TECHNION
Israel Institute
of Technology

Code is available

NEURAL INFORMATION
PROCESSING SYSTEMS

Alias-Free ViT:
Fractional Shift Invariance via Linear Attention

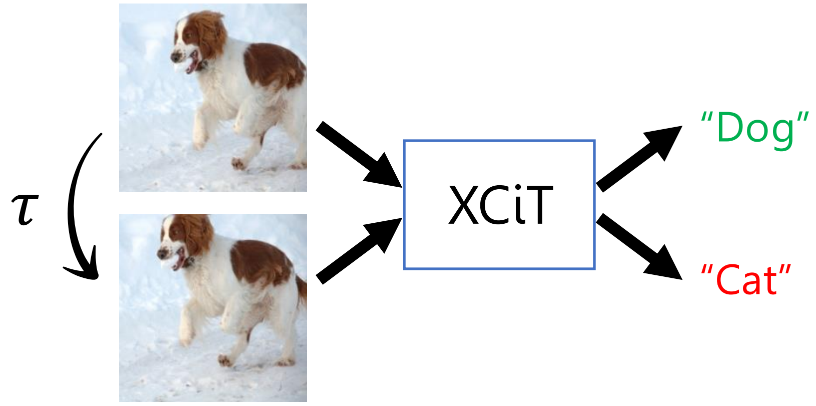

Hagay Michaeli


Daniel Soudry

The research of DS was funded by the European Union
(ERC, A-B-C-Deep, 101039436)

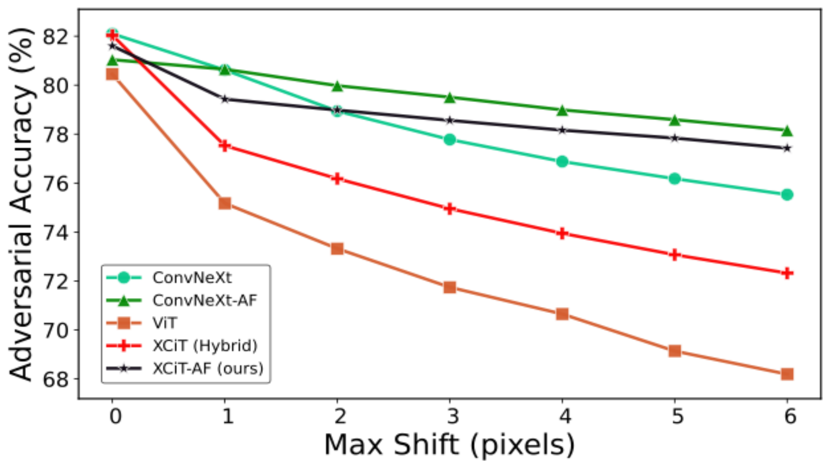
Wed 3 Dec 11 am

Motivation: built-in prior



- Shift-invariance
  - A key prior assumption for image classifiers
  - Motivating the first convolutional models

- However, modern Vision Transformers (ViTs) are very sensitive to small image translations



- Convnets are less sensitive to image translations, but still not perfectly robust
  - Aliasing in pooling and nonlinearities break shift-equivariance[1]
  - Alias-Free convnets[2] have provable shift-invariance

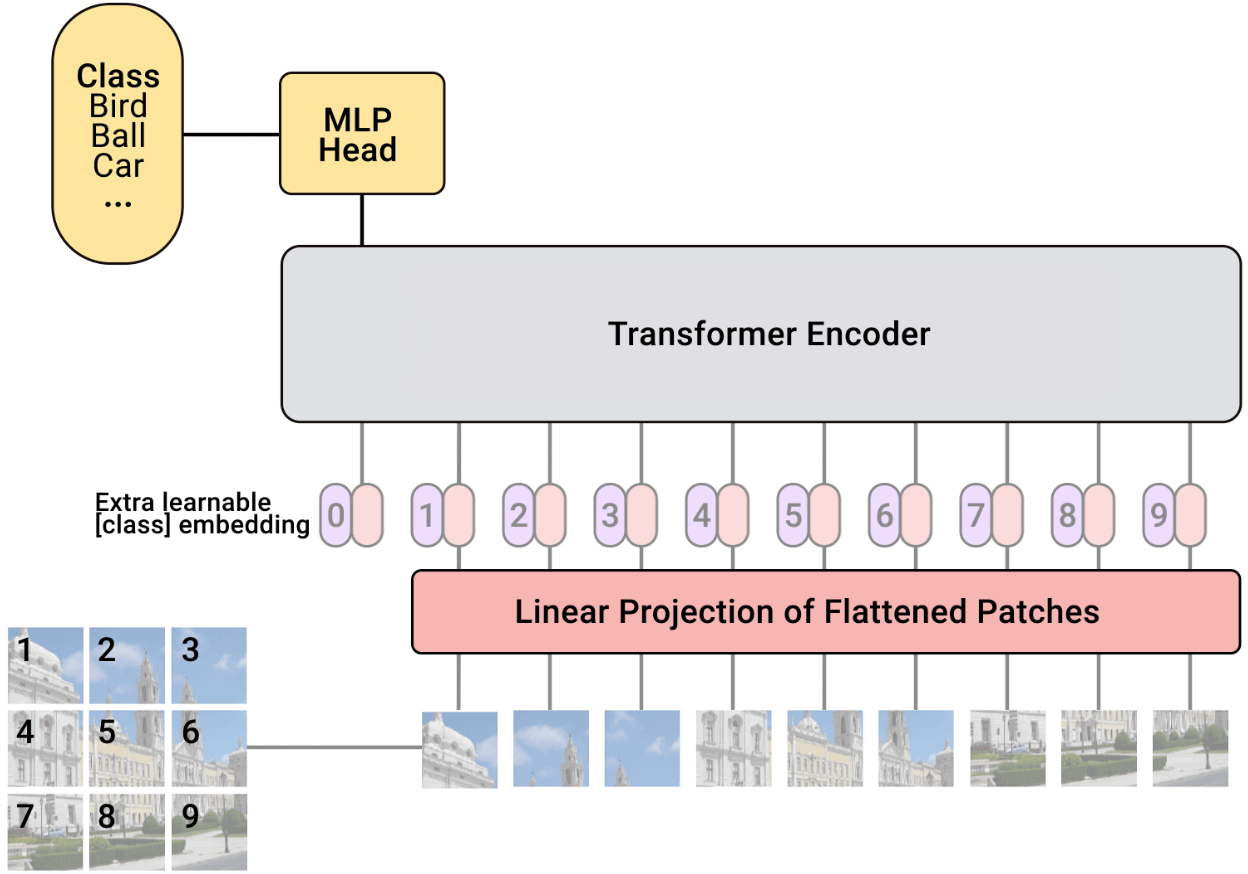
Goal: Use Alias-Free framework to construct shift-invariant ViT

[1] Making convolutional networks shift-invariant again. Zhang, ICML 2019
[2] Alias-Free Convnets: Fractional Shift Invariance via Polynomial Activations. Michaeli et al, CVPR 203

Contribution

- We present a class of Shift-equivariant attention layers (SEA)
  - Includes linear attention and cross-covariance attention
- We design an Alias-Free Vision Transformer (AFT)
  - Using cross-covariance attention and Alias-free nonlinearities
  - Competitive in image classification
- We improve translation robustness
  - ~99% consistency for fractional-cyclic shifts
  - Significant improvement in adversarial robustness to practical translation-attacks

Background: Vision Transformers[1]



[1] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Dosovitskiy et al, ICLR 2021
Image by Phil Wang - MIT License.

Shift-Invariant ViT

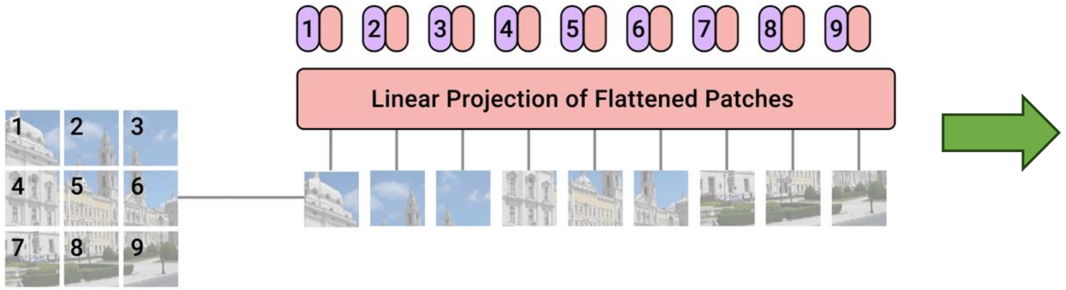
Shift-Invariant ViT – Patch Embedding

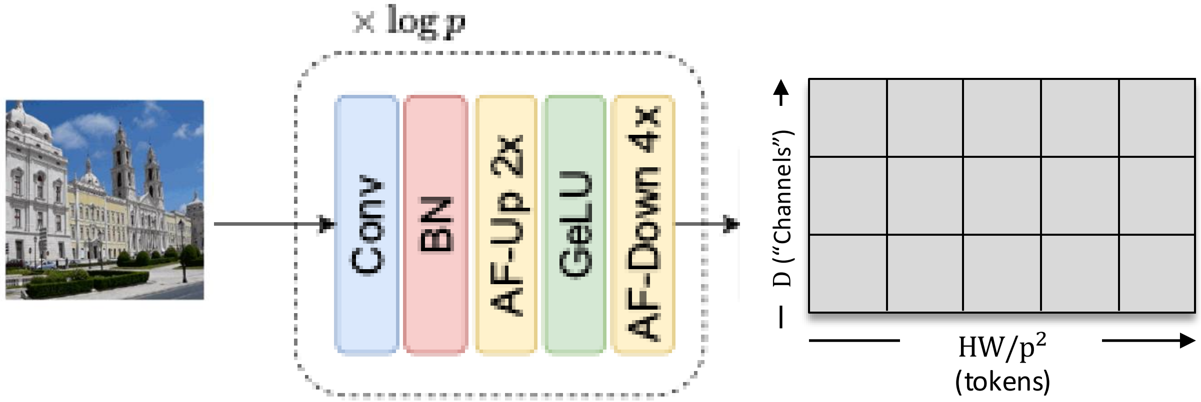
Observation:
Patch-embedding ↔ Strided convolution

Aliasing in PE and POE break shift-equivariance

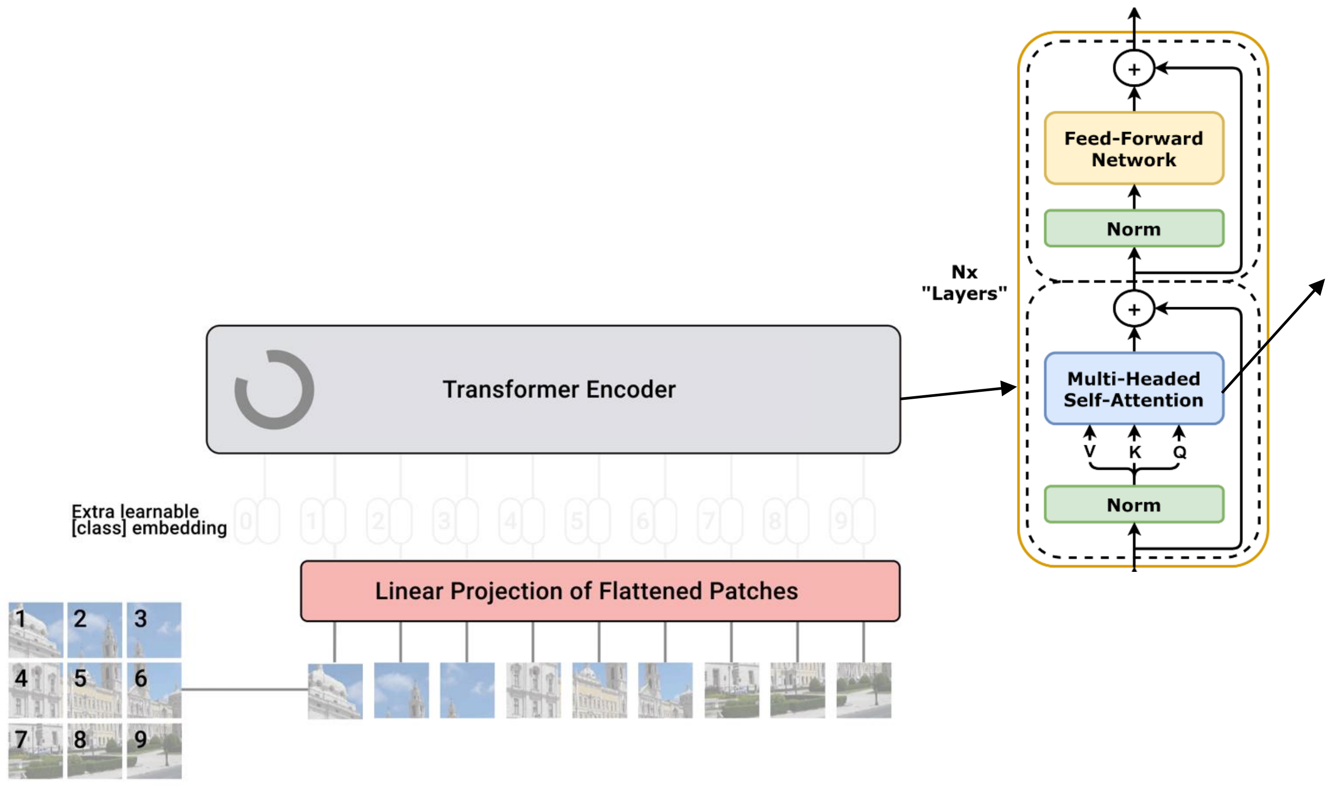
Solution:
Alias-Free Convolutional Patch Embedding

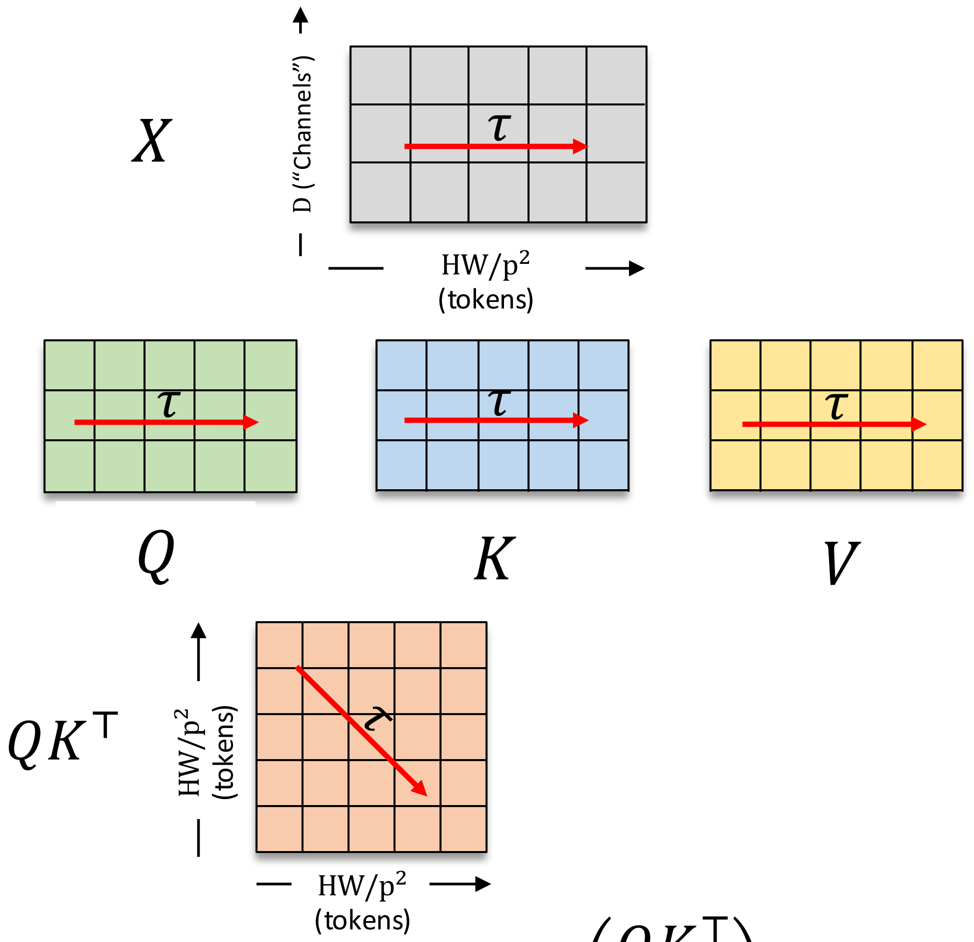
- Gradual, alias-free downsampling
- Alias-free nonlinearities
- No positional encoding





Shift-Invariant ViT – Attention





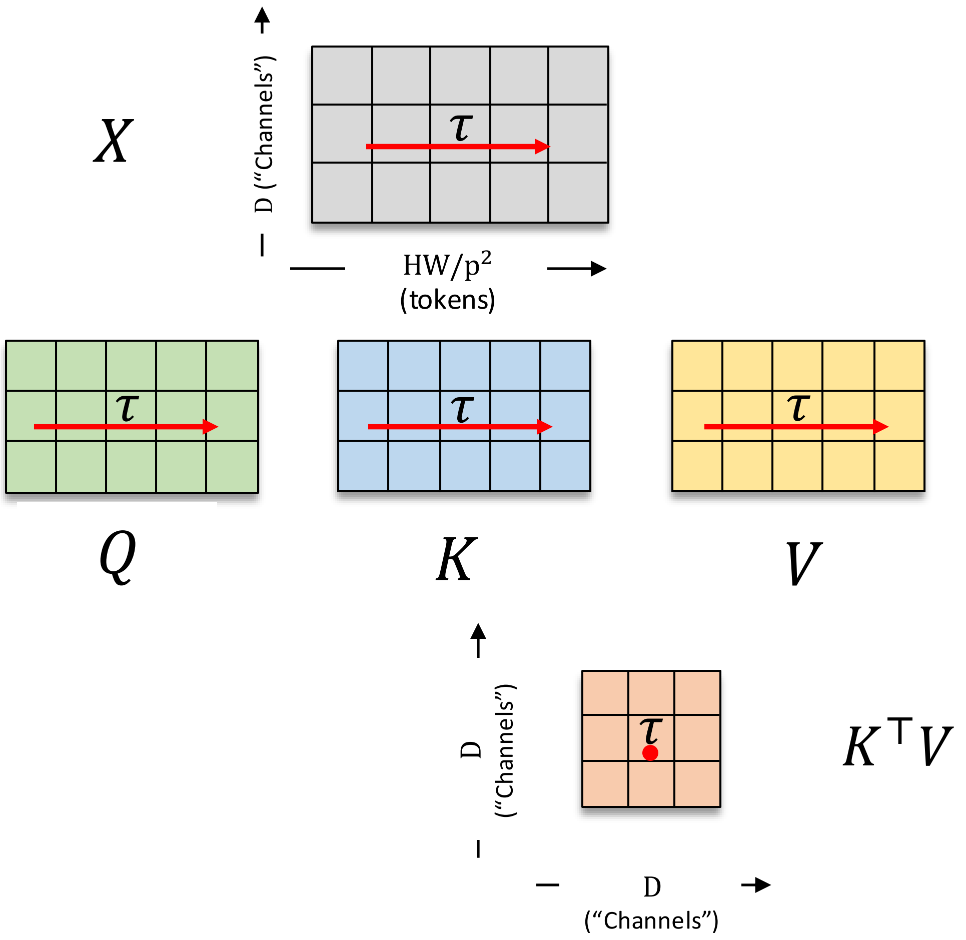
$$SA(X) = \text{softmax}\left(\frac{QK^\top}{\sqrt{D}}\right) V$$

Shift-Invariant ViT – Attention

Proposition 1:
Q, K, V are shift-equivariant

Proposition 2:
$K^\top V$ is shift-invariant

Proposition 3:
$SEA(X) = Q\, f(K^\top V)$ is shift-equivariant



$$SEA(X) = Qf(K^\top V)$$

Alias-Free XCiT[1]

Convolutional patch embedding → AF Convolutional patch embedding

XCA blocks
- Cross-covariance attention (SEA)
- Layer Norm
- GeLU

→

AF-XCA blocks
- Cross-covariance attention (SEA)
- Alias-Free Normalization[2]
- Reduced-aliasing GeLU[3]

Class attention → Shift-invariant global representation
- Global average pooling
- Shift-invariant class attention

[1] XCiT: Cross-covariance image transformers. Ali et al, NeurIPS 2021
[2] Alias-Free Convnets: Fractional Shift Invariance via Polynomial Activations. Michaeli et al, CVPR 203
[3] Alias-Free Latent Diffusion Models: Improving Fractional Shift Equivariance of Diffusion Latent Space. Zhou et al, CVPR 2025

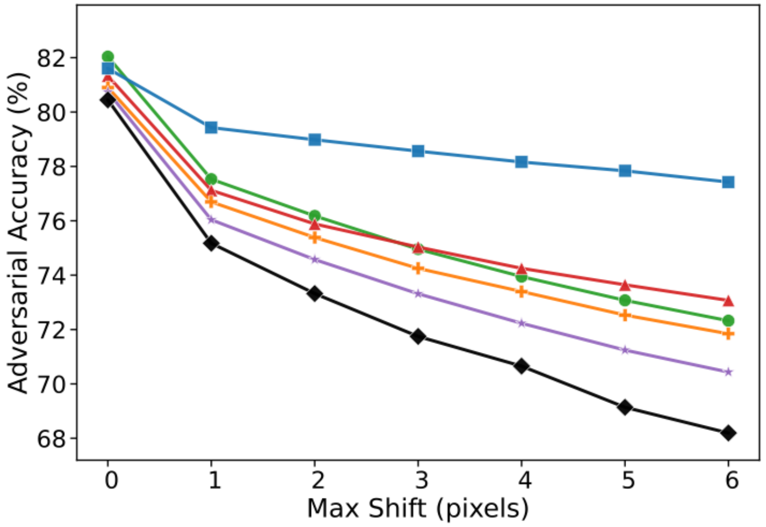
Results: ImageNet accuracy and consistency

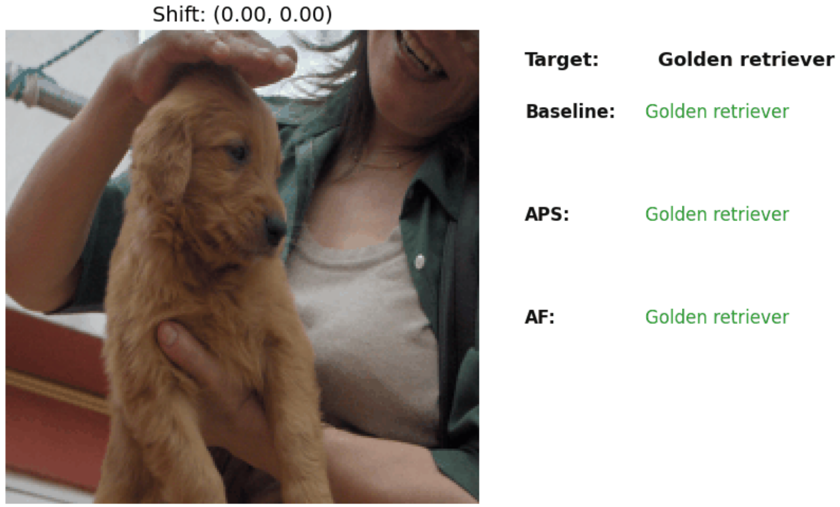
- Accuracy on par with baseline
- ~ 99% consistency to cyclic integer and fractional shifts

	Model	Test Accuracy	Integer shift consistency	Half-pixel shift consistency
XCiT-Nano	Baseline	70.4	83.7	82.0
	APS	68.4	**100.0**	87.5
	AF (ours)	**70.5**	99.2	**98.7**
XCiT-Small	Baseline	**82.0**	91.4	89.8
	APS	81.3	**100.0**	94.0
	AF (ours)	81.8	99.5	**99.4**

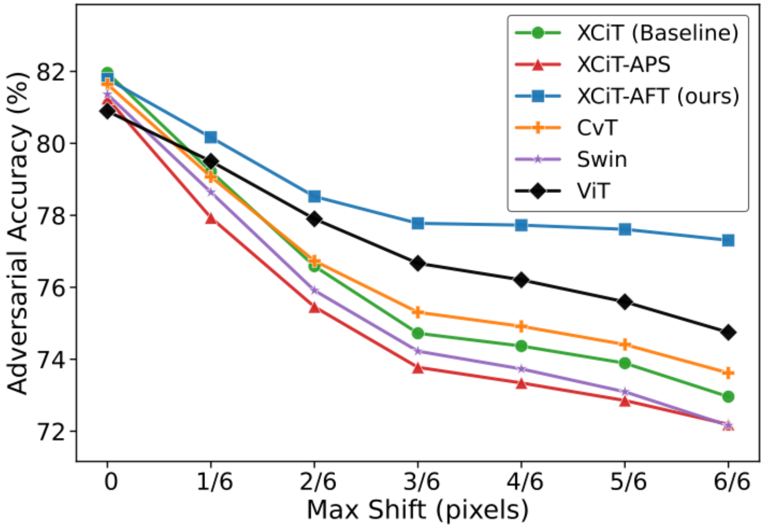
Results: Practical translations

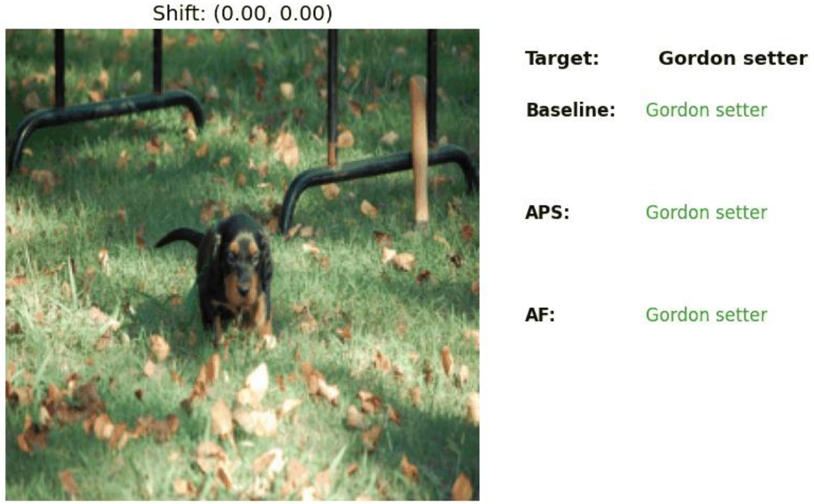
- Crop-shifts:
  - Imitating camera translations





- Bilinear-interpolation fractional shifts:
  - Imitating small, sub-pixel translations





Summary

- Problem & goal
  - ViTs are very sensitive to image translation comparing to Convnets
  - Build alias-free shift-equivariant transformer encoder
- Approach
  - Shift-Equivariant Attention (SEA): $SEA(X) = Q \cdot f(K^{\top} V)$
    - Includes linear and cross-covariance attention

  - Alias-Free ViT (AFT):
    - Shift-Equivariant Attention
    - Alias-free patch embedding, activations, and normalization
- Results
  - Competitive ImageNet accuracy
  - ~99% consistency under fractional cyclic shifts
  - Stronger robustness to realistic translations (crop / sub-pixel)