Alias-Free ViT: Fractional Shift Invariance via Linear Attention

Hagay Michaeli
Technion
Haifa, Israel
hagaymi@campus.technion.ac.il

Daniel Soudry
Technion
Haifa, Israel
daniel.soudry@gmail.com

Abstract

Transformers have emerged as a competitive alternative to convnets in vision tasks, yet they lack the architectural inductive bias of convnets, which may hinder their potential performance. Specifically, Vision Transformers (ViTs) are not translation-invariant and are more sensitive to minor image translations than standard convnets. Previous studies have shown, however, that convnets are also not perfectly shift-invariant, due to aliasing in downsampling and nonlinear layers. Consequently, anti-aliasing approaches have been proposed to certify convnets translation robustness. Building on this line of work, we propose an Alias-Free ViT, which combines two main components. First, it uses alias-free downsampling and nonlinearities. Second, it uses linear cross-covariance attention that is shift-equivariant to both integer and fractional translations, enabling a shift-invariant global representation. Our model maintains competitive performance in image classification and outperforms similar-sized models in terms of robustness to adversarial translations.[1]

1 Introduction

Transformers, primarily designed for language modeling [58], have become dominant in vision tasks [26, 34]. Since they were originally designed for sequential data, their underlying attention mechanism is not sensitive to the locality of information in visual data. As a result, Vision Transformers (ViTs) exhibit a lack of shift-invariance, a shortcoming that becomes evident in cases where small image translations lead to significant deviations in output [25, 51].

To mitigate this gap, many studies have been conducted on the integration of convolutional priors, such as spatial hierarchy and shift-invariance, into ViT architectures. For example, approaches include hierarchical patch merging [39], the incorporation of convolutional layers [61], and the design of relative positional encodings [62]. Furthermore, some works have drawn parallels between self-attention and dynamic convolutions [1, 9, 27], motivating reinterpretations of the attention mechanism through a convolutional lens.

Despite being more spatially aware than transformers, convnets are not perfectly shift-invariant due to aliasing introduced by strided convolutions and pooling layers [4, 67]. This has led to a line of research that aims to restore shift-invariance, by methods including anti-aliasing filters [21, 23, 32, 40, 67, 71] and adaptive sampling techniques [6, 46]. Building on these advances, recent works have adapted such methods for transformer architectures. For example, Adaptive Polyphase Sampling (APS) has been employed to achieve cyclic shift-equivariance in ViTs [13, 47]. However, despite the latter approach efficiently guaranteeing shift-invariance for integer pixel cyclic shifts, it falls short in fractional (i.e., sub-pixel) shifts and “standard” shifts (i.e. imitating camera translation) [11, 40, 51].

[1] Our code is available at github.com/hmichaeli/alias_free_vit.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

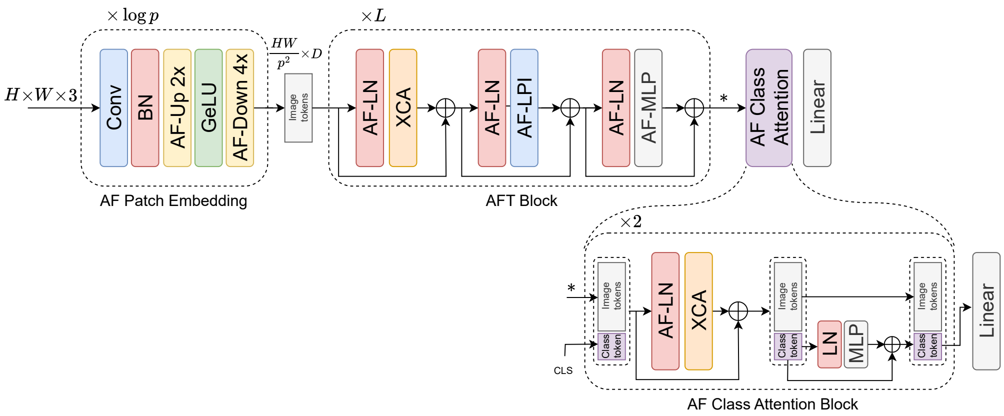


Figure 1: **Overview of the Alias-Free Vision Transformer (AFT) architecture.** The input image is first processed by an alias-free patch embedding module composed of convolutional layers (Conv), batch normalization (BN), and alias-free activation, composed of upsampling (AF-Up), GELU and downsampling (AF-Down). The result is reshaped to a token matrix form and fed through L Alias-Free Transformer blocks, each consisting of alias-free layer normalization (AF-LN), cross-covariance attention (XCA), alias-free local patch interaction (AF-LPI), and alias-free MLP (AF-MLP) layers, interconnected by residual connections. The result is concatenated with a learnable class token embedding and fed into two Alias-Free Class Attention blocks composed of an XCA layer and an MLP applied on the class token. The final representation is the updated class token, which is fed into a final linear classifier. Detailed explanations of each component are provided in Section 3.

As aliasing is primarily related to downsampling layers, which are not frequently used in ViTs, only few studies have been conducted on integrating aliasing-reduction techniques to achieve shift-invariance in ViTs. Qian et al. [42] propose plugging a low-pass filter post self-attention to reduce aliasing, however, this only provides a partial solution, as it does not resolve the inherent lack of shift-equivariance in self-attention and aliasing in other nonlinearities. Recent works study aliasing in latent diffusion models [2, 64, 70] that typically include attention layers, in order to improve their consistency i.e. in video generation. However, these works do not address the main issue in the self-attention operation. A similar problem may also hinder transformer neural operators which have recently become popular [28, 38, 52], as aliasing has been shown to be related to discretization errors [5, 44, 69].

Another emerging direction focuses on linear and softmax-free attention mechanisms, initially proposed to reduce the quadratic complexity of standard attention in large language models (LLMs) [7, 33, 59]. In the vision domain, models such as XCiT [16] and SimA [35] demonstrate that alternative attention formulations, e.g., using cross-covariance or linear attention, can maintain competitive performance without directly computing a full attention map. Beyond improved efficiency, we now show that such mechanisms enable designing a transformer-based architecture that is shift-equivariant, similar to convnets.

Contributions. In this paper

- We present in Section 2 a certain class of shift-equivariant attention layers (SEA), including linear attention and cross-covariance attention, which is useful for vision tasks.
- We design in Section 3 a shift-invariant, alias-free Vision Transformer (AFT) using cross-covariance attention and alias-free nonlinearities and show it has a competitive performance in image classification.
- We show in Section 4 that the AFT is robust ($\sim 99\%$ consistency) to fractional cyclic shifts, and significantly more robust to practical translations than other similar models, albeit with increased computational overhead due to the alias-free components.

2 Methods

2.1 Preliminaries: the Vision Transformer

The Vision Transformer (ViT) [15] transfers the Transformer architecture [58] from text to images by interpreting an image as a sequence of visual tokens. Given an input $x \in \mathbb{R}^{C\times H\times W}$, the image is partitioned into $N = \frac{HW}{p^2}$ non-overlapping patches of resolution $p \times p$. Each patch is flattened and projected by a learned linear layer into a D-dimensional embedding, forming $X \in \mathbb{R}^{N\times D}$. For classification, a learnable "class" token is prepended to X and later serves as a global representation, propagated into the classification module. Learnable *absolute* positional encodings $\mathcal{P} \in \mathbb{R}^{(N+1)\times D}$ are then added to compensate for the permutation-invariance of self-attention.

The resulting sequence $\tilde{X} = X + \mathcal{P}$ is processed by L identical Transformer encoder blocks. Each block is composed of a Multi-Head Self-Attention (MHSA) module with a two-layer Feed-Forward Network (FFN), interleaved with Layer Normalization (LN) and residual connections:

$$\hat{X}^{(\ell)} = \tilde{X}^{(\ell-1)} + \mathrm{MHSA}\Big(\mathrm{LN}\left(\tilde{X}^{(\ell-1)}\right)\Big), \tag{1}$$

$$\tilde{X}^{(\ell)} = \hat{X}^{(\ell)} + \mathrm{FFN}\Big(\mathrm{LN}\left(\hat{X}^{(\ell)}\right)\Big), \tag{2}$$

where $\ell = 1, \ldots, L$ and $\tilde{X}^{(0)} = \tilde{X}$.

Multi-Head Self-Attention. Each of the h heads linearly projects the input into queries, keys and values, $Q = XW_q,\ K = XW_k,\ V = XW_v$, with $W_q, W_k, W_v \in \mathbb{R}^{D\times d_h}$ and $d_h = D/h$. Self-Attention is then computed as

$$\mathrm{SA}\left(X\right) = \mathrm{softmax}\Big(QK^\top / \sqrt{d_h}\Big)\, V\,. \tag{3}$$

The outputs of all heads are concatenated and projected back to $\mathbb{R}^D$ by a final linear layer.

Feed-Forward Network. The FFN first expands the embedding dimension to $4D$, applies a GELU activation [30], and projects back:

$$\mathrm{FFN}\left(X\right) = W_2\, \mathrm{GELU}\left(W_1 X + b_1\right) + b_2, \tag{4}$$

with $W_1 \in \mathbb{R}^{4D\times D}$ and $W_2 \in \mathbb{R}^{D\times 4D}$.

2.2 Linear Attention

Notably, Equation (3) requires computing $QK^\top \in \mathbb{R}^{N\times N}$ explicitly, which has a quadratic cost in the number of tokens. This has motivated many linear complexity variants of self-attention [7, 33, 59]. In vision, SimA removes softmax entirely by maintaining stability using ℓ_1–normalized Q and K [35]. XCiT mixes channels instead of tokens with cross-covariance attention (XCA) [16],

$$\mathrm{XCA}(X) = V\, \mathrm{softmax}\left(\hat{K}^\top\, \hat{Q}/\tau\right), \tag{5}$$

where $\hat{Q}, \hat{K}$ are ℓ_2-normalized along the token dimension and $\tau > 0$ is a learnable temperature.

2.3 Alias-Free Vision Transformer

Next, we describe our proposed model, replacing every ViT component that is not shift-equivariant with a modification that restores equivariance. For brevity, the analysis is given for a one-dimensional signal. The same principles can be applied in the two-dimensional case by viewing the sequence of N tokens in a two-dimensional representation, namely $X \in \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}\times D}$.[2]

Similar to Karras et al. [32] and Michaeli et al. [40], we view the input as a discrete sampling of an underlying band-limited continuous signal. The main difference from a standard convnet emerges

[2] In practice two-dimensional signals tokens are also stacked into a one-dimensional sequence (formally the row-stack). Nevertheless, shift-equivariance in the two dimensions is still maintained under this representation.

after tokenization: the patch-embedding matrix $X \in \mathbb{R}^{N \times D}$ stores the sequence length N along its first axis, and the embedding dimension D in the second axis. Thus, the spatial and channel roles are swapped compared to convnet feature maps, where the channel index comes first and spatial indices follow. Throughout our analysis, we therefore interpret the D columns of X as D "channels" of a length-N one-dimensional signal. Maintaining shift-equivariance then amounts to ensuring that each column transforms equivariantly under fractional translations in the continuous domain.

Shift invariance and equivariance. We reuse the definitions of Michaeli et al. [40]. Let $x[n]$ be a discrete signal, T its sample spacing, and $z(t)$ the unique $\frac{1}{2T}$-band-limited signal with $x[n] = z(nT)$. For any (possibly non-integer) shift $\Delta \in \mathbb{R}$,

$$\tau_{\Delta} x[n] \; = \; z(nT + \Delta).$$

An operator f is *shift-equivariant* if $f(\tau_{\Delta} x) = \tau_{\Delta} f(x)$ and *shift-invariant* if $f(\tau_{\Delta} x) = f(x)$ for all x and Δ.

In some cases, we may claim that a value is shift-equivariant. This is a slight abuse of the definition to say the overall operator computing this value is shift-equivariant w.r.t. the input signal.

Patch embedding. Algebraically, the tokenization process described in Section 2.1 is equivalent to a convolution with kernel size p, stride p, and D output channels. By separating this stride-p convolution into a stride-1 convolution followed by an alias-free downsampling [21], the composite operator becomes shift-equivariant [32, 40, 67]. However, this requires inserting a single low-pass filter with cut-off $1/p$, which would severely attenuate high-frequency content, especially for large patches. Instead, we employ a convolutional patch-embedding that replaces the single stride-p layer with a short convnet of progressively smaller strides, often used in hybrid models [16, 24, 61, 66]. Here, we can avoid aliasing by plugging a low-pass filter with a larger cut-off before each downsampling layer, and by using an alias-free activation function [32, 40]. This gradual approach still enables the network to learn high-frequency features, despite the anti-aliasing components.

Positional encoding. Absolute positional encoding injects the global index of each token and therefore breaks shift-equivariance. Several relative schemes have been proposed that depend only on pairwise offsets and thus preserve shift-equivariance at the token level [8, 39, 50]. These methods, however, do not guarantee equivariance to pixel-level translations, as they cause the patch contents themselves to change. Notably, the convolutional patch embedding already breaks permutation invariance of the tokens, possibly reducing the need for additional positional signals. Moreover, recent studies demonstrate that hybrid transformers can learn effectively without any explicit positional encoding [3, 68]. Consistent with this observation, we find empirically (Section 4.4) that positional encoding may be unnecessary in architectures that include convolutional layers inside the transformer blocks, such as XCiT [16].

Shift-Equivariant Attention. We next show a class of attention operations that, by removing the softmax in Equation (3), are shift-equivariant. This primarily includes the linear attention, which, although still less common, is attractive for its lower complexity [33, 54] and yields competitive vision results [35]. Formally,

$$\text{SEA}\left(X\right) \; = \; Qf\left(K^{\top}V\right) \, , \tag{6}$$

where we keep the existing notation $Q = XW_q, \; K = XW_k, \; V = XW_v$, and let $f : \mathbb{R}^{D \times D} \rightarrow \mathbb{R}^{D \times D}$ be an arbitrary function.

We now establish the desired property in three steps.

Proposition 1. *Q, K and V are shift-equivariant.*

Proof. Each column of Q, K or V is a fixed linear combination of the columns of X. As the columns of X are merely the channels of the same signals, they translate together, and any linear combination of them is also shift-equivariant.

□

Proposition 2. *$f\left(K^{\top}V\right)$ is shift-invariant.*

Proof sketch. Consider the matrix entry $\left(K^\top V\right)_{ij} = K_i^\top V_j$, where K_i and V_j denote columns. By Parseval's theorem,

$$K_i^\top V_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{K}_i(\omega)\, \hat{V}_j^*(\omega)\, d\omega.$$

A fractional translation by τ multiplies both Fourier transforms by the same phase factor $e^{j\omega\tau}$, which cancels in the product. Hence, every entry of $K^\top V$ remains unchanged due to a translation (see formal proof in appendix A). An important observation is that since $K^\top V$ is shift-*invariant*, any matrix operation can be applied on it without compromising this property.

Proposition 3. *$V' = Qf\left(K^\top V\right)$ is shift-equivariant.*

Proof. Column j of V' satisfies

$$V'_j \;=\; \sum_{i=1}^{D} Q_i\, f\left(K^\top V\right)_{ij},$$

i.e. a sum of *shift-equivariant* columns Q_i (Proposition 1) scaled by *shift-invariant* coefficients $f\left(K^\top V\right)_{ij}$ (Proposition 2). The resulting column is therefore shift-equivariant. □

As mentioned above, since $K^\top V$ is shift-*invariant*, any matrix operation f can be applied on it without compromising the overall shift-*equivariance*. By the same argument, $K^\top Q$ is also shift-invariant; thus a row-wise softmax on $K^\top Q$ as used in XCA (Equation (5)) is permissible. Note that in XCA, the columns of Q and K are ℓ_2-normalized along the token dimension, which as well maintains shift-equivariance [40].

MLP. The MLP linear layers apply the same linear transformation on all tokens, maintaining shift-equivariance as in Proposition 1. However, pointwise nonlinearities induce aliasing that breaks fractional shift-equivariance. This can be solved similarly to [40] by an alias-free activation function, which includes upsampling before the nonlinearity and downsampling back after low-pass filtering.

Layer normalization. Per-token LayerNorm rescales each column differently and breaks equivariance. The same problem has been addressed by Michaeli et al. [40], by using a global variant, namely

$$\hat{X}_{ij} = \frac{X_{ij} - \mu}{\sqrt{\sigma^2 + \epsilon}}, \quad \mu_i = \tfrac{1}{D} \sum_j X_{ij},\ \sigma^2 = \tfrac{1}{ND} \sum_{i,j} (X_{ij} - \mu_i)^2. \tag{7}$$

where μ is computed per token and σ^2 per layer.

Class token. Prepending the "class" token interferes with the signal representation of the columns of the embedding matrix, and breaks shift-equivariance. A simple solution is to instead construct a global representation using global average pooling over the embedding dimension after the last transformer block, i.e.

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i^{(L)} \in \mathbb{R}^D, \tag{8}$$

similar to the common approach in convnets and a few other ViTs [24, 39, 65]. This, assuming shift-equivariance is maintained, yields a shift-invariant global representation [40].

Class Attention. Some ViTs append the class token only after L patch-only blocks and update it through a few class-attention (CA) layers [10, 16, 56]. These layers usually fix the patch embeddings and update only the class token embedding by attending it to itself and to the frozen patch embeddings. We find that when using SEA after concatenating the class token, Propositions 1 to 3 still hold w.r.t. the patch tokens, and additionally the class token remains shift-invariant (see proof in appendix A). Retaining similarity to the original class attention blocks, we only propagate the class embedding through the MLP, which also prevents aliasing in the nonlinearity. We use this approach instead of global average pooling since we find it performs slightly better.

3 Implementation

We implement the Alias-Free Vision Transformer (AFT) based on XCiT architecture [16]. XCiT replaces standard self-attention with cross-covariance attention, described in Equation (5). This, by Proposition 3, preserves shift-equivariance. The remaining modifications focus on eliminating aliasing in the patch-embedding, local patch interaction (LPI), and MLP blocks. The overall model is presented in Figure 1.

Conv Patch Embedding. XCiT patch embedding (PE) module is a sequence of blocks composed of strided convolution, batch normalization, and GELU. Similar to [40], we replace the strided convolution with a stride of 1 and insert an alias-free downsampling at the end of the block, implemented by truncation of high-frequencies in the Fourier domain, using Fast Fourier Transform (FFT) [22, 45]. We replace zero-padding with circular padding. We use alias-free activation functions by wrapping the GELU activation with upsampling and alias-free downsampling layers [32, 40]. In contrast to [40], we find that replacing the GELU with a polynomial activation to get perfect shift-invariance degrades the model performance significantly. Conversely, we find that keeping GELU affects the translation-robustness marginally. We do not add positional encodings and do not prepend a class token at this stage; the class token is appended only before the class attention blocks.

AFT Block. Following the patch embedding, XCiT is composed of a sequence of Blocks, each consisting of three components: cross-covariance attention (XCA), local patch interaction (LPI), and MLP. As argued in Section 2.3, the XCA is already shift-equivariant. The LPI consists of two convolutional layers, batch normalization, and activation layers, and we treat each of them as in the PE, forming an alias-free version we denote AF-LPI. The MLP applies a shared two-layer FFN on each token, and can be viewed as a convolutional layer with kernel size 1. Therefore, the only required modification to form the alias-free variant (AF-MLP) is converting the activation function into an alias-free activation. We replace all LayerNorm instances, applied before XCA, LPI, and MLP, with the alias-free layer norm described in Section 2.3.

Classification. After the last AFT block we append a learnable class token and apply two *AF class attention* blocks. Each block consists of AF-LN, an XCA layer, and an MLP applied only to the class token. Residual connections are used around XCA and the MLP, mirroring the AFT blocks. By the SEA properties, the patch tokens remain shift-equivariant and the class token is shift-invariant (see Proposition 4). The final prediction is obtained by a linear classifier applied to the class token.

4 Experiments

We evaluate our Alias-Free Transformer (AFT) on the ImageNet dataset [12] and compare its accuracy and shift consistency with the baseline XCiT model. We additionally compare our method with the adaptive polyphase sampling (APS) approach [13, 47], which we implement by replacing the strided-convolutions in the PE with stride-1 convolutions followed by APSPool [6], and using the standard class attention blocks for classification (maintaining integer shift-invariance). We use the nano and small XCiT versions with 12 layers and patch-size 16, processing inputs of size 224×224.

We train all models for 400 epochs, following the XCiT training recipe [16], using PyTorch [41], on a single machine with 8 × NVIDIA RTX A6000. We observe a slight improvement for the AF models when training with a smaller batch size; therefore, we reduce the batch size from 1024 to 512 for the AF versions (See additional details in Appendix D.1).

In Sections 4.1 and 4.2 we evaluate the baseline, APS, and AFT models using cyclic translations and implement m/n-fractional translation by translating in m pixels the n-upsampled image using sinc-interpolation, as our and the APS models were initially designed under those assumptions. In Section 4.3 we use more "realistic" types of translations, and add additional publicly available ViTs to the comparison.

4.1 Accuracy and shift consistency

The results in Table 1 (left) show the classification accuracy and consistency, defined as the percentage of validation samples whose prediction did not change after a random translation. The alias-free

Table 1: **ImageNet accuracy and cyclic shift consistency. Left**: Shift consistency is defined as the percentage of test samples that did not change their prediction following a random translation. The alias-free models have similar accuracy to the baseline models, and much higher consistency in both integer and half-pixel translations. **Right**: Adversarial accuracy is defined as the percentage of correctly classified samples in the worst case at the corresponding grid (Equation (9)). See Appendix B for additional results.

Model	Test accuracy	Integer shift consist.	Half-pixel shift consist.	Adversarial integer grid	Adversarial half-pixel grid
XCiT-Nano (Baseline)	70.4	83.7	82.0	50.9	52.9
XCiT-Nano-APS	68.4	**100.0**	87.5	68.4	62.9
XCiT-Nano-AF (ours)	**70.5**	99.2	**98.7**	**69.9**	**69.5**
XCiT-Small (Baseline)	**82.0**	91.4	89.8	70.9	71.3
XCiT-Small-APS	81.3	**100.0**	94.0	**81.3**	78.2
XCiT-Small-AF (ours)	81.8	99.5	**99.4**	81.3	**81.1**

models have similar accuracy to the baseline models, and much higher consistency in both integer and half-pixel translations. The APS models, on the other hand, achieve near-100% consistency under integer translations, as expected. However, they have a more modest improvement in consistency to half-pixel shifts.

4.2 Adversarial robustness

To show a practical implication of shift consistency, we ask whether an adversary, free to translate the image within a prespecified grid, can find any shift that flips the label. We define adversarial accuracy as the fraction of images that are classified correctly in the worst-case at this grid. In Table 1 (right), we show results for cyclic integer and half-pixel translations, reporting adversarial accuracy over the following grids

$$T_{\text{integer}} = \left\{ (i,j) \,|\, -6 \le i,j \le 6 \right\} \qquad T_{\text{half}} = \left\{ \left(\tfrac{i}{2}, \tfrac{j}{2}\right) \,\middle|\, -6 \le i,j \le 6 \right\} \tag{9}$$

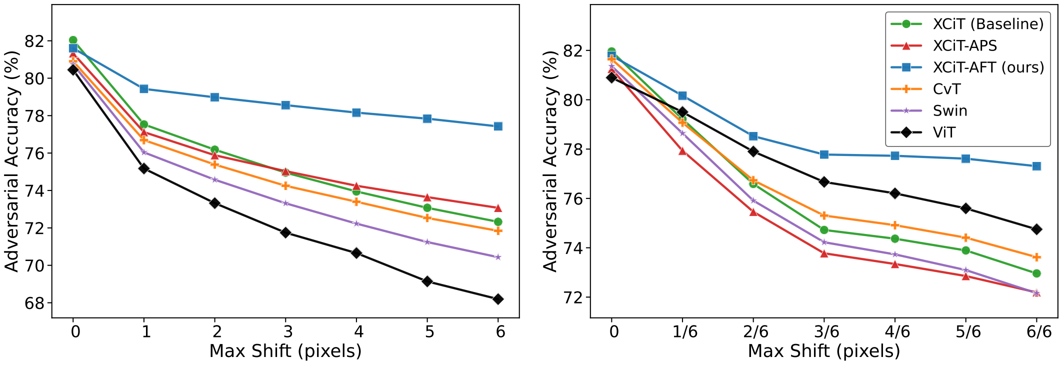
The AFT models maintain high accuracy under both integer and half-pixel attacks, having 2% relative accuracy reduction in the nano version and less than 1% reduction in the small model. This is in contrast to the baseline models, with relative accuracy reductions of 25% and 14% in the nano and small models respectively. As expected, the APS models maintain their accuracy under the integer grid adversarial attacks. Their accuracy under half-pixel grid attacks decreases slightly in comparison to the baseline models. The reason for this is that the APS is invariant to any two half-pixel translations, as these differ in exact integer translations. We expect the APS accuracy to decrease more under arbitrary fractional translations, as can be seen in Section 4.3 and [40]. See Appendix B for additional results.

4.3 Robustness to realistic shifts

The experiments above use cyclic translations, which may leave unnatural image artifacts in realistic cases (where the input is not a sample of some periodic signal). We therefore test two more realistic perturbations and measure adversarial accuracy exactly as in Section 4.2. See Appendix C for visualizations of the used translations.

- **Crop-shifts.** The image is first center-cropped, then cropped in offsets (δ_x, δ_y) with $|\delta_x|, |\delta_y| \le s$. This mimics a camera translation that moves content out of view instead of wrapping it around.
- **Bilinear fractional shifts.** To simulate sub-pixel motion, we translate the image by $(\delta_x/6, \delta_y/6)$ with $|\delta_x|, |\delta_y| \le s$, using bilinear interpolation. Here, we leave an edge of one pixel of the original image in each direction to avoid edge artifacts.

We compare the baseline, APS and AFT XCiT-Small models with other publicly available trained models, in similar scale as XCiT-small (26M) (indicates number of parameters): CvT-13 (20M), Swin-T (28M), and ViT-Base (86M). We repeat these experiments with s (max-shift) in the range 0



(a) Adversarial "Crop-shifts"

(b) Adversarial Bilinear fractional shifts

Figure 2: **ImageNet adversarial accuracy under realistic translations.** Adversarial accuracies under (a) "Crop-shifts," simulating camera translations, and (b) "Bilinear fractional shifts," simulating realistic sub-pixel image translations. The Alias-Free Transformer (AFT) consistently outperforms baseline XCiT, APS, and other vision transformer variants (CvT, Swin and ViT), demonstrating superior robustness against realistic translations.

to 6. The results in Figure 2 show our model has improved robustness to both these types of realistic shifts, despite not being specifically designed for them. Among the additional baselines, the vanilla ViT degrades the most under crop-shifts, whereas under bilinear fractional shifts it is surprisingly competitive and in fact better than the hybrid backbones (XCiT, Swin, CvT).

4.4 Ablation study

We conduct an ablation study on XCiT-Nano to evaluate the impact of each of the alias-free modifications on the model performance. We train the baseline model on ImageNet with one specific change, and report the results in Table 2. Surprisingly, replacing the layer norm with alias-free layer-norm and replacing class attention with average pooling cause a much larger degradation in accuracy in comparison to the marginal degradation in the final alias-free model. On the other hand, removing the positional encoding leads to a small improvement in accuracy, emphasizing that it may be unnecessary in hybrid architectures.

Table 2: **Ablation study on alias-free components of XCiT-Nano (ImageNet).** Evaluation of isolated alias-free modifications to the baseline model. Alias-free layer normalization (AF-LayerNorm) and replacing class-attention with average pooling (AvgPool) result in notable accuracy degradation individually. Removing positional encoding slightly improves performance. The final combined alias-free model retains near-baseline accuracy.

Model	Accuracy	Change (%)
Baseline	70.4	–
Cyclic convolution	70.4	+0.0%
AvgPool	69.1	–1.8%
AF-LayerNorm	69.6	–1.1%
No positional encoding	70.7	+0.4%
AF (AvgPool)	70.4	+0.0%
AF (AF Class Attention)	70.6	+0.3%

Table 3: **Training runtime.** Train time was measured on 8 × NVIDIA RTX A6000 using batch size 1024 for the baseline model and batch size 512 in the APS and AF models, due to memory constraints.

Model	Train time [hours]
XCiT-Small (Baseline)	69
XCiT-Small-APS	98
XCiT-Small-AF (ours)	487

5 Related work

A few studies have shown a broad effect of aliasing in deep neural networks, e.g., breaking shift-equivariance in convnets [4, 67], inconsistencies in image generation [32, 64, 70], and breaking continuous-discrete equivalence in neural operators [5, 19, 55, 69].

Shift invariant convnets. For a long time, convolutional neural networks have held dominance in vision thanks to their useful inductive biases, including translation invariance. However, previous studies have shown their output can in fact change in a large extent due to small translations [4, 17]. This has led to extensive research to find the root causes and resolve this problem. Azulay and Weiss [4], Zhang [67] have identified shift-invariance breaks as a result of aliasing in downsampling and nonlinear layers. Consequently, other studies suggested solving this problem by plugging a low-pass filter before downsampling [23, 29, 67], and preventing aliasing in nonlinearities by using smooth activation functions [31, 40, 57] and by applying activations after upsampling [32, 40, 60]. Other works suggested maintaining shift-invariance in convnets by downsampling on adaptive grids [6, 46]. Specifically, the adaptive sampling method (APS) [6, 46] has been shown to retain perfect consistency to integer cyclic translations, while the anti-aliasing approach maintains consistency in fractional translations as well. Worth mentioning here are recent works that propose transforming the input into a “canonic” shift-invariant representation [11, 51], which theoretically makes equivariance of the following neural network unnecessary.

Shift invariant transformers. Vision transformers have gained dominance despite not having the convnet priors and being even more sensitive to image translations. Some studies have proposed hierarchical ViT architectures similar to convnets [14, 18, 39, 49, 62], and “hybrid” architectures including convolutional layers directly [61]. Furthermore, some works have drawn parallels between self-attention and dynamic convolutions [1, 9, 27], motivating reinterpretations of the attention mechanism through a convolutional lens. Few studies have taken inspiration from these studies aiming to retain shift-invariance in convnets and implemented their ideas into ViTs. Qian et al. [42] proposes plugging a low-pass filter post self-attention to reduce aliasing, partially improving consistency similar to Zhang [67]. Ding et al. [13], Rojas-Gomez et al. [47] adapt the adaptive sampling method (APS) into ViT layers, certifying consistency to integer cyclic translations. Other studies proposed more general framework for group equivariant attention [48, 63]. Yet, to the best of our knowledge, no other work has dealt with the invariance of ViTs to fractional shifts.

Linear attention. The standard Transformer architecture relies on softmax-based attention [58], characterized by quadratic computational complexity in the number of tokens. To overcome scalability limitations, linear and kernel-based attention mechanisms have been proposed [7, 33], substantially reducing complexity while maintaining performance. For example, linear attention leverages a linear approximation of the softmax kernel, achieving significant efficiency gains [59]. In vision, models like SimA [35] and XCiT [16] utilize simplified normalization schemes to replace the expensive softmax operation, enabling to avoid a direct computation of full attention maps.

6 Discussion and limitations

In this paper, we propose a shift-invariant alias-free vision transformer by introducing a class of shift-equivariant attention operations. We show that the AFT maintains competitive accuracy and superior robustness to fractional shifts, compared to other ViTs. We next discuss a few of our model limitations.

Polynomial activation function Michaeli et al. [40] propose replacing nonlinear activation functions, such as GELU, with polynomial approximations. This mathematically ensures the overall activation layer (including upsampling) is shift-equivariant w.r.t. continuous domain, namely perfectly consistent to fractional shifts. In our experiments, we observe that this leads to a significant reduction in performance, which is caused specifically due to the activation replacement in the patch-embedding stage (see Appendix B). On the other hand, we observe that the filtered activation function using GELU leads to a rather small reduction in consistency. Notably, the certified consistency is limited to cyclic shifts and fractional shifts performed by sinc-interpolation, both induce artifacts that do not appear in natural images. We find that similar to the AFC, our model also has significant improvement in robustness to realistic translations despite the imperfect consistency to cyclic shifts.

Runtime performance The alias-free modifications we perform in our model to attain shift invariance, despite not requiring any additional parameter, cause a substantial runtime increase, as shown in Table 3. This is mainly due to the downsampling and upsampling, which are implemented in the Fourier domain using FFT, similar to other works [22, 23, 40, 43, 70], seemingly underoptimized for GPU as of today [20, 53].

Acknowledgments and Disclosure of Funding

The research of DS was funded by the European Union (ERC, A-B-C-Deep, 101039436). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (ERCEA). Neither the European Union nor the granting authority can be held responsible for them. DS also acknowledges the support of the Schmidt Career Advancement Chair in AI.

References

[1] Jean-Marc Andreoli. Convolution, attention and structure embedding. May 2019. URL `https://arxiv.org/abs/1905.01289v5`. arXiv: 1905.01289.

[2] Md Fahim Anjum. Advancing Diffusion Models: Alias-Free Resampling and Enhanced Rotational Equivariance. November 2024. URL `https://arxiv.org/abs/2411.09174v1`. arXiv: 2411.09174.

[3] Bouzid Arezki, Fangchen Feng, and Anissa Mokraoui. Convolutional Transformer-Based Image Compression. In *2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 154–159, September 2023. doi: 10.23919/SPA59660.2023.10274433. URL `http://arxiv.org/abs/2409.04118`. arXiv:2409.04118 [eess].

[4] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20:1–25, 2019. URL `https://youtu.be/MpUdRacvkWk`. arXiv: 1805.12177v4 ISBN: 1805.12177v4.

[5] Francesca Bartolucci, Emmanuel de Bézenac, Bogdan Raonić, Roberto Molinaro, Siddhartha Mishra, and Rima Alaifari. Representation Equivalent Neural Operators: a Framework for Alias-free Operator Learning. *Advances in Neural Information Processing Systems*, 36, May 2023. ISSN 10495258. URL `https://arxiv.org/abs/2305.19913v2`. arXiv: 2305.19913 Publisher: Neural information processing systems foundation.

[6] Anadi Chaman and Ivan Dokmanić. Truly shift-invariant convolutional neural networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3772–3782, November 2020. ISSN 10636919. doi: 10.48550/arxiv.2011.14214. URL `https://arxiv.org/abs/2011.14214v4`. arXiv: 2011.14214 Publisher: IEEE Computer Society ISBN: 9781665445092.

[7] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking Attention with Performers, November 2022. URL `http://arxiv.org/abs/2009.14794`. arXiv:2009.14794 [cs].

[8] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional Positional Encodings for Vision Transformers, February 2023. URL `http://arxiv.org/abs/2102.10882`. arXiv:2102.10882 [cs].

[9] Jean Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the Relationship between Self-Attention and Convolutional Layers. *8th International Conference on Learning Representations, ICLR 2020*, November 2019. URL `https://arxiv.org/abs/1911.03584v2`. arXiv: 1911.03584 Publisher: International Conference on Learning Representations, ICLR.

[10] Marco Cotogni, Fei Yang, Claudio Cusano, Andrew D. Bagdanov, and Joost van de Weijer. Exemplar-free Continual Learning of Vision Transformers via Gated Class-Attention and Cascaded Feature Drift Compensation, July 2023. URL http://arxiv.org/abs/2211.12292. arXiv:2211.12292 [cs].

[11] Berken Utku Demirel and Christian Holz. Shifting the Paradigm: A Diffeomorphism Between Time Series Data Manifolds for Achieving Shift-Invariancy in Deep Learning, February 2025. URL http://arxiv.org/abs/2502.19921. arXiv:2502.19921 [cs].

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, June 2009. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206848.

[13] Peijian Ding, Davit Soselia, Thomas Armstrong, Jiahao Su, and Furong Huang. Reviving Shift Equivariance in Vision Transformers. June 2023. URL https://arxiv.org/abs/2306.07470v1. arXiv: 2306.07470.

[14] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows, January 2022. URL http://arxiv.org/abs/2107.00652. arXiv:2107.00652 [cs].

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021 - 9th International Conference on Learning Representations*, October 2020. URL https://arxiv.org/abs/2010.11929v2. arXiv: 2010.11929 Publisher: International Conference on Learning Representations, ICLR.

[16] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jegou. XCiT: Cross-Covariance Image Transformers, June 2021. URL http://arxiv.org/abs/2106.09681. arXiv:2106.09681 [cs].

[17] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the Landscape of Spatial Robustness. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:3218–3238, December 2017. doi: 10.48550/arxiv.1712.02779. URL https://arxiv.org/abs/1712.02779v4. arXiv: 1712.02779 Publisher: International Machine Learning Society (IMLS) ISBN: 9781510886988.

[18] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale Vision Transformers. *Proceedings of the IEEE International Conference on Computer Vision*, pages 6804–6815, April 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.00675. URL https://arxiv.org/abs/2104.11227v1. arXiv: 2104.11227 Publisher: Institute of Electrical and Electronics Engineers Inc. ISBN: 9781665428125.

[19] V. Fanaskov and I. Oseledets. Spectral Neural Operators, May 2022. URL http://arxiv.org/abs/2205.10573. arXiv:2205.10573 [cs, math].

[20] Daniel Y. Fu, Hermann Kumbong, Eric Nguyen, and Christopher Ré. FlashFFTConv: Efficient Convolutions for Long Sequences with Tensor Cores, November 2023. URL http://arxiv.org/abs/2311.05908. arXiv:2311.05908 [cs].

[21] Julia Grabinski. FrequencyLowCut Pooling-Plug & Play against Catastrophic Overfitting. Technical report. URL https://github.com/GeJulia/flc_pooling. arXiv: 2204.00491v2.

[22] Julia Grabinski, Steffen Jung, Janis Keuper, and Margret Keuper. FrequencyLowCut Pooling - Plug and Play Against Catastrophic Overfitting. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13674, pages 36–57. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-19780-2 978-3-031-19781-9. doi: 10.1007/978-3-031-19781-9_3. URL https://link.

springer.com/10.1007/978-3-031-19781-9_3. Series Title: Lecture Notes in Computer Science.

[23] Julia Grabinski, Janis Keuper, and Margret Keuper. Fix your downsampling ASAP! Be natively more robust via Aliasing and Spectral Artifact free Pooling, July 2023. URL http://arxiv.org/abs/2307.09804. arXiv:2307.09804 [cs].

[24] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. LeViT: A Vision Transformer in ConvNet's Clothing for Faster Inference. pages 12259–12269, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Graham_LeViT_A_Vision_Transformer_in_ConvNets_Clothing_for_Faster_Inference_ICCV_2021_paper.html.

[25] Suriya Gunasekar. Generalization to translation shifts: a study in architectures and augmentations, July 2022. URL https://arxiv.org/abs/2207.02349v1. arXiv: 2207.02349.

[26] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A Survey on Visual Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (1):87–110, December 2020. doi: 10.1109/TPAMI.2022.3152247. URL http://arxiv.org/abs/2012.12556. arXiv: 2012.12556v6 Publisher: IEEE Computer Society.

[27] Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming Ming Cheng, Jiaying Liu, and Jingdong Wang. On the Connection between Local Attention and Dynamic Depth-wise Convolution. *ICLR 2022 - 10th International Conference on Learning Representations*, 3, June 2021. URL https://arxiv.org/abs/2106.04263v5. arXiv: 2106.04263 Publisher: International Conference on Learning Representations, ICLR.

[28] Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A General Neural Operator Transformer for Operator Learning, June 2023. URL http://arxiv.org/abs/2302.14376. arXiv:2302.14376 [cs].

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. February 2015. URL http://arxiv.org/abs/1502.01852. arXiv: 1502.01852.

[30] Dan Hendrycks and Kevin Gimpel. GAUSSIAN ERROR LINEAR UNITS (GELUS). arXiv: 1606.08415v4.

[31] Md Tahmid Hossain, Shyh Wei Teng, Ferdous Sohel, and Guojun Lu. Anti-aliasing Deep Image Classifiers using Novel Depth Adaptive Blurring and Activation Function, October 2021. URL https://arxiv.org/abs/2110.00899v1. arXiv: 2110.00899.

[32] 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*, volume 34, pages 852–863. Curran Associates, Inc., 2021. URL https://arxiv.org/abs/2106.12423.

[33] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Francois Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. 2020. arXiv: 2006.16236v3.

[34] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in Vision: A Survey. *ACM Computing Surveys*, 54(10), January 2021. ISSN 15577341. doi: 10.1145/3505244. URL https://arxiv.org/abs/2101.01169v5. arXiv: 2101.01169 Publisher: Association for Computing Machinery.

[35] Soroush Abbasi Koohpayegani and Hamed Pirsiavash. SimA: Simple Softmax-Free Attention for Vision Transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2607–2617, January 2024.

[36] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, December 2013. doi: 10.1109/ICCVW.2013.77. URL https://ieeexplore.ieee.org/document/6755945.

[37] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images.

[38] Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for Partial Differential Equations' Operator Learning. Technical report. URL `https://github.com/BaratiLab/OFormer`.

[39] 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. *Proceedings of the IEEE International Conference on Computer Vision*, pages 9992–10002, March 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.00986. URL `https://arxiv.org/abs/2103.14030v2`. arXiv: 2103.14030 Publisher: Institute of Electrical and Electronics Engineers Inc. ISBN: 9781665428125.

[40] Hagay Michaeli, Tomer Michaeli, and Daniel Soudry. Alias-Free Convnets: Fractional Shift Invariance via Polynomial Activations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16333–16342, 2023. doi: 10.48550/arXiv.2303.08085. ISSN: 23318422.

[41] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017.

[42] Shengju Qian, Hao Shao, Yi Zhu, Mu Li, and Jiaya Jia. Blending Anti-Aliasing into Vision Transformer. *Advances in Neural Information Processing Systems*, 34:5416–5429, December 2021. URL `https://github.com/amazon-research/anti-aliasing-transformer`.

[43] Md Ashiqur Rahman and Raymond A. Yeh. Truly Scale-Equivariant Deep Nets with Fourier Layers, November 2023. URL `http://arxiv.org/abs/2311.02922`. arXiv:2311.02922 [cs].

[44] Bogdan Raonić, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional Neural Operators for robust and accurate learning of PDEs. February 2023. URL `https://arxiv.org/abs/2302.01178v3`. arXiv: 2302.01178 ISBN: 2302.01178v3.

[45] Oren Rippel, Jasper Snoek, and Ryan P. Adams. Spectral Representations for Convolutional Neural Networks, June 2015. URL `http://arxiv.org/abs/1506.03767`. arXiv:1506.03767 [stat].

[46] Renan A. Rojas-Gomez, Teck-Yian Lim, Alexander G. Schwing, Minh N. Do, and Raymond A. Yeh. Learnable Polyphase Sampling for Shift Invariant and Equivariant Convolutional Networks. October 2022. doi: 10.48550/arxiv.2210.08001. URL `https://arxiv.org/abs/2210.08001v1`. arXiv: 2210.08001.

[47] Renan A. Rojas-Gomez, Teck-Yian Lim, Minh N. Do, and Raymond A. Yeh. Making Vision Transformers Truly Shift-Equivariant. May 2023. URL `https://arxiv.org/abs/2305.16316v1`. arXiv: 2305.16316.

[48] David W. Romero and Jean-Baptiste Cordonnier. Group Equivariant Stand-Alone Self-Attention For Vision, March 2021. URL `http://arxiv.org/abs/2010.00977`. arXiv:2010.00977 [cs].

[49] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles, June 2023. URL `http://arxiv.org/abs/2306.00989`. arXiv:2306.00989 [cs].

[50] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-Attention with Relative Position Representations, April 2018. URL `http://arxiv.org/abs/1803.02155`. arXiv:1803.02155 [cs].

[51] Ofir Shifman and Yair Weiss. Lost in Translation: Modern Neural Networks Still Struggle With Small Realistic Image Transformations. April 2024. URL `https://arxiv.org/abs/2404.07153v1`. arXiv: 2404.07153.

[52] Benjamin Shih, Ahmad Peyvan, Zhongqiang Zhang, and George Em Karniadakis. Transformers as Neural Operators for Solutions of Differential Equations with Finite Regularity, May 2024. URL `http://arxiv.org/abs/2405.19166`. arXiv:2405.19166 [cs].

[53] Matteo Spanio and Antonio Rodà. TorchFX: A modern approach to Audio DSP with PyTorch and GPU acceleration, April 2025. URL `http://arxiv.org/abs/2504.08624`. arXiv:2504.08624 [eess].

[54] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive Network: A Successor to Transformer for Large Language Models. July 2023. URL `https://arxiv.org/abs/2307.08621v4`. arXiv: 2307.08621.

[55] Karn Tiwari, N. M. Anoop Krishnan, and Prathosh A. P. CoNO: Complex Neural Operator for Continuous Dynamical Systems, October 2023. URL `http://arxiv.org/abs/2310.02094`. arXiv:2310.02094 [nlin, physics:physics].

[56] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with Image Transformers, April 2021. URL `http://arxiv.org/abs/2103.17239`. arXiv:2103.17239 [cs] version: 2.

[57] Cristina Vasconcelos, Hugo Larochelle, Vincent Dumoulin, Nicolas Le Roux, and Ross Goroshin. An Effective Anti-Aliasing Approach for Residual Networks. November 2020. doi: 10.48550/arxiv.2011.10675. URL `https://arxiv.org/abs/2011.10675v1`. arXiv: 2011.10675.

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. June 2017. arXiv: 1706.03762.

[59] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-Attention with Linear Complexity, June 2020. URL `http://arxiv.org/abs/2006.04768`. arXiv:2006.04768 [cs].

[60] Emmy S. Wei. Aliasing-Free Convolutional Nonlinear Networks Using Implicitly Defined Functions. March 2022. URL `https://hal.archives-ouvertes.fr/hal-03475613`.

[61] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22–31, October 2021.

[62] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10041, 2021.

[63] Renjun Xu, Kaifan Yang, Ke Liu, and Fengxiang He. $E(2)$-Equivariant Vision Transformer, July 2023. URL `http://arxiv.org/abs/2306.06722`. arXiv:2306.06722 [cs].

[64] Cuihong Yu, Cheng Han, and Chao Zhang. DMFFT: improving the generation quality of diffusion models using fast Fourier transform. *Scientific Reports*, 15(1):10200, March 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-94381-8. URL `https://www.nature.com/articles/s41598-025-94381-8`. Publisher: Nature Publishing Group.

[65] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. MetaFormer Is Actually What You Need for Vision, July 2022. URL `http://arxiv.org/abs/2111.11418`. arXiv:2111.11418 [cs].

[66] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis E H Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. URL `https://github.com/yitu-opensource/T2T-ViT`. arXiv: 2101.11986v3.

[67] Richard Zhang. Making convolutional networks shift-invariant again. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:12712–12722, 2019. URL `https://richzhang.github.io/antialiased-cnns/`. arXiv: 1904.11486 ISBN: 9781510886988.

[68] Tianxiao Zhang, Wenju Xu, Bo Luo, and Guanghui Wang. Depth-Wise Convolutions in Vision Transformers for Efficient Training on Small Datasets. *Neurocomputing*, 617:128998, February 2025. ISSN 09252312. doi: 10.1016/j.neucom.2024.128998. URL `http://arxiv.org/abs/2407.19394`. arXiv:2407.19394 [cs].

[69] Jianwei Zheng, Wei Li, Ni Xu, Junwei Zhu, Xiaoxu Lin, and Xiaoqin Zhang. Alias-Free Mamba Neural Operator. Technical report. URL `https://github.com/ZhengJianwei2/Mamba-Neural-Operator`.

[70] Yifan Zhou, Zeqi Xiao, Shuai Yang, and Xingang Pan. Alias-Free Latent Diffusion Models: Improving Fractional Shift Equivariance of Diffusion Latent Space. *CVPR*, 2025.

[71] Xueyan Zou, Fanyi Xiao, Zhiding Yu, and Yong Jae Lee. Delving Deeper into Anti-aliasing in ConvNets. *International Journal of Computer Vision 2022*, pages 1–15, August 2020. ISSN 15731405. doi: 10.1007/S11263-022-01672-Y/FIGURES/11. URL `http://arxiv.org/abs/2008.09604`. arXiv: 2008.09604 Publisher: Springer.

A Full proofs

Proposition 2. $f\left(K^\top V\right)$ is shift-invariant.

Proof. The problem can be simplified by considering an arbitrary entry in $K^\top V$, since

$$\left(K^\top V\right)_{i,j} = K_i^\top V_j \,,$$

where $K_i,\, V_j$ are the i-th and j-th columns of K and V, representing 1-dimensional signals.

By Parseval theorem, this inner product between signals equals to their inner product in Fourier domain:

$$K_i^\top V_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{K}_i\left(\omega\right) \hat{V_j}^*\left(\omega\right) d\omega \,. \tag{10}$$

Denote K^τ and V^τ the queries and values of a τ-shifted input signal w.r.t. the input signal of K and V. Following Proposition 1, a translation by τ of the input x yields a translation by τ of both K and V. The Fourier transform of the translated signals differs by a phase which cancels out in the inner product, thus we get the same product:

$$\begin{aligned}
K_i^{\tau\top} V_j^\tau &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{K_i^\tau}\left(\omega\right) \hat{V_j^\tau}^*\left(\omega\right) d\omega && (11)\\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{K}_i\left(\omega\right) e^{j\omega\tau} \left(\hat{V}_j\left(\omega\right) e^{j\omega\tau}\right)^* d\omega && (12)\\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{K}_i\left(\omega\right) \hat{V_j}^*\left(\omega\right) d\omega && (13)\\
&= K_i^\top V_j \,. && (14)
\end{aligned}$$

□

Class Attention. Below we formalize the statement regarding shift-invariance of the Class Attention layer in Section 2.3.

Proposition 4. *Let $Q = XW_q$, $K = XW_k$ and $V = XW_v$ be the query, key and value matrices of a patch sequence $X \in \mathbb{R}^{N\times D}$. Append the sequence with a class token,*

$$\tilde{Q} \;=\; \left[Q^\top,\, q_{\text{cls}}\right]^\top, \qquad \tilde{K} \;=\; \left[K^\top,\, k_{\text{cls}}\right]^\top, \qquad \tilde{V} \;=\; \left[V^\top,\, v_{\text{cls}}\right]^\top, \tag{15}$$

with learnable vectors $q_{\text{cls}}, k_{\text{cls}}, v_{\text{cls}} \in \mathbb{R}^D$. Our CA layer applies the SEA update

$$\tilde{V}' \;=\; \mathrm{SEA}\Big(\tilde{Q},\, \tilde{K},\, \tilde{V}\Big) \;=\; \tilde{Q}\, f\Big(\tilde{K}^\top \tilde{V}\Big) \;\in\; \mathbb{R}^{(N+1)\times D}, \tag{16}$$

where $f\colon \mathbb{R}^{(D\times D)} \to \mathbb{R}^{D\times D}$ is any matrix function. Denote the output by $\tilde{V}' = \left[V'^\top,\, v'_{\text{cls}}\right]^\top$ with $V' \in \mathbb{R}^{N\times D}$. Then

1. ***Patch equivariance:*** *V' is shift-equivariant.*
2. ***Class invariance:*** *v'_{cls} is shift-invariant.*

Proof. Let Q^τ K^τ and V^τ be the keys and values obtained from the τ-translated input X, and define

$$\tilde{Q}^\tau \;=\; \Big[\left(Q^\tau\right)^\top,\, q_{\text{cls}}\Big]^\top, \quad \tilde{K}^\tau \;=\; \Big[\left(K^\tau\right)^\top,\, k_{\text{cls}}\Big]^\top, \quad \tilde{V}^\tau \;=\; \Big[\left(V^\tau\right)^\top,\, v_{\text{cls}}\Big]^\top. \tag{17}$$

Shift-invariance of the attention weights. Since the translation τ acts only on patch tokens, we get

$$\tilde{K}^{\tau\,\top} \tilde{V}^\tau \;=\; \left(K^\tau\right)^\top V^\tau + k_{\text{cls}} v_{\text{cls}}^\top \;=\; K^\top V + k_{\text{cls}} v_{\text{cls}}^\top \;=\; \tilde{K}^\top \tilde{V}, \tag{18}$$

where the middle equality uses Proposition 2. The rank-one class term $k_{\text{cls}} v_{\text{cls}}^\top$ is constant (independent of the input translation), hence $f(\tilde{K}^\top \tilde{V})$ is shift-invariant.

It holds that

$$\tilde{V}' = \tilde{Q} f\left(\tilde{K}^\top \tilde{V}\right) = \left[Q^\top f\left(\tilde{K}^\top \tilde{V}\right), q_{\text{cls}}^\top f\left(\tilde{K}^\top \tilde{V}\right)\right]^\top \tag{19}$$

Patch equivariance. From Equation (19), the patch tokens post CA are $V' = Q^\top f\left(\tilde{K}^\top \tilde{V}\right)$, where $f\left(\tilde{K}^\top \tilde{V}\right)$ is shift-invariant, therefore V' is shift-equivariant similar to Proposition 3.

Class invariance. From Equation (19), the class token post CA is $q_{\text{cls}}^\top f\left(\tilde{K}^\top \tilde{V}\right)$, which is shift-invariant. □

B Additional results

B.1 Additional datasets

We evaluate all models from section 4.1 on three additional classification benchmarks — CIFAR-10, CIFAR-100 [37], and Stanford Cars [36] — under two protocols: (i) fine-tuning ImageNet-pretrained checkpoints and (ii) training from scratch. In both protocols, we train each model using the ImageNet recipe of Table 8 with 1,000 epochs, where in the fine-tuning protocol, we initialize the model weights using the checkpoints from section 4.1. We report top-1 accuracy together with cyclic shift consistency for integer and half-pixel translations, defined exactly as in Section 4.1.

Across all datasets, AFT maintains near-perfect shift-equivariance (above 99% consistency in integer and half-pixel shifts). Additionally, when fine-tuned, the XCiT-Small-AF model is slightly but consistently more accurate than the baseline on all three datasets, suggesting that the shift-invariance prior can benefit larger transformers when adapting to small datasets.

Table 4: **CIFAR and Stanford Cars (fine-tuning): accuracy and cyclic shift consistency.** We fine-tune ImageNet-pretrained checkpoints on CIFAR-10/100 and Stanford Cars. Metrics are top-1 test accuracy and consistency to integer and half-pixel cyclic translations (as in Section 4.1).

	CIFAR-10			**CIFAR-100**			**Stanford Cars**		
Model	Test accuracy	Integer shift consist.	Half-pixel shift consist.	Test accuracy	Integer shift consist.	Half-pixel shift consist.	Test accuracy	Integer shift consist.	Half-pixel shift consist.
XCiT-Nano (Baseline)	**98.0**	98.0	98.1	**84.3**	89.4	89.9	**92.3**	95.2	95.1
XCiT-Nano-APS	97.7	**100.0**	99.4	84.1	**100.0**	96.9	92.2	**100.0**	97.0
XCiT-Nano-AF (ours)	97.7	99.9	**99.9**	**84.3**	99.6	**99.4**	92.1	99.9	**99.8**
XCiT-Small (Baseline)	98.2	98.4	98.5	85.6	87.4	89.6	92.6	95.4	96.3
XCiT-Small-APS	98.3	**100.0**	99.5	85.4	**100.0**	96.2	92.2	**100.0**	97.2
XCiT-Small-AF (ours)	**98.4**	99.9	**99.9**	**85.8**	99.6	**99.5**	**93.0**	99.9	**99.9**

Table 5: **CIFAR and Stanford-Cars (from scratch): accuracy and cyclic shift consistency.** Models are trained from scratch on each dataset using the ImageNet training setup with 1,000 epochs; metrics as in Section 4.1.

	CIFAR-10			**CIFAR-100**			**Stanford Cars**		
Model	Test accuracy	Integer shift consist.	Half-pixel shift consist.	Test accuracy	Integer shift consist.	Half-pixel shift consist.	Test accuracy	Integer shift consist.	Half-pixel shift consist.
XCiT-Nano (Baseline)	**97.2**	98.0	98.0	**82.4**	90.3	90.2	**86.5**	91.6	91.7
XCiT-Nano-APS	97.2	**100.0**	99.2	82.3	**100.0**	97.1	84.1	**100.0**	93.3
XCiT-Nano-AF (ours)	96.5	99.9	**99.9**	81.3	99.5	**99.4**	85.3	99.7	**99.6**
XCiT-Small (Baseline)	**98.3**	98.7	98.7	**85.3**	91.3	91.3	89.6	95.1	94.5
XCiT-Small-APS	98.0	**100.0**	99.4	85.1	**100.0**	95.7	**90.5**	**100.0**	96.5
XCiT-Small-AF (ours)	97.6	99.9	**99.9**	83.4	99.6	**99.6**	88.5	99.8	**99.8**

B.2 Global average pooling vs AF Class Attention

In section 2 we propose two mechanisms to get a shift-invariant global representation out of the AFT — a global average pooling over the embedding dimension and an alias-free class attention that leverages SEA to maintain a shift-invariant class token. We compare the final *AF class attention* (AFCA) head with a *global average pooling* (AvgPool) head within the AFT architecture. As shown in Table 6, both AFCA and AvgPool maintain near-perfect shift consistency. AFCA demonstrates consistent improvement over AvgPool in top-1 accuracy, most visible in the Small variant.

Table 6: **AF class attention vs. global average pooling (AFT).** Top-1 accuracy and cyclic shift consistency (integer and half-pixel) on ImageNet. AvgPool and AFCA retain near-perfect equivariance, while AFCA provides a consistent accuracy gain, most notably for the Small variant.

Model	Test accuracy	Integer shift consist.	Half-pixel shift consist.
XCiT-Nano-AF (AvgPool)	70.35	99.0	98.6
XCiT-Nano-AF (AF-CA)	70.48	99.2	99.4
XCiT-Small-AF (AvgPool)	80.70	99.5	99.4
XCiT-Small-AF (AF-CA)	81.81	99.5	99.4

B.3 Polynomial vs GELU comparison

Similar to the Alias-Free ConvNet (AFC) of Michaeli et al. [40], certified shift-invariance in the Alias-Free Transformer (AFT) can be achieved by replacing the filtered GELU activations with polynomial approximations. We therefore train an alias-free XCiT-Nano variant whose activations are polynomials with learnable coefficients per embedding channel, following Michaeli et al. [40]. We use polynomials of degree 2 in the AFT blocks and degree 3 in the patch-embedding stage (PE), which remains alias-free thanks to the downsampling layers following the activations in the PE. The results in Table 7 show that the full polynomial model ("Poly") has near-100% shift consistency, with a small gap that can be attributed to numerical errors, similar to the case in the APS model (see Table 1). However, we observe that unlike in the AFC, polynomial activations lead to a significant reduction in accuracy.

Interestingly, when the four GELU activations in the PE are retained and only the block activations are replaced ("GELU (PE), Poly (Blocks)"), most of the lost accuracy is recovered. This may indicate that polynomial activations limit the representational capacity of the convolutional PE which is much shallower than the convnet tested in AFC.

Table 7: **Effect of polynomial activations on ImageNet performance and shift consistency.** Top-1 accuracy and consistency (%) of XCiT-Nano with the standard filtered GELU, full polynomial replacement (Poly), and a hybrid that keeps GELU in the patch-embedding (PE)

Model	Test accuracy	Integer shift consist.	Half-pixel shift consist.
GELU	70.4	98.8	98.4
Poly	65.8	99.7	99.6
GELU (PE), Poly (Blocks)	68.5	99.4	98.7

C Translation visualization

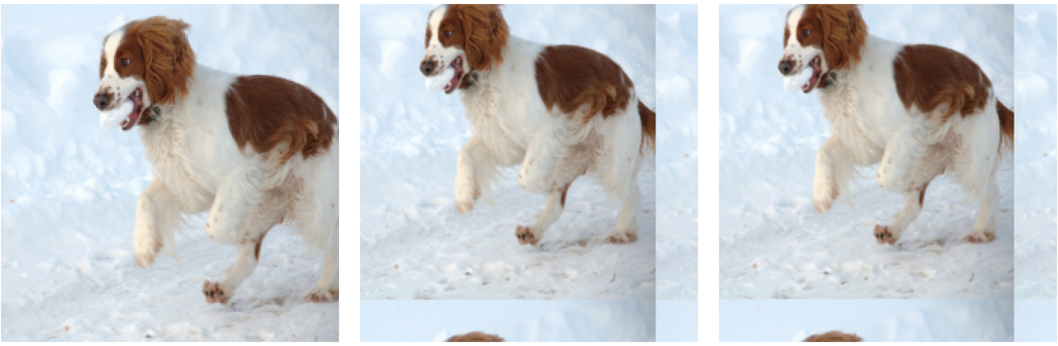
We provide visual examples of the translations described in the paper in Figures 3 and 4.

D Experiments details

As mentioned in Section 4, we used the same training settings as in XCiT [16], except for lowering the batch size for the AF model. We report the used training hyperparameters in Table 8.

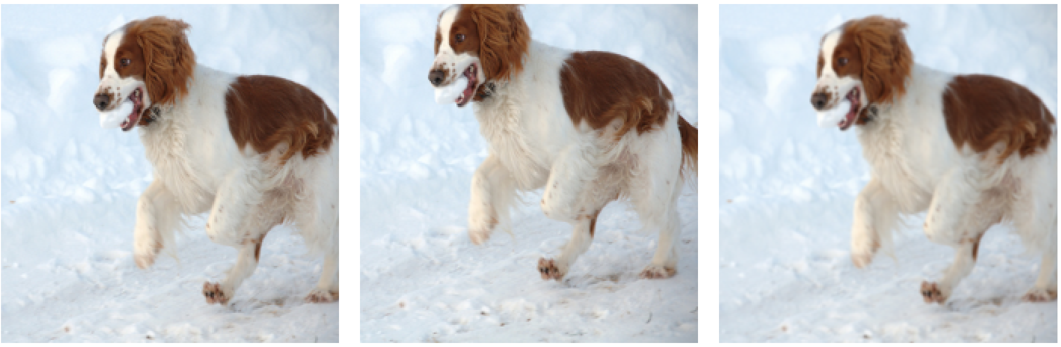
D.1 Batch size choice

Aiming to avoid expensive hyperparameter tuning, we used the XCiT original recipe [16]. However, in our early experimentations, we observed fluctuations in the training curves of the AF and APS models, which were alleviated by reducing the batch size. To decide fair batch sizes, we trained the *Nano* variants of Baseline, APS, and AF with batch sizes 512 and 1024 and picked, for each method, the configuration that performed best. The Baseline favored 1024 (top-1 70.4 vs. 70.1 at 512), while



(a) Original image (b) Circular integer shift (c) Circular half-pixel shift

Figure 3: **Visualization of cyclic shifts.** (a) Original ImageNet [12] validation-set image. (b) Circular shift of 16 pixels in horizontal and vertical axes. (c) Circular shift of 16.5 pixels in horizontal and vertical axes. The original image is upsampled by a factor 2, circularly shifted by 33 pixels, and downsampled by factor 2.



(a) Original image (b) Crop-shift (c) Bilinear fractional shift

Figure 4: **Visualization of realistic shifts.** (a) Original ImageNet [12] validation-set image — 224×224 center crop of the original 256×256 image. (b) Crop-shift of the original image of 16 pixels in the horizontal and vertical axes. The cropped area is shifted by 16 pixels with respect to the cropped area in the original image. (c) Bilinear fractional shift of 0.5 pixels in horizontal and vertical axes. We use a 226×226 center crop of the original 256×256 image and simulate a fractional-pixel shift using a grid-sample with a fractional offset.

Table 8: **Hyperparameters.** Unless stated otherwise, the same settings apply to all models.

Category	**Parameter**	**Value**
Optimizer	Optimizer	AdamW
	(β_1, β_2)	(0.9, 0.999)
	Weight decay	0.05
Learning rate scheduling	Base LR	1×10^{-3} (Baseline), 5×10^{-4} (AF, APS)
	Warm-up epochs	5
	LR decay	Cosine
	Min LR	1×10^{-5}
Data	Resolution	224×224
	Batch size	1024 (Baseline), 512 (AF, APS)
Regularization	Layer scale (ϵ init)	1.0
	Stochastic depth	0.0 (Nano), 0.05 (Small)

APS and AF favored 512 (APS: 68.7 vs. 67.2 at 1024; AF: 70.4 vs. 70.1 at 1024). We therefore used batch size 1024 for the Baseline and 512 for APS and AF throughout the paper. Note that the figures above for the AF and APS models are with an AvgPool head. We applied the same choice to the Nano and Small variants with CA and AFCA heads without further tuning.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] ,

   Justification: We describe the proposed alias-free model in sections 2 and 3 and evaluate its performance in section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the proposed model compared to other studies in the field in section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes] ,

Justification: The main theoretical results are presented in propositions 1 to 3, each followed by a proof or proof sketch (in this case we provide a full proof in the supplementary material).

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We explain in detail all parts of the proposed model in sections 2 and 3. Some elements (including the baseline model, layers in our proposed model, models we evaluate for comparison, and training recipe) are based on other public manuscripts to which we refer.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We intend to release the code and trained models upon acceptance of the paper. The model implementation and training details are explained in the paper, and the only dataset we used (ImageNet) is publicly available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training regime is explained in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We train all models once due to high computation costs. The rest of the results are deterministic.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the type of hardware we used in Section 4 and report training runtime in Section 6.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The proposed method is an image classification model trained on standard datasets (ImageNet) consisting mostly of animals and objects, and has no higher risk of negative societal impact than other studies in the field.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release data. The proposed method is an image-classification model trained on standard datasets (ImageNet) and has no high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All relevant assets, including models, datasets, and the implementation framework, are mentioned with a proper citation and have standard open-source licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The model and training details are fully explained in the paper. We intend to release the code and trained models upon acceptance of the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The paper does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.