Semantic Segmentation Using DINOv2 ViT and U-Net on Pascal VOC 2012

Semantic segmentation involves labeling each pixel in an image with a corresponding class. This project applies semantic segmentation to natural scene images from the **Pascal VOC 2012** dataset using a hybrid model that combines a **pre-trained DINOv2 Vision Transformer (ViT)** with a **U-Net-style decoder**.

Dataset

The **Pascal VOC 2012** dataset consists of real-world images with pixel-level annotations across **21 classes** (including background). These include objects such as people, animals, vehicles, and indoor items.

- **Classes**: 20 foreground classes + 1 background class.
- **Preprocessing**:
    - All input images are resized to **518×518** to ensure compatibility with the ViT backbone (`vit_base_patch14_dinov2`).
    - Masks are resized using **nearest neighbor interpolation** to preserve discrete class labels.
    - Dataset split used: `train` and `val` provided by Pascal VOC 2012.
- Dataset link: [PASCAL VOC 2012]

Model Architecture

This model is a modified **U-Net** architecture composed of:

Encoder (Backbone)

- **DINOv2 Vision Transformer** (`vit_base_patch14_dinov2`)
- Outputs features of shape `[B, 768, 37, 37]` after patch projection and positional encoding.

- **Frozen during training**, except for the last transformer block (`blocks.11`) to reduce computational cost and overfitting.

Decoder

- The decoder progressively up samples the 37×37 feature map using convolutional blocks:

Layer	Channels	Size
up1	768 → 512	37×37 → 74×74
up2	512 → 256	74×74 → 148×148
up3	256 → 128	148×148 → 296×296
up4	128 → 64	296×296 → 518×518

- A final **1×1 convolution** reduces the feature maps to **21 channels**, representing per-pixel class logits.

Training Details

- **Loss Function**: `Lovasz Softmax Loss` — designed to optimize the mean IoU directly, more robust for segmentation tasks with class imbalance.
- **Optimizer**: `AdamW` with a learning rate of `1e-4`.
- **Training Strategy**:
  - Only the decoder and final transformer block are trained.
  - Input batch size: 2 (due to memory limits of ViT + 518×518 inputs).
- **Epochs**: 8

Evaluation Metrics

- **Pixel Accuracy**: Measures the ratio of correctly classified pixels.
- **Mean Intersection over Union (mIoU)**: Evaluates the overlap between predicted and ground-truth masks for each class.

Results

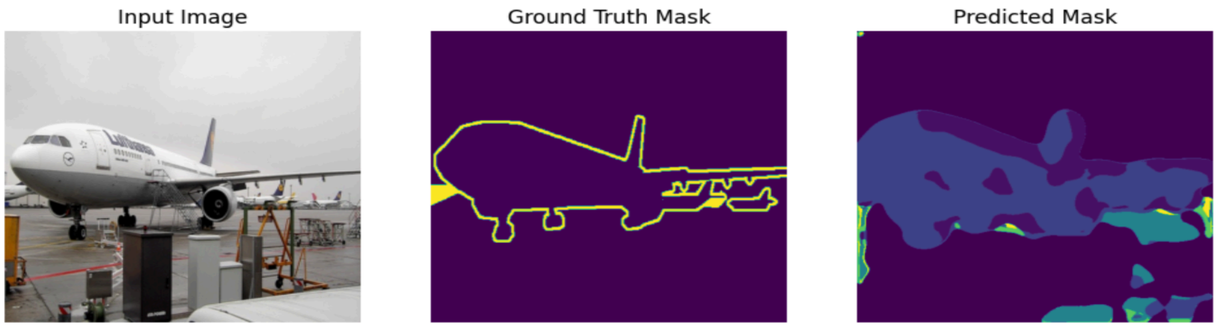
Epoch	Train Loss	Val Accuracy	Val mIoU
1	0.6927	0.7481	0.0868
2	0.5166	0.8150	0.1073
3	0.4497	0.8839	0.1196
4	0.3999	0.8839	0.1202
5	0.3499	0.9125	0.1290
6	0.3193	0.9350	0.1361
7	0.2882	0.9426	0.1376
8	0.2659	0.9368	0.1373

The model shows **steady improvements in pixel accuracy and mIoU**, showing that it successfully learns meaningful segmentation patterns from the images.

Visual Results

Below are examples from the validation set, showing:

At first epoch,



At 6th epoch,

At 8th epoch,

Challenges and Scope of Improvement

- **Backbone Frozen**: Only the last block of the transformer is fine-tuned; fully unfreezing may yield better performance.

- **No Skip Connections**: U-Net skip connections could help preserve fine-grained spatial details.

- **Loss Function Options**: Alternative losses like Dice or Focal Loss may improve class-wise performance for underrepresented classes.

- **Training Schedule**: Using learning rate schedulers (e.g., cosine annealing or OneCycleLR) could boost convergence and generalization.

Presentation Link - https://youtu.be/h-K6XyY-x-w?si=bvMmiE8Rc742R1_K