EE604 2025-26 Even
Course Project

# StanNet: Fully Complex-valued CNNs for Generic Image Classification

Nishant Pandey
220724, IIT Kanpur
nishantp22@iitk.ac.in

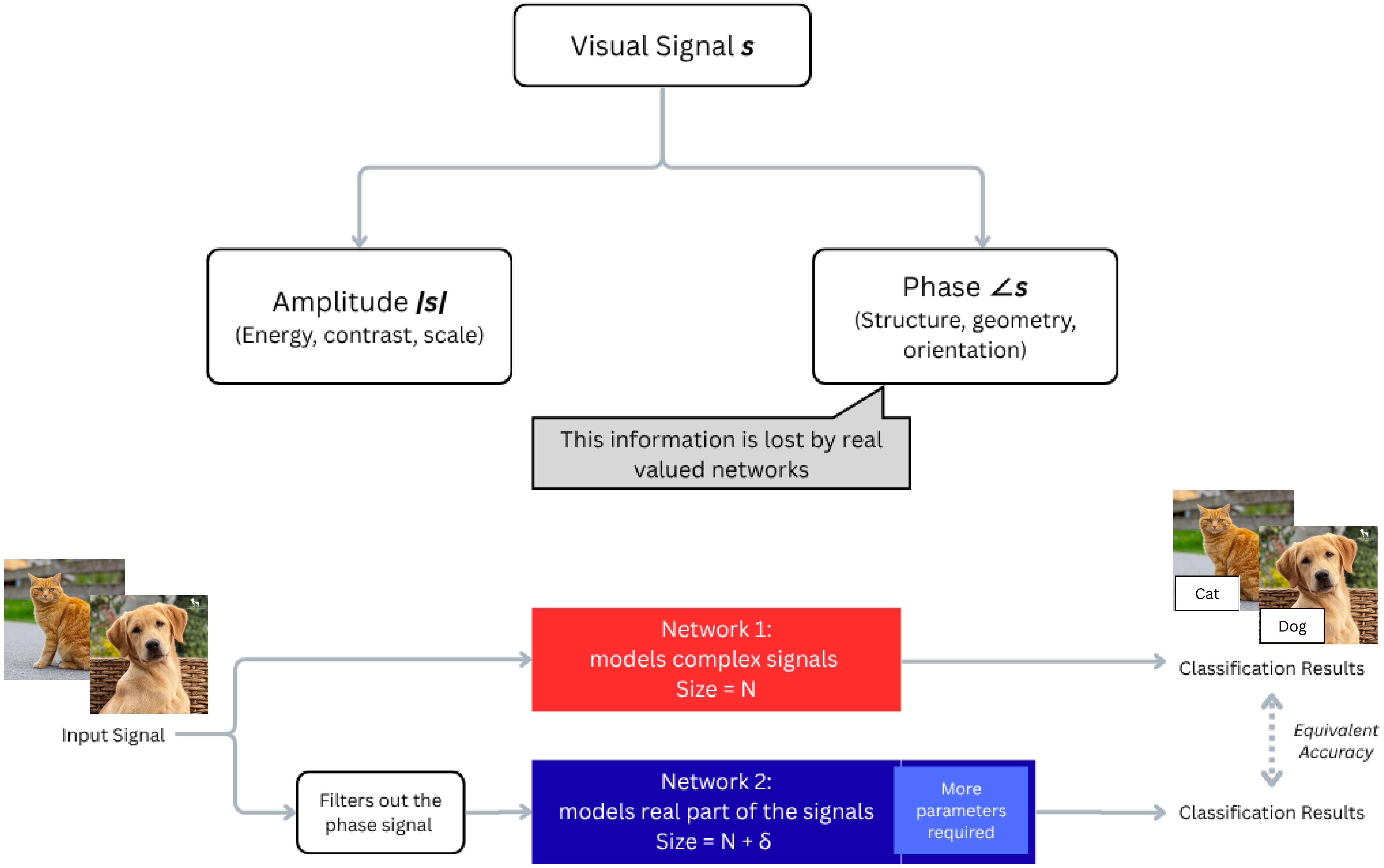
Mohd Mufeed Amir
220660, IIT Kanpur
mmamir22@iitk.ac.in

Tanmay Siddharth
221129, IIT Kanpur
tanmays22@iitk.ac.in

Sumit Rojaria
221104, IIT Kanpur
sumitr22@iitk.ac.in

Aditya Raj Mishra
220078, IIT Kanpur
adityarm22@iitk.ac.in

# The loss of information that comes with discarding phase information in visual signals




Real valued network need **larger models** than complex counterparts to **emulate the patterns** in phase information

**Objective:**
Build classifiers that **operate on complex inputs** and keep phase information **throughout the pipeline**

**Input:**
A colored image processed into a **complex color representation**

**Output:**
Classification of the image into one of the target classes while **preserving magnitude and phase contributions** when producing logits

**Constraints:**
Achieve at par results with **lesser model parameters** than SOTA real valued models

## CNN Foundation

Inaugral works that leverage *stacked convolutions, non-linearities*, and *GPU training*
Further improved by *Residual Learning*

Key papers surveyed by us propose:

- *Deep ReLU CNN with local response normalization, dropout, and multi-GPU training*
- ***Residual shortcuts** which let very deep networks learn identity mappings plus residuals*
- *Training a ViT by distilling knowledge with a **learned distillation token** and **strong augmentations/ regularization***

## Vision Transformers

Treat images as *sequences of patch tokens* and learn global context with *self-attention*

Key papers surveyed by us propose:

- *Splitting an image into fixed-size patches, adding **positional embeddings** and a **class token**, and training a pure transformer encoder that fine-tunes effectively on standard classification benchmarks*
- *Swin Transformer, which applies **windowed self-attention with shifted windows***

## Convolutional modernizations

*Architectural refinements* plus *contemporary training pipelines* sustain competitive classification accuracy in comparison to Transformers

Key paper surveyed by us propose:
*ConvNeXt, which modernize ResNetstyle backbones with design choices inspired by ViT-era practices*
*The model attributes gains to architectural tweaks and training recipes, underscoring the importance of **unified design–optimization co-evolution** for classification*

## Spectral Gating

*Frequency-selective operations* within neural network architectures, allowing models to emphasize global low frequency structures and suppress high frequency noise directly in the Fourier domain

Key paper surveyed by us propose:

- *Learnable frequency-domain masks for instance segmentation*
- *Training neural architectures in the **spectral (reciprocal) domain** to leverage frequency information for enhanced generalization and robustness*

## Complex-valued Networks

*Complex convolutions, normalizations, and activations* to capture magnitude–phase structure. Surpass real-valued baselines where phase and amplitude interactions are informative for classification

Key paper surveyed by us propose:

- Complex convolution, complex batch normalization, and complex activations with Wirtinger calculus foundations
- *Fully complex-valued CNNs with a **complex color model** and a **complex-valued loss** impose complex information flow end-to-end*

## Key Takeaway toward a novel method:

Spectral Gating

+

Complex Valued Networks

↓

Enhanced frequency information extraction from signal phase

# **StanNet:** a purpose-built, fully complex-valued CNN

**Input:**
Complex color representation of Images

**Network Trunk:**
Feature maps and intermediate layers remain complex

**Final Classifier Head:**
Preserves magnitude and phase contributions when producing logits.

Novelty elements

Rest of the architecture blocks
*complex batch-norm, CPReLU activations, complex dropout, and stochastic depth*

Combines:
- compact stem
- Four stages of complex “ComplexMixerBlocks”
- Progressive channel expansion

Complex output head that performs adaptive pooling and a 1×1 complex projection to classes

StanNet blocks are not just complex analogues of real convolutions,
**They introduce mechanisms that explicitly exploit frequency and phase structure**

# Novel primitives that make StanNet, *StanNet*

**Spectral Gate** *An intra-block module that explicitly selects and reweights frequency content of complex feature*

**Where its used?**
per channel and implemented with differentiable Fourier transforms so gradients flow through the spectral domain back into preceding layers

**Main goals:**
1. To provide a learned bias towards global/low-frequency structure when helpful
2. To suppress high-frequency noise that often harms generalisation.

**Intuition:**

Stronger techniques like **Attention** require a **larger complexity of O(N2).**
SpectralGate is primarily based on **FFT of the entire image**, which is an **O(N logN)** operation

SpectralGate also helps us *manipulate the Phase of the input*, which is crucial for learning spatial differences. This is not possible in the real domain.

In standard CNNs, we need **multiple CNN layers, more parameters, larger kernels, an even attention blocks** to **learn global representations.**
SpectralGate avoids this by **directly dealing with the Fourier spectrum of the entire image in one shot**.

The gate can directly learn to *manipulate the frequency for each channel*, based on the learned parameters

# Novel primitives that make StanNet, *StanNet*

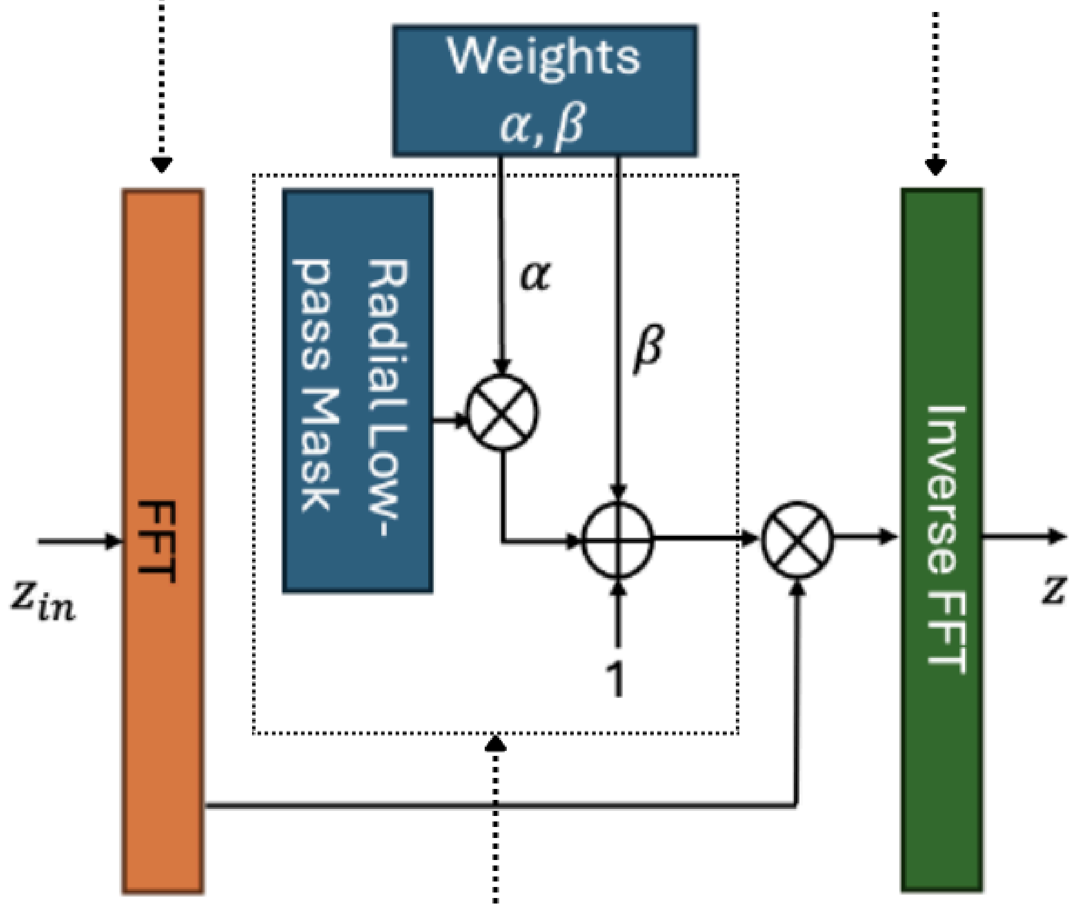
**Spectral Gate** *An intra-block module that explicitly selects and reweights frequency content of complex feature*

For a single complex channel feature map z ∈ CH×W
compute its discrete Fourier transform (DFT)

$$\hat{z}(u,v) \;=\; \mathcal{F}\{z\}(u,v)$$

Return to the spatial domain

$$z_{\mathrm{sp}}(h,w) \;=\; \mathcal{F}^{-1}\big(\widetilde{M}(\cdot)\cdot\hat{z}(\cdot)\big)(h,w).$$




A mask with learnable parameters is applied

$$\widetilde{M} = 1+\beta+\alpha M$$

We use a smooth learnable radial low pass mask, parameterised by cutoff *c* and Temperature *T*>0

$$M(\omega;c,\tau) \;=\; \sigma\Big(-\frac{r-c}{\tau}\Big) \;=\; \frac{1}{1+\exp\big(\frac{r-c}{\tau}\big)}$$

the PDF indicates a radial, smooth low-pass mask with cutoff/sharpness control.

# Novel primitives that make StanNet, *StanNet*

## Magnitude Phase Cross-Gate

*separates amplitude and phase processing and learns distinct gating functions for each*

**Where its used?**
Operates on complex batch normalized output from Spectral gate, and its output is passed into stochastic depth mechanism path

**Main goals:**
1. Allow selective amplification/suppression of magnitudes,
2. Allow controlled phase modulation (offset) conditioned on both magnitude and local phase context.

**Intuition:**
Once we process the output from the SpectralGate module and identify global representations, the **MPCrossGate module helps us learn local representations**.
Specifically, it helps us learn how much each local feature should contribute to the next layer based on:

Magnitude: It *learns which feature channels are meaningful* based on their **relative energy at that spatial location**.
For example: strong activations in certain channels might signal presence of a specific texture or pattern.

Phase: It learns which spatial regions have coherent structure. Nearby pixels with consistent phase means that we may have aligned edges, patterns, boundaries, shapes, and random/ uncoordinated phases implies the presence of noise.

## Novel primitives that make StanNet, *StanNet*

**Magnitude Phase Cross-Gate** *separates amplitude and phase processing and learns distinct gating functions for each*

To avoid learning difficulties due to phase discontinuities at $\pm\pi$ we embed phase using its sine and cosine:

$$\mathbf{e}_\phi = [\cos\phi,\ \sin\phi],$$

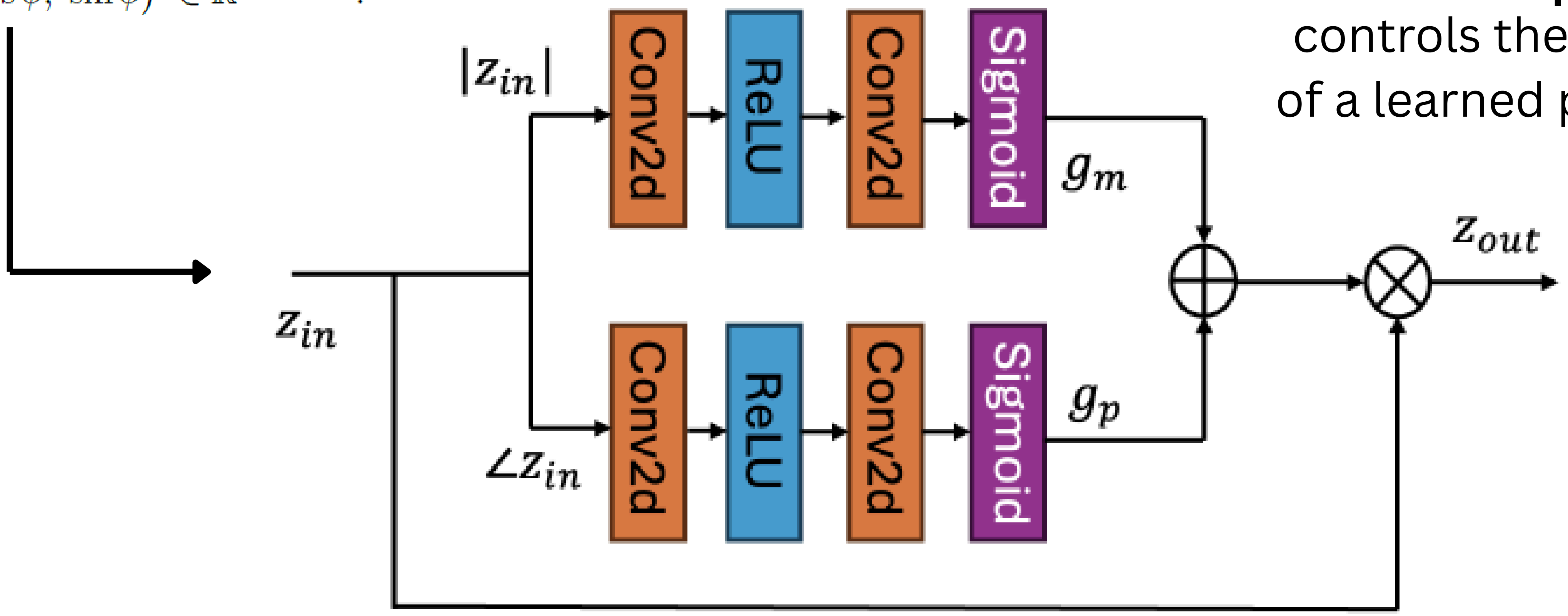
and form a joint real embedding per spatial location:

$$\mathcal{E}(z) = \text{concat}(m,\ \cos\phi,\ \sin\phi) \in \mathbb{R}^{3\times H\times W}.$$

controls amplitude scaling

$$g_m = \sigma(\mathcal{G}_m(\mathcal{E}(z))) \in (0,1)^{H\times W},$$
$$g_\phi = \sigma(\mathcal{G}_\phi(\mathcal{E}(z))) \in (0,1)^{H\times W},$$

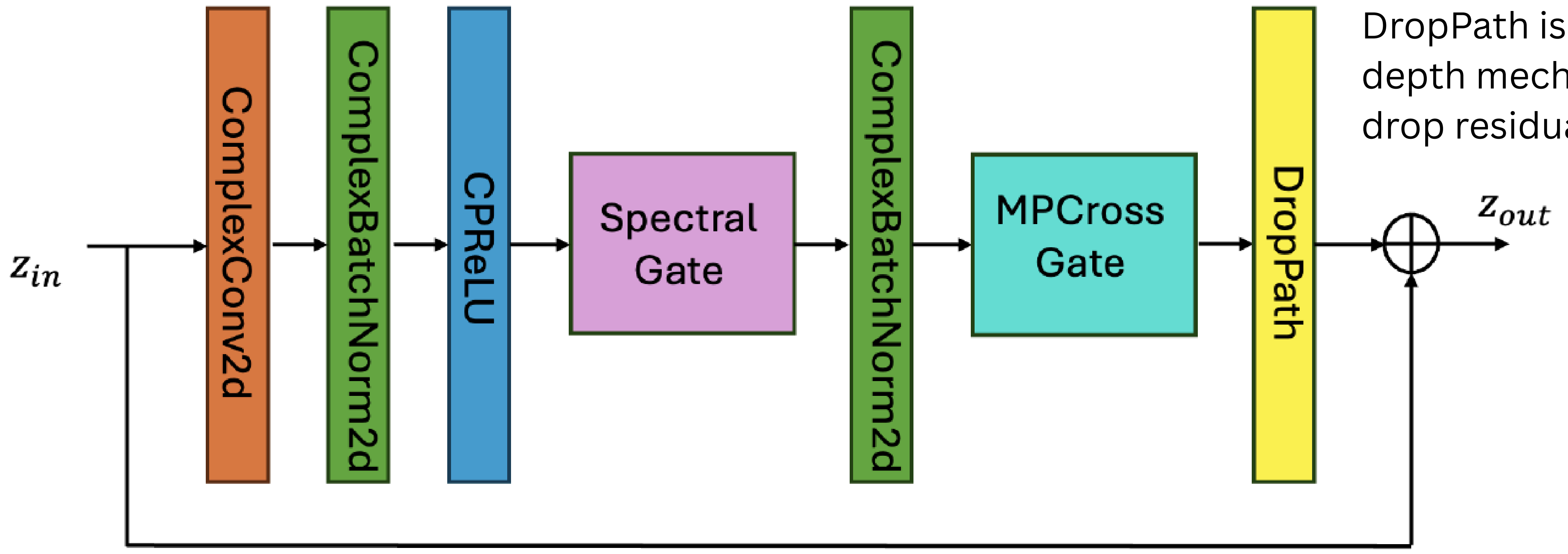
controls the magnitude of a learned phase offset




$$= \frac{(g_m + g_p)}{2} * \mathbf{X}_{in}$$

# Putting the novelties together

## Complex Mixer Module



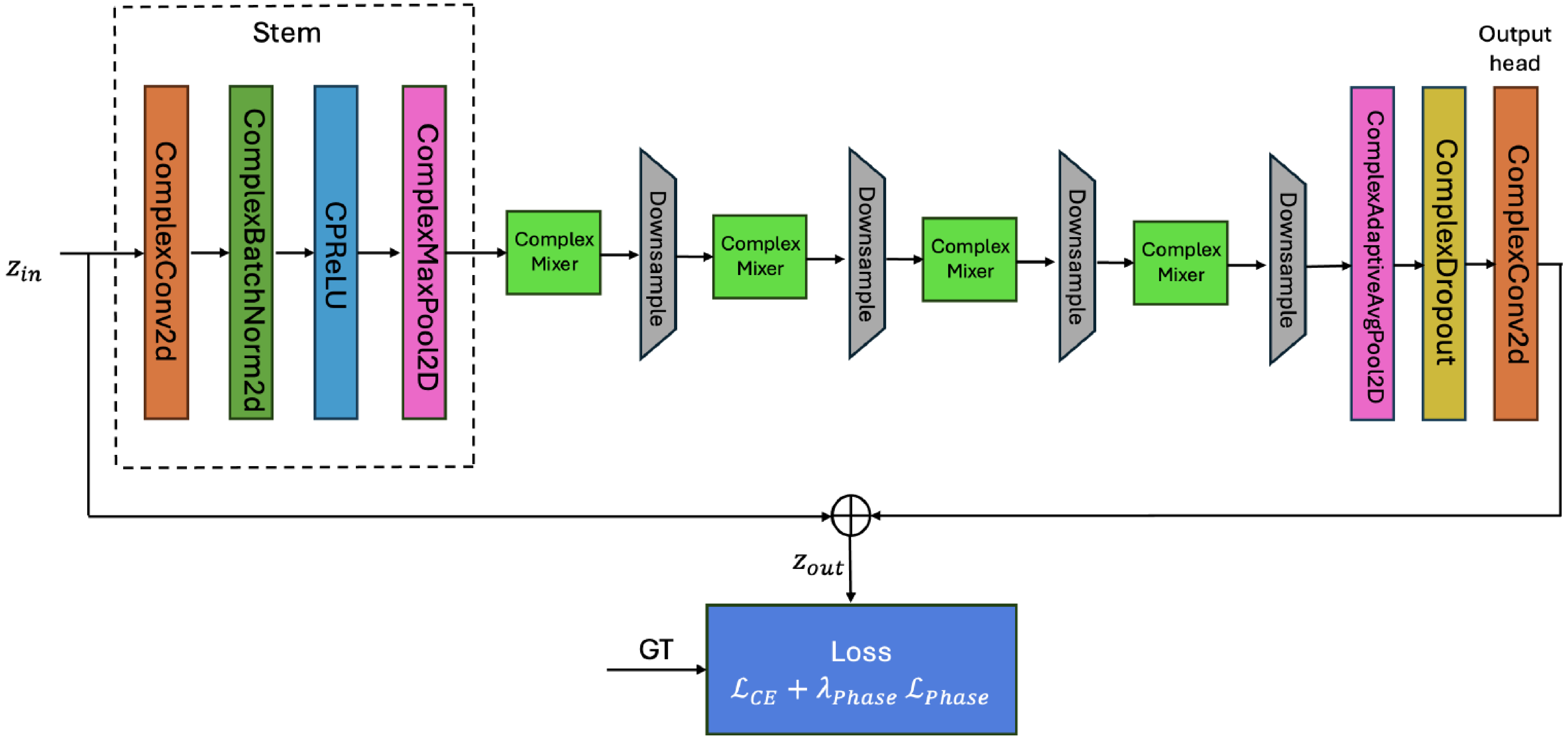

DropPath is the standard stochastic depth mechanism which randomly drop residual connections

This composition keeps the block **compact** while giving it explicit spectral and magnitude/phase **routing capacities**

## Losses, and the complete pipeline

StanNet is trained with a standard cross-entropy classification loss on logits formed from pooled magnitudes.
Optionally a phase-consistency regulariser is added
(the complex head converts pooled complex descriptors to magnitudes before the real linear classifier)

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{phase}} \mathcal{L}_{\text{phase}},$$



# Dataset 1: **CIFAR10**

*The standard dataset is considered for training our proposed architecture*

Implementation details
- Device used: Apple M2 system with 16GB unified memory
- Other architectures considered: ResNet18, ResNet50, AlexNet, VGG19 and VGG16
- epochs: 30
- learning rate = 3e-4
- input image size = 224 x 224
- validation split ratio = 0.2
- For masks in spectral gate: $\tau = 0.8$, $c = 0.25$

*Note: rest of the implementation details are in the code

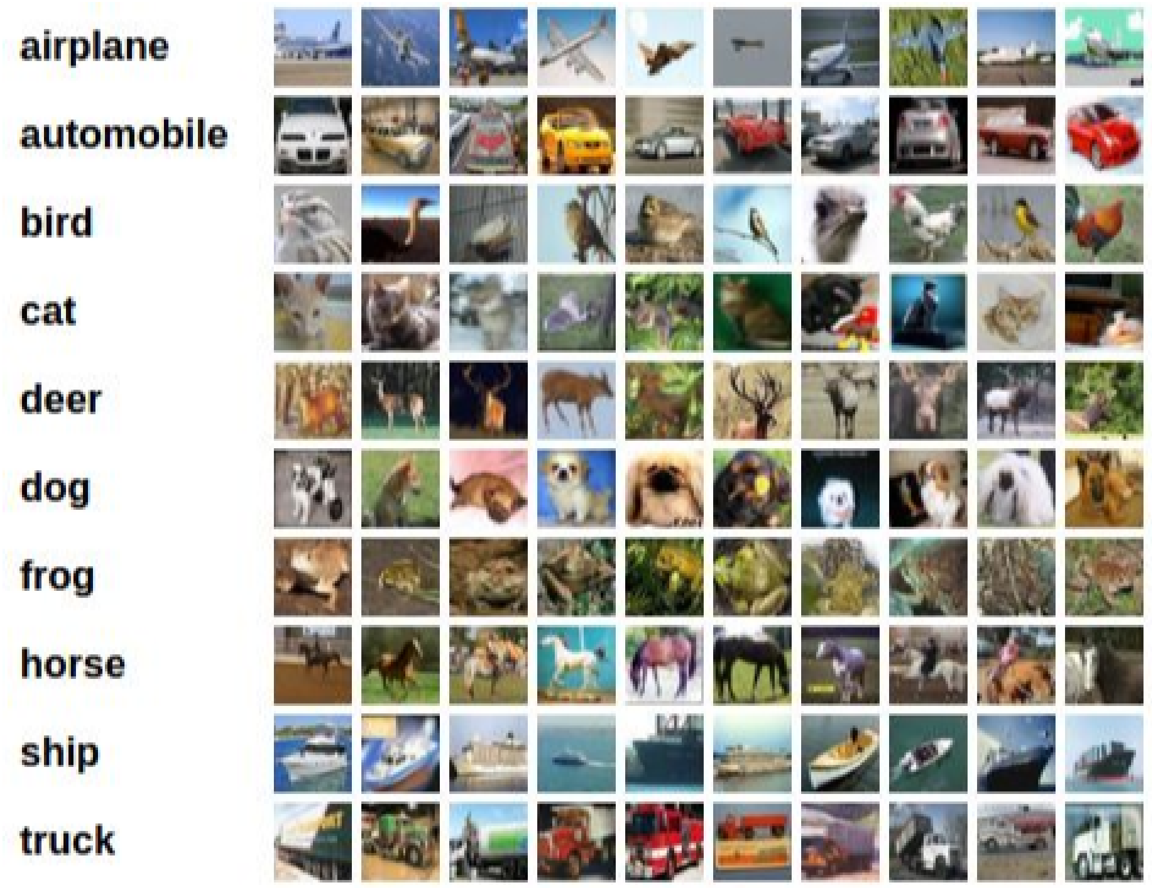



Table 1. Validation Accuracy on CIFAR10

| Model | Validation Accuracy(%) |
|---|---|
| ResNet18 | 86.18 |
| ResNet50 | 86.09 |
| VGG19 | 85.87 |
| **StanNet** | **84.62** |
| VGG16 | 84.19 |
| AlexNet | 81.37 |

**Result:**

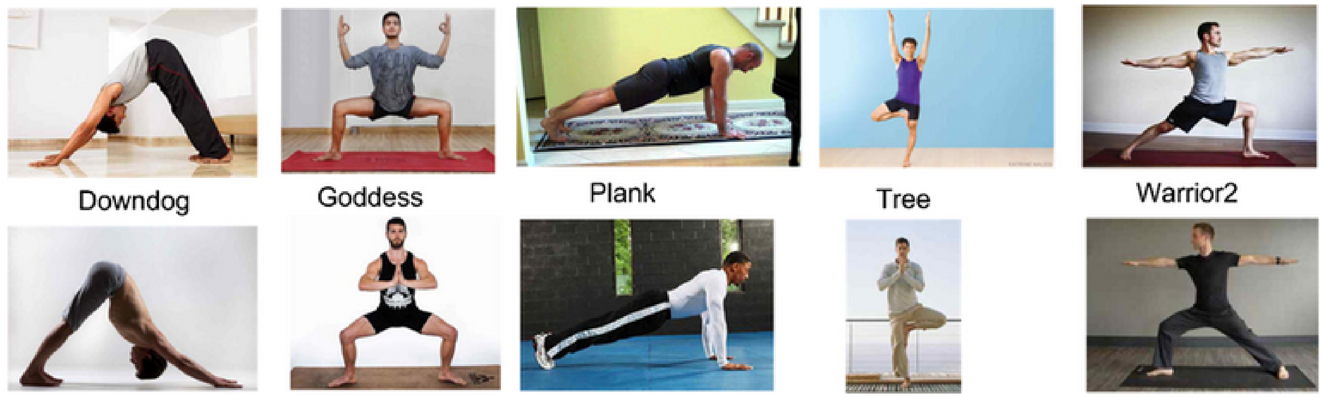
**Comparable accuracies**

(despite our model being **lightweight** and having a much **simpler architecture** than standard contenders)

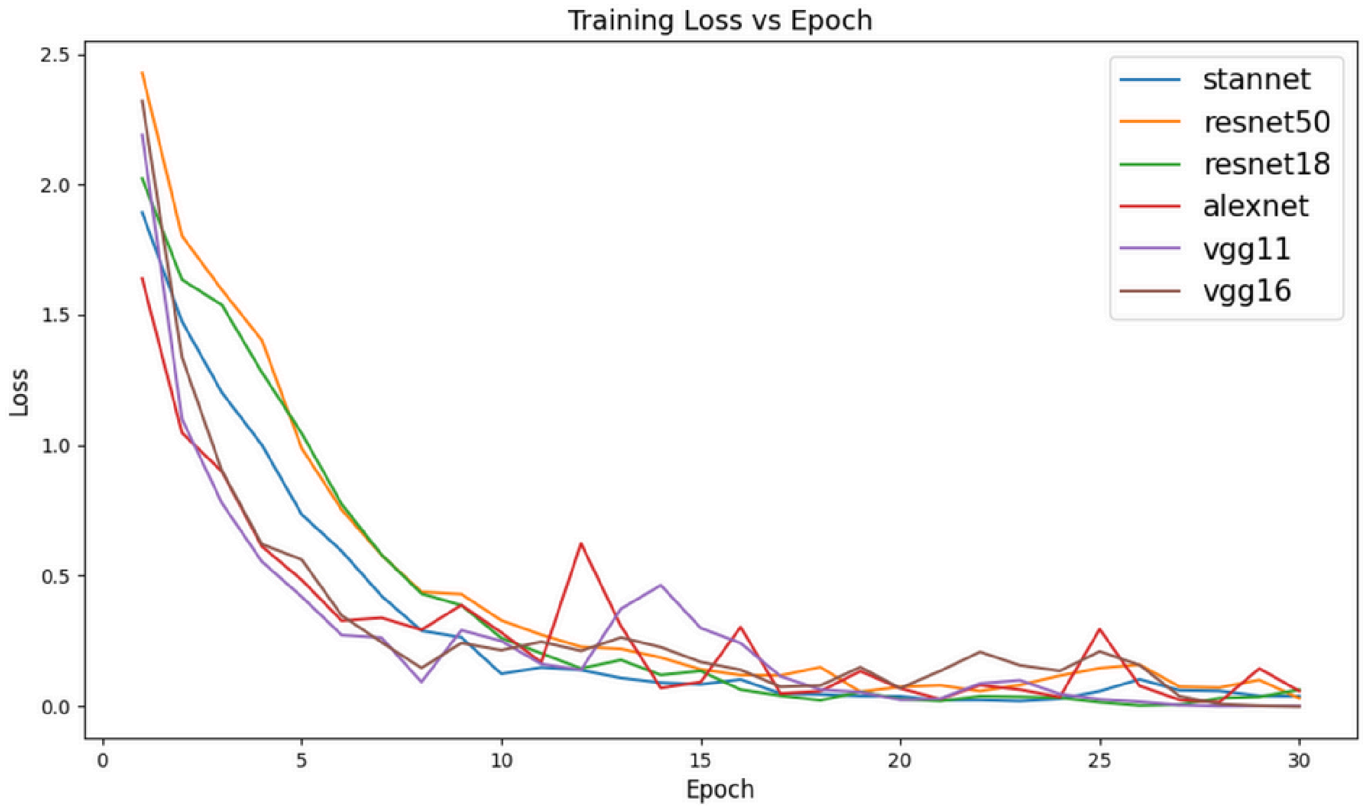
potential changes which can further improve the accuracy:
- fine tuning of hyperparameters,
- exploring more skip/residual connections,
- adding more layers/blocks,
- making our radial lowpass mask also learnable
- modifying the loss to incorporate more aspects of the complex representation.

# Dataset 2: **Yoga Poses**

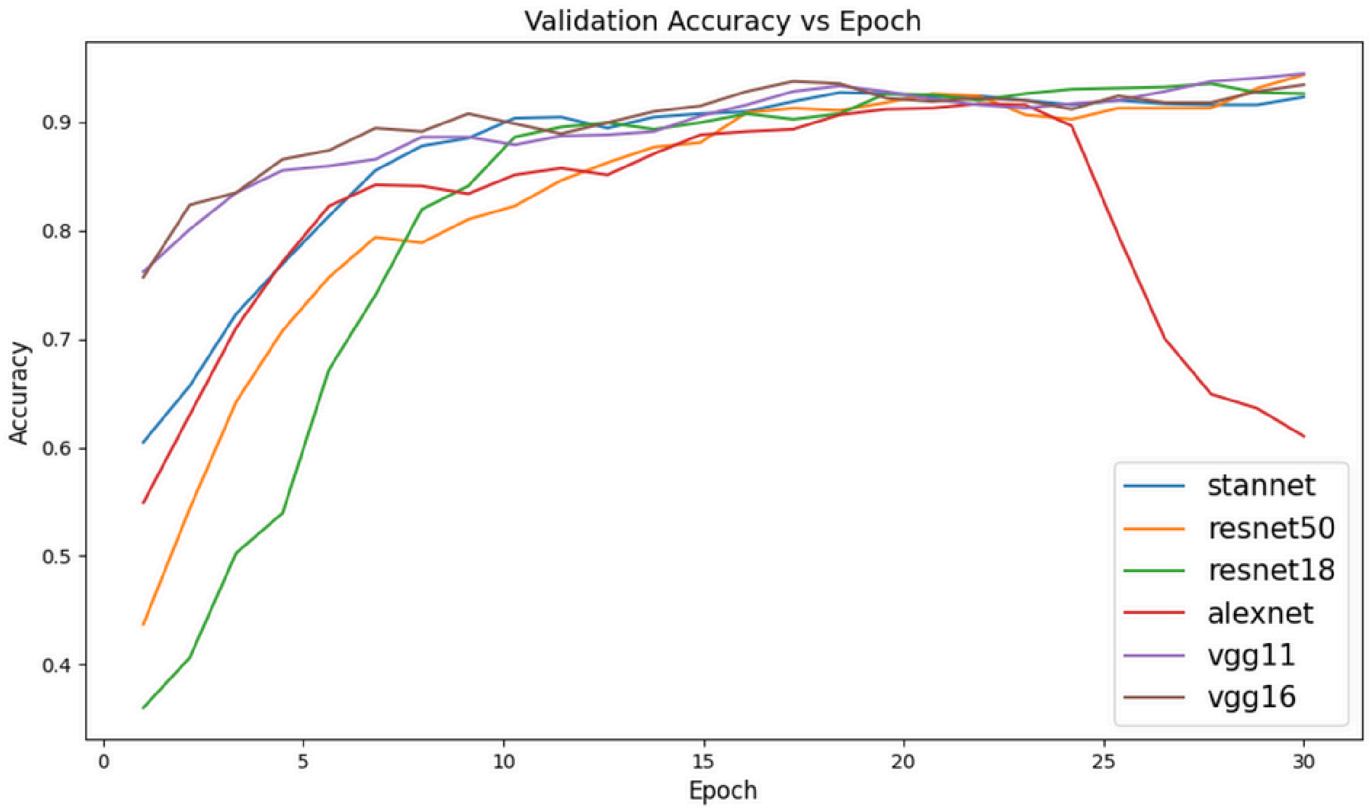
*A rather smaller dataset helps us get decents results with less intensive training*




Implementation details, same as CIFAR-10 dataset




- Better performance and faster convergence over ResNet18, ResNet50
- Better than AlexNet in consistently reducing loss
- But not better than VGG11 or VGG16




- Better performance and faster performance over ResNet18, ResNet50, and AlexNet
- VGG11 or VGG16 are still superior

Table 2. Number of Parameters

| Model | Number of Parameters ($\times 10^7$) |
|---|---|
| **StanNet** | **1.06** |
| VGG11 | 1.86 |
| ResNet18 | 2.23 |
| VGG16 | 2.96 |
| ResNet50 | 4.70 |
| AlexNet | 5.74 |

For **similar, or even better accuracies** on the validation set, our model:

- The **least number of parameters**
- The **most compact** and least resource hungry

Overall, our model consistently gives respectable **accuracies of 90%+** for this task, while being the **most lightweight** out of all the models.

# GUI Demonstration