Inter IIT Tech Meet 10.0

Team ID - 8

Bosch's Model Extraction Attack for Video Classification

Table of Contents

Problem Statement

To develop an efficient strategy to extract the video-based models in the black-box and grey-box setting for:

- Video Swin-T Model for Action Classification on Kinetics-400 dataset
- MoViNet-A2-Base Model for Video Classification on Kinetics-600 dataset

The true method of knowledge is experiment.

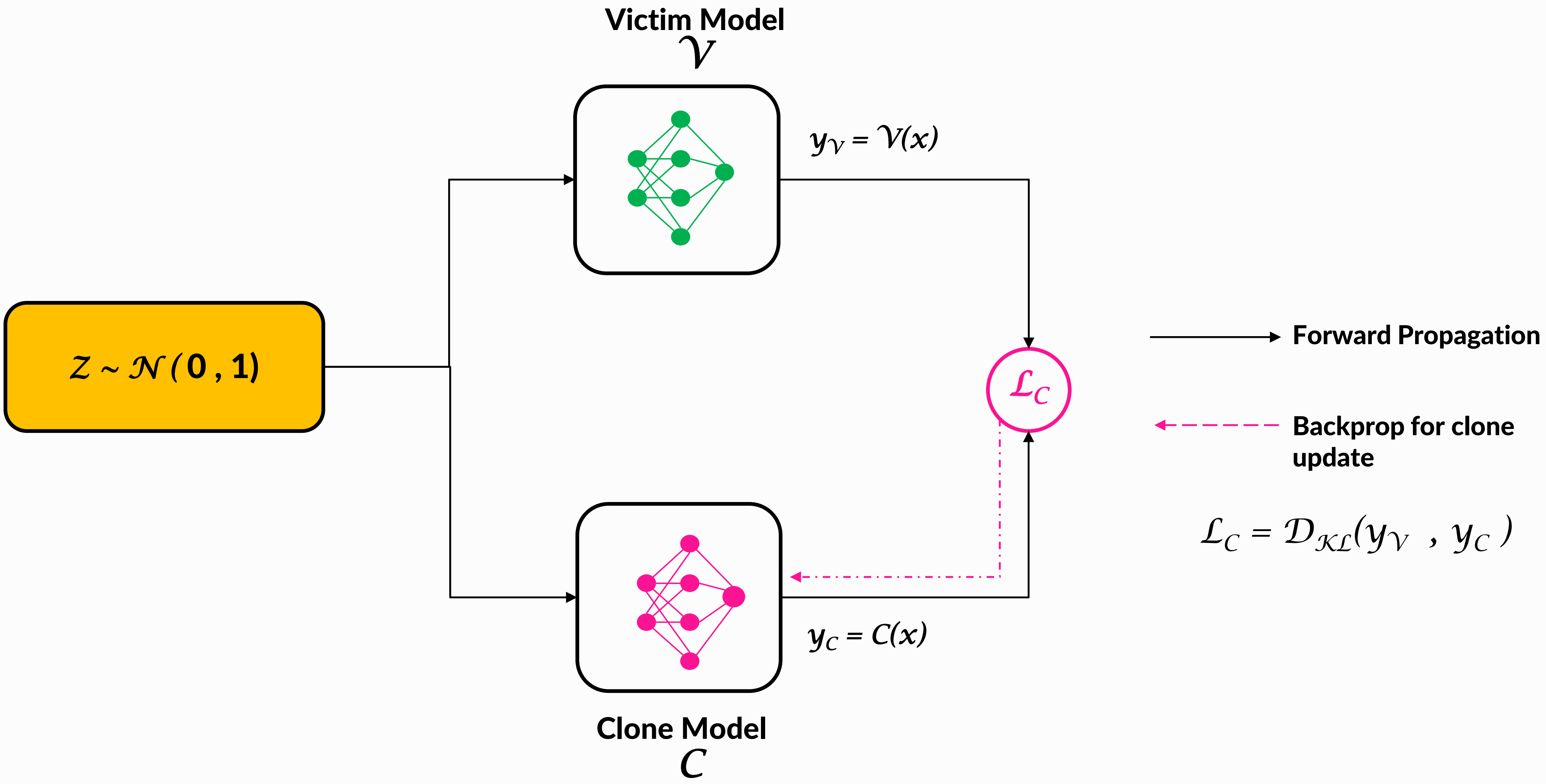
William Blake

Black Box Approach

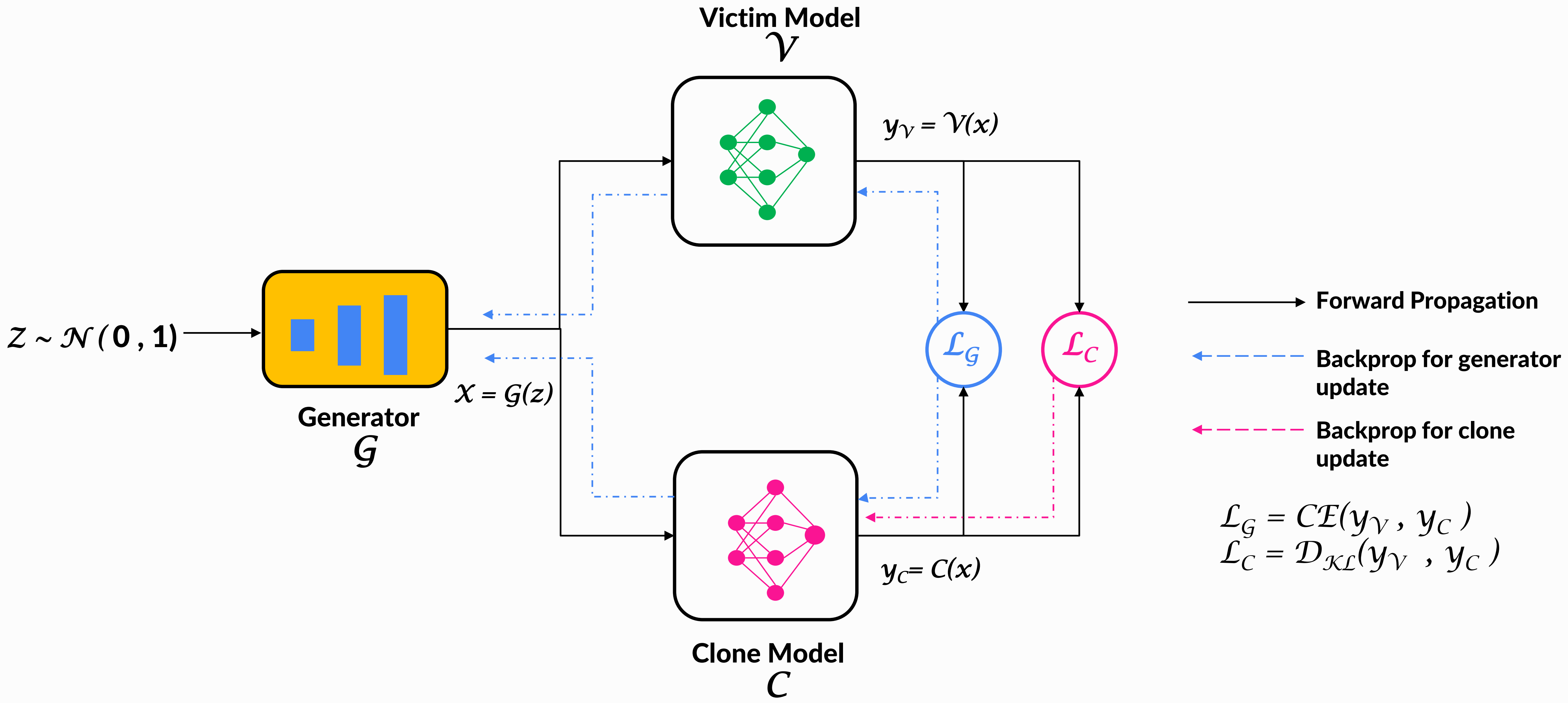
Black Box Approach

1. Extraction Strategies
2. Models Used
3. Results

1. Random Normal Sampling



2. Training Generator and Clone Together



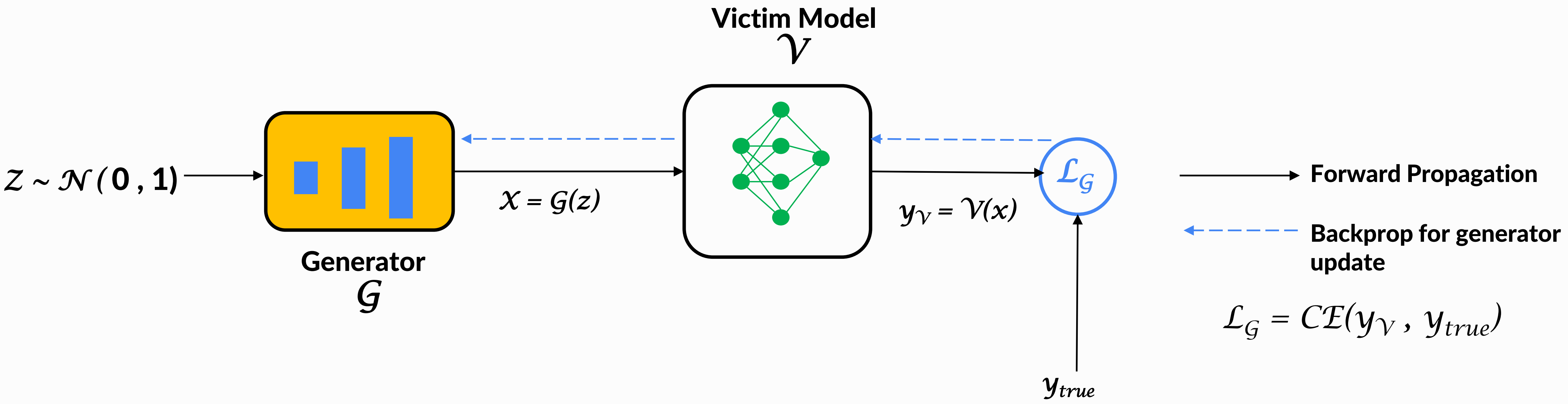
$$\mathcal{L}_{\mathcal{G}} = \mathcal{CE}(y_{\mathcal{V}}, y_C)$$
$$\mathcal{L}_C = \mathcal{D}_{\mathcal{KL}}(y_{\mathcal{V}}, y_C)$$

2. Training Generator and Clone Together

- Build upon the approach presented in **MAZE**[1] and **DFME**[2]
- Add a **generator** to help make meaningful queries
- Generator is based on **DVD-GAN** architecture
- Generator weights updated using **zeroth-order gradient estimates** of the victim
- Clone is updated **simultaneously**

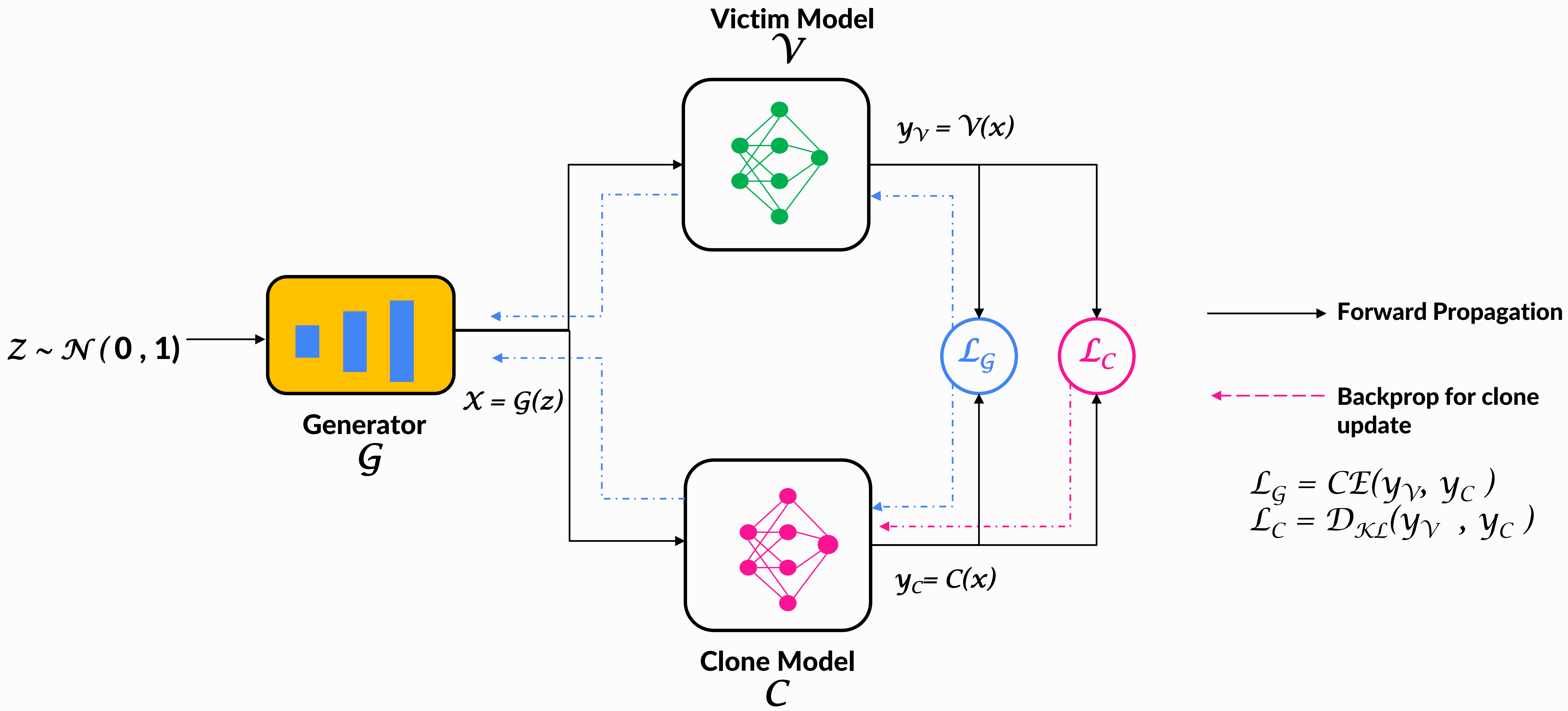
[1]MAZE: Model Stealing Attack using Zeroth-Order Gradient Estimation
[2]DFME: Data-free Model Extraction

3. Training Generator and Clone Independently



$$\mathcal{L}_{\mathcal{G}} = C\mathcal{E}(y_{\mathcal{V}}, y_{true})$$

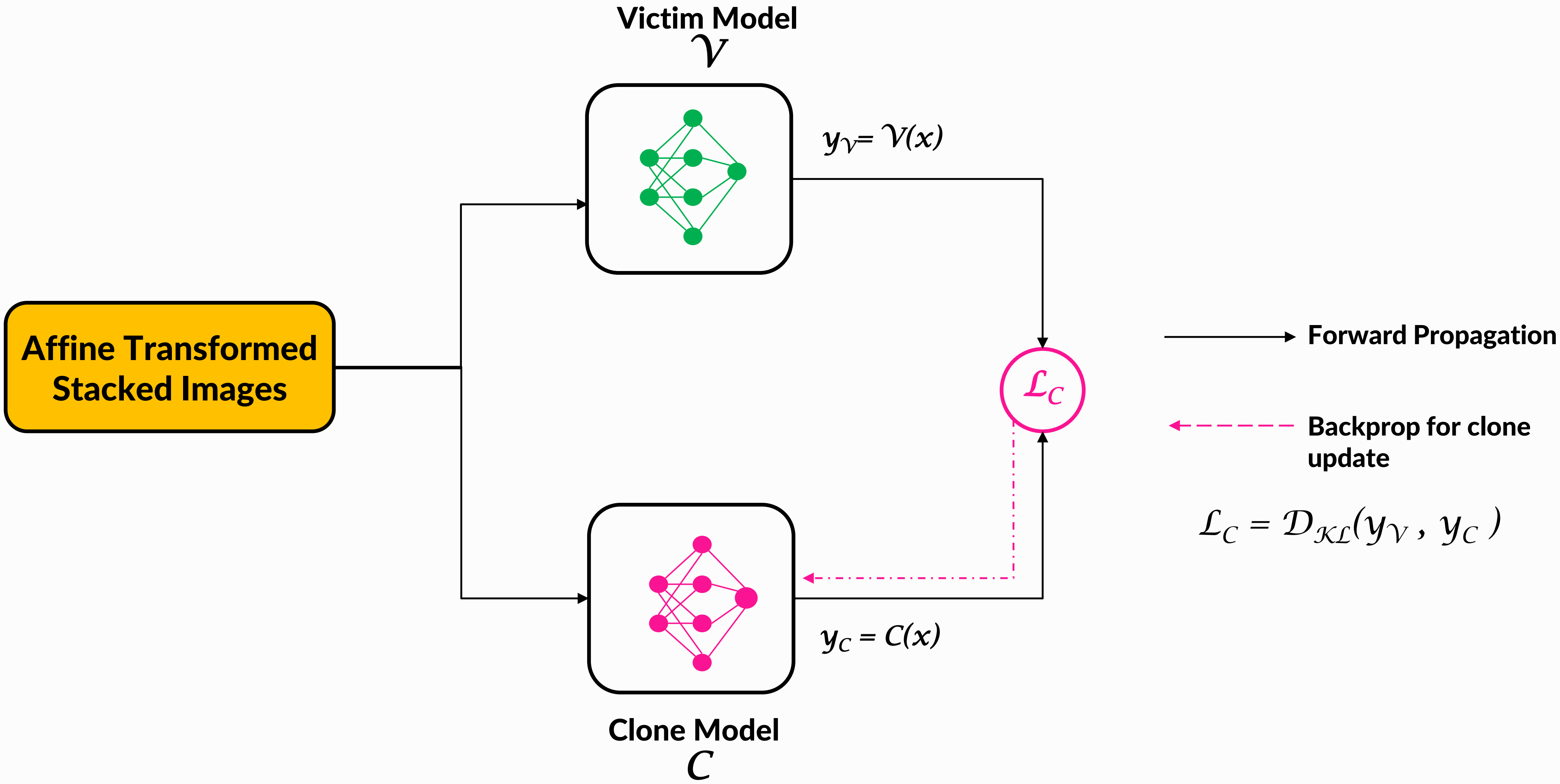
3. Training Generator and Clone Independently



3. Training Generator and Clone Independently

- Generator is made **conditional** and is trained **independently** using teacher predictions
- Trained generator is then used in a manner like the **previous approach**
- The generator is still being trained along with the clone

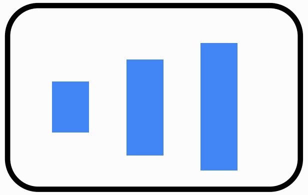
4. Stacking Affine Transformed Images



4. Stacking Affine Transformed Images

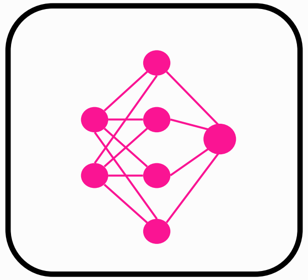
Generator Model

- **DVD-GAN Generator**
  - SOTA results in video generation for higher resolutions with higher temporal coherence between the generated frames on Kinetics datasets.
  - Conditional generator for video generation satisfied the necessary requirements for the second training paradigm of pretraining a generator



Generator
$\mathcal{G}$

Clone Model



Clone Model
C

- **ResNet 3D**
  - Simple architecture with readily available code
  - Less compute-intensive
- **ResNet (2+1)D**
  - Lightweight architecture compared to transformers
  - Among Top-20 in Video classification related tasks

Experimental Results Obtained for Swin-T

Experimental technique	Clone Model	Top-5 Accuracy	Top-1 Accuracy
Random normal sampling	ResNet3D	1.26	0.27
Training generator along with clone	ResNet3D	2.69	0.41
Training conditional GAN independently	ResNet3D	**4.85**	**0.84**
Stacking affine-transformed images	R(2+1)D	1.22	0.30

Final Results Obtained for Black Box

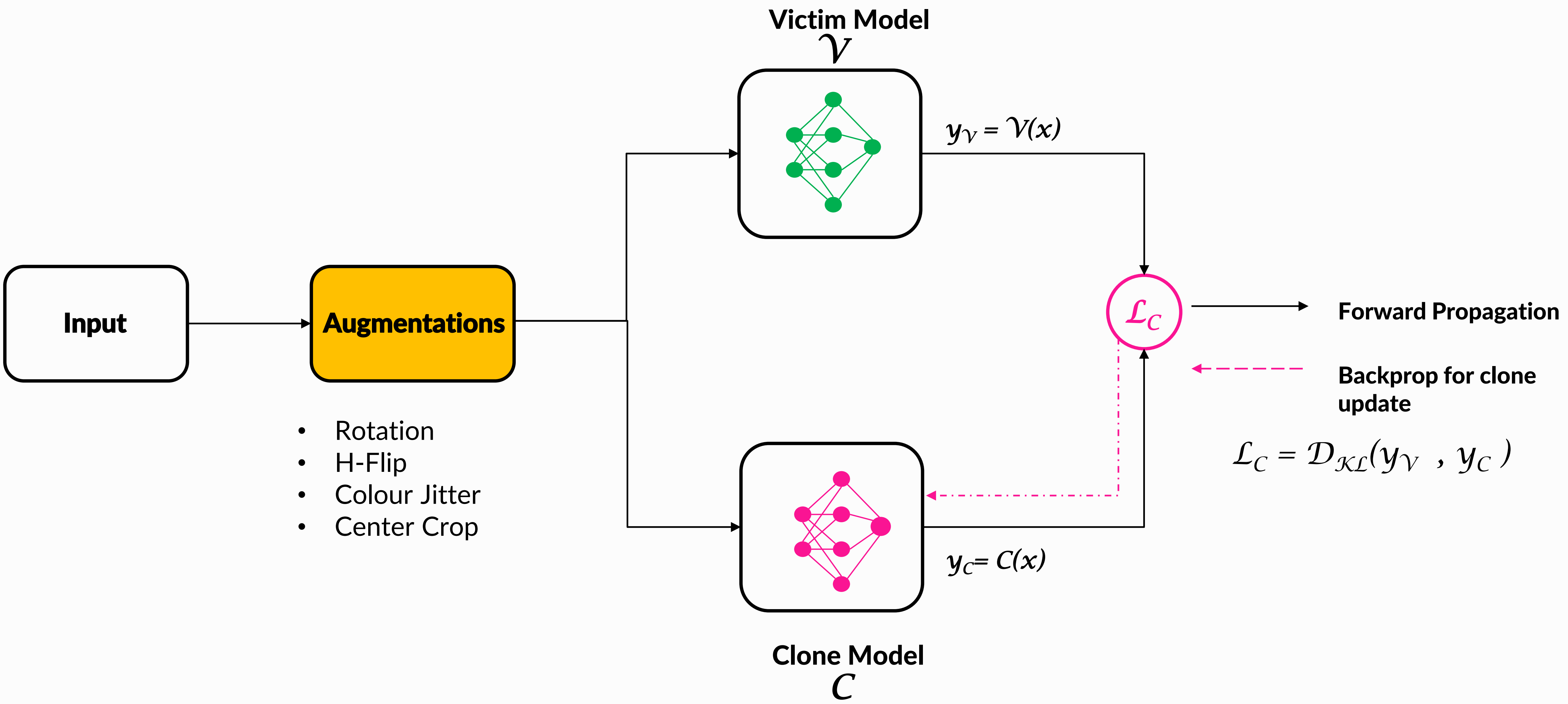
Victim Model	Clone Model	Top-5 Accuracy	Number of Queries
Video Swin Transformer	R(2+1)D	4.85	~1M
MoViNet-A2-Base	R(2+1)D	4.13	~1M

Grey Box Approach

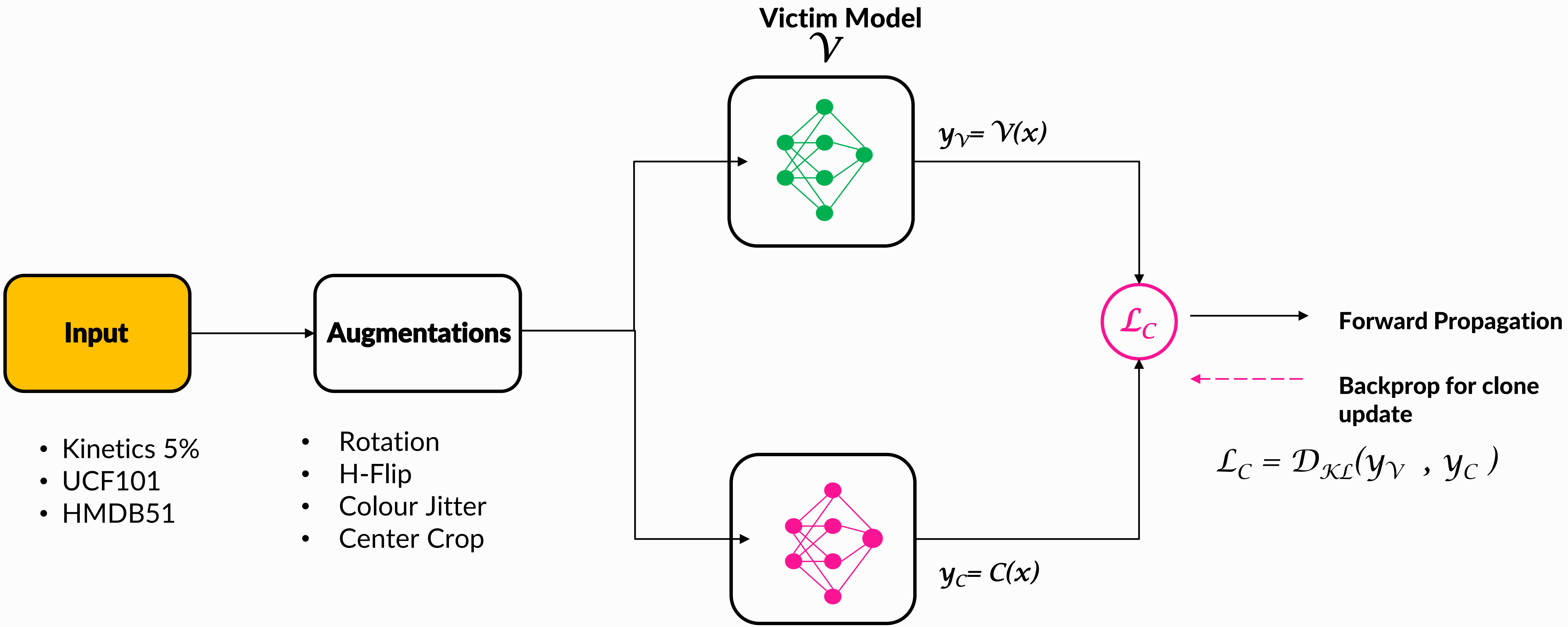
Grey Box Approach

1. Extraction Strategies
2. Models used
3. Results

1. Augmenting Kinetics



2. Concatenated Datasets



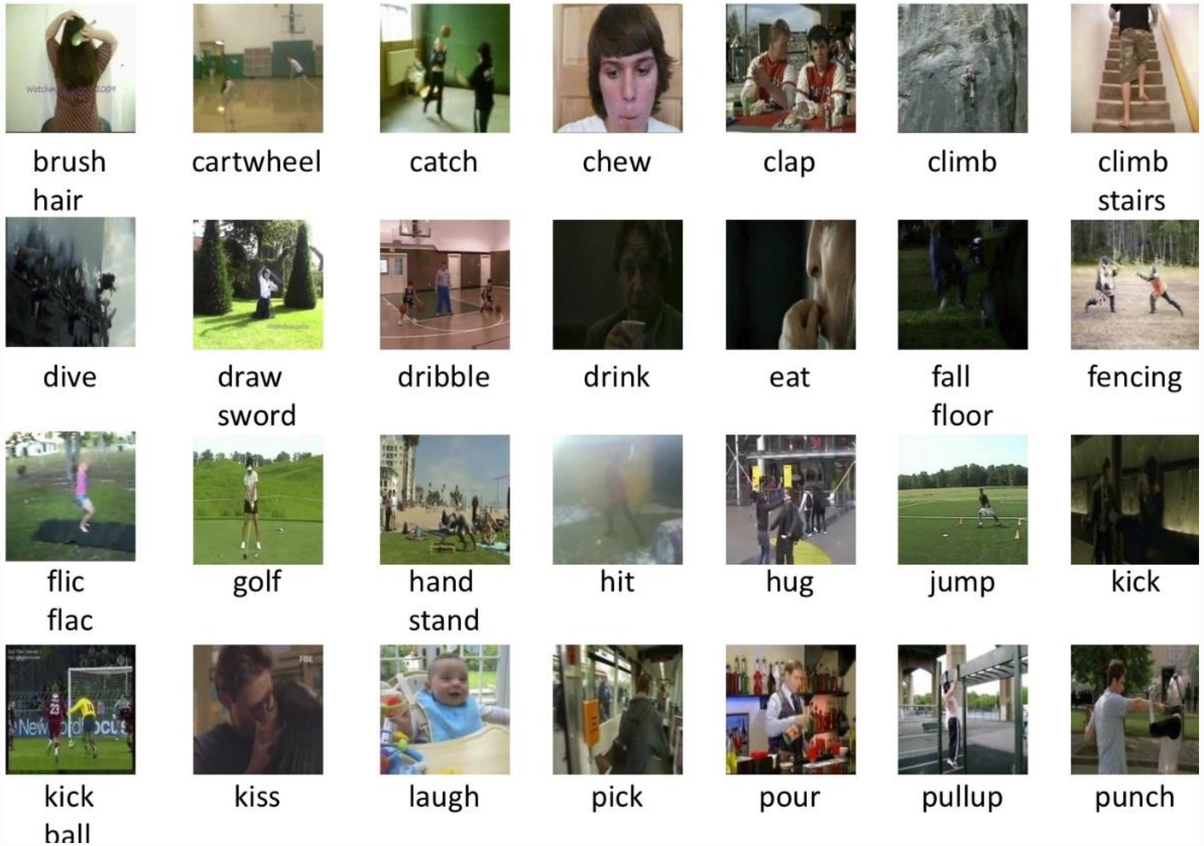
- Kinetics 5%
- UCF101
- HMDB51

- Rotation
- H-Flip
- Colour Jitter
- Center Crop

$$\mathcal{L}_C = \mathcal{D}_{\mathcal{KL}}(y_{\mathcal{V}}\ ,\ y_C)$$

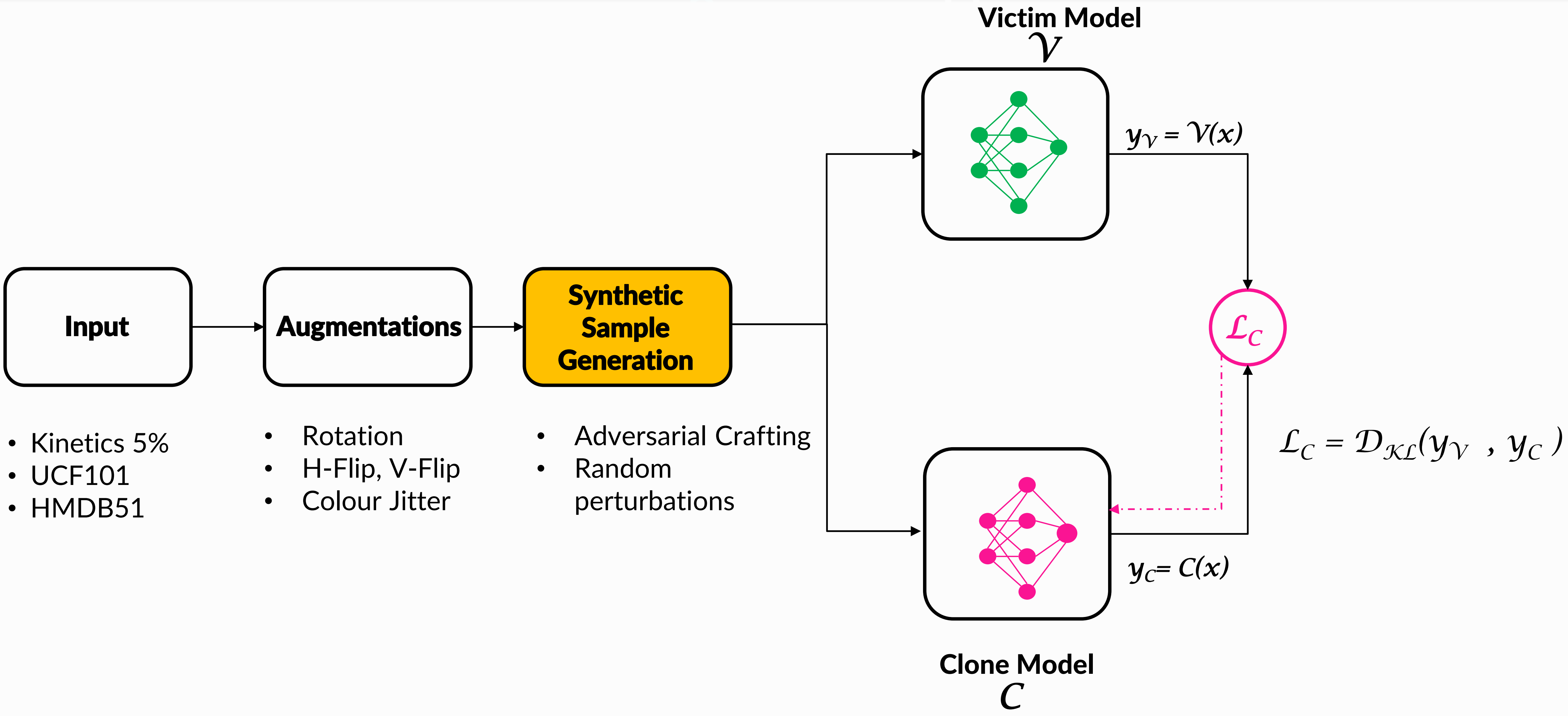
2. Concatenated Datasets

UCF101



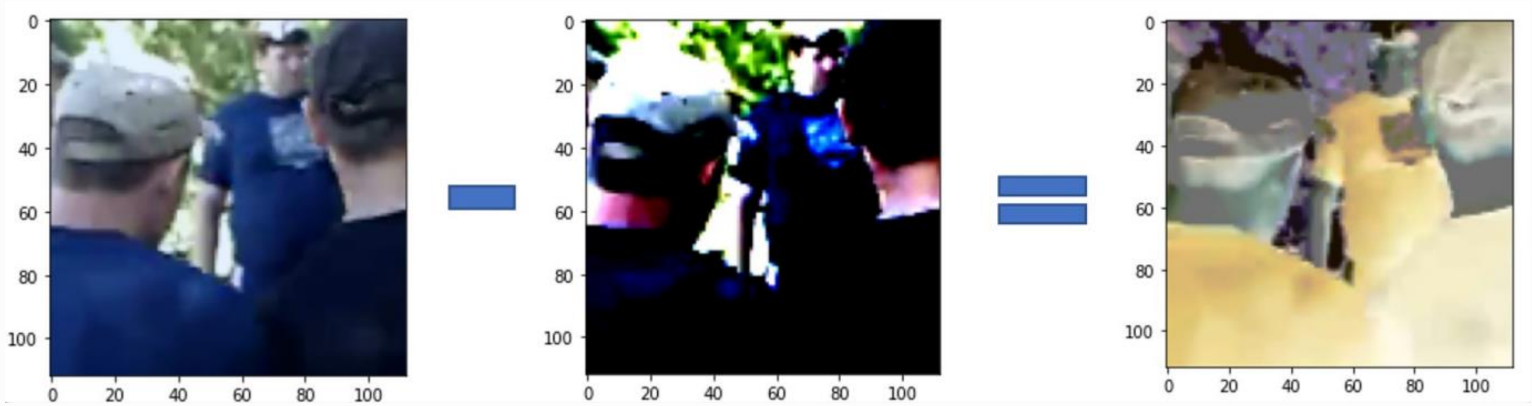
HMDB51

3. Combining PRADA Approach



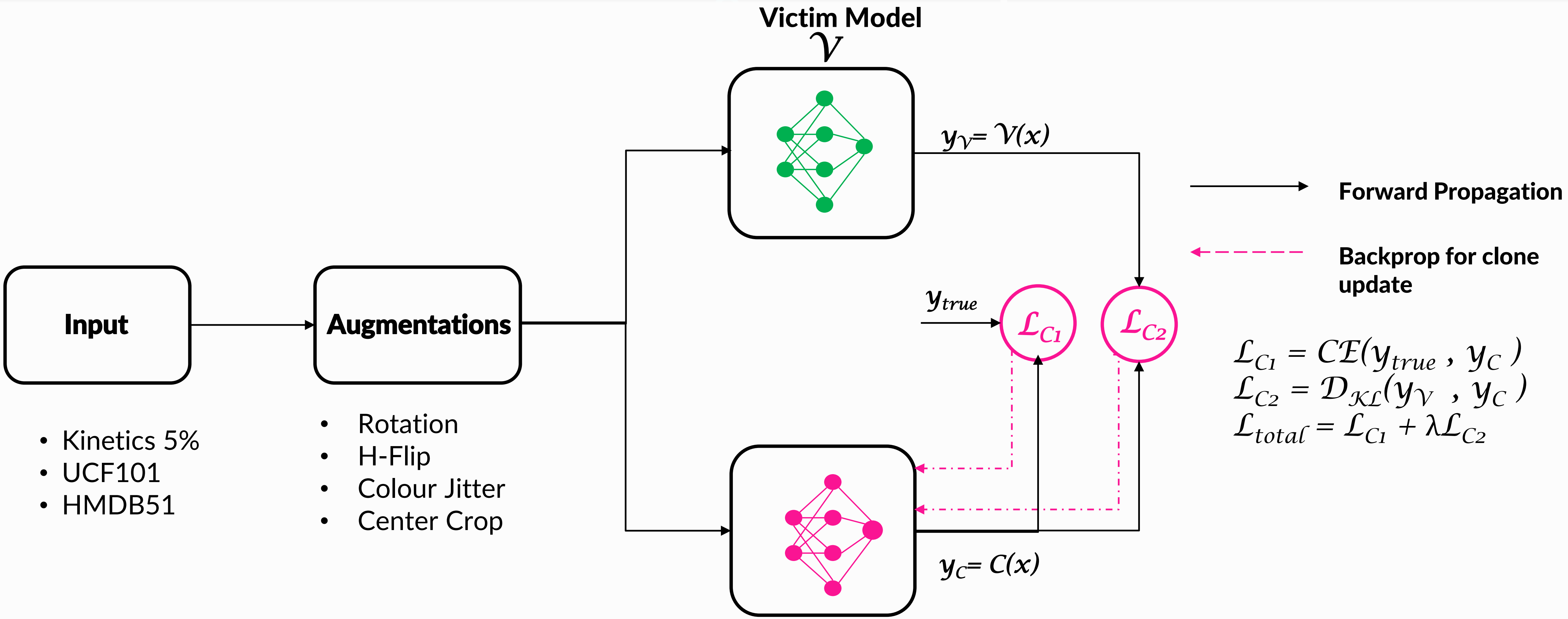
3. Combining PRADA Approach

- Extended the attack strategy proposed in **PRADA**[1] for videos
- Increased coverage of the input space by leveraging **synthetic sample generation**
- **FGSM[2]-like attack** through clone produced novel videos for training
- **Random perturbations** further improved the variety of queries to the victim



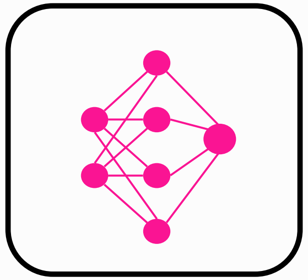
[1]PRADA: Protecting Against DNN Model Stealing Attacks
[2]FGSM: Fast Gradient-Sign Method

4. Combining KD Techniques



$$\mathcal{L}_{C1} = \mathcal{CE}(y_{true}\ ,\ y_C\ )$$
$$\mathcal{L}_{C2} = \mathcal{D}_{\mathcal{KL}}(y_{\mathcal{V}}\ \ ,\ y_C\ )$$
$$\mathcal{L}_{total} = \mathcal{L}_{C1} + \lambda\mathcal{L}_{C2}$$

Clone Model



Clone Model
C

- **C3D**
  - One of the early architectures in video classification.
  - Pretrained on Sports-1M
- **ResNet (2+1)D**
  - Pretrained on IG65M
  - Among Top-20 in Video Classification related tasks

Experimental Results Obtained for Swin-T

Experimental technique	Clone Model	Top-5 Accuracy	Top-1 Accuracy
Augmented Kinetics	C3D	27.5	8.4
Augmented Kinetics	R(2+1)D	42.5	19.1
Concatenated dataset	R(2+1)D	51.8	30.6
Combining PRADA approach	R(2+1)D	34.2	12.67
Combining KD techniques	R(2+1)D	**54.8**	**31.4**

Final Results Obtained for Grey Box

Victim Model	Clone Model	Top-5 Accuracy	Number of Queries
Video Swin Transformer	R(2+1)D	54.8	~0.4M
MoViNet-A2-Base	R(2+1)D	50.4	~0.4M

Scope for Improvement

Black Box

- Increasing number of queries multifold
- Selecting a good prior data distribution
- Stabilizing the generator training

Grey Box

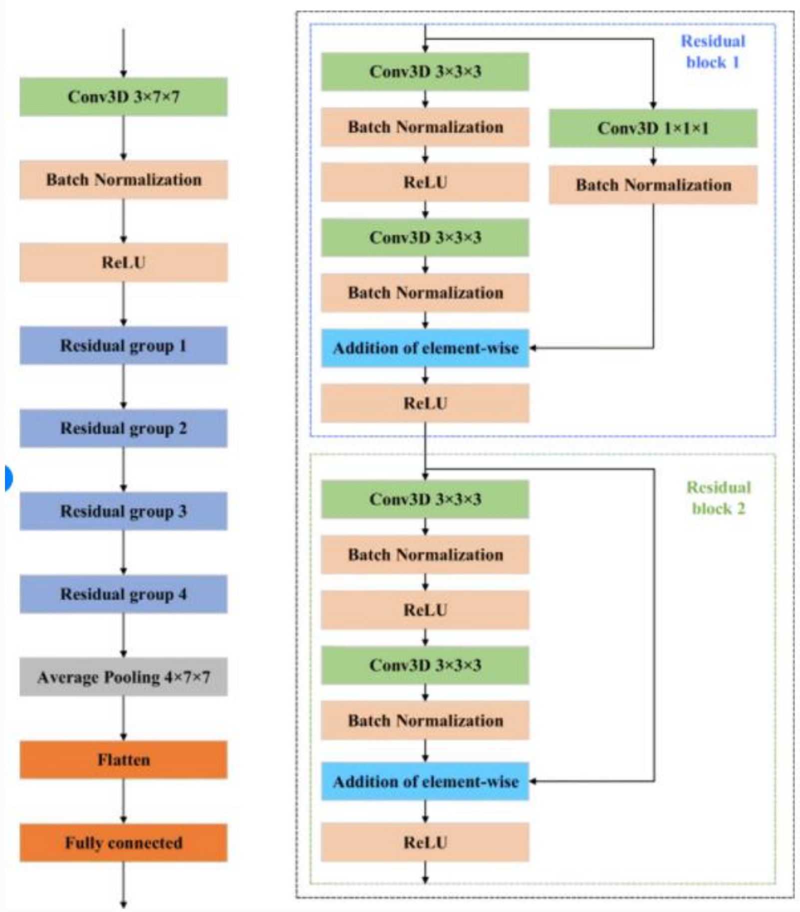
- Extended training duration and faster hardware
- Use generator to create synthetic data from existing distribution
- Use a transformer model as clone
- Use adversarial crafting in better way

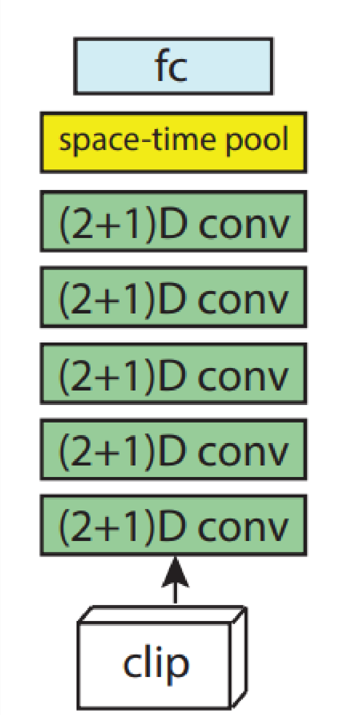
References

1. Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, “Video swin transformer,” arXiv preprint arXiv:2106.13230, 2021
2. Dan Kondratyuk, Liangzhe Yuan and B. Gong, “Movinets: Mobile video networks for efficient video recognition,” arXiv preprint arXiv:2103.11511, 2021
3. M. Juuti, S. Szyller, S. Marchal, and N. Asokan, “Prada: Protecting against dnn model stealing attacks,” in 2019 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2019.
4. D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6450–6459, 2018
5. S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(1):221–231, 2013.
6. S. Kariyappa, A. Prakash, and M. K. Qureshi, “Maze: Data-free model stealing attack using zeroth-order gradient estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021
7. J.-B. Truong, P. Maini, R. J. Walls, and N. Papernot, “Data-free model extraction,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021

Thank You!

Appendix





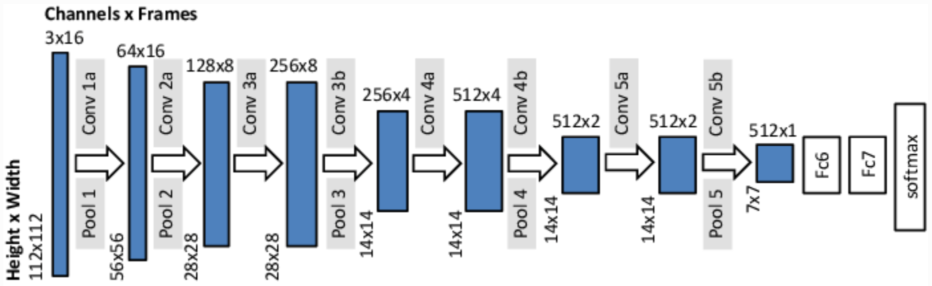


Fig. 3: C3D architecture with eight convolution layers, five max pooling layers and two fully connected layers.

R3D

R(2+1)D

C3D

Appendix

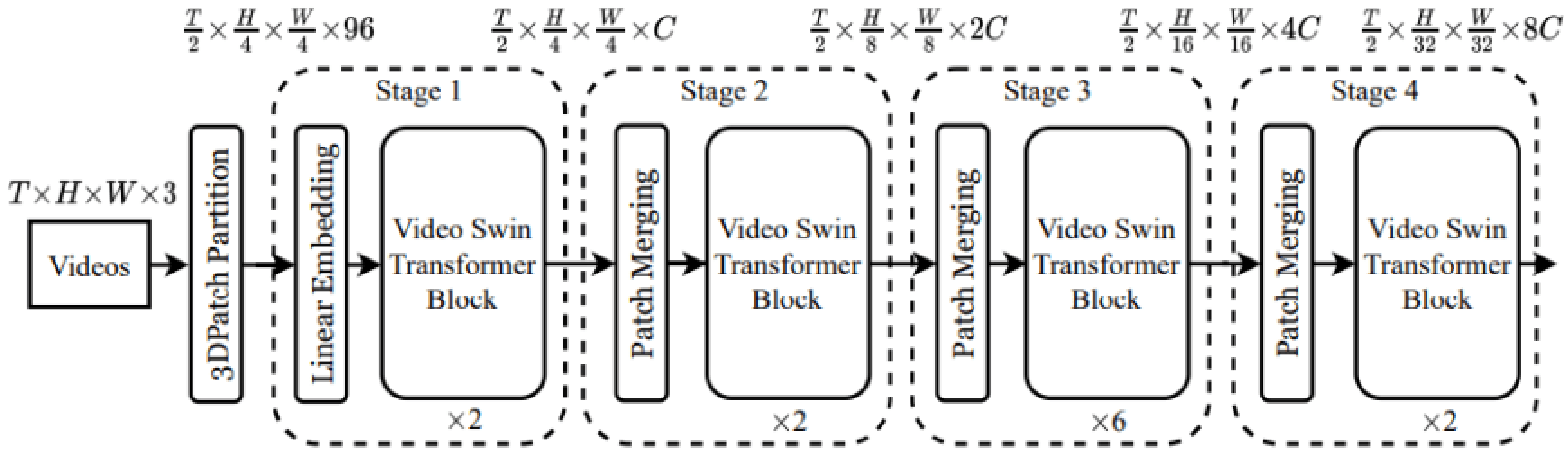
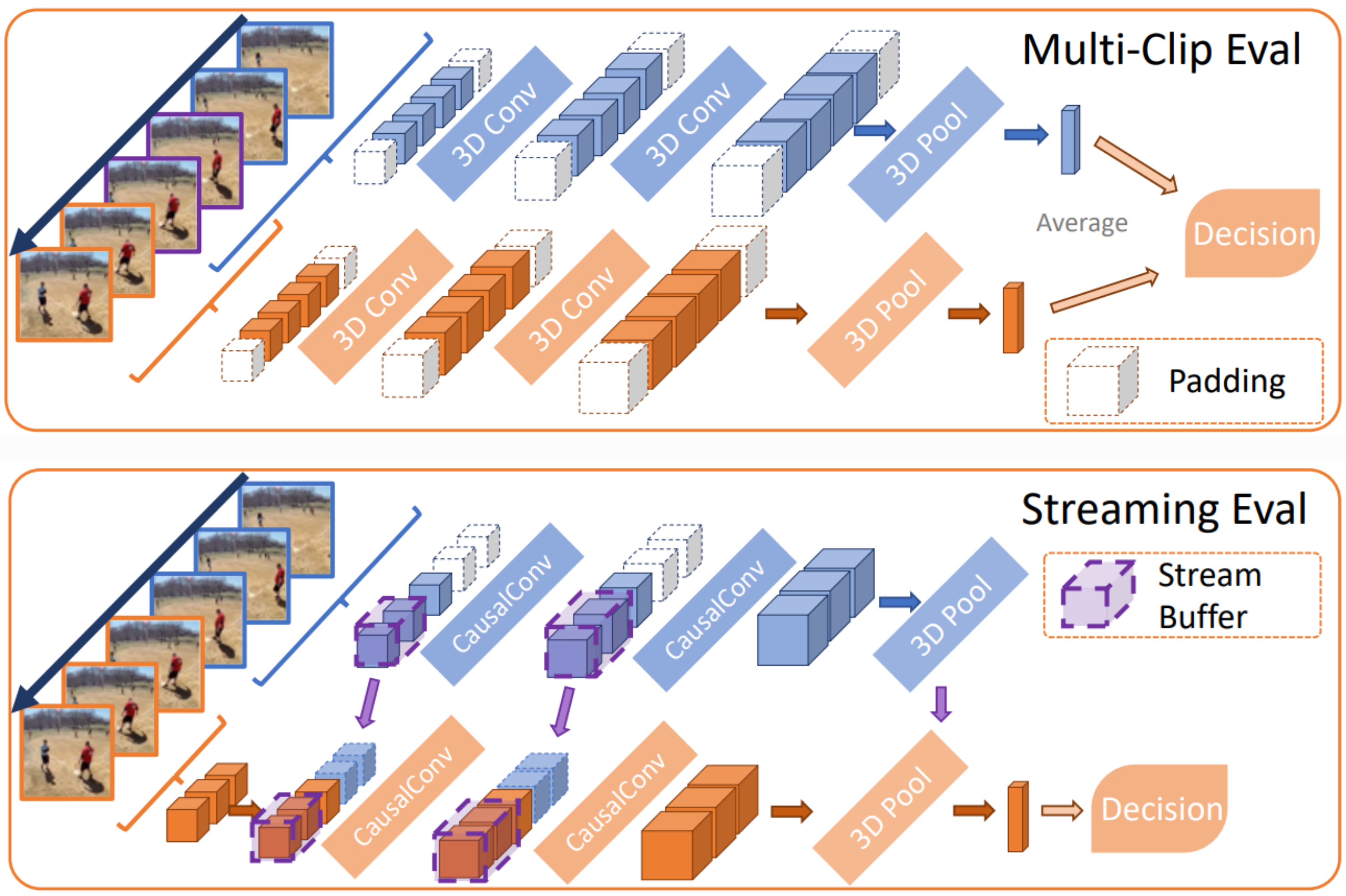


Figure 1: Overall architecture of Video Swin Transformer (tiny version, referred to as Swin-T).

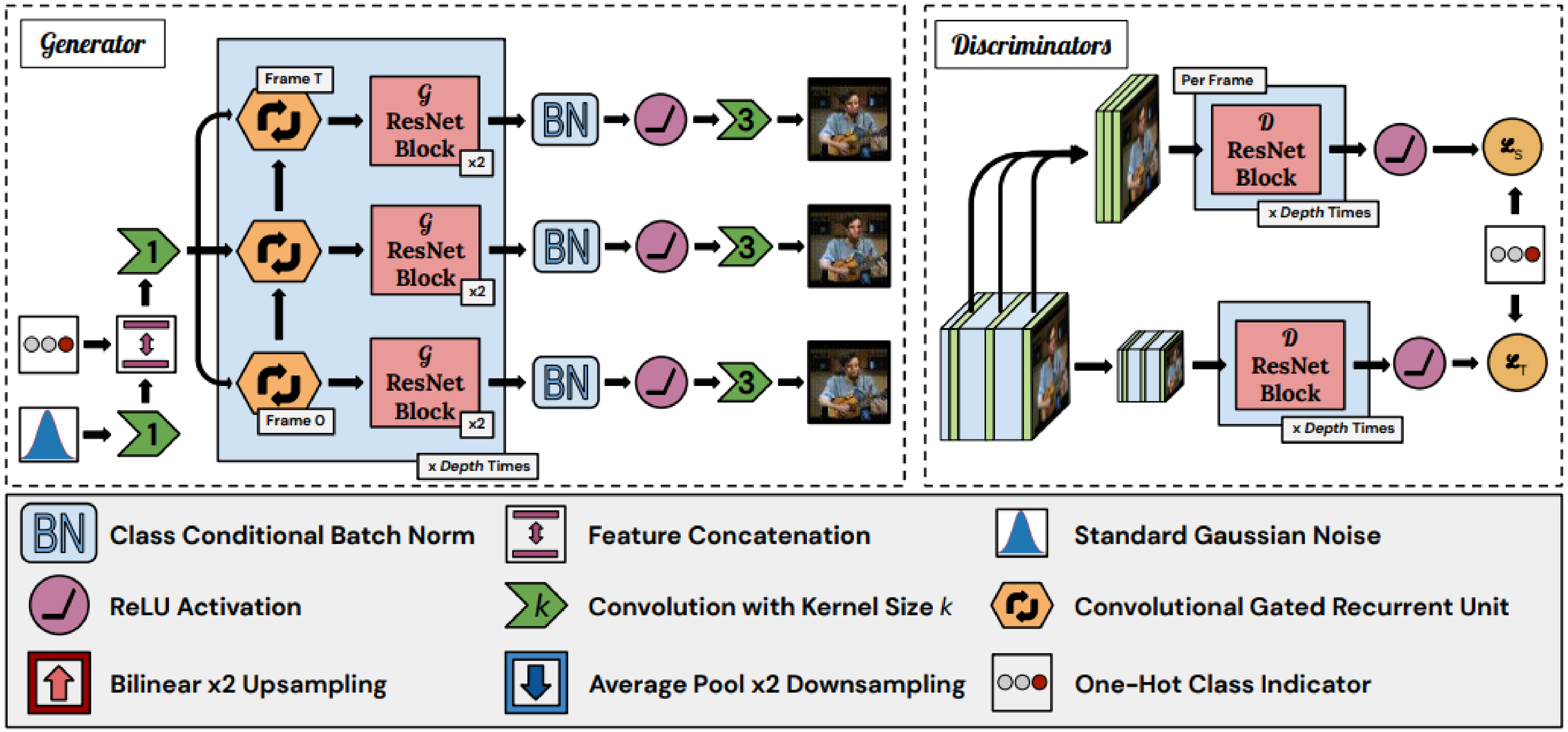
Video Swin-T

Appendix



MoViNet

Appendix



DVD-GAN