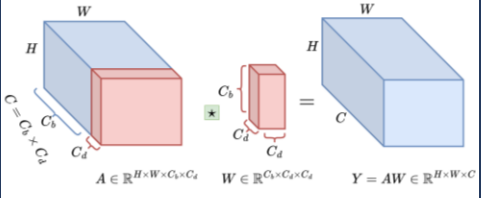
Microsoft

Scattering Vision Transformer: Spectral Mixing Matters

Badri Patro and Vijay Agneeswaran
Microsoft



SVT Model

Project Page

Badri N. Patro

Vijay Agneeswaran

Agenda

- Goals
- Introduction
- Related Work
- Method
- Data
- Results
- Business Use-Case

Goal

Given an input image, the model need to prediction a category out of 1000 predefined category.

Input Image

SVT Transformer

Predicted Top-5 Categories

256 : Newfoundland dog prob = 92.7%
247 : Saint Bernard prob = 0.1%
244 : Tibetan mastiff prob = 0.1%
260 : chow, chow chow prob = 0.0%
257 : Great Pyrenees prob = 0.0%

To make **efficient transformer in terms of parameter, Computation and** make more **robust feature representation**

Introduction

- Issues in Transformer
- Proposed Solution

Issues in Vision Transformer

Computational Complexity increase quadratic with Sequence Length.

- Thus, existing solutions commonly employ down-sampling operations (e.g., average pooling) over keys/values to dramatically reduce the computational cost.
- Unfortunately, such operations are non-invertible and can result in information loss.
- Memory (#Parameters)
- Computation Cost (#Gflops)
- Latency

Issue of capturing fine-grained information within images effectively

Proposed Solution

SCATTERING TRANSFORMATION: DTCWT

SPECTRAL GATING NETWORK : TBM AND EBM

CHANNEL AND TOKEN MIXING

Contribution

We introduce a novel invertible scattering network based on DTCWT transformation into vision transformers to decompose image features into low-frequency and high-frequency features

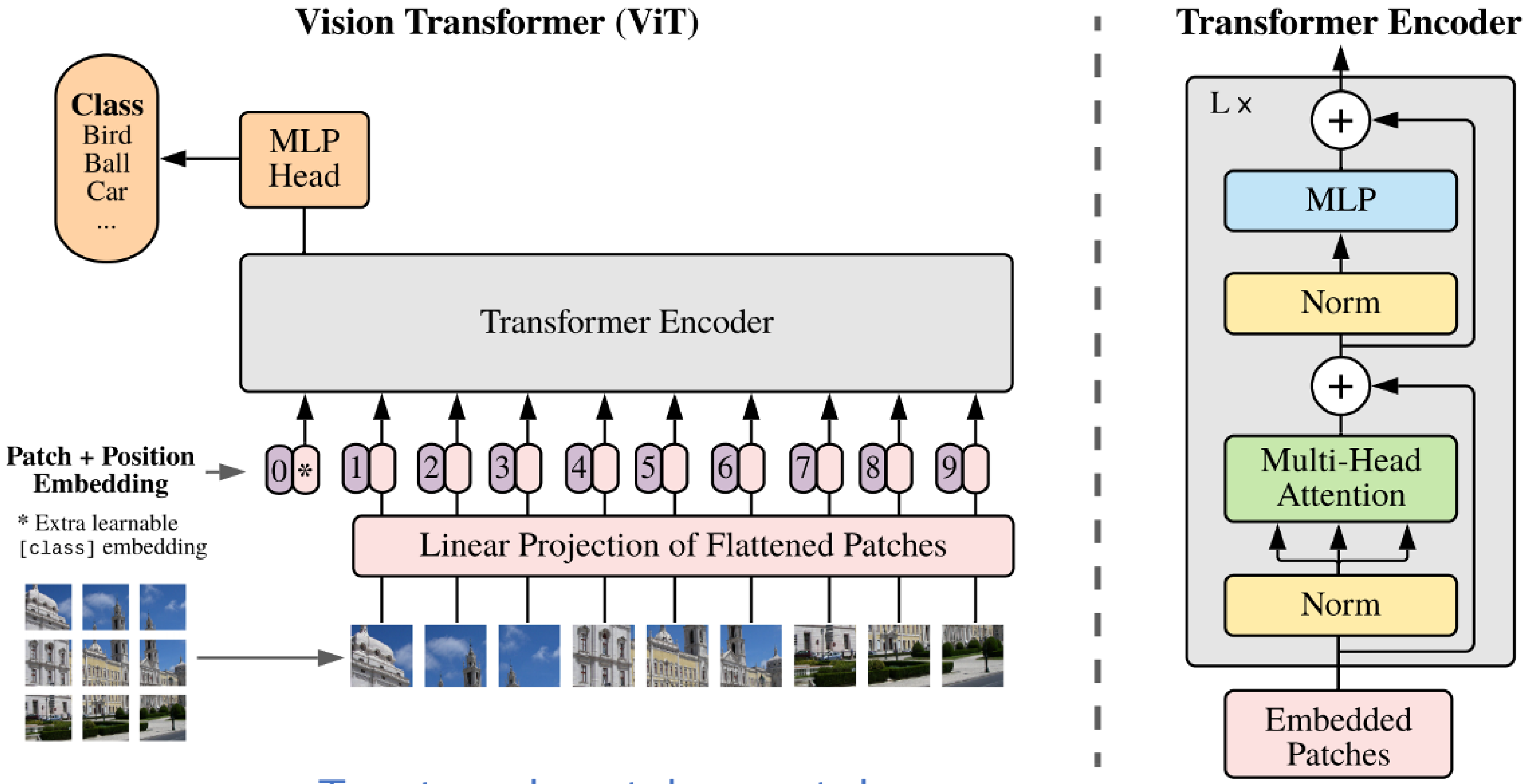
We proposed a novel SGN, which uses TBM to mix low-frequency components and EBM to mix high-frequency components.

We use an efficient way of mixing high-frequency components using channel and token mixing with the help of Einstein multiplication

We use tensor multiplication in low-frequency components and Einstein multiplication in high-frequency components leading to an efficient implementation of SVT, both in the number of parameters and computational complexity.

Related work

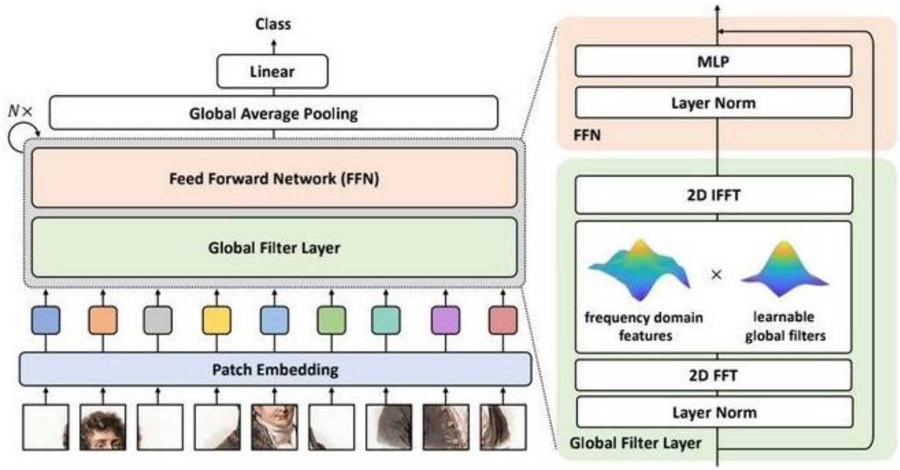
Vision Transformer



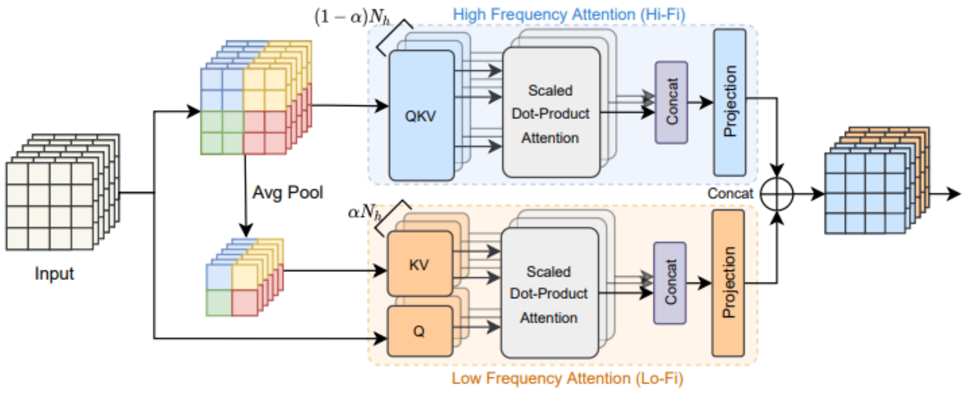
Treat each patch as a token (like a word) in NLP

Recent Developments and Views on Computer Vision x Transformer | by Akihiro FUJII | Towards Data Science

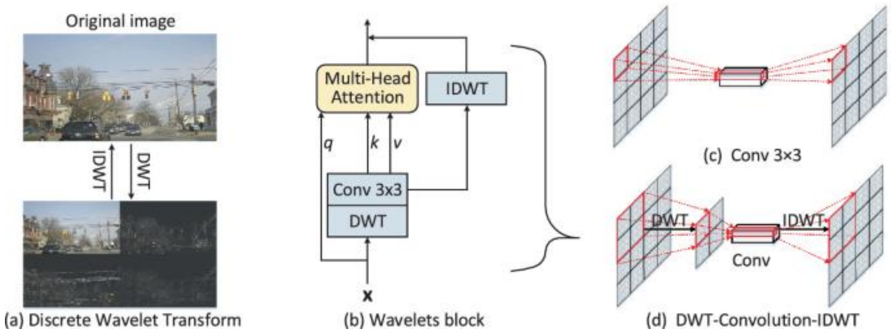
Spectral Vision Transformers



GFNet,[Rao et al. NeurIPS 2021]



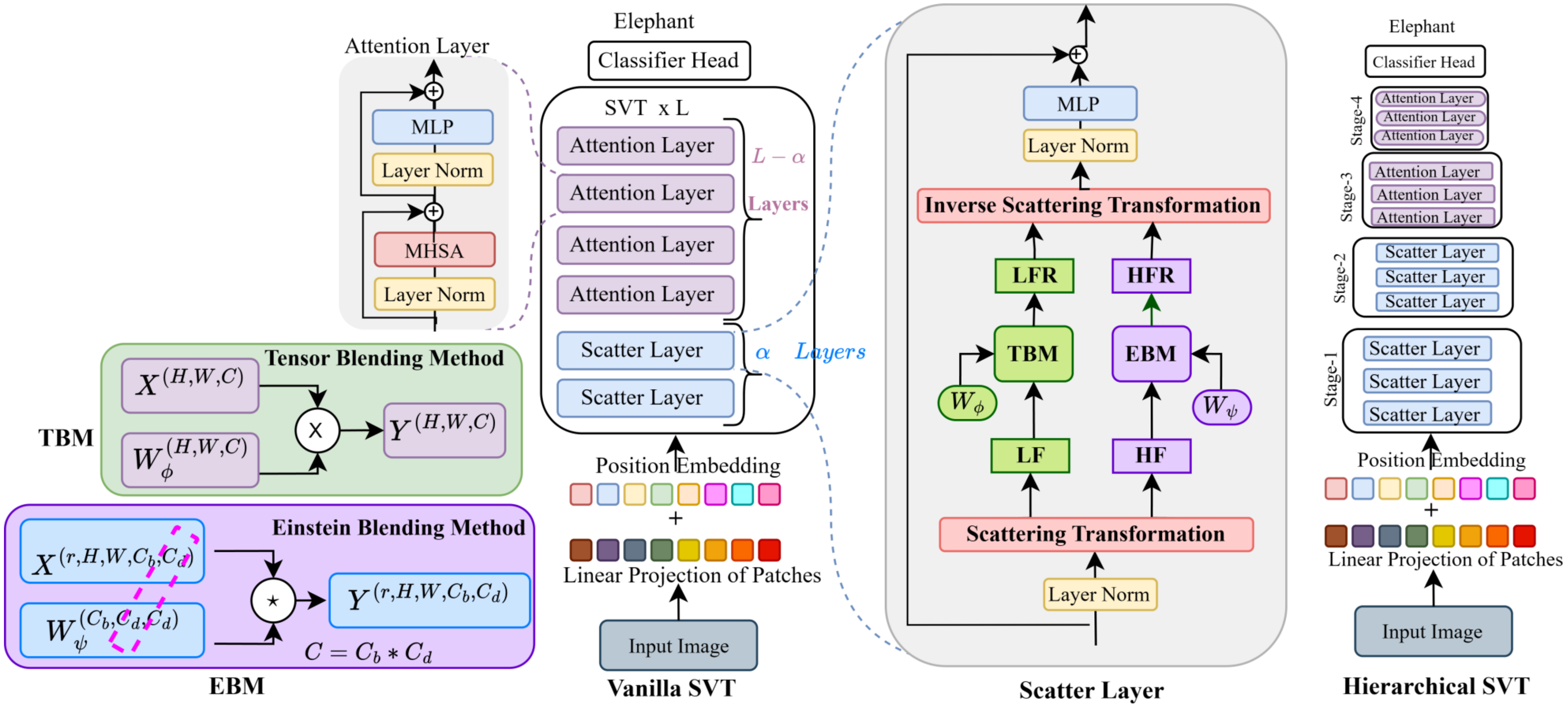
Hilo Attention [Pan et. al., NeurIPS 2022]



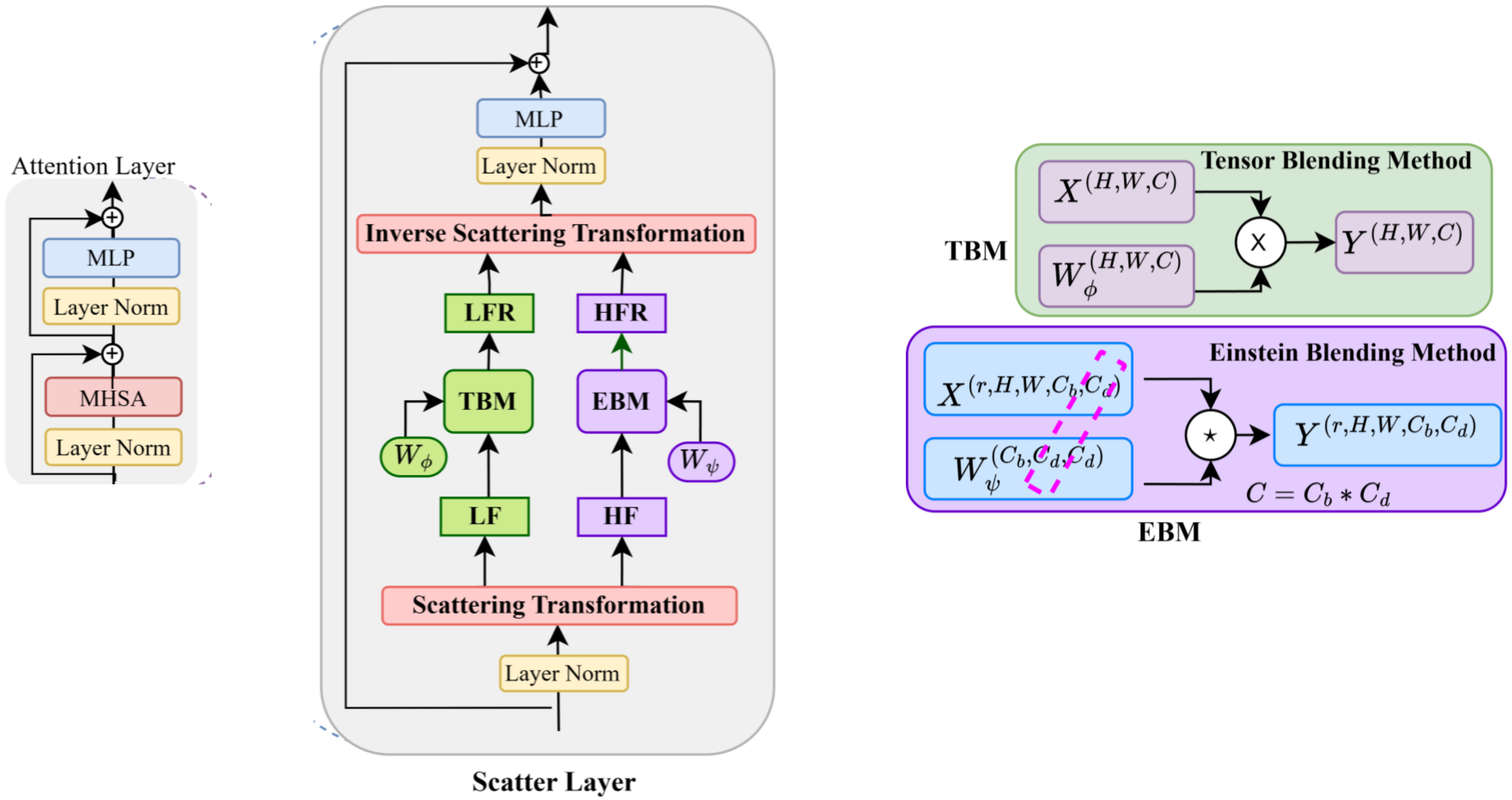
WaveViT ,[Yao et al. ECCV 2022]

Methodology

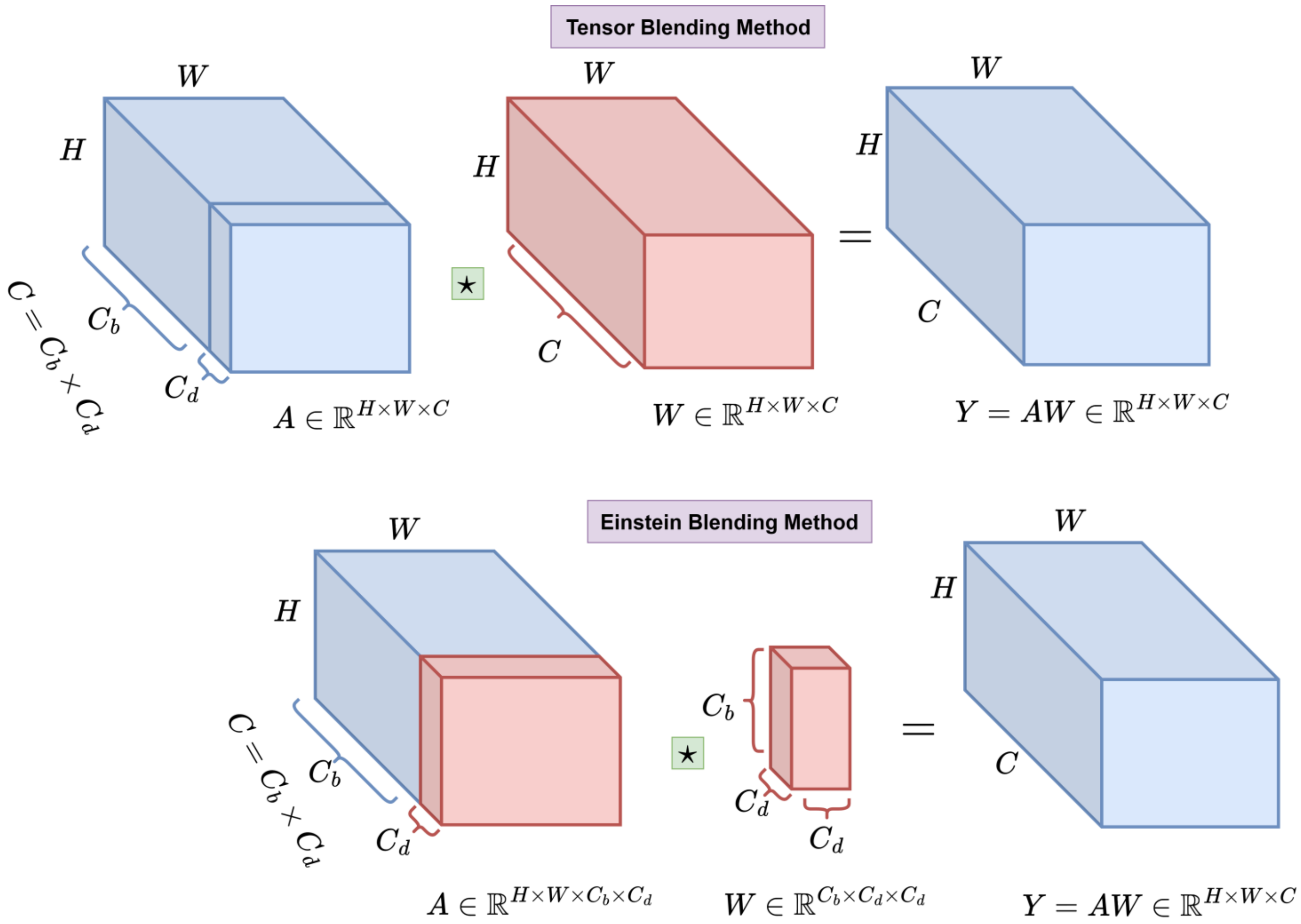
SVT Model Diagram



SVT Model Diagram



Einstein Blending Method



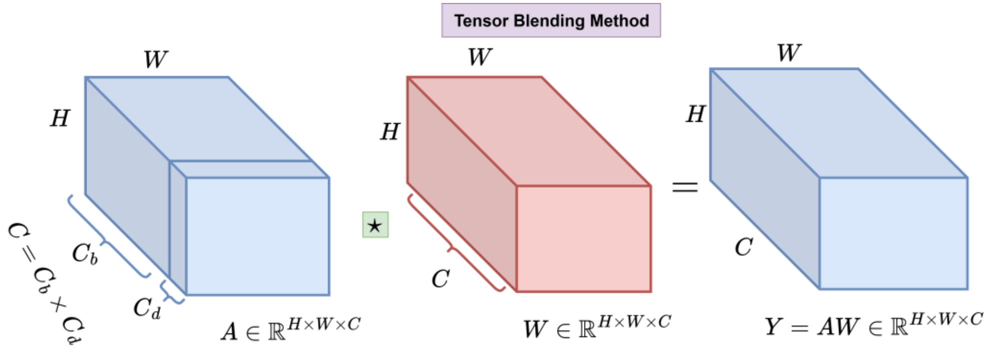
TBM

Dual Tree Complex Wavelet Transform (DTCWT)

$$\mathbf{X}_F(u,v) = \mathbf{X}_\phi(u,v) + \mathbf{X}_\psi(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} c_{M,h,w}\phi_{M,h,w} + \sum_{m=0}^{M-1}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\sum_{k=1}^{6} d^k_{m,h,w}\psi^k_{m,h,w}$$

Tensor Blending Method

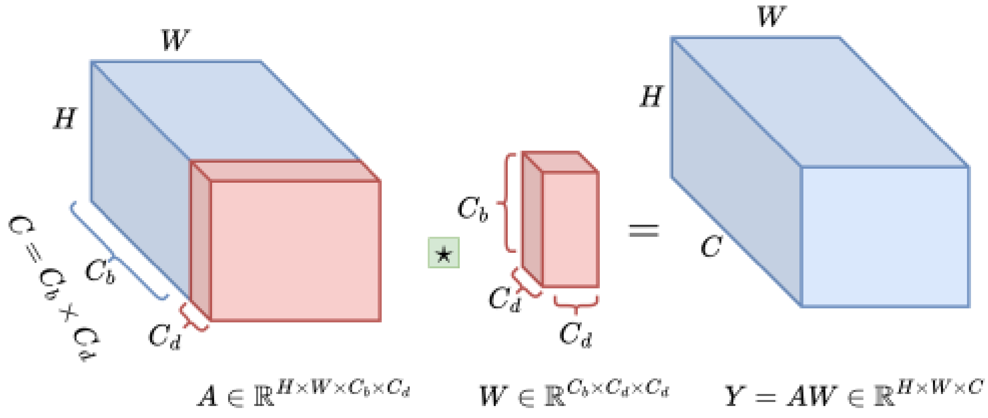
$$\mathcal{M}_\phi = [\mathbf{X}_\phi \odot \mathcal{W}_\phi], \qquad \text{where } (\mathbf{X}_\phi, \mathcal{W}_\phi) \in \mathcal{R}^{C\times H\times W}, \text{ and } \mathbf{M}_\phi \in \mathcal{R}^{C\times H\times W},$$



EBM

Einstein Blending Method

$$\mathbf{Y}^{H\times W\times C_b\times C_d} = \mathbf{A}^{H\times W\times C_b\times C_d} \circledast \mathbf{W}^{C_b\times C_d\times C_d}$$



EBM for Channel Mixing

$$\mathbf{S}_{\psi_c}^{2\times k\times H\times W\times C_b\times C_d} = \mathbf{X}_{\psi}^{2\times k\times H\times W\times C_b\times C_d} \circledast \mathbf{W}_{\psi_c}{}^{C_b\times C_d\times C_d} + b_{\psi_c}$$

EBM for Token Mixing

$$\mathbf{S}_{\psi_t}^{2\times k\times C\times W\times H} = \mathbf{S}_{\psi_c}^{2\times k\times C\times W\times H} \circledast \mathbf{W}_{\psi_t}{}^{W\times H\times H} + b_{\psi_t}$$

Data

Data

ImageNet Dataset Details:

- Train Set examples =**1200K**
- Test Set examples = **50K**

The CIFAR-10, and CIFAR 100 Dataset Details:

- Train Set examples =**50K**
- Test Set examples = **10K**

Flower Dataset

- Training set: The training set consists of **1,020 images** of flowers, with at least 10 images per category.
- Test set: The test set consists of the remaining **7,169 images** of flowers, with at least 20 images per category.

Stanford CAR Dataset

- Training set: The training set consists of 8,144 images of cars from 98 different classes.
- Test set: The test set consists of 8,041 images of cars from the remaining 98 classes.
- Link : SVT_Data.docx (sharepoint.com)

Results

SOTA on ImageNet

Method	Params	GFLOPs	Top-1	Top-5
Small				
ResNet-50 [22]	25.5M	4.1	78.3	94.3
BoTNet-S1-50 [48]	20.8M	4.3	80.4	95.0
Cross-ViT-S [6]	26.7M	5.6	81.0	-
Swin-T [37]	29.0M	4.5	81.2	95.5
ConViT-S [15]	27.8M	5.4	81.3	95.7
T2T-ViT-14 [68]	21.5M	4.8	81.5	95.7
RegionViT-Ti+ [5]	14.3M	2.7	81.5	-
SE-CoTNetD-50 [34]	23.1M	4.1	81.6	95.8
Twins-SVT-S [10]	24.1M	2.9	81.7	95.6
CoaT-Lite Small [64]	20.0M	4.0	81.9	95.5
PVTv2-B2 [58]	25.4M	4.0	82.0	96.0
LITv2-S [40]	28.0M	3.7	82.0	-
MViTv2-T [33]	24.0M	4.7	82.3	-
Wave-ViT-S [66]	19.8M	4.3	82.7	96.2
CSwin-T [13]	23.0M	4.3	82.7	-
DaViT-Ti [12]	28.3M	4.5	82.8	-
SVT-H-S	21.7M	3.9	83.1	96.3
iFormer-S[47]	20.0M	4.8	83.4	96.6
CMT-S [18]	25.1M	4.0	83.5	-
MaxViT-T [54]	31.0M	5.6	83.6	-
Wave-ViT-S* [66]	22.7M	4.7	83.9	96.6
SVT-H-S*	**22.0M**	**3.9**	**84.2**	**96.9**
Base				
ResNet-101 [22]	44.6M	7.9	80.0	95.0
BoTNet-S1-59 [48]	33.5M	7.3	81.7	95.8
T2T-ViT-19 [68]	39.2M	8.5	81.9	95.7
CvT-21 [60]	32.0M	7.1	82.5	-
GFNet-H-B [44]	54.0M	8.6	82.9	96.2
Swin-S [37]	50.0M	8.7	83.2	96.2
Twins-SVT-B [10]	56.1M	8.6	83.2	96.3
SE-CoTNetD-101 [34]	40.9M	8.5	83.2	96.5
PVTv2-B3 [58]	45.2M	6.9	83.2	96.5
LITv2-M [40]	49.0M	7.5	83.3	-
RegionViT-M+ [5]	42.0M	7.9	83.4	-
MViTv2-S [33]	35.0M	7.0	83.6	-
CSwin-S [13]	35.0M	6.9	83.6	-
DaViT-S [12]	49.7M	8.8	84.2	-
VOLO-D1* [69]	26.6M	6.8	84.2	-
CMT-B [18]	45.7M	9.3	84.5	-
MaxViT-S [54]	69.0M	11.7	84.5	-
iFormer-B[47]	48.0M	9.4	84.6	97.0
Wave-ViT-B* [66]	33.5M	7.2	84.8	97.1
SVT-H-B*	**32.8M**	**6.3**	**85.2**	**97.3**

Method	Params	GFLOPs	Top-1	Top-5
Large				
ResNet-152 [22]	60.2M	11.6	81.3	95.5
ResNeXt101 [63]	83.5M	15.6	81.5	-
gMLP-B [36]	73.0M	15.8	81.6	-
DeiT-B [52]	86.6M	17.6	81.8	95.6
SE-ResNet-152 [24]	66.8M	11.6	82.2	95.9
Cross-ViT-B [6]	104.7M	21.2	82.2	-
ResNeSt-101 [71]	48.3M	10.2	82.3	-
ConViT-B [15]	86.5M	16.8	82.4	95.9
PoolFormer-M48 [67]	73.0M	11.8	82.5	-
T2T-ViTt-24 [68]	64.1M	15.0	82.6	95.9
TNT-B [20]	65.6M	14.1	82.9	96.3
CycleMLP-B4 [7]	52.0M	10.1	83.0	-
DeepViT-L [74]	58.9M	12.8	83.1	-
RegionViT-B [5]	72.7M	13.0	83.2	96.1
CycleMLP-B5 [7]	76.0M	12.3	83.2	-
ViP-Large/7 [23]	88.0M	24.4	83.2	-
CaiT-S36 [53]	68.4M	13.9	83.3	-
AS-MLP-B [35]	88.0M	15.2	83.3	-
BoTNet-S1-128 [48]	75.1M	19.3	83.5	96.5
Swin-B [37]	88.0M	15.4	83.5	96.5
Wave-MLP-B [49]	63.0M	10.2	83.6	-
LITv2-B [40]	87.0M	13.2	83.6	-
PVTv2-B4 [58]	62.6M	10.1	83.6	96.7
ViL-Base [72]	55.7M	13.4	83.7	-
Twins-SVT-L [10]	99.3M	15.1	83.7	96.5
Hire-MLP-Large [19]	96.0M	13.4	83.8	-
RegionViT-B+ [5]	73.8M	13.6	83.8	-
Focal-Base [65]	89.8M	16.0	83.8	96.5
PVTv2-B5 [58]	82.0M	11.8	83.8	96.6
SE-CoTNetD-152 [34]	55.8M	17.0	84.0	97.0
DAT-B [61]	88.0M	15.8	84.0	-
LV-ViT-M* [26]	55.8M	16.0	84.1	96.7
CSwin-B [13]	78.0M	15.0	84.2	-
HorNet-B_{GF} [43]	88.0M	15.5	84.3	-
DynaMixer-L [59]	97.0M	27.4	84.3	-
MViTv2-B [33]	52.0M	10.2	84.4	-
DaViT-B [12]	87.9M	15.5	84.6	-
CMT-L [18]	74.7M	19.5	84.8	-
MaxViT-B [54]	120.0M	23.4	85.0	-
VOLO-D2* [69]	58.7M	14.1	85.2	-
VOLO-D3* [69]	86.3M	20.6	85.4	-
Wave-ViT-L* [66]	57.5M	14.8	85.5	97.3
SVT-H-L*	**54.0M**	**12.7**	**85.7**	**97.5**

Similar Architect

Table 3: This shows a performance comparison of SVT with similar Transformer Architecture with different sizes of the networks on ImageNet-1K. ⋆ indicates additionally trained with the Token Labeling objective using MixToken[26].

Network	Params	GFLOPs	Top-1	Top-5
Vanilla Transformer Comparison				
FFC-ResNet-50 [8]	26.7M	-	77.8	-
FourierFormer [38]	-	-	73.3	91.7
GFNet-Ti [44]	7M	1.3	74.6	92.2
SVT-T	9M	1.8	**76.9**	**93.4**
FFC-ResNet-101 [8]	46.1M	-	78.8	-
Fnet-S [31]	15M	2.9	71.2	-
GFNet-XS [44]	16M	2.9	78.6	94.2
GFNet-S [44]	25M	4.5	80.0	94.9
SVT-XS	19.9M	4.0	**79.9**	**94.5**
SVT-S	32.2M	6.6	**81.5**	**95.3**
FFC-ResNet-152 [8]	62.6M	-	78.9	-
GFNet-B [44]	43M	7.9	80.7	95.1
SVT-B	57.6M	11.8	**82.0**	**95.6**
Hierarchical Transformer Comparison				
GFNet-H-S [44]	32M	4.6	81.5	95.6
LIT-S [41]	27M	4.1	81.5	-
iFormer-S[47]	20	4.8	83.4	96.6
Wave-ViT-S⋆ [66]	22.7M	4.7	83.9	96.6
SVT-H-S	21.7M	3.9	83.1	96.3
SVT-H-S⋆	22.0M	3.9	**84.2**	**96.9**
GFNet-H-B [44]	54M	8.6	82.9	96.2
LIT-M [41]	48M	8.6	83.0	-
LITv2-M [40]	49.0M	7.5	83.3	-
iFormer-B[47]	48	9.4	84.6	97.0
Wave-MLP-B [49]	63.0M	10.2	83.6	-
Wave-ViT-B⋆ [66]	33.5M	7.2	84.8	**97.0**
SVT-H-B⋆	32.8M	6.3	**85.2**	**97.3**
LIT-B [41]	86M	15.0	83.4	-
LITv2-B [40]	87.0M	13.2	83.6	-
HorNet-B_{GF} [43]	88.0M	15.5	84.3	-
iFormer-L[47]	87.0M	14.0	84.8	**97.0**
Wave-ViT-L⋆ [66]	57.5M	14.8	85.5	97.3
SVT-H-L⋆	54.0M	12.7	**85.7**	**97.5**

Model Performance

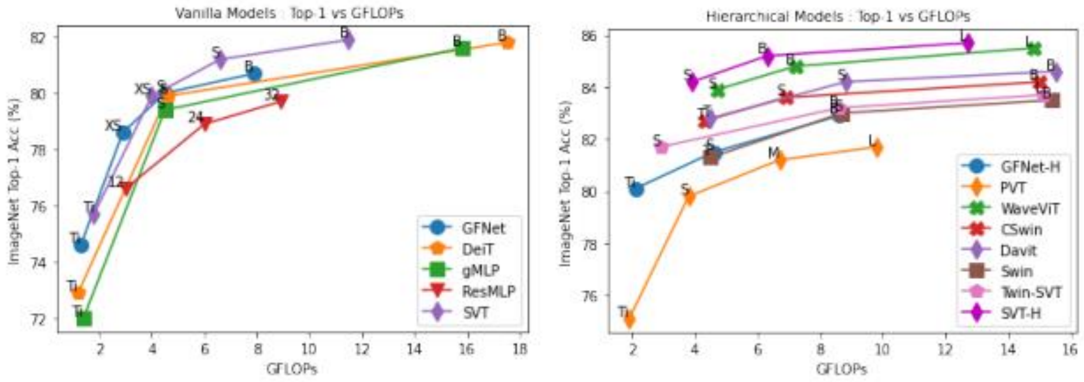


Figure 2: Comparison of ImageNet Top-1 Accuracy (%) vs GFLOPs of various models in Vanilla and Hierarchical architecture.

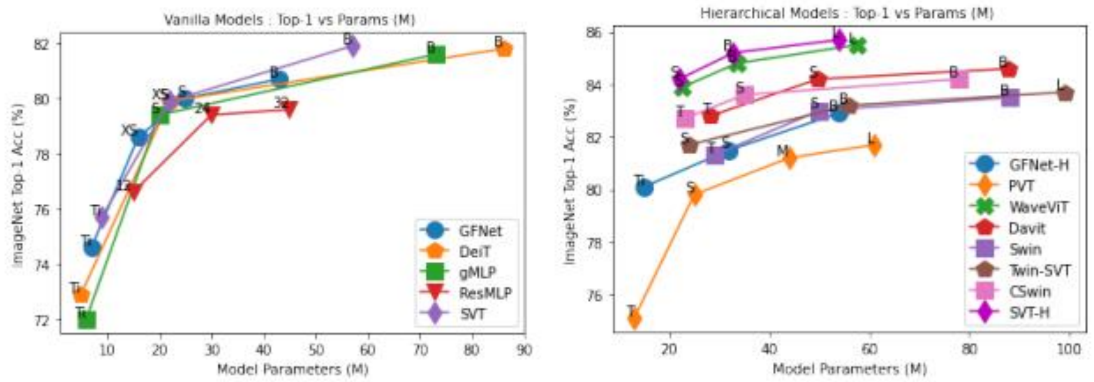


Figure 3: Comparison of ImageNet Top-1 Accuracy (%) vs Parameters (M) of various models in Vanilla and Hierarchical architecture.

Ablation Analysis

Backbone	Low Frequency		High Frequency		Params (M)	FLOPs (G)	Top-1 (%)	Top-5 (%)
	Token	Channel	Token	Channel				
SVT_{TTTT}	T	T	T	T	25.18	4.4	83.97	96.86
SVT_{EETT}	E	E	T	T	21.90	4.1	83.87	96.67
SVT_{EEEE}	E	E	E	E	21.87	3.7	83.70	96.56
SVT_{TTEE}	T	T	E	E	22.01	3.9	84.20	96.82
SVT_{TTEX}	T	T	E	✗	21.99	4.0	84.06	96.76
SVT_{TTXE}	T	T	✗	E	22.25	4.1	84.12	96.91

Table 7: SVT model comprises low-frequency component and High-frequency component with the help of scattering net using Dual tree complex wavelet transform. Each frequency component is controlled by parameterized by weight matrix using Patch mixing and/or Channel Mixing. this table shows details about all combinations and SVT_{TTEE} is outperforms among them.

Ablation Analysis

Table 4: This table shows the ablation analysis of various spectral layers in SVT architecture such as FN, FFC, WGN, and FNO. We conduct this ablation study on the small-size networks in stage architecture. This indicates that SVT performs better than other kinds of networks.

Model	Params (M)	FLOPs (G)	Top-1 (%)	Top-5 (%)	Invertible loss(↓)
FFC	21.53	4.5	83.1	95.23	–
FN	21.17	3.9	84.02	96.77	–
FNO	21.33	3.9	84.09	96.86	3.27e-05
WGN	21.59	3.9	83.70	96.56	8.90e-05
SVT	22.22	3.9	**84.20**	**96.93**	6.64e-06

Transfer Learning

Table 5: **Results on transfer learning datasets.** We report the top-1 accuracy on the four datasets.

Model	CIFAR 10	CIFAR 100	Flowers 102	Cars 196
ResNet50 [22]	-	-	96.2	90.0
ViT-B/16 [14]	98.1	87.1	89.5	-
ViT-L/16 [14]	97.9	86.4	89.7	-
Deit-B/16 [52]	99.1	90.8	98.4	92.1
ResMLP-24 [51]	98.7	89.5	97.9	89.5
GFNet-XS [44]	98.6	89.1	98.1	92.8
GFNet-H-B [44]	99.0	90.3	98.8	93.2
SVT-H-B	**99.22**	**91.2**	**98.9**	**93.6**

Task Learning

Table 6: The performances of various vision backbones on COCO val2017 dataset for the downstream instance segmentation task such as Mask R-CNN 1x [21] method. We adopt Mask R-CNN as the base model, and the bounding box and mask Average Precision (*i.e.*, AP^b and AP^m) are reported for evaluation

Backbone	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet50 [22]	38.0	58.6	41.4	34.4	55.1	36.7
Swin-T [37]	42.2	64.6	46.2	39.1	61.6	42.0
Twins-SVT-S [10]	43.4	66.0	47.3	40.3	63.2	43.4
LITv2-S [40]	44.9	-	-	40.8	-	-
RegionViT-S [5]	44.2	-	-	40.8	-	-
PVTv2-B2 [58]	45.3	67.1	49.6	41.2	64.2	44.4
SVT-S	**46.0**	**68.1**	**50.4**	**41.9**	**65.0**	**45.1**

Filter Characterisation

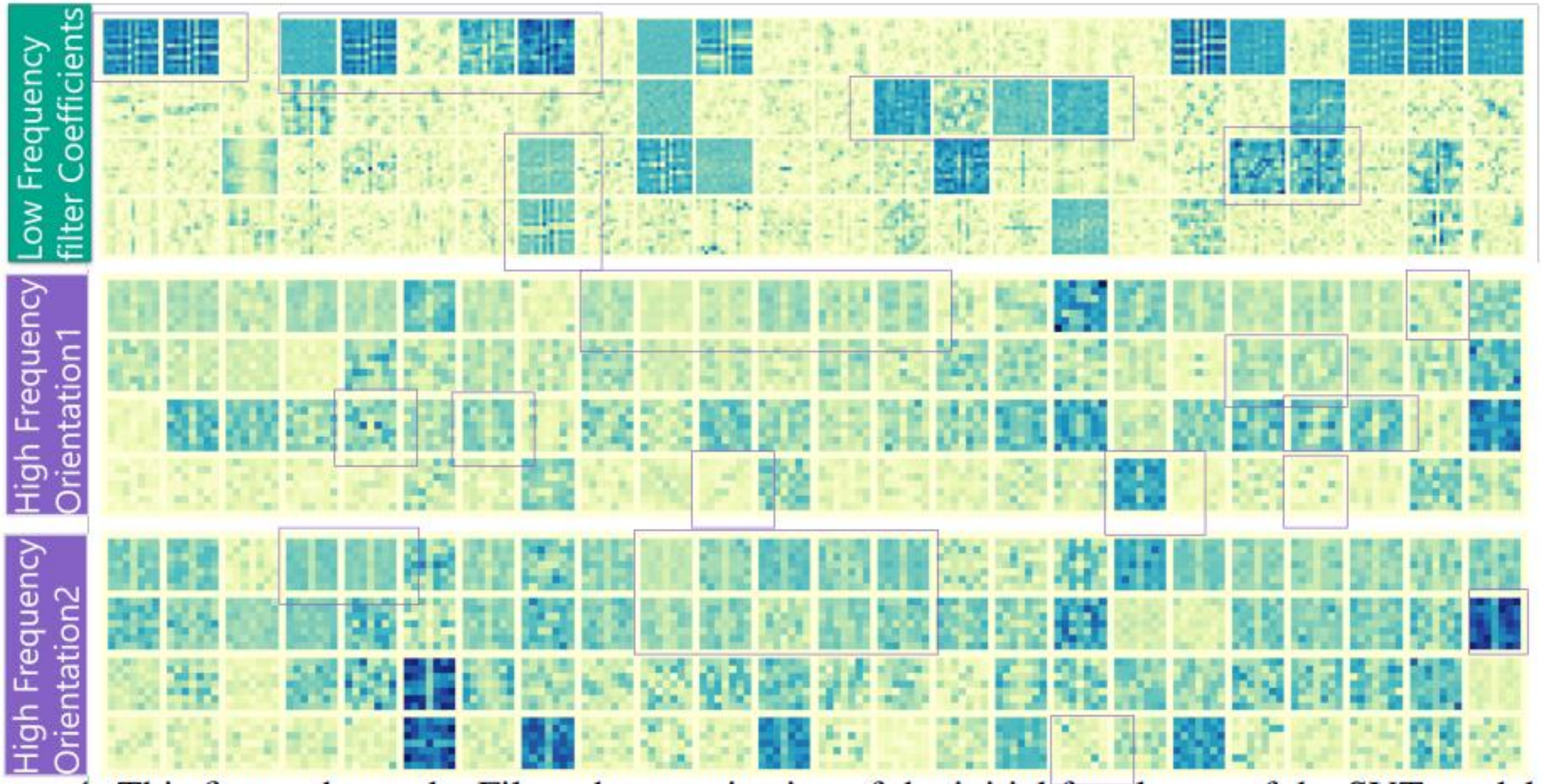


Figure 4: This figure shows the Filter characterization of the initial four layers of the SVT model. It clearly shows that High-frequency filter coefficient c captures local filter information such as lines, edges, and different orientations of an Image. The Low-frequency filter coefficient captures the shape with the maximum energy part in the image.

Business Use-Case

Business Use-Case

- Extreme Classification: Fine grain Classification
- Medical Imaging: X-rays, CT scans, MRI scans
- Satellite imaging: Infra-red Images
- Audio and Speech Applications
- Signal Processing Applications

Q&A

Thank You