Table 1: Training Network Architecture for Object Detection and Region Extraction (YOLOv8)

This network, based on YOLOv8, is designed to extract regions of interest from full fish body captures using a deep learning architecture with over 11 million parameters. It processes images through a series of layers, combining them across stages to enhance detail, and ends with a detection mechanism that identifies and outlines specific areas with bounding boxes. The model efficiently analyzes the entire image to pinpoint key regions, making it ideal for precise object localization in fish imagery.

Layer	Type	Output Shape	Parameters
0	Conv	(32, 320, 320)	928
1	Conv	(64, 160, 160)	18,560
2	C2f	(64, 160, 160)	29,056
3	Conv	(128, 80, 80)	73,984
4	C2f	(128, 80, 80)	197,632
5	Conv	(256, 40, 40)	295,424
6	C2f	(256, 40, 40)	788,480
7	Conv	(512, 20, 20)	1,180,672
8	C2f	(512, 20, 20)	1,838,080
9	SPPF	(512, 20, 20)	656,896
10-14	Upsample + Concat	(384, 40, 40)	591,360
15	C2f	(128, 80, 80)	148,224
16-20	Conv + Concat	(768, 20, 20)	1,230,464
21	C2f	(512, 20, 20)	1,969,152
22	Detect	-	2,116,822

Total parameters: 11,136,374

Conv: Convolutional layer
C2f: A custom module, likely a variant of CSP (Cross Stage Partial Network) block
SPPF: Spatial Pyramid Pooling
Upsample: For increasing spatial dimensions of feature maps
Concat: For concatenating feature maps from different layers
Detect: The final detection layer that produces bounding boxes and class predictions

Table 2: Training Twins Network Architecture for Similarity Evaluation

This neural network, built as a sequential model, employs a tower architecture to evaluate the similarity between two input images, utilizing a pre-trained ResNet50V2 backbone and additional layers, with a total of approximately 23.8 million parameters. Each image in the pair is processed independently through the same tower—passing through ResNet50V2 to extract feature maps, followed by global average pooling to summarize spatial data, batch normalization for stability, dropout to reduce overfitting, and a dense layer to produce a 128-dimensional descriptor. By generating identical 128-dimensional feature vectors for both images using the shared tower, the model enables direct comparison of these descriptors to assess image similarity.

```
# Model: "sequential"
# _________________________________________________________________
#  Layer (type)                Output Shape              Param #
# =================================================================
#  resnet50v2 (Functional)     (None, 7, 7, 2048)        23564800
#
#  global_average_pooling2d (  (None, 2048)              0
#  GlobalAveragePooling2D)
#
#  batch_normalization (Batch  (None, 2048)              8192
#  Normalization)
#
#  dropout (Dropout)           (None, 2048)              0
#
#  dense (Dense)               (None, 128)               262272
#
# =================================================================
# Total params: 23835264 (90.92 MB)
# Trainable params: 23785728 (90.74 MB)
# Non-trainable params: 49536 (193.50 KB)
```