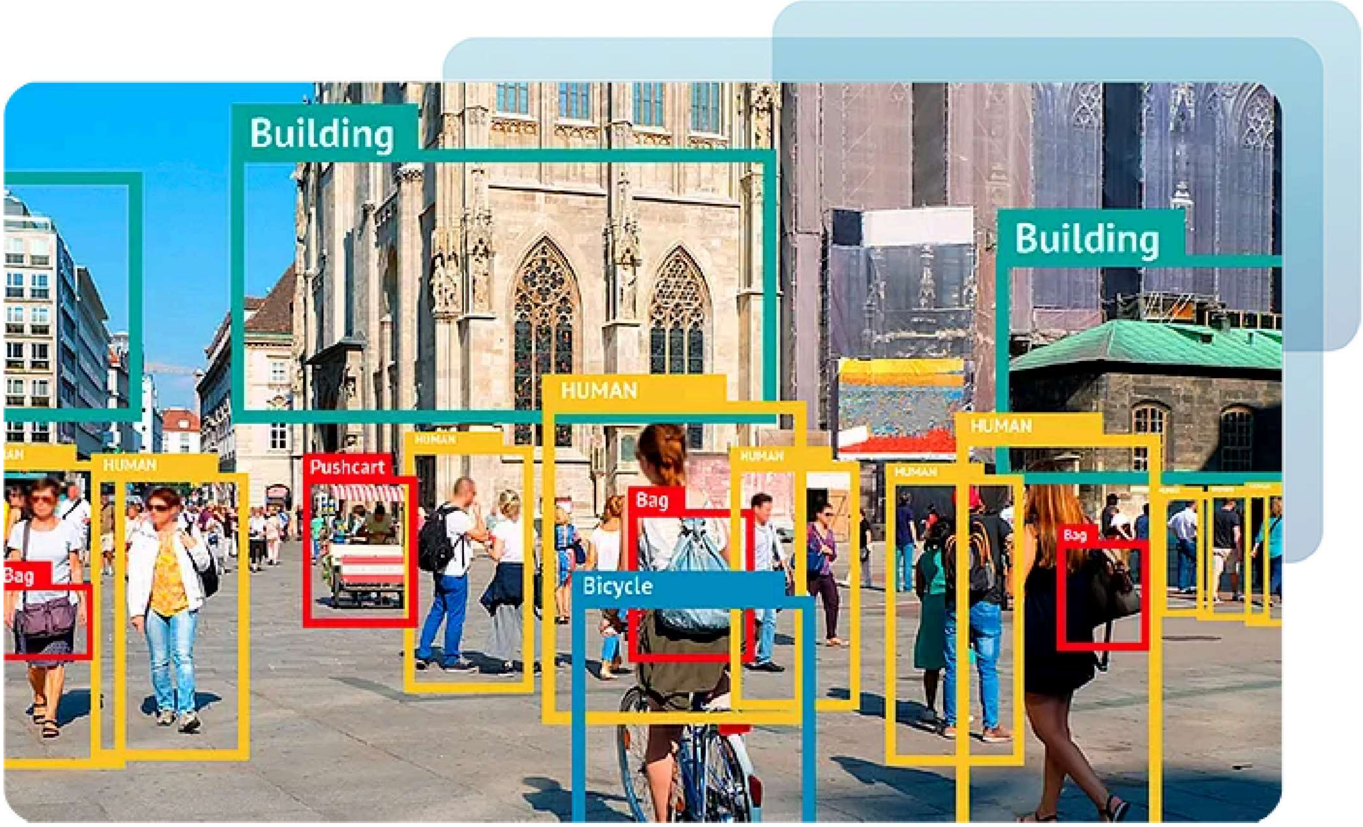
✦ Member-only story

Top 10 Object Detection Models in 2024

Aarafat Islam · Follow

Published in Tech Spectrum · 5 min read · Sep 30, 2024

2



Object detection is a fundamental task in computer vision that involves identifying and localizing objects within an image. Deep learning has revolutionized object detection, allowing for more accurate and efficient detection of objects in images and videos. In 2024, there are several deep-learning models that are making significant advancements in object detection. Here are the top 10 deep-learning models for object detection in 2024:

1. YOLOv10

YOLOv10 is a state-of-the-art deep learning model for object detection that uses a more efficient backbone network and a new set of detection heads. YOLOv10 can detect objects in real-time with high accuracy and can be trained on large datasets.

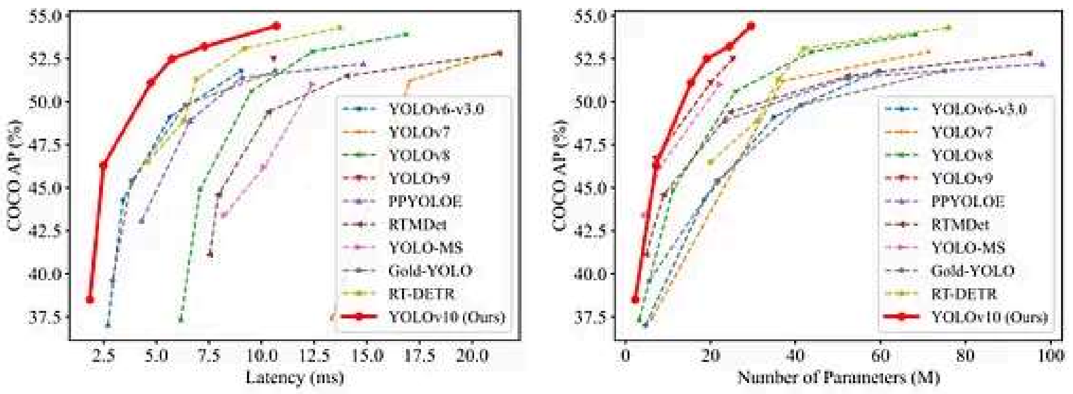


Figure 1: Comparisons with others in terms of latency-accuracy (left) and size-accuracy (right) trade-offs. We measure the end-to-end latency using the official pre-trained models.

> ***Paper:*** *https://arxiv.org/pdf/2405.14458*

```
Pros:
  1. Very fast and efficient object detection
  2. High accuracy on large datasets
```

```
   3. Runs on low-end devices

Cons:
   1. Can struggle with small object detection
   2. Requires a large dataset for optimal performance
```

2. EfficientDet

EfficientDet is a deep-learning model for object detection that uses an efficient backbone network and a new set of detection heads. EfficientDet is designed to be efficient and accurate and can detect objects in real-time with high accuracy.

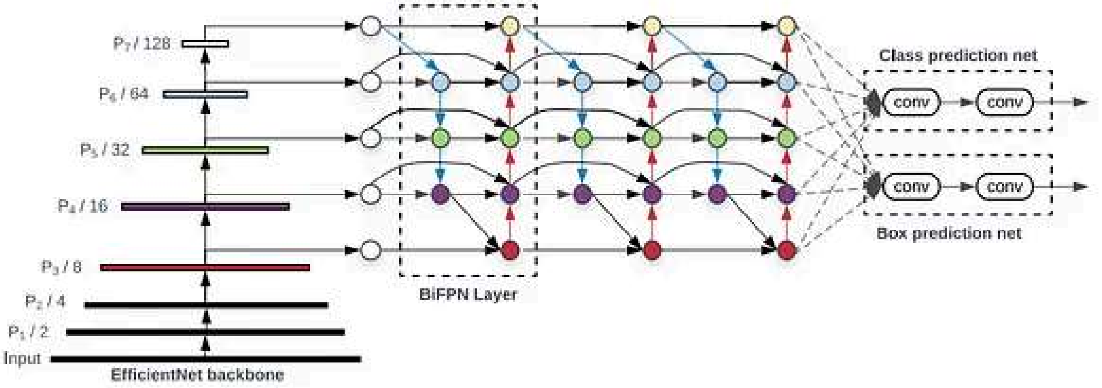


Figure 3: **EfficientDet architecture** – It employs EfficientNet [36] as the backbone network, BiFPN as the feature network, and shared class/box prediction network. Both BiFPN layers and class/box net layers are repeated multiple times based on different resource constraints as shown in Table 1.

> ***Paper:*** *https://openaccess.thecvf.com/content_CVPR_2020/html/Tan_EfficientDet_Scalable_and_Efficient_Object_Detection_CVPR_2020_paper.html*

```
Pros:
1. State-of-the-art performance on several benchmark datasets
2. Efficient and accurate object detection
```

```
3. Can be trained on large datasets

Cons:
1. Requires a large number of computational resources
2. Can be challenging to train on smaller datasets
```

3. RetinaNet

RetinaNet is a deep learning model for object detection that uses a feature pyramid network and a new focal loss function. RetinaNet is designed to address the imbalance between foreground and background examples in object detection, leading to improved accuracy.

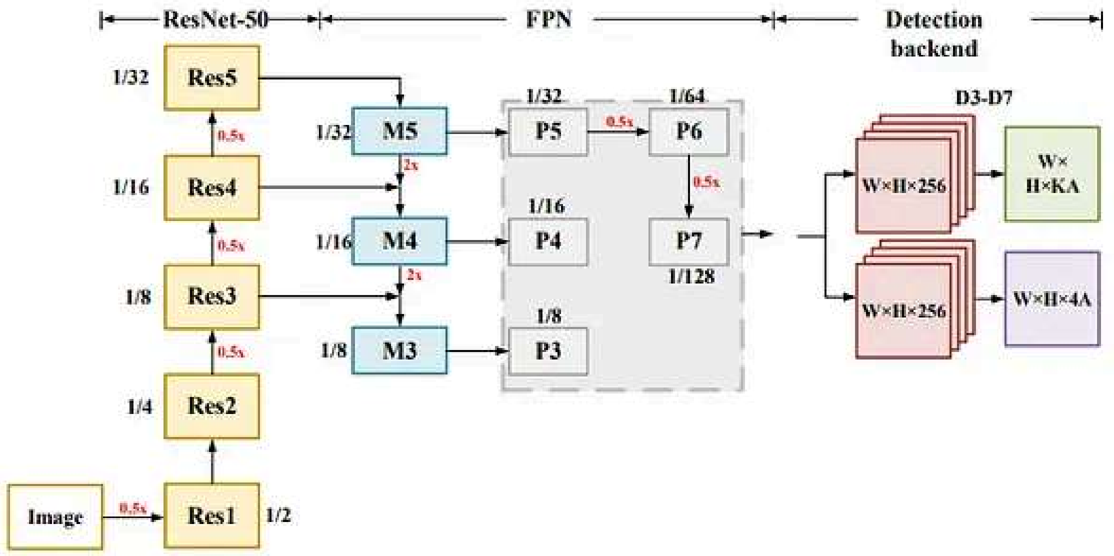


Figure 2: RetinaNet (ResNet50-FPN-800x800) network architecture.

> ***Paper:*** *https://arxiv.org/pdf/1905.10011*

```
Pros:
  1. Improved accuracy in object detection
  2. Efficient and can run on low-end devices
  3. Easy to train and use
```

```
Cons:
  1. Can struggle with small object detection
  2. Can require a large amount of data for optimal performance
```

4. DETR v2

DETR v2 is a deep learning model for object detection that uses a transformer-based architecture. DETR v2 uses a set prediction approach to simultaneously predict the class and location of each object.

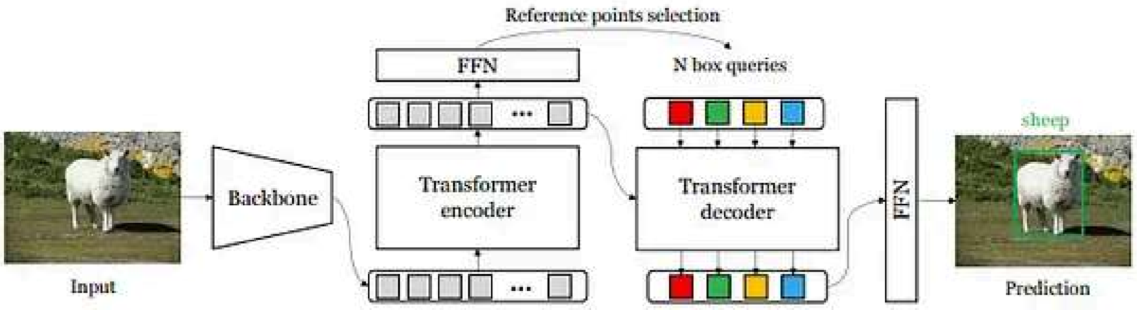


Fig. 1. The overall architecture. (1) We use a conventional CNN backbone (*e.g.*, ResNet-50) to extract the feature of an input image. (2) The learned embedding is fed into the transformer encoder layers to model the global dependencies between the inputs. (3) Then, we predict candidate boxes from the encoder embedding and select reference points according to the classification score of these boxes. The input of the decoder is the concatenation of the box query and the content query. The box query consists of the embedding of the selected reference points and the transformation predicted from the corresponding encoder embedding of these reference points. The content query is initialized from the selected candidate boxes. (4) We pass each output embedding of the decoder to a shared feed-forward network (FFN) that predicts either a detection (class and bounding box) or a ”no object” class.

> ***Paper:*** *https://arxiv.org/pdf/2207.08914*

```
Pros:
  1. High accuracy and simplicity in object detection
  2. Can handle highly overlapping objects
  3. No anchor boxes or non-maximum suppression required
Cons:
  1. Can be computationally expensive
  2. Requires a large amount of data for optimal performance
```

5. CenterNet++

CenterNet++ is a deep learning model for object detection that uses a heatmap to predict the center of each object. CenterNet++ then uses a second network to predict the size and orientation of the object.

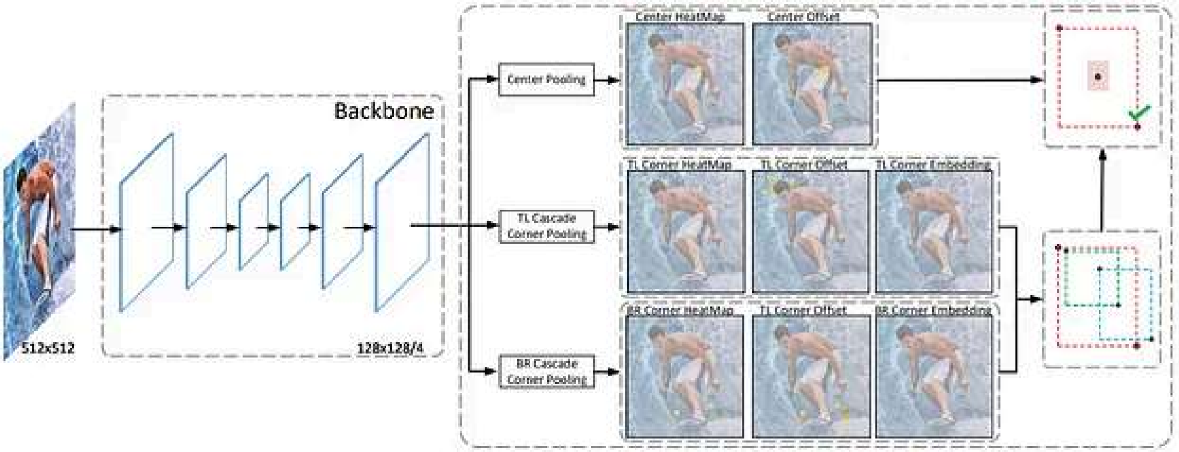


Fig. 2. Single-resolution detection framework of CenterNet. A convolutional backbone network applies cascade corner pooling and center pooling to output two corner heatmaps and a center keypoint heatmap, respectively. Note that the heatmaps are multi-class heatmaps, which means that the number of channel of each heatmap equals to the number of the classes in the dataset. Similar to CornerNet, a pair of detected corners and the similar embeddings are used to detect a potential bounding boxes. Then the detected center keypoints are used to determine the final bounding boxes.

> ***Paper:*** *https://arxiv.org/pdf/2204.08394*

```
Pros:
  1. High accuracy and efficiency in object detection
  2. Can handle occluded and small objects
Cons:
  1. Can be computationally expensive
  2. Can struggle with highly overlapping objects
```

6. FCOS

FCOS is a deep learning model for object detection that uses a fully

benchmark datasets.

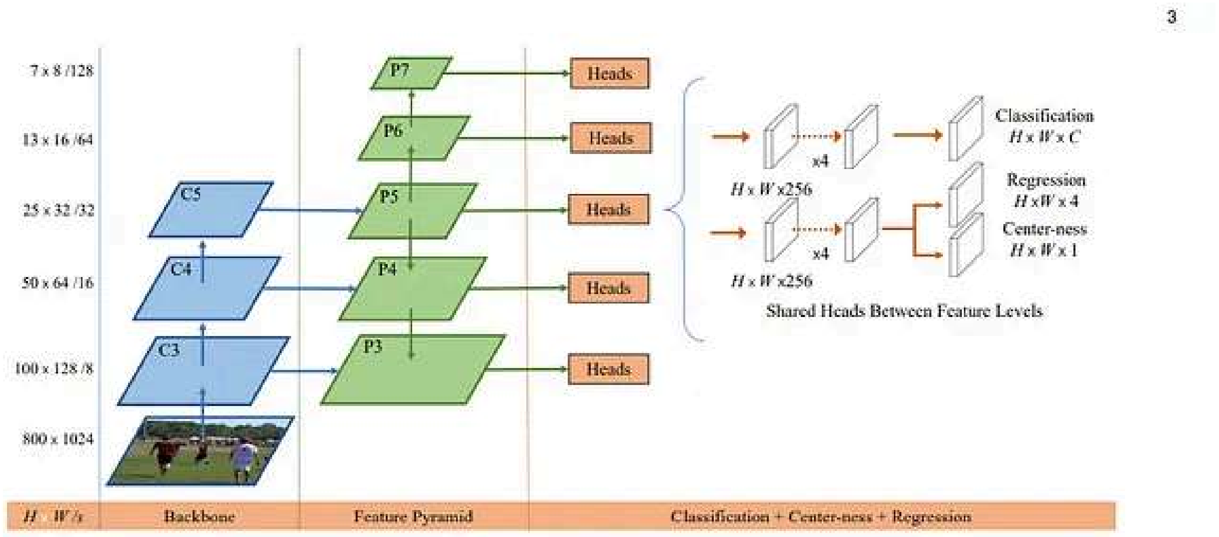


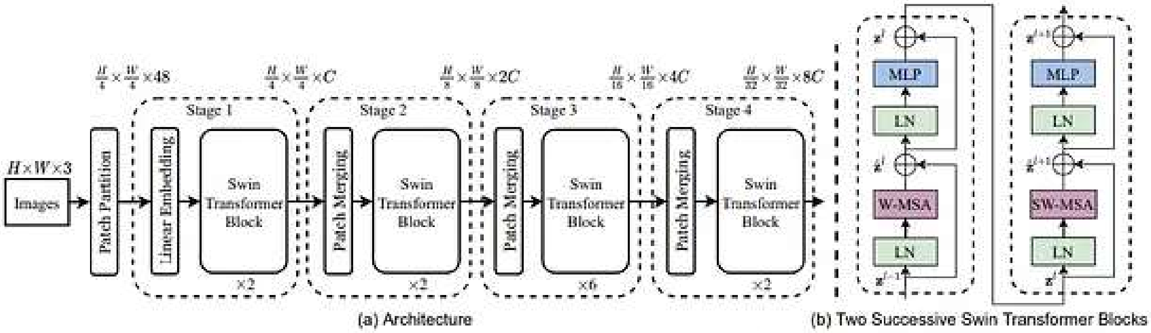
Fig. 2. **The network architecture of FCOS**, where C3, C4, and C5 denote the feature maps of the backbone network and P3 to P7 are the feature levels used for the final prediction. $H \times W$ is the height and width of feature maps. '/s' ($s = 8, 16, \ldots, 128$) is the down-sampling ratio of the feature maps at the level to the input image. As an example, all the numbers are computed with an 800×1024 input.

> *Paper: https://arxiv.org/pdf/2006.09214*

```
Pros:
  1. State-of-the-art performance on several benchmark datasets
  2. High accuracy and efficiency in object detection
  3. No anchor boxes or non-maximum suppression required
Cons:
  1. Can be computationally expensive
  2. Can require a large dataset for optimal performance
```

7. Swin Transformer

Swin Transformer is a deep learning model for object detection that uses a transformer-based architecture. Swin Transformer uses a set prediction approach to simultaneously predict the class and location of each object.



(a) Architecture

(b) Two Successive Swin Transformer Blocks

Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

> ***Paper:*** *https://openaccess.thecvf.com/content/ICCV2021/papers/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.pdf*

8. DINO

DINO is a deep learning model for object detection that uses a transformer-based architecture. DINO uses a set prediction approach to simultaneously predict the class and location of each object.

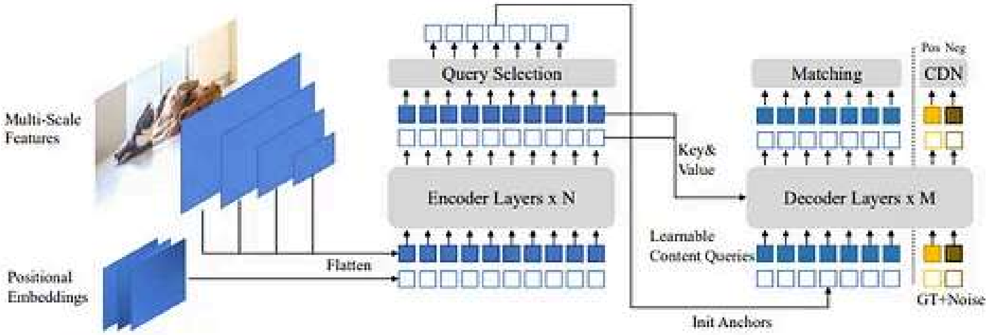


Figure 2: The framework of our proposed DINO model. Our improvements are mainly in the Transformer encoder and decoder. The top-K encoder features in the last layer are selected to initialize the positional queries for the Transformer decoder. Our decoder also contains a Contrastive DeNoising (CDN) part with both positive and negative examples.

> ***Paper:*** *https://openreview.net/pdf?id=3mRwyG5one*

9. ViTAE

ViTAE is a deep learning model for object detection that uses a transformer-based architecture. ViTAE uses a set prediction approach to simultaneously predict the class and location of each object.

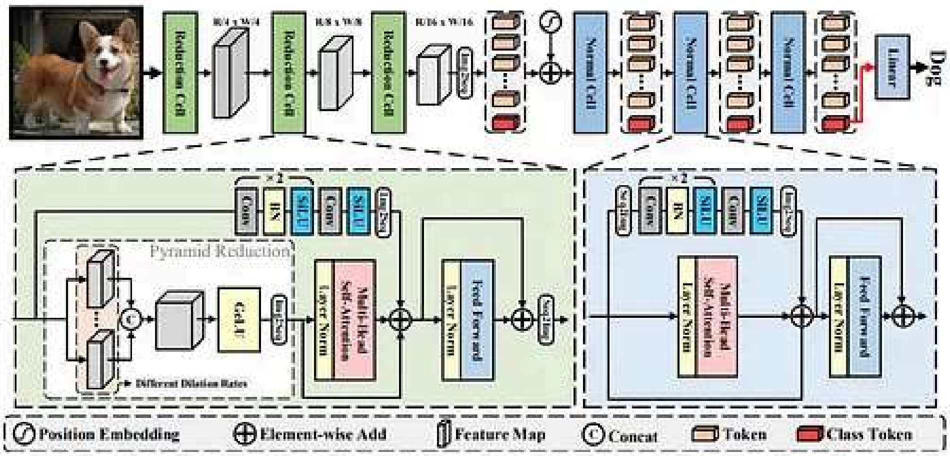


Figure 2: The structure of the proposed ViTAE. It is constructed by stacking three RCs and several NCs. Both types of cells share a simple basic structure, *i.e.*, an MHSA module and a parallel convolutional module followed by an FFN. In particular, RC has an extra pyramid reduction module using atrous convolutions with different dilation rates to embed multi-scale context into tokens.

> *Paper: https://proceedings.neurips.cc/paper_files/paper/2021/file/efb76cff97aaf057654ef2f38cd77d73-Paper.pdf*

10. BEiT

BEiT is a deep learning model for object detection that uses a transformer-based architecture. BEiT uses a set prediction approach to simultaneously predict the class and location of each object.

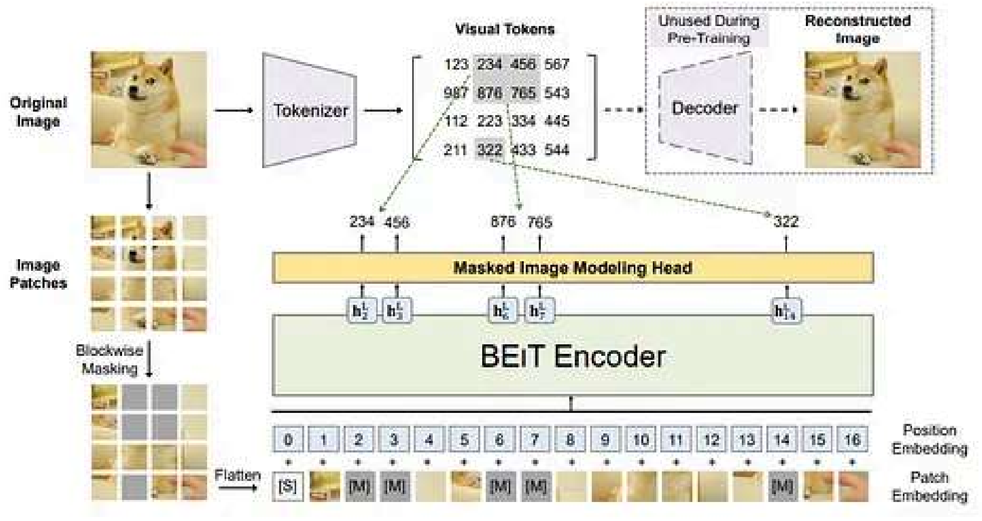


Figure 1: Overview of BEiT pre-training. Before pre-training, we learn an "image tokenizer" via autoencoding-style reconstruction, where an image is tokenized into discrete visual tokens according to the learned vocabulary. During pre-training, each image has two views, i.e., image patches, and visual tokens. We randomly mask some proportion of image patches (gray patches in the figure) and replace them with a special mask embedding [M]. Then the patches are fed to a backbone vision Transformer. The pre-training task aims at predicting the visual tokens of the *original* image based on the encoding vectors of the *corrupted* image.

> *Paper: https://openreview.net/pdf?id=p-BhZSz59o4*

```
Pros:
  1. High accuracy and simplicity in object detection
  2. Can handle highly overlapping objects
  3. No anchor boxes or non-maximum suppression required
Cons:
    1. Can be computationally expensive
    2. Requires a large amount of data for optimal performance
```

Object Detection Computer Vision Image Processing Deep Learning

Artificial Intelligence

Written by Aarafat Islam

Follow ○

312 Followers · Editor for Tech Spectrum

🌎 A Philomath | Predilection for AI, DL | Blockchain | Researcher | Technophile | True Optimist | Endeavors to make impact on the world! ✨

More from Aarafat Islam and Tech Spectrum

Aarafat Islam in The Pythoneers

Building a Blockchain from Scratch with Python!

Learn the Basics of Blockchain Technology and Create Your Own Blockchain in Python...

✦ Feb 18, 2023 296 4

Aarafat Islam in Tech Spectrum

666+ LLM Prompts!

Must-try LLM prompts for instant creativity.

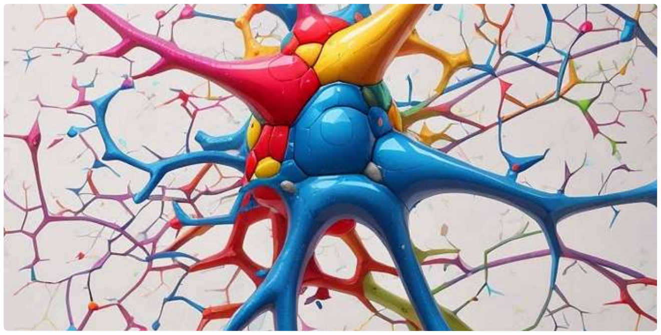
✦ Sep 25 40

Aarafat Islam in Tech Spectrum

The Cultural Impact of Artificial Intelligence!

Exploring the Positive and Negative Impacts of AI on Media, Creativity, Social Interactions...

✦ Feb 4, 2023 133 1



Aarafat Islam

A Comprehensive Guide to Activation Functions in Deep...

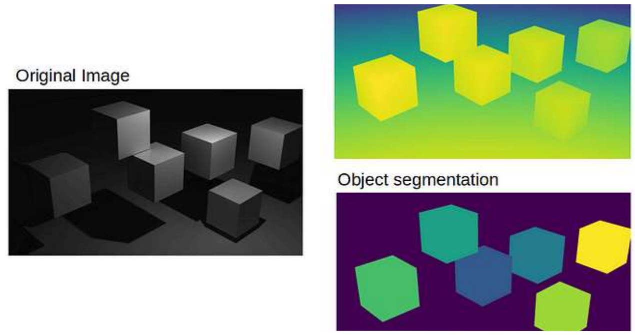
“Activation functions are the spark of intelligence in neural networks.”

✦ Sep 25, 2023 56

See all from Aarafat Islam

See all from Tech Spectrum

Recommended from Medium

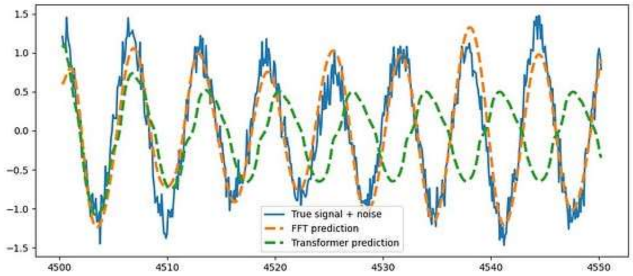


Merwansky

The Power of Synthetic Image Datasets: How Blender is...

Synthetic Datasets with Blender, Part IV/ V

★ Sep 10 12



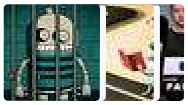
Harys Dalvi in Towards Data Science

Can Transformers Solve Everything?

Looking into the math and the data reveals that transformers are both overused and...

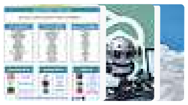
★ Oct 1 530 13

Lists



AI Regulation

6 stories · 584 saves



ChatGPT

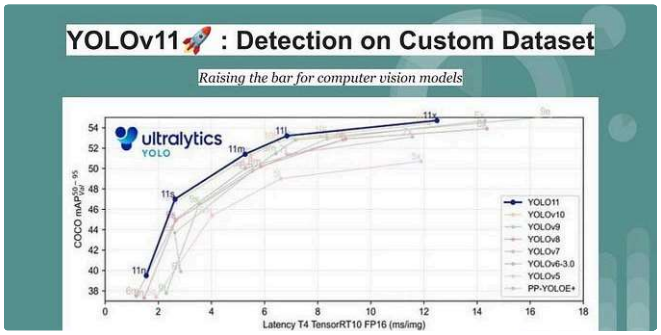
21 stories · 827 saves

Natural Language Processing

1741 stories · 1337 saves

Generative AI Recommended Reading

52 stories · 1417 saves



Vipin in The Deep Hub

YOLOv11🚀 : Sign Language Letter Detection on Custom Dataset

Raising the bar for computer vision models

✦ Oct 1 👏 211

AI Papers Academy

Sapiens by Meta AI: Foundation for Human Vision Models

In this post we dive into Sapiens, a new family of computer vision models by Meta AI that…

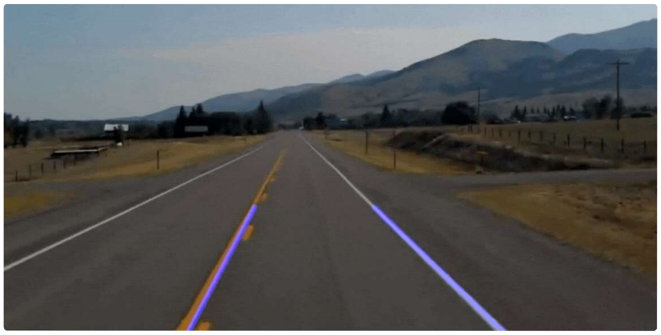
✦ Aug 26 👏 83

Kevin Akbari

Using GANs for Anomaly Detection

Generative Adversarial Networks (GANs) have garnered significant attention in recent…

✦ May 28 👏 10



md

Part V: Finding Road Marking Using OpenCV

In this section, we will detect road markings using OpenCV and other libraries. First, we…

✦ Sep 30 👏 39

See more recommendations