Self Supervised Learning

Bird Images Clustering

Phiphat Chomchit

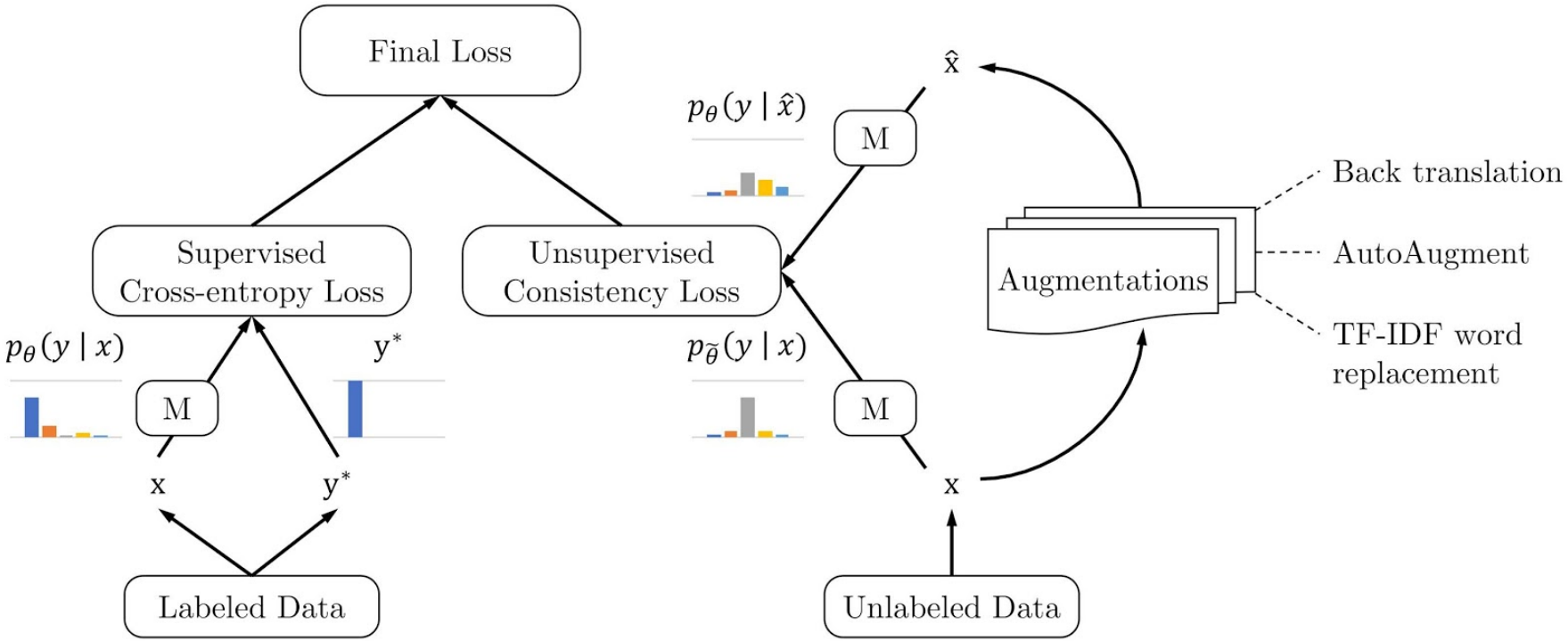
CMU

Self Supervised Learning?

Contrastive self-supervised learning has outperformed supervised pretraining on many downstream tasks like segmentation and object detection.

What if we can get labels for free for unlabelled data and train unsupervised dataset in a supervised manner? We can achieve this by framing a supervised learning task in a special form to predict only a subset of information using the rest. In this way, all the information needed, both inputs and labels, has been provided. This is known as self-supervised learning.

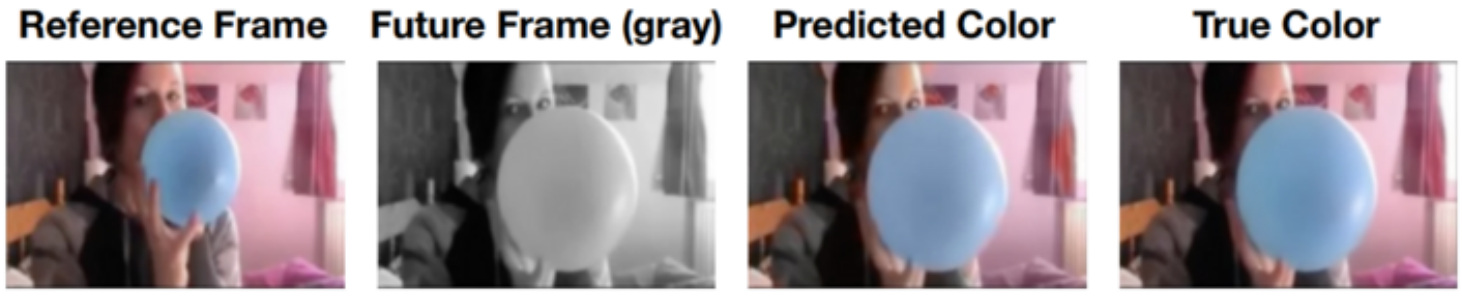
How it work?


Final Loss
Supervised
Cross-entropy Loss
Unsupervised
Consistency Loss
$p_\theta(y \mid x)$
y^*
M
x
y^*
Labeled Data
$p_\theta(y \mid \hat{x})$
M
$\hat{x}$
$p_{\tilde{\theta}}(y \mid x)$
M
x
Unlabeled Data
Augmentations
Back translation
AutoAugment
TF-IDF word
replacement

Data Augmentation

1. Colorization
2. Placing image patches in the right place
3. Inpainting

Colorization



Placing image patches in the right place

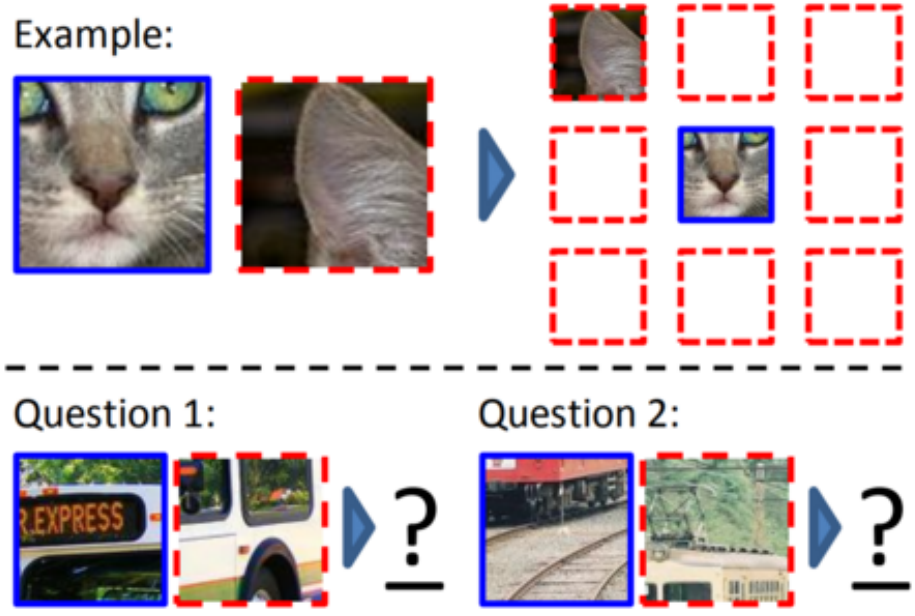
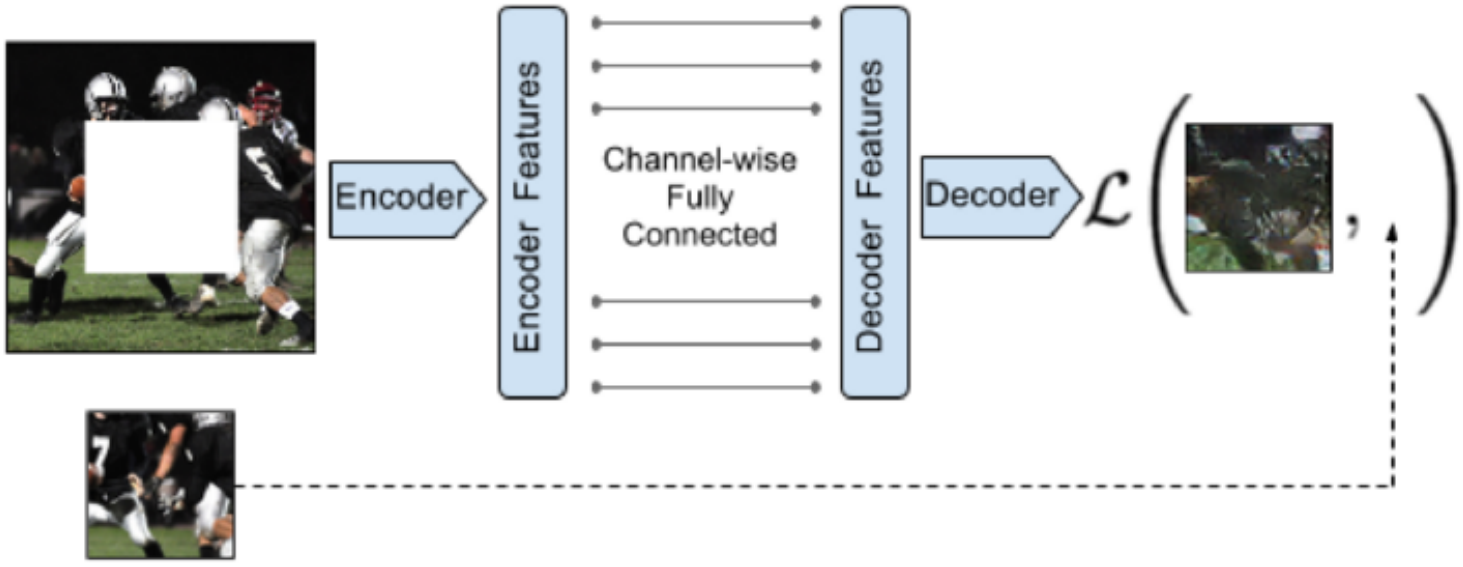


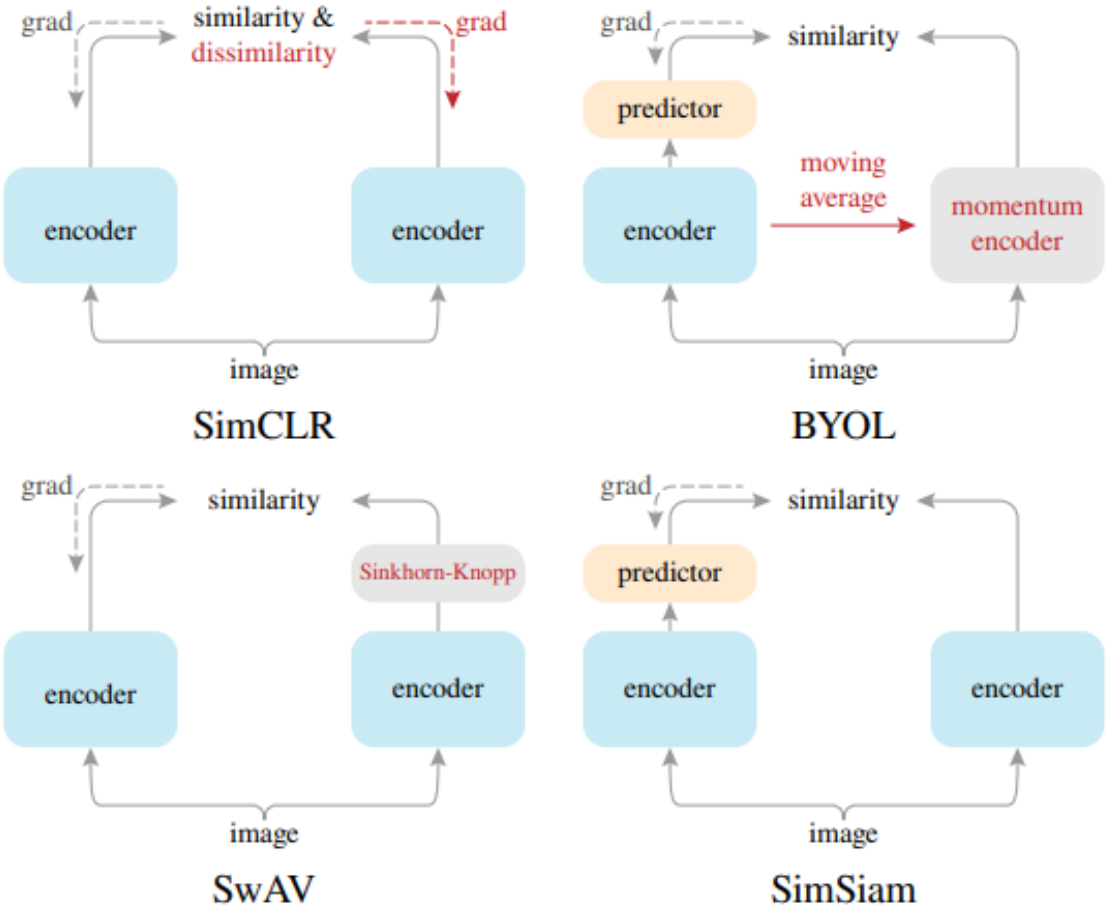
Figure 1. Our task for learning patch representations involves randomly sampling a patch (blue) and then one of eight possible neighbors (red). Can you guess the spatial configuration for the two pairs of patches? Note that the task is much easier once you have recognized the object!

Answer key: Q1: Bottom right Q2: Top center

Inpainting


Encoder
Encoder Features
Channel-wise
Fully
Connected
Decoder Features
Decoder
$\mathcal{L}$
,

Method


grad
similarity &
dissimilarity
grad
encoder
encoder
image
SimCLR
grad
similarity
predictor
moving
average
encoder
momentum
encoder
image
BYOL
grad
similarity
Sinkhorn-Knopp
encoder
encoder
image
SwAV
grad
similarity
predictor
encoder
encoder
image
SimSiam

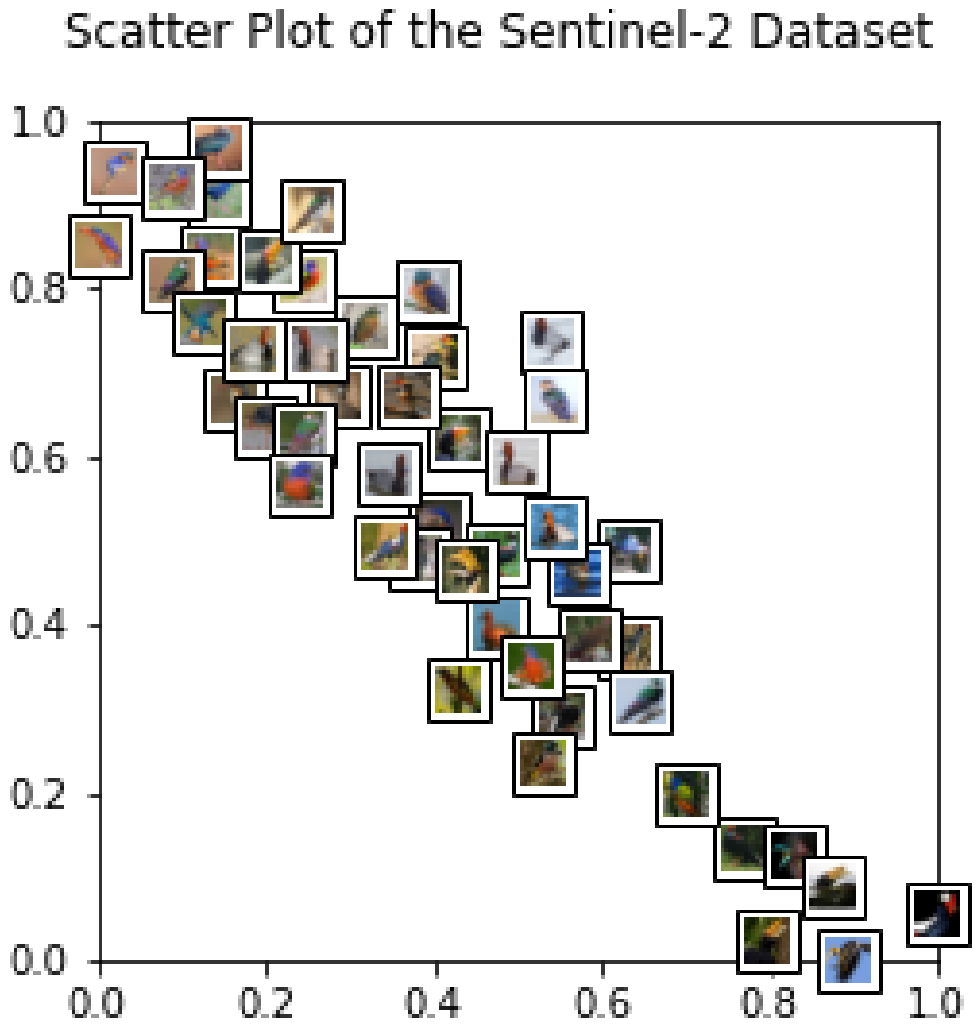
Fomulation

$$\mathcal{L}(\theta, \eta) = \mathbb{E}_{x,\mathcal{T}}\left[\|\mathcal{F}_\theta(\mathcal{T}(x)) - \eta_x\|_2^2\right]$$

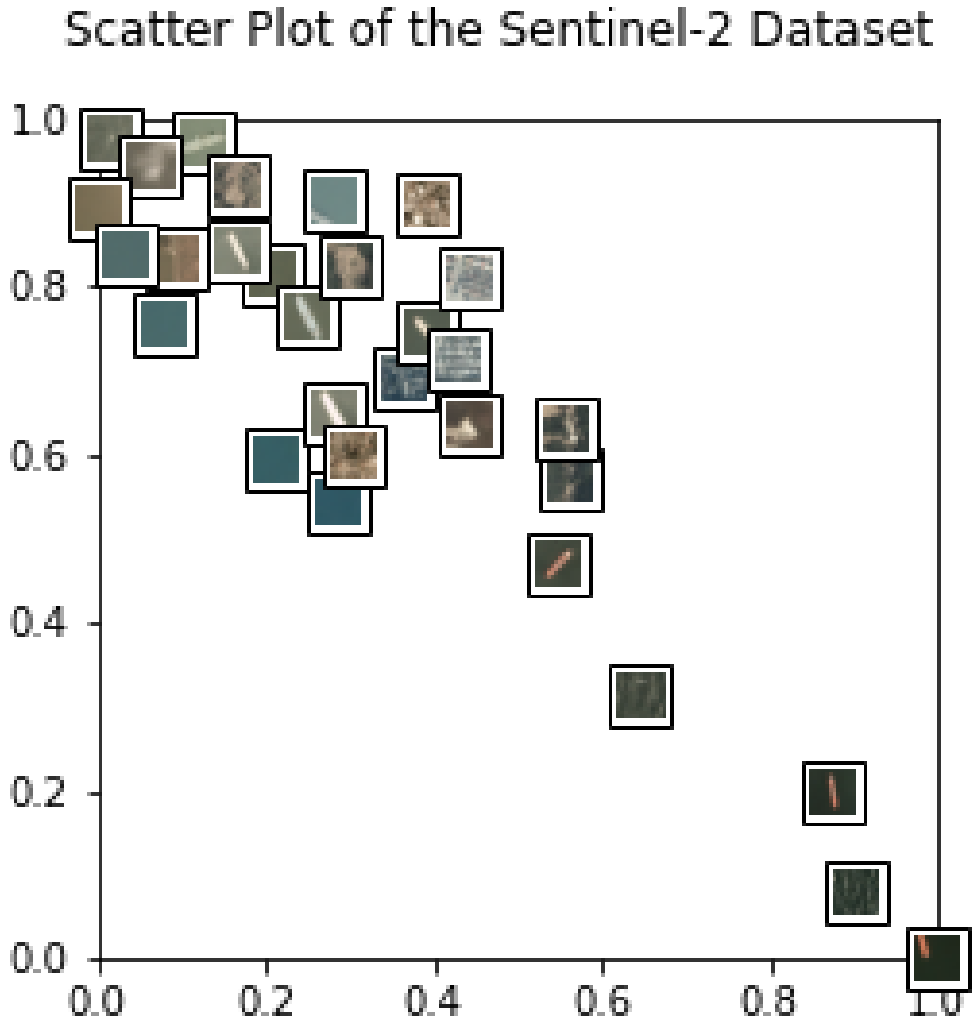
$$\min_{\theta,\eta} \mathcal{L}(\theta, \eta)$$

1. $\mathcal{F}$ is a network parameterized by θ.
2. $\mathcal{T}$ is the augmentation.
3. x is an image.
4. The expectation $\mathbb{E}[\cdot]$ is over the distribution of images and augmentations. For the ease of analysis, here we use the mean squared error $\|\cdot\|_2^2$
5. η_x is the representation of the image x

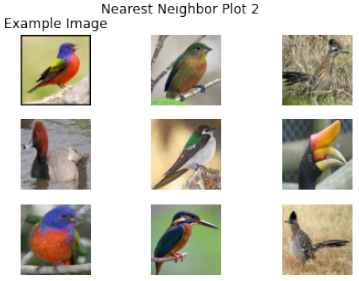
Clustering Image with Self-Supervised Learning (Birds)

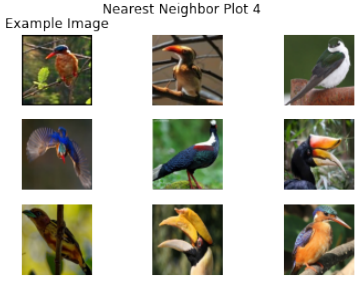


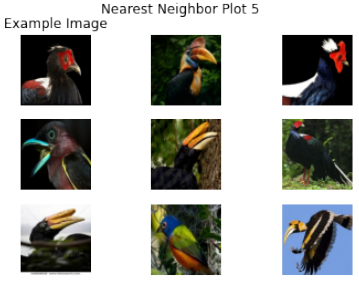
Clustering Image with Self-Supervised Learning (Lands)

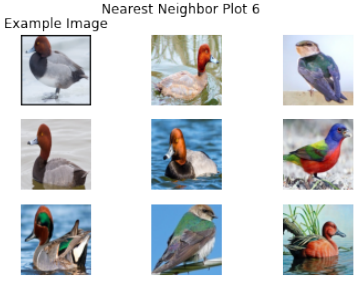


Solution (Birds)

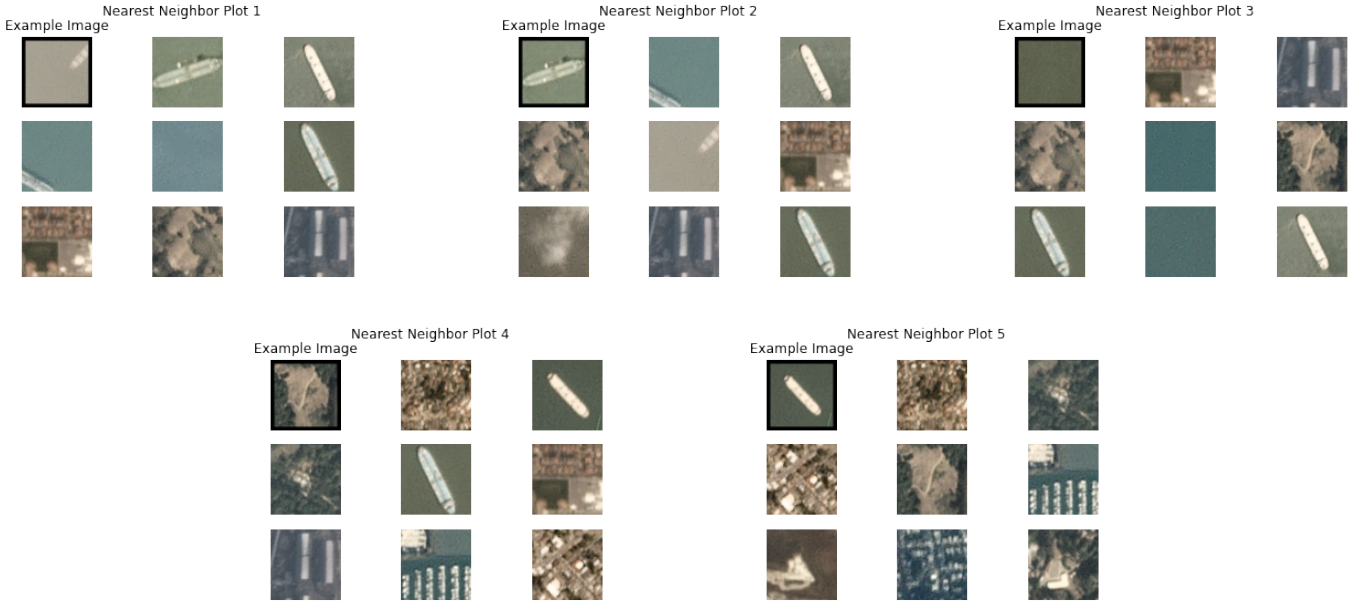








Solution (Lands)



Any Quation?

Thnk you!