


# TF-Vision Model Garden

 Disclaimer: All datasets hyperlinked from this page are not owned or distributed by Google. The dataset is made available by third parties. Please review the terms and conditions made available by the third parties before using the data.

## Introduction

TF-Vision modeling library for computer vision provides a collection of baselines and checkpoints for image classification, object detection, and segmentation.

## Image Classification

### ImageNet Baselines

#### ResNet models trained with vanilla settings

- Models are trained from scratch with batch size 4096 and 1.6 initial learning rate.
- Linear warmup is applied for the first 5 epochs.
- Models trained with l2 weight regularization and ReLU activation.

Model	Resolution	Epochs	Top-1	Top-5	Download
ResNet-50	224x224	90	76.1	92.9	<a href="#">config</a>
ResNet-50	224x224	200	77.1	93.5	<a href="#">config</a>
ResNet-101	224x224	200	78.3	94.2	<a href="#">config</a>
ResNet-152	224x224	200	78.7	94.3	<a href="#">config</a>

#### ResNet-RS models trained with various settings

We support state-of-the-art [ResNet-RS](#) image classification models with features:

- ResNet-RS architectural changes and Swish activation. (Note that ResNet-RS adopts ReLU activation in the paper.)
- Regularization methods including Random Augment, 4e-5 weight decay, stochastic depth, label smoothing and dropout.
- New training methods including a 350-epoch schedule, cosine learning rate and EMA.
- Configs are in this [directory](#).

Model	Resolution	Params (M)	Top-1	Top-5	Download
ResNet-RS-50	160x160	35.7	79.1	94.5	<a href="#">config</a>   <a href="#">ckpt</a>
ResNet-RS-101	160x160	63.7	80.2	94.9	<a href="#">config</a>   <a href="#">ckpt</a>
ResNet-RS-101	192x192	63.7	81.3	95.6	<a href="#">config</a>   <a href="#">ckpt</a>
ResNet-RS-152	192x192	86.8	81.9	95.8	<a href="#">config</a>   <a href="#">ckpt</a>
ResNet-RS-152	224x224	86.8	82.5	96.1	<a href="#">config</a>   <a href="#">ckpt</a>
ResNet-RS-152	256x256	86.8	83.1	96.3	<a href="#">config</a>   <a href="#">ckpt</a>

ResNet-RS-200	256x256	93.4	83.5	96.6	<a href="#">config</a>   <a href="#">ckpt</a>
ResNet-RS-270	256x256	130.1	83.6	96.6	<a href="#">config</a>   <a href="#">ckpt</a>
ResNet-RS-350	256x256	164.3	83.7	96.7	<a href="#">config</a>   <a href="#">ckpt</a>
ResNet-RS-350	320x320	164.3	84.2	96.9	<a href="#">config</a>   <a href="#">ckpt</a>

### Vision Transformer (ViT)

We support [ViT](#) and [DEiT](#) implementations in a TF Vision [project](#). ViT models trained under the DEiT settings:

model	resolution	Top-1	Top-5
ViT-s16	224x224	79.4	94.7
ViT-b16	224x224	81.8	95.8
ViT-l16	224x224	82.2	95.8

## Object Detection and Instance Segmentation

### Common Settings and Notes

- We provide models adopting [ResNet-FPN](#) and [SpineNet](#) backbones based on detection frameworks:
  - [RetinaNet](#) and [RetinaNet-RS](#)
  - [Mask R-CNN](#)
  - [Cascade RCNN](#) and [Cascade RCNN-RS](#)
- Models are all trained on [COCO](#) train2017 and evaluated on [COCO](#) val2017.
- Training details:
  - Models finetuned from [ImageNet](#) pretrained checkpoints adopt the 12 or 36 epochs schedule. Models trained from scratch adopt the 350 epochs schedule.
  - The default training data augmentation implements horizontal flipping and scale jittering with a random scale between [0.5, 2.0].
  - Unless noted, all models are trained with l2 weight regularization and ReLU activation.
  - We use batch size 256 and stepwise learning rate that decays at the last 30 and 10 epoch.
  - We use square image as input by resizing the long side of an image to the target size then padding the short side with zeros.

### COCO Object Detection Baselines

#### RetinaNet (ImageNet pretrained)

Backbone	Resolution	Epochs	FLOPs (B)	Params (M)	Box AP	Download
R50-FPN	640x640	12	97.0	34.0	34.3	<a href="#">config</a>
R50-FPN	640x640	72	97.0	34.0	36.8	<a href="#">config</a>   <a href="#">ckpt</a>

#### RetinaNet (Trained from scratch) with training features including:

- Stochastic depth with drop rate 0.2.
- Swish activation.

Backbone	Resolution	Epochs	FLOPs (B)	Params (M)	Box AP	Download

SpineNet-49	640x640	500	85.4	28.5	44.2	<a href="#">config</a>   <a href="#">TB.dev</a>
SpineNet-96	1024x1024	500	265.4	43.0	48.5	<a href="#">config</a>   <a href="#">TB.dev</a>
SpineNet-143	1280x1280	500	524.0	67.0	50.0	<a href="#">config</a>   <a href="#">TB.dev</a>

#### Mobile-size RetinaNet (Trained from scratch):

Backbone	Resolution	Epochs	FLOPs (B)	Params (M)	Box AP	Download
MobileNetv2	256x256	600	-	2.27	23.5	<a href="#">config</a>
Mobile SpineNet-49	384x384	600	1.0	2.32	28.1	<a href="#">config</a>   <a href="#">ckpt</a>

## Instance Segmentation Baselines

#### Mask R-CNN (Trained from scratch)

Backbone	Resolution	Epochs	FLOPs (B)	Params (M)	Box AP	Mask AP	Download
ResNet50-FPN	640x640	350	227.7	46.3	42.3	37.6	<a href="#">config</a>
SpineNet-49	640x640	350	215.7	40.8	42.6	37.9	<a href="#">config</a>
SpineNet-96	1024x1024	500	315.0	55.2	48.1	42.4	<a href="#">config</a>
SpineNet-143	1280x1280	500	498.8	79.2	49.3	43.4	<a href="#">config</a>

#### Cascade RCNN-RS (Trained from scratch)

Backbone	Resolution	Epochs	Params (M)	Box AP	Mask AP	Download
SpineNet-49	640x640	500	56.4	46.4	40.0	<a href="#">config</a>
SpineNet-96	1024x1024	500	70.8	50.9	43.8	<a href="#">config</a>
SpineNet-143	1280x1280	500	94.9	51.9	45.0	<a href="#">config</a>

## Semantic Segmentation

- We support [DeepLabV3](#) and [DeepLabV3+](#) architectures, with Dilated ResNet backbones.
- Backbones are pre-trained on ImageNet.

#### PASCAL-VOC

Model	Backbone	Resolution	Steps	mIoU	Download
DeepLabV3	Dilated Resnet-101	512x512	30k	78.7	
DeepLabV3+	Dilated Resnet-101	512x512	30k	79.2	

#### CITYSCAPES

Model	Backbone	Resolution	Steps	mIoU	Download

DeepLabV3+	Dilated Resnet-101	1024x2048	90k	78.79	
------------	--------------------	-----------	-----	-------	--

## Video Classification

### Common Settings and Notes

- We provide models for video classification with backbones:
  - SlowOnly in [SlowFast Networks for Video Recognition](#).
  - ResNet-3D (R3D) in [Spatiotemporal Contrastive Video Representation Learning](#).
  - ResNet-3D-RS (R3D-RS) in [Revisiting 3D ResNets for Video Recognition](#).
  - Mobile Video Networks (MoViNets) in [MoViNets: Mobile Video Networks for Efficient Video Recognition](#).
- Training and evaluation details (SlowFast and ResNet):
  - All models are trained from scratch with vision modality (RGB) for 200 epochs.
  - We use batch size of 1024 and cosine learning rate decay with linear warmup in first 5 epochs.
  - We follow [SlowFast](#) to perform 30-view evaluation.

### Kinetics-400 Action Recognition Baselines

Model	Input (frame x stride)	Top-1	Top-5	Download
SlowOnly	8 x 8	74.1	91.4	<a href="#">config</a>
SlowOnly	16 x 4	75.6	92.1	<a href="#">config</a>
R3D-50	32 x 2	77.0	93.0	<a href="#">config</a>
R3D-RS-50	32 x 2	78.2	93.7	<a href="#">config</a>
R3D-RS-101	32 x 2	79.5	94.2	-
R3D-RS-152	32 x 2	79.9	94.3	-
R3D-RS-200	32 x 2	80.4	94.4	-
R3D-RS-200	48 x 2	81.0	-	-
MoViNet-A0-Base	50 x 5	69.40	89.18	-
MoViNet-A1-Base	50 x 5	74.57	92.03	-
MoViNet-A2-Base	50 x 5	75.91	92.63	-
MoViNet-A3-Base	120 x 2	79.34	94.52	-
MoViNet-A4-Base	80 x 3	80.64	94.93	-
MoViNet-A5-Base	120 x 2	81.39	95.06	-

### Kinetics-600 Action Recognition Baselines

Model	Input (frame x stride)	Top-1	Top-5	Download
SlowOnly	8 x 8	77.3	93.6	<a href="#">config</a>

R3D-50	32 x 2	79.5	94.8	<a href="#">config</a>
R3D-RS-200	32 x 2	83.1	-	-
R3D-RS-200	48 x 2	83.8	-	-
MoViNet-A0-Base	50 x 5	72.05	90.92	<a href="#">config</a>
MoViNet-A1-Base	50 x 5	76.69	93.40	<a href="#">config</a>
MoViNet-A2-Base	50 x 5	78.62	94.17	<a href="#">config</a>
MoViNet-A3-Base	120 x 2	81.79	95.67	<a href="#">config</a>
MoViNet-A4-Base	80 x 3	83.48	96.16	<a href="#">config</a>
MoViNet-A5-Base	120 x 2	84.27	96.39	<a href="#">config</a>