Semantic Segmentation on Pascal VOC 2007: U-Net, DeepLabV3+, and SAM2

Abstract

I studied multi-class semantic segmentation on Pascal VOC 2007 using U-Net, DeepLabV3+, and SAM2. Due to computation constraints, all models were trained and evaluated at 256×256 resolution. I hypothesize that higher input resolution would improve boundary quality and small-object recognition. I compared performance across models and losses using standard metrics (mIoU, mean Dice, HD95, Pixel Accuracy, per-class performance), included qualitative mosaics and best/worst cases, and conducted ablations on backbone size (U-Net ResNet-18 vs ResNet-34; DeepLab ResNet-50 vs ResNet-101) and loss (cross-entropy vs Dice+CE). The default training setup used AdamW with lr=1e-4 and weight_decay=1e-4.

Dataset and Setup

- Dataset: Pascal VOC 2007 (20 foreground classes + background). I followed the provided splits and used the validation set as our test set.
- Preprocessing: images resized to 256×256.
- Losses: Two supervision objectives were compared:
  - Cross-Entropy (CE): the standard multi-class pixel-wise loss used as the baseline.
  - Dice + CE (hybrid): a balanced combination of CE (weight = 0.5) and a soft multi-class Dice term (weight = 0.5) that mitigates class-imbalance effects and stabilizes overlap metrics. The Dice term computes per-class overlaps from softmax probabilities, applies a smoothing constant = 0.1, and ignores void pixels (index 255). A light label-smoothing (ε = 1.0) was included in CE to regularize noisy boundaries.
- Optimizer: AdamW (lr=1e-4, weight_decay=1e-4). I explored a cosine scheduler and data augmentation; neither yielded material improvements, so results are omitted for brevity.
- Evaluation: Mean Dice, mIoU, 95th-percentile Hausdorff distance (HD95), Pixel Accuracy; per-class IoU and Accuracy are reported in artifacts.

Models

- U-Net (encoders: ResNet-18, ResNet-34, ResNet-50)
- DeepLabV3+ (backbones: ResNet-50, ResNet-101)
- SAM2 (fine-tuned variants): frozen SAM2.1 Hiera-L image encoder + lightweight segmentation head consisting of two Conv(3×3)-BN-ReLU blocks (in_ch→256→256) followed by a 1×1 convolution to 21 class logits, then bilinear upsampling to 256×256. Alternative heads (ASPP, deeper 4–6 block stacks, attention gates, mini U-Net with skip upsampling) were tested but did not yield consistent gains under identical training budget.

Results (Main Comparison)

Metrics are reported on the validation-as-test split. Rates are in %; HD95 in pixels.

Model	Pixel Acc	mIoU	Mean Dice	HD95

Model	Pixel Acc	mIoU	Mean Dice	HD95
U-Net (ResNet-18, CE)	78.69	13.88	19.20	15.54
U-Net (ResNet-34, CE)	83.06	30.42	42.99	12.90
U-Net (ResNet-34, Dice+CE)	84.38	31.52	44.62	12.59
DeepLabV3+ (ResNet-50, CE)	86.60	42.43	56.15	19.62
DeepLabV3+ (ResNet-101, CE)	86.67	42.41	56.23	24.51
DeepLabV3+ (ResNet-50, Dice+CE)	87.23	46.12	62.08	21.37
DeepLabV3+ (ResNet-101, Dice+CE)	86.95	45.88	61.35	22.10
SAM2 (frozen encoder, CE)	80.14	22.37	33.42	14.11
SAM2 (frozen encoder, Dice+CE)	81.62	24.73	35.61	13.58

Artifacts: see `metrics/` JSONs for full per-class breakdowns and confusion matrices (`*_cm.png`).

Observations

Overall performance.

DeepLabV3+ clearly achieves the strongest overall segmentation quality among all models. At 256×256, it surpasses U-Net by a wide margin on both mIoU and mean Dice, reflecting superior region consistency and finer delineation of object boundaries. U-Net performs reasonably well given its simplicity, but its plain decoder struggles to recover thin or highly textured structures.

Within U-Net, upgrading the encoder from ResNet-18 to ResNet-34 yields a substantial gain (+16.5 mIoU), confirming the value of richer encoder features even under limited input resolution. Extending further to ResNet-50 provided no measurable benefit—likely due to diminishing returns from deeper layers when spatial detail is already lost at 256 px inputs.

For DeepLab, moving from ResNet-50 to ResNet-101 brings almost no improvement, suggesting that the available information at this scale saturates model capacity. Both DeepLab variants maintain high accuracy (~87 %) and strong overlap (≈45 mIoU), showing that contextual aggregation through atrous convolutions and the decoder contributes more than raw depth to performance at this resolution.

SAM2, despite its large frozen Hiera-L encoder, lags behind task-specific architectures, indicating that fixed prompt-centric features transfer poorly to dense, fixed-class segmentation without adaptation.

Effect of loss.

Adding Dice to cross-entropy improves DeepLab-50 on mIoU (+3.7 pts, 42.43→46.12) and mean Dice (+5.9 pts, 56.15→62.08), with a modest HD95 rise (19.62→21.37), implying sharper but slightly rougher boundaries.

DeepLab-101 shows similar gains (42.41→45.88 mIoU; 56.23→61.35 Dice), while its HD95 drop (24.51→22.10) suggests mildly improved boundary stability for the deeper encoder.

For SAM2 (frozen encoder), Dice+CE yields small but consistent improvements, though overall performance remains behind task-specific architectures.

Training efficiency.

Training time per epoch correlates with encoder size and decoder complexity. SAM2 is the slowest due to its large frozen Hiera-L backbone and heavy feature extraction pipeline, even though only the decoder is updated.

DeepLabV3+ trains moderately fast—its atrous backbone adds computation, but the decoder is lightweight.

U-Net is the most efficient, converging quickly within 40–50 epochs; DeepLab typically requires slightly longer to stabilize (~50–60 epochs) for the same learning rate and optimizer settings.

Per-class tendencies.

Across top runs, large contiguous objects (background, person, car, bus/train, dog) achieve high IoU and Dice; mid-sized textured classes (cow, horse, cat) remain moderately reliable.

Small, thin, or cluttered objects (bottle, chair, potted plant, dining table, sofa, tv/monitor, bicycle) contribute most confusion—often missed or fragmented.

DeepLab reduces small-object fragmentation more effectively than U-Net, likely owing to its atrous spatial pyramid context, whereas U-Net tends to over-smooth fine structures.

SAM2 frozen variants under-segment small or adjacent objects, often merging them into background.

Qualitative Visualizations

Representative mosaics, confusion matrices, and best/worst examples are included per model. Best/worst are ordered by image-level IoU.

U-Net (ResNet-34, Dice+CE)

Mosaic:

Sample 35

Sample 11

Sample 17

Sample 24

Sample 50

Sample 67

Best examples:

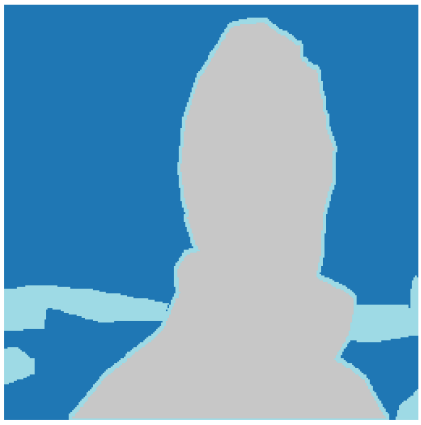
Input

Ground Truth

Prediction (IoU_person=0.957)

Input

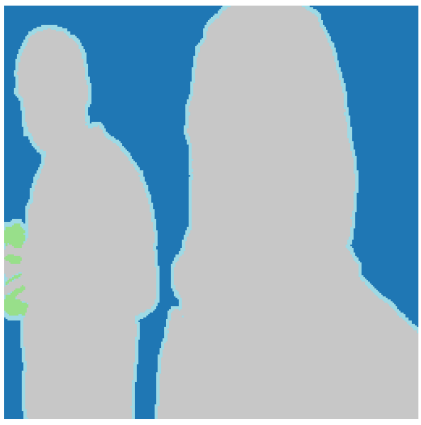
Ground Truth



Prediction (IoU_person=0.948)

Input

Ground Truth

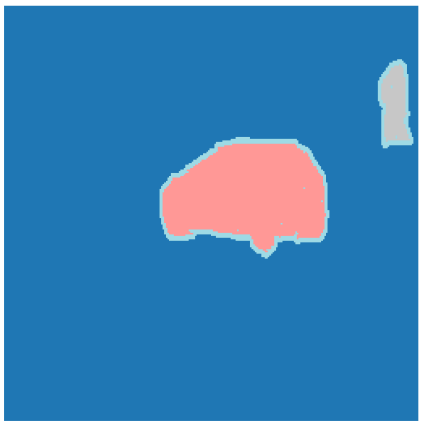


Prediction (IoU_person=0.932)

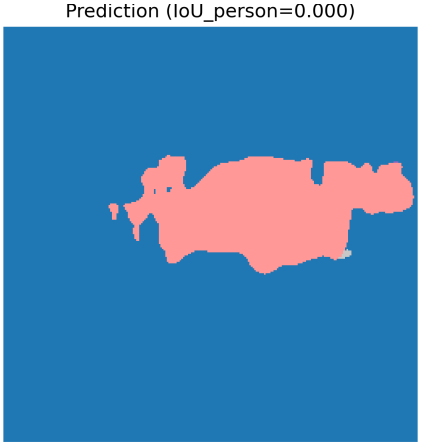
Worst examples:

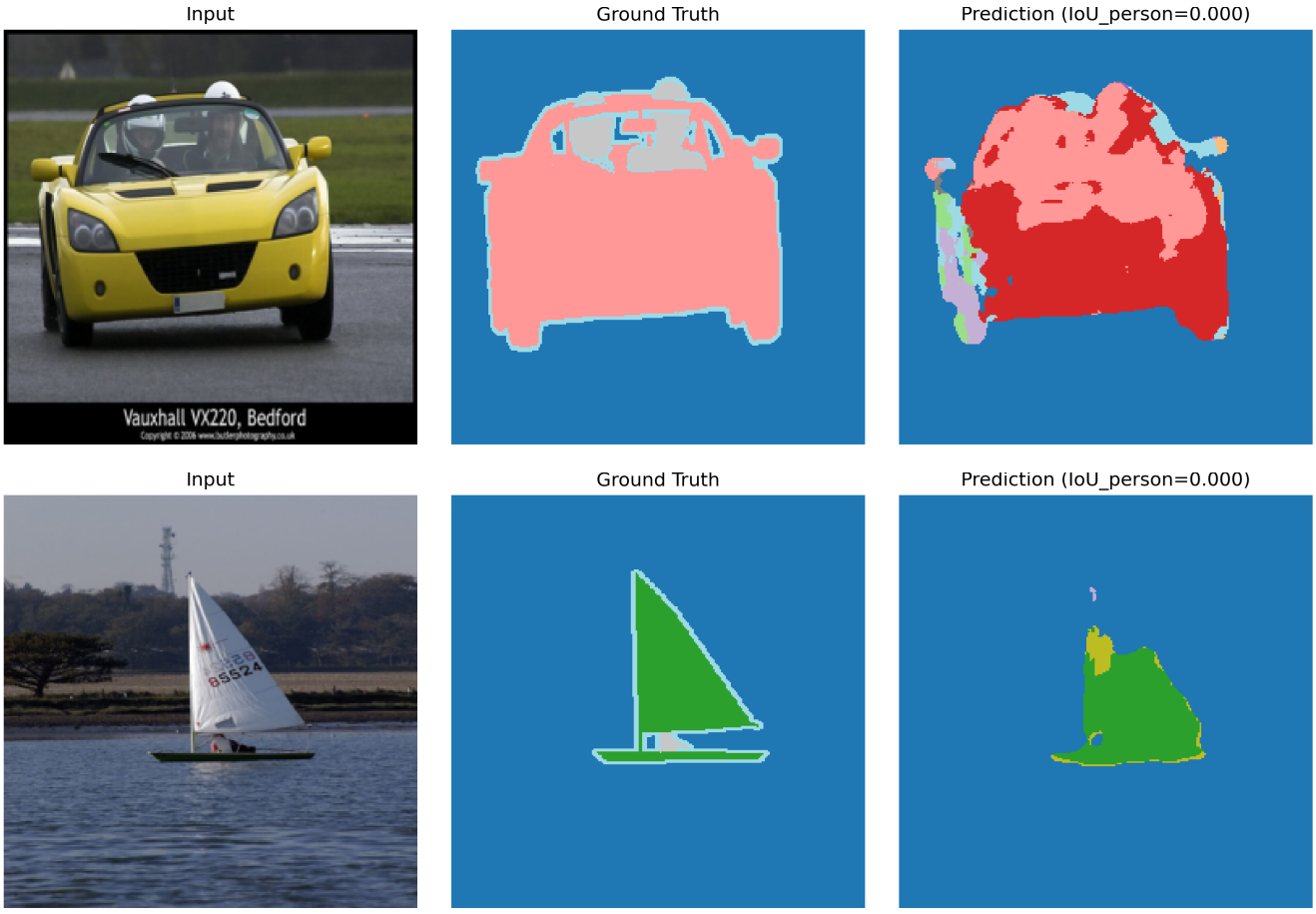
Input

Ground Truth

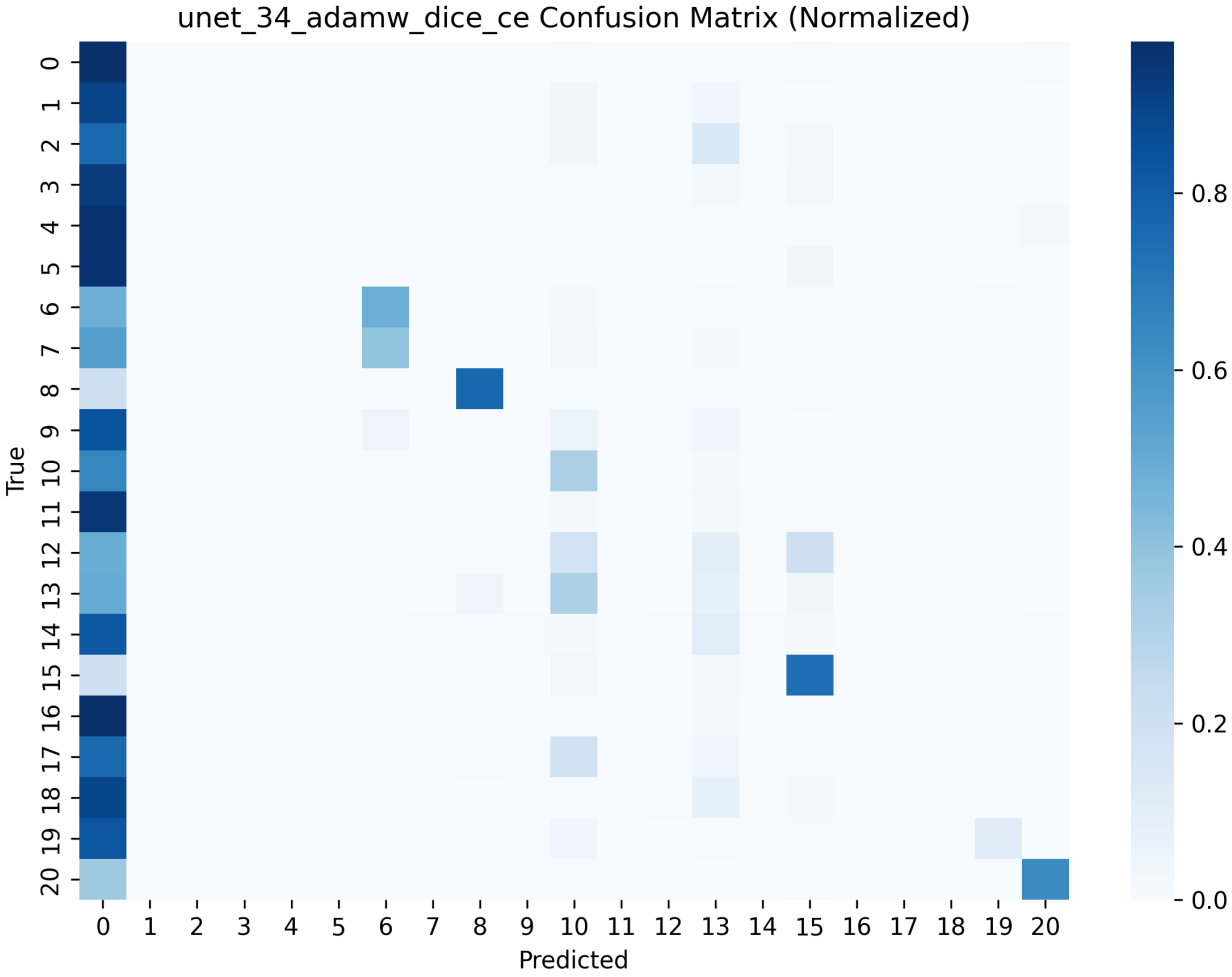


Prediction (IoU_person=0.000)





Confusion Matrix:



DeepLabV3+ (ResNet-50, Dice+CE)

Mosaic:

Sample 50

Sample 53

Sample 18

Sample 24

Sample 58

Sample 17

Best examples:

Input

Ground Truth

Prediction (IoU_person=0.936)

Input

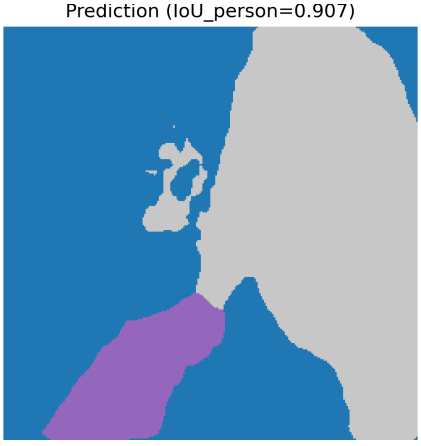
Ground Truth

Prediction (IoU_person=0.930)

Input

Ground Truth

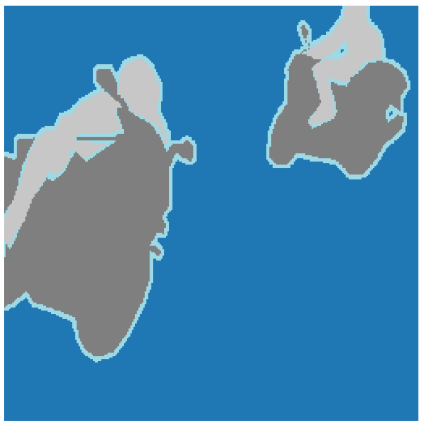
Prediction (IoU_person=0.907)



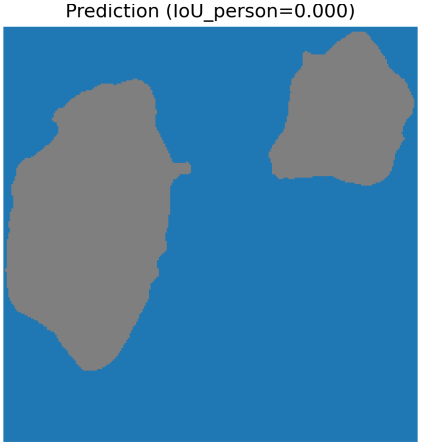
Worst examples:

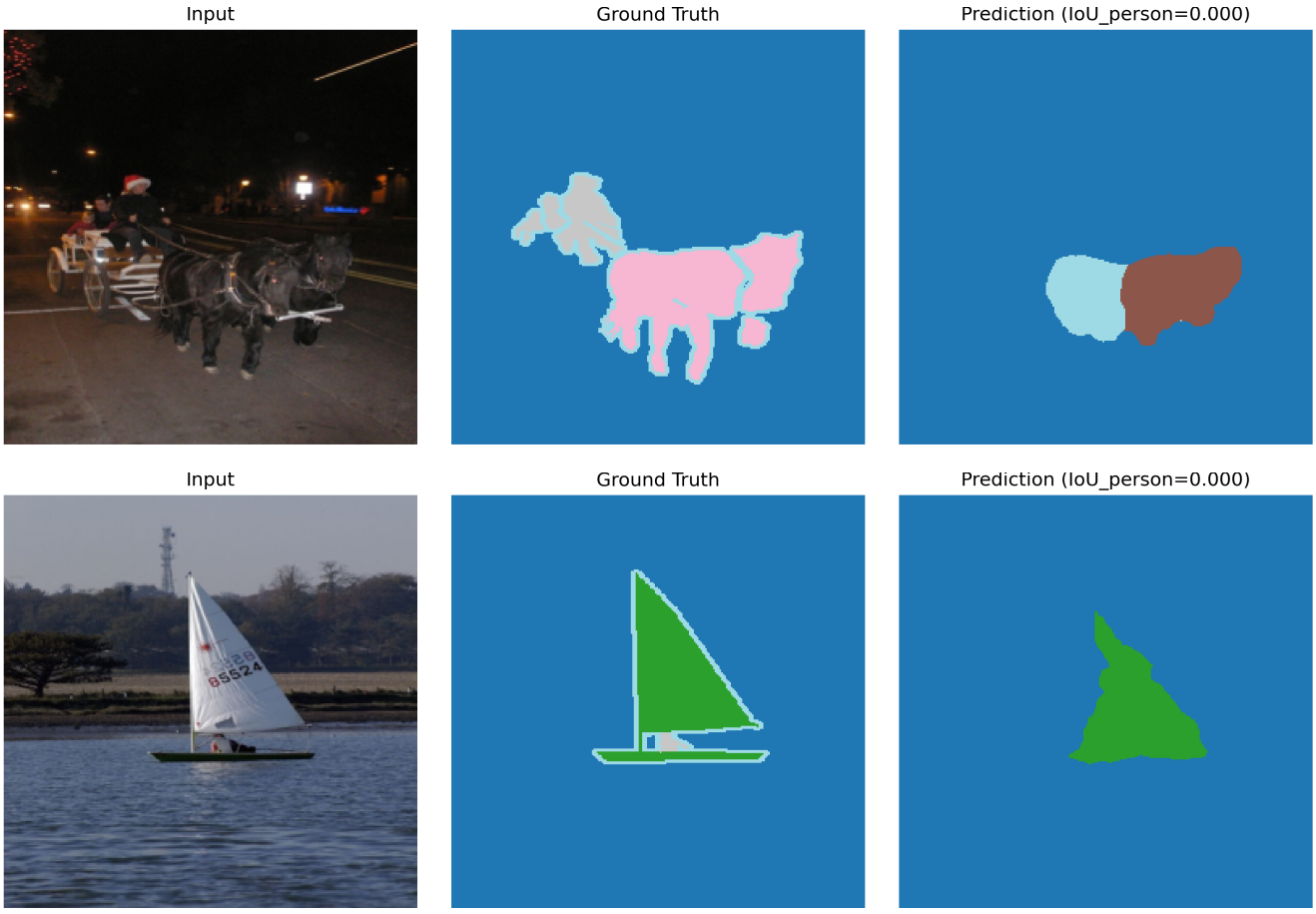
Input

Ground Truth

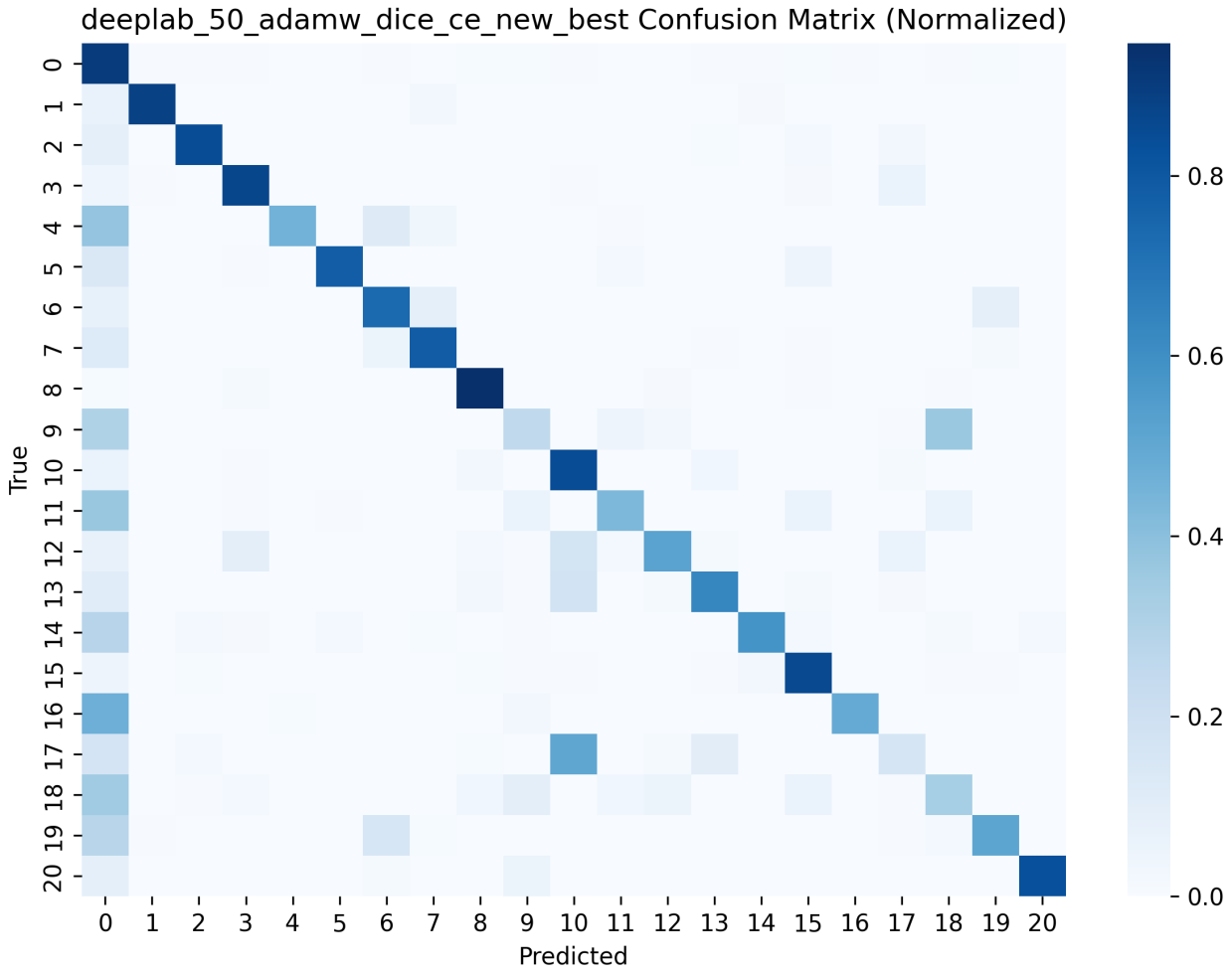


Prediction (IoU_person=0.000)



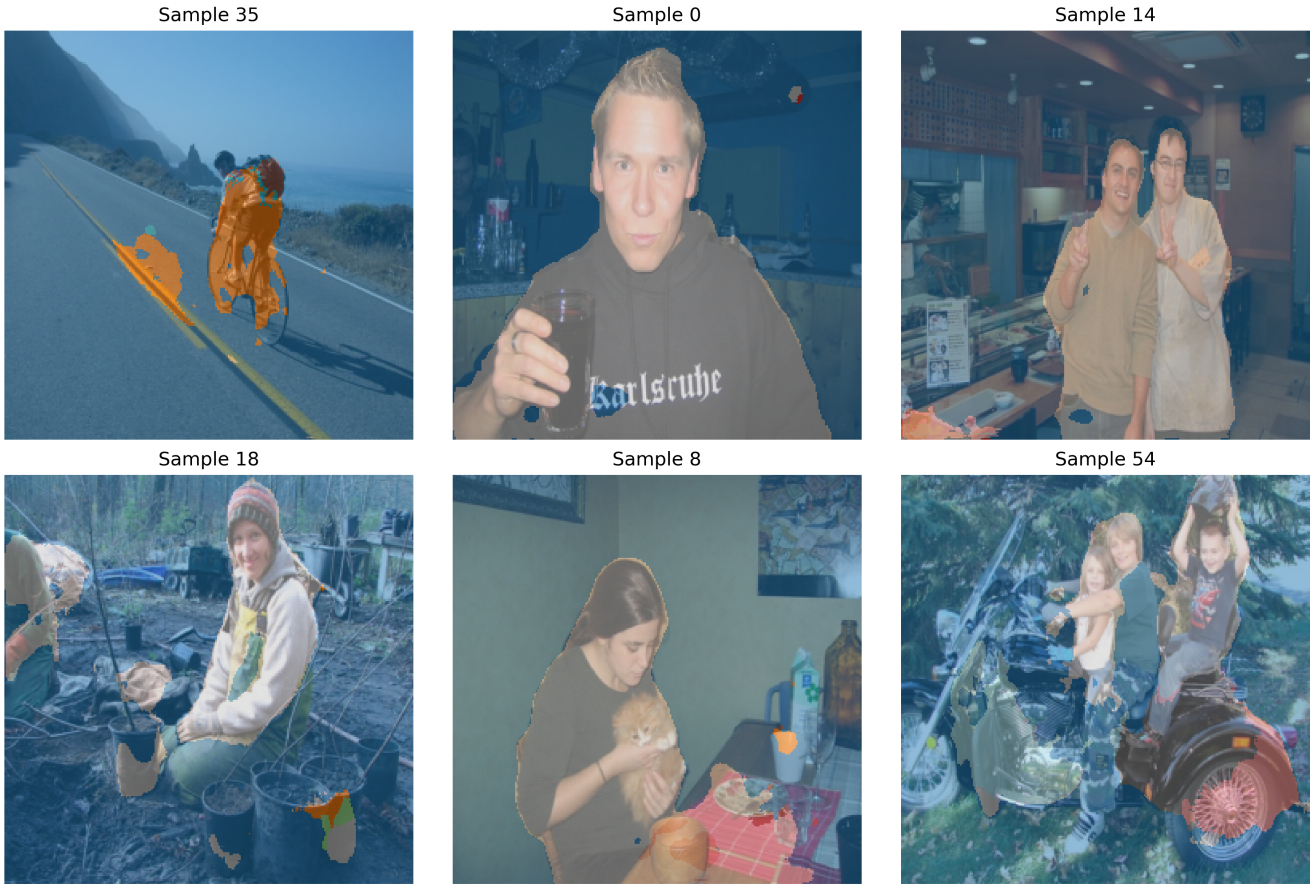


Confusion Matrix:

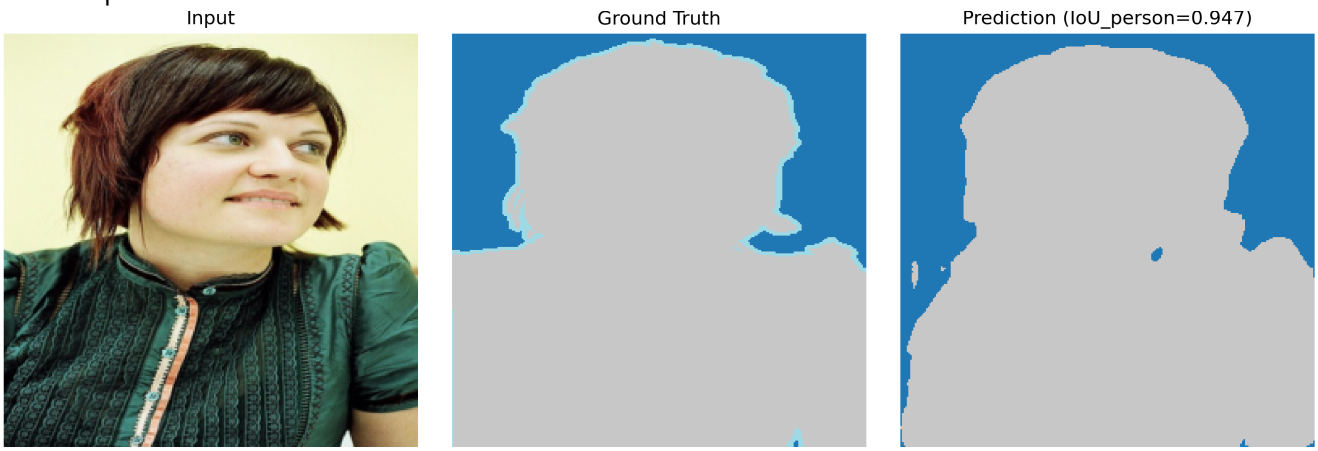


SAM2 (Frozen Encoder, Dice+CE)

Mosaic:

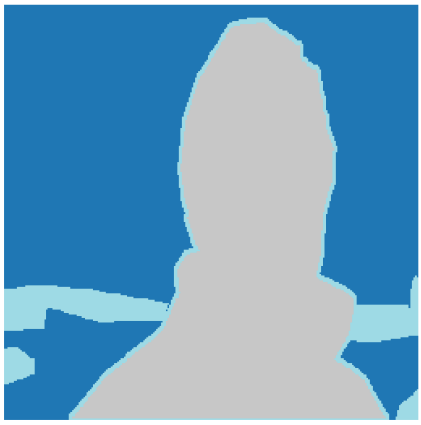


Best examples:



Input

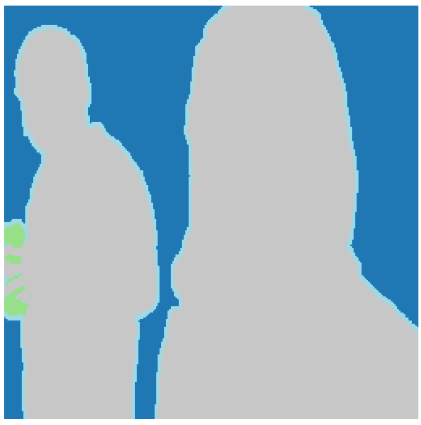
Ground Truth



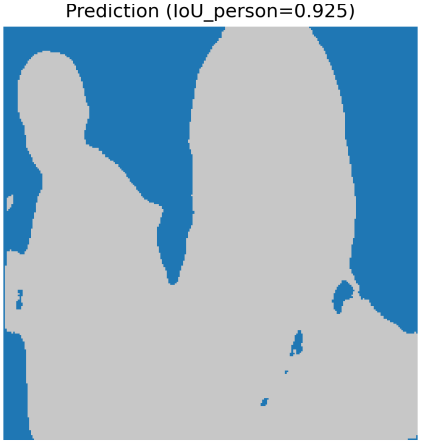
Prediction (IoU_person=0.946)

Input

Ground Truth



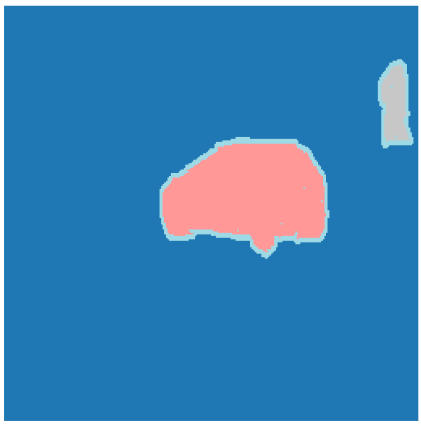
Prediction (IoU_person=0.925)



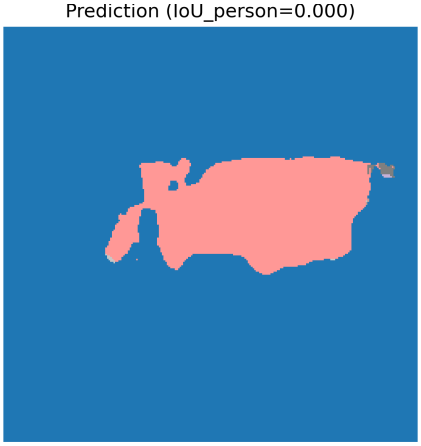
Worst examples:

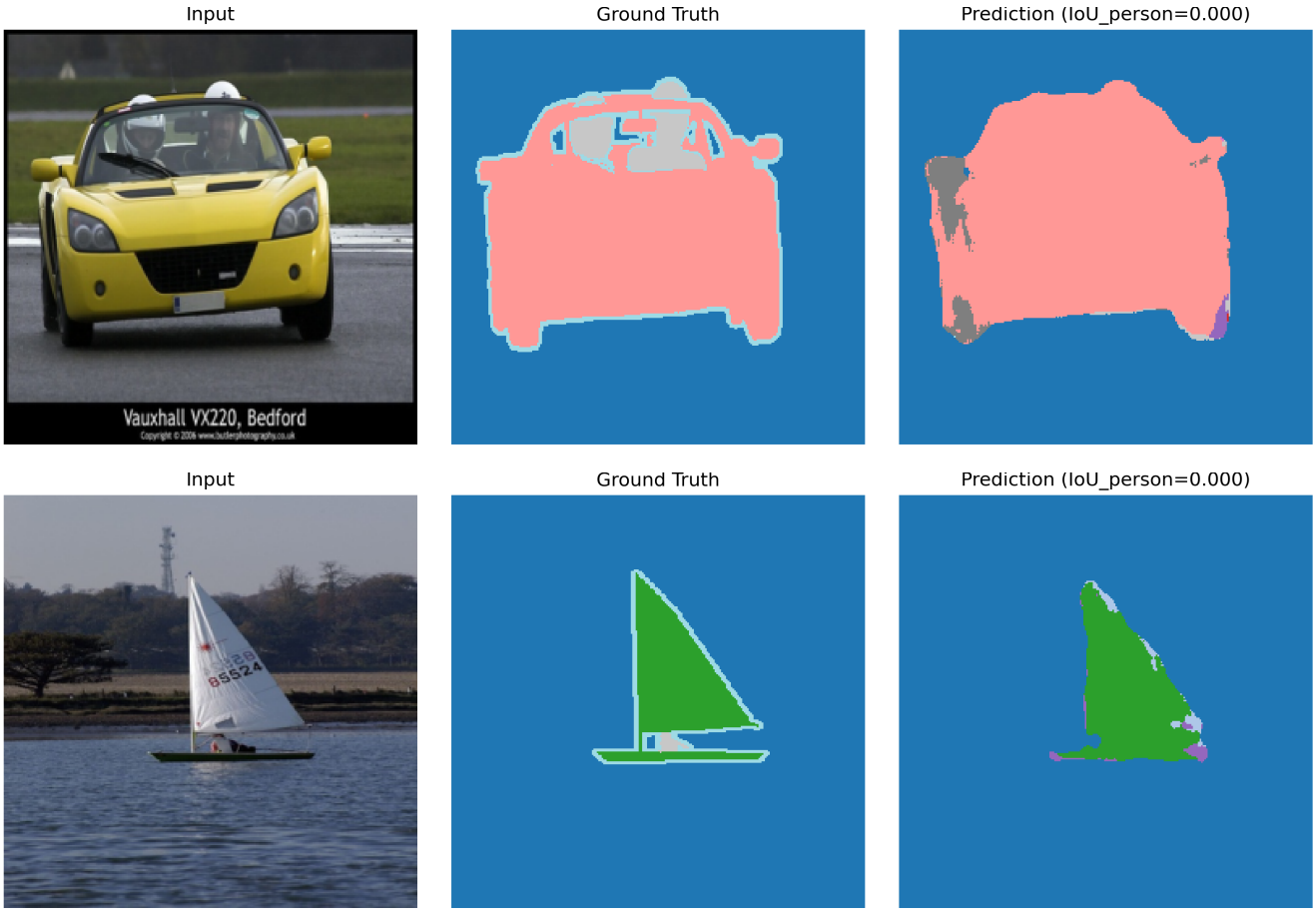
Input

Ground Truth

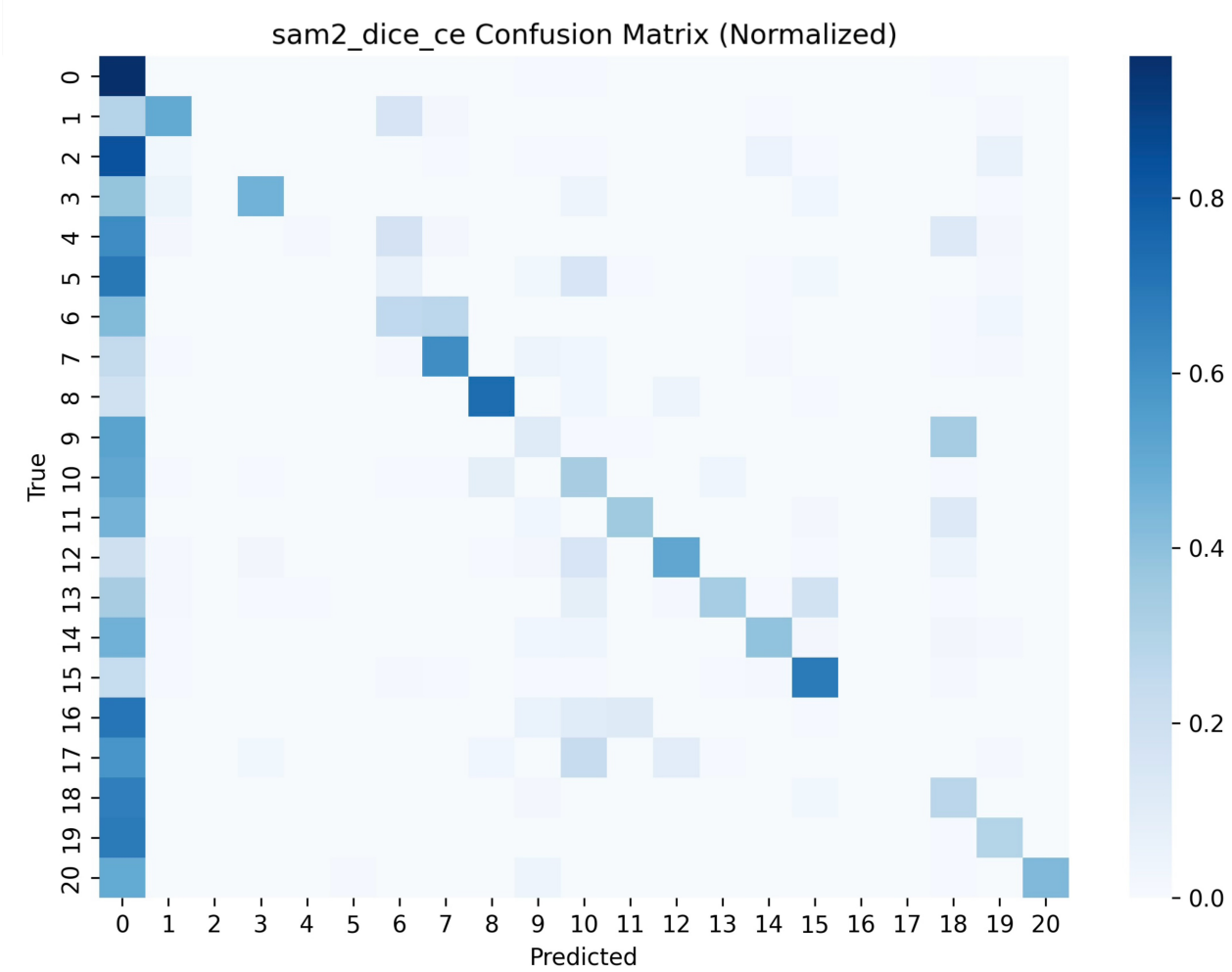


Prediction (IoU_person=0.000)





Confusion Matrix:



Qualitative error modes: Best cases show clean foreground separation and coherent interiors; DeepLab preserves limb and wheel detail better than U-Net. Worst cases collapse fine objects (bottle, plant, chair legs) into background or produce holes. SAM2 also shows that higher resolution or prompt guidance is needed for fine detail recovery.

Notes on the human class ("person"): DeepLab variants consistently deliver higher IoU for the person class compared to U-Net-34 CE under identical 256×256 conditions, corroborating quantitative results.

Ablation Studies

We include two focused ablations, keeping optimizer and hyperparameters fixed (AdamW, lr=1e-4, weight_decay=1e-4).

(A) Backbone size

- U-Net: ResNet-18 → ResNet-34 improves mIoU from 13.9% to 30.4% and mean Dice from 19.2% to 43.0%, with lower HD95. Larger encoder capacity helps recover finer details even at 256×256.
- DeepLabV3+: ResNet-50 vs ResNet-101 shows virtually no change in mIoU (~42.4% both). At this input size/training budget, the deeper backbone doesn't translate to measurable gains.

(B) Loss: CE vs Dice+CE

- U-Net-34: Dice+CE provides a modest improvement over CE (mIoU 30.42% → 31.52%; Dice 42.99% → 44.62%) with nearly unchanged HD95 (12.90 → 12.59), indicating slightly better overlap without boundary penalty.
- DeepLab-50: Dice+CE improved mIoU (46.12% vs 42.43%) and mean Dice (62.08% vs 56.15%) with a modest HD95 increase (21.37 vs 19.62), indicating sharper masks at modest boundary cost.
- DeepLab-101: Dice+CE improved mIoU (45.88% vs 42.41%) and mean Dice (61.35% vs 56.23%) while slightly reducing HD95 (22.10 vs 24.51), suggesting deeper encoder stabilizes boundaries.
- SAM2 (frozen encoder): Dice+CE yields a minor but consistent gain (mIoU 22.37 → 24.73; Dice 33.42 → 35.61) and a slight HD95 reduction (14.11 → 13.58), implying improved region consistency and smoother boundaries.

Discussion and Lessons Learned

At 256×256 resolution, several consistent trends emerge across architectures and loss functions.

Model architecture. DeepLabV3+ clearly surpasses U-Net on overlap (mIoU/Dice) and boundary quality (HD95), benefiting from its atrous pyramid context and decoder refinement. Increasing encoder depth (ResNet-50 → 101) yields only marginal gains, suggesting that input resolution, not backbone capacity, is the dominant constraint. In contrast, U-Net's simple upsampling decoder limits fine structure recovery, leading to over-smoothed boundaries and missed thin objects.

Loss design. Adding Dice to cross-entropy consistently improves overlap metrics by mitigating class-imbalance effects. The benefit is modest for U-Net-34 but more pronounced for DeepLab variants, which already have strong localization. Dice sharpens mask edges but can slightly raise HD95; this trade-off is mild and acceptable when combined with a stable optimizer such as AdamW.

Overlap–boundary trade-off. Dice tends to boost region overlap while sometimes exaggerating boundary noise. DeepLab-50 shows a small HD95 increase (19.6→21.4), whereas DeepLab-101 exhibits a minor HD95

decrease (24.5→22.1), implying that deeper encoders may stabilize boundary predictions under hybrid losses. Reporting mIoU, Dice, and HD95 jointly gives a more balanced view of segmentation quality.

Class difficulty and resolution effects. Large contiguous categories (background, person, vehicle) are reliably segmented; mid-size textured classes (cat, horse, dog) remain moderate; thin or small objects (bottle, chair, potted plant, bicycle) are the major failure modes. Downsampling to 256×256 severely limits small-object recall and boundary fidelity. Remedies include higher-resolution training, class-balanced or focal losses, and multi-scale inference.

SAM2 behavior. In this regime, SAM2 with a frozen encoder underperforms task-specific architectures. Its prompt-centric pretraining and fixed encoder hinder adaptation to fixed-class segmentation. Although the Dice+CE variant improves modestly, fine-detail recovery remains weak. Effective adaptation would likely require partially unfreezing late encoder blocks (with lr≈1e-5), training longer at higher resolution, and integrating prompt or text embeddings.

Practical takeaways. Under constrained compute, purpose-built dense-prediction models such as DeepLabV3+ and U-Net remain superior to prompt-based foundations. Dice+CE provides a reliable, lightweight improvement across models. SAM2's advantage lies not in low-resolution finetuning but in flexible, high-resolution prompt-guided inference—an avenue for future exploration.

Limitations and Future Work

- Computational budget constrained inputs to 256×256. Future work should evaluate 512×512 or multi-scale inputs.
- Explored learning rate scheduling and data augmentation; neither produced consistent improvements in this scenario, and thus the relevant results are omitted in this report.
- Investigate decoders with attention or OCR heads, and test class-balanced losses.
- SAM2-specific: progressively unfreeze late encoder blocks; integrate class prompt embeddings (text or learned tokens); add contrastive/distillation losses; move to higher resolution + multi-scale inference; extend schedule with adaptive LR restarts.

Reproducibility and Artifacts

- Metrics and confusion matrices: see `metrics/` (`*_metrics.json`, `*_cm.png`).
- Training curves: see `*_history.csv` per model in `metrics/`.
- Qualitative images: see `visualizations/`