Advanced Frame Prediction and Segmentation

Yaozhong Huang
yh2563@nyu.edu

Ning Yang
ny675@nyu.edu

Chengning Zhang
cz2885@nyu.edu

Abstract

This report presents an innovative approach to video frame prediction and segmentation, combining state-of-the-art techniques and novel algorithms. The focus is on predicting the 22nd frame of a video using only the first 11 frames and producing accurate semantic segmentation masks. The effectiveness of the Temporal Attention Unit (TAU) and Masked R-CNN models in achieving these objectives is highlighted.

1 Literature Review

Recent advancements in video analysis have primarily focused on two aspects: predictive frame analysis and object instance segmentation. Predictive frame analysis involves generating future frames based on a limited set of initial frames, while object instance segmentation focuses on identifying and delineating each object within a frame. Both are crucial for applications ranging from autonomous vehicles to video surveillance [1].

1.1 Predictive Frame Analysis

The field of predictive frame analysis has evolved significantly over the years, transitioning from traditional approaches to more sophisticated neural network-based methods. Early efforts primarily utilized Recurrent Neural Networks (RNNs) and their variants like Long Short-Term Memory (LSTM) networks. These methods were initially promising due to their ability to handle sequential data, yet they often struggled with capturing long-range temporal dependencies, a critical aspect for accurate frame prediction in videos [1].

The advent of Convolutional Neural Networks (CNNs) brought a substantial shift in this domain [1]. CNNs, with their inherent ability to process spatial hierarchies in images, provided a more robust framework for understanding and predicting visual data. However, the real breakthrough came with the integration of attention mechanisms into neural networks. This innovation allowed models to focus on relevant parts of the input data selectively, enhancing the ability to learn complex spatial-temporal relationships in videos. The Temporal Attention Unit (TAU) represents a significant advancement in this regard. By separating temporal attention into intra-frame statical attention and inter-frame dynamical attention, TAU addresses the inefficiencies of recurrent units in handling long-term dependencies [1]. This development has enabled more accurate prediction of future frames by effectively capturing both spatial and temporal dynamics.

1.1.1 TAU Architecture

The Temporal Attention Unit (TAU) is designed for enhancing spatiotemporal predictive learning. Its architecture is characterized by a division of temporal attention into two components:

1. **Intra-Frame Statical Attention:** Focuses on attention mechanisms within individual frames to identify and emphasize significant features.
2. **Inter-Frame Dynamical Attention:** Extends the attention mechanism across frames to capture temporal dynamics and relationships.

Preprint. Under review.

TAU processes a sequence of frames, applying both intra-frame and inter-frame attention to extract critical spatial and temporal features for frame prediction.

1.2 Object Instance Segmentation

Object instance segmentation has also seen remarkable progress. Initial methods in this field were heavily reliant on region proposal networks, which would generate potential object-bound regions in an image, followed by classification and post-processing steps to delineate object boundaries. These methods, though effective, were often computationally intensive and lacked precision in segmentation [2].

The introduction of Masked R-CNN marked a significant turning point. Building upon the successes of Faster R-CNN, a leading model for object detection, Masked R-CNN introduced an additional branch for predicting segmentation masks. This architecture enabled the simultaneous detection and high-quality segmentation of objects in an image, significantly improving the precision and efficiency of instance segmentation. The adaptability of Masked R-CNN to various tasks, such as human pose estimation, further underscored its versatility and effectiveness. This model has set new standards in the field, providing a robust framework for future research and applications in instance-level recognition [2].

1.2.1 Masked-RCNN Architecture

Masked R-CNN architecture consists of several key components [2]:

1. **Backbone Network:** Typically a CNN like ResNet, used for extracting feature maps from input images.
2. **Region Proposal Network (RPN):** Proposes candidate object bounding boxes by scanning the feature maps.
3. **RoI Align:** Aligns the proposed regions to a fixed size, preserving spatial information.
4. **Classification and Bounding Box Regression Head:** Classifies objects and refines bounding boxes.
5. **Mask Prediction Head:** Parallel to the classification head, it generates a segmentation mask for each object instance.

Masked R-CNN simultaneously detects objects and generates pixel-wise segmentation masks, offering detailed image analysis capabilities.

2 Idea

2.1 Data

For training and testing our models, we use a synthetic dataset of 3D shape videos, which offers a diverse range of shapes, materials, and colors. This dataset provides a controlled environment to test the robustness and accuracy of our models under various conditions.

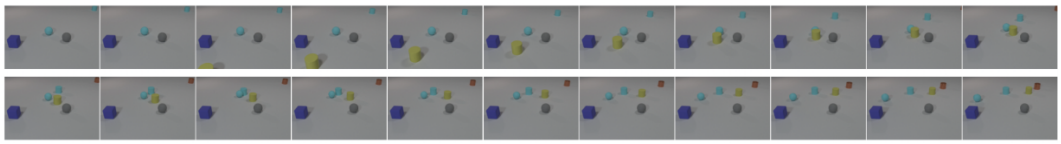

Figure 1: A sample video with 22 frames in the dataset

2.2 Models

The first part of our methodology involves using the Temporal Attention Unit (TAU) to predict future video frames. The TAU model is particularly suited for this task due to its ability to separate temporal

attention into intra-frame statical attention and inter-frame dynamical attention. This separation allows the model to efficiently capture long-term dependencies and dynamic changes across video frames, overcoming the limitations often encountered in recurrent models. In our specific implementation, TAU is trained on the first 11 frames of a video sequence to predict the 22nd frame. This choice of frame interval is designed to test the model's ability to extrapolate information over a substantial temporal gap, thus demonstrating its efficacy in predicting long-term dynamics in videos.

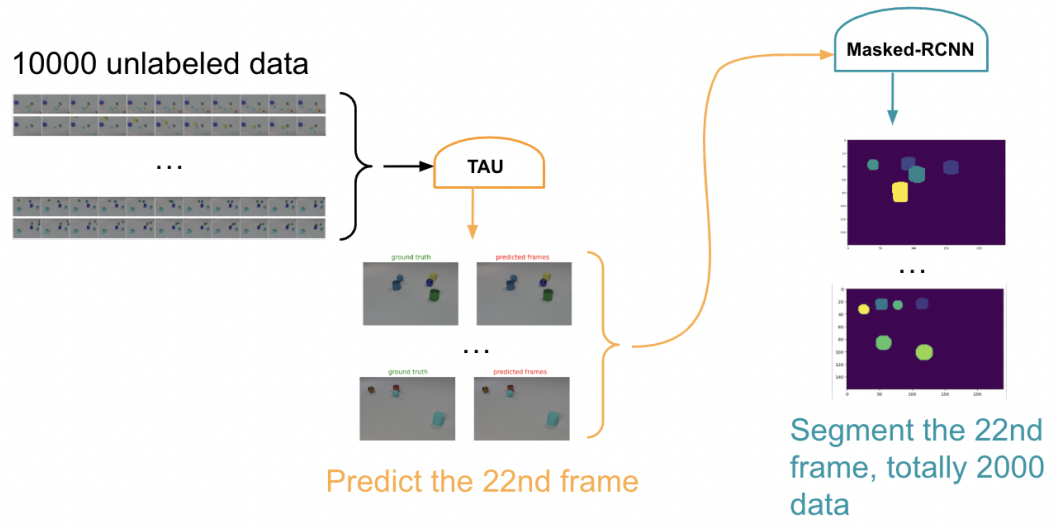

Figure 2: General workflow

The second component of our methodology is the implementation of Masked R-CNN for semantic segmentation. Once the future frame is predicted using TAU, Masked R-CNN is employed to generate accurate segmentation masks for each object instance within the frame. Masked R-CNN extends the capabilities of Faster R-CNN by adding a parallel branch for mask prediction, enabling the simultaneous detection and segmentation of objects. This model is particularly chosen for its high precision in segmenting complex and overlapping objects, a common challenge in video frames.

2.3 Evaluation

The performance of the models is evaluated using the Intersection Over Union (IOU) metric for both predictive frame analysis and object segmentation. IOU offers a clear and quantifiable measure of accuracy by comparing the overlap between the predicted and actual frames or segmentation masks.

3 Results and Visualization

3.1 Predictions from TAU

As shown in the Figure 3, the TAU appears to be capturing the general motion of the objects across frames. The predicted positions of the geometric shapes in the frames suggest that the model has learned to track the movement of objects over time. Also, there is a close resemblance between the ground truth and the TAU predictions in terms of object shape and orientation, indicating that TAU has effectively learned some spatial features of the objects.

However, as the sequence progresses, there is a noticeable blur or fading in the predicted frames. This could be due to the model's uncertainty about the objects' positions or due to the difficulty in capturing fine details over longer sequences. The model may struggle with predicting interactions or complex dynamics, such as when objects come close to each other or overlap, which can be observed in later frames where shapes begin to blur and colors start to blend.

Figure 3: Predicted frames from TAU

3.2 Predictions from Masked-RCNN

As shown in Figure 4, the segmentation masks for Frame 0, Frame 11, and Frame 21 show that the Masked R-CNN model is capable of consistently identifying and segmenting the objects across the sequence. Each object is represented by a different color in the segmentation masks, indicating that the model is distinguishing between different object instances.

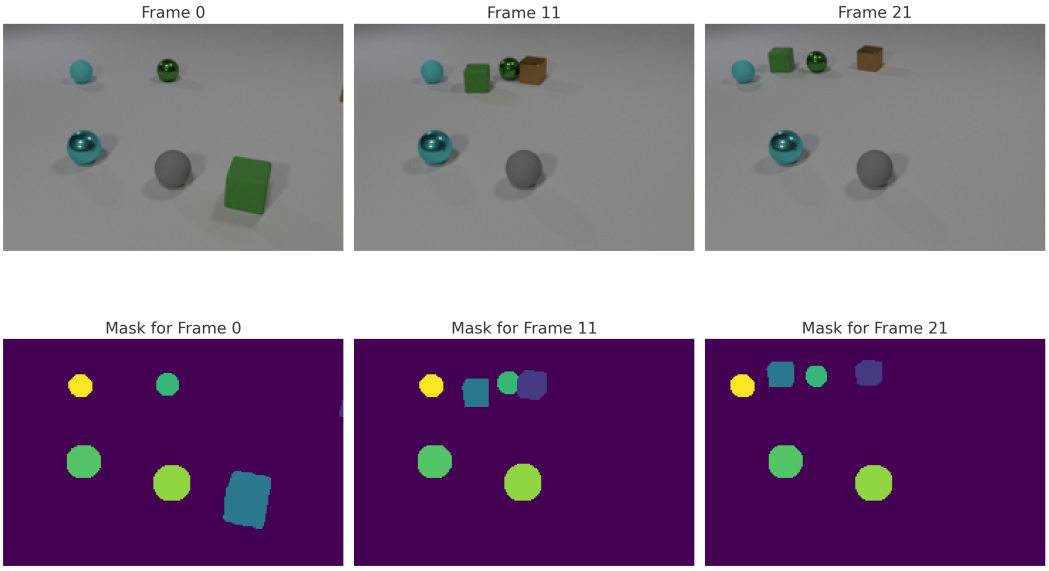

Figure 4: Predicted frames from Masked-RCNN

Given the context that the segmentation appears visually accurate, but the Intersection Over Union (IOU) on the test dataset is approximately 0.3, it is reasonable to infer that the lower IOU might be significantly influenced by the performance of the TAU model on predicting the 22nd frame. If the TAU model's predictions for the 22nd frame are not closely aligned with the ground truth in terms of object position and shape, even a well-performing segmentation model like Masked R-CNN will produce masks based on inaccurate predictions.

References

[1] Cheng Tan, Zhe Gao, Lechao Wu, Yue Xu, Jianfeng Xia, Shuai Li, and Stan Z Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. *arXiv preprint arXiv:2206.12126*, 2023.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017.