ETH zürich

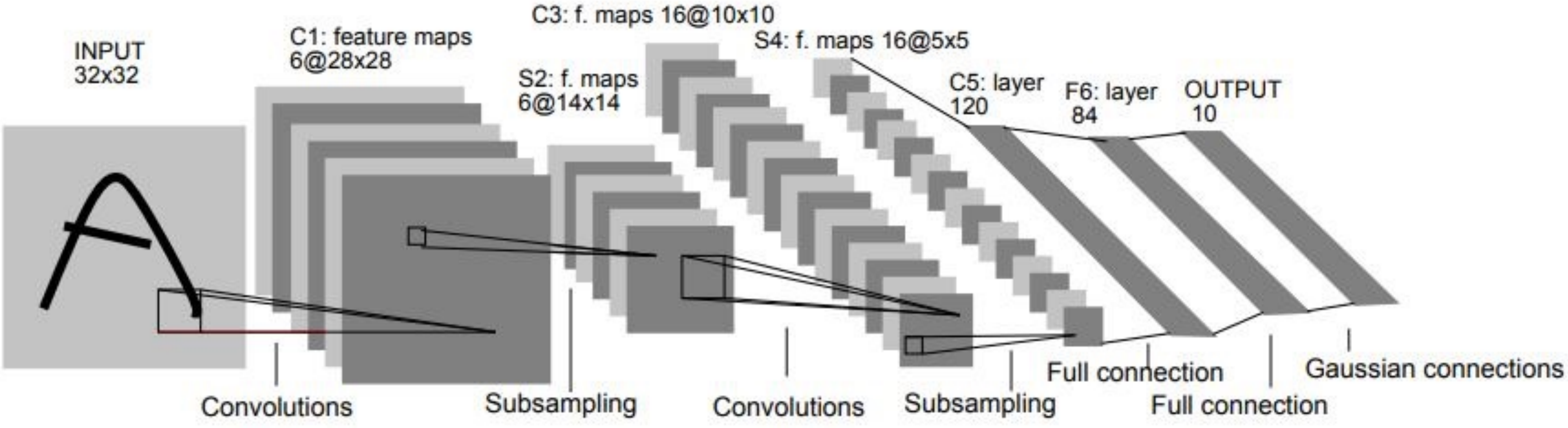
Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles

Thomas Kiefer

Seminar in Deep Neural Networks (FS 2024)
23 April 2024, ETH Zurich

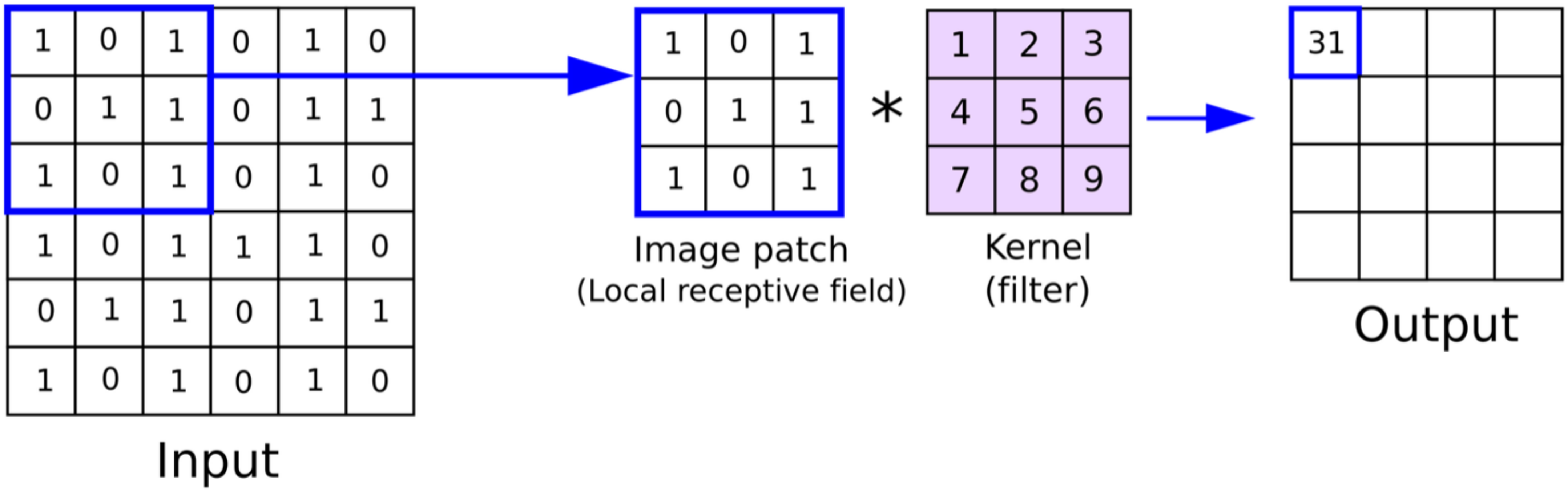
Architectures for Image Classification

LeNet-5



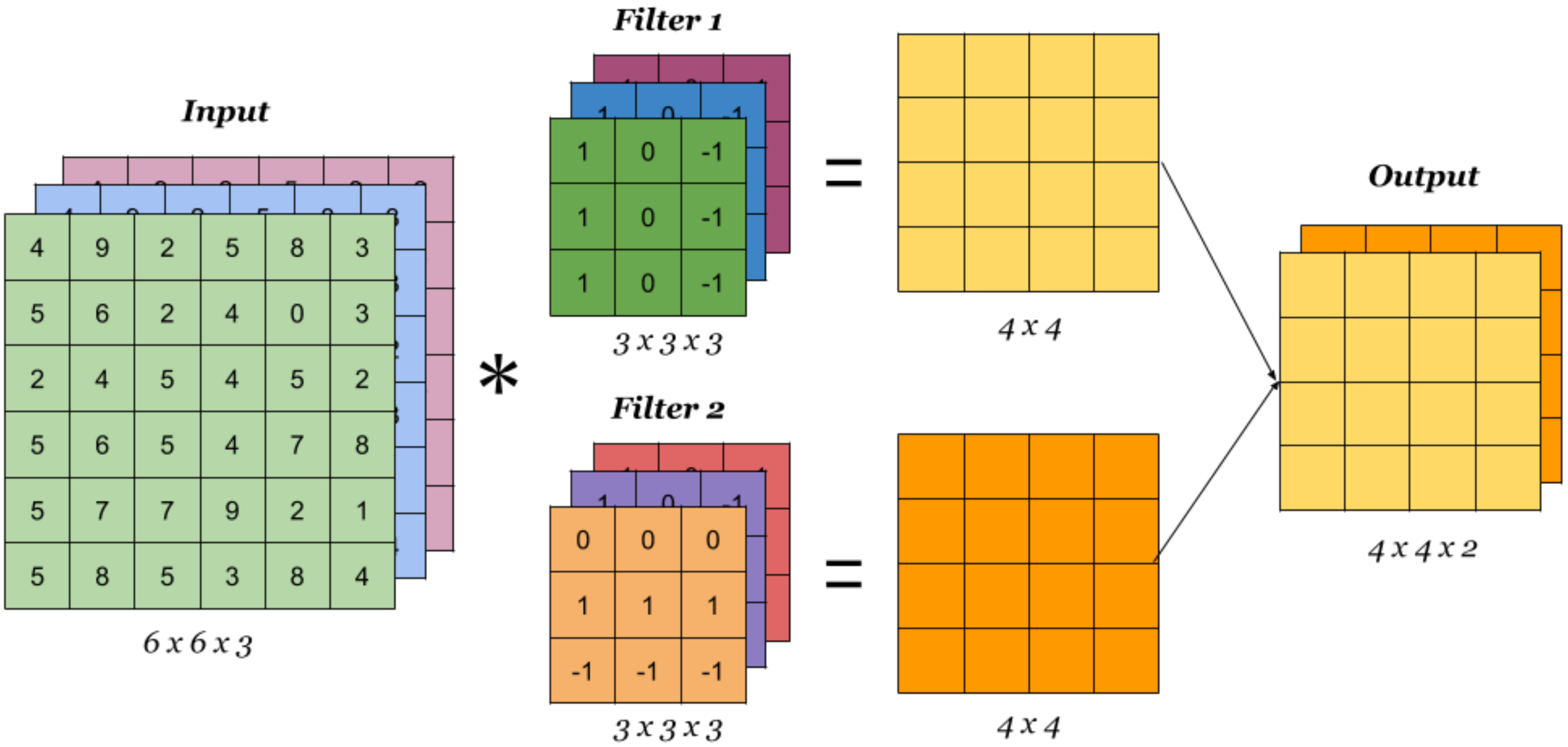
LeCun et al. (1998)

Convolutions



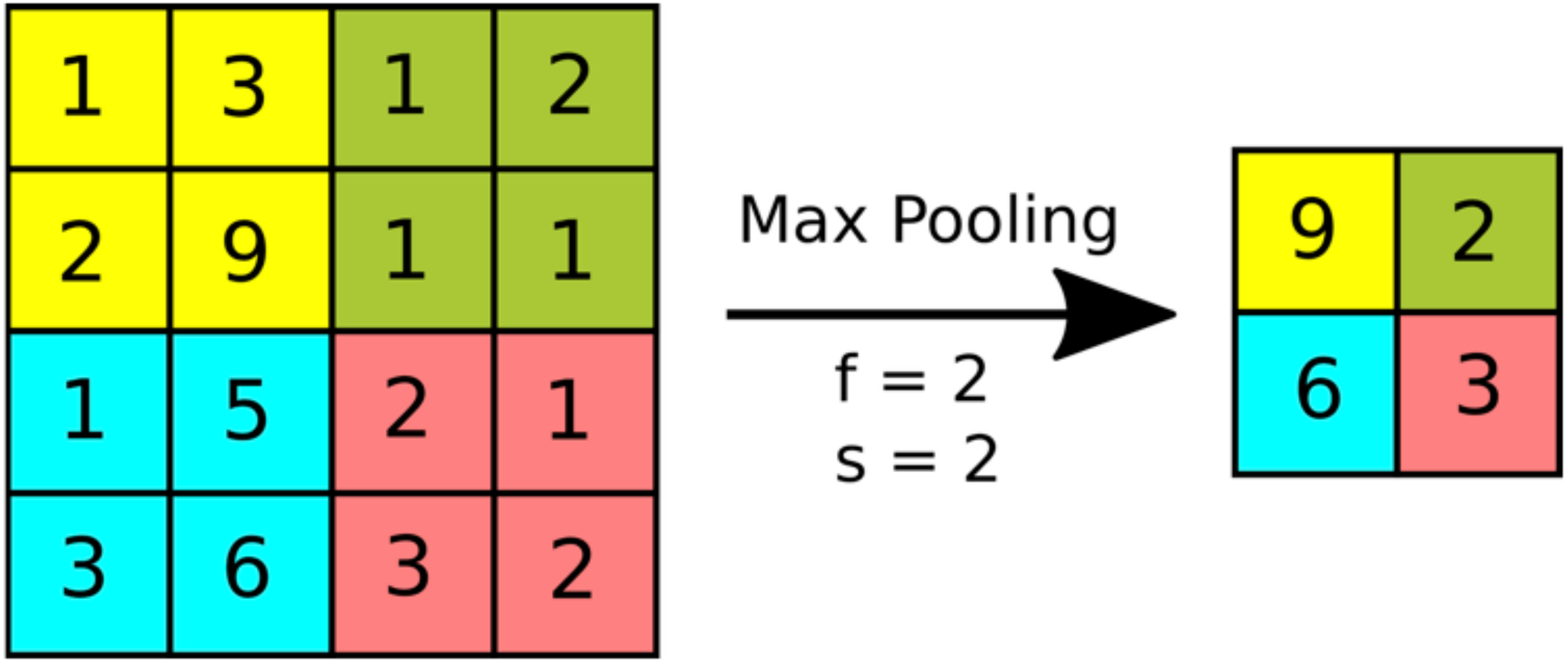
https://anhreynolds.com/blogs/cnn.html

Convolutions



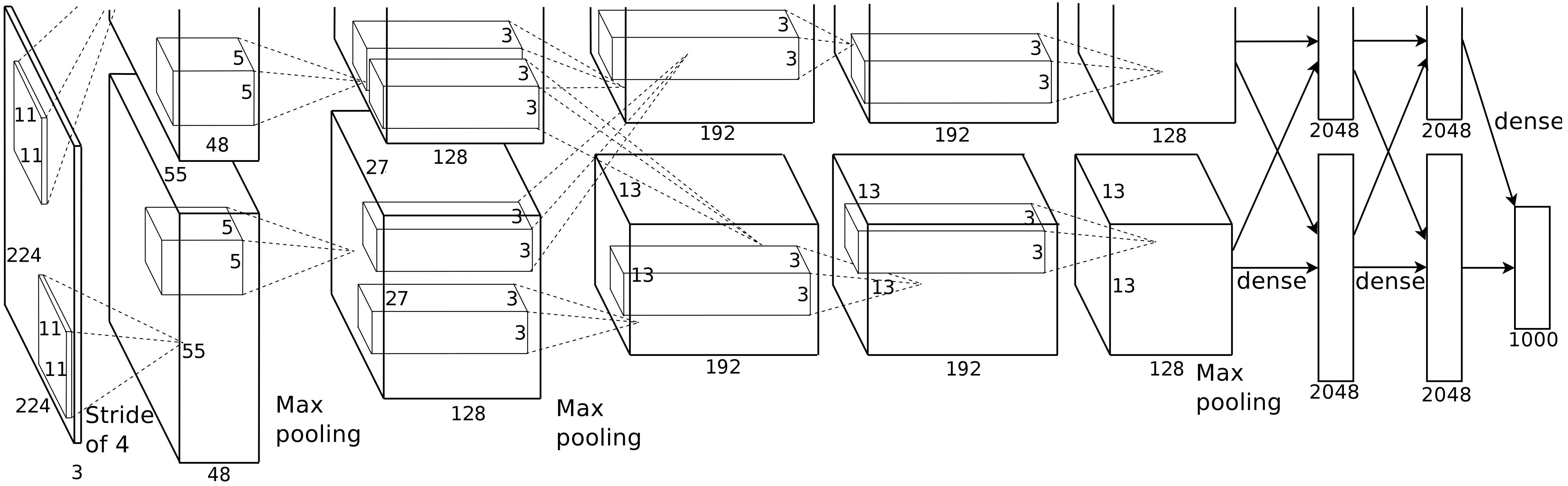
https://indoml.com/2018/03/07/student-notes-convolutional-neural-networks-cnn-introduction

Pooling



https://anhreynolds.com/blogs/cnn.html

AlexNet



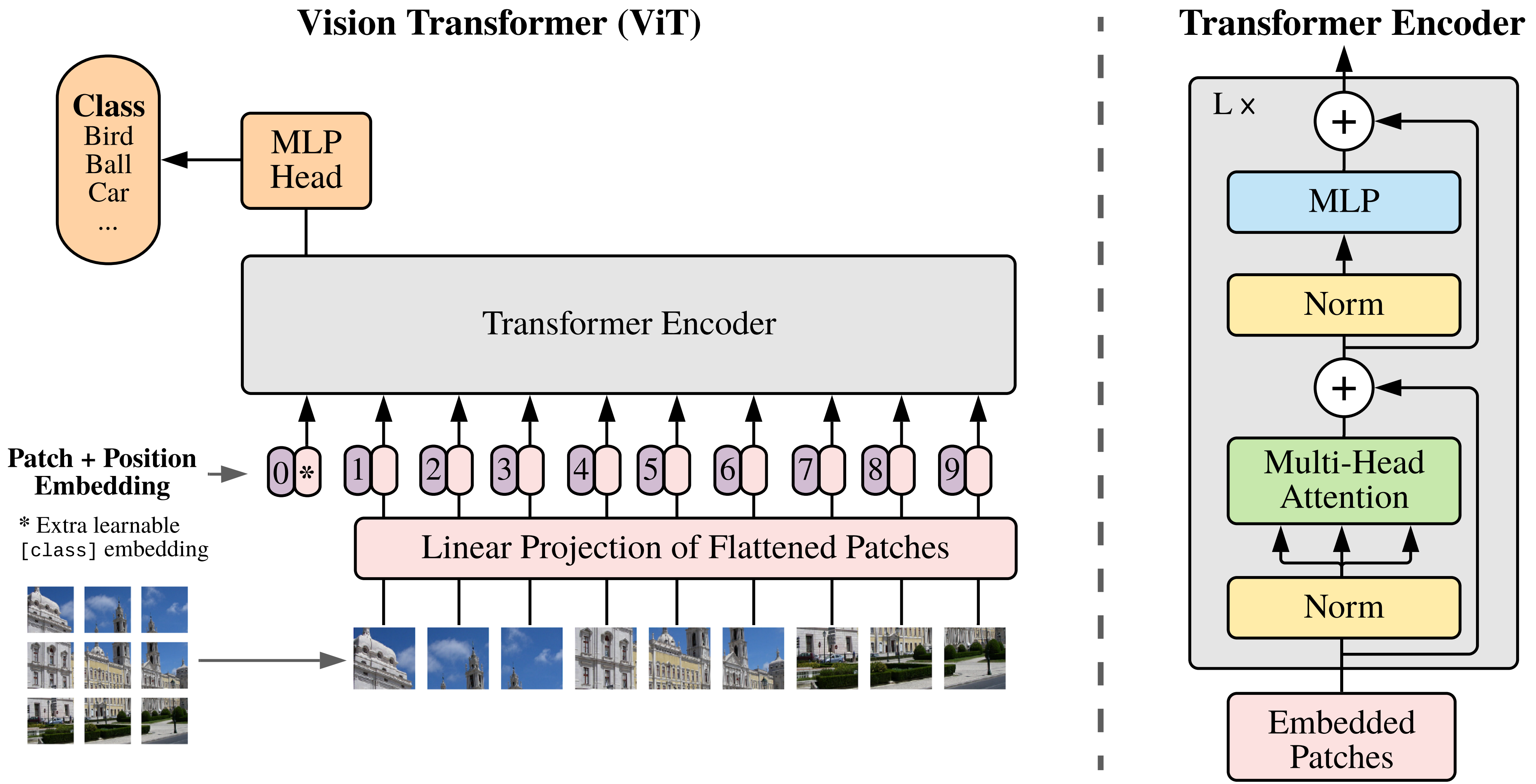
Krizhevsky et al. (2012)

Convolutional Neural Networks (CNNs)

- Local connectivity
- Parameter efficiency
- Translation invariance

- Handling of long-range dependencies

Vision Transformers (ViTs)



Dosovitskiy et al. (2020)

Vision Transformers (ViTs)

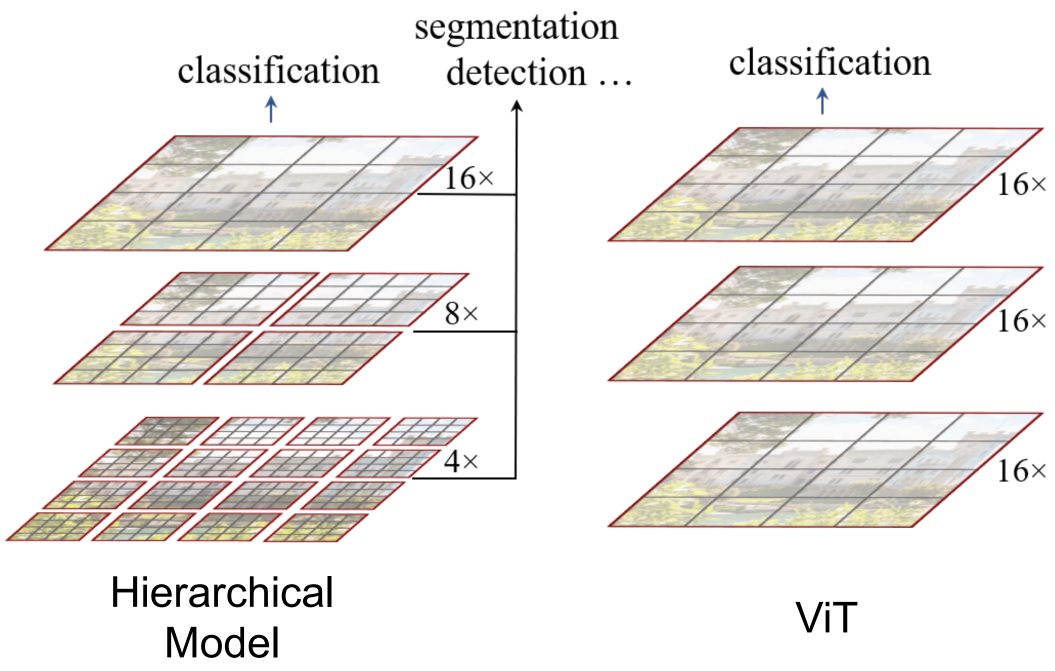
- Good accuracy
- Simple
- Cost of simplicity:
  - Inefficient use of parameters
  - Constant resolution and channel capacity

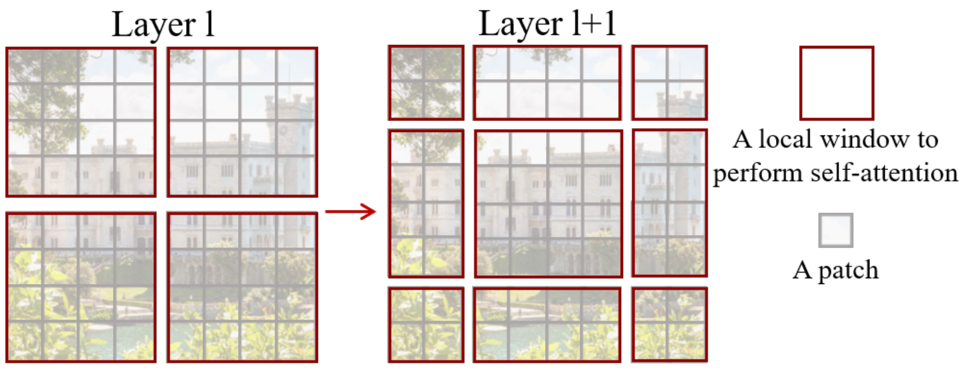
Hierarchical Models

- Start from higher resolution and simple features
- End at lower resolution and complex features

Examples:

- Swin (Shifted-Windows)





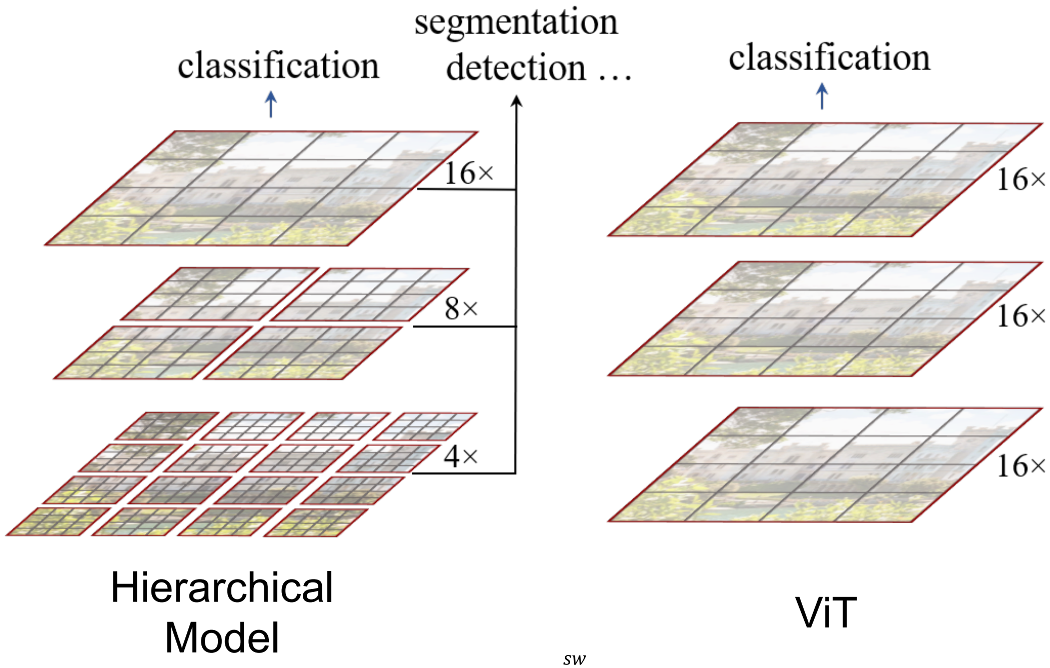
Liu et al. (2021)

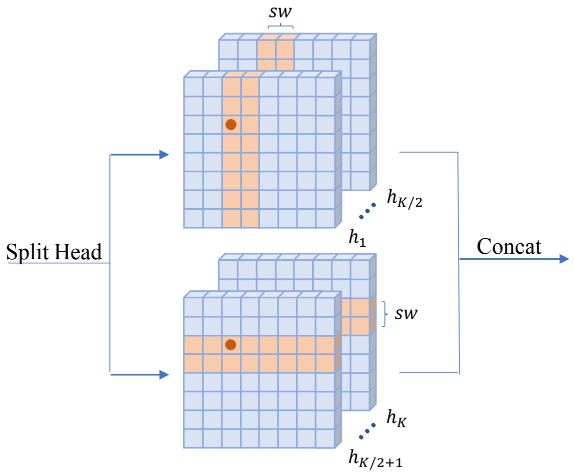
Hierarchical Models

- Start from higher resolution and simple features
- End at lower resolution and complex features

Examples:

- Swin (Shifted-Windows)
- CSWin (Cross-Shaped Windows)





Dynaic Stripe Window + Parallel Grouing Heads = CSWin

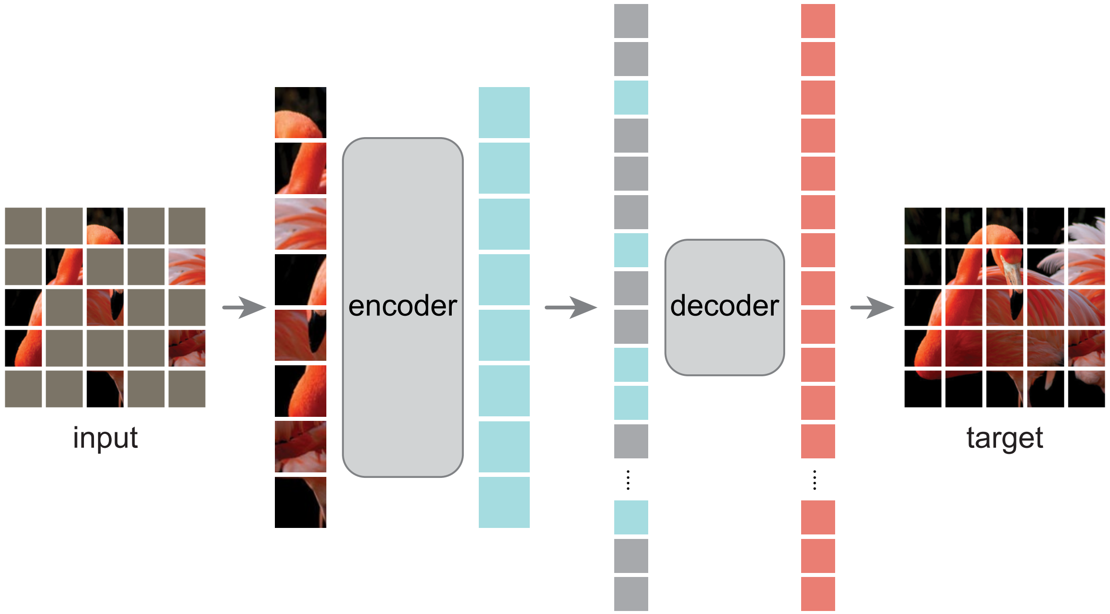
Dong et al. (2022)

Hierarchical Models

- More attractive FLOP counts compared to ViTs
- More inductive bias
- Slower

Hypothesis

- ViT lacks inductive bias
- Possible to **learn spatial bias** instead of manually adding it back
- Use **Masked Autoencoding (MAE)** as pretraining task



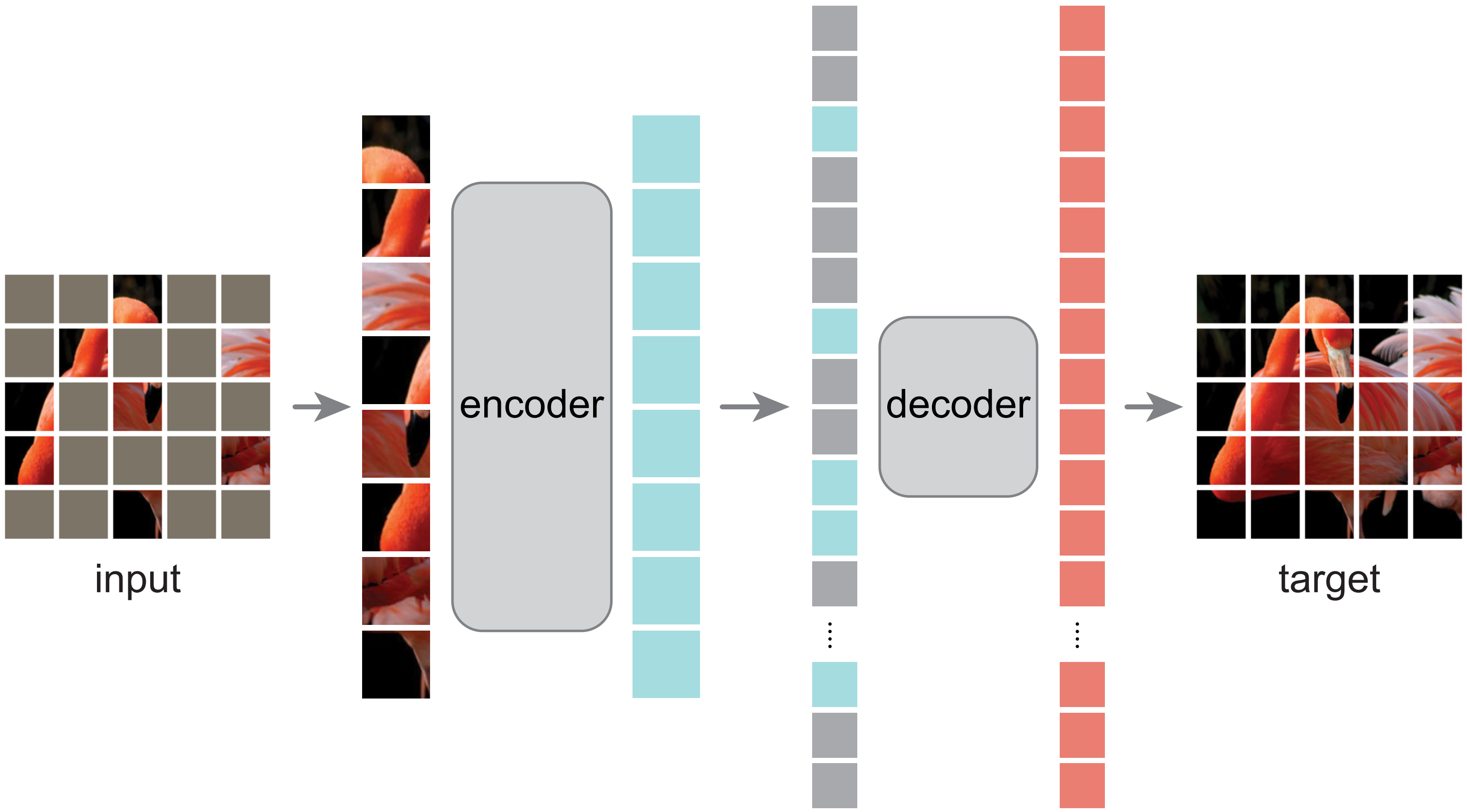
He et al. (2021)

Concepts

Masked Autoencoders (MAE)

He et al. (2021)

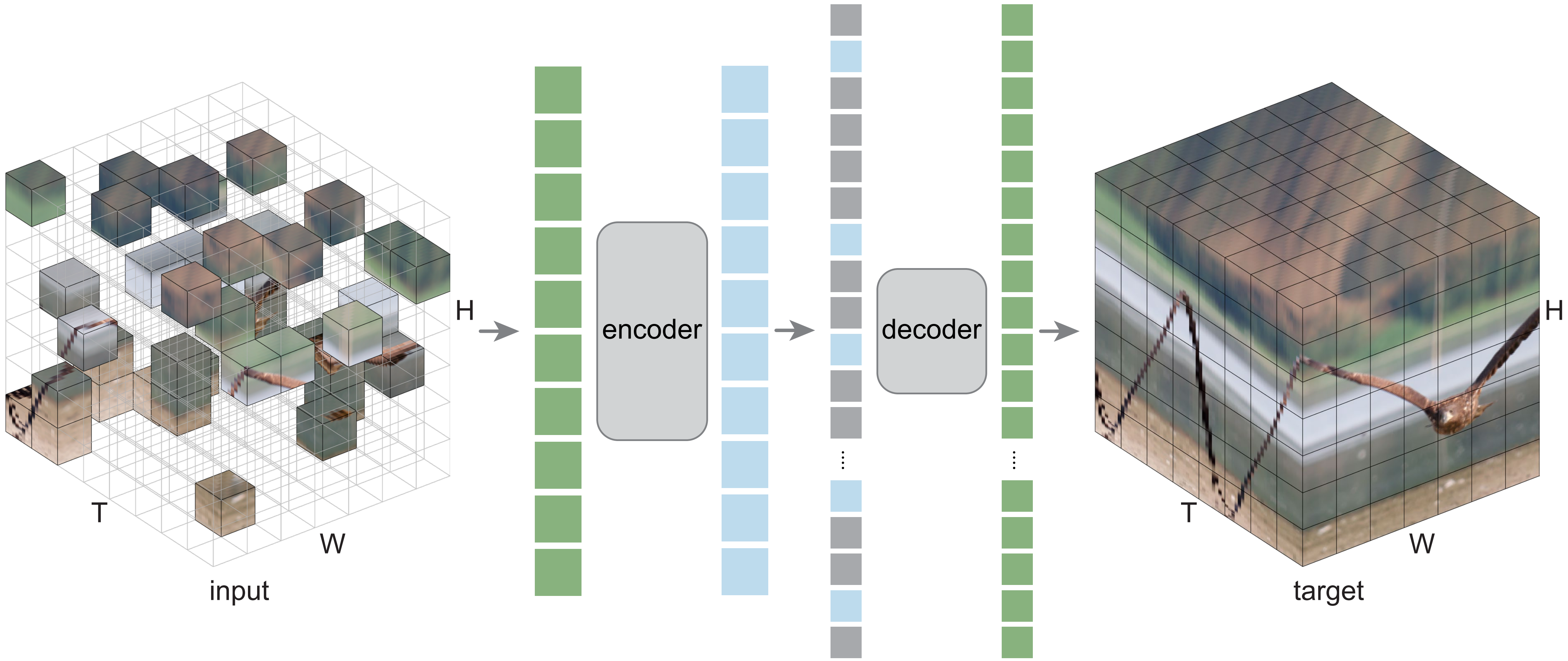
Masked Autoencoders (MAE)



He et al. (2021)

MAE on Video

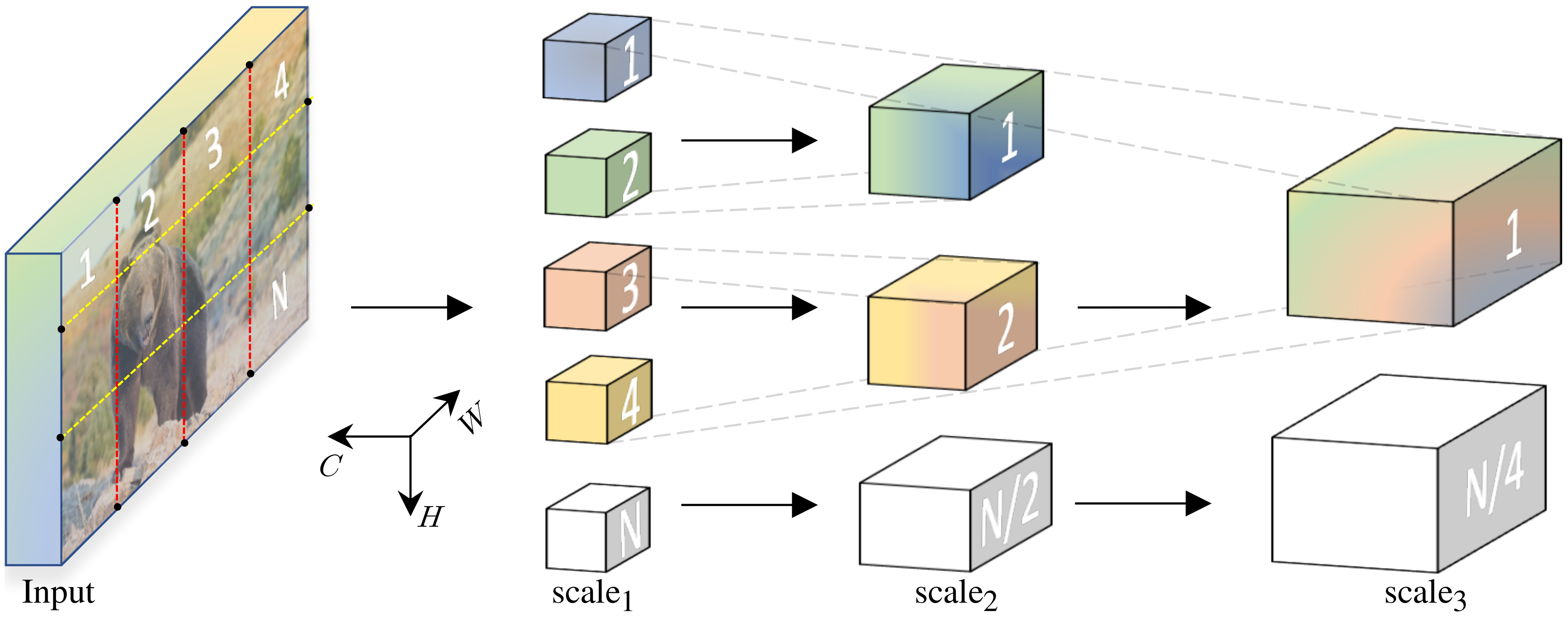
Previous Work



Feichtenhofer et al. (2022)

Multiscale Vision Transformers (MViT)

Previous work

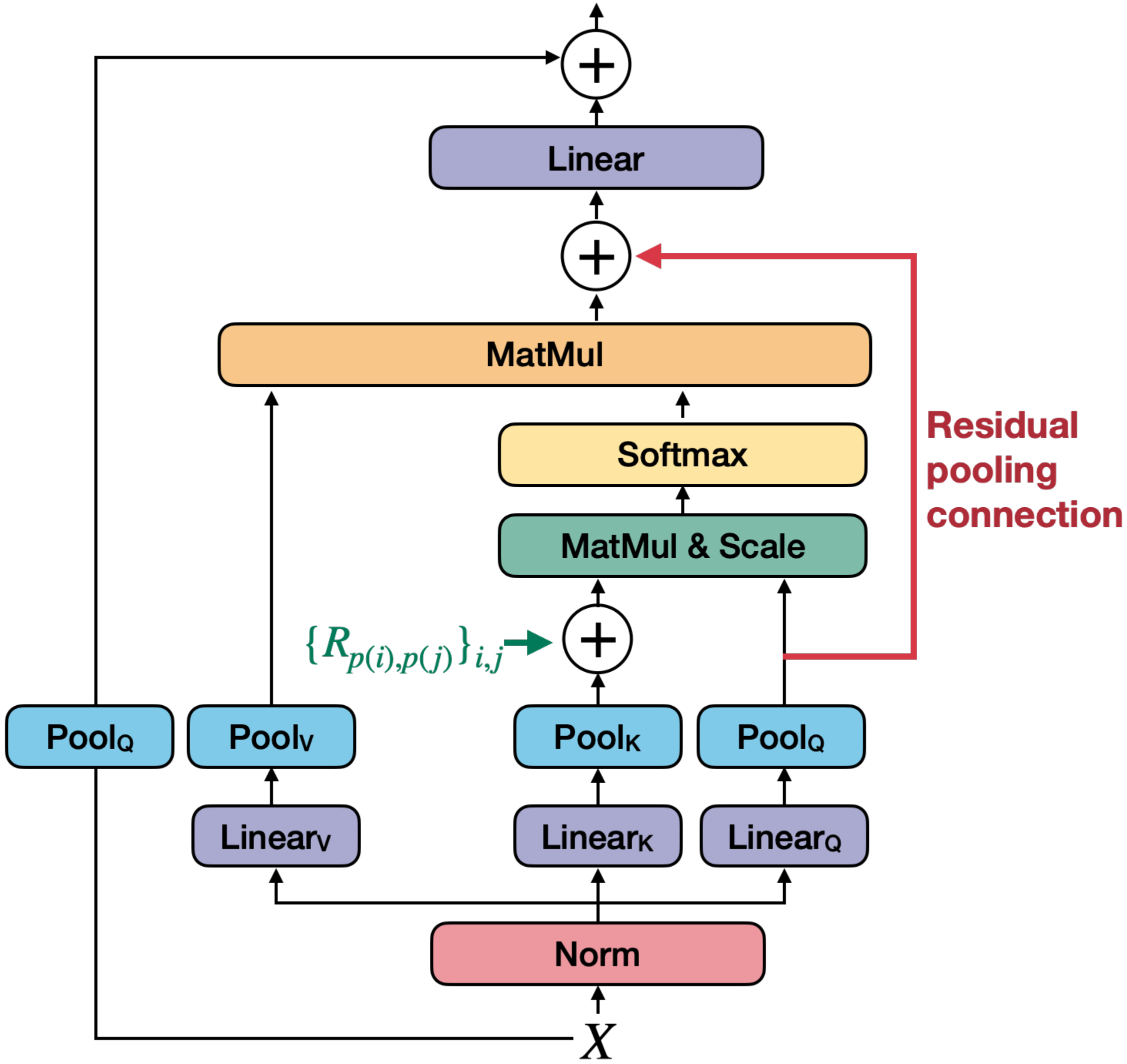


Fewer Channels
Higher Resolution
Simple Features

More Channels
Lower Resolution
Complex Features

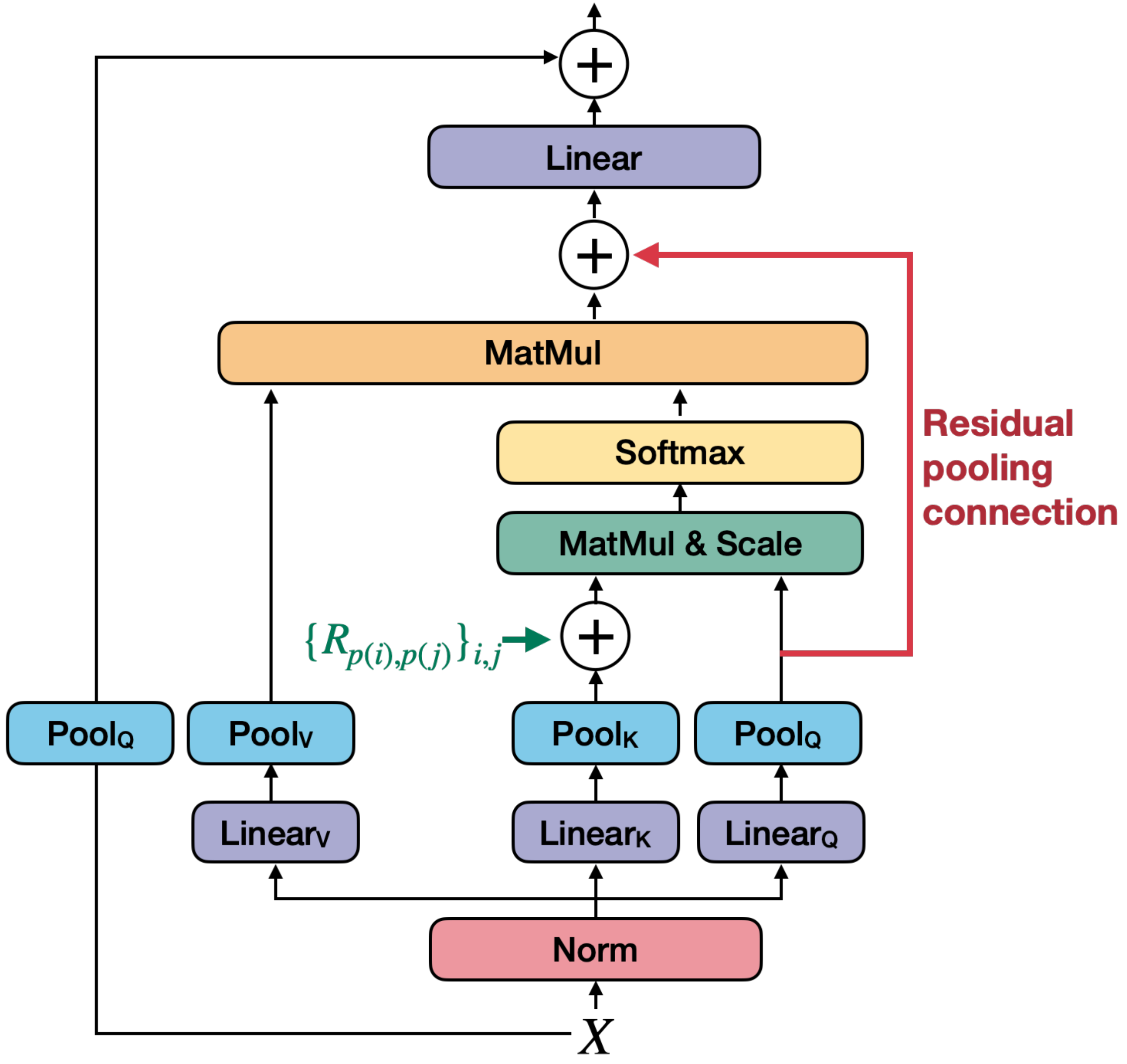
Fan et al. (2021)

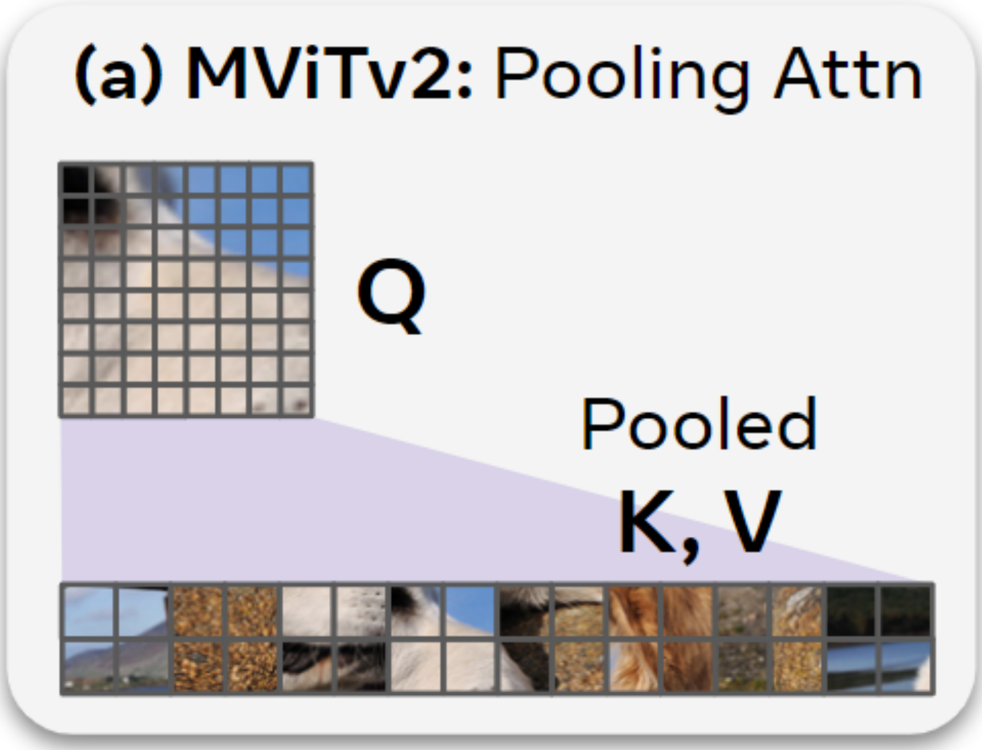
Multiscale Vision Transformers (MViT)



Li et al. (2021)

Multiscale Vision Transformers (MViT)

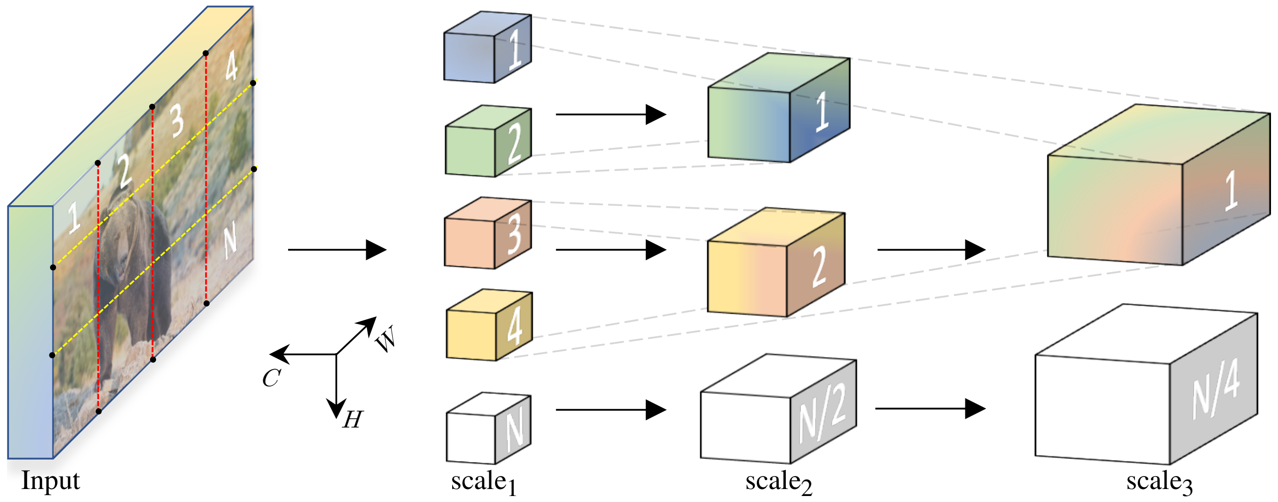




Li et al. (2021), Ryali et al. (2023)

Approach

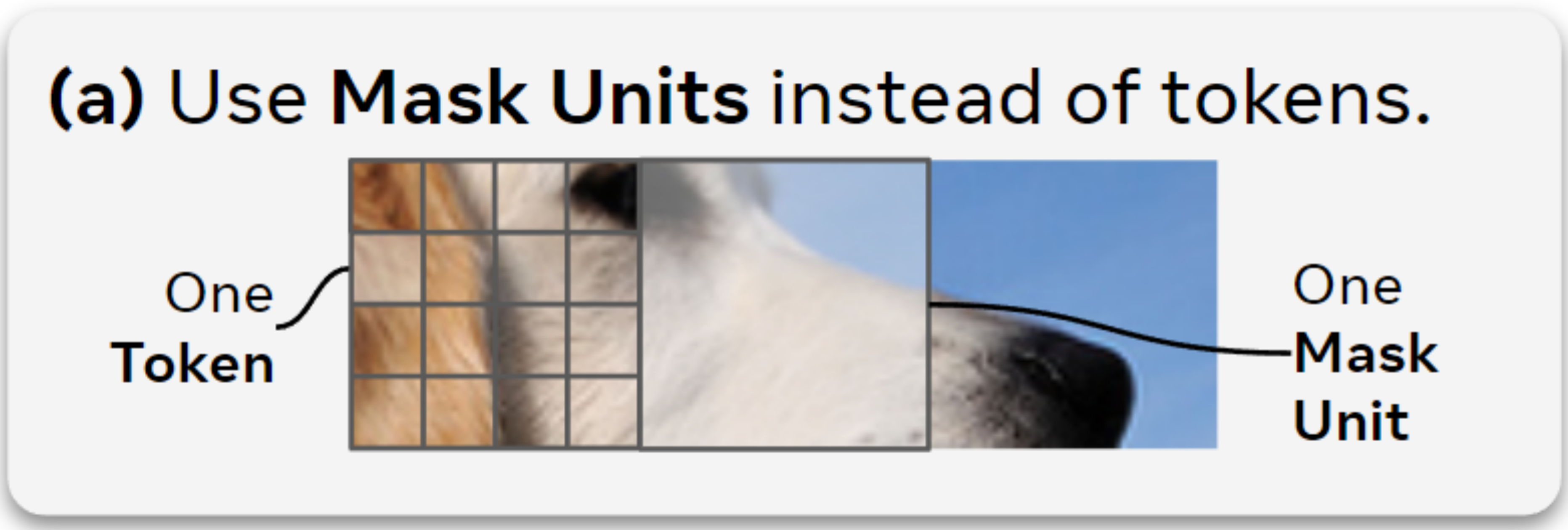
- Take **existing** hierarchical ViT (MViTv2)
- **Remove** non-essential components
- **Supply spatial bias** through pretraining with MAE



Fan et al. (2021)

MAE for Hierarchical Models

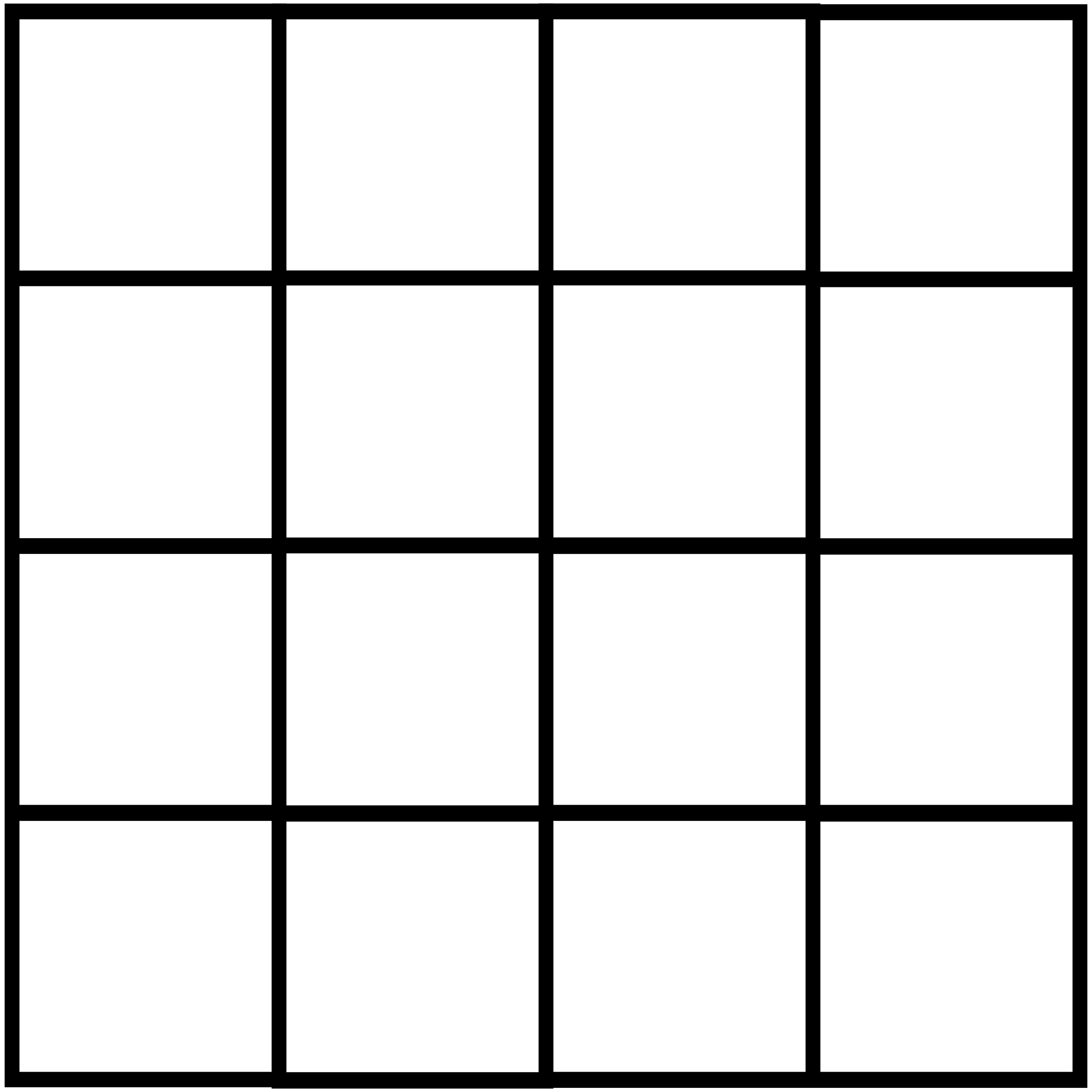
Adapting MAE



Ryali et al. (2023)

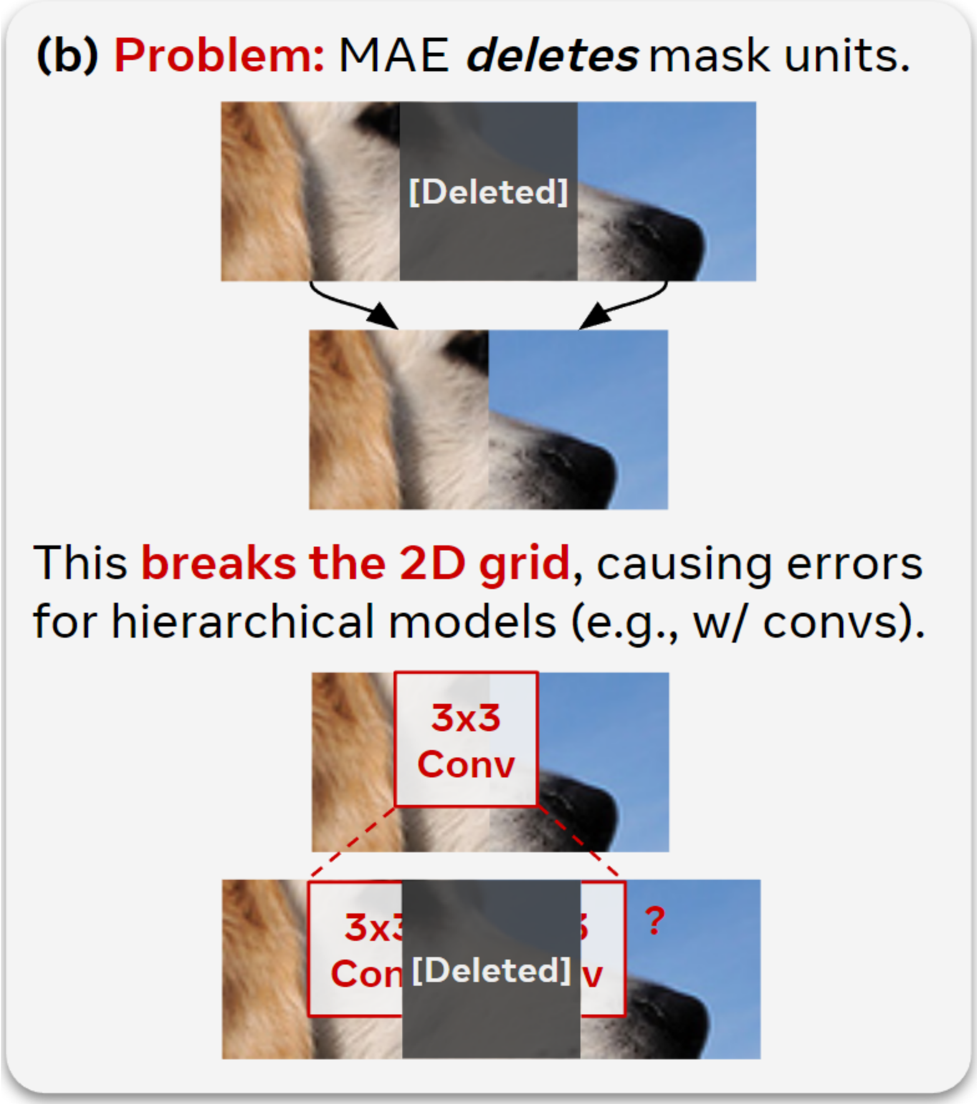
Adapting MAE

- MViTv2 downsamples by 2 × 2 three times
- Token size 4 × 4
- Mask unit size 32 × 32
- 8^2, 4^2, 2^2, 1^2 tokens in stages 1, 2, 3, 4 respectively



Ryali et al. (2023)

Adapting MAE

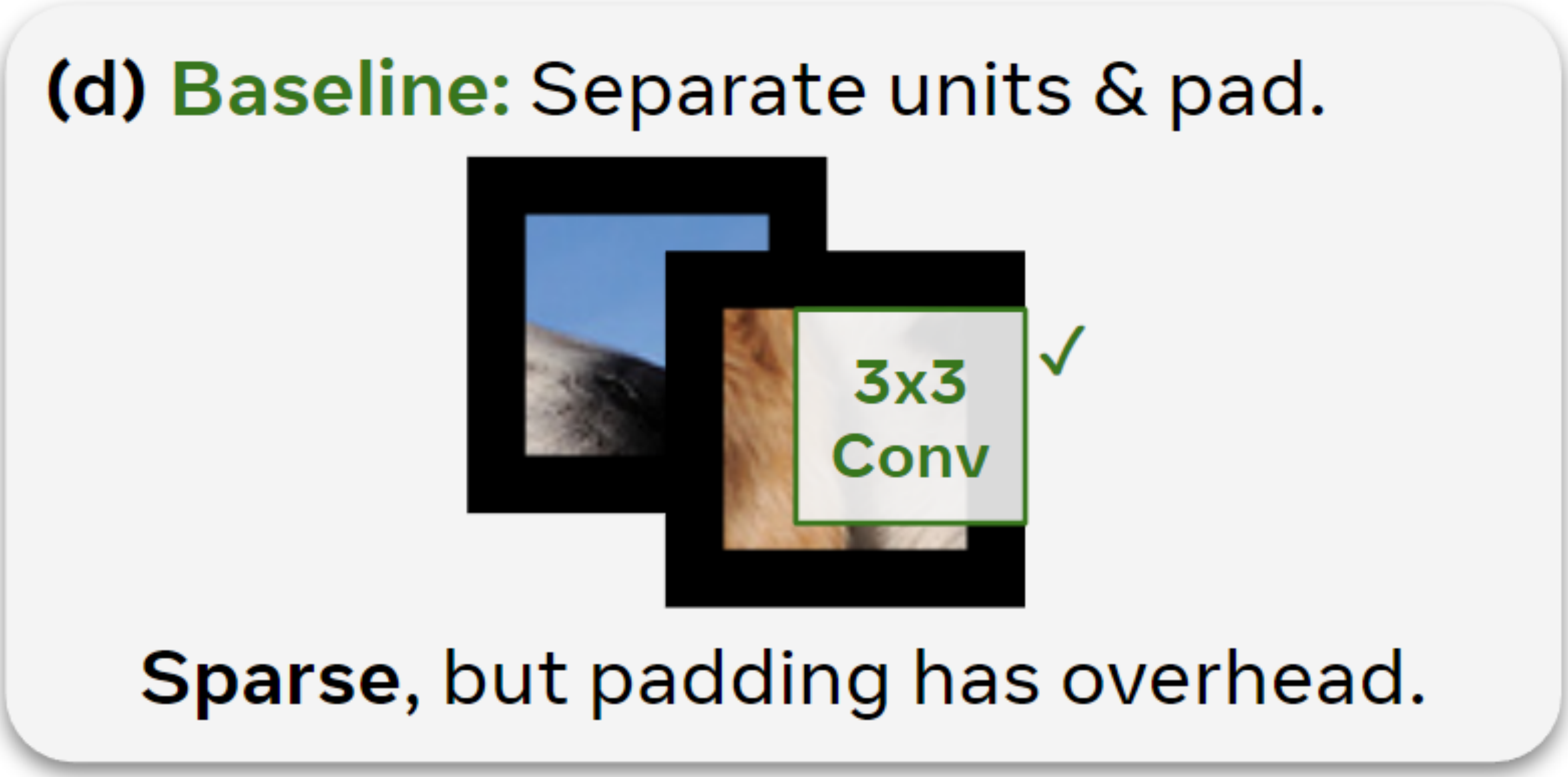


Ryali et al. (2023)

Adapting MAE

Ryali et al. (2023)

Adapting MAE



Ryali et al. (2023)

Adapting MAE

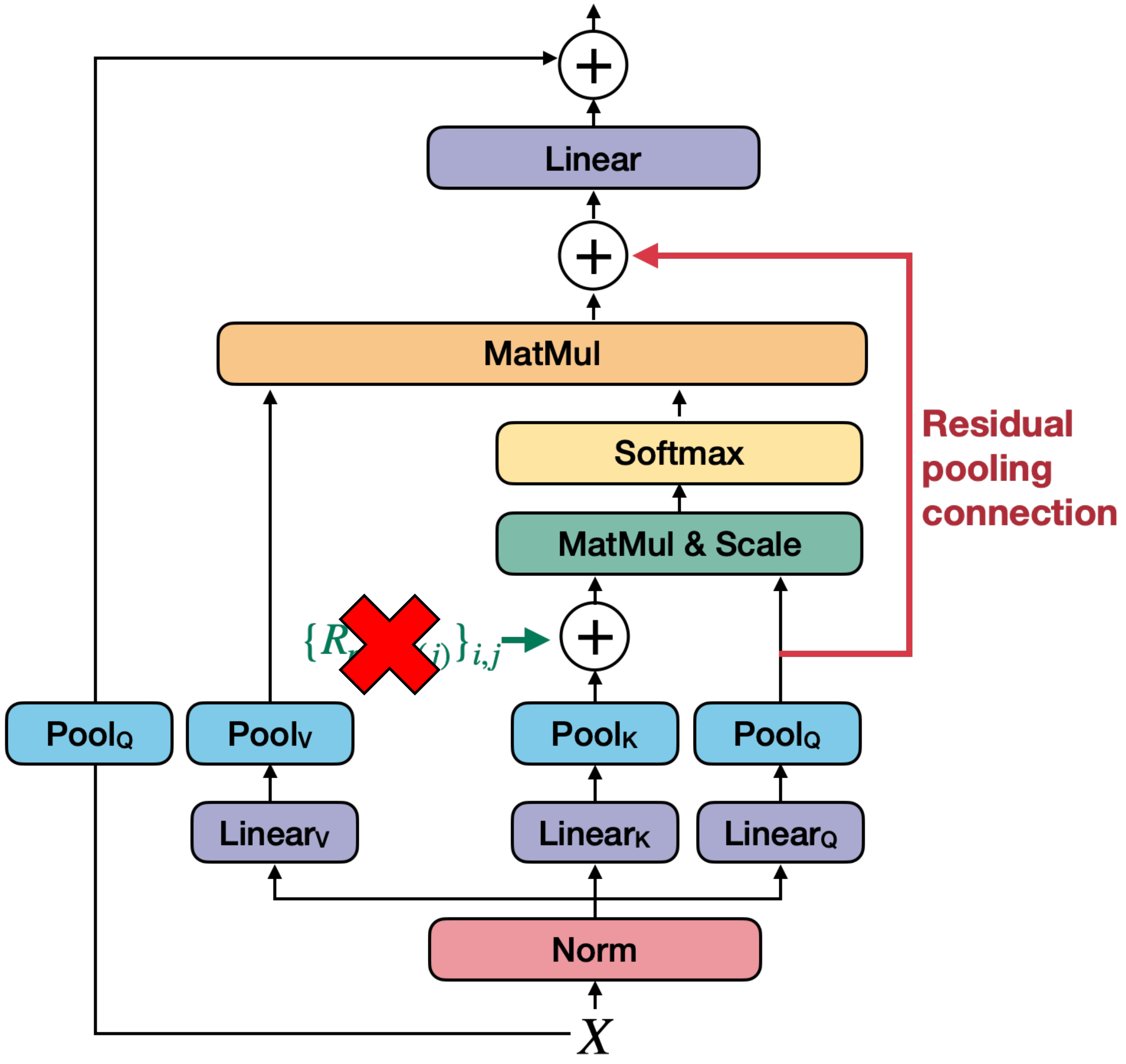
(e) Hiera: Just set *kernel size* = *stride*.

Sparse, no overhead, simple.

Ryali et al. (2023)

Removing non-essential components

Multiscale Vision Transformers (MViT)

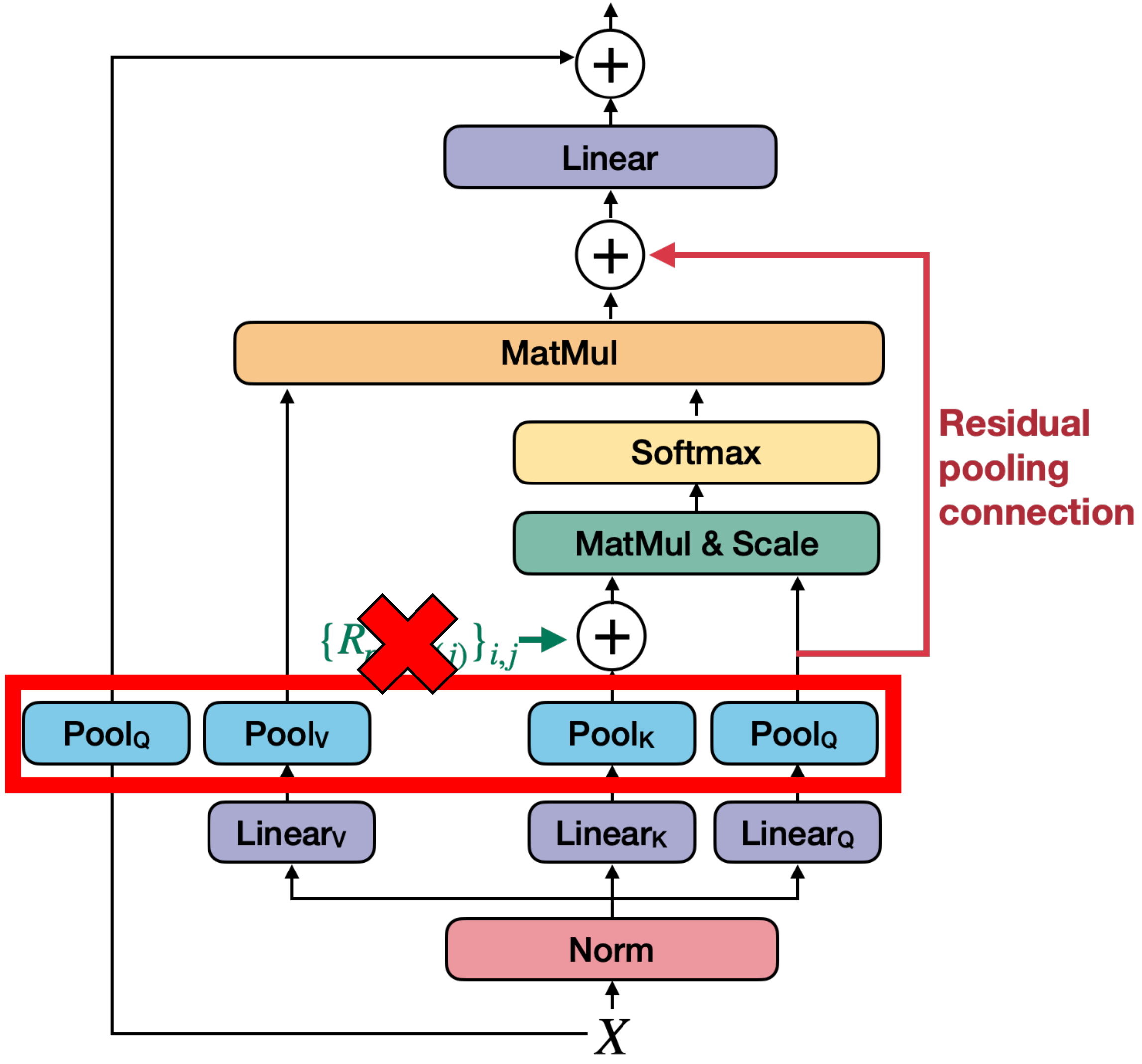


Li et al. (2021)

Relative Position Embeddings

Setting	Image acc.	Image im/s	Video acc.	Video clip/s
MViTv2-L Supervised	85.3	219.8	80.5	20.5
Hiera-L MAE				
a. replace rel pos with absolute *	<u>85.6</u>	253.3	<u>85.3</u>	20.7

Multiscale Vision Transformers (MViT)



- Previously: **Conv**
- Now: Maxpool

Li et al. (2021)

Relative Position Embeddings

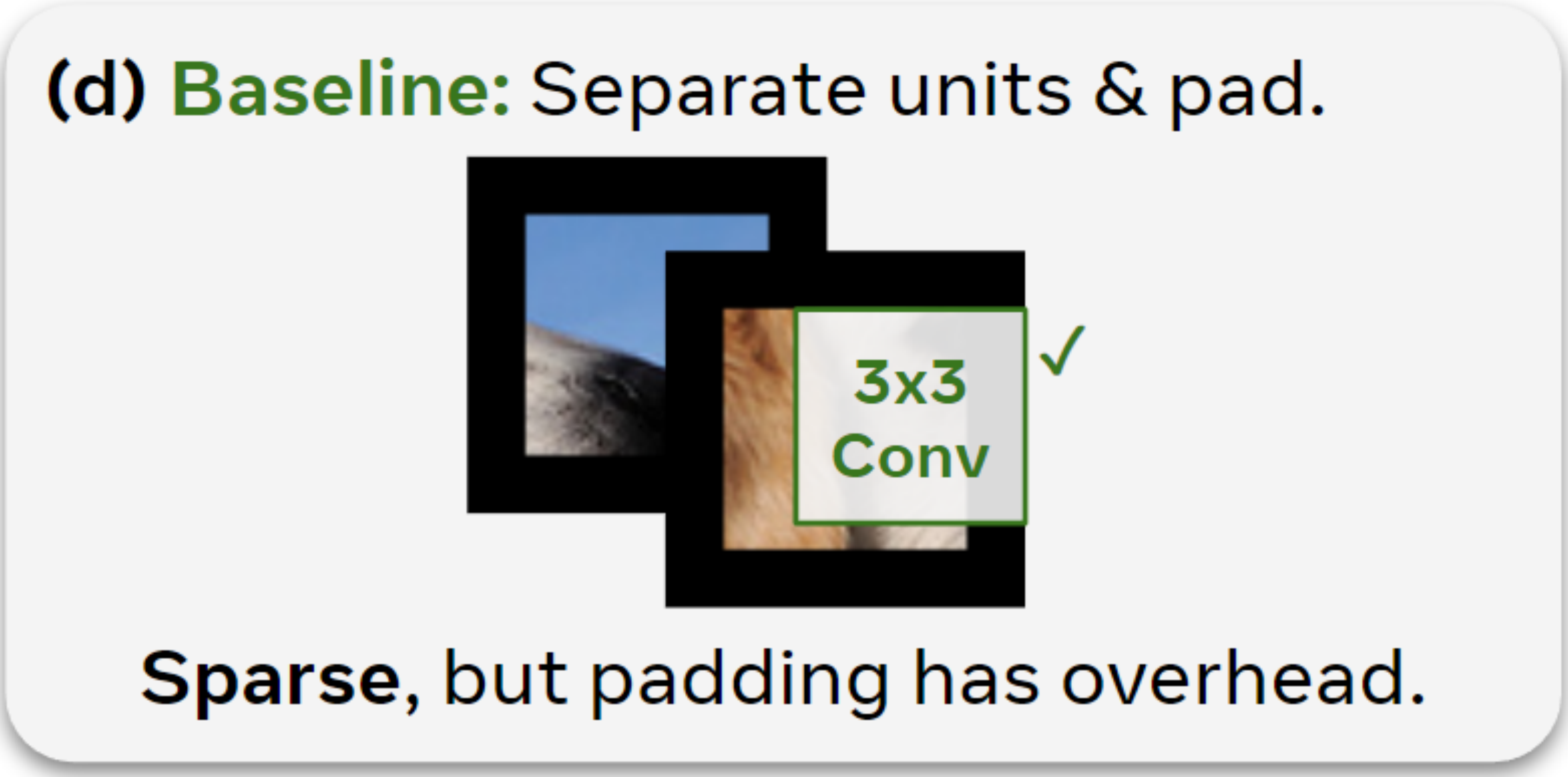
Setting	Image acc.	Image im/s	Video acc.	Video clip/s
MViTv2-L Supervised	85.3	219.8	80.5	20.5
Hiera-L MAE				
a. replace rel pos with absolute *	<u>85.6</u>	253.3	<u>85.3</u>	20.7
b. replace convs with maxpools *	84.4	99.9†	84.1	10.4†

Remove Convolutions

Setting	Image acc.	Image im/s	Video acc.	Video clip/s
MViTv2-L Supervised	85.3	219.8	80.5	20.5
Hiera-L MAE				
a. replace rel pos with absolute *	<u>85.6</u>	253.3	<u>85.3</u>	20.7
b. replace convs with maxpools *	84.4	$99.9^\dagger$	84.1	$10.4^\dagger$
c. delete stride=1 maxpools *	85.4	309.2	84.3	26.2

Ryali et al. (2023)

Adapting MAE



Ryali et al. (2023)

Adapting MAE

(e) Hiera: Just set *kernel size* = *stride*.

Sparse, no overhead, simple.

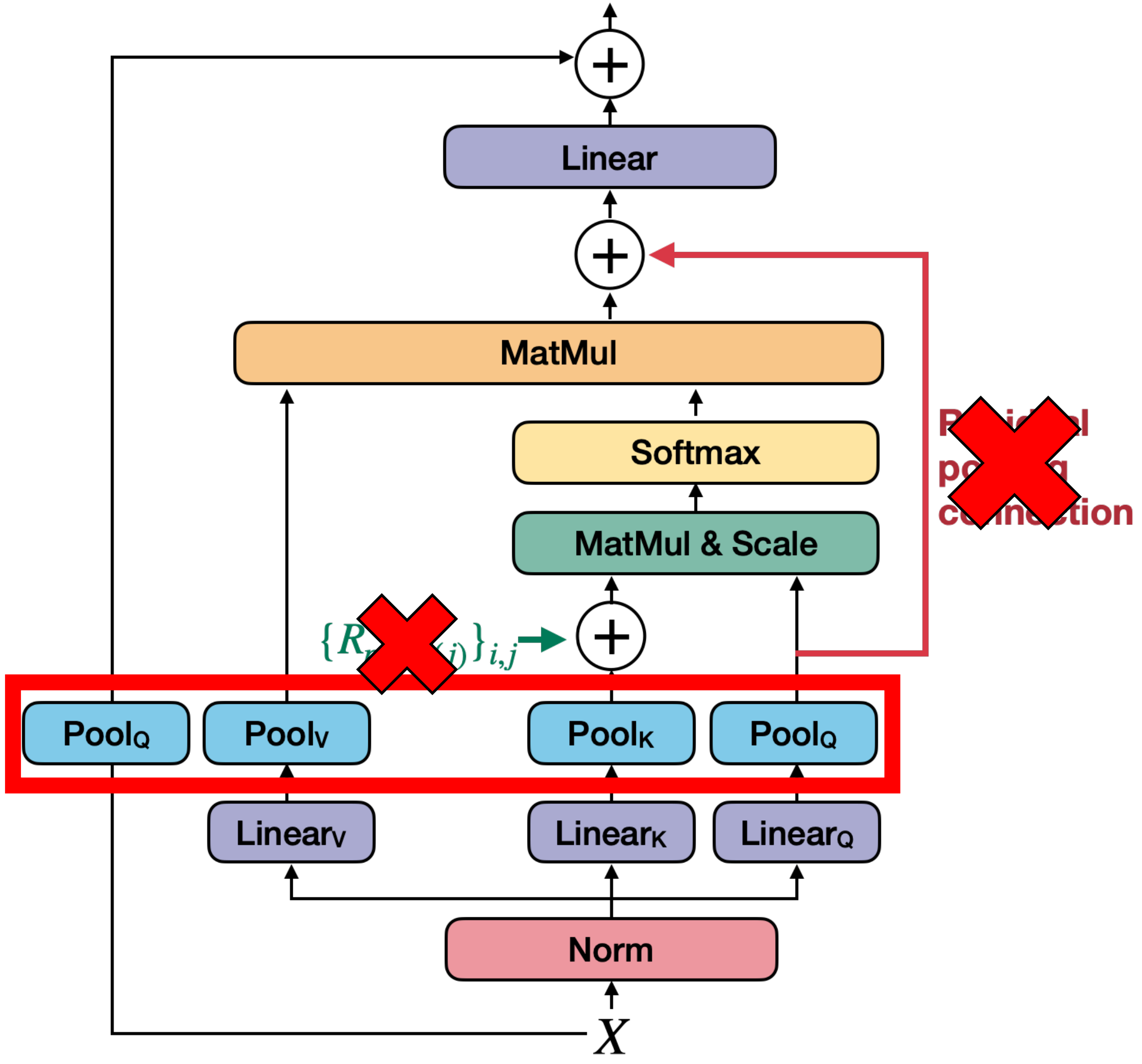
Ryali et al. (2023)

Remove Overlap

Setting	Image acc.	Image im/s	Video acc.	Video clip/s
MViTv2-L Supervised	85.3	219.8	80.5	20.5
Hiera-L MAE				
a. replace rel pos with absolute *	85.6	253.3	85.3	20.7
b. replace convs with maxpools *	84.4	99.9†	84.1	10.4†
c. delete stride=1 maxpools *	85.4	309.2	84.3	26.2
d. set kernel size equal to stride	**85.7**	369.8	**85.5**	29.4

Ryali et al. (2023)

Multiscale Vision Transformers (MViT)



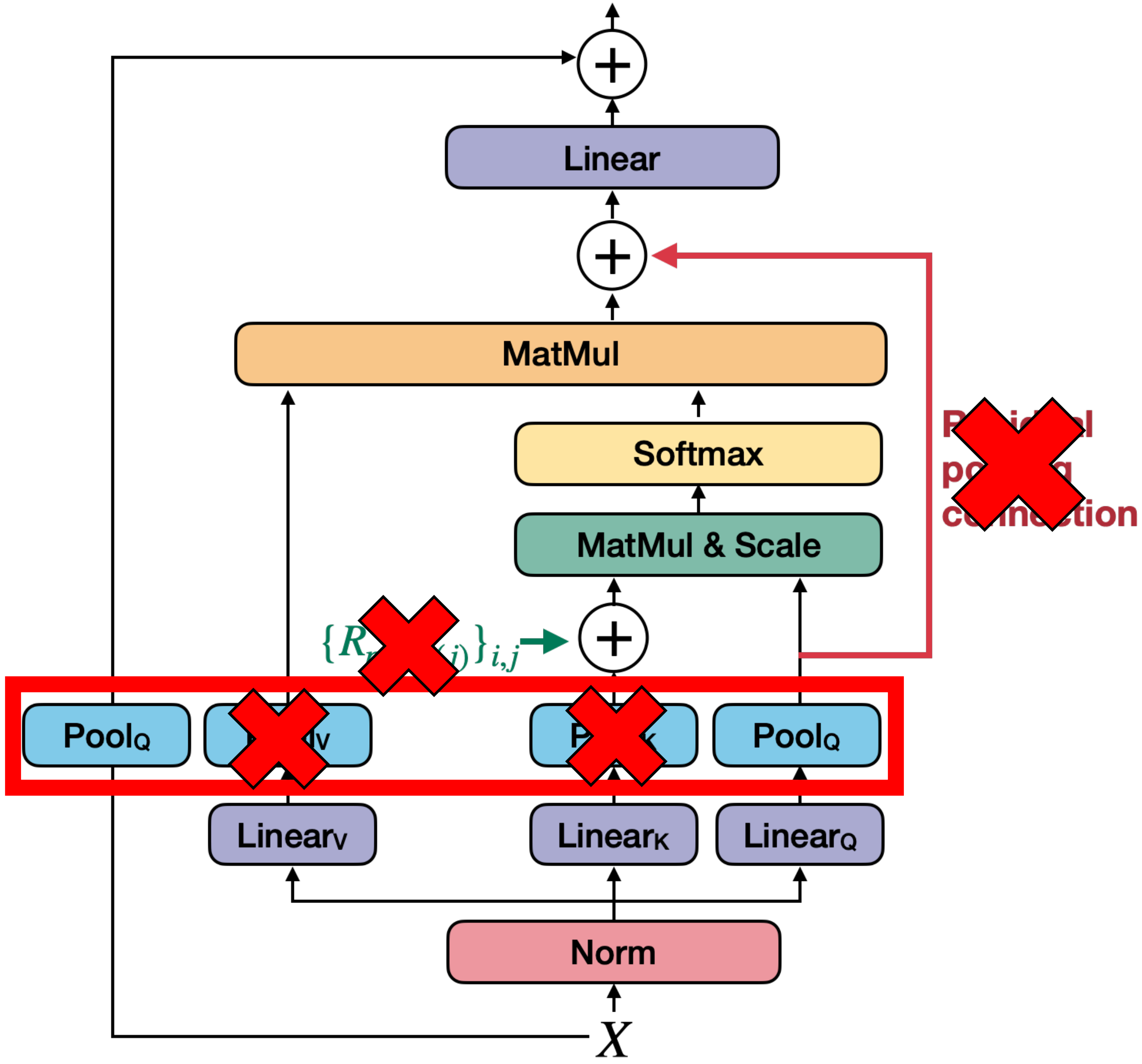
Li et al. (2021)

Remove Attention Residual

Setting	Image acc.	Image im/s	Video acc.	Video clip/s
MViTv2-L Supervised	85.3	219.8	80.5	20.5
Hiera-L MAE				
a. replace rel pos with absolute *	85.6	253.3	85.3	20.7
b. replace convs with maxpools *	84.4	99.9†	84.1	10.4†
c. delete stride=1 maxpools *	85.4	309.2	84.3	26.2
d. set kernel size equal to stride	**85.7**	369.8	**85.5**	29.4
e. delete q attention residuals	85.6	374.3	**85.5**	29.8

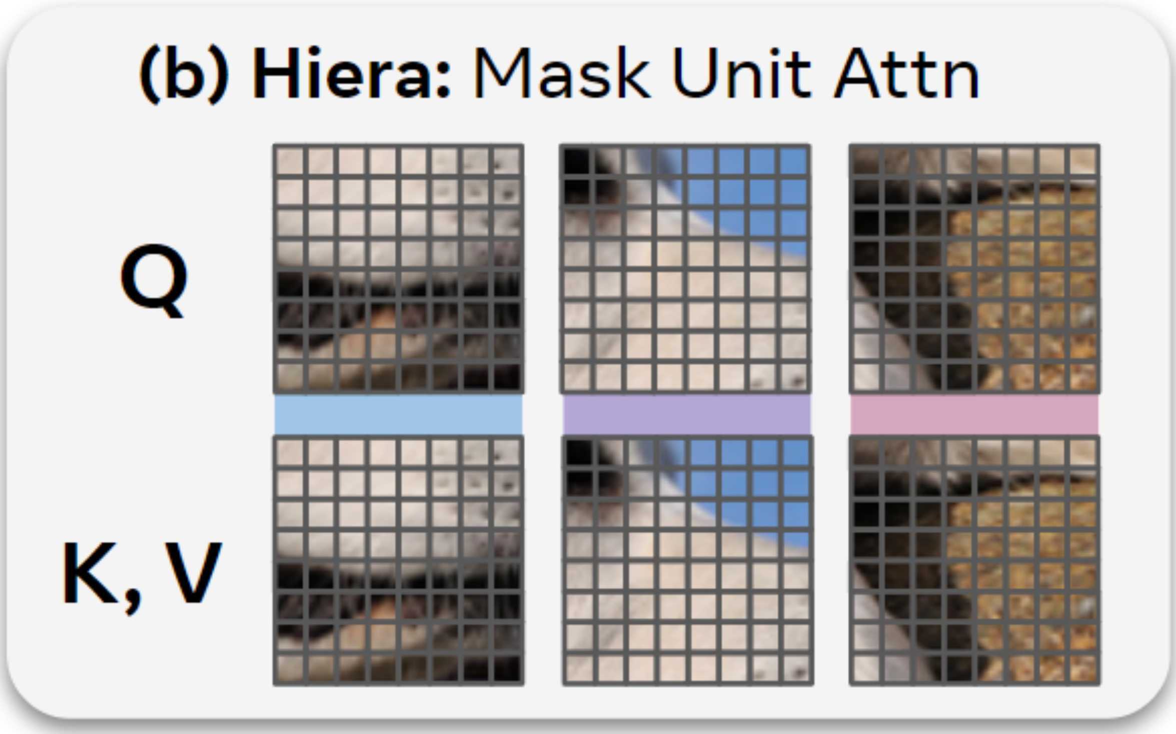
Ryali et al. (2023)

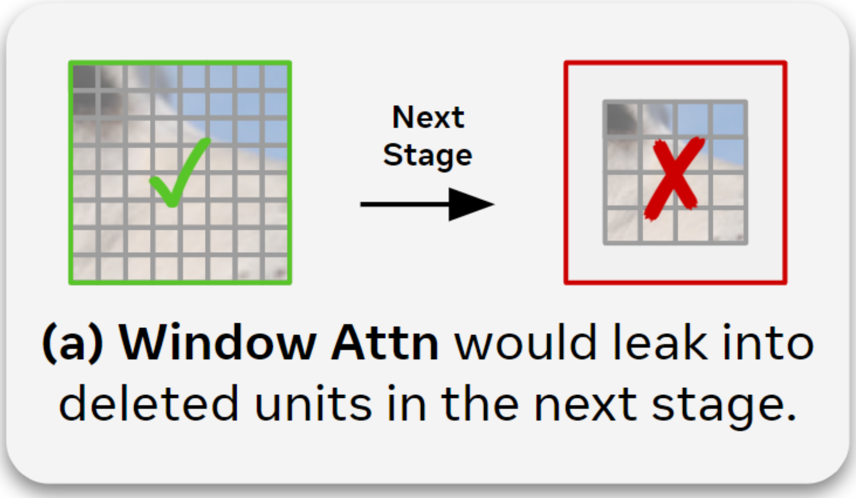
Multiscale Vision Transformers (MViT)

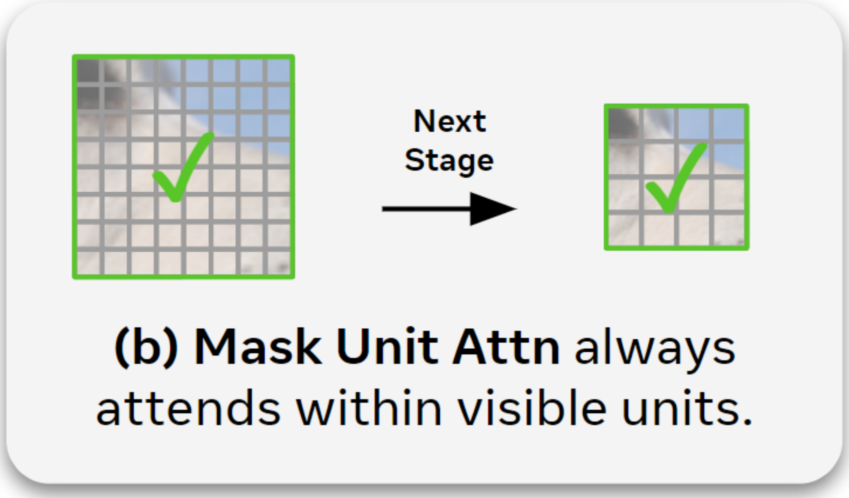


Li et al. (2021)

Mask Unit Attention







Ryali et al. (2023)

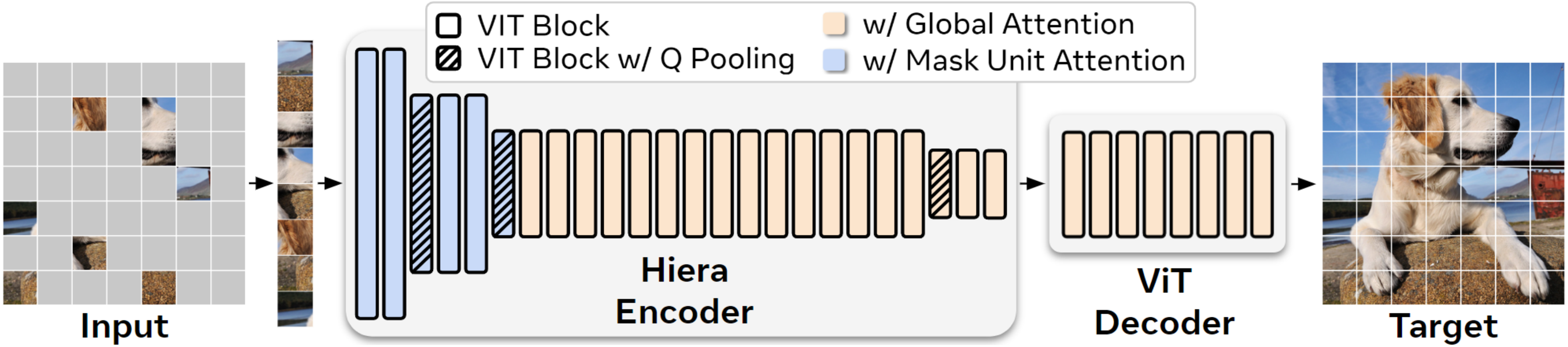
Mask Unit Attention

Setting	Image acc.	Image im/s	Video acc.	Video clip/s
MViTv2-L Supervised	85.3	219.8	80.5	20.5
Hiera-L MAE				
a. replace rel pos with absolute *	85.6	253.3	85.3	20.7
b. replace convs with maxpools *	84.4	99.9†	84.1	10.4†
c. delete stride=1 maxpools *	85.4	309.2	84.3	26.2
d. set kernel size equal to stride	**85.7**	369.8	**85.5**	29.4
e. delete q attention residuals	85.6	374.3	**85.5**	29.8
f. replace kv pooling with MU attn	85.6	**531.4**	**85.5**	**40.8**

Ryali et al. (2023)

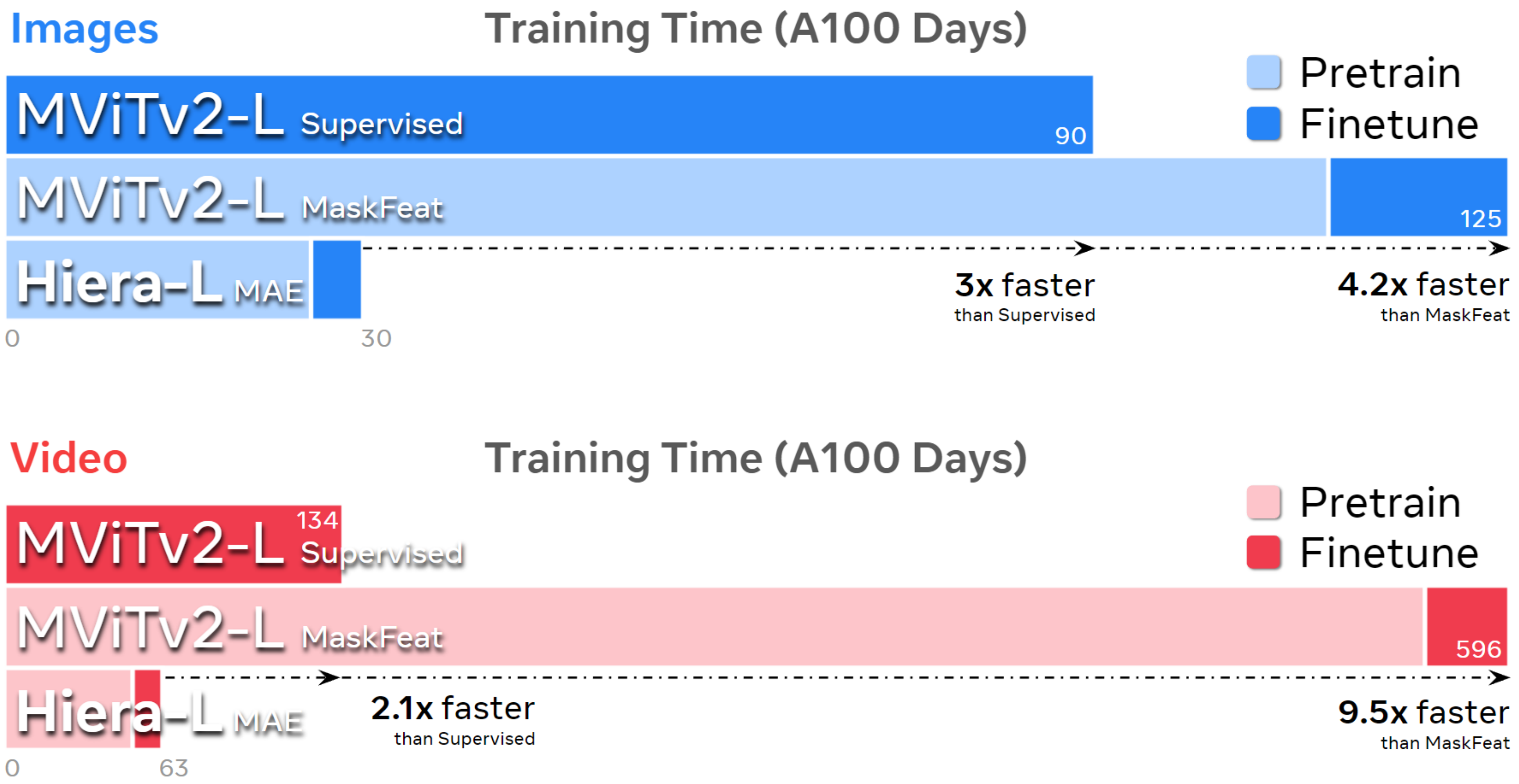
Hiera

Setup



Ryali et al. (2023)

Training Time



Ryali et al. (2023) [8]

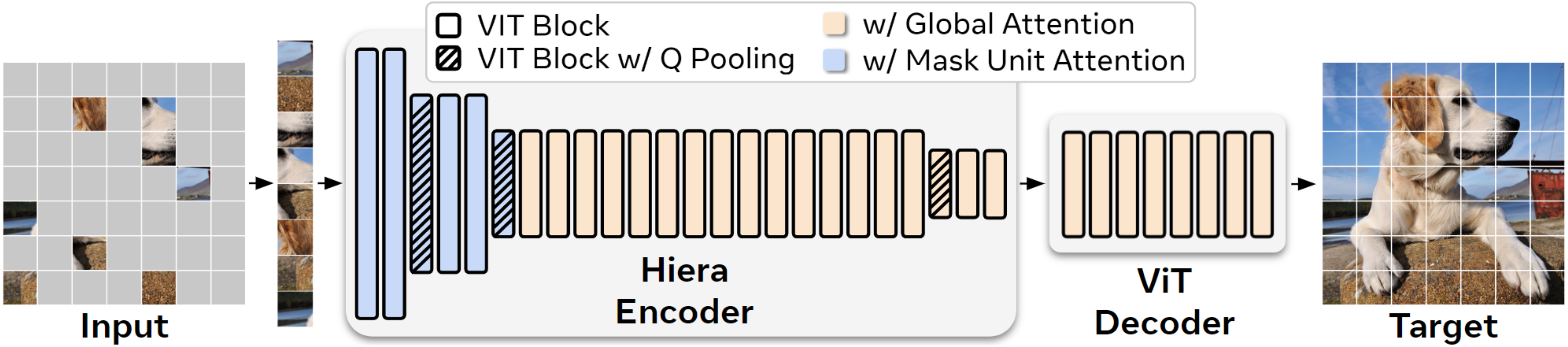
MAE Ablations

Multi-Scale Decoder

multi-scale	image	video
✗	85.0	83.8
✔	**85.6**	**85.5**

Ryali et al. (2023) [8]

Setup



Ryali et al. (2023)

Multi-Scale Decoder

multi-scale	image	video
✗	85.0	83.8
✔	**85.6**	**85.5**

Ryali et al. (2023) [8]

Mask ratio

mask	image	mask	video
0.5	85.5	0.75	84.9
0.6	**85.6**	0.9	**85.5**
0.7	85.3	0.95	84.4

Ryali et al. (2023) [8]

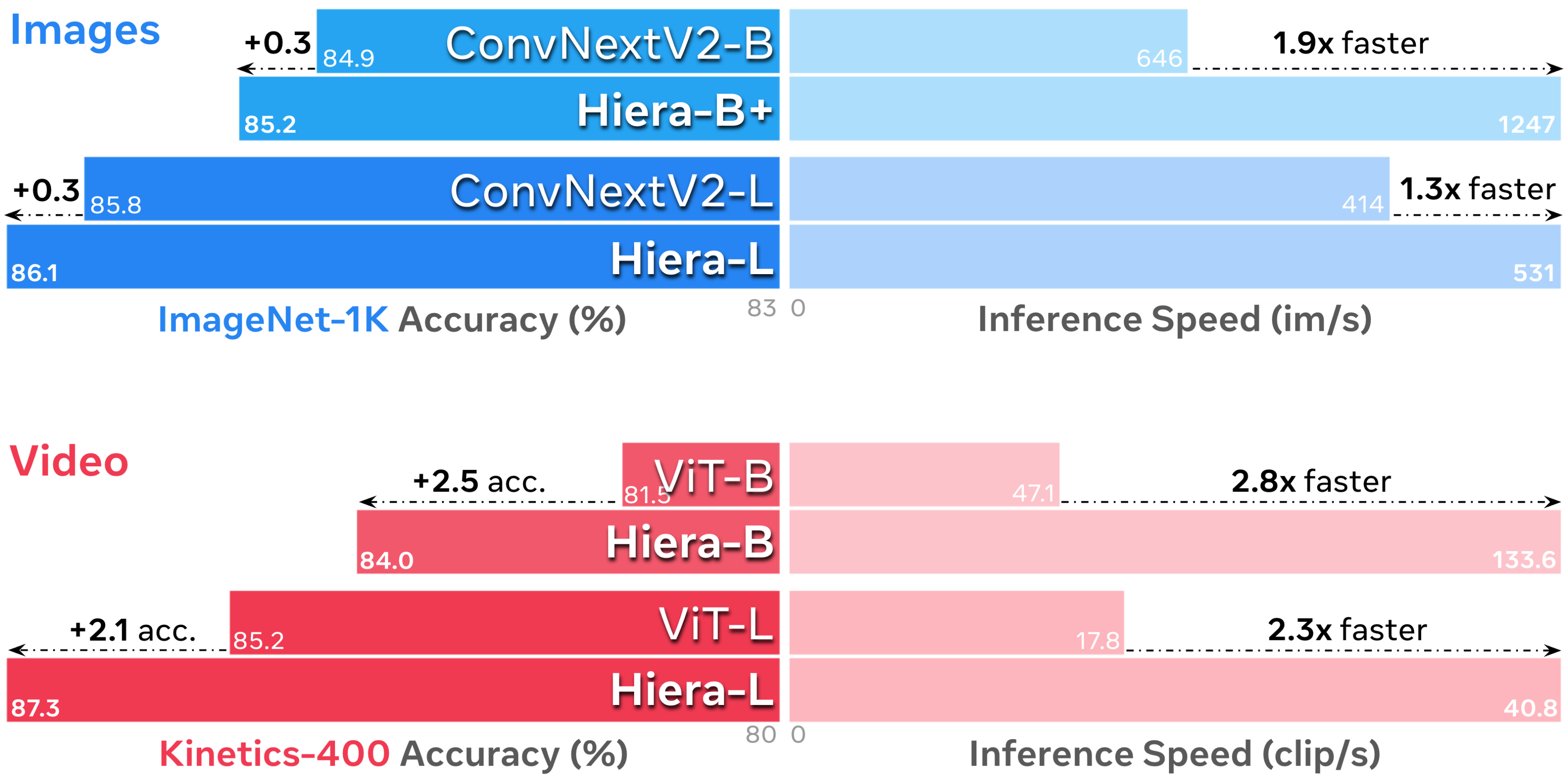
Pretraining schedule

epochs	image	video
400	85.6	84.0
800	85.8	85.5
1600	**86.1**	86.4
3200	**86.1**	**87.3**

Ryali et al. (2023) [8]

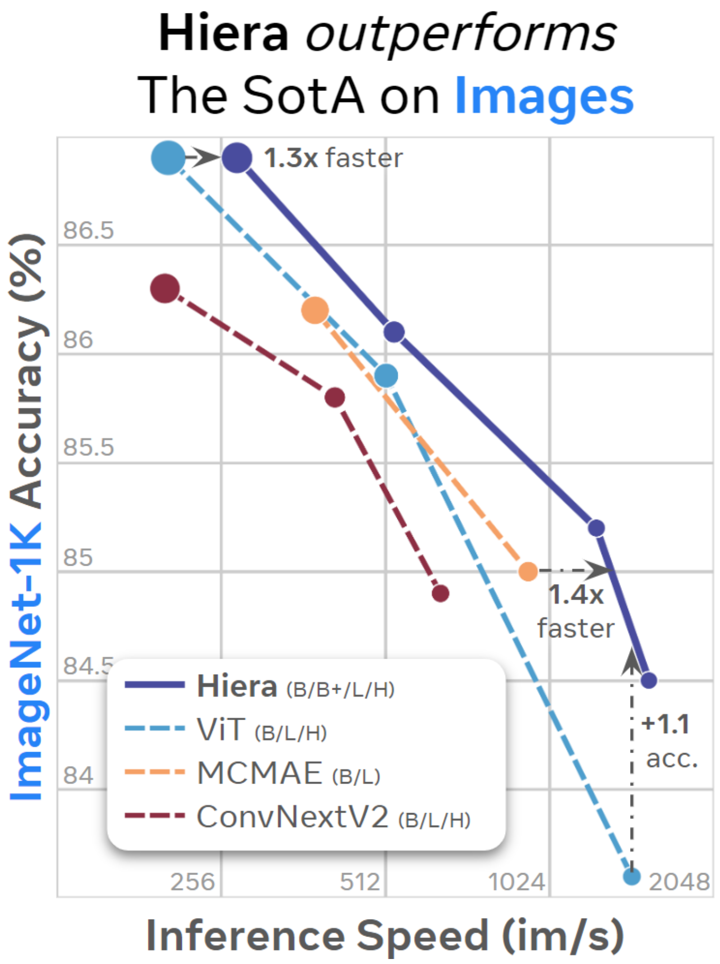
Results

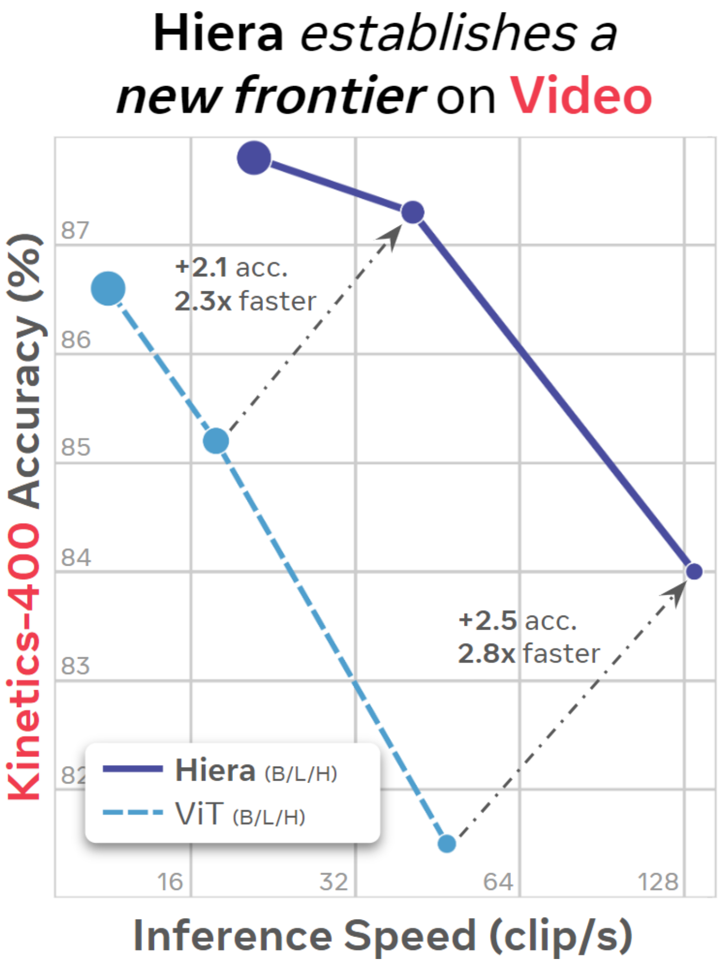
Results



Ryali et al. (2023) [8]

Results





Ryali et al. (2023) [8]

Conclusion

Conclusion

- Showed comprehensively that spatial bias can be learned through strong pretraining task such as MAE

- Either increased throughput or higher accuracy for similar model size (or both)

- Testing on even larger datasets might still be interesting

Questions?

Variants

model	#Channels	#Blocks	#Heads	FLOPs	Param
Hiera-T	[96-192-384-768]	[1-2-7-2]	[1-2-4-8]	5G	28M
Hiera-S	[96-192-384-768]	[1-2-11-2]	[1-2-4-8]	6G	35M
Hiera-B	[96-192-384-768]	[2-3-16-3]	[1-2-4-8]	9G	52M
Hiera-B+	[112-224-448-896]	[2-3-16-3]	[2-4-8-16]	13G	70M
Hiera-L	[144-288-576-1152]	[2-6-36-4]	[2-4-8-16]	40G	214M
Hiera-H	[256-512-1024-2048]	[2-6-36-4]	[4-8-16-32]	125G	673M

Ryali et al. (2023)

ImageNet-1K

backbone	pretrain	acc.	FLOPs (G)	Param
Swin-T		81.3	**5**	29M
MViTv2-T		82.3	**5**	**24M**
Hiera-T	MAE	**82.8**	**5**	28M
Swin-S		83.0	9	50M
MViTv2-S		83.6	7	**35M**
Hiera-S	MAE	**83.8**	**6**	**35M**
ViT-B		82.3	18	87M
Swin-B		83.3	15	88M
MViTv2-B		84.4	10	**52M**
ViT-B	BEiT, DALLE	83.2	18	87M
ViT-B	MAE	83.6	18	87M
ViT-B	MaskFeat	84.0	18	87M
Swin-B	SimMIM	83.8	15	88M
MCMAE-B	MCMAE	85.0	28	88M
Hiera-B	MAE	84.5	**9**	**52M**
Hiera-B+	MAE	**85.2**	13	70M
ViT-L		82.6	62	304M
MViTv2-L		85.3	42	218M
ViT-L	BEiT, DALLE	85.2	62	304M
ViT-L	MAE	85.9	62	304M
ViT-L	MaskFeat	85.7	62	304M
Swin-L	SimMIM	85.4	**36**	**197M**
MCMAE-L	MCMAE	**86.2**	94	323M
Hiera-L	MAE	86.1	40	214M
ViT-H		83.1	167	**632M**
ViT-H	MAE	**86.9**	167	**632M**
Hiera-H	MAE	**86.9**	**125**	673M

Ryali et al. (2023) [8]

Transfer Learning

backbone	iNat17	iNat18	iNat19	Places365
ViT-B	70.5	75.4	80.5	57.9
Hiera-B	73.3	77.9	83.0	58.9
Hiera-B+	**74.7**	**79.9**	**83.1**	**59.2**
ViT-L	75.7	80.1	83.4	59.4
Hiera-L	**76.8**	**80.9**	**84.3**	**59.6**
ViT-H	79.3	83.0	**85.7**	59.8
Hiera-H	**79.6**	**83.5**	**85.7**	**60.0**
ViT-H_{448}	83.4	86.8	88.3	60.3
Hiera-H_{448}	**83.8**	**87.3**	**88.5**	**60.6**

Ryali et al. (2023) [8]

K400

backbone	pretrain	acc.	FLOPs (G)	Param
ViT-B	MAE	81.5	180×3×5	87M
Hiera-B	MAE	84.0	**102**×3×5	**51M**
Hiera-B+	MAE	**85.0**	133×3×5	69M
MViTv2-L	-	80.5	**377**×1×10	218M
MViTv2-L	MaskFeat	84.3	**377**×1×10	218M
ViT-L	MAE	85.2	597×3×5	305M
Hiera-L	MAE	**87.3**	413×3×5	**213M**
ViT-H	MAE	86.6	1192×3×5	**633M**
Hiera-H	MAE	**87.8**	**1159**×3×5	672M

Ryali et al. (2023) [8]

Transferring to Action Detection (AVA v2.2)

backbone	pretrain	mAP	FLOPs (G)	Param
K400 pretrain				
ViT-L	supervised	22.2	598	304M
MViTv2-L$_{40,312}$	MaskFeat	<u>38.5</u>	2828	<u>218M</u>
ViT-L	MAE	37.0	<u>597</u>	305M
Hiera-L	MAE	**39.8**	**413**	**213M**
ViT-H	MAE	39.5	1192	**633M**
Hiera-H	MAE	**42.5**	**1158**	672M
K600 pretrain				
ViT-L	MAE	38.4	<u>598</u>	304M
MViTv2-L$_{40,312}$	MaskFeat	<u>39.8</u>	2828	<u>218M</u>
Hiera-L	MAE	**40.7**	**413**	**213M**
ViT-H	MAE	40.3	1193	**632M**
Hiera-H	MAE	**42.8**	**1158**	672M
K700 pretrain				
ViT-L	MAE	39.5	598	304M
Hiera-L	MAE	**41.7**	**413**	**213M**
ViT-H	MAE	40.1	1193	**632M**
Hiera-H	MAE	**43.3**	**1158**	672M

Ryali et al. (2023) [8]

Reconstruction target

target	image	video
pixel	85.6	85.5
HOG	**85.7**	**86.1**

Ryali et al. (2023) [8]

Drop path rate

dpr	image	video
0.0	85.2	84.5
0.1	85.6	85.4
0.2	**85.6**	**85.5**
0.3	85.5	85.2

Ryali et al. (2023) [8]

Decoder depth

depth	image	video
4	85.5	84.8
8	**85.6**	**85.5**
12	85.5	85.4

Ryali et al. (2023) [8]