads. These changes allow our model to capture both long range features, which help generate good structure, as well as the local features, allowing for attention to detail akin to convolutions. We demonstrate this through visual inspection, in Appendix C, and our attention maps, in Appendix D.

3.1. Motivation

Current CNN based GANs have some limitations that make it difficult for them to consistently produce high quality and convincing images. CNNs have difficulties in capturing long range features but do have strong local inductive biases and are equivariant to translations and rotations. Thus we need to incorporate architectures that maintain this equivariance, while having both local inductive biases and global inductive biases.

Transformers on the other hand have strong global inductive biases, are translationally equivariant, and are able to capture long range features across an entire scene. The downside of these is that they are computationally expensive. Building a pure transformer-based GAN would be infeasible, as the memory and computation required to generate attention weights grow quadratically and poorly scales with resolution. Neighborhood Attention gives us a transformer that is quasi-equivariance to translations and rotations, with minimal aliasing.

3.2. Hydra-NA: Different Heads, Different Views

At the core of StyleNAT is the flexible Hydra-NA architecture. Inspired by the multi-headed design in Vaswani et al. [53], we sought to increase the power of each head, by giving them new ways to view their landscape. While global attention heads can learn to pay attention to various parts of a scene, the reduced receptive field of localized attention mechanisms means that these perspectives can only be performed on that limited viewpoint.

To extend the power of these attention heads, StyleSwin splits attention heads into halves. One half uses a windowed self-attention (WSA) and the other uses a shifted window self-attention (SWSA). This method has limitations in that it still is unable to perceive the global landscape and is limited in the perspectives these heads can take on, even if given more partitions. Swin does not maintain translational and rotational equivariance [12, 52] and the partitions further exacerbate these problems. Hydra-NA takes a different approach to achieve a similar goal, by allowing each of the attention heads to have a unique perspective of the latent
3.3. StyleNAT

We use the same root style based architecture as StyleGAN to incorporate all the advantages of interpretability but seek to increase the inductive powers of our model through the use of Hydra-NA. At the first resolution, $4 \times 4$ we use Multithreaded Self Attention, as it is more computationally efficient and there is not much to learn within such a small feature space. Subsequent layers utilize NAT, and in general we choose a kernel size of 7 but differ in the corresponding dilations and partitioning.

In this respect our partitions have both dense kernels as well as sparse kernels, and even allow for progressive sparsity. With this framework we choose two main designs: a dual partition with a local/dense kernel combined with a global/sparse kernel and a progressive dilation design – increasing sparsity and the receptive field with the number of partitions.

4. Experiments

There is a large variance in dataset choice for evaluating generative models as well as vastly different model sizes and generation throughput. These issues, along with the limitations of evaluation metrics, make it difficult to compare models in a fair manner and in a way aligned with our goal: generating realistic imagery. Given our limited compute we believe FFHQ (Section 4.1) and LSUN Church (Section 4.2) are good proxies to the different problem types while maximizing the number of comparators under our constraints. These datasets are common GAN evaluation benchmarks that can help us understand how the generator works on more regular data as well as high variance data.

We do minimal hyper-parameter searching and generally keep hyper-parameters similar to StyleSwin. This means that we use the Two Time-Scale Update Rule (TTUR) [18] training method, have a discriminator learning rate of $2 \times 10^{-4}$, use Balanced Consistency Regularization (bCR) [60], r1 regularization [39], and a learning rate decay (per dataset). Given this, it is highly probable that our hyper-parameters are not optimal and we focused on searching the kernel and dilation space of Hydra-NA. We also perform an ablation study compared to StyleSwin (Section 4.3) to demonstrate the effects of the improvements we have made. Our experiments were performed on nodes with NVIDIA
A100 or A6000 GPUs. We expect follow-up research with more computational resource to be able to out-perform our results with better parameters and encourage them to open issues on our GitHub page.

Our partitioned design with progressive dilation demonstrates that we can bridge the complexity and compute gap, having the best of both worlds. We are able to achieve high quality results on both highly structured data, like FFHQ, as well as less structured data like LSUN Church. We are able to achieve this while maintaining a high throughput and without a large number of parameters. Note, the number of partitions has nearly no affect on throughput, slight due to lack of CUDA optimization, and does not affect MACs or parameters, making all our models at a given resolution approximately equal.

4.1. FFHQ

Flickr-Faces-HQ Dataset (FFHQ) [29] is a common face dataset that allows generative models to train on regular typed data. FFHQ has images sized 1024x1024 and represents high resolution imaging which are cropped to be centered on faces, making it uni-modal. Faces are fairly regular in structure with the most irregular structure being features like hair and background. This dataset serves as a substitute for data with regular structures and can give us an idea of how effective generation is for specific tasks. For these experiments we train on all 70k images and then evaluate from a random 50k subset against a 50k sample from the model.

Figure 4: Samples from FFHQ 256 with FID50k of 2.05.

For FFHQ-256 experiments we trained our networks for 940k iterations with a batch size of 8 (per GPU), for a total of 60.2 million images seen. We start or LR decay at 775k iterations achieving a FID50k of 2.046. For our best result we also include a random horizontal flip, which will double the number of potential images. As we are compute limited, it is possible that these scores could be improved and our loss had not diverged by 1M iterations. For this task we used only 2 partitions, and we got significant improvements by combining dense and sparse receptive fields. We show these results in our ablation study, Table 3. Our results represent the current state of the art for FFHQ-256 image generation. FFHQ-256 samples are shown in Figure 4.

For FFHQ-1024 we use a batch size of 4 (per GPU), for 900k iterations (28.8M images), horizontal flips, and we start our LR-Decay at 500k iterations obtaining FID50k of 4.174. All other hyper-parameters are identical to the 256 run and we only trained this model a single time. Similarly, our model did not appear to converge and our loss was within noise by 1M iterations. This model still significantly benefits from just two partitions but due to the substantial increase in computation required for these larger images, we have yet to be able to train other architectures (this is the only one we tried). Still, our results reflect the state of the art for transformer based GANs. Note that training on a single A100 node takes approximately a month of wall time and that writing 50k images to disk (to calculate FID) can take over 5.5hrs alone! Due to this compute cost, we did no tuning on this model and just trained once and reported the results. In our FFHQ results we notice that our results frequently capture the long range features that we are seeking. This provides evidence that our hypothesis about including non-local features increases the quality of the results. Since human perception does not correlate with most metrics [49], we have a detailed discussion in the appendix: in Appendix C we investigate visual fidelity and artifacts and in Appendix D we investigate the attention maps to explain these differences.

4.2. LSUN Church

Our second experiment is with LSUN [57] Church. This dataset contains churches, towers, cathedrals, and temples, and is less structural than FFHQ where images have substantially higher variance from one another. Particularly the main object in the scene is not in a consistent location and auxiliary objects – such as trees, cars, or people – are in arbitrary locations and quantity. This task demonstrates the ability to perform on diverse and complicated distributions of images. We found that two partitions was not sufficient to capture the complexity of this dataset, ablation in Table 3. Across generative literature it is clear that there is a pattern that models perform better on more structured data or less structured data, with no model being a one size fits all. The models with best performance on both datasets tend to have significantly more parameters, preventing a fair comparison. Specifically, we notice performance gaps in
Table 1: FID50k results based on training with entire 70k FFHQ dataset. A random 50k samples are used for each FID evaluation. We separate convolutional based methods from transformer based using a horizontal line. Throughput and the number of parameters are included for better comparison of networks. “Usage Metrics” are evaluated on the $256 \times 256$ resolution on the same machine, run by ourselves, to maintain fair comparisons. We were not able to load/run the official checkpoints of some models and measure their parameters or throughput. StyleNAT results (on any dataset) do not utilize fidelity increasing techniques such as the truncation trick [32].

ProjectedGAN, Unleashing (Transformers). ProjectedGAN has excellent Church performance, SOTA, but actually as a non-competitive FID score for FFHQ. Similarly Unleashing Transformers has a large gap, performing better on Church. This may be due to evaluation differences and limitations with FID, which we discuss in Appendix C. While we do not achieve state of the art results yet with our current limited hyperparameter exploration, we still maintain a competitive performance with an FID of 3.400. Samples from the Church generation can be seen in Figure 5.

4.3. Ablation and Discussion

Our closest competitor is StyleSwin, where we share a lot of similar architecture choices but with some distinct differences. Because of this, we use StyleSwin as our baseline for our ablation and build from there. We will compare against FFHQ as this is the most popular dataset.

Starting with StyleSwin, introduce Neighborhood Attention into the model, simply by replacing both WSA and SWSA modules in the generator with Neighborhood Attention (+NA). Next, we introduce Dilated NA (DiNA), replacing only one split of the attention heads with the dilated variant. In other words, we replace WSA with NA and SWSA with DiNA (+Hydra). Third, we introduce the random horizontal flip to demonstrate the effect of augmentation (+Flips). StyleSwin did not use this on FFHQ but did with other datasets, so we assume they did not see improvements. Finally, we introduce our four partitions to see if intermediate receptive fields help. We use a progressive dilation, with the first partition with dilation 1 (NA) and the last partition with maximum dilation according to feature map size, evenly growing between (+Prog Di). All results are tested on FFHQ-256 and shown in Table 3. These results demonstrate that our innovation is more than simply adding NA and that the partitioning and dilations are critical to the high performance. While appearing simple, we are the first to implement such a design for transformers.

While limited, we also include an ablation for our search space on LSUN Church investigating the number of parti-
Table 3: Ablation study comparing models on FFHQ-256 dataset. We start with StyleSwin, replace the Swin transformer with a Neighborhood Attention Transformer, then introduce Hydra, then horizontal flip augmentation, and finally a 4 partition Hydra design with progressive dilation.

| Ablation          | FID ↓ | Diff ↓ |
|-------------------|-------|--------|
| StyleSwin         | 2.81  | –      |
| + NA              | 2.74  | -0.07  |
| + Hydra           | 2.24  | -0.50  |
| + Flips           | **2.05** | -0.19 |
| + Prog Di (4)     | 2.55  | +0.50  |

Table 4: Comparison for number of head partitions when learning LSUN Church. Min heads represents the minimum number of heads in our transformer.

| Partitions | Min Heads | FID ↓ | Diff ↓ |
|------------|-----------|-------|--------|
| 2          | 4         | 23.33 | –      |
| 4          | 4         | 6.08  | -17.25 |
| 6          | 8         | 5.50  | -0.58  |
| 8          | 8         | 3.40  | -2.10  |

In several configurations, despite poor results we noticed our model would frequently create readable watermarks and citations, shown in Section 4.3, and produce what the authors thought were higher quality images. This suggests that the model may be able to learn high quality images and poor performance is due to non-optimal hyper-parameter choices, which we used a fairly limited search. We have a further discussion in the Appendix providing possible insight into this, analyzing via our attention visualization technique.

5. Conclusion

In this work we present StyleNAT, which shows a highly flexible generative network able to accomplish generation on multiple types of datasets. The Hydra-NA design allows for different partitions of heads to utilize different kernels and/or dilations. This design allows for a single atten-
This work was in part supported by the Intelligence Advanced Research Projects Activity (IARPA) under Contract No. 2002-21102100004. We also thank the University of Oregon, University of Illinois at Urbana-Champaign, and Picsart AI Research (PAIR) for their generous support that made this work possible.

Acknowledgments. This work was in part supported by the Intelligence Advanced Research Projects Activity (IARPA) under Contract No. 2002-21102100004. We also thank the University of Oregon, University of Illinois at Urbana-Champaign, and Picsart AI Research (PAIR) for their generous support that made this work possible.

References

[1] Rameen Abdal, Peihao Zhu, Niloy Jyoti Mitra, and Peter Wonka. Styleflow: Attribute-conditioned optimization of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 2021. 4

[2] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P. Breckon, and Chris G. Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vectorized codes. In *European Conference on Computer Vision (ECCV)*, 2022. 7

[3] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. 13

[4] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022. 13

[5] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv:2011.10650*, 2021. 1

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 2

[7] Pratul Tharwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv:2105.02533*, 2021. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[9] Rinon Gal, Dana Cohen Hochberg, Amit Bermano, and Daniel Cohen-Or. Swagan: A style-based wavelet-driven generative model. *ACM Transactions on Graphics (TOG)*, 2021. 4, 7

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 21

[11] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. *arXiv:2209.15001*, 2022. 2, 3

[12] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. *arXiv:2204.07143*, 2022. 2, 3, 4

[13] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 14

[14] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv:2104.05704*, 2021. 2

[15] Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. Efficient-vgvae: Less is more. *arXiv:2203.13751*, 2022. 1

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[17] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020. 21

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5

[19] Drew A Hudson and C. Lawrence Zitnick. Compositional transformers for scene generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 7

[20] Drew A Hudson and C. Lawrence Zitnick. Generative adversarial transformers. In *International Conference on Machine Learning (ICML)*, 2021. 1, 7

[21] Ahmed Imitiaz Humayun, Randall Balestriero, and Richard Baraniuk. MaGNET: Uniform sampling from deep generative network manifolds without retraining. In *International Conference on Learning Representations (ICLR)*, 2022. 4

[22] Ahmed Imitiaz Humayun, Randall Balestriero, and Richard Baraniuk. Polarity sampling: Quality and diversity control of pre-trained generative networks via singular values. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[23] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 7

[24] Animesh Karnewar and Oliver Wang. Msg-gan: Multi-scale gradients for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4
Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196, 2018. 4

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. arXiv:2006.06676, 2020. 4

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In Proc. NeurIPS, 2020. 21

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In Advances in Neural Information Processing Systems (NeurIPS), 2021. 4, 7, 14, 27

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 2, 4, 6

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 4, 7

Tuomas Kynkänen, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in Fréchet inception distance. In The Eleventh International Conference on Learning Representations, 2023. 13

Tuomas Kynkänen, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. arXiv:1904.06991, 2019. 7, 13

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In European Conference on Computer Vision (ECCV), 2022. 2

Jingyun Liang, Jiezheng Cao, Guoeli Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021. 3

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 2

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 2, 3, 14, 16

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 3

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 3

Lars M. Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In International Conference on Machine Learning (ICML), 2018. 5

Muhammad Ferjad Naem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In Hal Daumé III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 7176–7185. PMLR, 13–18 Jul 2020. 13

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In International Conference on Machine Learning (ICML), 2018. 2

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training, 2018. 2

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. Advances in Neural Information Processing Systems (NeurIPS), 2019. 2, 3

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. Advances in neural information processing systems, 31, 2018. 13

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In Advances in Neural Information Processing Systems (NeurIPS), 2021. 7

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. Advances in Neural Information Processing Systems, 34:17480–17492, 2021. 21

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. arXiv:2201.00273, 2022. 7

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv:2010.02502, 2021. 1

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models, 2023. 6, 13

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning (ICML), 2020. 2

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 2

Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 3, 4
A. Model Architecture

We include a table for the StyleNAT architecture for added clarity and reproducibility. Each level has 2 NAT layers which are split heads. The half the heads takes the normal NA kernel and the other half takes DiNA kernels. Dilation follows the pattern of using powers of 2 and starts when we start using the NA kernels. Table 5 shows the configurations used for both FFHQ results.

| Level | Kernel Size | Dilation | Dilated Size |
|-------|-------------|----------|--------------|
| 4     | -           | -        | -            |
| 8     | 7           | 1        | 7            |
| 16    | 7           | 2        | 14           |
| 32    | 7           | 4        | 28           |
| 64    | 7           | 8        | 56           |
| 128   | 7           | 16       | 112          |
| 256   | 7           | 32       | 224          |
| 512   | 7           | 64       | 448          |
| 1024  | 7           | 128      | 896          |

Table 5: StyleNAT 2-Split Model Architecture. First level uses Multi-headed Self Attention and not DiNA. This model is used for all FFHQ results.

| Level | Kernel Size | Dilation |
|-------|-------------|----------|
| 4     | -           | -        |
| 8     | 7           | 1        |
| 16    | 7           | 1.2      |
| 32    | 7           | 1.2, 4   |
| 64    | 7           | 1.2, 4, 8|
| 128   | 7           | 1.2, 4, 8, 16 |
| 256   | 7           | 1.2, 4, 8, 16, 32 |
| 512   | 7           | 1.2, 4, 8, 16, 32, 64 |
| 1024  | 7           | 1.2, 4, 8, 16, 32, 64, 128 |

Table 6: Example of progressive dilation with 8 heads, referred to “pyramid dilation.”

Figure 7: Full code for StyleNAT’s Hydra-NA module. Requires NATTEN package. Using NATTEN v0.14.6

A.1. Discussion of Other Configurations

We include a small discussion about our architecture search to help others better learn how to utilize StyleNAT/Hydra-NA and accelerate usage. Initially we had explored using progressively growing kernels in a pure NA based architecture but found that this scaled poorly as well as did not provide significant improvements over other models. Progressive kernels also has the downside of significantly increased computational load.

We also investigated replacement of StyleSwin where individual layers were replaced with pure NA layers, but did not run these to convergence. Each swap resulted in a minor
improvement in the initial loss curves but not by significant margins. A version where we replaced all version of swin layers can be seen in our ablation study (Table 3, +NAT) and we note that NAT alone only provides minor improvements from Swin, but that dilation (DiNA) offers significant improvements. We remind the reader here that using two NA partitions with the same kernel and dilation – as was done here – is equivalent to using a single partition.

We also did not find any significant improvements when using one layer with a small kernel and another layer with a larger kernel (on the same level but different steps). While we found NA based runs more consistent than StyleSwin runs, layer replacement runs were often within the margin of variance. This leads us to conclude that the increased performance is coming from NA’s higher expressivity.

Once we introduced the dilation into the partitioned head design did we find significant improvements within our generative process. This leads us to conclude that the advantage of StyleNAT is the ability to pay attention to both local and global features. We believe that this architecture extends the power of heads in transformer architectures. In a typical multi-headed transformer model, heads are able to pay attention to different features. In our model each of these heads is not only able to pay attention to different features, but can do so in different ways. In addition, one can think of this hydra structure as also a singular creature that can view the same object from multiple viewpoints and combine the information before making decisions. The advantage our of network is this ability to be able to view different things and from different perspectives. This is because we partition the heads, meaning each kernel/dilation configuration has multiple heads.

This flexibility is important for adapting to different datasets with different complexities. We demonstrate this with our two ablations. In our ablation on FFHQ, Table 3, we do not see continual improvements from added partitions (+ Prog Di (4)). But in our LSUN Church ablations, Table 4, we see that this partitioning is essential. We also show that the gain in fidelity is not due purely to the minimum number of heads as there is a much larger gain when moving to 8 partitions rather than 6 partitions, despite that both models have the same minimums. We will continue expanding upon these studies but with increased flexibility comes a larger solution space.

B. Hydra-NA

In Figure 7 we include the full PyTorch code for our Hydra style NA design. We include the full code added reproducibility/transparency and so others can test our design quickly and to reduce potential typos. This code account for uneven partitions and decides the number of splits by the number of kernels supplied. All comments, warnings, and asserts have been removed for clarity but are included in the official release. This code requires the NATTEN package (functions highlighted in dark yellow) and PyTorch (highlighted in red). We also expect this module to be beneficial for domains outside of generation.

C. Visual Quality and Metric Limitations

Interpreting quality of generative models is quite difficult and metrics like FID do not quite capture the nuances that humans perceive. Each generative model type leaves their own fingerprints on a generation and thus it is valuable to recognize and discuss these artifacts. In this section we investigate the artifacts generated from StyleGAN3, StyleSwin, and StyleNAT on FFHQ-1024 samples. All samples will be curated and selected for highest quality to maintain fairness. Note that with respect to FID for FFHQ-1024 images StyleGAN3 has FID 2.79, StyleSwin with 5.07, and StyleNAT with 4.17. We note this because FID also measures variance across sampling and that we’d expect StyleSwin to produce the worst samples. We’d additionally expect StyleSwin and StyleNAT to produce significantly worse examples than StyleGAN3. Due to our curating there is going to be sample bias and we notice that images with simple backgrounds or those with Bokeh like effects appear to have higher overall fidelity. This may just be a bias to FFHQ, but we notice this pattern across all three models.

The motivation for doing this is that we know that FID has biases that do not directly correspond to image fidelity [44, 32, 40], as well as we know that ImageNet plays an important role [31]. Since the goal of these models is to generate high fidelity images, we should be cognizant of the limitations of our numerical evaluations and ensure that we do our best to evaluate on the actual goal and not the evaluation biases. A recent work [49] compares many types of generative models with human evaluations and discusses some of these issues in more depth, demonstrating that on many datasets – including FFHQ – that human evaluation has little correlation to many metrics, including FID. [3, 4] give a good overview of the limitations of each metrics, many of which are used in [49].

We find that the generative artifacts discussed in these sections are relatively common and often unique to the particular architecture. Readers are highly encouraged to zoom in on photos and use a good monitor. These features can also be good indicators of deep fakes and a careful read of this section can help readers detect deep fakes and potentially identify the generative model. In particular, we find common regions for artifacts to appear, universally, are on the ears, eyebrows, eyelashes, neck, pupils (location, size, and color), as well as textures especially within these regions. Once some of these patterns are noticed they may be difficult to not notice and we warn readers to be careful of this bias.
C.1. StyleGAN3

We start with a sample from StyleGAN3 [28]. Figure 8 which has significantly higher fidelity than the other networks. Despite being from the curated dataset, with additional curating by us, we still see clear generative artifacts. We notice that on the forehead (Figure 8b) there is some banding pattern and in the glasses (Figure 8c) there are near hexagonal patterns.

Despite being “alias free” we find these artifacts prolific and reminiscent of the the internal representations they themselves show. Such artifacts are common in StyleGAN3 images and the banding pattern can be found in nearly every sample from their curated list. Particularly the banding in the skin texture is quite noticeable and an obvious sign unique to StyleGAN3, though we did not find this same pattern in StyleGAN2 samples. We find that StyleGAN3 images frequently have an artificial texture on the face. We also find that fine details may not be well captured here, with this sample lacking any eyelashes.

C.2. StyleSwin

For StyleSwin [58] we noticed a different set of patterns. Their network performed better on glasses but created some blocking features and hard lines. We believe that this is due to rectangular features that the Swin mechanism [36] has squared features. The attention maps in Appendix D make this argument clearer. Hassani et al. [13] also shows some of this squaring from ViTs, and specifically Swin, through the use of Saliency Maps as well as discusses the lack of translational and rotational equivariances and invariances. To show this we generated 50 images from StyleSwin’s FFHQ-1024 checkpoint and selected the best one, to keep a fair comparison of curated examples. In Figure 9 we can see some of the blocking that happens. Particularly in Figure 9b we can see that there are multiple rectangular shapes that are inside other rectangles. We note that this is not nearly as obvious of an error as StyleGAN3 and do consider this a more successful sample. We also focus on the right ear (Figure 9c) and we can see unrealistic patterns. There are also some minor errors around the eye with crow’s feet like structure with similar patterns to the ear. We did also notice that while StyleSwin captured some long range features, such as eye color and facial symmetry, that this still left some aspects wanting. The eye and pupil sizes are a glaring mistake and create an uncanny valley. Fine details also tend to be lost, such as the eyelashes. We believe that this is due to the limitations of the shifted window mechanism.

C.3. StyleNAT

For StyleNAT we notice far fewer of these errors and believe that the artifacts are more realistic with respect to actual human skin. Our sample in Figure 10 shows far less issues with these types of features. Skin texture looks far more realistic and in the foreheads of samples we only see soft lines which many viewers may think indistinguishable from natural creases (Figure 10b) or straggling hairs. We note that while these lines are difficult to see, they also are not realistic. We also notice some slight spotting and discoloration around the edges of eyes (Figure 10c) but again, this is far less noticeable than that from the other networks. At a high zoom we can also see artificial textures, especially on the lips (common on all networks). Readers may need a good monitor. We also see improved performance
on the eyelashes but believe that these are still wanting and look almost drawn at high zoom.

We notice that NA provides far better long range feature matching than compared to StyleSwin, and this demonstrates that our network better approaches the goals that we’d seek from a transformer based generator – obtaining consistent long range features.

D. Attention Maps

To help explain the observations we see throughout this work, we visualized the attention maps for both StyleSwin and StyleNAT. We note that neither of these networks can have attention maps generated in the usual manner. Our code for this analysis will also be included in our GitHub. For both versions we can’t extract the attention map from the forward network, as would be usually done, but instead extract both the query and key values. Exact methods are explained in the respective FFHQ sections. We believe that these maps demonstrate the inherent biases of the network and specifically demonstrate why StyleNAT, and critically the Hydra attention, result in superior performance. Swin’s shifted windows demonstrate a clear pooling, which may be beneficial in classification tasks, but not as much for generative tasks, which are more sensitive and unstable. They also provide explanations upon where both networks may be improved within future works and we believe this tool will be valuable to other researchers in other domains.

We will look at both FFHQ and LSUN Church to try to determine the differences and biases of the networks and attention mechanisms. For all of these we will generate a random 50 samples and select by hand representative images for the give tasks. It is important to take care that there is a lot of subjectivity here and that these maps should only
be used as guides into understanding our networks rather than explicit interpretations. Regardless, the attention maps are still a helpful tool in determining features and artifacts in generation, as we will see below. The patterns discussed were generally seen when looking at each of the sampled images during our curation.

Our attention maps suggest that these networks follow a fairly straightforward and logical method in building images. In general we see that lower resolutions focus on locating the region of the main objects within the scene while the higher resolutions have more focus on the details of the images. We see progressive generation of the images, that each resolution implicitly learns the final image in progressively detail. This suggests that the progressive training seen in StyleGAN-XL may also benefit both of these networks. The Style-based networks generally have two main feature layers (or blocks) per resolution level, which we similarly follow. Our maps also suggest that a logical generation method is performed at the resolution level. Where the first layer generating the structure, realigning the image after the previous up-sampling layer. The second layer generates more details at the resolution level. This may suggest that a simple means to increasing fidelity would be to make each resolution level deeper, which is also seen in StyleGAN-XL. In other words, fidelity directly correlates to the number of parameters, and thus it is necessary to incorporate that within our evaluation. The goals of this work is on architecture and the changes that they make, rather than overall fidelity. We leave that to larger labs with larger compute budgets.

D.1. FFHQ

For FFHQ we will look at specifically the 1024 dataset and we will select our best sample. We are doing this to help determine the differences in artifacts that we saw in Appendix A. Since many of these features are fine points we will want to see the high resolution attention maps to understand what the transformers are concentrating on and how the finer details are generated. Specifically, we use the same images that were used within the previous section to help us identify the specific issues we discussed.

D.1.1 (FFHQ) StyleSwin

For StyleSwin we extract the query and key values from each forward layer (note that there are two attentions per resolution level for StyleGAN based networks). We perform this for each split window which has shape $[B, n_h, w^2, C']$, where $B$ is the batch, $W$, $H$ are the height and width, $w$ is the window size, $n_h$ is the number of heads, and $C'$ is the number of channels. We concatenate along the split heads and then reverse the windowing operation by re-associating the windows with the height and width. Once this is done we can mean the pixel dimensions for the query and flatten them for the key (q is unsqueezed for proper shaping). We then can obtain a normal attention map where we have an image of dimensionality $B, n_h, H, W$.

We will look at this attention map for the same sample as in Figure 9. We break these into multiple figures so that they fit properly with Figure 11 representing the $1024 \times 1024$ resolution, Figure 12 the $512 \times 512$, Figure 13 representing both the $256 \times 256$ and $128 \times 128$, Figure 14 the $64 \times 64$ and $32 \times 32$, and finally Figure 15 representing the $16 \times 16$ and $8 \times 8$ resolutions. Note that the second half of the heads represents a shifted window, per the design specified in their paper. In these feature maps we see consistent blocking happening, which is indicative of the issues with the Swin Transformer [36]. This also confirms the artifacts and texture issues we saw in the previous section. These artifacts can even be traced down to the 64 resolution level, Figure 14. We believe that this is a particularly difficult resolution for this network as it has more blocking in the second layer than others.

We also notice that in the earlier feature maps that StyleSwin has difficulties in picking up long range features, such as ears and eyes. This likely confirms the authors’ observations of frequent heterochromism (common in GANs), mismatched pupil sizes, and differing ear shapes.

At higher resolutions (128 and above) we find that the network struggles with texture along the face despite establishing the general features. This is seen by half the heads being dark and the other half being bright, as is seen in Figs. 12 and 13. This does suggest some under-performance from the network, with one set of heads doing significantly more work when compared to the others. We also see higher blocking, especially in the first layer, at lower resolutions, indicating difficulties in acquiring the general scene structure. This warrants more flexibility, such as that offered by Hydra.

D.1.2 (FFHQ) StyleNAT

For StyleNAT we perform a similar operation as to StyleSwin. We similarly extract the queries and keys, mean over the query’s pixel dimensions (unsqueezing), and flattening the key’s pixel dimensions. We similarly get back an image of shape $[B, n_h, H, W]$, with similar dimension definitions. We break these into multiple figures so that they fit properly with Figure 16 representing the $1024 \times 1024$ resolution, Figure 17 the $512 \times 512$, Figure 18 representing both the $256 \times 256$ and $128 \times 128$, Figure 19 the $64 \times 64$ and $32 \times 32$, and finally Figure 20 representing the $16 \times 16$ and $8 \times 8$ resolutions. The first half of the heads has no dilation and the second half has dilations corresponding with the architecture specified in Table 5.
Figure 11: FFHQ StyleSwin attention maps at the level with 1024 resolution. Each layer has 4 heads with kernel size of 8 but half of them were trained with Shifted WSA. First level appears to concentrate on facial structure and texture. Second level appears to focus on symmetric features such as cheeks and eyes.

Figure 12: FFHQ StyleSwin attention maps at the level with 512 resolution. Each layer has 4 heads with kernel size of 8 but half of them were trained with Shifted WSA. Generative artifacts are clearly visible on forehead in most maps. Heads have vastly different concentration levels.
Figure 13: FFHQ StyleSwin attention maps at levels with 128 and 256 resolution. Every layer in each level has 4 heads with kernel size of 8 but half of them were trained with Shifted WSA. 128 resolution shows beginning indications of generative artifact.

We see that in the high resolution images that NA is learning textures and long range features across the face. This supports the claim that transformer mechanisms are adequately learning these long range features, as should be expected, and would support more facial symmetry that would be seen in human faces.

In the first layer we notice that more local features are being learned, which explains the better textures seen in our samples. Particularly we notice in Figs. 16 and 17 that the first 2 heads learn feature maps on the main part of the face.
Noting that the first two heads represent $7 \times 7$ kernels that are not dilated. We also noticed some swirling patterns in the second half of heads, which correspond to dilated neighborhood attention mechanisms. These correspond to the soft lines we saw above, and act like edge detectors. The second layer does a better job at finer detail and removes many of these, tough they are still visible on the chin. We notice that these particularly appear around hair and may be reasoning that the hair quality of our samples perform well. The long range, dilated, features all do tend to learn long
range features and aspects like backgrounds, as we would expect.

In the lower resolutions we see that these attention maps learn more basic features such as noses, ears, and eyes, which helps resolve many of the issues faced by CNN based GANs. Details such as eyes and mouth can be identified even at the 16 resolution image Figure 20! The authors noticed that while generating they observed lower rate of heterochromia (different eye colors), which are common mistakes of GANs. This is difficult to quantify as it would un-
likely be caught by metrics such as FID but we can see from the attention maps that the early focus on eyes suggests that this observation may not be purely speculative.

We believe that these attention maps demonstrate a strong case for StyleNAT and more specifically our Hydra Neighborhood Attention. That small kernels can perform well on localized features, like CNNs, but that our long range kernels can incorporate long range features that we’d want from transformers. We can also see from the feature maps that the mixture of heads does support our desire for added flexibility. This is done in a way that is still efficient computationally, having high throughputs, training speed, and a low requirement on memory.

D.2. Church Attention Maps

To help us understand the differences in performances specifically in the LSUN Church dataset we also wish to look at the attention maps to help give us some clues. We know that FID has limitations being that Inception V3 is trained on ImageNet-1k and uses a CNN based architecture. ImageNet-1k is primarily composed of biological figures and so does not have many objects that have hard corners like LSUN Church. Additionally, CNNs have a biased towards texture [10, 17], which can potentially make the metric less meaningful, especially on datasets like this. Since we had noticed that the Swin FFHQ attention maps had a bias to create blocky shapes and StyleNAT had a bias to create rounder shapes, we may wish to look into more detail to determine if these are biases of the architecture or that of the dataset. We find that this is true for StyleSwin but not of StyleNAT.

To understand why this dataset provides larger difficulties for these networks we not only select a good sample, but also a bad sample, hoping to find where the model loses coherence. We find that in general this happens fairly early on, with the networks having difficulties placing the “subjects” within the scene. We see higher fidelity maps in the better samples but find that overall these struggle far more than on the FFHQ task.

We believe that our results here show that FID is not reliable for the LSUN Church dataset, as well as demonstrates that Church is a significantly harder generation problem for these models than FFHQ is. This is claim is consistent with many of the aforementioned works, which present stronger cases and make similarly arguments for other metrics. These demonstrate the need to perform visual analysis as well as feature analysis to ensure that the model is properly aligned with the goal of high quality synthesis rather than with the biases of our metrics. Evaluation unfortunately remains a difficult task, where great detail and care is warranted.

Specifically, at low resolutions both networks have difficulties in capturing the general concept of the scene. We believe that this is due to the increased variance and diversity of this dataset, compared to FFHQ. While human centered faces share a lot of general features, such as a large oval centered in the image, this generalization is not true for the Church dataset, which a wide variety of differing building shapes, many different background objects to include (which we say FFHQ prefers simple backgrounds), and that the images are taken from many different distances.

StyleSwin samples are shown in Figure 21 and the StyleNAT samples are show in Figure 26. We believe that both these samples look on par with the quality of that of ProjectedGAN [46], which currently maintains SOTA on LSUN Church with an FID of 1.59, and thus are sufficiently “good” samples. We will look at the blocks sized 32 to 256, as we believe this is sufficient to help us understand the problems, but we could generate smaller maps.

It is unclear at this point if the fidelity could be increased simply by increasing the number of training samples or if additional architecture changes need to be made in order to resolve this (as suggested above). We will specifically note that even SOTA generation on this dataset, ProjectedGAN [46], does not produce convincing fakes, while this task has been possible on FFHQ for some time, albeit not consistently. This is extra interesting considering that the SOTA FID on LSUN Church is 1.59, with 3 networks being below a 2.0 while SOTA FFHQ-256 (the same size) is 2.05 (this work) and scores as high as 3.8 [27] frequently produce convincing fakes. We also remind the reader that while ProjectedGAN performs well on Church (1.59), it does not do so on FFHQ (3.46), see Tables 1 and 2.

D.2.1 (Church) StyleSwin

For the StyleSwin generated images, Figure 21, we can see that the good image looks nearly like a shutterstock image, almost reproducing a mirrored image and where the text is almost legible. But in this we also see large artifacts, like the floating telephone poll, the tree coming out of a small shed, or other distortions. We believe that this telephone poll may actually be part of a watermark, but are unsure. In the bad image, we see that there was a mode failure and specifically that the generation lost track of the global landscape. Even with this failure, we still do see church like structures, such as a large distorted window, making this image more of a surrealist interpretation of a church than a photograph. These images will thus provide good representations for understanding these two modes, of why good church images are still hard to generate and why they completely fail.

In short, we find that at all levels, there is higher blockiness and less detail captured by the attention mechanisms when compared to FFHQ. We see either very high or very low activations with the inability to focus on singular tasks. At low resolutions we see difficulties capturing structure
Figure 16: FFHQ StyleNAT attention maps at the level with 1024 resolution. Every layer has 4 heads with kernel size of 7, but the last 2 heads have a dilation of 128. First layer concentrates more on structure and second more on texture. Heads without dilations appear to focus more on texture and the face.

Figure 17: FFHQ StyleNAT attention maps at the level with 512 resolution. Every layer has 4 heads with kernel size of 7, but the last 2 heads have a dilation of 64. First layer appears to focus more on structure, with no dilations concentrating on the face. Dilated heads focus on head shape.
and that this error propagates through the model.

Figure 22 shows our full resolution images, and in the attention maps we can again see the same blocky/pixelated structures that we found in the FFHQ investigation, but at a higher rate. For the good images, in the first layer we see that the first head is performing an outline detection on the scene, almost like a sobel filter. We also see that the last head is delineating the boundary between the foreground
Figure 19: FFHQ StyleNAT attention maps at levels with 32 and 64 resolution. Every layer in each level has 16 heads with kernel size of 7. The second half of the heads have dilations 4 and 8, respectively. Main structure visible in these resolutions, including eyes and the separation of face and hair.

and background, particularly the sky. Interestingly this appears almost like Pointillism, which we see in many following maps as well. This is likely bias from the shifted windows. In the second layer we see more fine grained structure, but interestingly we do not see as much as we saw in the FFHQ version at the same resolution, Figure 13, which suggests that this is a more difficult task for this network. We can also see that this level is concentrating on the text.
Figure 20: FFHQ StyleNAT attention maps at levels with 8 and 16 resolution. Every layer in each level has 16 heads with kernel size of 7. The second half of the heads have dilations 1 and 2, respectively. The 8 resolution image looks to be focusing on placement of object within the scene, taking the general round shape and distinguishing subject from background.

at the bottom of the image, which is not as clearly visible in the smaller resolution maps. Notably, we do not clearly see the floating telephone pole or the wires in the sky. These could be formed from another part of the network, such as the RGB or MLP layers, but we have not investigated this. Further investigation is needed to understand the contributions of these layers.

Moving down to the 128 resolution images in Figure 23...
we see that the attention maps overall get much messier. For the good image we can see that the roof of the church is picked up by many of the heads. In the second layer, on the second head, we also see a clear filter looking at the tree and roof of the church, which we can also see a less clear selection in head 6 and the first layer at head 5. These same heads provide decent filters for the bad images, the layer 1 head 5 and layer 2 head 2 seeming to do the best at object filtering.

Looking at the 64 resolution Figure 24 and 32 resolutions Figure 30 we can more clearly see where the problems are happening. In FFHQ the 64 resolution maps, Figure 14, is where we start to first see our main object with relative details and the 32 resolution has a decent depiction of broad shape. We do not have as good of an indication within these maps, where the 64 resolution images do not have clearly identifiable building textures, let alone building shapes. This is even worse at the 32 resolution image level.

Interestingly, at this resolution it is difficult to distinguish which version would generate the good or bad image, which doesn’t seem distinguishable till at least the 128 resolution. These aspects suggest that the generation of this data is substantially harder for this model. This is extra interesting considering that the FID scores are fairly close for both of these datasets, with FFHQ being 2.81 and Church being 2.95. With more difficulties in capturing general structure the network then struggles to increase detail and this systematic issue cannot be resolved. This network saw trained on 1.5M iterations.

**D.2.2 (Church) StyleNAT**

StyleNAT performs significantly worse at LSUN Church, and it isn’t clear why this is. For these attention maps we will use the model that generated visible text and what the authors thought were higher quality. This network uses smaller kernel sizes of 3 and has a max dilation rate of 8. Thus the dilations are \([1],[1,2],[1,2,4],[1,2,4],[1,2,4],[1,2,4,8]\). This is the same configuration as when higher overfitting was observed. Since higher overfitting tends to correspond to higher fidelity we want to investigate what went right, to improve the work. The good image here is on par with that of Swin, and SOTA works, but the bad image is again a surrealistic work wherein we see a agglomeration of a “Church.”

The good image appears pixeled, has scan lines, and some other distortions such as the car being reflected and turned into a bush. The bad image seems to incorporate nearly every feature within the dataset, including churches, towers, temples, cathedrals, as well as many different trees all smashed together Cronenberg style.

In short, we find that StyleNAT is in fact able to generate hard lines, as this dataset is biased towards, but does tend to prefer smoother features. We also see that at even the early stages that the scene has difficulties capturing global coherence. This likely explains the instabilities we faced and why training often diverged fairly early on, with nearly a fifth of the number of iterations as FFHQ and nearly 10% of StyleSwin.

The 256 resolution attention maps, Figure 27, images we immediately see that some of the attention heads to not have rounder features, indicating that our network does not have a significant bias towards biological shapes. In the first level, the first three attention heads have what appear to be scan lines, which we do see manifest in the full image. We also see traces of this in the next three heads, as well as most of the heads in the second layer. It appears that in this case, this level is looking a lot at texture, similar to that in FFHQ Figure 18. An interesting feature here, clearer in the first layer in heads 4-8, is that the we see what looks like the skeleton of a tree with branches coming out, almost centered at where the actual tree is in the main picture (left). What is notable here is that neither the trunk nor the branches are visible in the generated image, and that the “imagined” trunk is a bit translated from where we may predict it would be on the “actual” tree.

In other maps that we generated, that aren’t shown, we noticed this pattern is extremely frequent when trees exist in the scene and there exists identical structure when the tree does not have foliage. This includes the circular shapes adjacent to the trunks. We did not notice this feature when only the foliage is visible, where the tree may look more like a bush, such as in the bad sample. We did not notice such skeletons as prevalent in the Swin version, although the best example can be seen in head 4 of the 256 layer in Figure 22, but this appeared to be an exception rather than the norm. We are careful to make a conclusion that the network has classified trees and understands their skeletal structure and note that a reasonable alternative explanation is that this trunk looking figure can just as easily be a guide for distinguishing the location of the tree.

There is also a notable difference in the attention maps between levels. In general we believe these show that the first layer is working more on the general structure of the scene while the second layer is improving detail. We also saw such correlations within the FFHQ analysis. We believe that this is a reasonable guess because the first level follows an upsampling layer and thus the network needs to first re-establish the structure of the scene before it can provide detail. We also believe that this happens within the Swin based generator as well. This can mean that potentially higher fidelity generators can add additional layers, and that this is more necessary at higher resolutions.

As for the reasons for the low quality generations, we notice that the scenes in the 32 and 64 resolution, Figs. 29
Figure 21: Church good and bad samples from StyleSwin. Good example has a clearly visible church and tree with a good distinction of foreground and background. Good example has a fairly legible citation but no other watermarks. Bad example has lost global structure but does maintain church like features.

and 30, levels have potentially suggestive attention maps. Particularly we notice that the bad quality image has much more chaotic attention maps. Interestingly, we also see the scan lines.

E. Limitations

This work has some significant compute limitations compared to other works. For example, most of the compute budget was allocated to the FFHQ-1024 experiment, where we only performed a single training run. With this, it would be naive to assume that these parameters are an optimal setting for this task. The only parameter we changed was when the learning rate decay started, which was done via watching the training logs. This is different than our competitors, where we see that StyleSwin changes the channel multiplier on the generator. StyleGAN3, and thus StyleGAN-XL, make significant changes to hyper-parameters between each experiment and resolution. It is also worth noting that Appendix G of the StyleGAN3 [28] shows that they spent a total of 91.77 V100 years for their project which is many times the compute budget of this project. With the increased flexibility of our method, we introduce additional parameters which should be changed per-dataset and potentially per resolution. This means that there is potentially still gains to be made within performance, especially on the Church dataset. Another limiting factor of this work is that of Church and other datasets. Due to a limited compute budget, we were unable to search these spaces and fully explore how our model can, or cannot, adapt to the differing environments.

Additionally, due to our compute limits we were only able to perform experiments based on FFHQ and Church. While two datasets may not be much, we believe that these adequately capture the complexity landscape and are sufficient to motivate utility of the work. After all, utility is not simply based on getting SOTA across the board.

E.1. Church Overfitting

Despite our poor FID results on Church, we found that the model had a propensity to overfit results. We found it rather odd that we frequently got good images like those in Figure 31 but that our FID scores were poor. We believe that this means that our generator is overfitting to the dataset, given the distributional nature of the FID calculation. Interestingly the watermarks and shutterstock ids are clearly visible in these images and are quite readable. We also include some attention maps for a sample where we had such features appear. We are able to identify the watermark down to the 64 resolution level, where a bar in the middle can be seen. For the photo citation, we can actually see that the model starts paying attention to this at the 16
Figure 22: Church StyleSwin 256 sized samples with bad and good samples. Blocky structure still exists akin to pointillism. Maps has filters reminiscent of edge filters, where the good sample can distinguish foreground and background. The tree and church are clearly visible and the good sample has a predictable final image. The floating telephone or watermark is not clearly identifiable here but we can see activations in the shutterstock citation at the bottom. Bad sample does not have as clear of an identification, and is more likely to have curved features. Similar to FFHQ the first 2 heads of the first layer have low activations while the other heads have disproportionately high.
Figure 23: Church Style Swin 128 sized samples with bad and good samples. Good sample has clear good filters and some heads have strong focus on the main objects in the scene. One head in the bad sample has this same clear filter. Maps have less detailed focus, activating on many different points within the scene. Maps have less structure and features at this resolution than we saw within the FFHQ examples. In FFHQ we had less pointillism, especially in the first layer, but this is extremely prominent here indicating a difficulty in attending to the scene. Attention activation is highly disproportionate at this resolution.
Figure 24: Church StyleSwin 64 sized samples with bad and good samples. Samples are difficult to differentiate at this level and we have lower interpretability. In FFHQ the face was locatable at this resolution and the second layer started to reduce the blocking. This resolution still appears to be concentrating on the main structure of the objects, but has large range contexts in both layers. In the good sample we have difficulties identifying the tree or church, but they are somewhat visible. Attentions are highly checkerboard, likely due to the shifting of windows. The bottom of the images indicates concentration on the shutterstock citation in both samples.
Figure 25: Church StyleSwin 32 sized samples with bad and good samples. Global structure is generally lost and would be difficult to predict produced sample from these maps. The church is identifiable in the second layer of the good sample, but attention is a bit scattered. First level is more sporadic compared to the second level, which is more connected. Similar to FFHQ the first layer has large checkerboard patterns and second layer is smoother, though less general structure is identifiable. This appears to indicate that the first layer is matching structure to the upscaling and the second layer concentrates on details. Coherence is likely lost after this resolution.
resolution level, the smallest we investigated, where we see a strong white bar at the bottom of the image. These results help suggest that these models may have overfit. Unlike the FFHQ samples, we did not notice a strong correlation between this feature and the actual fidelity of the image. But we did notice that if the citation, watermark, and “X” across the image all strongly correlate. We typically do not find images where one is present but the other two aren’t, sometimes only visible in the attention maps or my careful inspection. See the third column and fourth columns of the first row in Figure 31, where the “X” is barely visible, zooming may be required. This additionally suggests that there is some overfitting to the dataset, as these features are not decoupled.

As seen in Appendix D.2.2 we were able to see that many of these features were visible at low resolutions, implying the model has a preference for these types of images. In Figure 31 we show a clearer case from our randomly generated 50 images, from the previous procedure. We only show the final attention maps, in order, and note that the first head of the first layer clearly highlights the watermark, which is nearly readable in non-occluded areas. The subsequent heads place a negative activation to the region and our maps suggest that the model views the watermark as part of the foreground. This is because the same first head is the only head to significantly highlight the “church.” In the second layer the watermark is outlined in non-occluded regions in the first head. Neither the web address nor the numeric ID are clearly readable within the attention maps, but we do see positive and negative attention within these regions.

F. Ethical Considerations

There are many ethical dilemmas to consider when discussing generative models, especially models that are able to produce high fidelity images of people. We do not condone the usage of our work for nefarious purposes, including but not limited to: manipulation of others or political gain. This includes usage in bot accounts, using the work to generate false or misleading narratives, or acts of terrorism. We do encourage the usage in artistic manners as well as using these results of these models to supplement training data. We highly encourage the usage of these models for the benefit of all and for using them to advance our scientific understanding of the world. These models have both the potential to do great good as well as great harm. While we cannot control how users of our models and frameworks utilize them, we have a duty to stress the importance of thinking about how products using these models can harm our fellow human beings. We stress that engineers need to think deeply about how their products can be abused and to con-
Figure 27: Church StyleNAT 256 sized samples with bad and good samples. Generated images are highly predictable within both good and bad samples. Scanlines artifacts and hard lines are visible in both images, showing hard lines can be learned. “Tree trunk” like feature visible in good sample, with branches and swirls where foliage is located. Likely indicates a guide for the location of the tree in the scene rather than learning tree skeletal structures. Both maps can distinguish foreground and background. Bad sample looks more church like than the actual image. Notably the second layer, which we believe focuses on detail, has far lower activations in the bad sample. Despite progressive dilation, it is difficult to tell if heads are associated with local or global features, as was apparent in FFHQ.
Figure 28: Church StyleNAT 128 sized samples with bad and good samples. Final image fairly predictable in the good sample but the bad sample looks more akin to stacked housing apartments. Scanlines weaker in the good sample and we can see loss of coherence in the bad sample. In both samples the detail layer has lower activations with one head appearing to dominate. We believe this decoherence propagates, preventing network from learning enough detail before scaling. Similar difficulties within StyleSwin indicate that this dataset may be more challenging and that detail is more important in lower resolutions. The early heads, which have no dilations, also clearly struggle to capture fine details. This is exceptionally apparent in the second layer which is more oriented towards this task.
Figure 29: Church StyleNAT 64 sized samples with bad and good samples. The final images are not easily predictable at this resolution and we see little coherence. General shapes can be distinguished but this is not as strong as in FFHQ. First layer clearly focuses on general structure while the second on more detail. We continue to have difficulties associating head dilation with the corresponding receptive fields of the scene. The many bands suggest that there are difficulties in locating the object’s placement within the scene. Both samples have tall tower like structures within the attention maps despite not being in final image or maps of the subsequent resolution. There are a lot of similarities between both samples, especially within the first layer. This could indicate overfitting and a strong preference to a strategy.
Figure 30: Church StyleNAT 32 sized samples with bad and good samples. General structure is fairly coherent with blocky and tower like structures. Strong band at the bottom likely indicates attempt to generate shutterstock citation. In FFHQ we had clear placement of the subject within the scene at this level and even features like eyes and mouth. We see difficulties for this at this level, but do see towering structures. Unlike FFHQ this dataset has many differences in the general structure and location of main objects. This resolution has decent coherence for both samples but the lack of detail in the second layer may indicate how the lack of quality propagates within the network. Similar to StyleSwin these maps tend to put focus on the center of the image.
Figure 31: Examples of Church generated images overfitting. Watermark and the image ‘ID’ are readable despite poor FID. The third image in the first row has a rotated 3 and z, while the fourth has what looks like a £. The second row’s IDs are all clearly numbers. The existence of the watermark correlates with other similar shutterstock features but not necessarily with image quality. Samples are with FID 4.22 and 4.73 respectively.

Figure 32: Attention maps on watermark visible Church samples. Attention maps for 256 resolution show. We can see the shutterstock logo, especially in the first head, as well as the “X” and the general citation location. Our citation is not as high quality as the other examples, but it is clear the model memorized the web address and number like features.

Consider the potential costs and benefits and to encourage these discussions within our communities. We are releasing these model and source code with the hope that they will be used for the benefit of all and not for immoral reasons (including those unmentioned herein).