Deep Learning Final Report

Yuanhe Guo
New York University
yg2709@nyu.edu

Jianing Zhang
New York University
jz5212@nyu.edu

Haotong Wu
New York University
hw2933@nyu.edu

Abstract

In this project, we tackle the problem of predicting future video segmentations, aiming to forecast the 22nd frame from the first 11 frames across 48 categories with varying shapes, colors, and materials. Our two-stage approach employs U-Net to generate segmentation masks from unlabeled data, providing structured inputs for the subsequent proposed SimMP2 model to improve future frame prediction accuracy. This integration leverages U-Net's precise localization and SimMP2's prediction capabilities, enhancing segmentation and forecasting.

1 Introduction

In the field of computer vision, image segmentation and prediction are pivotal tasks with widespread applications. U-Net Ronneberger et al. [2015], featuring a distinctive U-shaped architecture, has proven remarkably successful for image segmentation. This fully convolutional network combines a contracting path to capture multi-scale context and an expanding path for precise localization, enabling efficient transfer of contextual information across the network, crucial for achieving accurate segmentations.For image prediction, SimVP Gao et al. [2022] excels in efficiently predicting future video frames. It employs a streamlined architecture that encodes input frames into a latent space with linear dynamics, facilitating modeling and prediction before decoding back into the image space. This approach accelerates prediction and enhances scalability.
We propose a novel approach that leverages the complementary strengths of U-Net and SimVP. U-Net generates segmentation masks from unlabeled datasets. These masks serve as inputs for our SimVP-based SimMP2 to predict future states. This integration harnesses U-Net's localization capabilities to enhance SimMP2's input quality, enabling more accurate future state predictions while reducing complexity and enhancing focus on critical frame elements.

2 Related Works

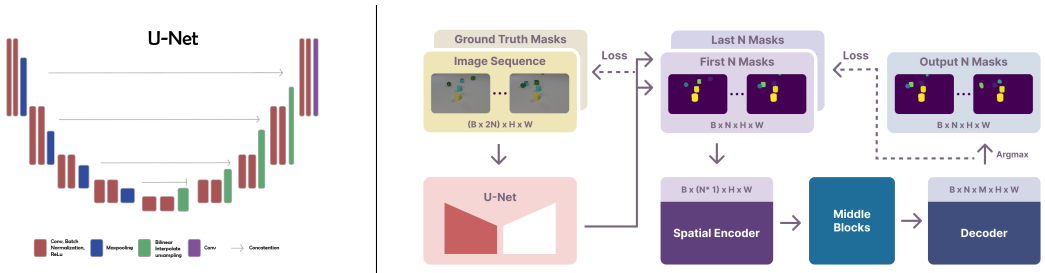
Image Segmentation In the field of image segmentation, more and more model architectures are emerging. A group of works utilize Region-based Convolutional Network (RCNN) He et al. [2017], Ren et al. [2015] for complex image datasets requiring detailed object localization. However, precise pixel-wise classification results in resource-intensive and time-consuming training.

Recently evolving transformer-based architecture like Mask2Former Cheng et al. [2022] achieved SOTA results in tasks like panoptic, instance, and semantic segmentation. However it suffers from over-fitting on small and simple dataset.

Video Prediction The task of video prediction requires model's ability to learn the underlying dynamics of movements. SOTA result for Moving MNIST dataset Srivastava et al. [2015] is achieved by SimVP Gao et al. [2022], which utilizes a simple CNN-based structure.

Another branch of works focus on learning the underlying physics in video, primarily for Visual Q&A tasks. SlotFormer Wu et al. [2022] take advantage of transformer-based auto regressive method for object-centric representations, achieving remarkable results in video prediction.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) (b)

Figure 1: (a) Detailed Architecture of U-Net. (b) Overview of our proposed pipeline, where B is batch size, T is length of sequence, H and W are image size, M is the number of mask classes. In our task, $T = 11, H = 160, W = 240, M = 49$. We use cross entropy for both losses.

For our task, one vanilla approach is combining a video predictor with an image segmentation model. A segmentation mask yields from the predicted 22th frame based on the first 11 video sequence. However, video prediction models tend to predict blurry images given insufficient training data, which limits the segmantation mask quality, since subsequent image segmentation models depend heavily on predicted frames. To address this problem, the order of video predictor and image segmentation model could be switched. Also known as pseudo labeling, a image segmentation model trained on a small labeled dataset is used to generate masks for a large unlabeled dataset. Then the video predictor could be trained to directly predict future masks based on video sequence input.

3 Method

Our approach takes the idea of pseduo labeling one step further. After getting pseudo labels, we train a model that takes in a sequence of 11 segmentation masks and predict the 22th mask. The mask prediction model no longer need to cope with complex shadow and material information, thus could potentially lead to better mask prediction result. We will discuss our image segmentation model and mask prediction model in detail in the following section.

3.1 U-Net

U-Net, originally designed for biomedical image segmentation by Ronneberger et al. Ronneberger et al. [2015], features a symmetric encoder-decoder architecture that significantly enhances precise localization. Our adaptation of U-Net for general image segmentation tasks maintains the core elements of its design. The input layer of our model accepts three-channel images, beginning with a double convolution layer with 64 base channels, which doubles at each contracting stage. This is followed by max pooling and another double convolution that reduces the spatial dimensions while simultaneously increasing the number of features. During up-sampling, we use either bilinear interpolation or transpose convolutions to restore the image resolution, incorporating skip connections to merge features from the down-sampling stages effectively. The output layer utilizes a 1x1 convolution to map these features to 49 classes. The network balances accuracy with computational efficiency. In the forward pass, feature maps from both the contracting and expanding paths are concatenated to preserve essential information and prevent data loss during encoding.

3.2 SimMP2

For future mask prediction, we propose SimMP2 (Simple Mask-2-Mask Prediction). Adapted from SimVP Gao et al. [2022], SimMP2 takes mask sequence with only one channel, and predict logits of M channels, where $M = 49$ for our competition. The structure shown in Fig. 1b consists of three parts: spatial encoder, middle blocks and decoder. We extract spatial features with shape $(T, \hat{C}, \hat{H}, \hat{W})$ by treating input mask sequence as one-channel images of $(T, 1, H, W)$. The middle blocks, also known as translator, follows the exact same CNN-based structure in SimVP for learning temporal information. The decoder predicts a sequence of logits with shape $(T, 49, H, W)$ for future masks. Detailed training configurations could be found in Sec. 4.2.

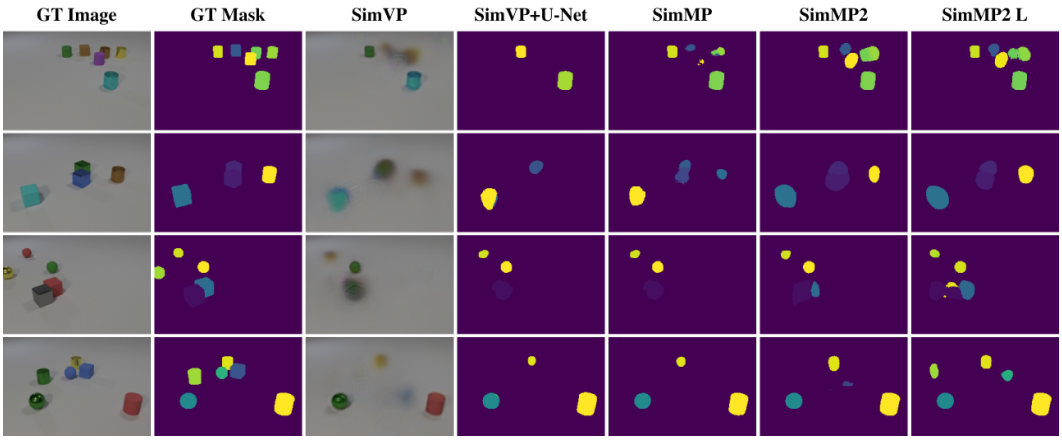


Figure 2: Qualitative Comparison. The left two columns are ground truth images and masks for the 22th frame in validation data. All models use gsta as middle block with hidden dimension being 256, except for SimMP2 L whose hidden dimension is 512.

4 Experiments

4.1 Data Description

Our project's dataset, tailored for video prediction and segmentation analysis, comprises 13,000 unlabeled videos of 22 frames each, with a resolution of 160x240 pixels and three color channels (RGB). Additionally, it includes 1,000 labeled training videos and 1,000 labeled validation videos, identical in structure to the unlabeled set. Crucially, each labeled video contains fully annotated segmentation masks spanning 49 distinct classes for every frame, enabling model validation and aligning the training process with real-world data. This extensive and detailed labeling is vital for training our U-Net model to accurately segment video frame features.

4.2 Model Training

U-Net Training specifications for U-Net includes: optimizer Adam with a learning rate of 0.001 on cross-entropy loss. During training, we performed gradient descent and calculated IoU metrics to monitor and enhance model performance, saving the best-performing model based on validation IoU.

SimMP2 Before training SimMP2, we processed the unlabeled videos by feeding them to our trained U-Net for pseudo labels. We combined them with ground-truth labels in training dataset. During training, future sequence of 11 masks were predicted by SimMP2 based on the first 11 masks, and then applied with cross entropy loss with cosine annealing scheduler. Thanks to the simple yet effective design, the middle block of SimMP2 could be fitted with a variety of model structures. The original SimVP Gao et al. [2022] claims gsta (gated spatial-temporal attention) achieves the best performance, which also holds true in our case. While for the gsta structure, we trained SimMP2 and SimMP2 L(arge) with two configurations: setting both input channels to be 64, and hidden channels to be 256 and 512 accordingly. We used SimMP2 L for the final submission.

4.3 Evaluation

In fig. 2 we compare SimMP2 and SimMP2 L with two baselines, SimVP+U-Net and SimMP. The SimVP+U-Net approach combines a SimVP model trained on unlabeled dataset with UNet whose weight is finetuned on image predicted by SimVP. SimMP, similar to SimMP2, are trained on pseudo labels for directly predicting mask sequence, but takes image seqence as input. Besides, we visualize images predicted by SimVP on the third column. In all examples, SimVP fails to clearly predict moving objects. These blurred parts are mistaken as backgrounds by UNet, resulting in the missing of multiple object masks. In contrast, SimMP2 L for final submission achieves remarkable result in Row 1&2. For Row 3&4, SimMP2 L fails to identify occlusion, resulting in significant derivation from ground truths. However, it still out performs other methods by predicting more potential classes.

	SimVP+U-Net	SimMP	SimMP2	SimMP2 L
Cross Entropy ↓	0.2550	0.2036	**0.2012**	0.2067
MSE ↓	35.0883	31.0084	28.7396	**28.0020**
SSIM ↑	0.8238	0.8425	0.8580	**0.8615**
IoU ↑	0.2275	0.3308	0.4128	**0.4224**

Table 1: Quantitative Comparison. Models has the same configuration as these in Fig. 2. All metrics are calculated on validation data.

Table. 1 evaluates models' performance on different metrics. Our SimMP2-based methods significantly out perform baseline approaches, while SimMP2 L only demonstrates a moderate advantange over SimMP2. Note that although SimMP2 L achieves higher IoU, the cross entropy loss used for training is higher than SimMP2. This indicates that using IoU as a regularization during training could potentially lead to even better performance.

5 Conclusion

This project used a new approach that integrates U-Net and SimMP models to address the challenges of video frame prediction and segmentation. The method effectively utilizes the segmentation mask generated by U-Net as an input to the SimMP model, which improves the accuracy of the prediction. Experimental results show that the method provides significant improvement over other architectures for image segmentation and video prediction tasks. Future work will focus on refining these models, further tuning the parameters to improve their accuracy and efficiency, and potentially exploring more sophisticated machine learning techniques to solve more complex problems.

References

Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Alexander G. Schwing. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1611, June 2022.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3170–3180, June 2022.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, October 2017.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, October 2015.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 843–852, Lille, France, 07–09 Jul 2015. PMLR.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. *arXiv preprint arXiv:2210.05861*, 2022.