Masked Image Modelling for Representation Learning in Earth Observation

Hugo Hernández Hernández

Helmholtz AI Conference 2023

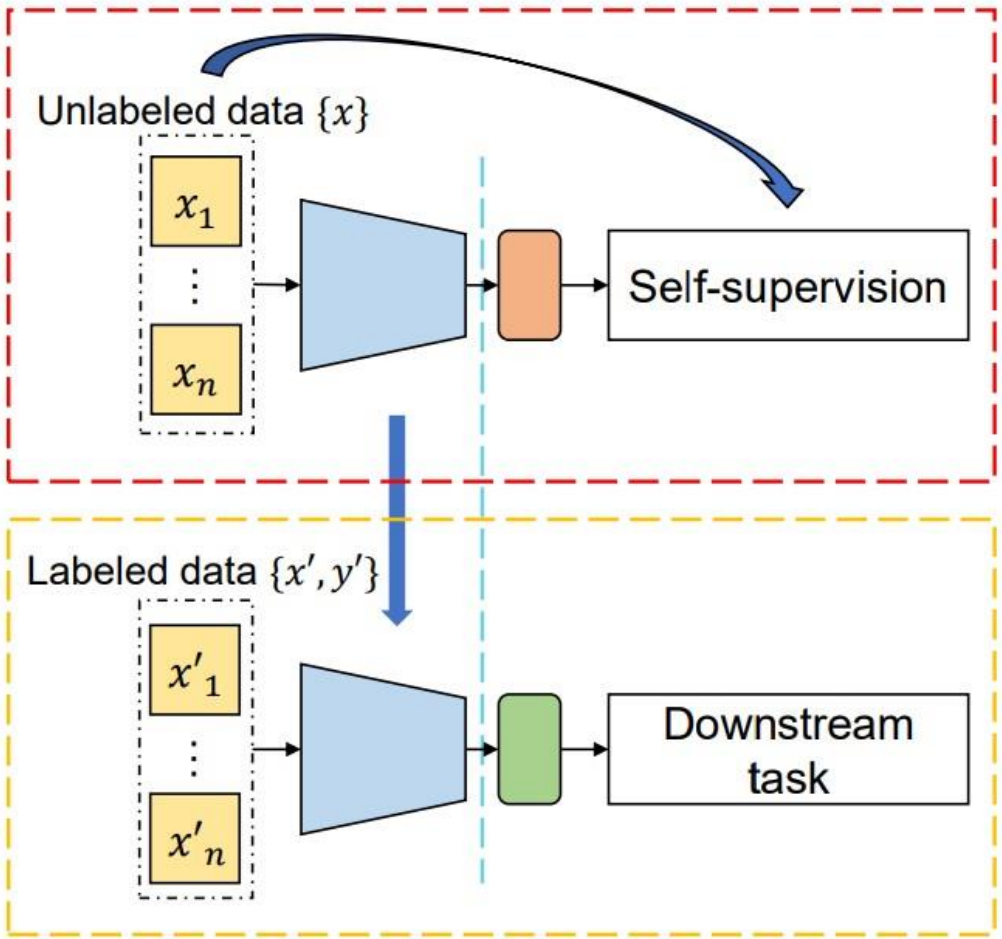
Hamburg, 12.06.2023

Contents

Introduction

1. Introduction | Background | Self-supervised learning

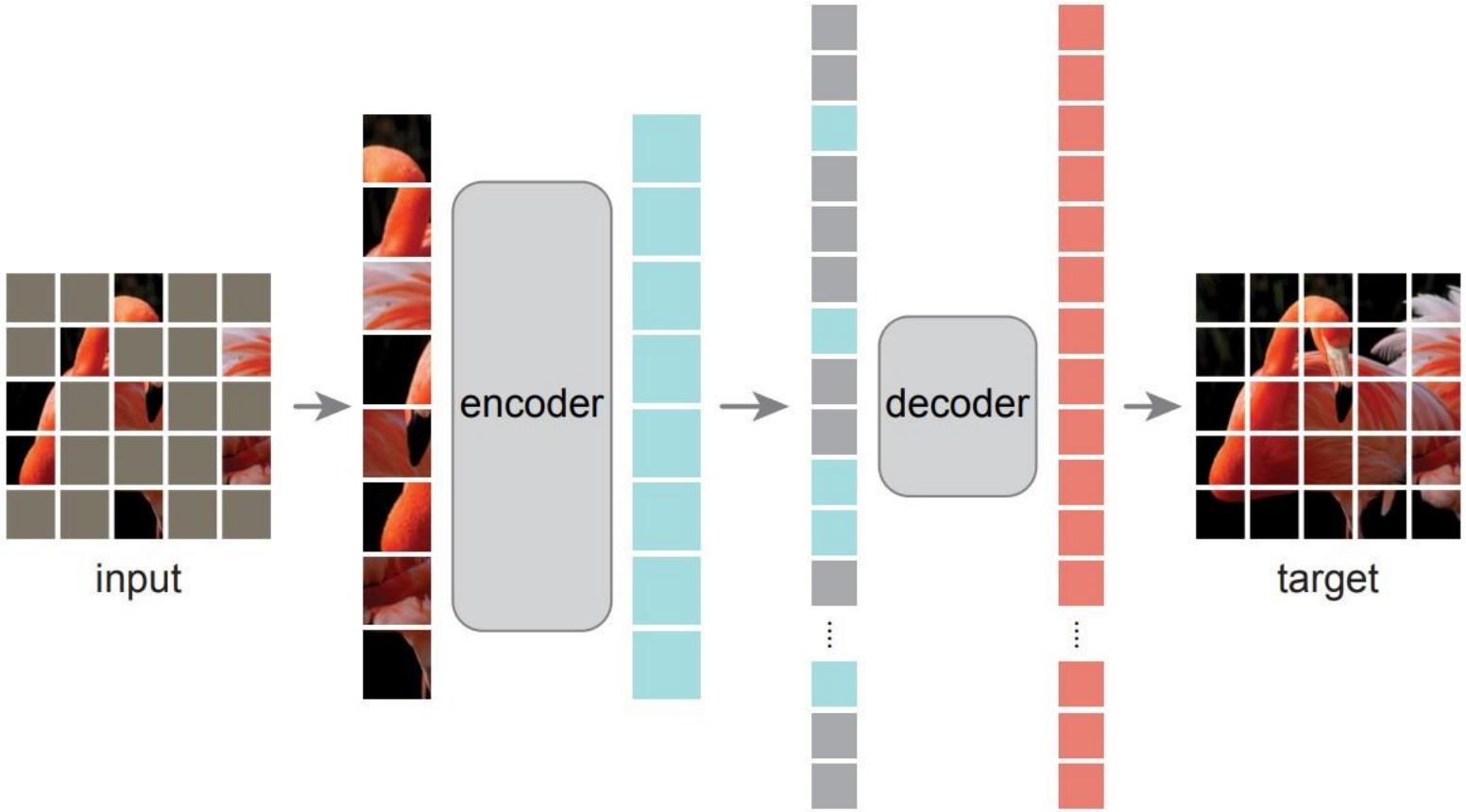
- SSL aims to learn by itself how to learn massive amount of data
- Masked Image Modeling as part of SSL



(Wang et al., 2022)

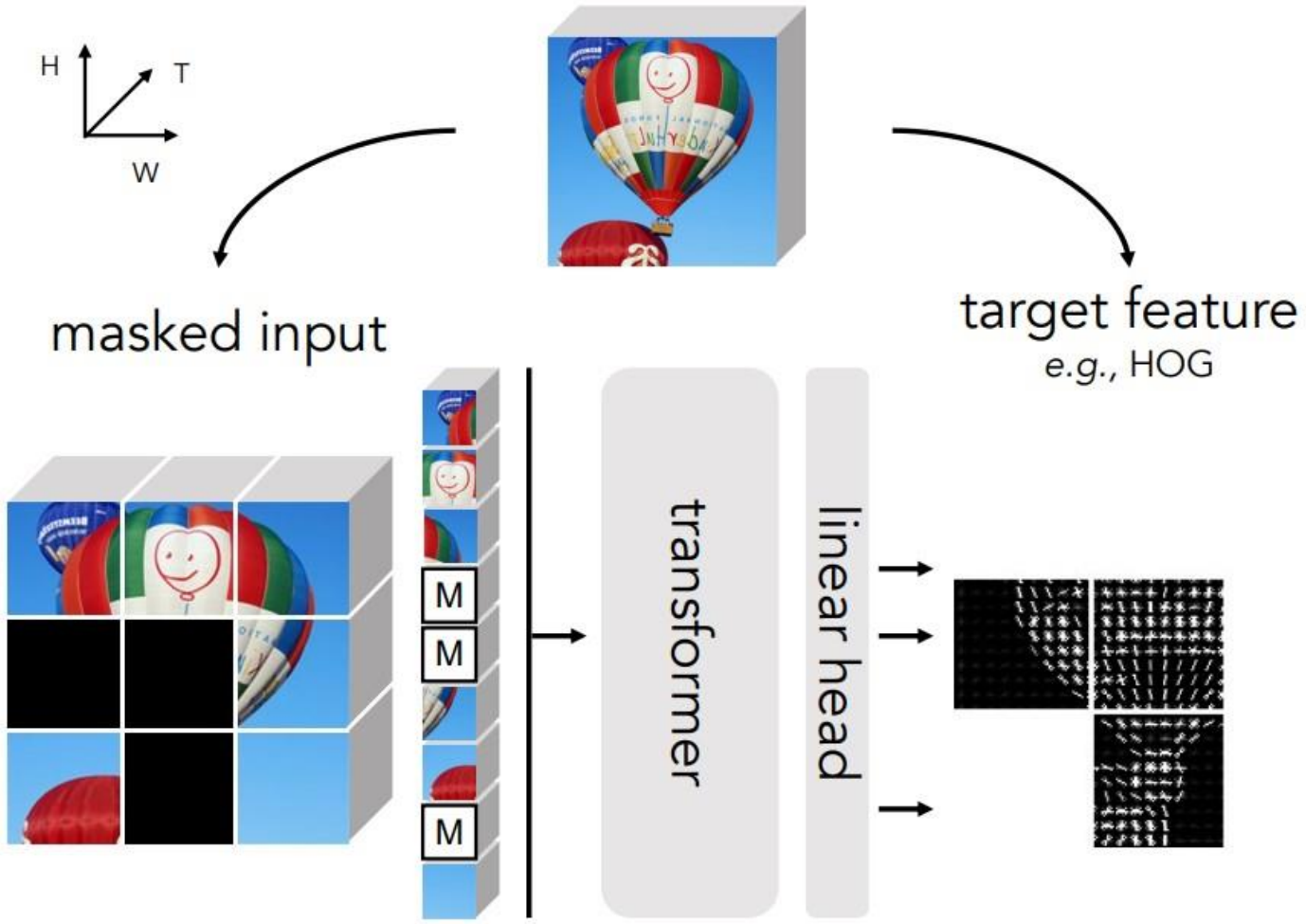
Methodology

2. Methodology | MIM pre-training | Masked Autoencoders



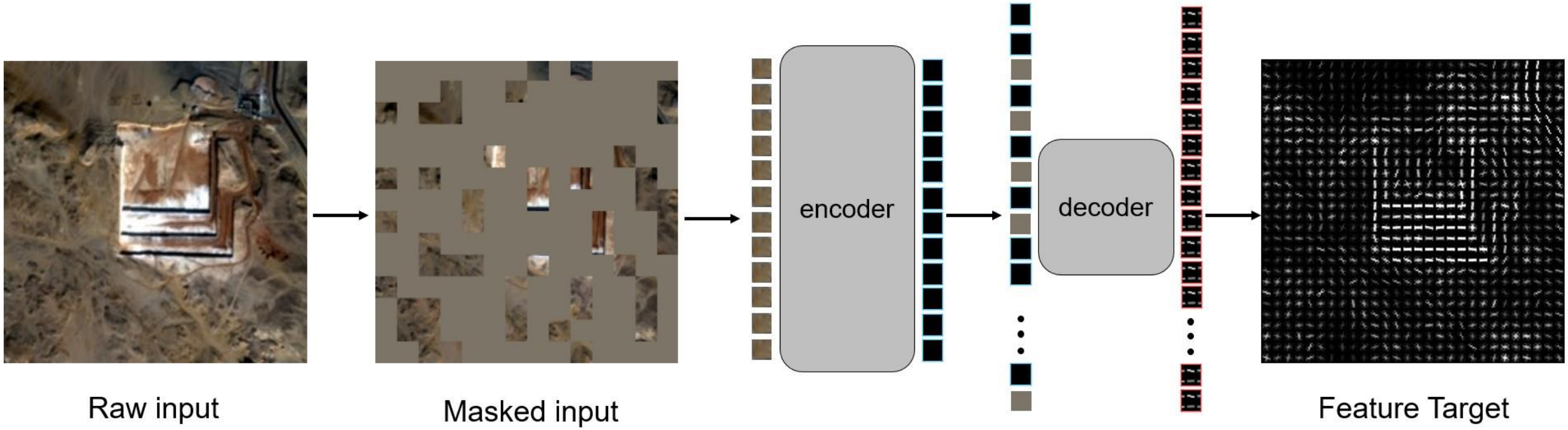
(He et al., 2021)

2. Methodology | MIM pre-training | Masked Feature Prediction

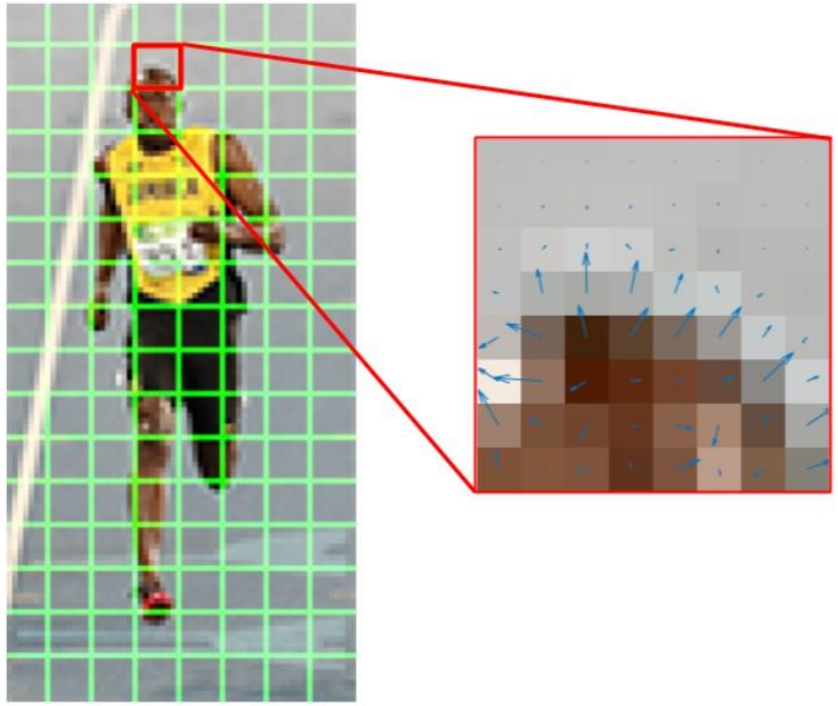


(Wei et al., 2021)

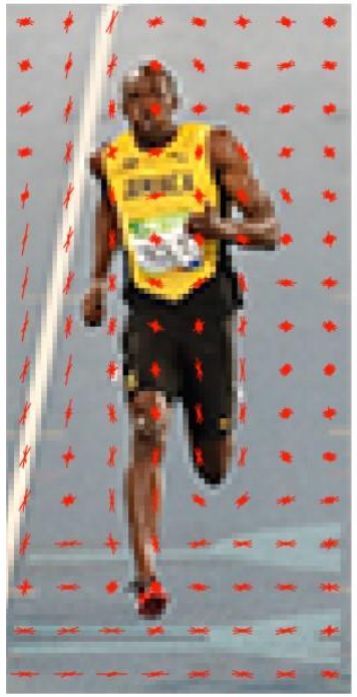
2. Methodology | MIM pre-training | MAE + MFP



2. Methodology | MIM pre-training | Histogram Oriented Gradients



Representation of gradients



(Satya Mallick, 2016)

Visualization of HOG

Experiments

3. Experiments | EO datasets

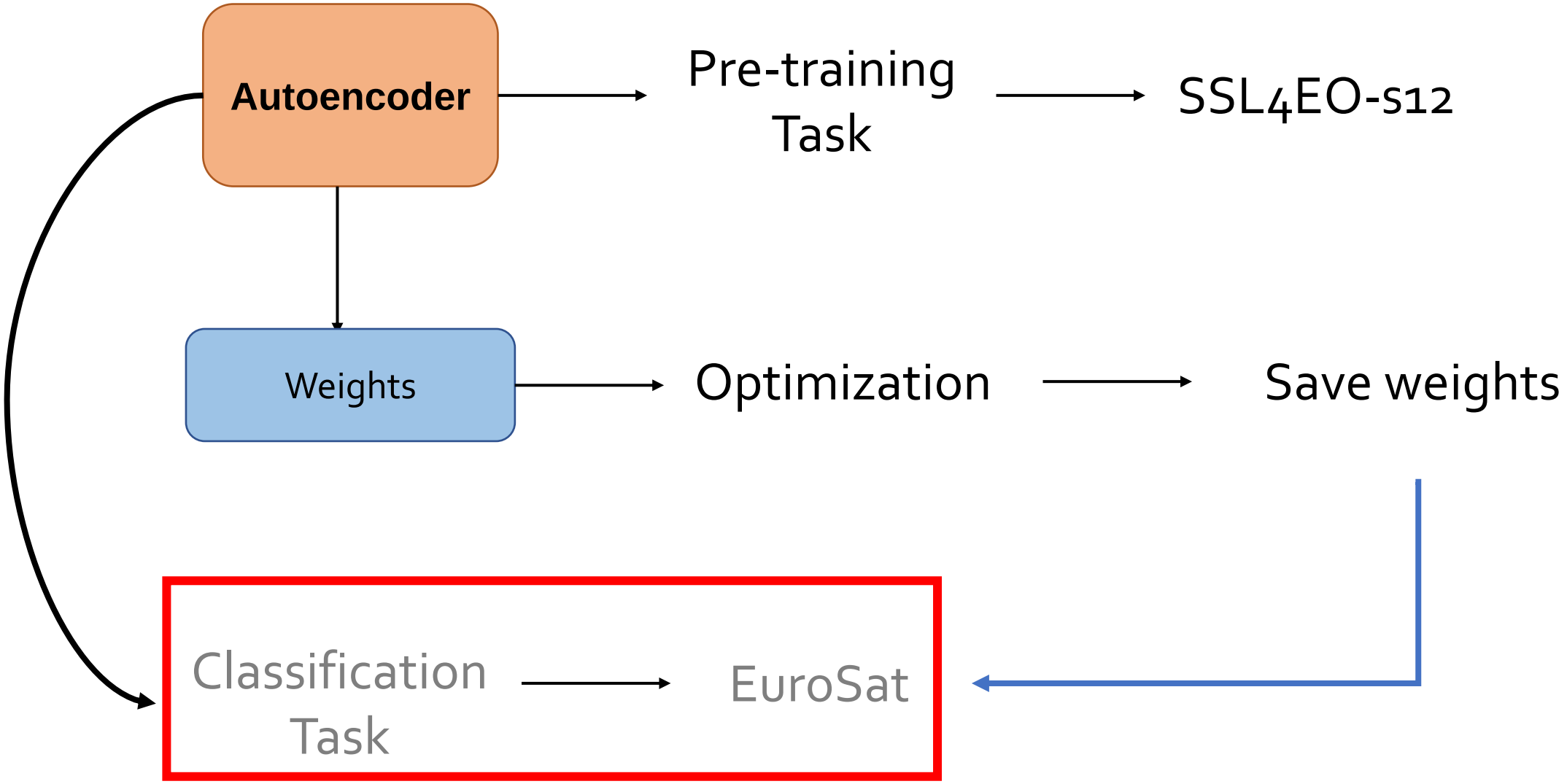
SSL4EO-s12

- 3 million Sentinel-2 and sentinel-1 images
- 250k locations sampled
- Patch scale 264x264 px
- Multi-spectral images over 13 bands

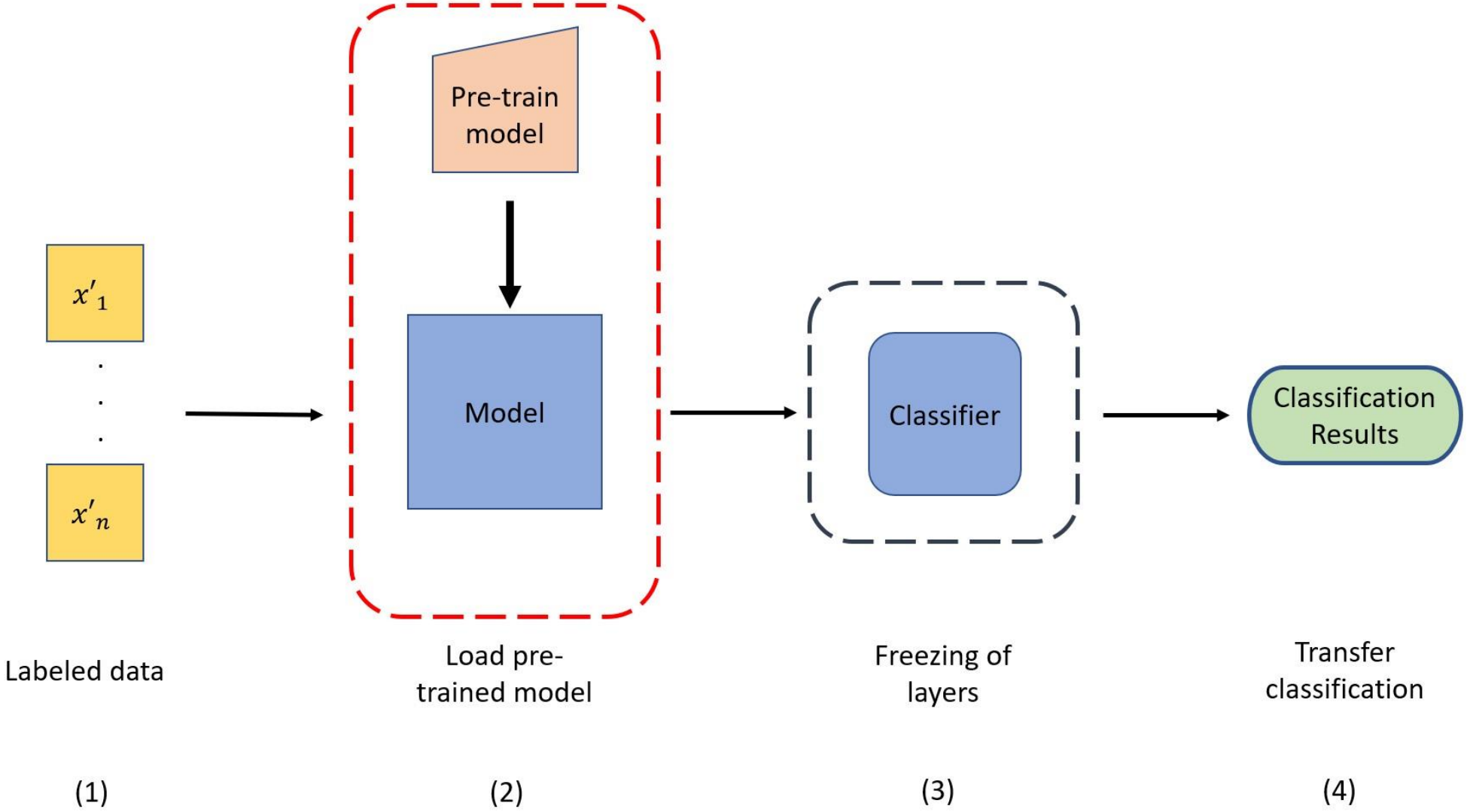
EuroSAT

- 27k labeled and georeferenced images
- Locations distributed all over Europe
- Patch scale 64x64 px
- Multi-spectral images over 13 bands

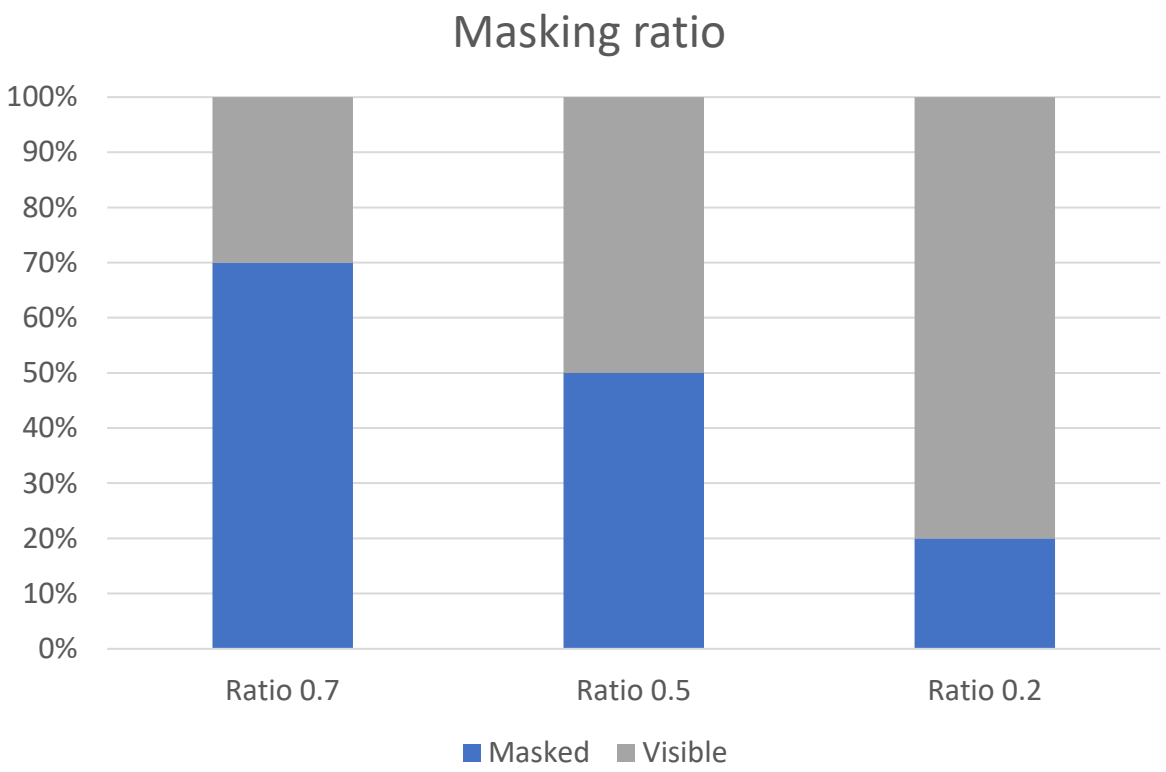
3. Experiments | Self-supervised pre-training

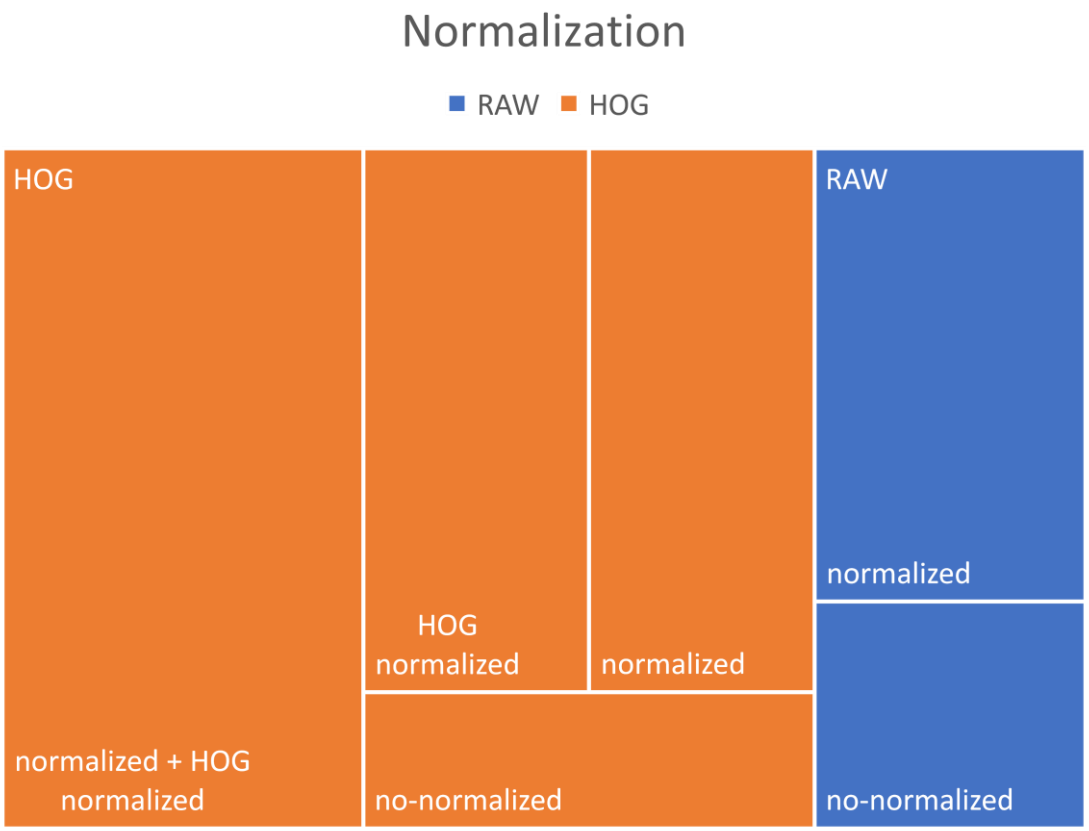


3. Experiments | Downstream Task

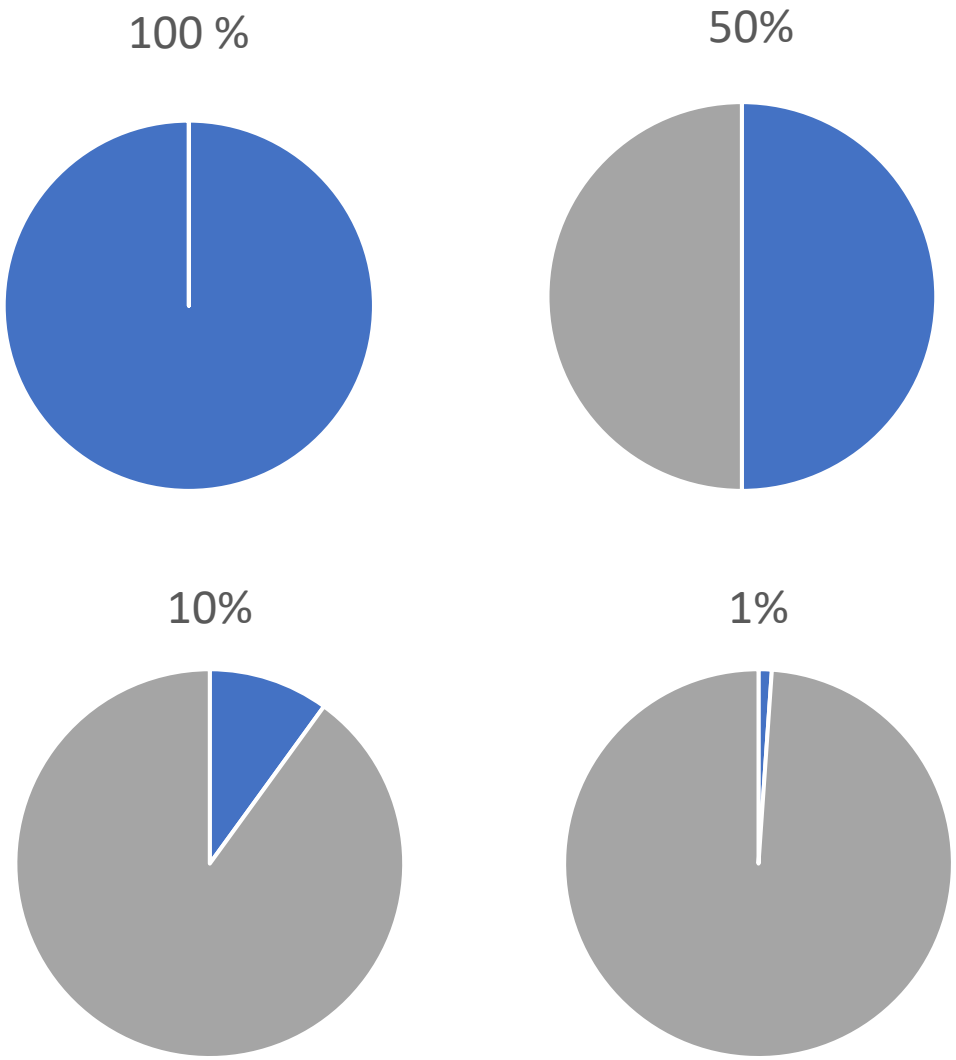


3. Experiments | Ablation studies





3. Experiments | Evaluation strategies



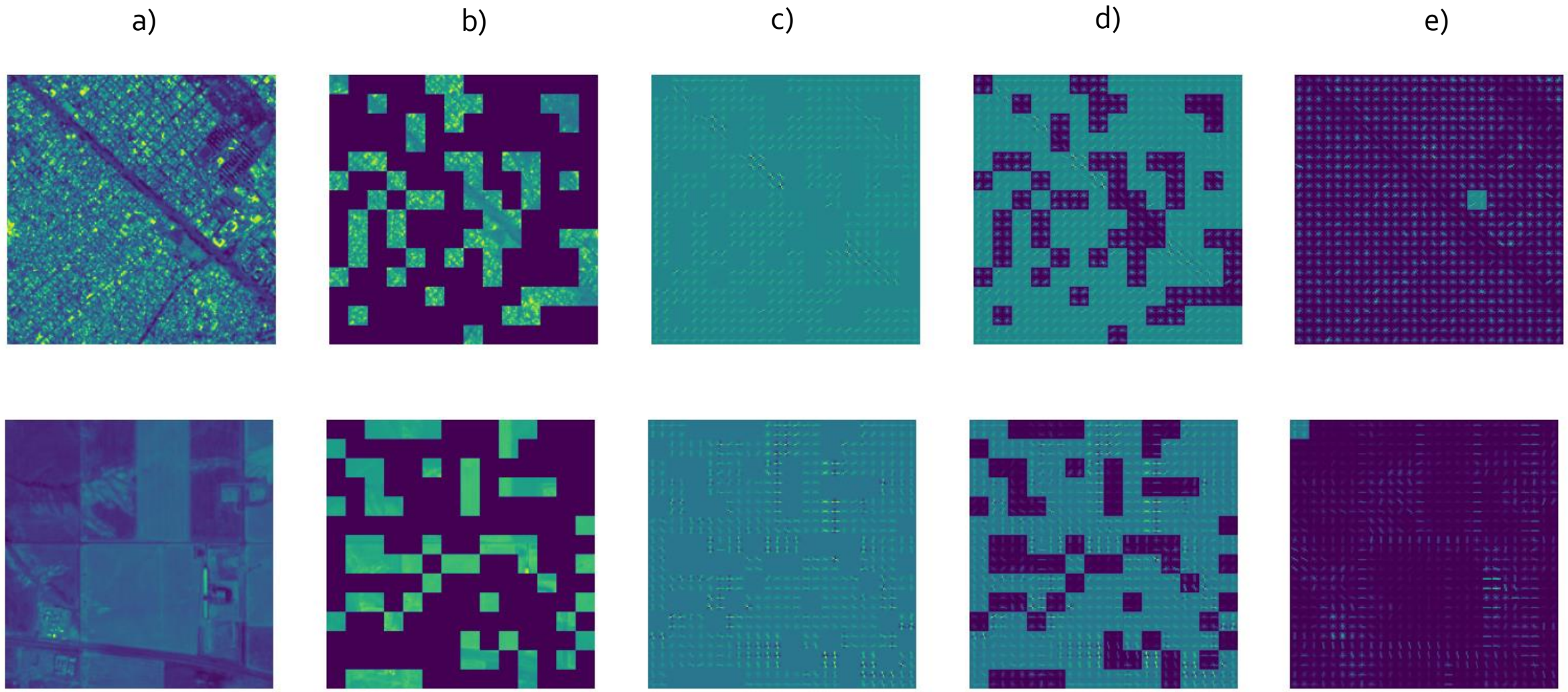
- Labels variation

- Supervised Learning

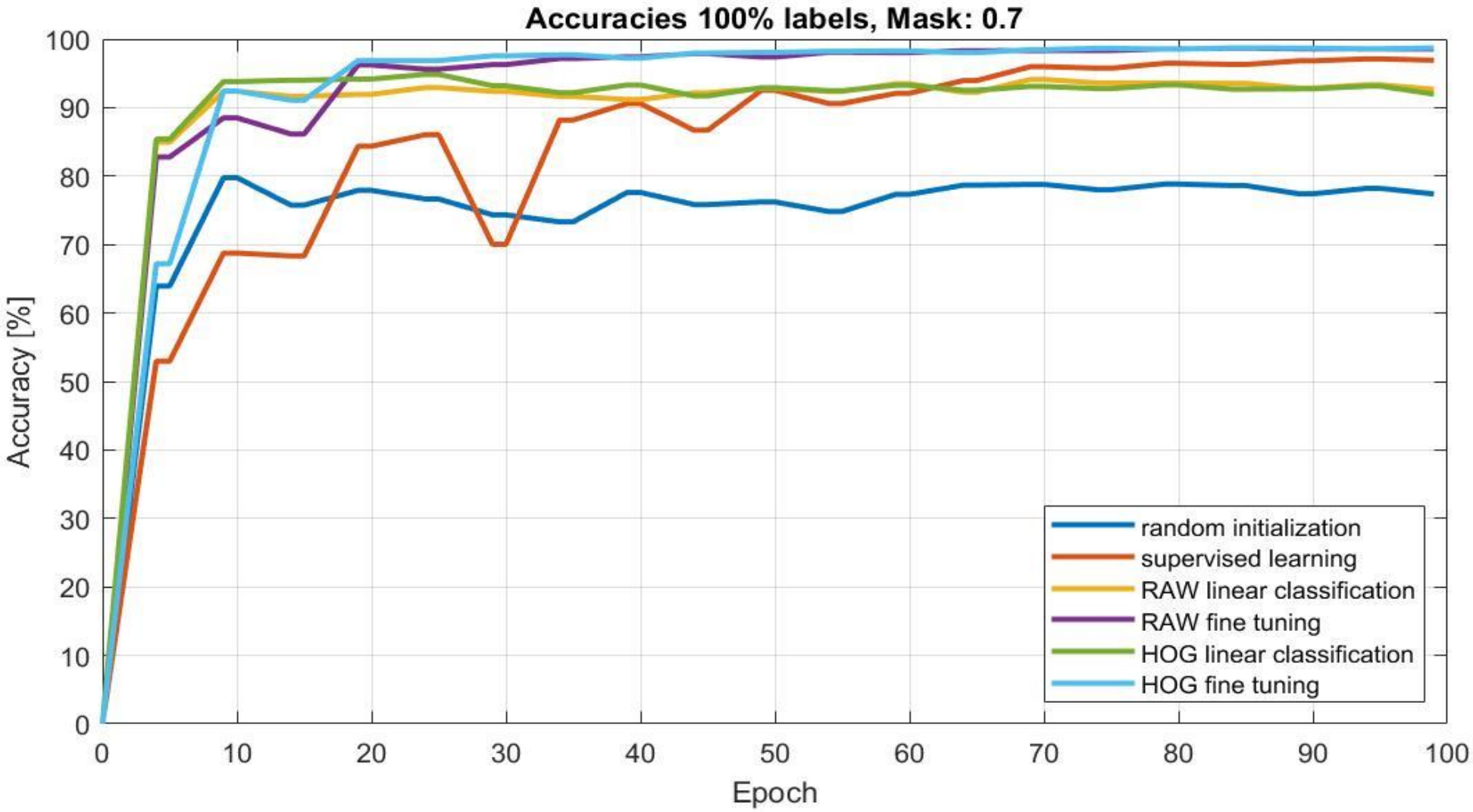
- Random Initialization

Results

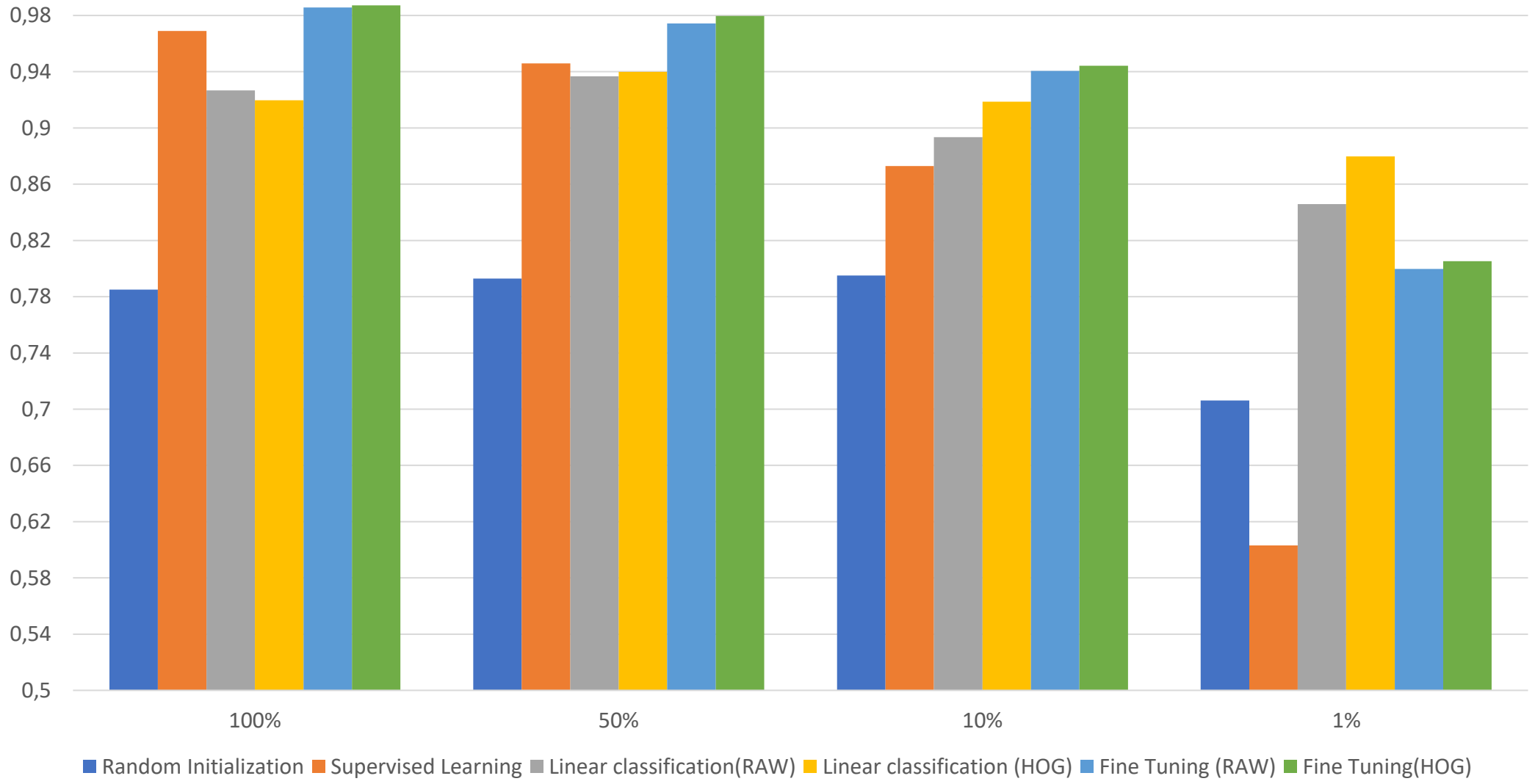
4. Results | SSL pre-training | Image reconstruction



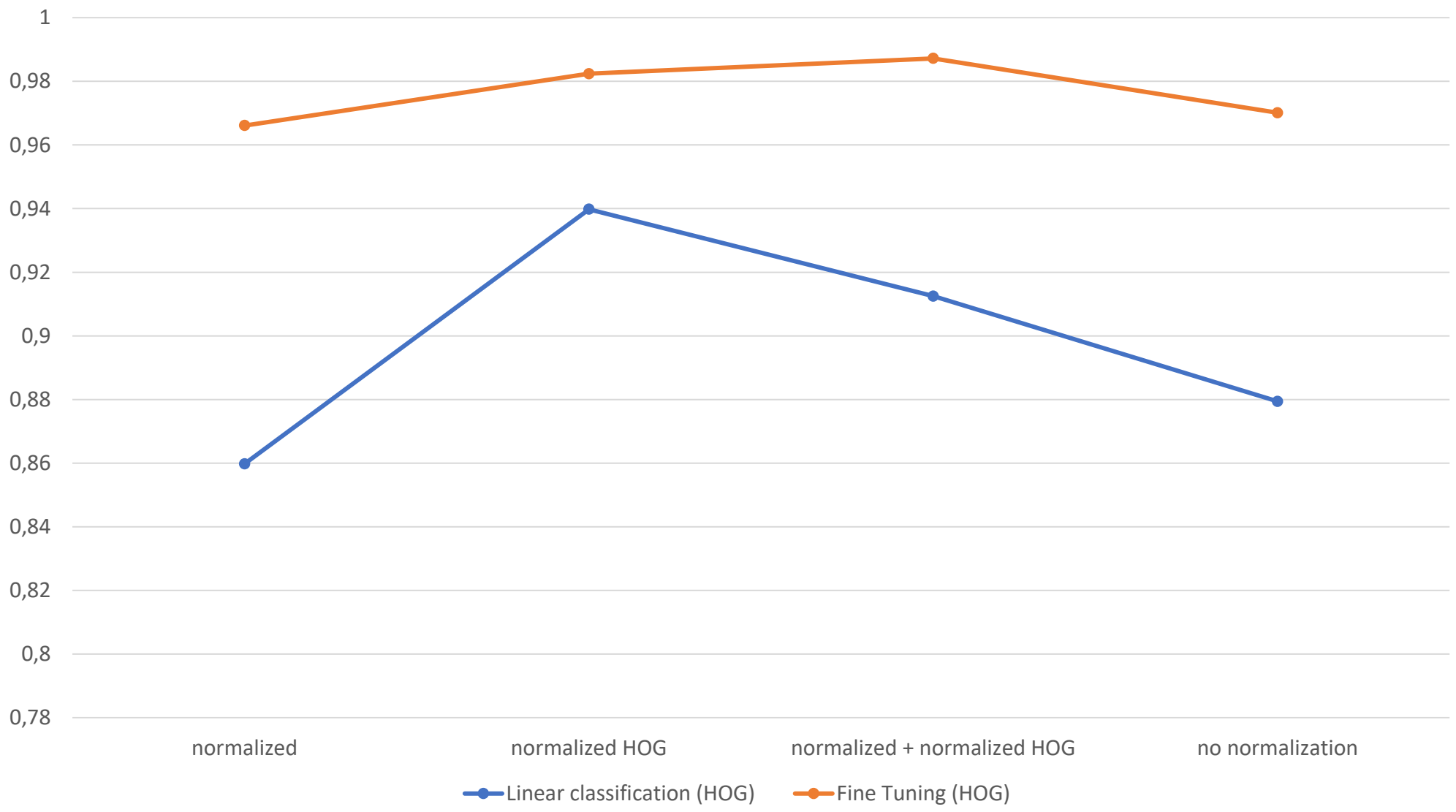
4. Results | Downstream task | Classification accuracies



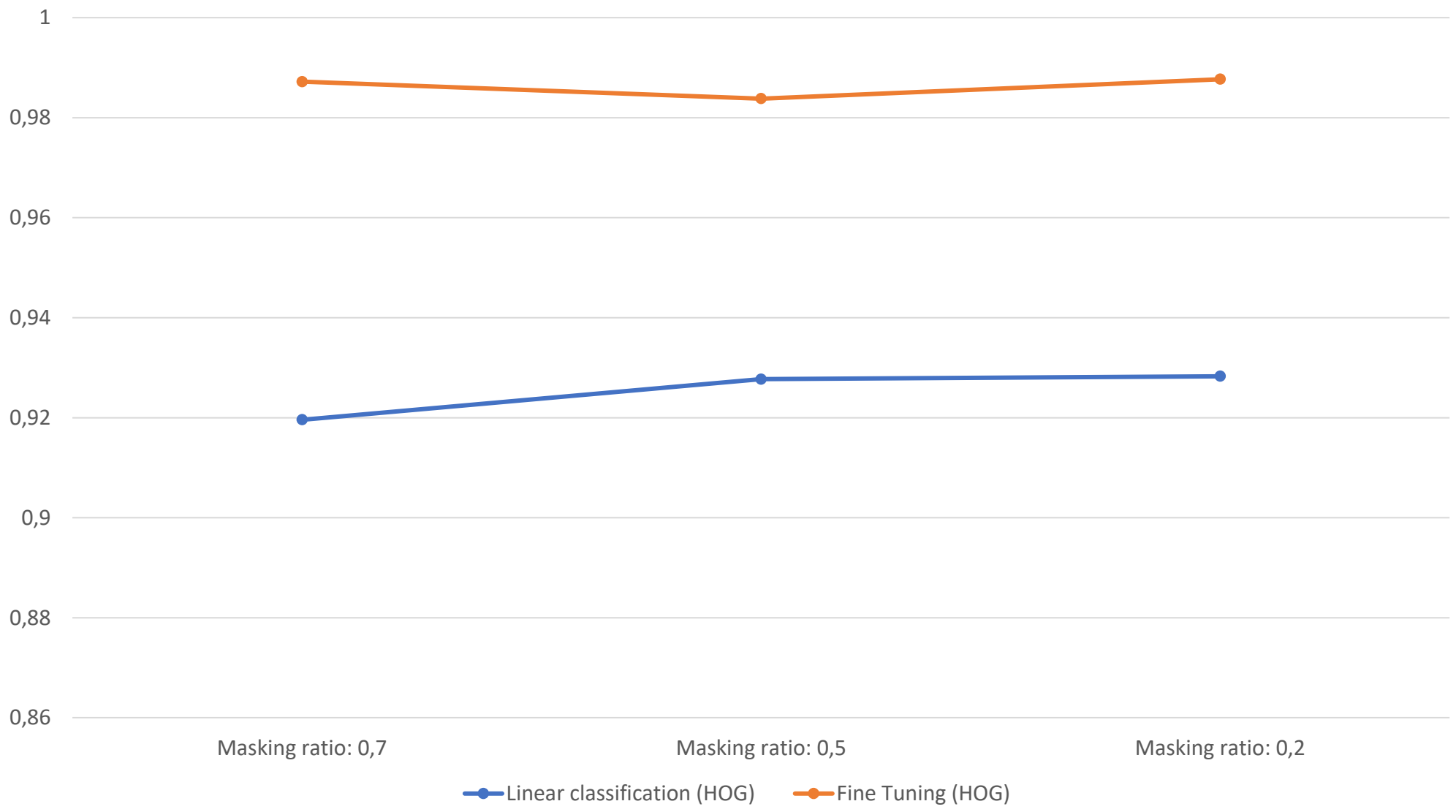
4. Results | Downstream task | Evaluation strategies



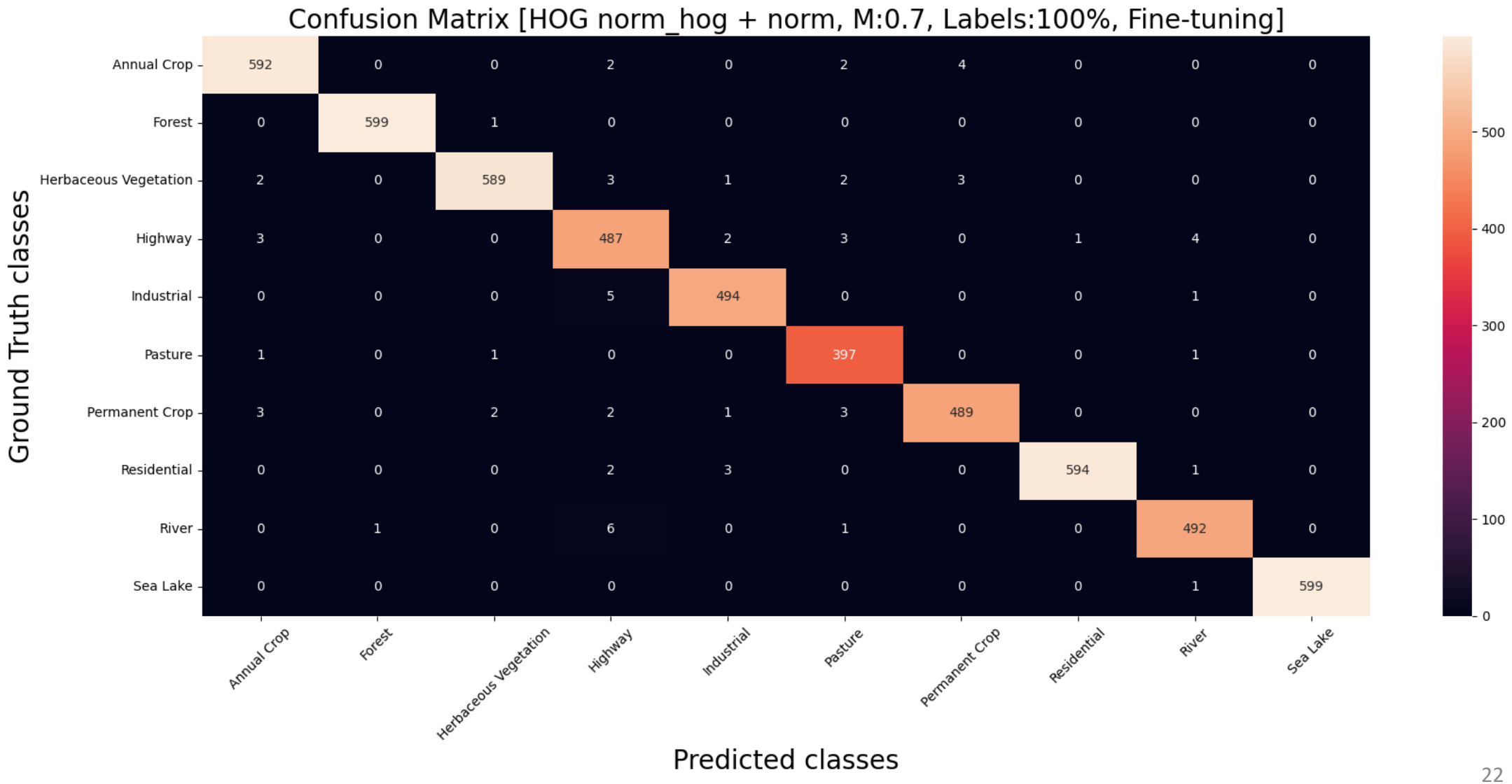
4. Results | Downstream task | Normalization



4. Results | Downstream task | Masking ratio



4. Results | Downstream task | Confusion Matrix 100%



4. Results | Downstream task | Misclassified images

(a)

(b)

(c)

(d)

(e)

5. Conclusions and Outlook

- Hybrid model performed in MSI remote sensing using SSL
- Classification accuracy surpasses state of the art, feature descriptors slightly improve performance
- Performance do not present a big improvement with respect pixelwise analysis
- Open possibility to explore new feature descriptors

Thanks for your attention!

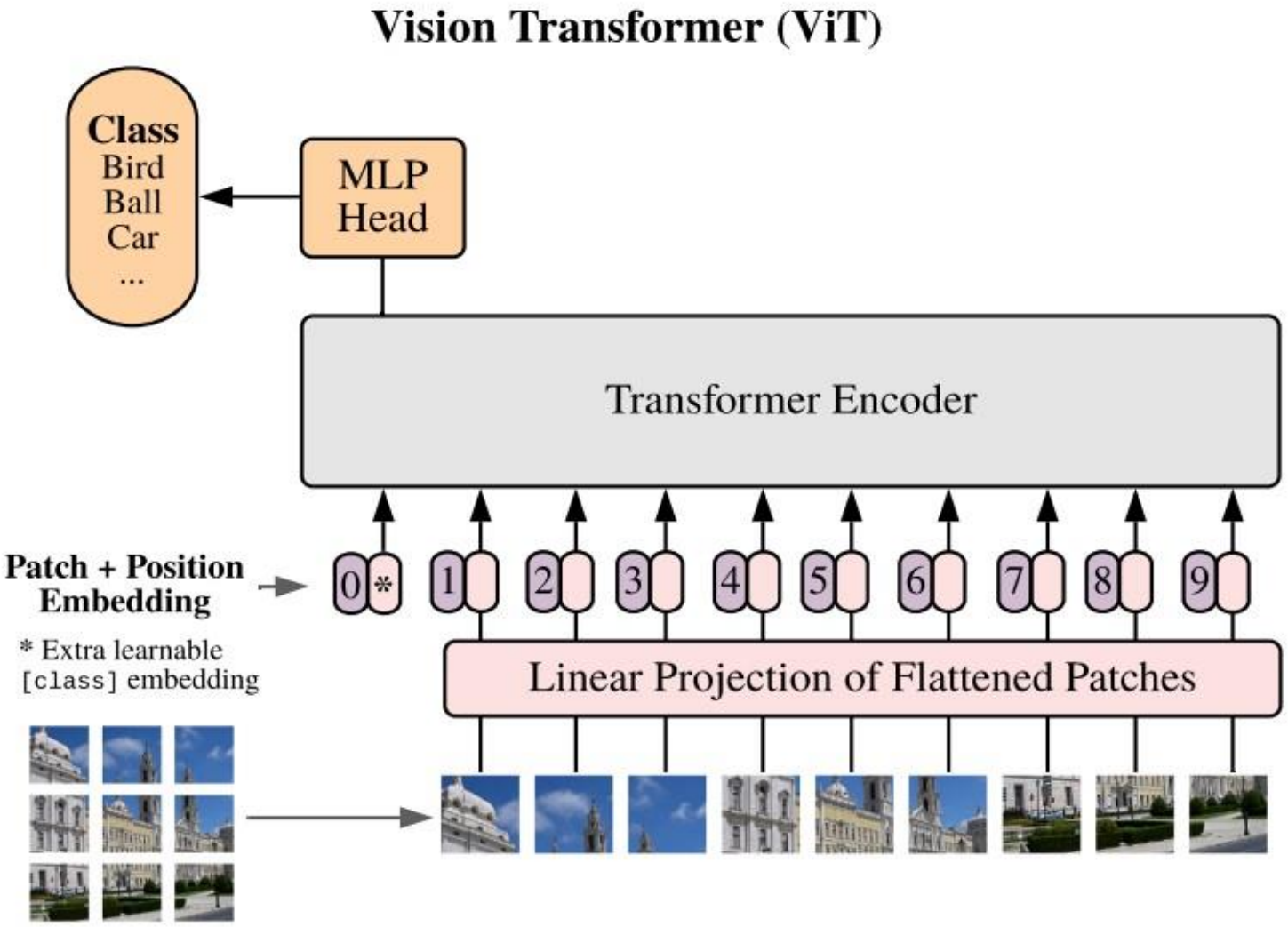
Hugo Hernández Hernández

Appendix

Appendix | Motivation

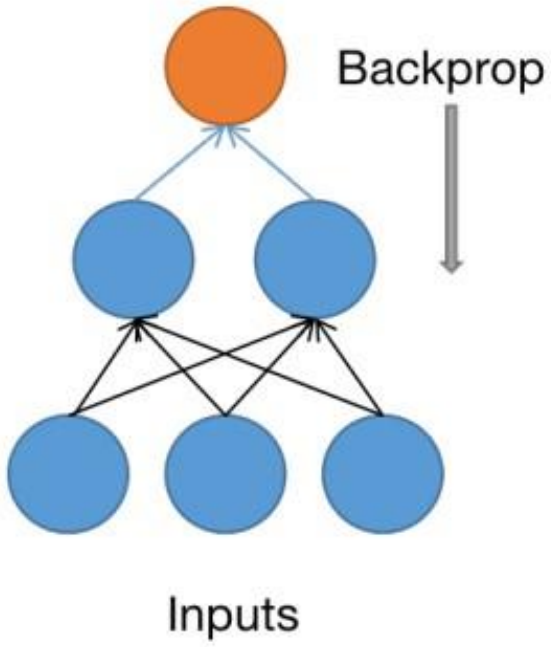
- Human annotated datasets are costly and time-consuming
- Earth observation deals with massive-scale datasets, self-supervised learning can address this issue

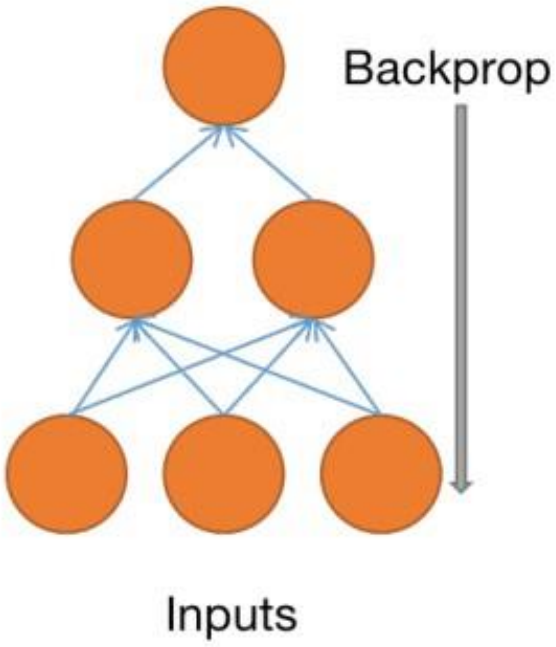
Appendix | Vision Transformers



(Dosovitskiy et al., 2020)

Appendix | Downstream task

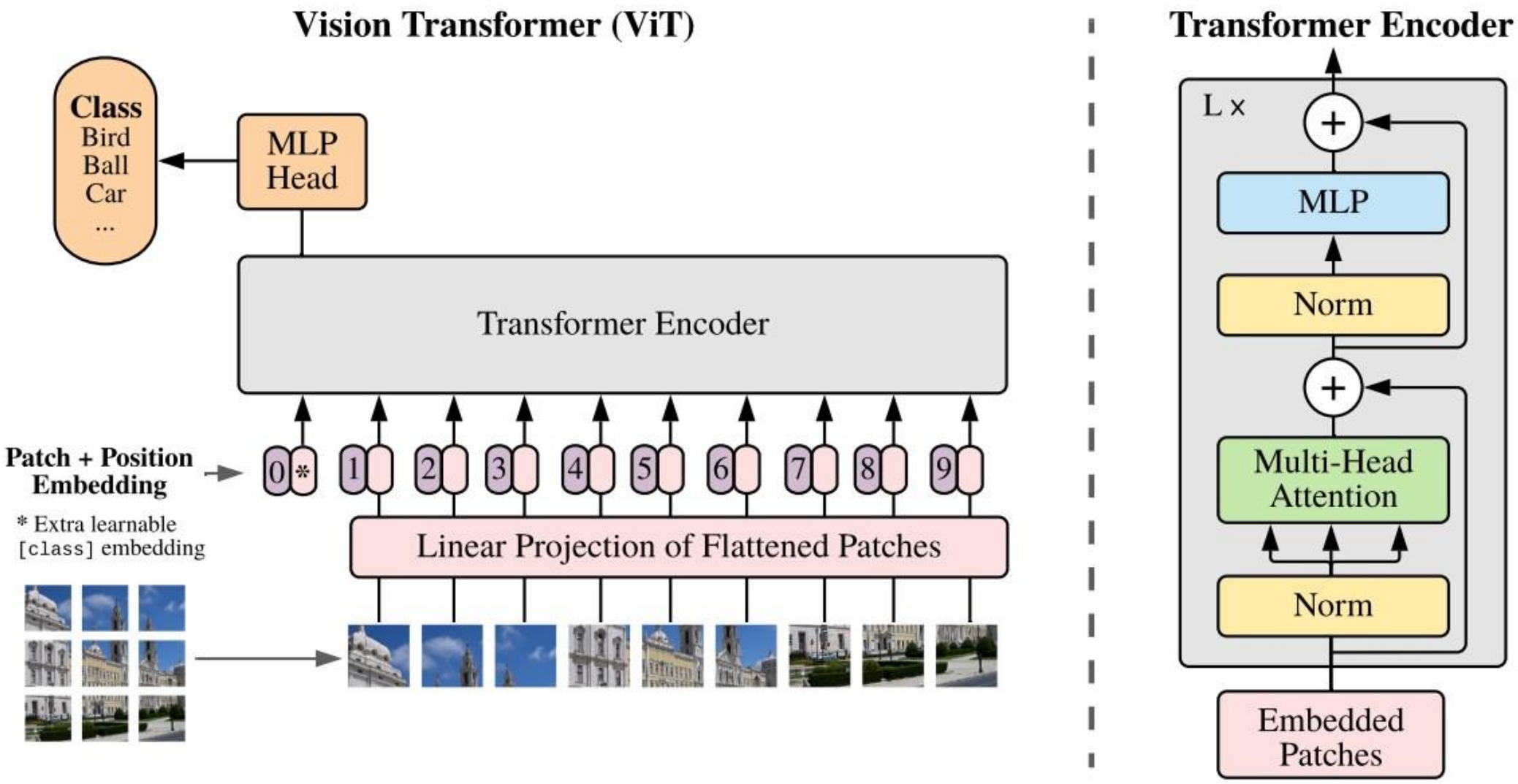




(Gao et al., 2022)

Linear Classification

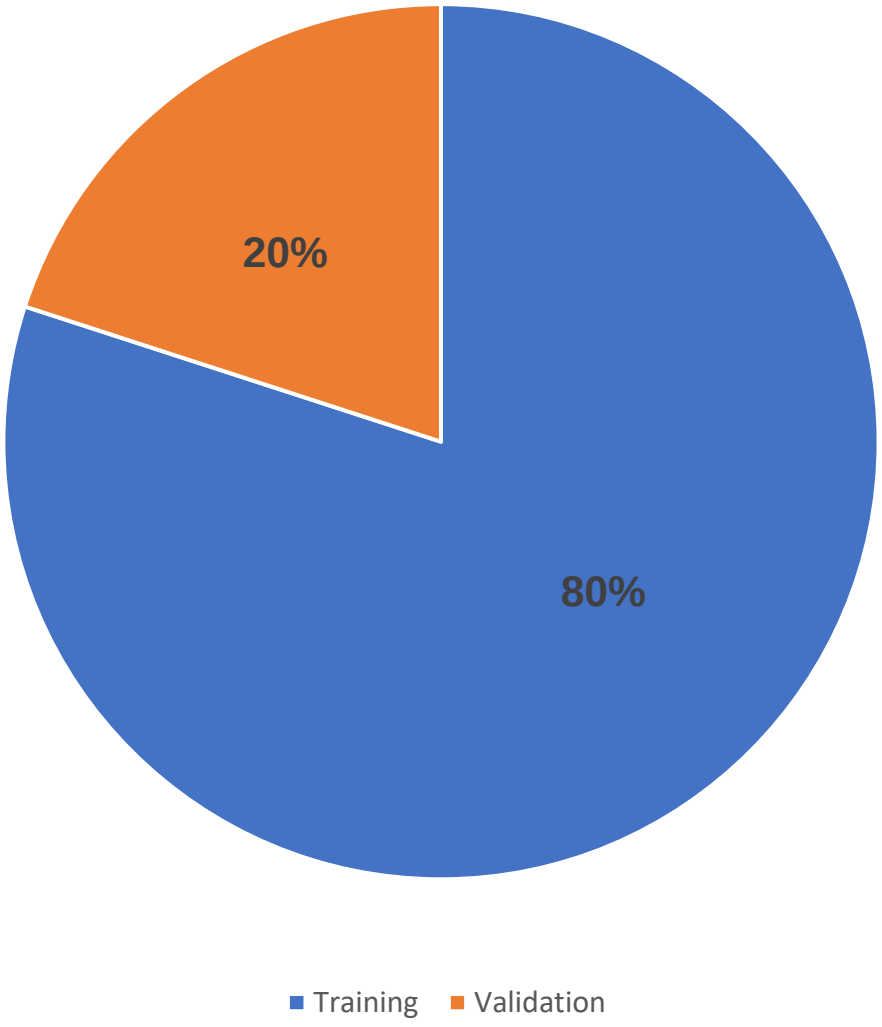
Fine-tuning



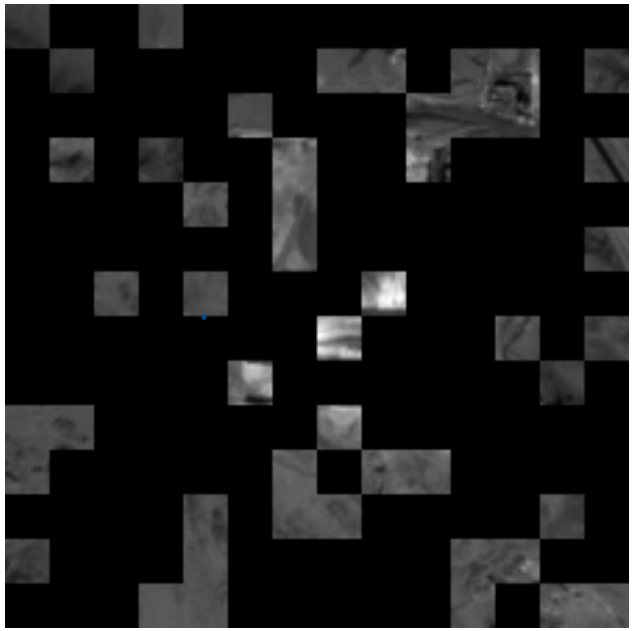
(Dosovitskiy et al., 2020)

Appendix | Data preparation

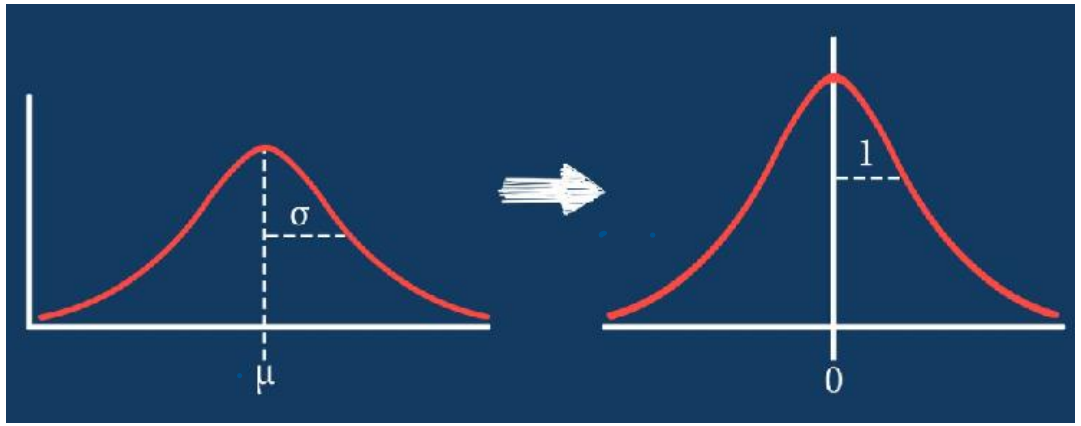
- 27k images
- 10 categories



Appendix | Ablation study



Masking ratio



(Towards AI, 2020)

Normalization

Appendix | SSL pre-training | Training loss

Mask/Norm	A	B	C	D	E	F
0.7	4.55E-04	0.038	0.664	0.034	5.74E-06	5.75E-06
0.5		2.78E-04	0.593	0.030		
0.2		1.72E-04	0.525	0.026		

- A – RAW normalized
- B – RAW no normalized
- C – HOG normalized + hog-normalized
- D – HOG hog-normalized
- E – HOG normalized
- F – HOG no normalized

Appendix | Downstream task | Evaluation strategies

		100%	50%	10%	1%
Random Initialization		0.7840	0.7929	0.795	0.7062
Supervised Learning		0.969	0.9459	0.8729	0.6031
Linear Classif.	**RAW**	0.9267	0.9367	0.8935	0.8459
	HOG	0.9196	0.9398	0.9187	0.8798
Fine Tuning	**RAW**	0.9857	0.9742	0.9405	0.7998
	HOG	**0.9872**	0.9796	0.9442	0.8053

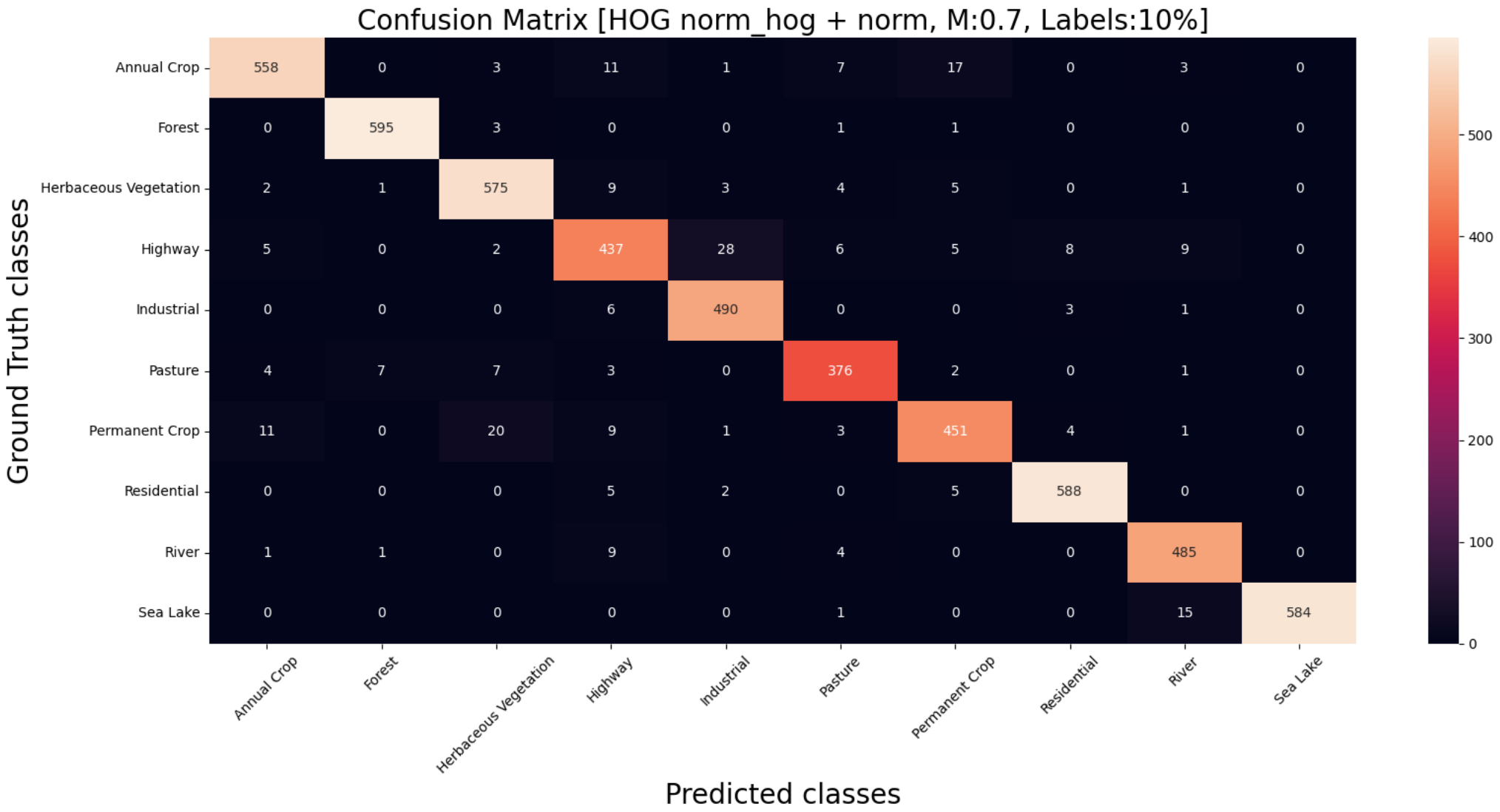
Appendix | Downstream task | Normalization

		norm	norm.hog	n+norm.hog	-
Linear Classif.	**HOG**	0.8598	0.9398	0.9125	0.8794
Fine Tuning	**HOG**	0.9661	0.9824	**0.9872**	0.9701

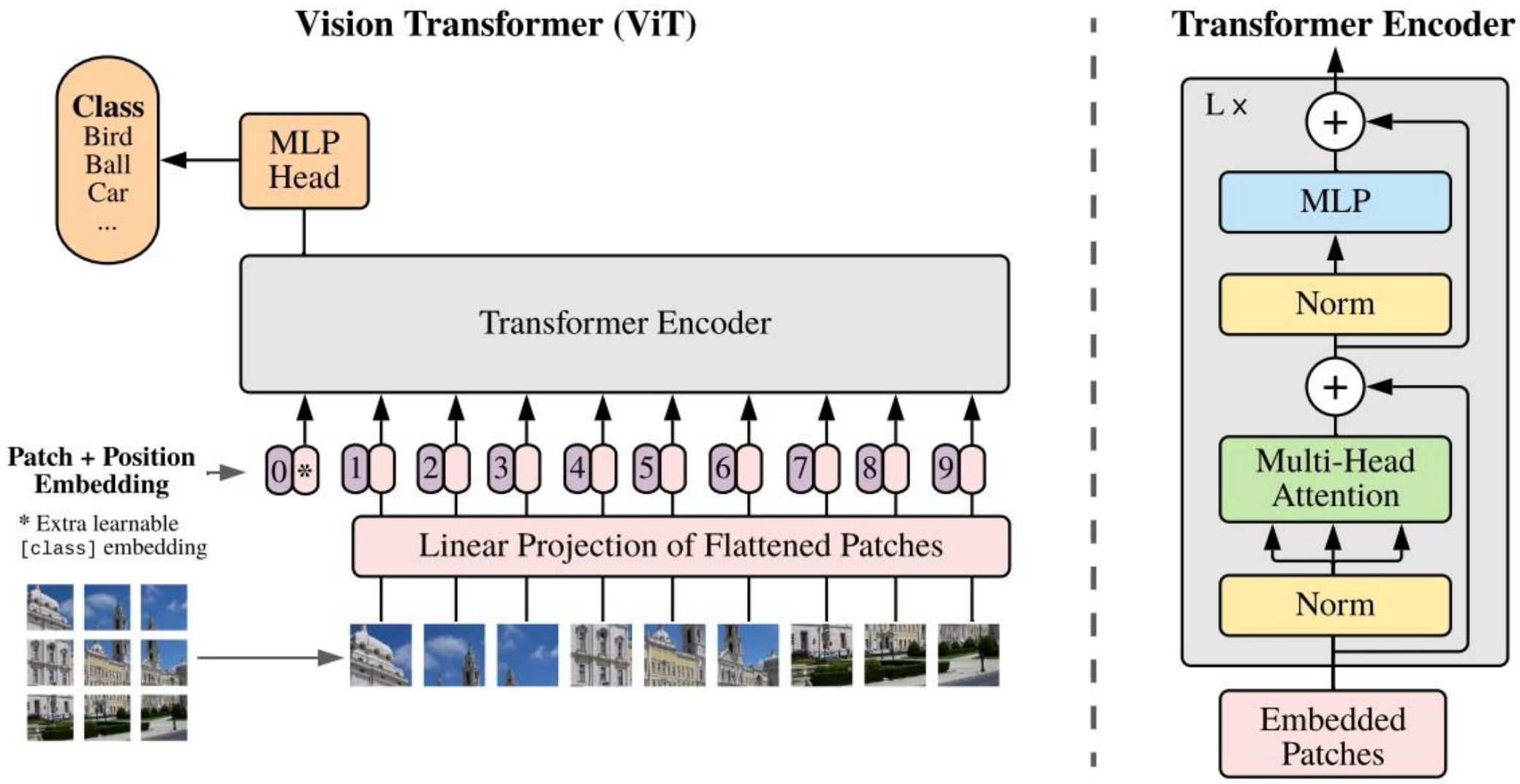
Appendix | Downstream task | Masking ratio

		M: 0.7	M: 0.5	M: 0.2
Linear Classif.	**HOG**	0.9196	0.9277	0.9283
Fine Tuning	**HOG**	0.9872	0.9838	**0.9877**

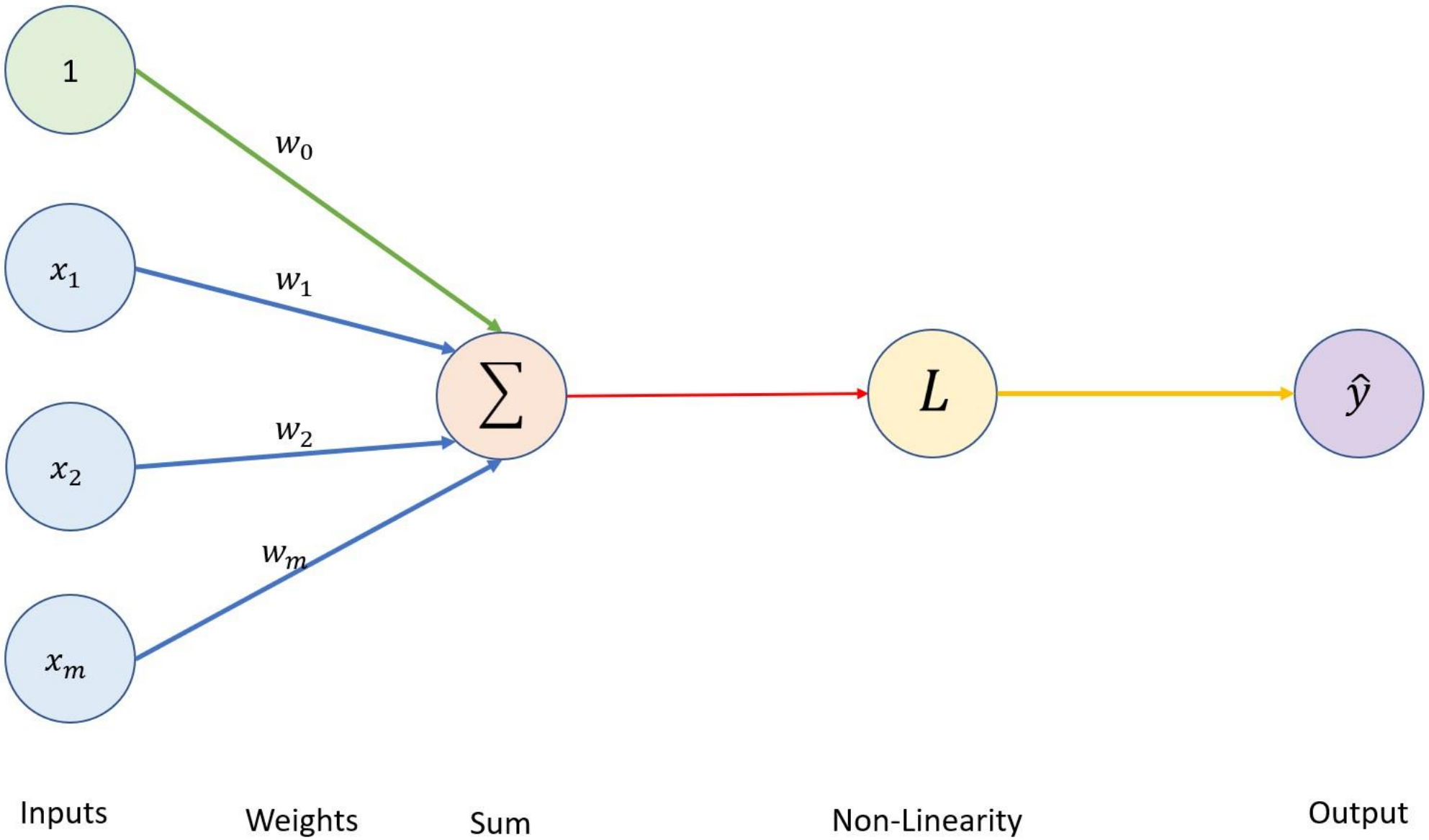
4. Results | Downstream task | Confusion Matrix 10%

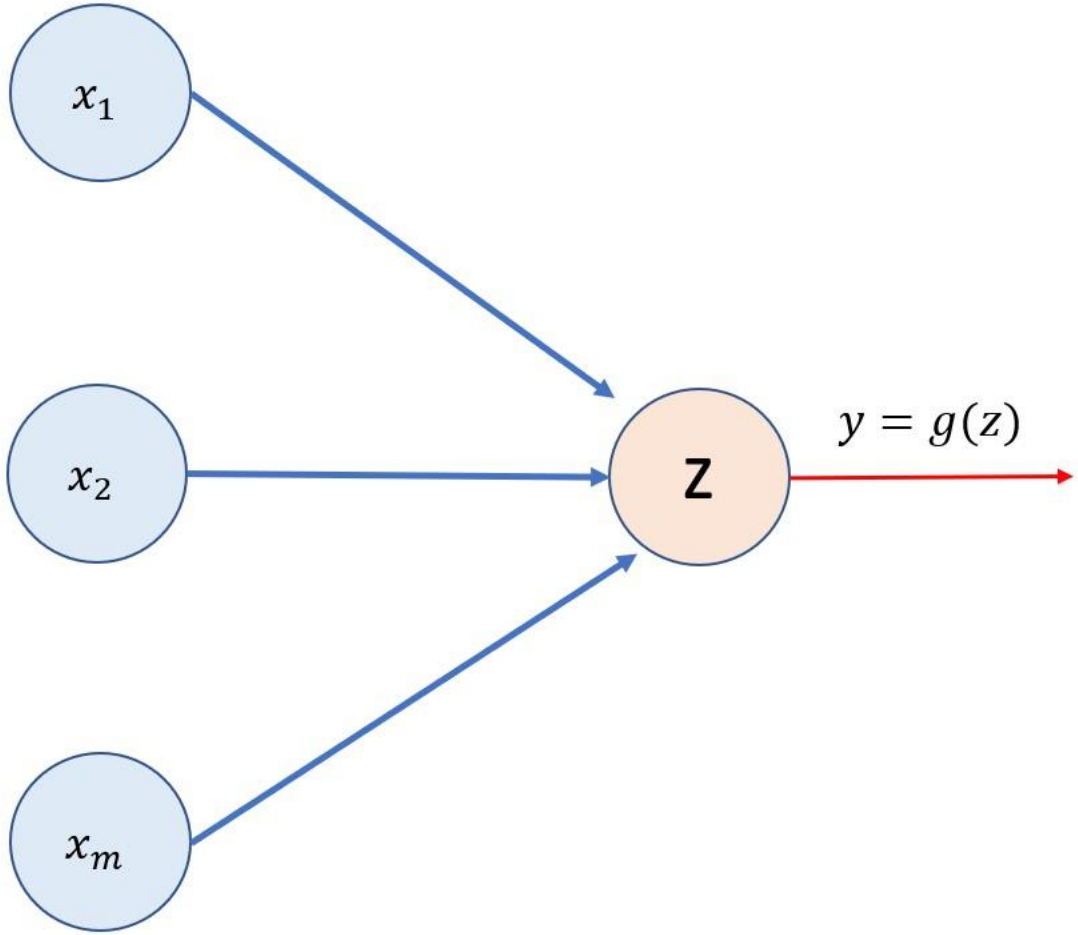


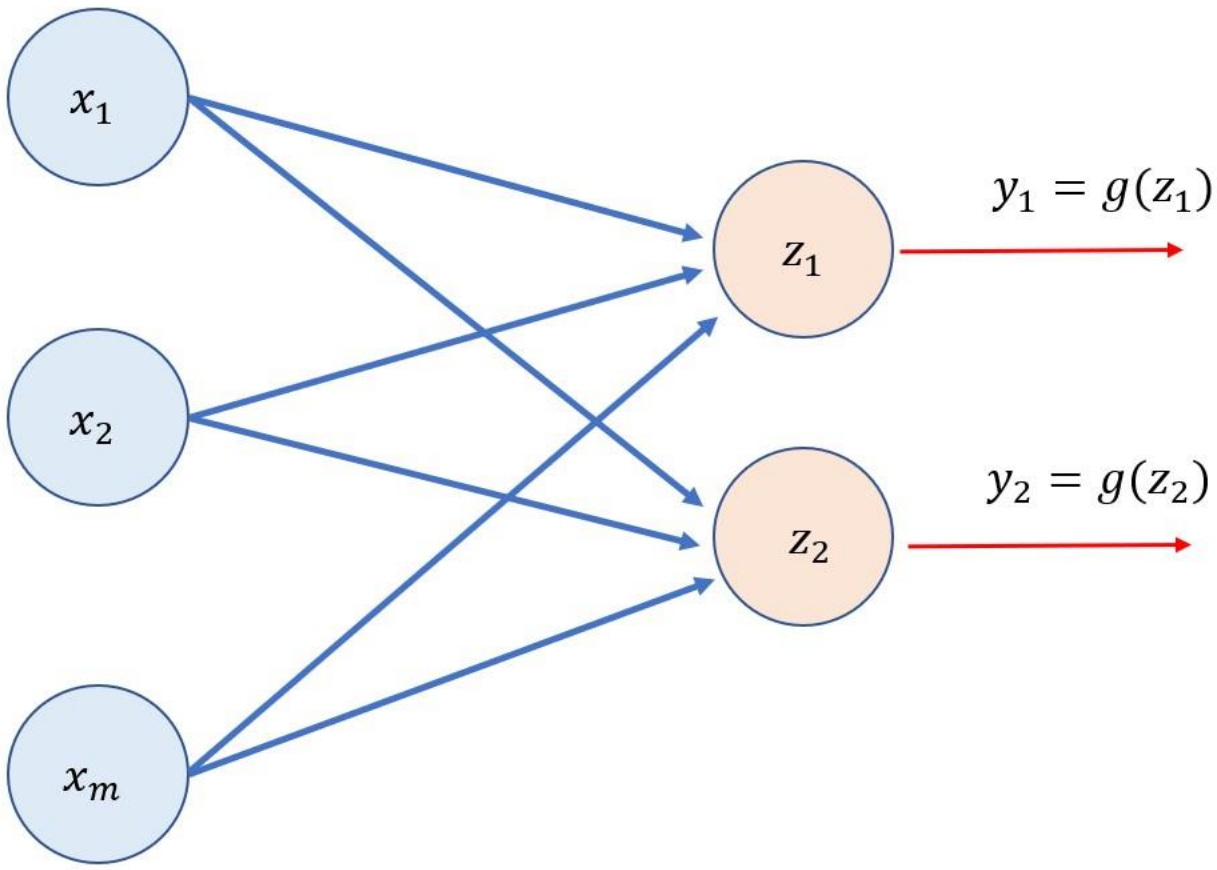
Appendix | Vision Transformer complete

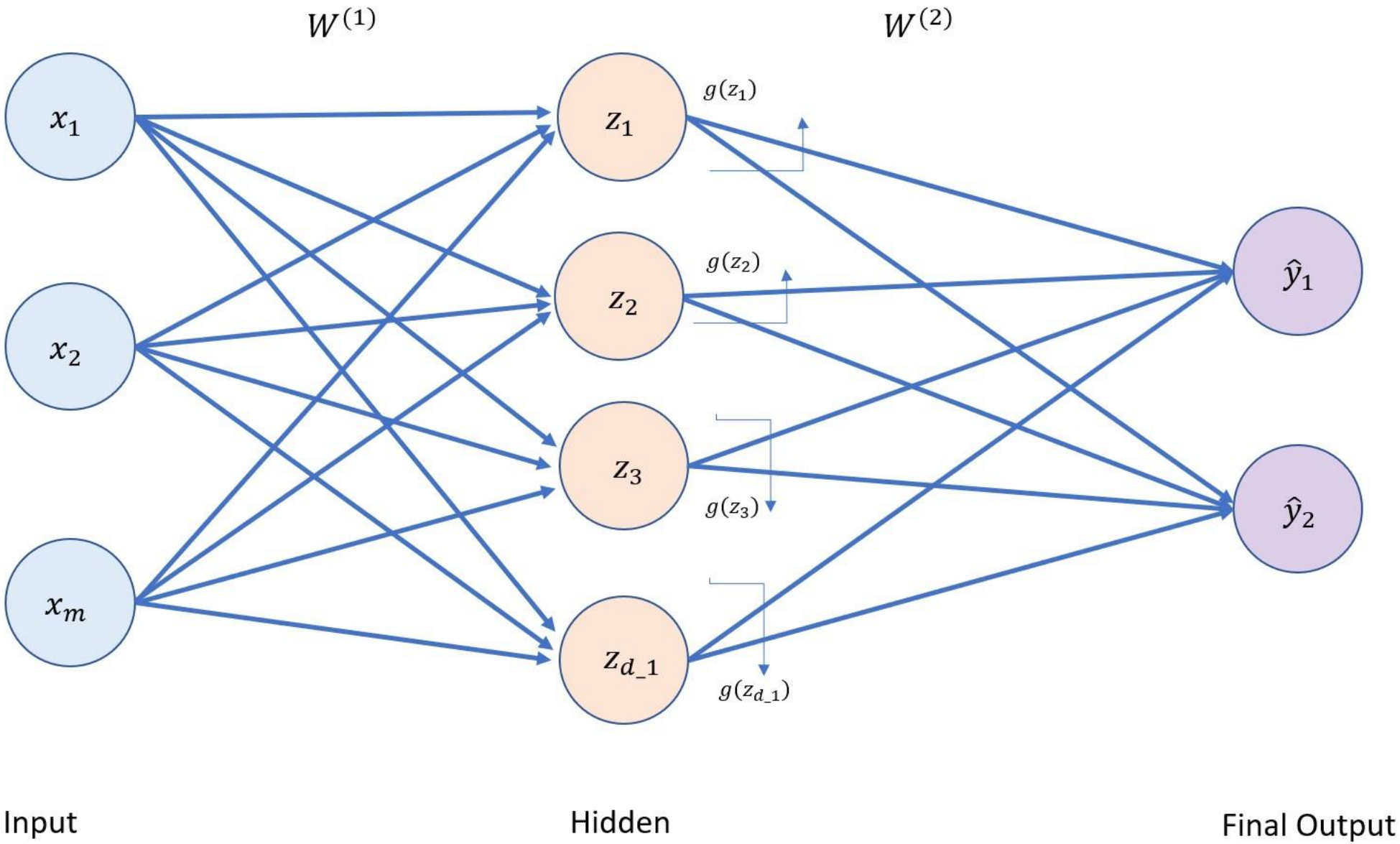


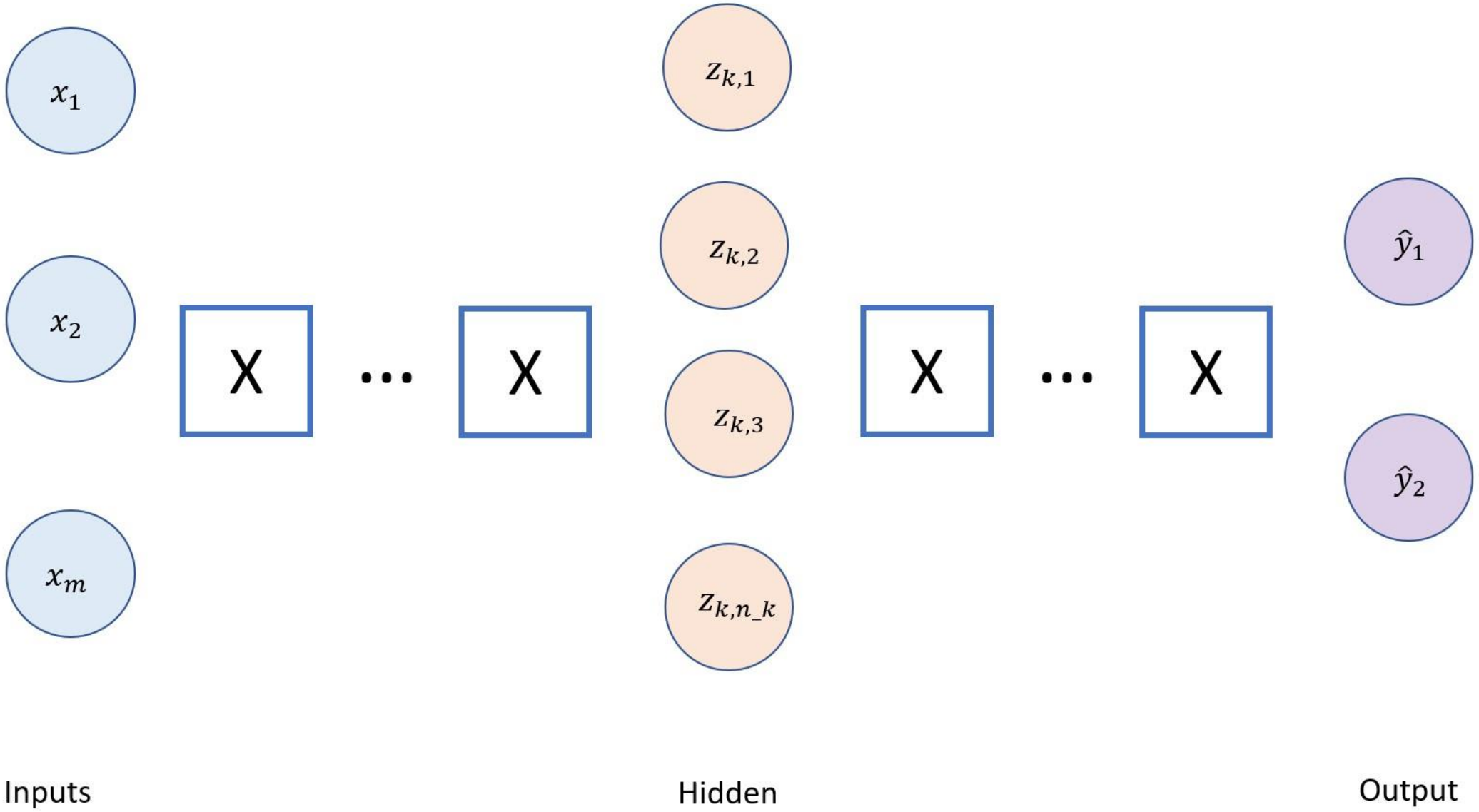
(Dosovitskiy et al., 2020)

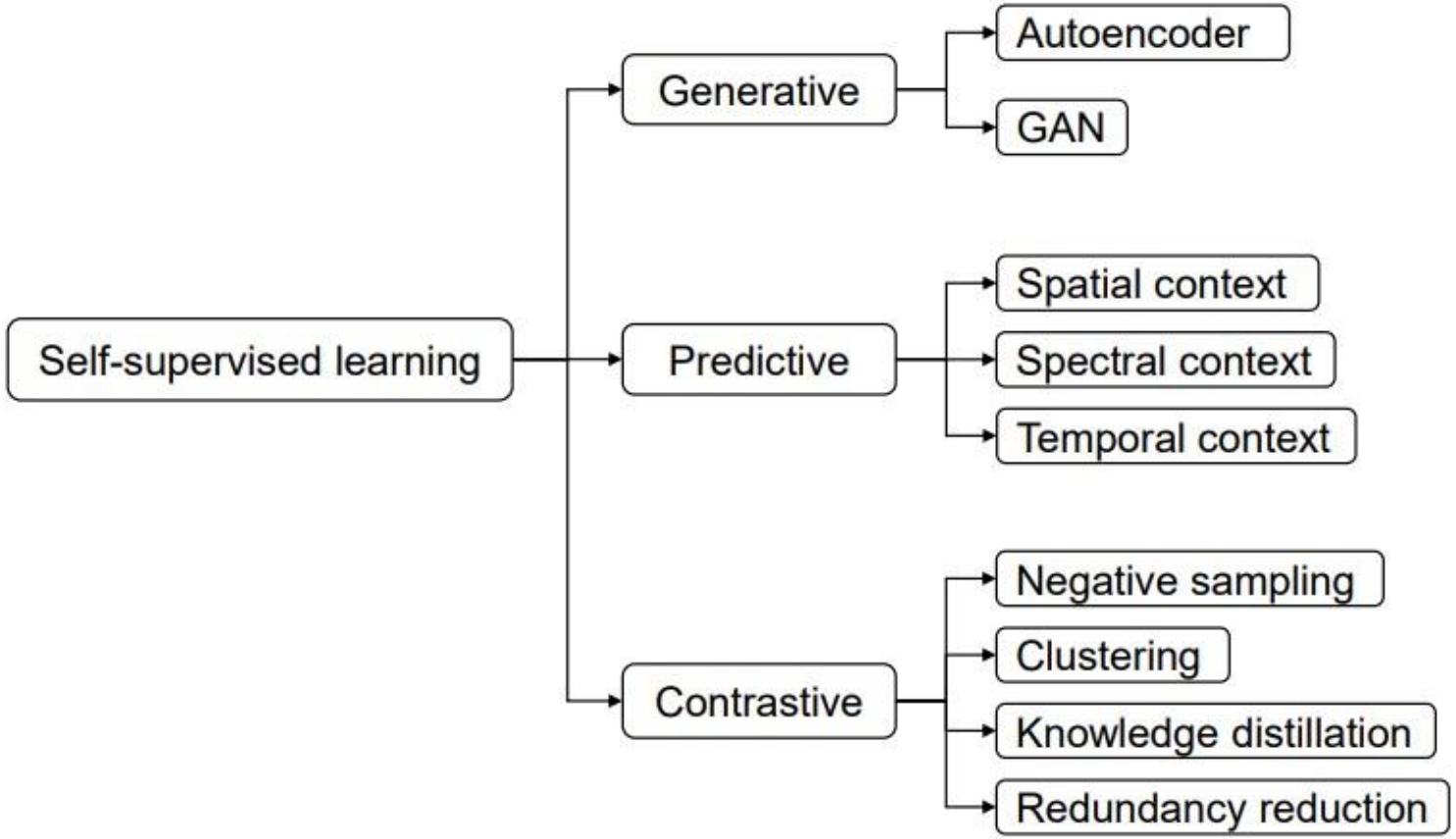

1
w_0
x_1
w_1
x_2
w_2
w_m
x_m
$\sum$
L
$\hat{y}$
Inputs
Weights
Sum
Non-Linearity
Output


x_1
x_2
x_m
Z
$y = g(z)$

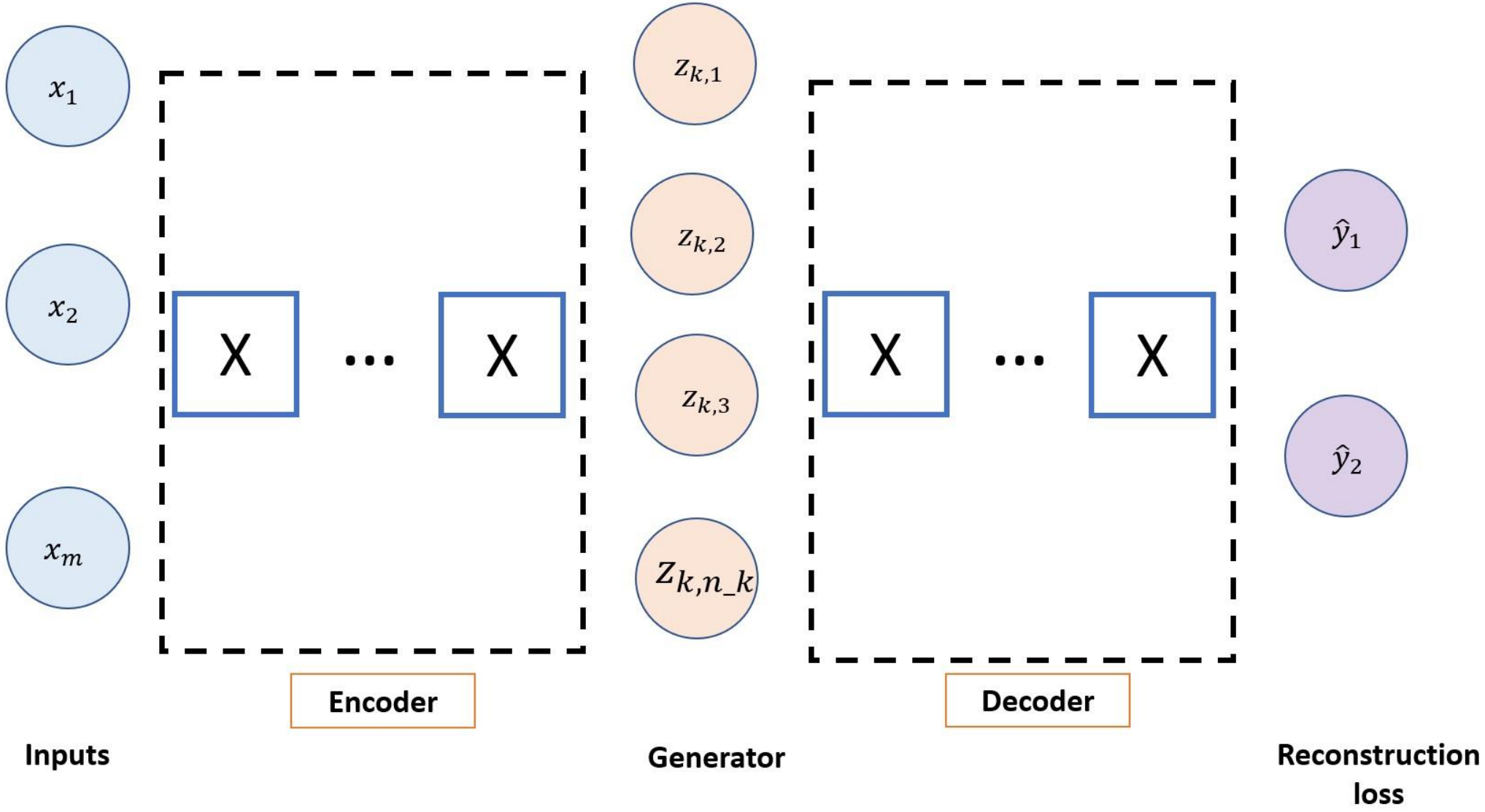

x_1
x_2
x_m
z_1
z_2
$y_1 = g(z_1)$
$y_2 = g(z_2)$


$W^{(1)}$
$W^{(2)}$
x_1
x_2
x_m
z_1
z_2
z_3
z_{d_1}
$g(z_1)$
$g(z_2)$
$g(z_3)$
$g(z_{d\_1})$
$\hat{y}_1$
$\hat{y}_2$
Input
Hidden
Final Output


x_1
x_2
x_m
X
...
X
$z_{k,1}$
$z_{k,2}$
$z_{k,3}$
z_{k,n_k}
X
...
X
$\hat{y}_1$
$\hat{y}_2$
Inputs
Hidden
Output



(Wang et al., 2020)


x_1
x_2
x_m
X
...
X
Encoder
Inputs
$z_{k,1}$
$z_{k,2}$
$z_{k,3}$
z_{k,n_k}
Generator
X
...
X
Decoder
$\hat{y}_1$
$\hat{y}_2$
Reconstruction loss

Definition of Loss Function

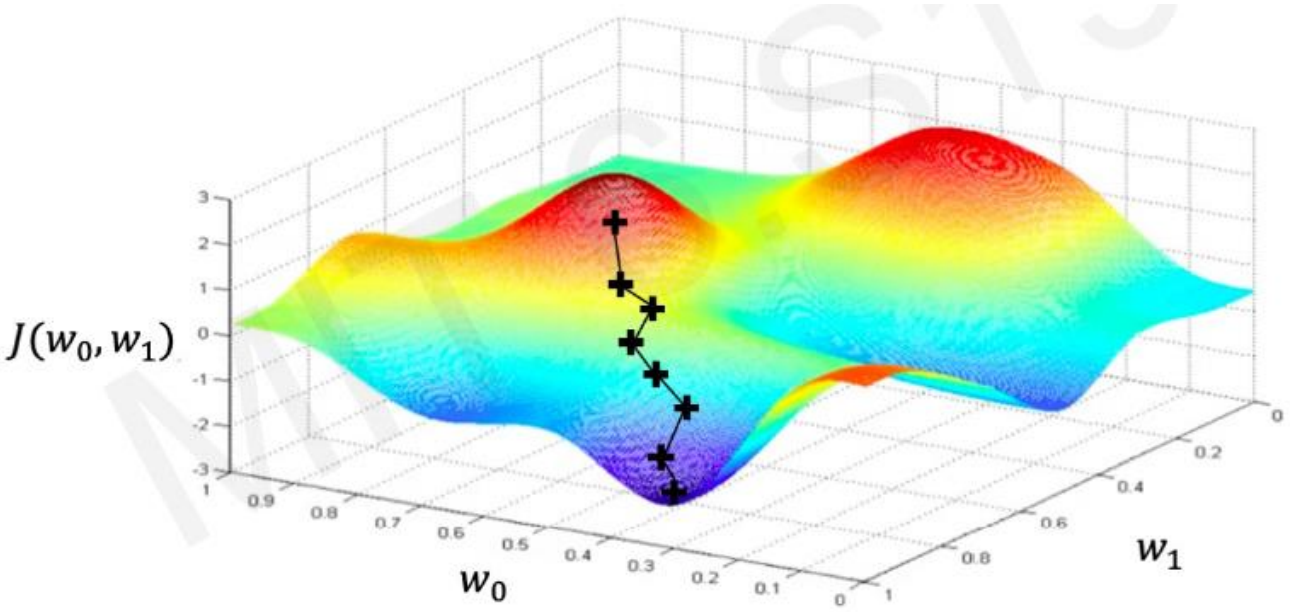
- The loss of our network measures the cost incurred from incorrect predictions

$$L(\underbrace{f\left(x^{(i)}, W\right)}_{\text{Predicted}}, \underbrace{y^{(i)}}_{\text{Actual}})$$

- Empirical Loss: Measures the total loss over our entire dataset
- Binary Cross Entropy Loss: Can be used with models that output a probability between 0 and 1
- Mean Squared Error Loss: Can e used with regression models that output continuous real numbers

Training a Neural Network

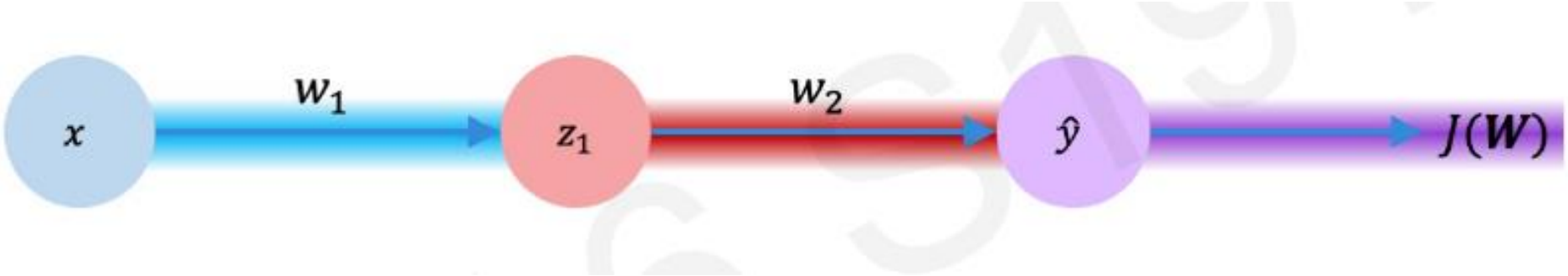
- How can we use the loss function to train weights?
- We want to find the network weights that achieve the lowest loss
- We want to compute the lowest point and compute the gradient (iterative process)



[Amini, 2020]

Computing Gradients: Backpropagation

- How much small change in w is going to affect our loss J



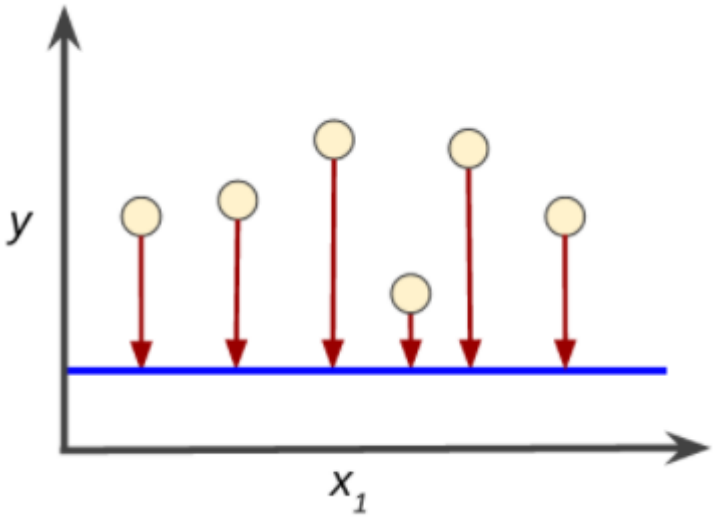
[Amini, 2020]

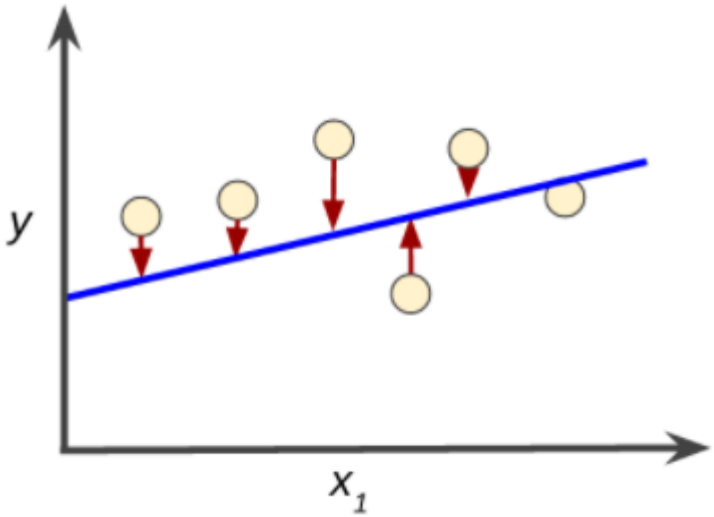
$$\frac{\partial J(\boldsymbol{W})}{\partial w_1} = \frac{\partial J(\boldsymbol{W})}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial z_1} * \frac{\partial z_1}{\partial w_1}$$

Repeat for every weight in the network using gradients from later layers

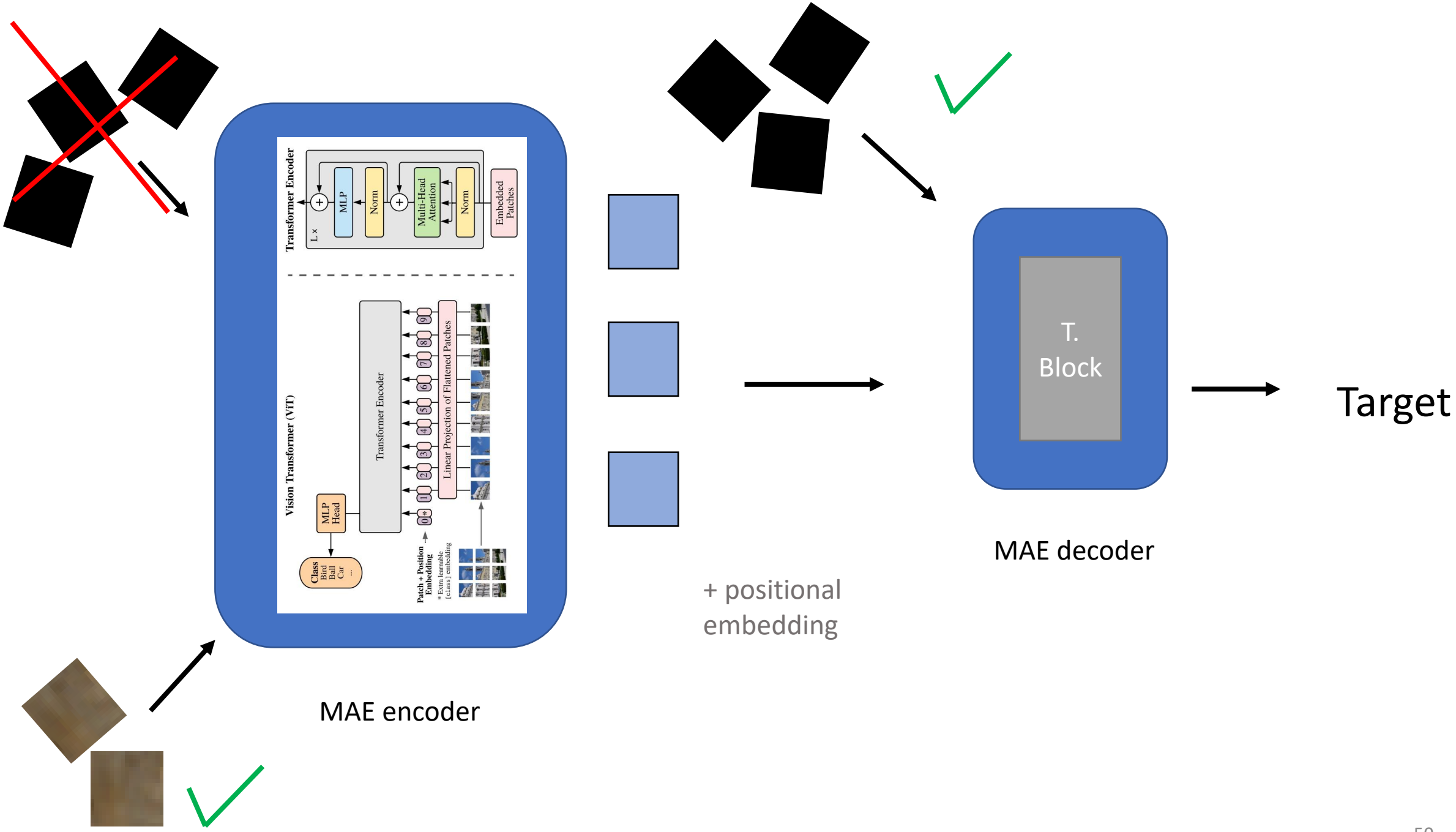
Training loss

- Training a model means determining good values for all the weights and the bias from labeled examples
- Loss is the penalty for a bad prediction
- Loss is a number indicating how bad the model's prediction was on a single example
- If the model's prediction is perfect, the loss is zero, otherwise, the loss is greater
- Goal of training a model is to find a set of weights and biases that have loss, on average, across all examples



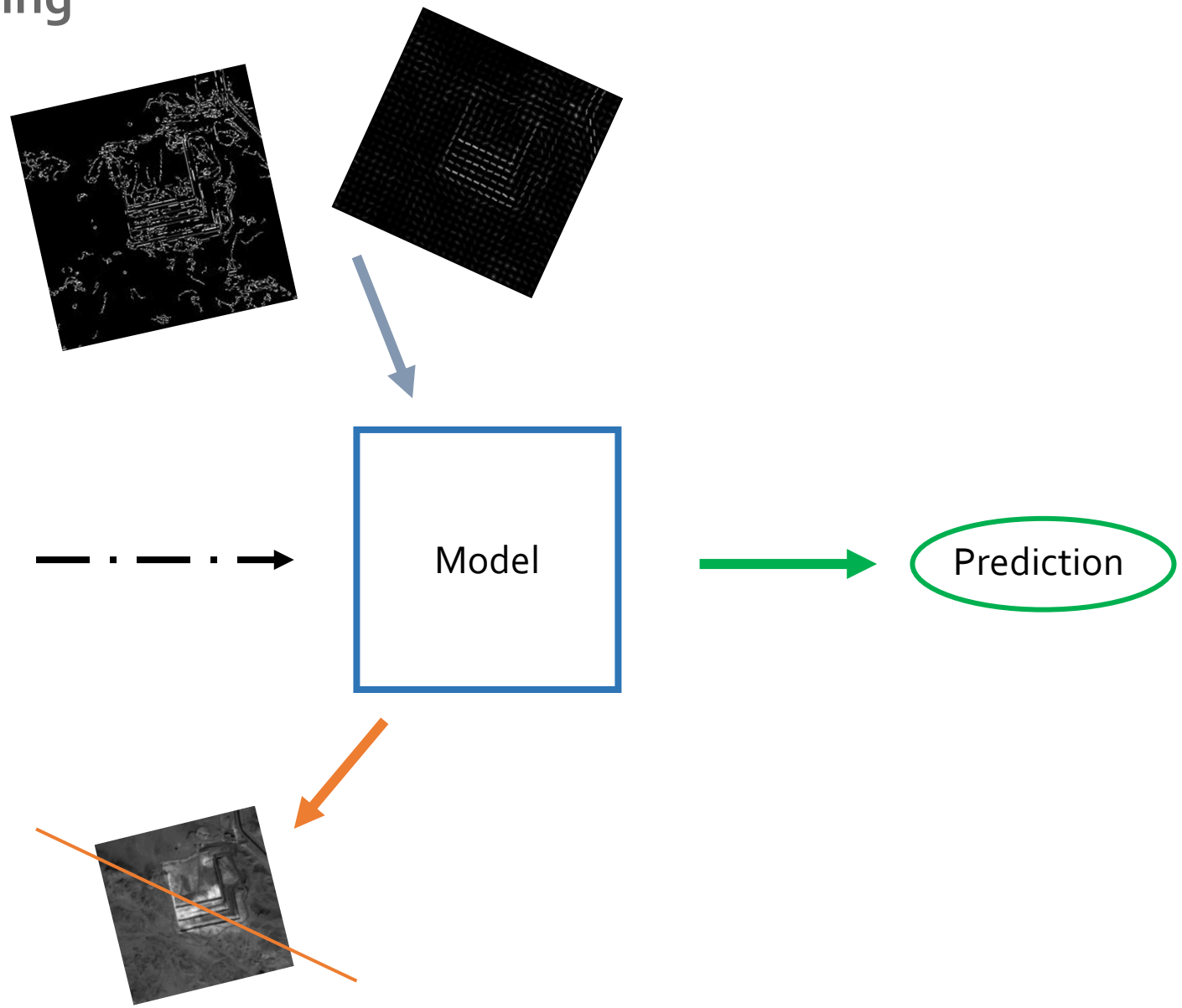


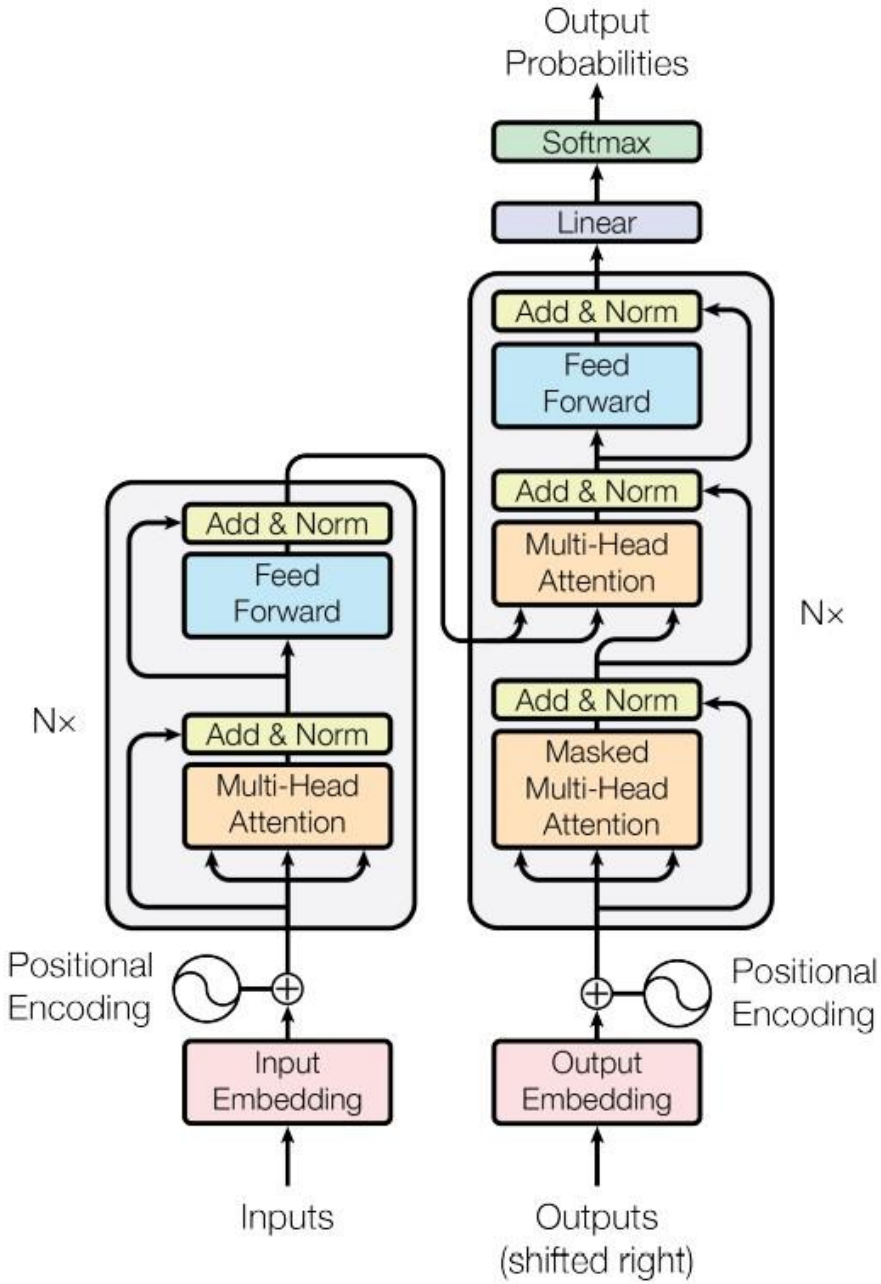
[Machine Learning Crash Course Google, 2020]


Vision Transformer (ViT)
Transformer Encoder
MLP Head
Class
Bird
Ball
Car
...
0*
1
2
3
4
5
6
7
8
9
Patch + Position Embedding
* Extra learnable [class] embedding
Linear Projection of Flattened Patches
Transformer Encoder
L ×
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches
MAE encoder
+ positional embedding
T. Block
MAE decoder
Target

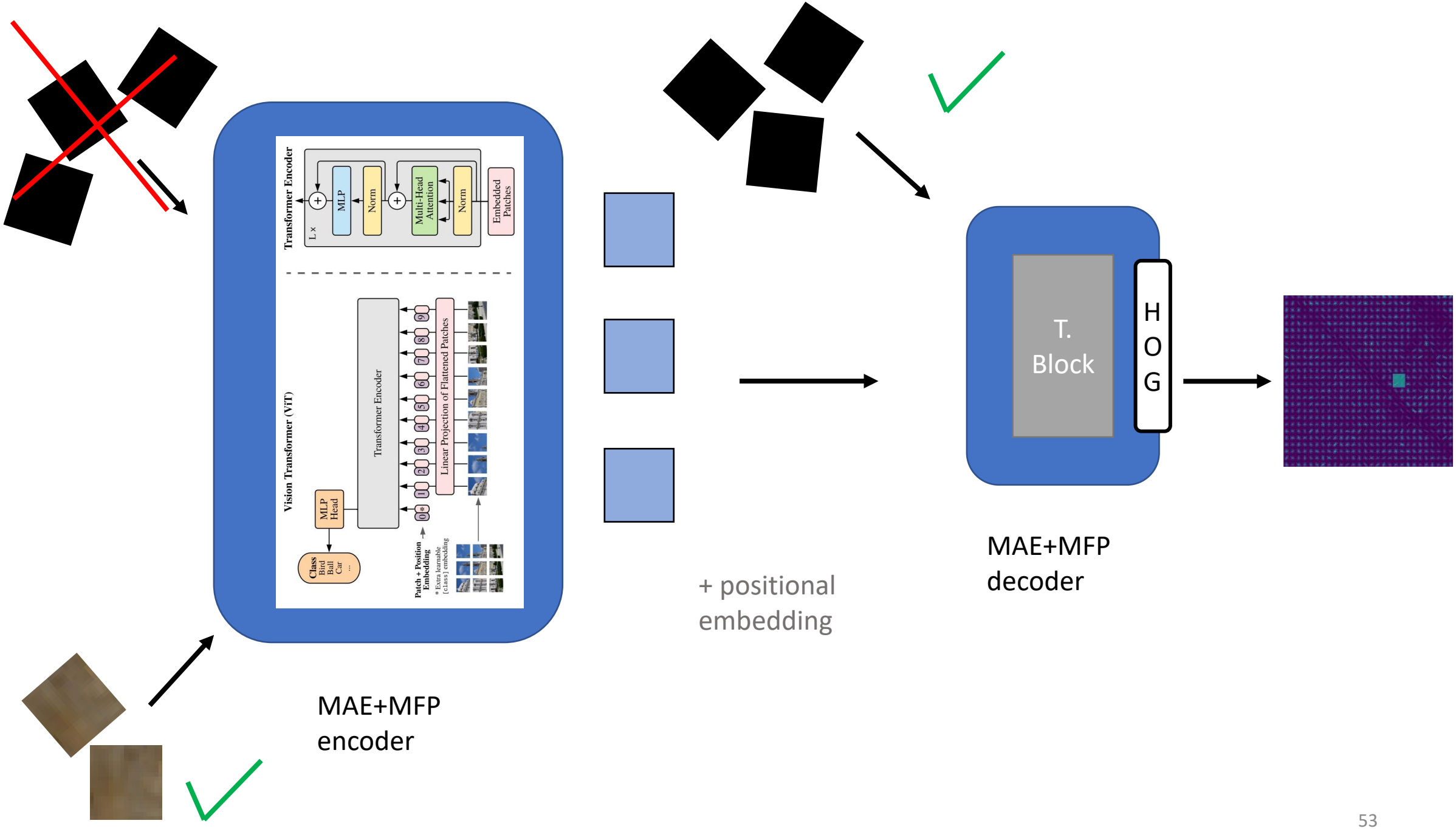
Targets replacement MFP pre-training

- Pretraining: MAE pre-training
- Change of reconstruction targets: MFP pre-training

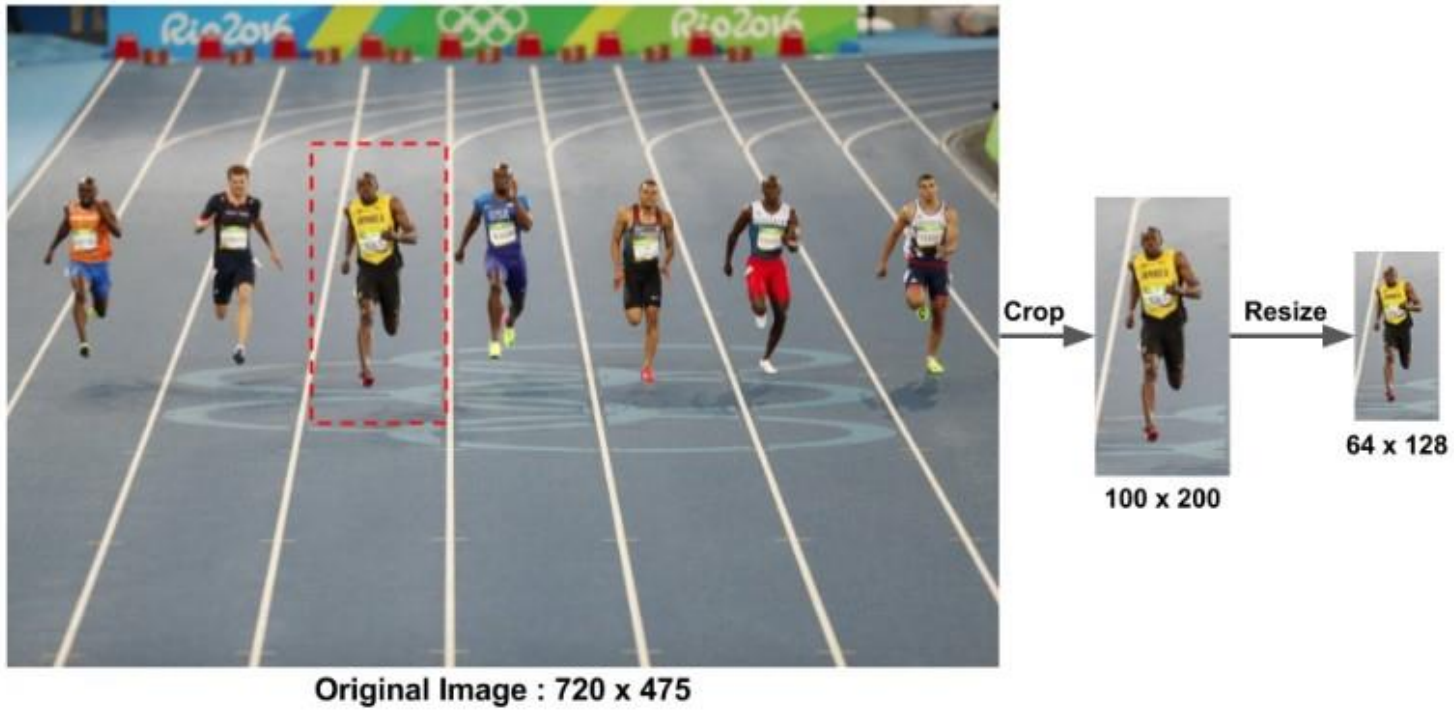




(Vaswani et al., 2017)


Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches
Vision Transformer (ViT)
MLP Head
Class
Bird
Ball
Car
...
Transformer Encoder
Linear Projection of Flattened Patches
Patch + Position Embedding
* Extra learnable [class] embedding
0*
1
2
3
4
5
6
7
8
9
MAE+MFP encoder
+ positional embedding
T. Block
H O G
MAE+MFP decoder

Step 1: Preprocessing



(Satya Mallick, 2016)

Step 2: Calculate the Gradient Images

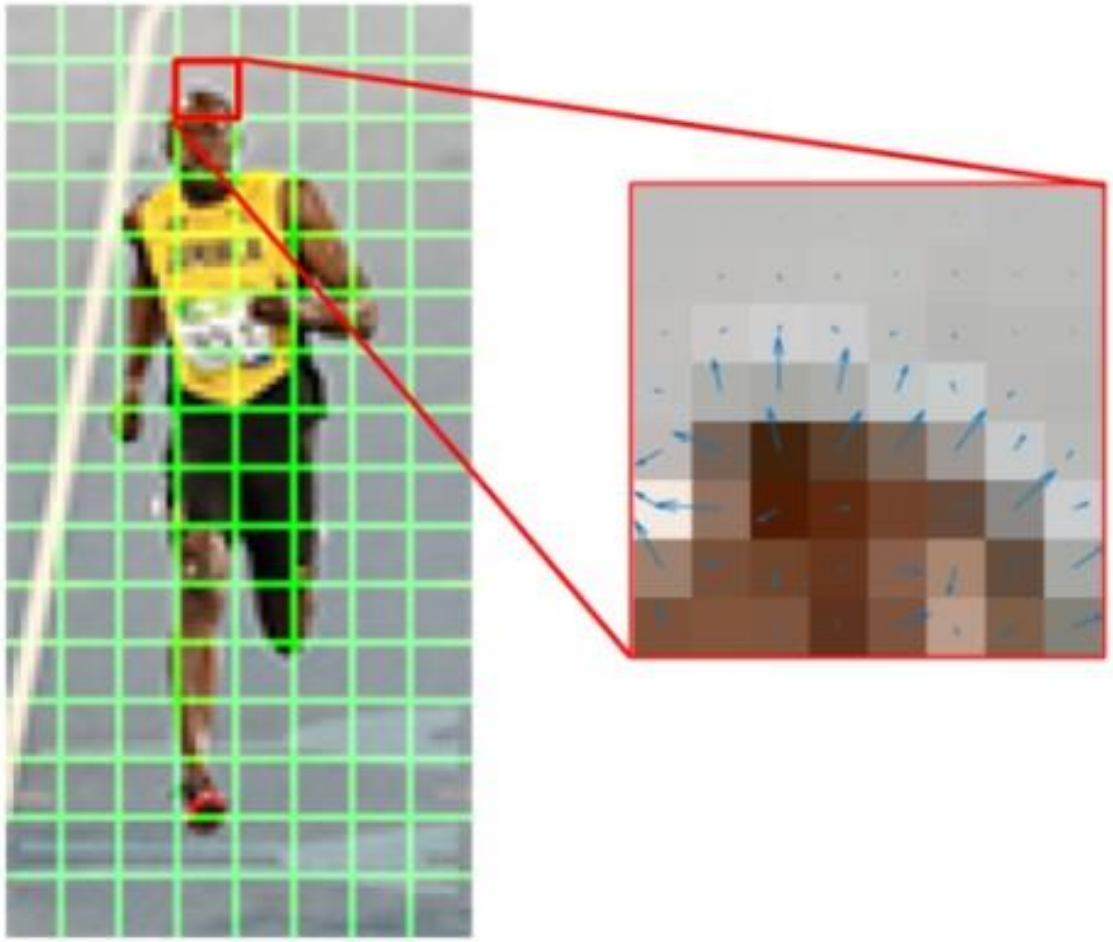
$$g = \sqrt{g_x^2 + g_y^2}$$

$$\theta = arctan\frac{g_y}{g_x}$$

Left : Absolute value of x-gradient. Center : Absolute value of y-gradient. Right : Magnitude of gradient.

(Satya Mallick, 2016)

Step 3: Calculate Histogram of Gradients



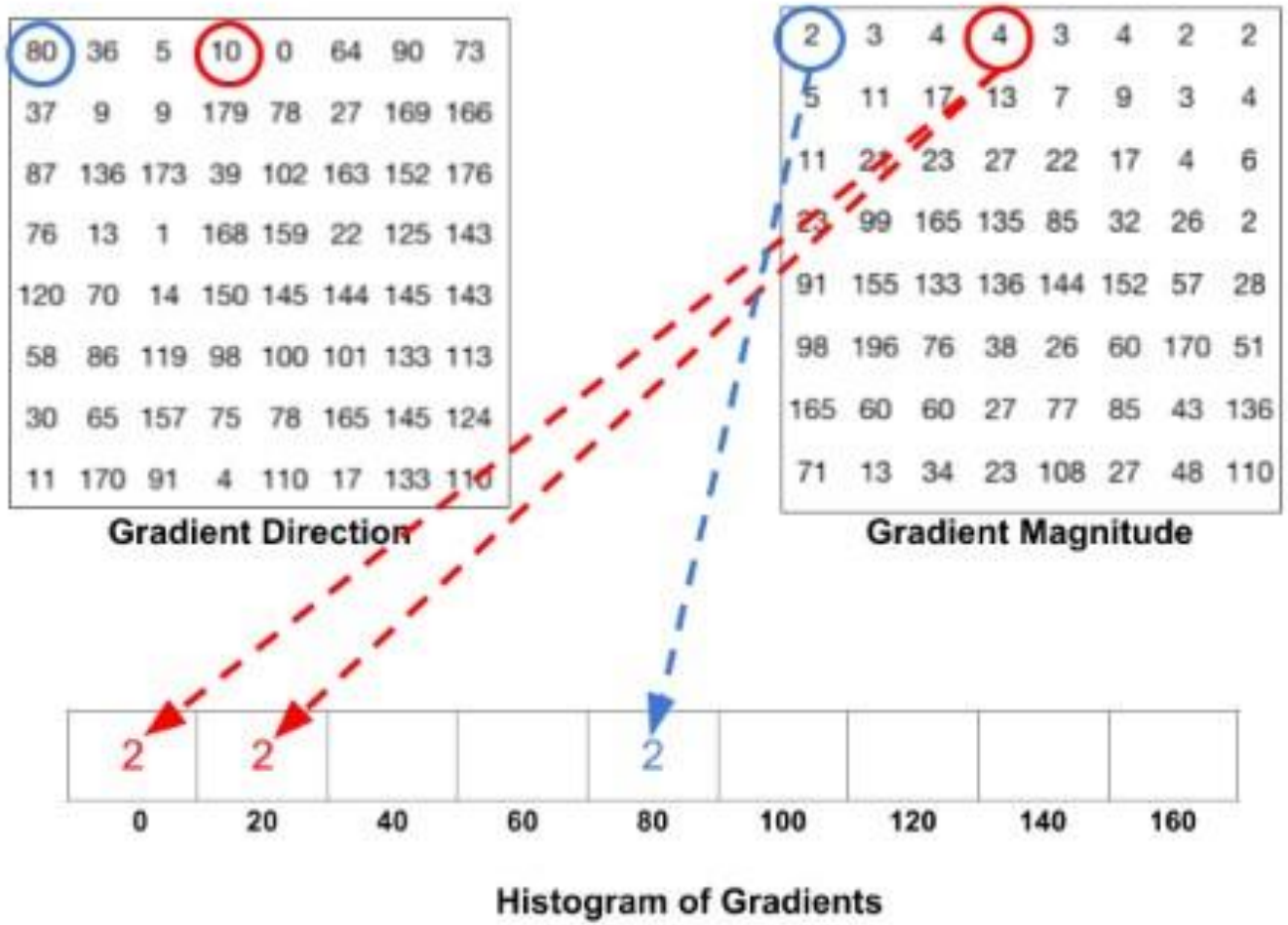
2	3	4	4	3	4	2	2
5	11	17	13	7	9	3	4
11	21	23	27	22	17	4	6
23	99	165	135	85	32	26	2
91	155	133	136	144	152	57	28
98	196	76	38	26	60	170	51
165	60	60	27	77	85	43	136
71	13	34	23	108	27	48	110

Gradient Magnitude

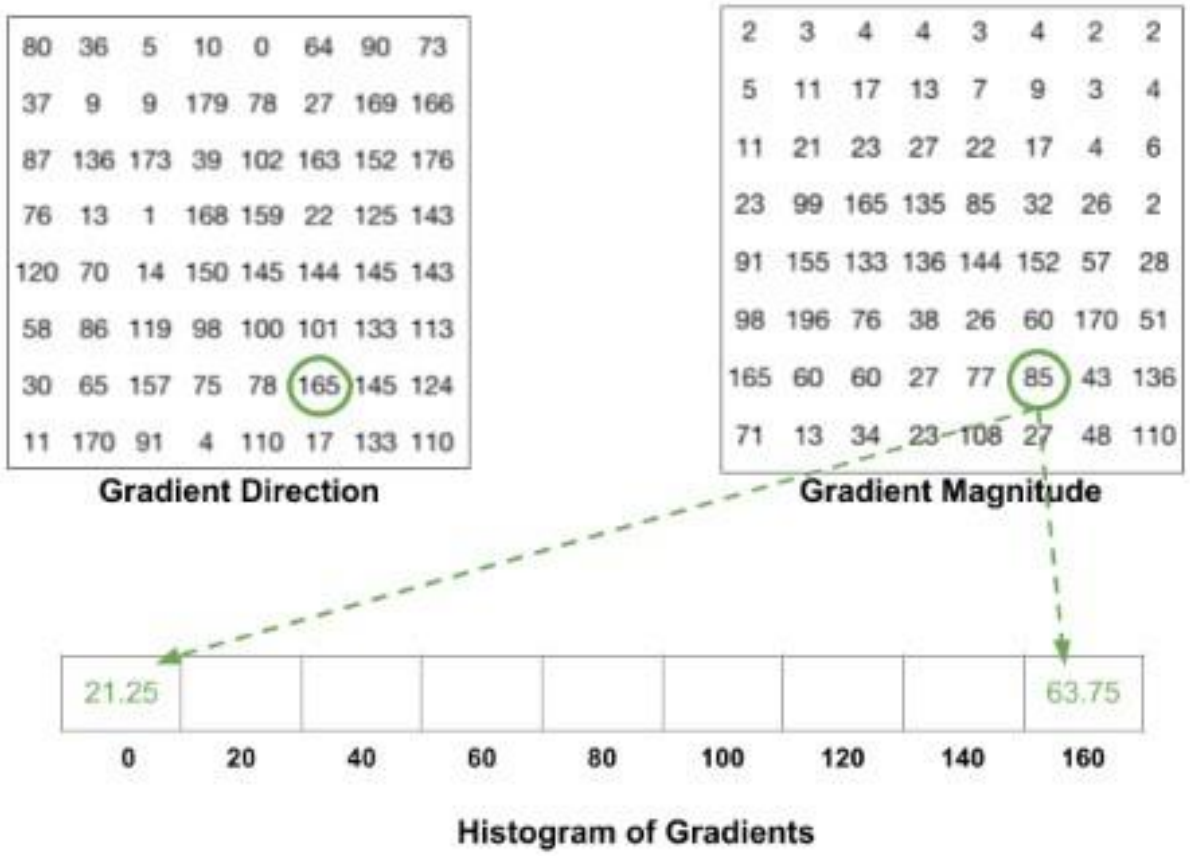
80	36	5	10	0	64	90	73
37	9	9	179	78	27	169	166
87	136	173	39	102	163	152	176
76	13	1	168	159	22	125	143
120	70	14	150	145	144	145	143
58	86	119	98	100	101	133	113
30	65	157	75	78	165	145	124
11	170	91	4	110	17	133	110

Gradient Direction

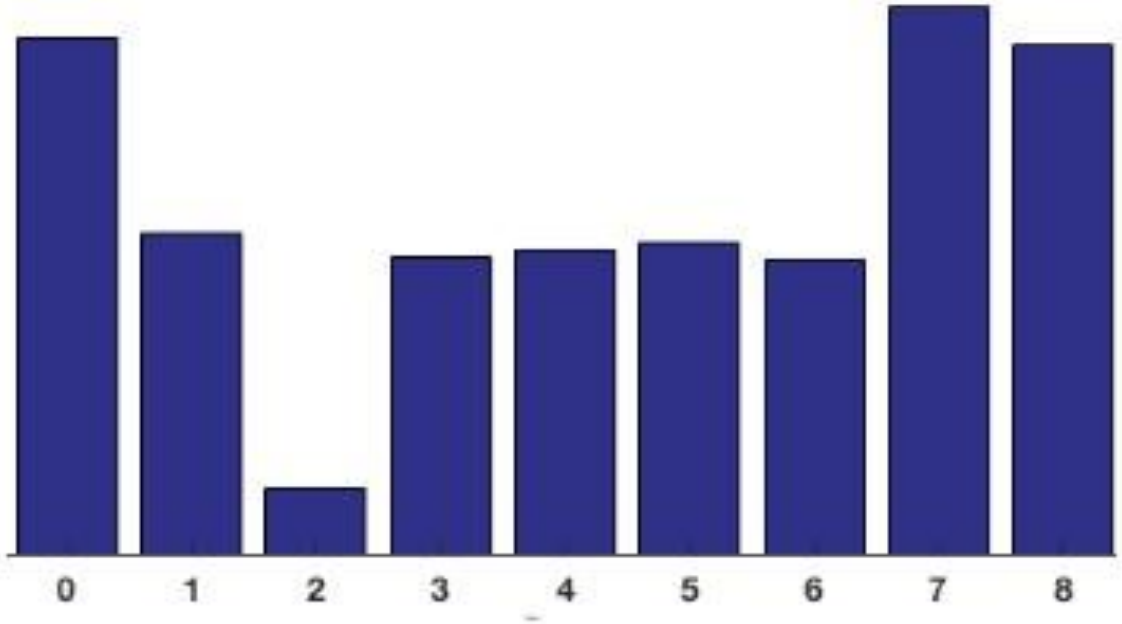
(Satya Mallick, 2016)



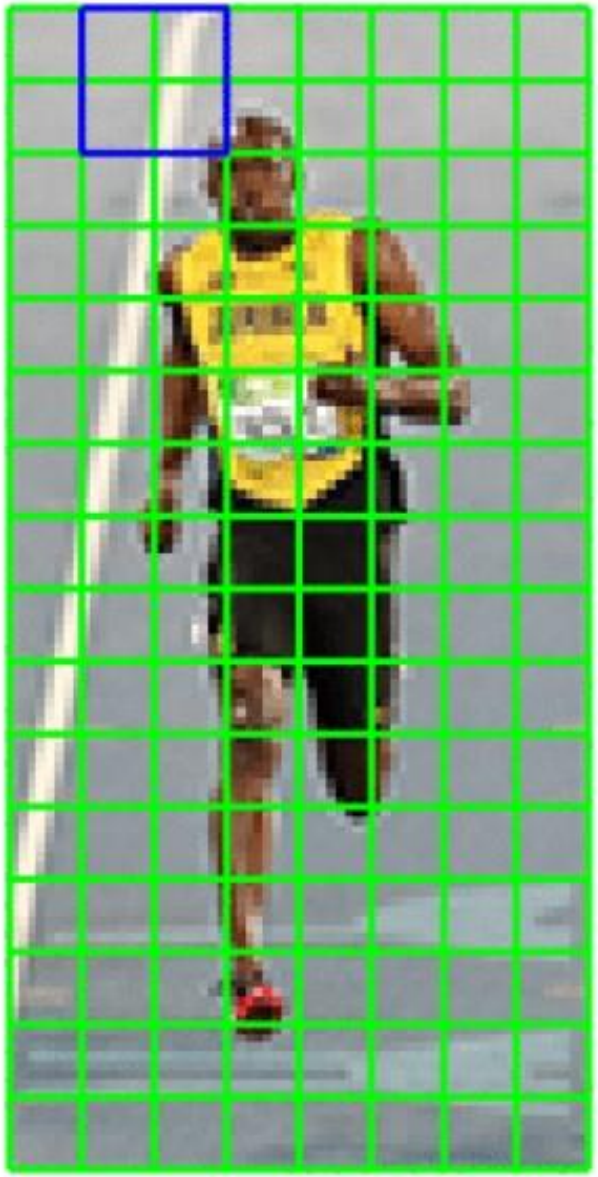
(Satya Mallick, 2016)



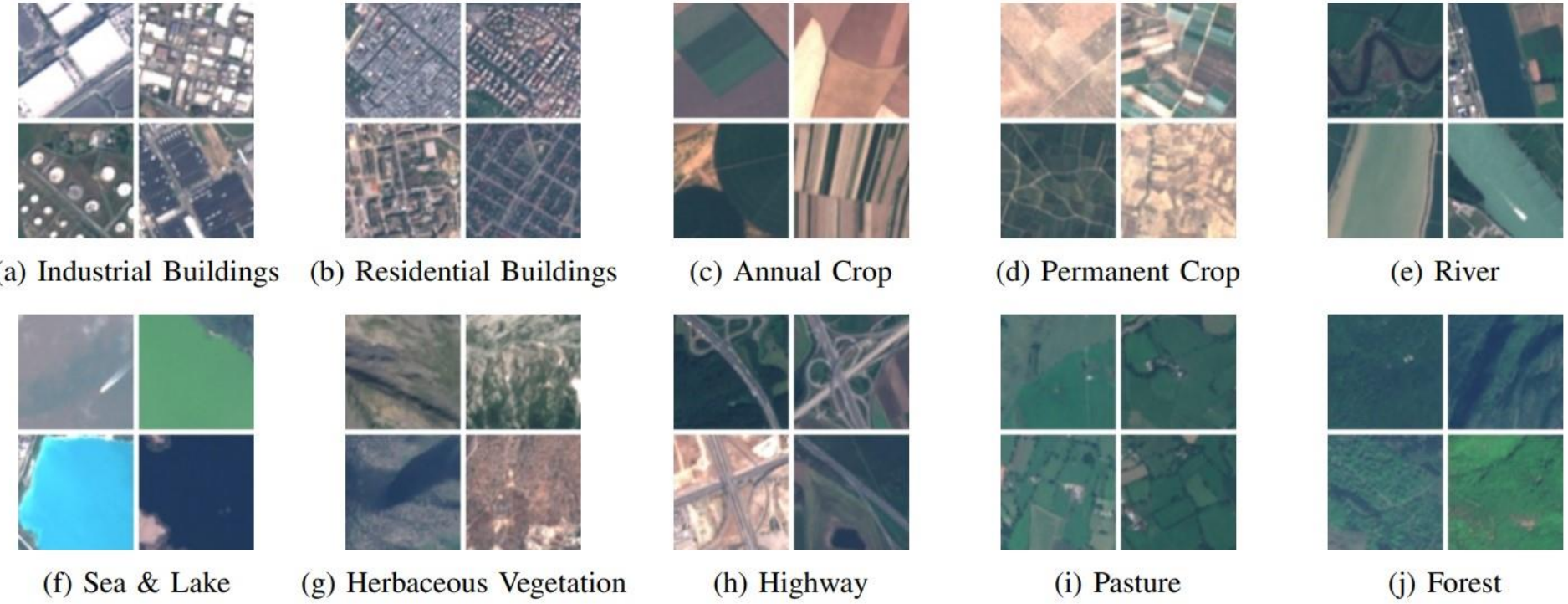
(Satya Mallick, 2016)



(Satya Mallick, 2016)



(Satya Mallick, 2016)



(a) Industrial Buildings (b) Residential Buildings (c) Annual Crop (d) Permanent Crop (e) River

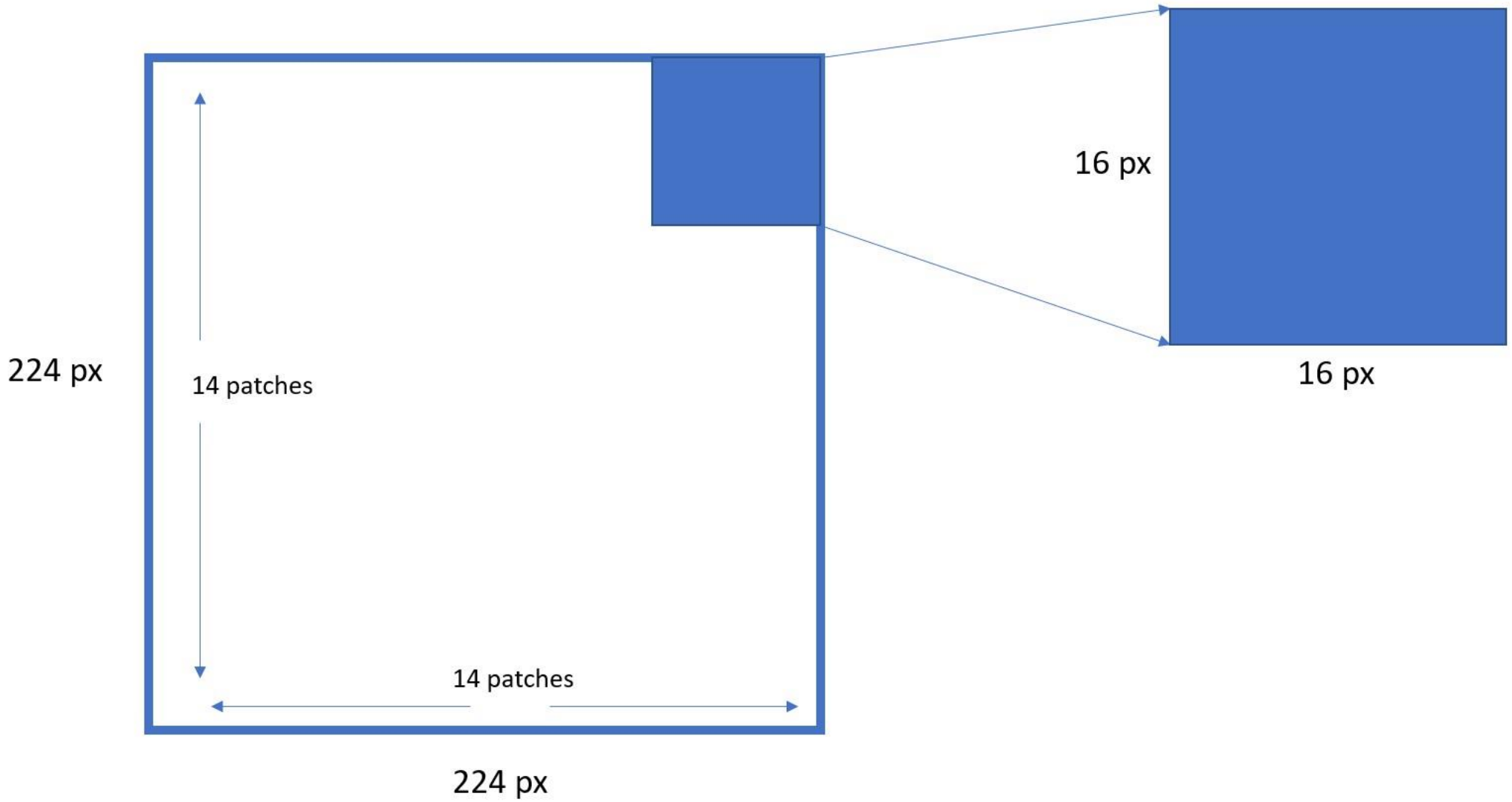
(f) Sea & Lake (g) Herbaceous Vegetation (h) Highway (i) Pasture (j) Forest

(Helber. P, et.al, 2015)

13 Bands covered by Sentinel's 2 Multispectral band

Band	Spatial Resolution m	Central Wavelength nm
B01 - Aerosols	60	443
B02 - Blue	10	490
B03 - Green	10	560
B04 - Red	10	665
B05 - Red edge 1	20	705
B06 - Red edge 2	20	740
B07 - Red edge 3	20	783
B08 - NIR	10	842
B08A - Red edge 4	20	865
B09 - Water vapor	60	945
B10 - Cirrus	60	1375
B11 - SWIR 1	20	1610
B12 - SWIR 2	20	2190

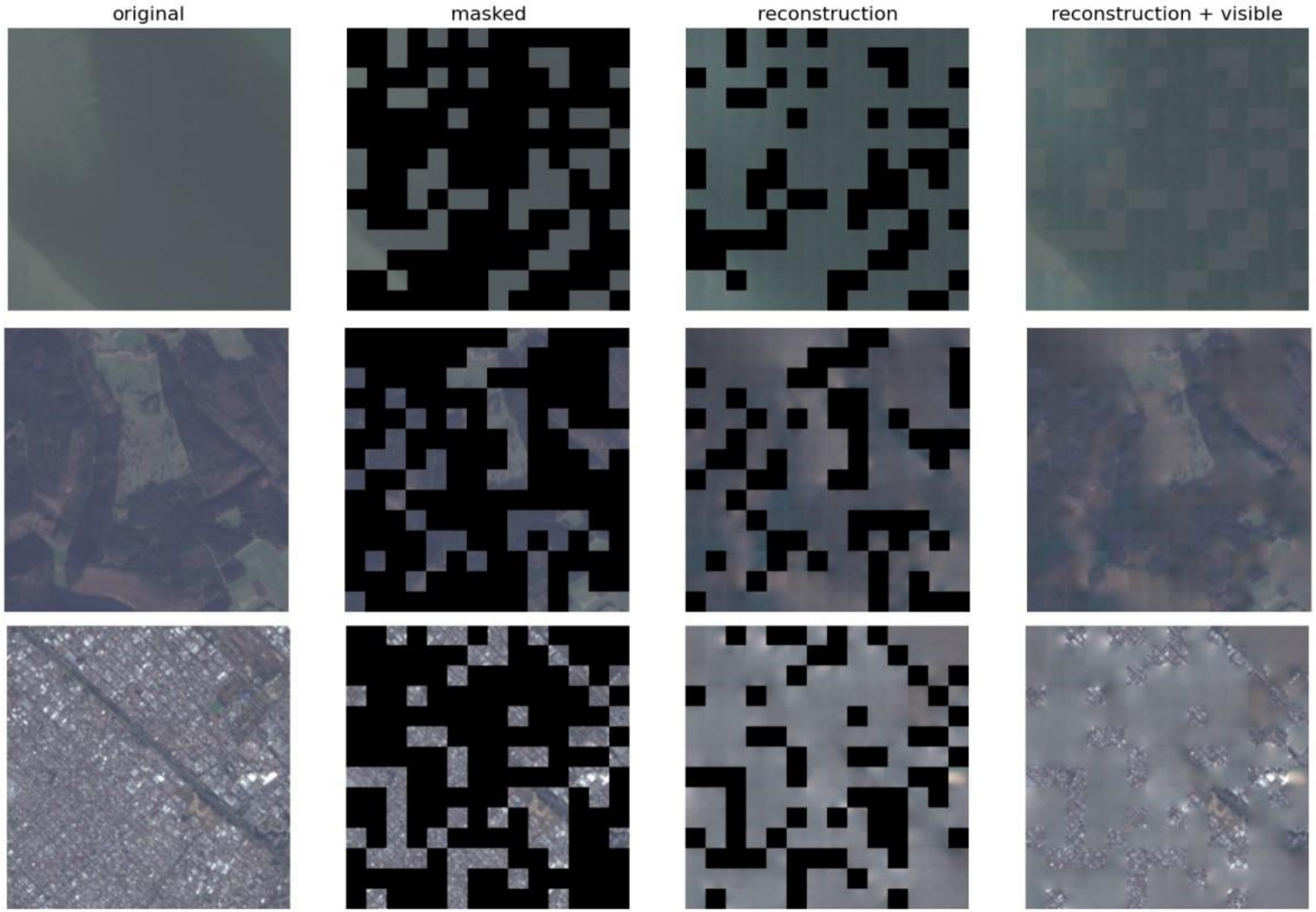
(Helber. P, et.al, 2015)

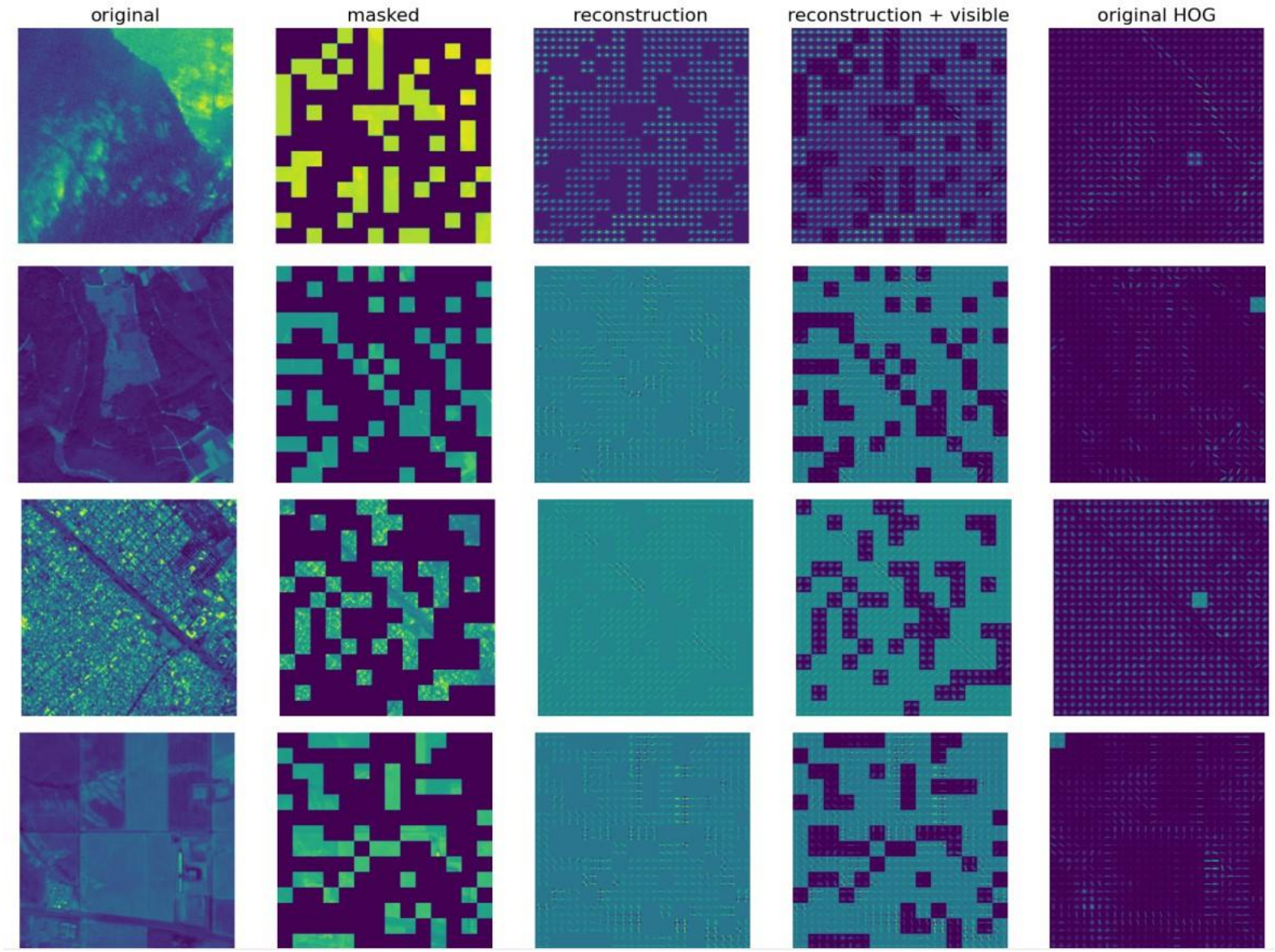

16 px
16 px
224 px
14 patches
14 patches
224 px

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

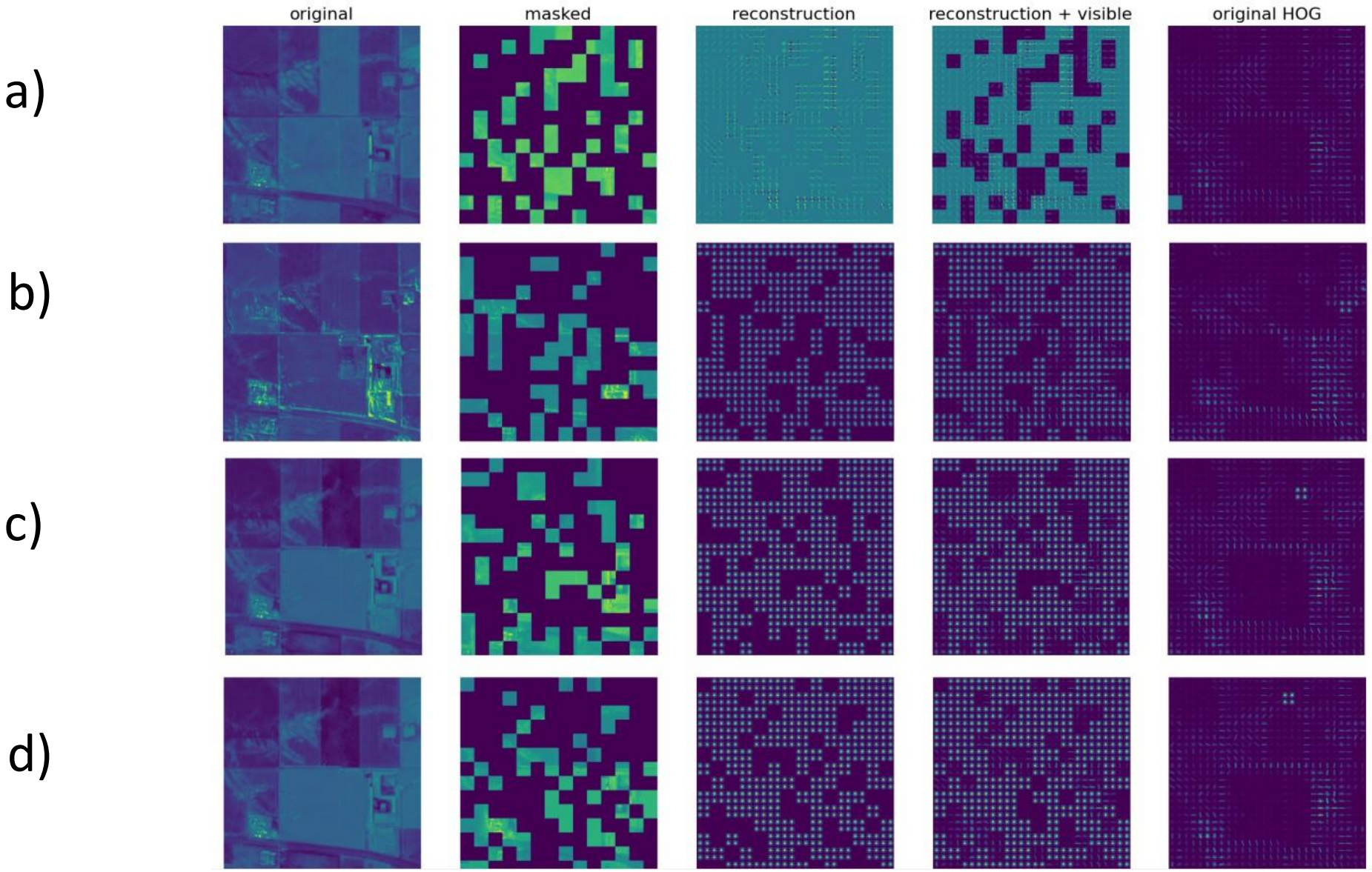
Table 1: Details of Vision Transformer model variants.

Model	**Layers**	**Hidden size D**	**Heads**	**Image size (px)**
ViT small	12	384	6	224×224

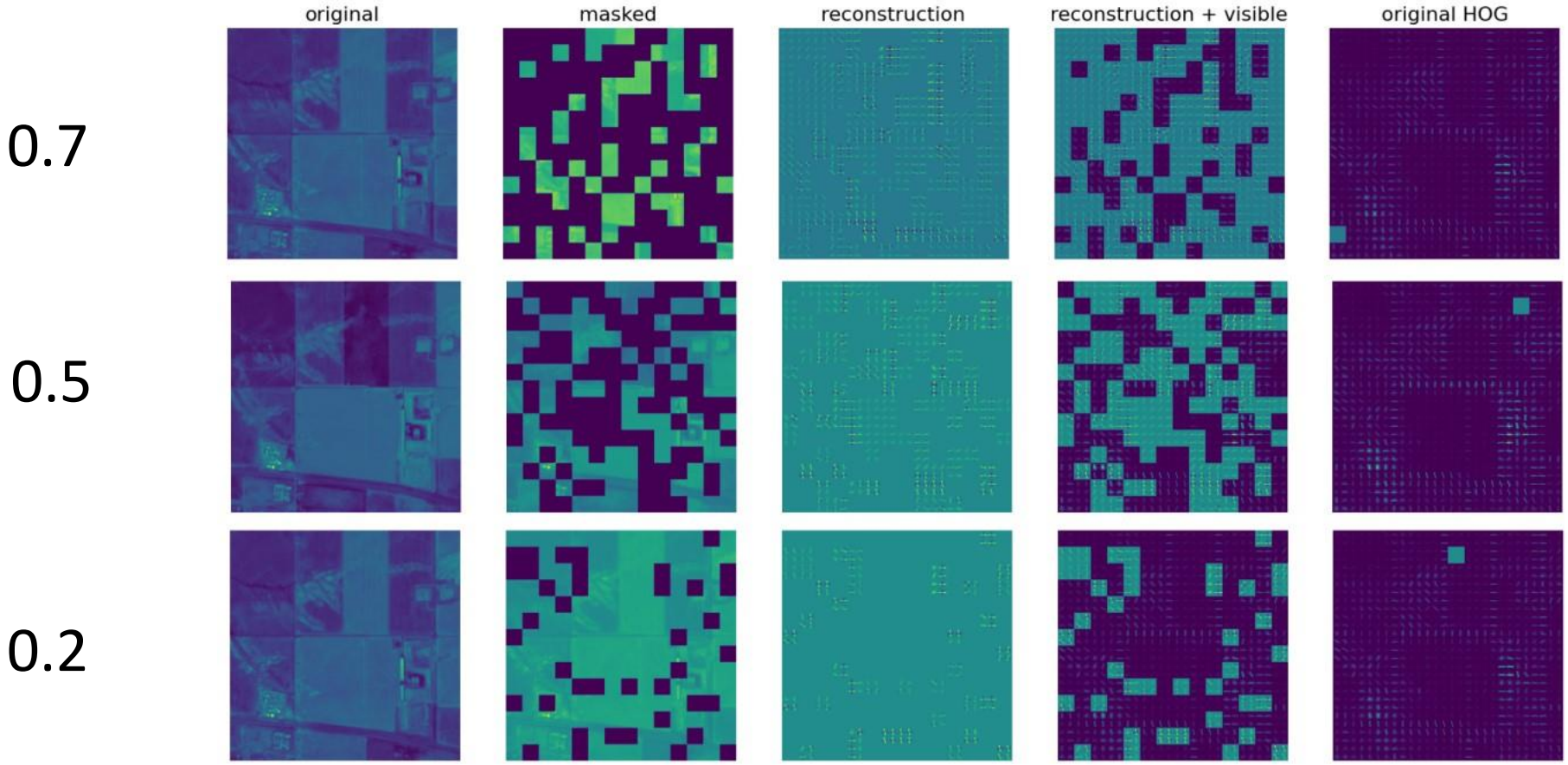

original
masked
reconstruction
reconstruction + visible

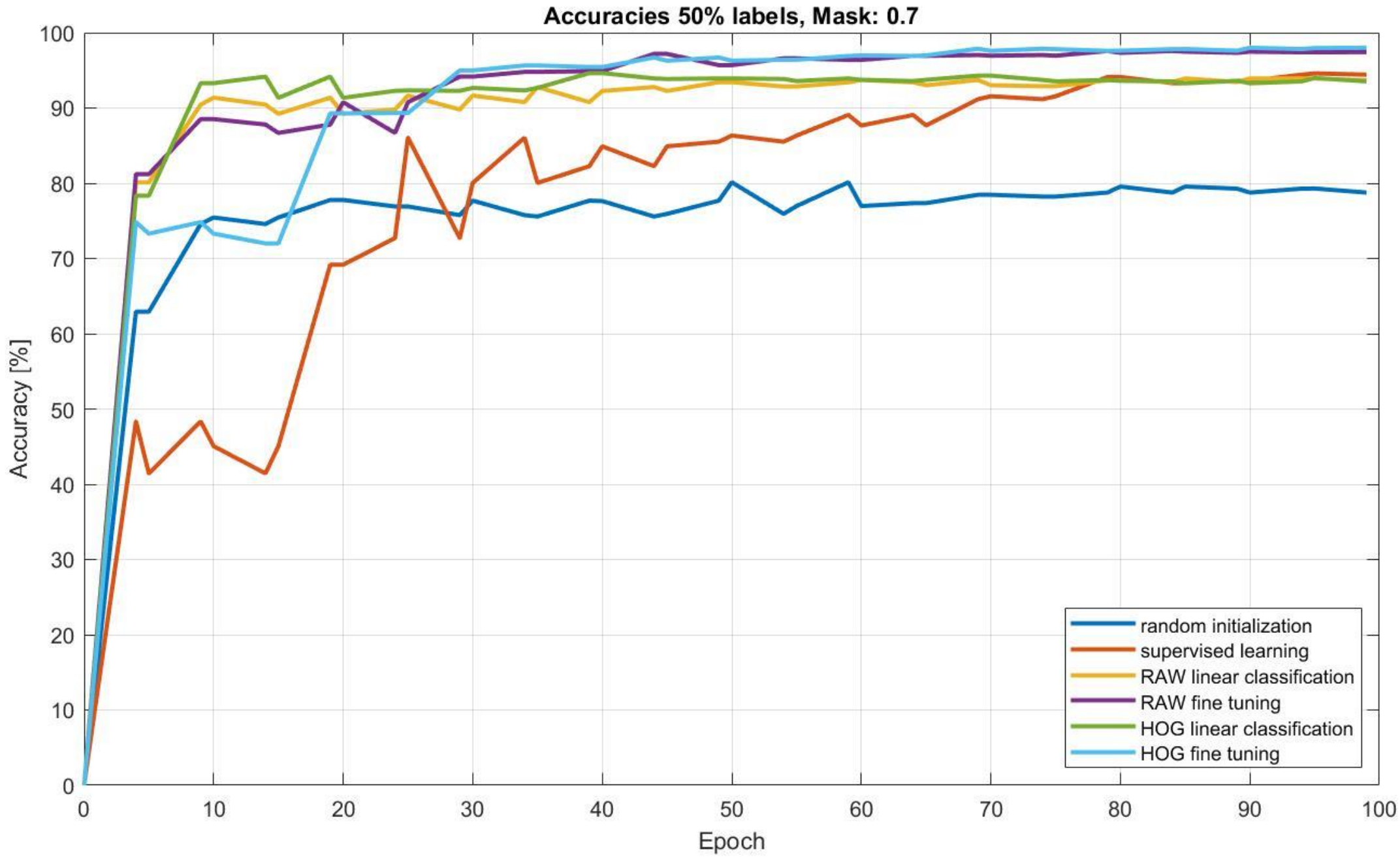

original
masked
reconstruction
reconstruction + visible
original HOG

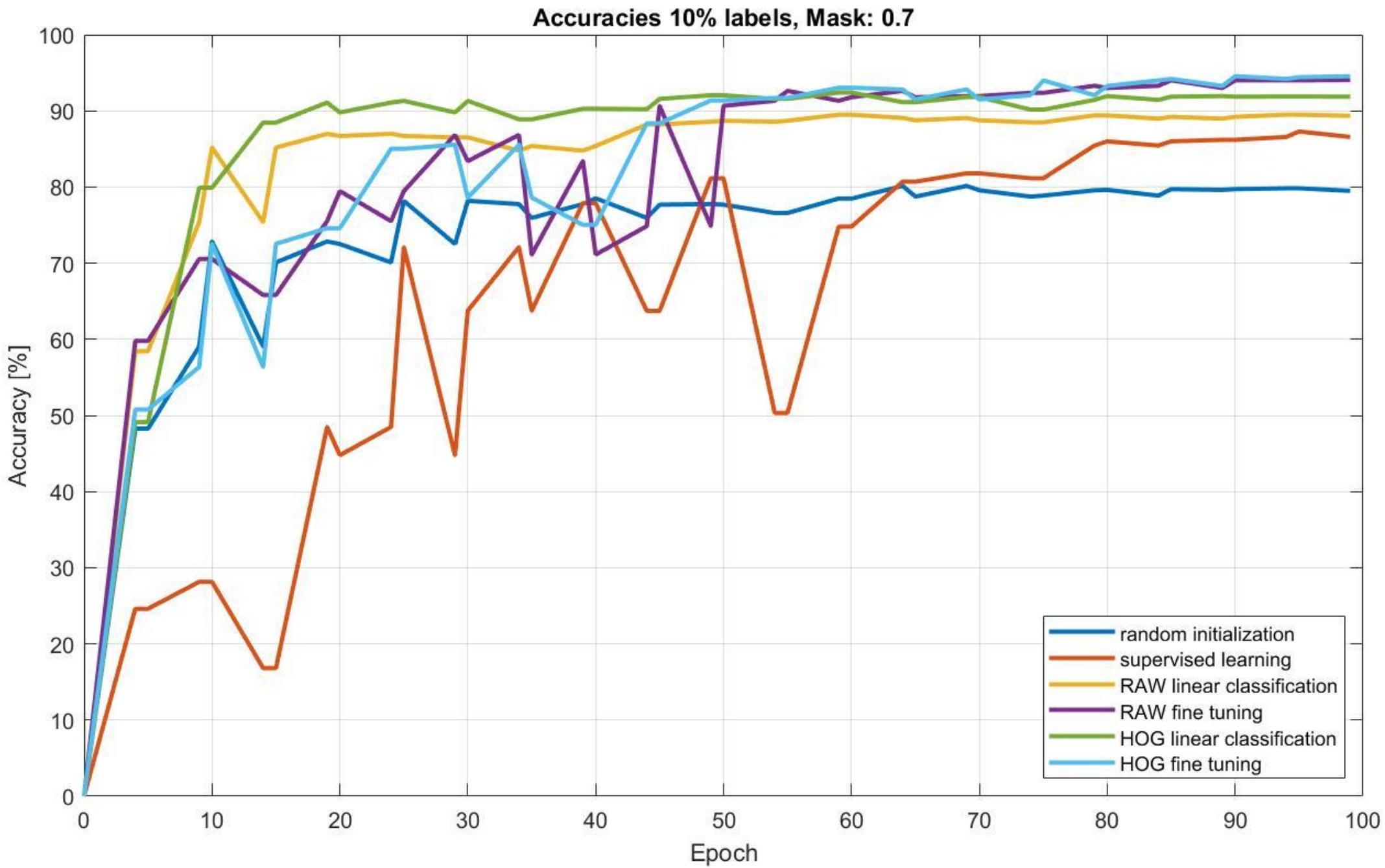
Normalization variation reconstruction

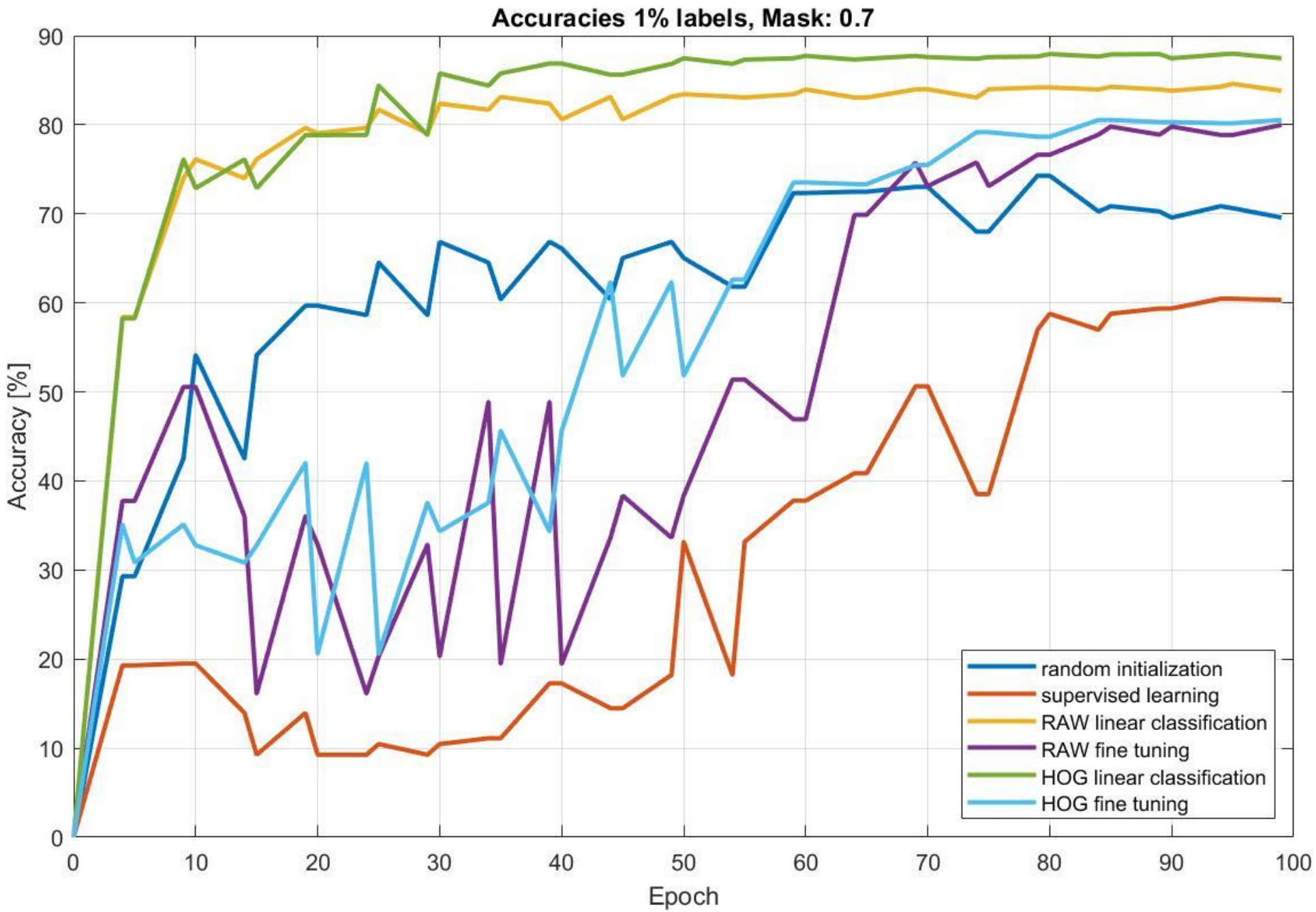


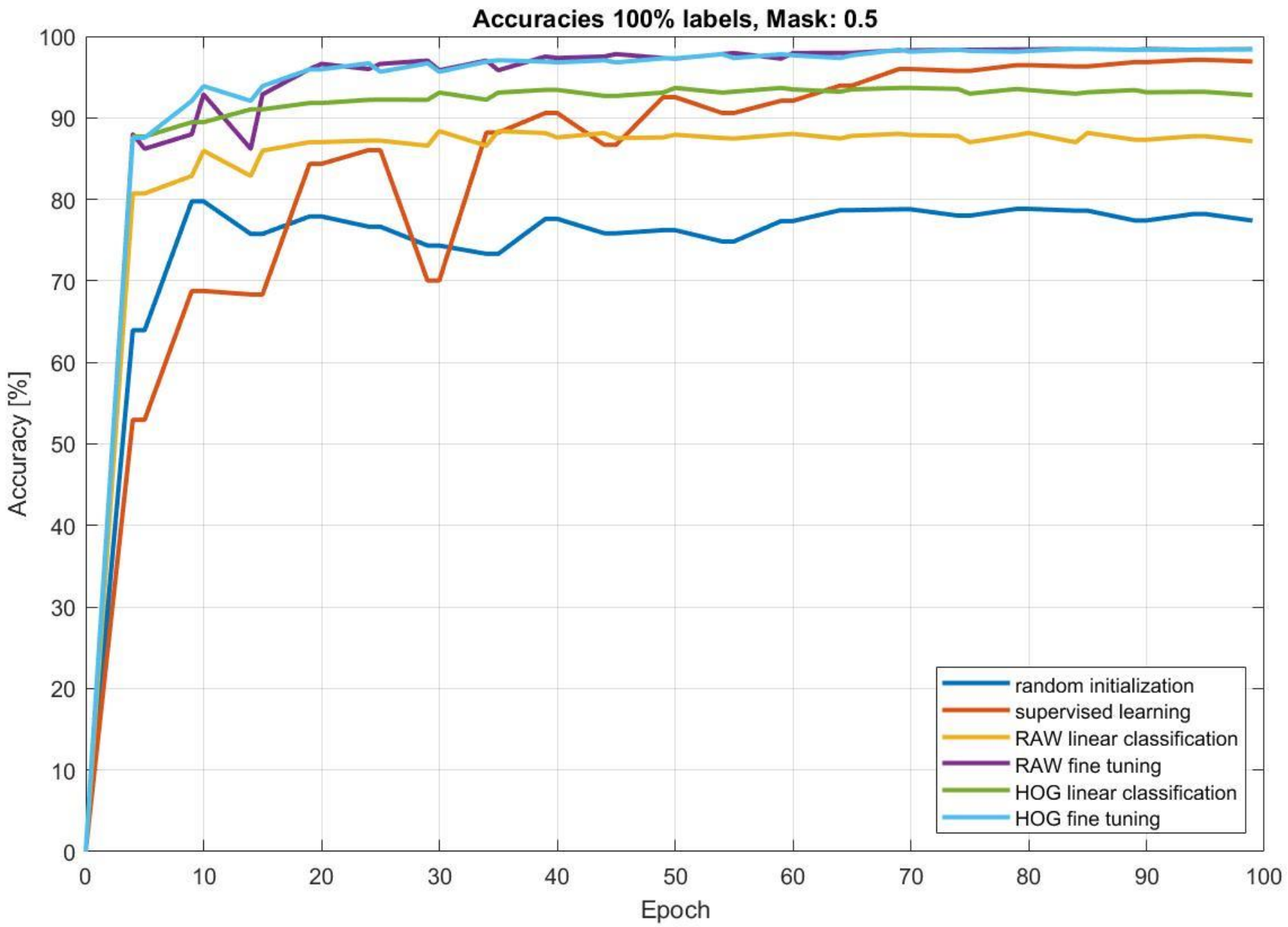
Masking variation reconstruction




Accuracies 50% labels, Mask: 0.7
Accuracy [%]
Epoch
0
10
20
30
40
50
60
70
80
90
100
random initialization
supervised learning
RAW linear classification
RAW fine tuning
HOG linear classification
HOG fine tuning


Accuracies 10% labels, Mask: 0.7
Accuracy [%]
Epoch
0
10
20
30
40
50
60
70
80
90
100
random initialization
supervised learning
RAW linear classification
RAW fine tuning
HOG linear classification
HOG fine tuning


Accuracies 1% labels, Mask: 0.7
Accuracy [%]
Epoch
0
10
20
30
40
50
60
70
80
90
100
random initialization
supervised learning
RAW linear classification
RAW fine tuning
HOG linear classification
HOG fine tuning


Accuracies 100% labels, Mask: 0.5
Accuracy [%]
Epoch
0
10
20
30
40
50
60
70
80
90
100
random initialization
supervised learning
RAW linear classification
RAW fine tuning
HOG linear classification
HOG fine tuning

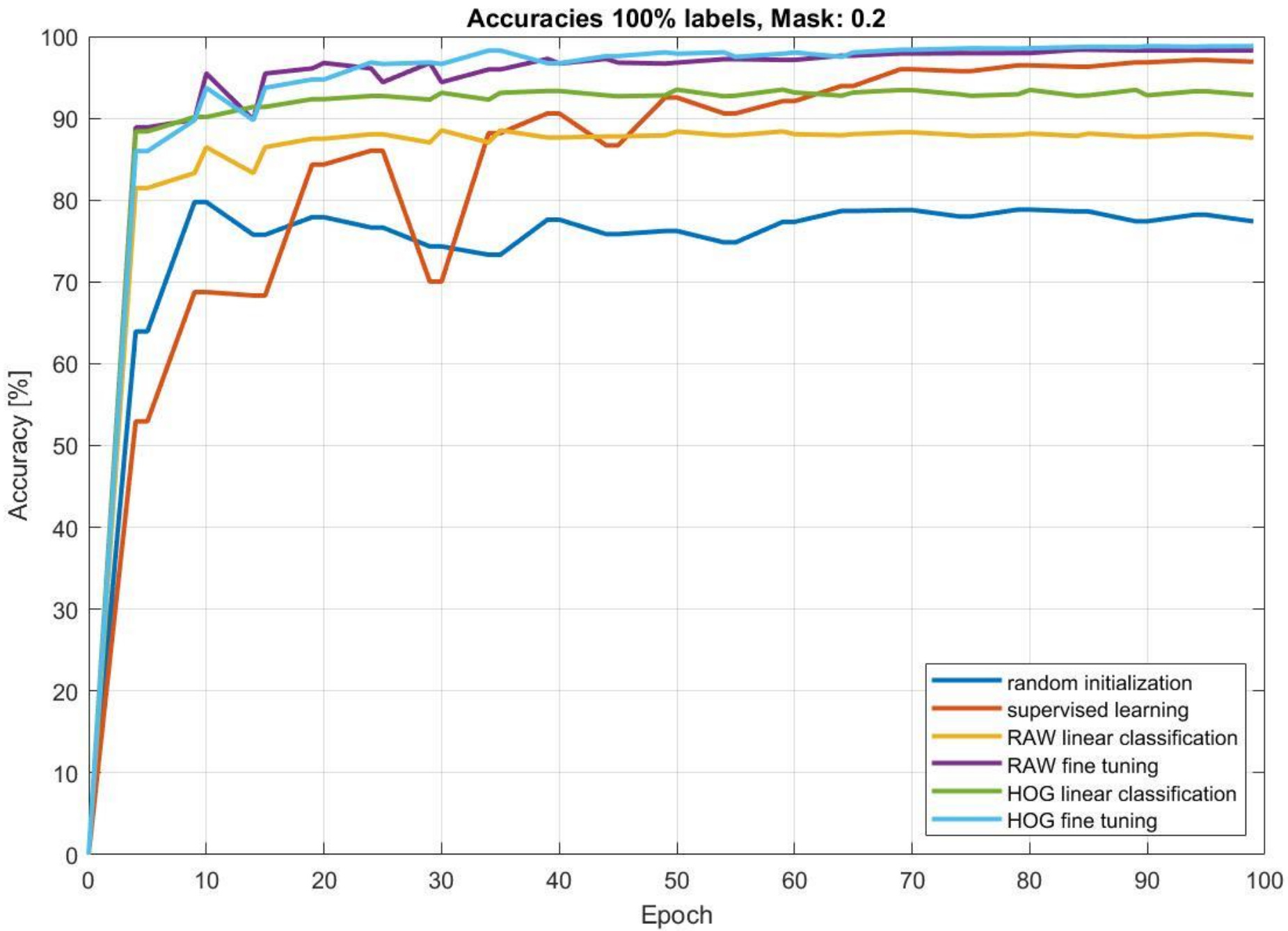

Accuracies 100% labels, Mask: 0.2
Accuracy [%]
Epoch
0
10
20
30
40
50
60
70
80
90
100
random initialization
supervised learning
RAW linear classification
RAW fine tuning
HOG linear classification
HOG fine tuning

Table 5.5 Confusion Matrix: Accuracies per class for 100% of the labels

-	RAW norm	HOG norm+norm.hog	Sup. Learning
Annual Crop	99.0	98.83	97.16
Forest	99.5	99.83	99.67
Herbaceous Vegetation	98.16	98.67	98.16
Highway	97.4	97.6	94.6
Industrial	97.8	90.0	97.6
Pasture	99.0	98.0	96.0
Permanent Crop	98.2	97.4	97.0
Residential	99.5	99.5	98.67
River	99.0	98.8	98.0
Sea Lake	100.0	99.67	99.0

Table 5.6 Confusion Matrix: Accuracies per class for 10% of the labels

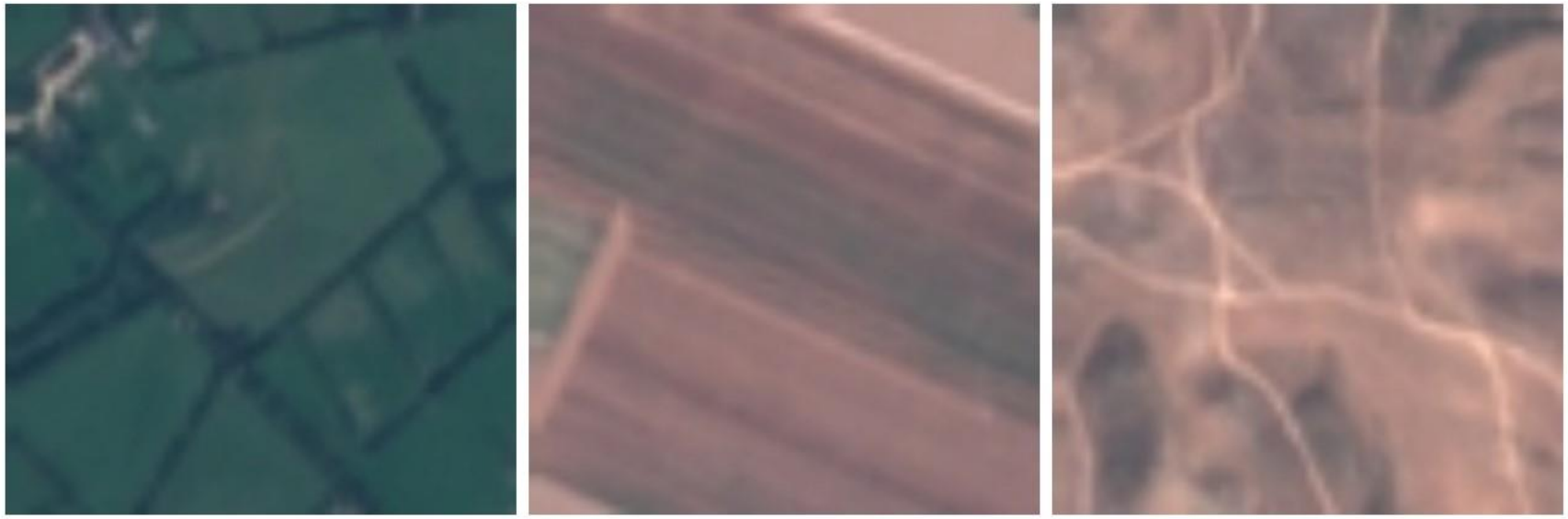
-	RAW norm	HOG norm+norm.hog	Sup. Learning
Annual Crop	91.83	93.0	87.33
Forest	99.0	99.17	96.67
Herbaceous Vegetation	94.5	95.83	91.5
Highway	87.0	87.4	75.2
Industrial	96.8	98.0	93.4
Pasture	93.75	94.0	89.25
Permanent Crop	89.8	90.2	79.2
Residential	95.67	98.0	90.83
River	97.4	97.0	95.6
Sea Lake	95.17	97.33	96.17

(a) (b) (c)

Figure 5.22 (a) Highway misclassified as permanent crop, (b) Highway misclasified as permanent crop, (c) Highway misclassified as Industrial zone

(a) (b) (c)

Figure 5.23 (a) Industrial zone misclassifeid as river, (b) Residential zone misclassified as permanent crop, (c) Industrial zone misclassified as highway

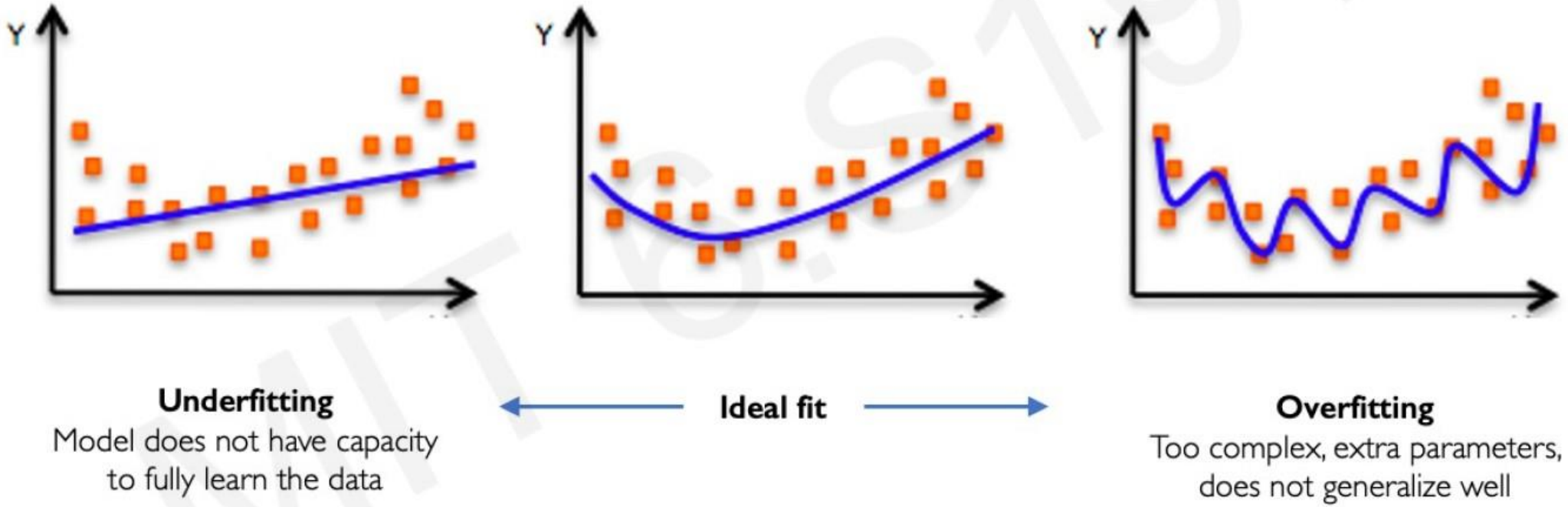


(a) (b) (c)

Figure 5.24 (a) Pasture misclassified as Herb. Vegetation, (b) Annual crop misclassified as permanent crop, (c) Herb. vegetation misclassified as permanent crop

(a) (b) (c)

Figure 5.25 (a) Permanent crop misclassified as highway, (b) Sea lake misclassified as river, (c) River misclassified as highway

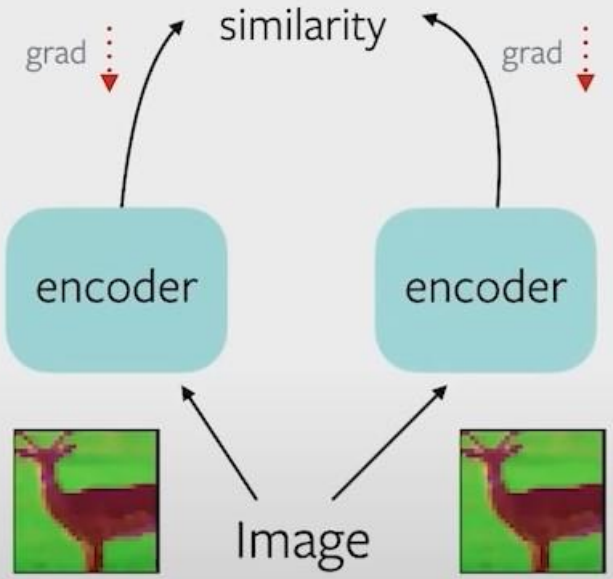


(Alexander Amini, 2022)

Trivial Solutions

$$f_\theta(I) = f_\theta(\text{augment}(I))$$

$$f_\theta(I) = \text{constant}$$



Satisfies the invariance property, but not useful

(Ishan Misra, 2021)