YOLOv11 Architecture Blocks

1. Conv Block

Structure:

- Conv → BatchNorm → SiLU

Details:

- Kernel: `3×3`, Stride: `1` or `2`
- Stride 2 → Downsampling
- Preserves/increases channel size

Purpose in YOLOv11:

- Basic unit of feature extraction
- Captures low-level textures and edges
- Downsamples early layers to reduce resolution and increase receptive field

2. C3k2 Block (C3 with k=2)

Structure:

- CSP-style split into two paths
- Main: 2 Bottleneck layers; Shortcut: direct
- Concat → Conv

Details:

- Optional shortcut in bottlenecks
- `hidden_channels = expansion × out_channels`

Purpose in YOLOv11:

- Balances efficiency and performance
- Reduces redundancy by reusing partial features
- Maintains rich features with fewer parameters than ResNet blocks

3. SPPF Block (Spatial Pyramid Pooling - Fast)

Structure:

- 3 sequential MaxPool(`5×5`, stride=1)
- Concat original + 3 pooled outputs → Conv

Details:

- Captures multi-scale receptive fields
- Efficient alternative to traditional SPP

Purpose in YOLOv11:

- Encodes global context into deeper layers
- Improves object detection at different scales

- Helps in dense and cluttered scenes

4. C2PSA Block (Cross Stage + Self-Attention)

Structure:

- 2 branches: one with convolutions + attention, one identity
- Concat → Conv

Details:

- Uses SE or Transformer-style attention
- Enhances feature map selectively

Purpose in YOLOv11:

- Focuses on salient regions in the scene
- Improves detection of small and occluded objects
- Boosts feature expressiveness before detection head

5. nn.Upsample

Structure:

- No parameters, uses interpolation
- Nearest-neighbor by default

Details:

- Upscales features (e.g. 40×40 → 80×80)

Purpose in YOLOv11:

- Facilitates feature fusion in FPN/Neck
- Combines coarse and fine resolution info
- Improves small object detection

6. Concat

Structure:

- Merges feature maps along channel dimension

Details:

- Used in Neck: combines upsampled and skip features

Purpose in YOLOv11:

- Preserves both low-level and high-level features
- Strengthens semantic flow through layers
- Enhances precision in object boundaries

7. Detect Layer

Structure:

- Receives 3 feature maps: P3/8, P4/16, P5/32
- Each → 1×1 Conv → Prediction tensor
- Prediction shape: `anchors × (x, y, w, h, obj, classes)`

Details:

- Multi-scale detection output
- Anchor-based or anchor-free

Purpose in YOLOv11:

- Outputs final bounding boxes and class scores
- Handles objects of different scales efficiently
- Crucial layer for prediction inference

Additional Notes

- **Output size, kernel size, stride** Syntax of arg type1
- **Output channels, shortcut, expansion** Syntax of type2