Semantic Segmentation with Pascal VOC 2007 Dataset

Qianyu Fan

1. Dataset

The dataset used in this project is the PASCAL VOC 2007 segmentation dataset, publicly available on Kaggle. It contains twenty object classes and one background class. The original dataset provides 209 training images and 213 validation images. In this project, the validation set is designated as the test set, while the original training set is further split into training and validation subsets using an 8:2 ratio. All images and corresponding segmentation masks are resized to 256×256 pixels to ensure consistent input dimensions during model training.

2. Methods

2.1. U-Net

Ronneberger et al. [3] introduced the U-Net architecture in 2015, our model is built upon this foundational design (Figure 1).

To refine the original design, our implementation employs lightweight dual-convolution blocks, each comprising two 3×3 convolutional layers followed by batch normalization and ReLU activation, enhancing feature representation and training stability. Downsampling is achieved via max pooling to progressively capture higher-level semantics, while the decoder restores spatial resolution using bilinear upsampling, providing a computationally efficient alternative to transposed convolutions. Skip connections between encoder and decoder stages preserve fine-grained information throughout the network. A final 1×1 convolution generates pixel-wise class logits.

This refined U-Net maintains the representational strengths of the original architecture while improving computational efficiency and training stability, making it well suited for segmentation tasks under constrained computational resources.

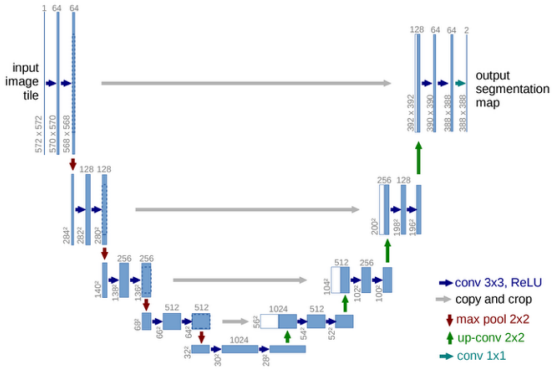

Figure 1. U-Net Architecture from Original Paper

2.2. DeepLabV3+

DeepLabv3+ [1], introduced by Chen et al. (2018), enhance semantic segmentation by employing atrous convolution and an encoder-decoder structure, respectively, setting new benchmarks for image segmentation efficiency and precision.

We employ the DeepLabV3+ model (Figure 2) with a ResNet-50 backbone to leverage its advanced segmentation capabilities. The ResNet-50 encoder, known for its residual connections, enhances model capacity and training stability. This configuration uses weights pre-trained on the ImageNet dataset, providing a strong initialization for feature extraction necessary for accurate and detailed segmentation.

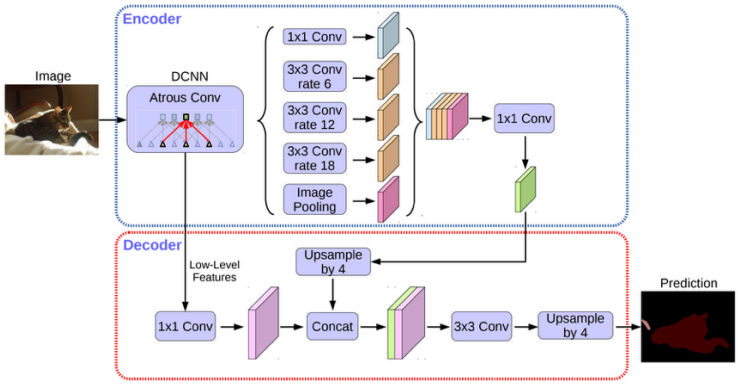

Figure 2. DeepLabV3+ Architecture from Original Paper

2.3. SAM2

The Segment Anything Model 2 (SAM2) [2] is a state-of-the-art segmentation model introduced by Meta AI in 2024, designed for general-purpose image segmentation. SAM2 employs a promptable architecture consisting of an image encoder, a flexible prompt encoder, and a lightweight mask decoder. The image encoder extracts high-level visual features from the input image, while the prompt encoder encodes user-provided prompts such as points, boxes, or masks. The mask decoder then predicts segmentation masks based on the combined image and prompt embeddings, enabling accurate and efficient mask generation for diverse objects and regions.

In our approach, we leverage SAM2, specifically `sam2.1_hiera_large`, for class-specific segmentation. SAM2 is used in inference mode only, without any gradient updates. For each class in an image, a coarse bounding box is generated based on approximate object regions. This bounding box serves as the input prompt for SAM2, guiding the model to focus on the relevant object or region, and enabling it to generate the segmentation mask for the corresponding class. This approach provides minimal guidance while remaining fully automated, allowing SAM2 to produce accurate class-specific masks.

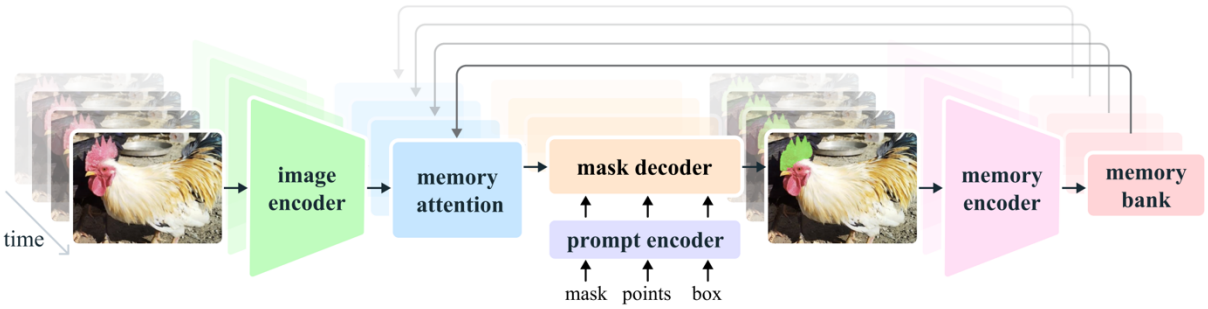

Figure 3. SAM2 architecture

3. Experiments

3.1 Loss Function

To train the U-Net and DeepLabV3+ semantic segmentation models, we adopt a hybrid loss function that combines the pixel-wise multi-class cross-entropy loss with a soft Dice loss term. This combination encourages both accurate pixel classification and improved overlap between predicted and ground-truth segmentation masks, especially beneficial for classes with limited foreground representation.

Formally, for a predicted segmentation output $\hat{Y} \in R^{C\times H\times W}$ and corresponding ground-truth mask $Y \in \{0, \dots, C-1\}^{H\times W}$, our training loss is defined as: $\mathcal{L} = \mathcal{L}_{\text{CE}}(\hat{Y}, Y) + \lambda\, \mathcal{L}_{\text{Dice}}(\hat{Y}, Y)$ where λ=0.02 is a weighting hyperparameter that balances the contribution of the Dice loss.

3.1.1 Dice Loss

The Dice loss function is universally applied for our methods to evaluate the similarity between two binary masks. Consider two binary masks, *M1* and *M2*, which correspond to the set of pixels each mask covers. The Dice coefficient for these masks is calculated as follows:

$$\text{Dice}(M_1, M_2) = \frac{2|M_1 \cap M_2|}{|M_1| + |M_2|}$$

where $|\cdot|$, denotes the number of pixels in the set. The coefficient ranges between 0 and 1, where a Dice coefficient of 1 indicates perfect overlap between the masks and 0 indicates no overlap.

For *M1* as the predicted mask and *M2* as the true mask, the Dice loss is defined as:

$$\mathcal{L}_{Dice} = 1 - \text{Dice}(M_1, M_2)$$

For multi-class segmentation with *C* classes, the soft Dice loss is computed per class and averaged:

$$\mathcal{L}_{Dice} = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_i p_{i,c}\ g_{i,c}}{\sum_i p_{i,c} + \sum_i g_{i,c} + \epsilon}$$

Our objective is to minimize this weighted combination of cross-entropy loss and Dice loss to improve the accuracy of the predicted mask in matching the true mask.

3.1.2 Cross-entropy

Pixel-wise multi-class cross-entropy [4] is the standard loss for semantic segmentation, measuring the divergence between the predicted class distribution at each pixel and the true class label. It is defined as:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{1}(Y_i = c)\log\widehat{p_{i,c}}$$

3.2 Evaluation Metrics

3.2.1 Dice Coefficient

The Dice coefficient measures the degree of overlap between the predicted mask *M1* and the ground truth mask *M2*, treating prediction as a set similarity problem. It is defined as:

$$\text{Dice}(M_1, M_2) = \frac{2|M_1 \cap M_2|}{|M_1| + |M_2|}$$

where $|\cdot|$, denotes the number of pixels in the set. The coefficient ranges between 0 and 1, where a Dice coefficient of 1 indicates perfect overlap between the masks and 0 indicates no overlap.

The Dice coefficient is particularly sensitive to the size of the overlapping region and is robust for imbalanced datasets. For multi-class segmentation, Dice is computed per class and averaged.

3.2.2 Intersection over Union (IoU)

Intersection-over-Union (IoU) is a metric used to evaluate the accuracy of an object detector on a particular dataset.

$$IoU = \frac{|M_1 \cap M_2|}{|M_1 \cup M_2|} = \frac{|M_1 \cap M_2|}{|M_1| + |M_2| - |M_1 \cap M_2|}$$

IoU provides a value between 0 and 1, where 0 means no overlap and 1 means perfect overlap. A higher IoU score indicates a more accurate model. In practice, a threshold (like 0.5) is often used to decide whether predictions are correct.

3.2.3 95th percentile Hausdorff distance (HD95)

The Hausdorff distance evaluates segmentation quality at the boundary level, rather than region overlap. For two sets of contour points P (prediction) and G (ground truth), the directed Hausdorff distance is defined as:

$$d(P, G) = \max_{p \in P} \min_{g \in G} |p - g| d(P, G)$$

To obtain a more stable and robust measure, we adopt the 95th-percentile Hausdorff distance (HD95), which computes the 95th percentile of the point-wise distances instead of the maximum. Lower HD95 values indicate better boundary alignment between the predicted segmentation and the ground truth.

3.2.4 Pixel Accuracy

Pixel accuracy measures the proportion of correctly classified pixels:

$$\text{Pixel Accuracy} = \frac{\text{\# correctly predicted pixels}}{\text{\# total pixels}}$$

3.3 Experiment Setup

To ensure consistency, reproducibility, and fair comparison across different segmentation models, all experiments were performed on Google Colab Pro, using a A100 GPU. Random seeds were fixed across PyTorch, NumPy, and the Python environment to ensure reproducibility of results.

All images in the Pascal VOC 2007 dataset were resized to 256 × 256 for efficient training and evaluation. For U-Net and DeepLabV3+, nearest-neighbor interpolation was applied to the masks, and void pixels (label = 255) were ignored. SAM2 evaluations were performed using the original image resolution to better align with SAM2's design, which benefits from higher-resolution inputs.

Model Training and Inference Parameters:

- **U-Net**: Batch Size: 4, Image Size: 256×256, learning rate: 0.001, Optimizer: Adam, Loss Function: CE+0.02×Dice, Epochs: 30. Total training time: 2.5 minutes.
- **DeepLabV3+:** Batch Size: 4, Image Size: 256×256, learning rate: 0.001, Optimizer: Adam, Loss Function: CE+0.02×Dice, Epochs: 30, Backbone: ResNet-50. Total training time: 4 minutes.

Both U-Net and DeepLabV3+ were trained using the hybrid loss function introduced in Section 3.1.

- **SAM2**: Model: `sam2.1_hiera_large`, Checkpoint: official pretrained weights, Inference mode only, Prompt: bounding box as input prompt encoder.

4. Results and Analysis

4.1 Model Results

Table 1 presents the overall quantitative performance of U-Net, DeepLabV3+, and SAM2 on the Pascal VOC 2007 validation set. SAM2 consistently outperforms both U-Net and DeepLabV3+. Its mean dice coefficient of 0.8007 and mean IoU of 0.7255 reflect highly accurate region-level predictions, while its HD95 value of 21.8963 indicates precise alignment with ground-truth boundaries. Pixel accuracy also reaches 0.9420, again outperforming all baselines. DeepLabV3+ achieves moderate performance, with a mean IoU of 0.4366 and pixel accuracy of 0.8701, reflecting its capacity to capture multi-scale features and global context. U-Net, on the other hand, performs extremely poorly on this challenging multi-class dataset, reaching only 0.0426 mean IoU and 0.6711 Pixel Accuracy, consistent with known limitations of its simple encoder–decoder architecture when applied to natural images containing diverse objects and complex backgrounds.

Table 1. Model results

Model	Mean Dice	Mean IoU	HD95	Pixel Accuracy
U-Net	0.0533	0.0426	73.1726	0.6711
DeepLabV3+	0.5741	0.4366	50.8635	0.8701
SAM2	0.8007	0.7255	21.8963	0.9420

A deeper examination of per-class IoU and pixel accuracy, shown in Table 2, reveals consistent trends across individual object categories. SAM2 produces the strongest results across nearly every class, particularly for large and well-defined objects such as bus, train, cat, tvmonitor, and dog, where IoU scores commonly exceed 0.85 or even 0.90. This indicates excellent generalization across a wide range of object geometries and contexts. DeepLabV3+ also performs well on many common categories, especially animals and vehicles, but shows noticeably weaker results on small or thin structures, including chair, potted plant, bicycle, and sheep, where its IoU drops significantly. These categories typically require fine-grained boundary precision and robust contextual reasoning, which remain challenging for conventional CNN-based architectures. U-Net's performance is the weakest among the three models: it produces zero IoU for most classes except background and person, indicating a substantial inability to generalize to diverse objects in natural scenes without substantial architectural enhancements or pretraining.

Table 2. Per-class IoU and Accuracy for models

	IoU			Pixel Accuracy		
Class	U-Net	DeepLabV3+	SAM2	U-Net	DeepLabV3+	SAM2
background	0.7379	0.8848	0.5258	0.8625	0.9592	0.6237
aeroplane	0	0.6271	0.8637	0	0.6932	0.9856
bicycle	0	0.1921	0.3376	0	0.3530	0.9353
bird	0	0.6689	0.7452	0	0.7390	0.9746
boat	0	0.2831	0.7882	0	0.3041	0.9817
bottle	0	0.2138	0.5724	0	0.2494	0.9610
bus	0	0.6023	0.9117	0	0.7709	0.9724
car	0	0.5743	0.6658	0	0.7219	0.9498
cat	0	0.7771	0.8991	0	0.8934	0.9605
chair	0	0.1180	0.5352	0	0.1359	0.9381
cow	0	0.2602	0.8480	0	0.6363	0.9780
diningtable	0	0.3308	0.5573	0	0.4214	0.8906
dog	0	0.4761	0.8402	0	0.5381	0.9845
horse	0	0.3758	0.8143	0	0.5351	0.9834
motorbike	0	0.4785	0.8497	0	0.6018	0.9648
person	0.1568	0.7272	0.6991	0.6794	0.8566	0.9404
pottedplant	0	0.1611	0.3884	0	0.1702	0.8974
sheep	0	0.1430	0.7648	0	0.1472	0.9184
sofa	0	0.2253	0.8564	0	0.4718	0.9774
train	0	0.4870	0.9175	0	0.5454	0.9793
tvmonitor	0	0.5628	0.8547	0	0.6725	0.9858

4.2 Qualitative Segmentation Maps

Figure 4 presents qualitative segmentation maps for three randomly selected examples containing the person class, comparing ground-truth masks against predictions from U-Net, DeepLabV3+, and SAM2. The visual trends closely mirror the quantitative findings reported earlier. U-Net frequently fails to recover the correct object shape, often producing fragmented masks, missing limbs, or incorrectly merging foreground and background regions. This is particularly noticeable in the first row, where the predicted U-Net mask spreads across the table and food items, demonstrating its limited capacity to separate complex foreground–background boundaries in real-world scenes.

DeepLabV3+ provides substantially improved results. Across all three images, it captures the rough outline of the person class and correctly identifies major body regions such as the head, torso, and arms. However, the predicted masks remain somewhat coarse: boundaries are often smoothed or misaligned, and fine structural details, such as hair contours or the edges of clothing, are not preserved. These observations match the per-class IoU patterns seen in Table 2, where DeepLabV3+ performs reasonably well on common object categories but struggles with precision in detailed boundary regions.

In contrast, SAM2 produces the most accurate and visually coherent segmentation masks. Its predictions closely adhere to ground-truth boundaries, capturing subtle curvature along the head

and shoulders and successfully separating the person from complex backgrounds. SAM2's strong performance persists even in challenging lighting conditions or when the person appears in non-canonical poses, as shown in the first row. The masks generated by SAM2 preserve sharp, detailed edges and exhibit minimal over- or under-segmentation, reflecting the model's superior boundary alignment and robust generalization ability.

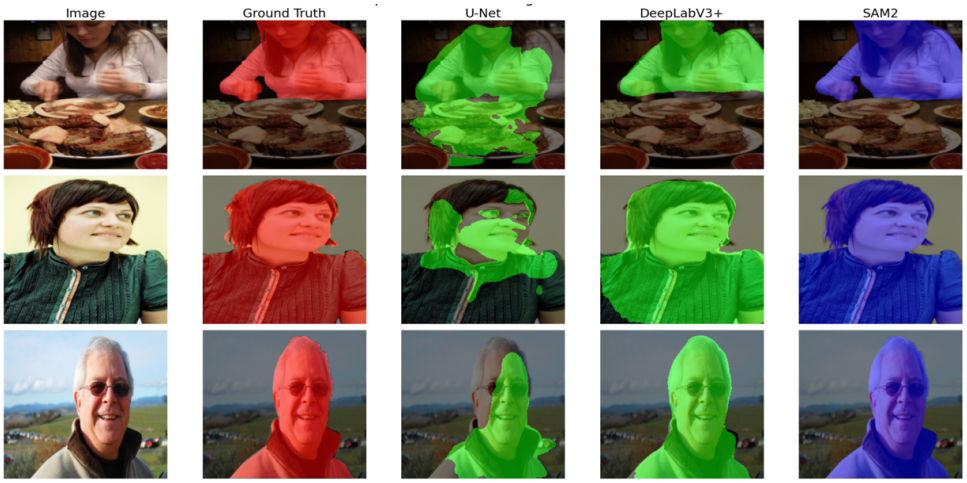

Figure 4. Comparison of Predictions for all Segmentation Models
Left to right: input image, ground-truth mask, U-Net prediction, DeepLabV3+ prediction, and SAM2 prediction

4.3 Qualitative Error Analysis (Best and Worst Results)

To further understand the behavior of each model beyond the average-case qualitative comparisons, we additionally examine the top three best-performing and top three worst-performing segmentation results from person class for U-Net, DeepLabV3+, and SAM2, as shown in Figure 5. These examples were selected based on per-image IoU scores and provide deeper insight into the strengths and failure modes of each approach.

For U-Net, the best predictions reveal that the model can correctly segment simple scenes containing a single, clearly centered object with smooth boundaries and minimal background clutter. In these cases, U-Net produces coarse but generally correct masks that align with the person or animal regions in the image. However, the worst predictions show substantial failure: U-Net often collapses the object shape into an amorphous blob, leaks across object boundaries, or misses large portions of the object altogether. This behavior is especially pronounced in images with multiple objects, textured backgrounds, or occlusions. The model tends to overfit local textures rather than learning robust semantic representations, which explains the extremely low mean IoU and poor per-class performance observed earlier.

DeepLabV3+ demonstrates noticeably stronger and more stable qualitative performance. In its best examples, DeepLabV3+ correctly captures both global object shape and finer structural details.

For instance, it accurately delineates the contours of people, animals, and objects such as boats or windows. These results highlight the benefit of atrous convolutions and multi-scale context aggregation. Nevertheless, the model's worst predictions reveal characteristic weaknesses: DeepLabV3+ often struggles with thin structures (e.g., legs, hair, or object edges), and in more cluttered or low-contrast environments, it may partially mis-segment the object or confuse small regions with background. Unlike U-Net, these errors do not reflect a total model breakdown, but they do demonstrate limitations in fine-grained boundary precision and object differentiation under challenging geometric conditions.

In the top-performing examples, SAM2 produces masks that are almost indistinguishable from the ground truth, capturing detailed silhouette structure and cleanly separating foreground from background even in visually complex scenes. This is consistent with its superior Dice, IoU, and HD95 scores. However, in its worst predictions, SAM2 tends to misclassify the target region as background, resulting in masks that contain almost no foreground pixels. Rather than producing noisy or fragmented segmentations, SAM2's errors typically manifest as an over-suppression of the object, where the entire foreground region is omitted and replaced by background. This failure mode reflects a misinterpretation or insufficient activation of the coarse box prompt, rather than boundary-level inaccuracies.

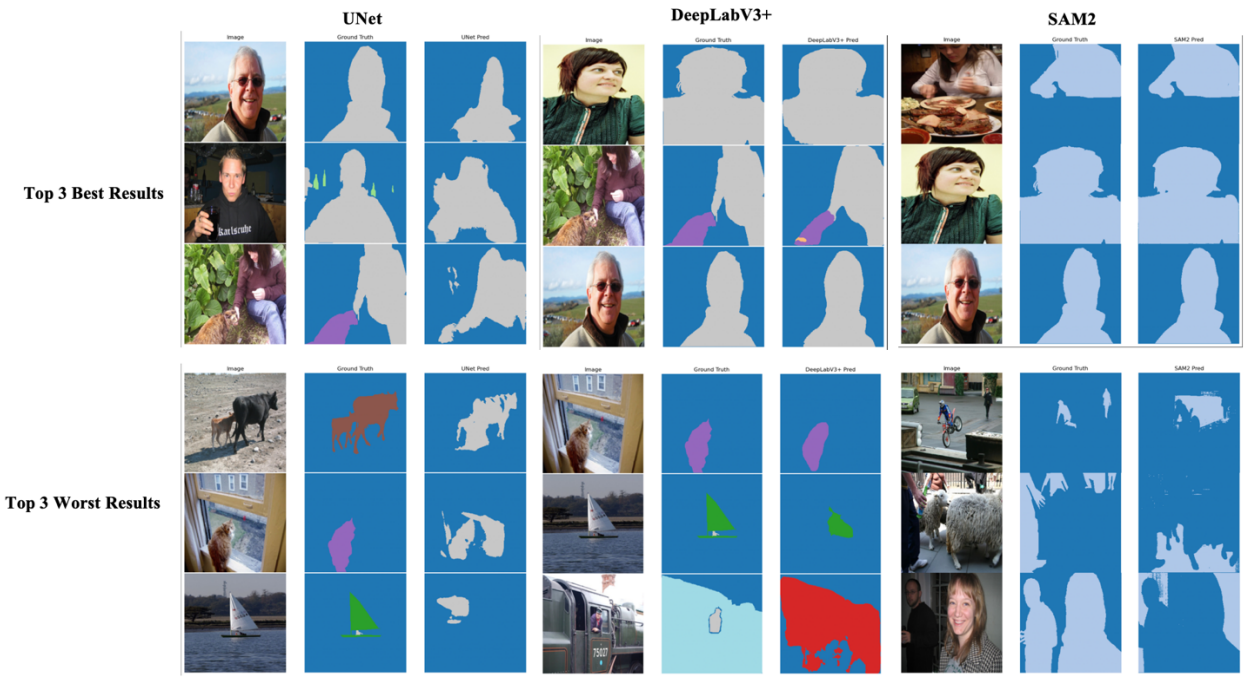

Figure 5. Top 3 Best and Worst Results for all Segmentation Models

4.4 Training and Loss Analysis

Training time in Table 3 presents another significant difference between the models. U-Net trains the fastest, completing in approximately 2.5 minutes, but its extremely poor performance makes it impractical for VOC-style segmentation tasks. DeepLabV3+ requires roughly 4 minutes of training and achieves much better results, illustrating the important gains provided by stronger encoders and atrous spatial pyramid pooling. In contrast, SAM2 requires no training at all; its results arise entirely from zero-shot, prompt-based inference using its large-scale pretrained vision foundation model. Despite having zero training cost, SAM2 dramatically outperforms both supervised baselines, demonstrating the effectiveness of large-scale pretraining and promptable mask generation.

Table 3. Training Time Comparison

Model	Total Training Time
U-Net	2.5 minutes
DeepLabV3+	4 minutes
SAM2	No training (zero-shot with prompts)

The loss curves in Figure 6 for U-Net and DeepLabV3+ highlight clear differences in training stability. U-Net shows a general downward trend in both training and validation loss, but the validation curve fluctuates noticeably. This behavior is typical when using Dice loss, which are highly sensitive to small changes in predicted mask overlap, especially when the foreground region is small, leading to unstable validation behavior. In contrast, DeepLabV3+ exhibits smoother convergence: its training loss decreases steadily due to the stronger ResNet-50 encoder backbone, while the validation loss plateaus early, indicating mild overfitting but overall more stable optimization than U-Net. These patterns are consistent with the quantitative results, where U-Net underfits the VOC dataset while DeepLabV3+ learns more robust features but with limited generalization.

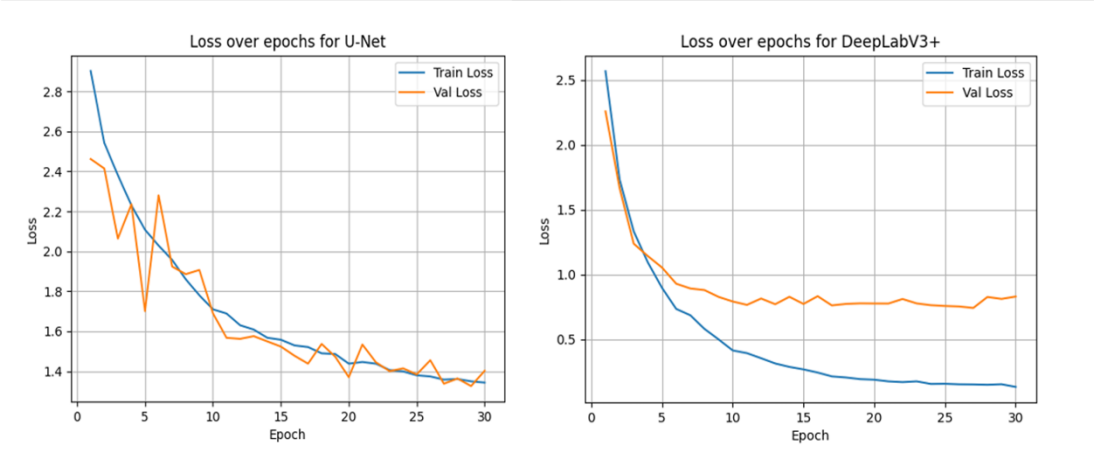

Figure 6. Loss vs Epochs for U-Net and DeepLabV3+

5. Observations and ablation studies

5.1 Model Size

To better understand the limitations of the baseline U-Net model and explore whether its performance on Pascal VOC can be improved, we evaluated U-Net with different encoder backbones: ResNet-18 encoder and ResNet-50 encoder. The results in Table 4 show a clear and consistent trend: stronger encoders substantially improve segmentation quality. The baseline U-Net achieves only 0.0533 mean dice coefficient and 0.0426 mean IoU, confirming that its default encoder is insufficient for capturing the complex, high-frequency features present in VOC images. Replacing the encoder with ResNet-18 yields notable gains, mean dice coefficient increases from 0.0533 to 0.1314, mean IoU more than doubles from 0.0426 to 0.0972, and HD95 decreases from 73.17 to 56.87, indicating tighter and more accurate boundary alignment. The ResNet-50 backbone further amplifies these improvements, reaching 0.1952 mean dice coefficient, 0.1414 mean IoU, and 49.59 HD95, demonstrating that deeper encoders can extract richer semantic representations and improve robustness.

Table 4. Results of U-Net with different encoder backbones

Model	Mean Dice	Mean IoU	HD95	Pixel Accuracy
U-Net (baseline)	0.0533	0.0426	73.1726	0.6711
U-Net (Resnet18)	0.1314	0.0972	56.8749	0.7945
U-Net (Resnet50)	0.1952	0.1414	49.5879	0.8097

Per-class performance (Table 5) confirms the same pattern. The baseline U-Net achieves non-zero IoU on background and person classes, with most categories, particularly animals, vehicles, and small objects, remaining completely unrecognized. ResNet-18 provides partial recovery on several classes, including bus, cat, and person, suggesting that deeper feature extractors help U-Net distinguish more complex object shapes. ResNet-50 extends this further, producing the most consistent per-class improvements and achieving meaningful IoU on categories such as cat, person, motorbike, and tvmonitor. However, even with a ResNet-50 encoder, U-Net still struggles with many small or irregularly shaped objects, reinforcing the inherent limitations of U-Net's decoder architecture relative to more modern segmentation models.

Table 5. Per-class IoU and Accuracy for U-Net with different encoder backbones

	IoU			Pixel Accuracy		
Class	U-Net (baseline)	U-Net (Resnet18)	U-Net (Resnet50)	U-Net (baseline)	U-Net (Resnet18)	U-Net (Resnet50)
background	0.7379	0.8474	0.8724	0.8625	0.9777	0.9685
aeroplane	0	0	0.0001	0	0	0.0001
bicycle	0	0	0	0	0	0
bird	0	0.0015	0.0016	0	0.0015	0.0016
boat	0	0	0.0001	0	0	0.0001
bottle	0	0.0002	0.0004	0	0.0002	0.0004
bus	0	0.2742	0.0452	0	0.5883	0.0647
car	0	0.0248	0.0128	0	0.0317	0.0134
cat	0	0.27	0.5167	0	0.3837	0.9193
chair	0	0.0276	0	0	0.0306	0
cow	0	0.0023	0.0013	0	0.0023	0.0013
diningtable	0	0.0006	0.1445	0	0.0006	0.2465
dog	0	0.0429	0.0719	0	0.074	0.0877
horse	0	0.0337	0.1876	0	0.036	0.2083
motorbike	0	0.0218	0.2817	0	0.0226	0.4791
person	0.1568	0.474	0.5136	0.6794	0.8698	0.8862
pottedplant	0	0	0	0	0	0
sheep	0	0	0	0	0	0
sofa	0	0	0	0	0	0
train	0	0.0032	0.1112	0	0.0036	0.1701
tvmonitor	0	0.018	0.209	0	0.0182	0.3368

Qualitative comparisons, illustrated in Figure 7, visually confirm these trends. The baseline U-Net produces overly smooth, blob-like masks that fail to follow object boundaries. The ResNet-18 variant captures the general shape but still demonstrates noticeable leakage and incomplete foreground coverage. The ResNet-50 encoder yields the most coherent masks, recovering finer details in the head, shoulders, and torso that the baseline model entirely misses. Even so, the predictions remain less accurate and less detailed than those produced by DeepLabV3+ or SAM2, emphasizing that backbone improvements alone cannot fully compensate for architectural limitations.

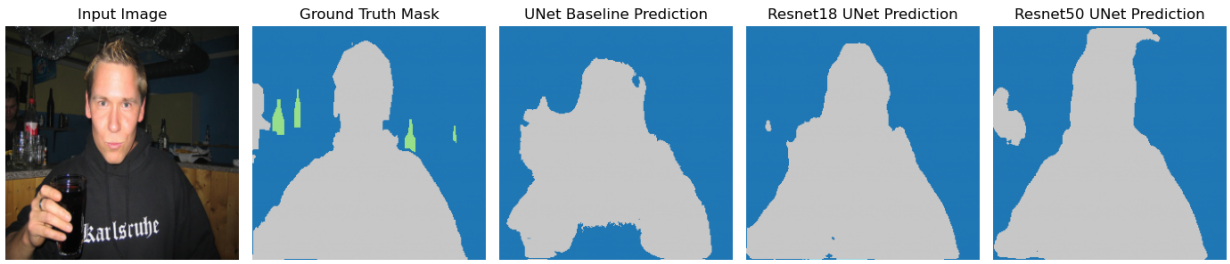

Figure 7. Comparison of Segmentation Map for U-Net with different encoder backbones

The differences in training dynamics across backbones further support these observations. As seen in Figure 8, all three U-Nets exhibit the characteristic fluctuations caused by Dice-based loss functions, which are sensitive to small changes in predicted overlap. However, models with deeper backbones display smoother and more stable convergence. Both the ResNet-18 and ResNet-50 variants show a faster reduction in training loss and a more gradually decreasing validation loss compared to the baseline U-Net. This indicates that deeper encoders not only improve representational capacity but also facilitate more stable optimization. Still, the persistent gap

between training and validation curves suggests that all U-Net variants experience some degree of overfitting on VOC, likely due to the limited dataset size relative to model capacity.

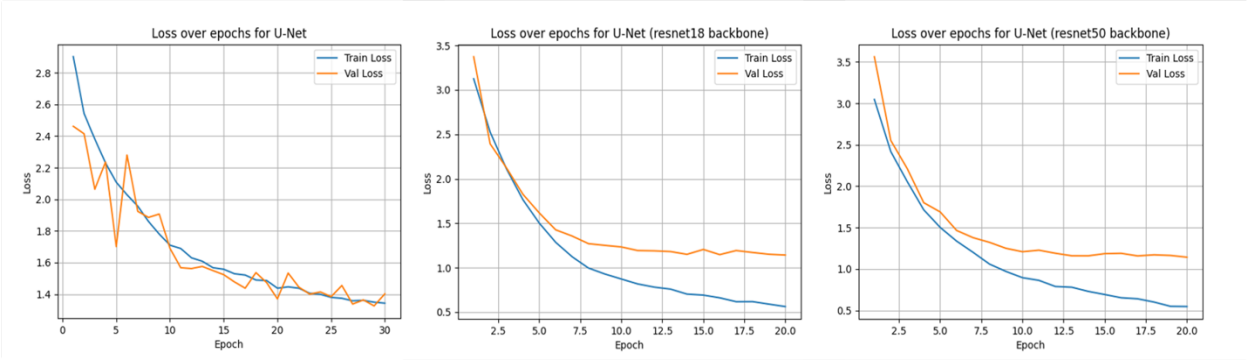

Figure 8. Loss vs Epochs for U-Net with different encoder backbones

Overall, the stronger encoders noticeably improve U-Net's segmentation accuracy and boundary quality. However, even the best U-Net variant still lags behind DeepLabV3+ and remains far below SAM2, underscoring the value of modern architectures and large-scale pretraining for robust real-world segmentation.

5.2 Data Augmentation

To evaluate whether data augmentation improves U-Net's ability to generalize on the Pascal VOC dataset, we trained two variants of the baseline model: one without augmentation and one with a standard augmentation pipeline including color jitter, and Gaussian blur. Quantitatively, augmentation did not improve U-Net's performance. As shown in Table 6, the model trained with augmentation performs slightly worse across all metrics, with mean dice coefficient decreasing from 0.0533 to 0.0521, mean IoU decreasing from 0.0426 to 0.0405, and pixel accuracy dropping from 0.6711 to 0.6394. HD95 increases substantially (from 73.17 to 85.62), indicating poorer boundary alignment and more scattered foreground predictions.

Table 6. Results of U-Net with and without augmentation

Model	Mean Dice	Mean IoU	HD95	Pixel Accuracy
U-Net (baseline)	0.0533	0.0426	73.1726	0.6711
U-Net (with augmentation)	0.0521	0.0405	85.6220	0.6394

The per-class results reinforce this trend. While the augmented model recovers weak signal for a few additional classes (such as bottle, bus, motorbike or train), its IoU and pixel accuracy generally deteriorate across the person class, where IoU falls from 0.1568 to 0.1227. These results suggest that augmentation introduces additional appearance variation that the shallow U-Net encoder cannot effectively absorb, causing the model to overfit noisy patterns rather than learning more robust semantic structure.

Table 7. Per-class IoU and Accuracy for U-Net with and without augmentation

	IoU		Pixel Accuracy	
Class	U-Net (baseline)	U-Net (augment)	U-Net (baseline)	U-Net (augment)
background	0.7379	0.6999	0.8625	0.8222
aeroplane	0	0	0	0
bicycle	0	0	0	0
bird	0	0	0	0
boat	0	0	0	0
bottle	0	0.0178	0	0.0726
bus	0	0.0086	0	0.0162
car	0	0	0	0
cat	0	0	0	0
chair	0	0	0	0
cow	0	0	0	0
diningtable	0	0	0	0
dog	0	0	0	0
horse	0	0	0	0
motorbike	0	0.0004	0	0.0004
person	0.1568	0.1227	0.6794	0.5355
pottedplant	0	0	0	0
sheep	0	0	0	0
sofa	0	0	0	0
train	0	0.0001	0	0.0001
tvmonitor	0	0	0	0

In Figure 9, U-Net baseline already produces overly smooth and blob-like masks, but with augmentation, the predictions become even less stable, sometimes collapsing into small, isolated patches or spreading incorrectly across large background regions. Instead of improving boundary accuracy or object completeness, augmentation appears to exacerbate U-Net's difficulty in modeling complex natural scenes.

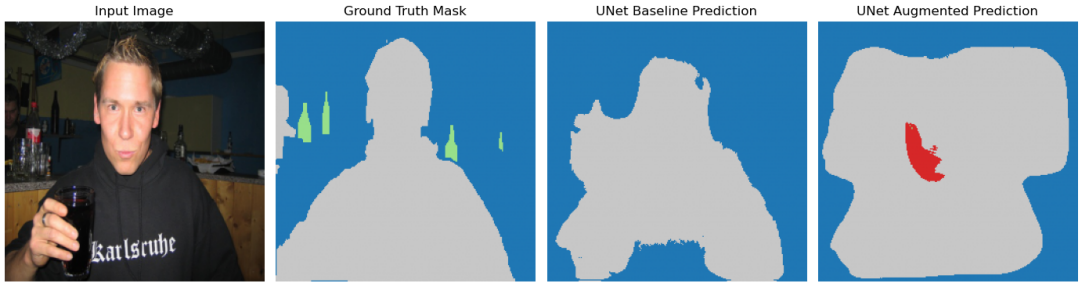

Figure 9. Comparison of Segmentation Map for U-Net with and without augmentation

This behavior is also reflected in the training dynamics. The loss curves in Figure 10 show that the augmented model exhibits substantially larger fluctuations in validation loss, especially in early epochs, consistent with the sensitivity of Dice loss to noisy or inconsistent training samples. Although training loss decreases monotonically, the elevated and unstable validation loss indicates

that augmentation increases the difficulty of the learning task without providing meaningful regularization benefits for this architecture.

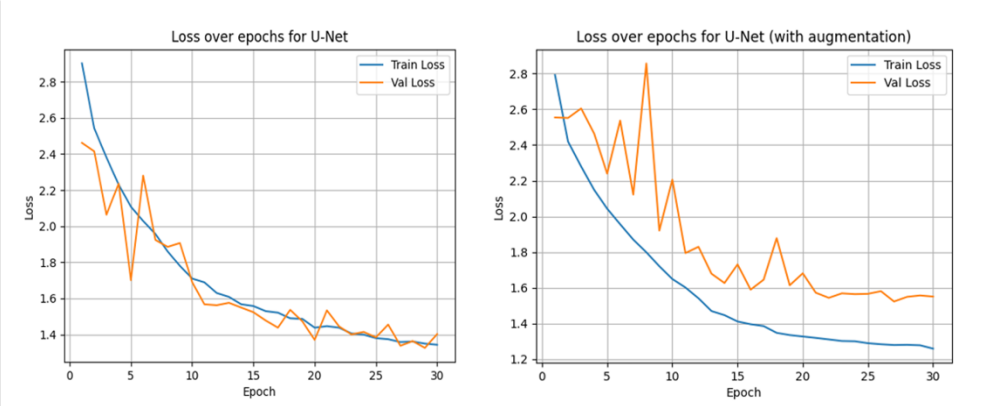

Figure 10. Loss vs Epochs for U-Net with and without augmentation

Overall, these results demonstrate that data augmentation does not meaningfully improve segmentation performance for U-Net on VOC and can in fact be detrimental when the model's capacity is insufficient to handle the added variability.

5.3 Loss Functions

To evaluate how different loss functions affect U-Net's learning behavior on the Pascal VOC dataset, we compared the baseline model trained with a combination of cross-entropy and a small Dice loss term (CE + 0.02 Dice) against a model trained with cross-entropy only. Quantitatively, using CE alone does not improve segmentation performance in Table 8. Although CE-only yields a slightly higher mean dice coefficient (0.0575 vs. 0.0533) and a marginal improvement in mean IoU (0.0458 vs. 0.0426), its HD95 becomes substantially worse (105.19 vs. 73.17), indicating poorer boundary precision and more diffuse mask predictions. Pixel accuracy increases modestly, but this largely reflects correct background classification rather than improved foreground segmentation.

Table 8. Results of U-Net with different loss functions (CE+Dice vs CE only)

Model	Mean Dice	Mean IoU	HD95	Pixel Accuracy
U-Net (bsseline)	0.0533	0.0426	73.1726	0.6711
U-Net (CE only)	0.0575	0.0458	105.1922	0.7224

Per-class results highlight the same pattern in Table 9. CE-only produces a small improvement on a few classes, such as person and diningtable, but performance on nearly all other categories remains zero, similar to the baseline. The qualitative comparison in Figure 11 further illustrates the difference: while the CE-only model produces smoother masks, it tends to shrink the

segmented region and fails to capture the full extent of the foreground object, reflecting weak overlap with ground truth. This suggests that cross-entropy alone biases the network toward predicting the dominant background class, especially in datasets like VOC where foreground objects occupy a relatively small portion of each image.

Table 9. Per-class IoU and Accuracy for U-Net with different loss functions (CE+Dice vs CE only)

	IoU		Pixel Accuracy	
Class	U-Net (baseline)	U-Net (CE only)	U-Net (baseline)	U-Net (CE only)
background	0.7379	0.7552	0.8625	0.9353
aeroplane	0	0	0	0
bicycle	0	0	0	0
bird	0	0	0	0
boat	0	0	0	0
bottle	0	0	0	0
bus	0	0	0	0
car	0	0	0	0
cat	0	0	0	0
chair	0	0	0	0
cow	0	0	0	0
diningtable	0	0.0028	0	0.0029
dog	0	0	0	0
horse	0	0	0	0
motorbike	0	0.0041	0	0.0042
person	0.1568	0.1994	0.6794	0.5364
pottedplant	0	0	0	0
sheep	0	0	0	0
sofa	0	0	0	0
train	0	0	0	0
tvmonitor	0	0	0	0

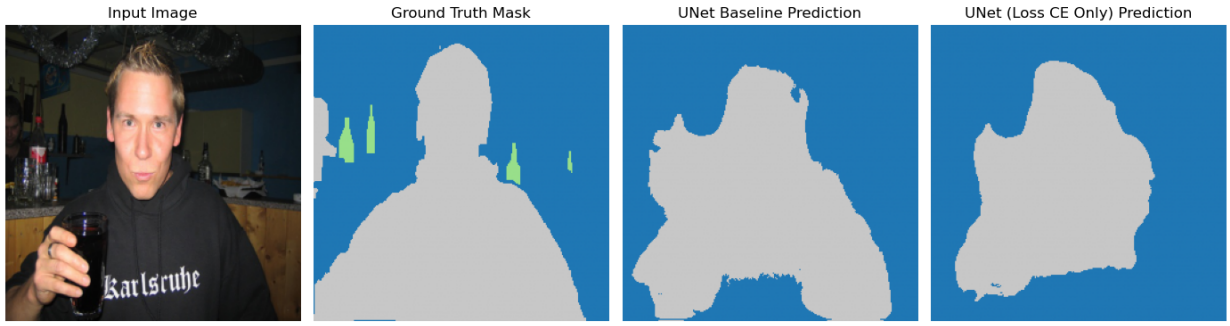

Figure 11. Comparison of Segmentation Map for U-Net with different loss functions (CE+Dice vs CE only)

CE-only exhibits smoother convergence and fewer fluctuations than CE + Dice, because Dice loss is more sensitive to small variations in mask overlap. However, this sensitivity is precisely what makes Dice loss useful: Dice emphasizes foreground pixels and directly optimizes for overlap, helping to counteract class imbalance by giving more weight to correctly segmented object regions rather than background-dominated accuracy. Without this term, the model tends to overfit the background and produces more conservative, under-segmented predictions, leading to higher HD95 and qualitatively poorer masks.

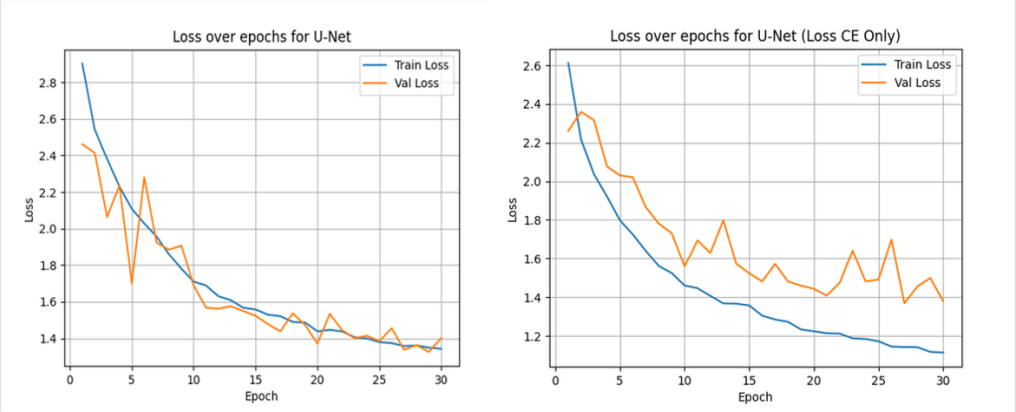

Figure 12. Loss vs Epochs for U-Net with different loss functions (CE+Dice vs CE only)

Overall, although CE-only training produces smoother optimization behavior, incorporating a small Dice component is essential for encouraging U-Net to capture minority foreground regions and avoid excessive bias toward predicting background. Dice loss provides better boundary alignment and foreground sensitivity, even if overall performance remains limited by U-Net's architectural constraints on complex natural images.

6. Interpretations and Lessons learned

Across all experiments, clear differences emerge in the segmentation capabilities of U-Net, DeepLabV3+, and SAM2, revealing both architectural strengths and fundamental limitations. U-Net consistently struggles in natural image settings: its shallow encoder is unable to capture the rich visual diversity of the Pascal VOC dataset, resulting in coarse, unstable predictions that frequently miss object boundaries or collapse entirely into background. DeepLabV3+ performs substantially better, benefiting from a strong ResNet-50 backbone and multi-scale context aggregation through ASPP. Its predictions are generally coherent and capture major object regions, though boundary precision and fine structures remain challenging. SAM2, in contrast, delivers high-fidelity masks that closely match the ground truth even in visually complex scenes. Its strong performance, achieved without any task-specific training, highlights the remarkable generalization ability afforded by large-scale pretraining and prompt-driven mask refinement. Together, the qualitative and quantitative results show a clear performance hierarchy: U-Net < DeepLabV3+ < SAM2.

The ablation studies deepen these insights by diagnosing why U-Net performs so poorly and exploring whether architectural or training modifications can improve its behavior. Replacing U-Net's shallow encoder with deeper ResNet-18 and ResNet-50 backbones produces consistent improvements across all metrics, confirming that encoder capacity is the primary bottleneck. However, even with ResNet-50, U-Net remains far below DeepLabV3+ and SAM2, suggesting that its decoder design and limited multi-scale reasoning are additional limiting factors. Data

augmentation did not help U-Net; in fact, it often hurt performance, likely because the model lacks the representational capacity needed to effectively absorb the increased variability. Similarly, training with cross-entropy alone yielded smoother optimization but led to worse boundary alignment and severe under-segmentation, illustrating the importance of Dice loss for addressing class imbalance and encouraging overlap-sensitive learning. Taken together, these ablations show that although U-Net can be moderately improved through stronger encoders and more careful loss design, its fundamental architectural simplicity prevents it from handling challenging multi-class scenes.

These findings lead to several broader lessons for future work. First, modern semantic segmentation tasks benefit significantly from architectures that incorporate strong pretrained backbones, multi-scale contextual features, or attention mechanisms. Traditional encoder–decoder models without pretraining are no longer competitive on complex datasets. Second, class imbalance remains a central challenge, and loss functions such as Dice or focal losses can help ensure that minority classes are not overwhelmed by background-dominant gradients. Third, large-scale pretraining, whether supervised, self-supervised, or multimodal, dramatically improves generalization, as demonstrated by SAM2's robust zero-shot performance. Future work could explore combining U-Net-style decoders with pretrained vision transformers or experimenting with SAM-style promptable interfaces for supervised segmentation tasks. Finally, building hybrid approaches that incorporate prompt-based refinement or foundation model guidance may offer a practical pathway to strong performance even with limited compute or training data.

References

[1] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
[2] Ravi, Nikhila, et al. "Sam 2: Segment anything in images and videos." *arXiv preprint arXiv:2408.00714* (2024).
[3] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Cham: Springer international publishing, 2015.
[4] Mao, Anqi, Mehryar Mohri, and Yutao Zhong. "Cross-entropy loss functions: Theoretical analysis and applications." International conference on Machine learning. pmlr, 2023.