"*Temporal Cycle consistency:* for a video-to-video translation."

Kirubel Abebe Senbeto

A Thesis Submitted to the Department of Computing School of Electrical Engineering and Computing

Presented in Partial Fulfilment of the Requirement for the Degree of Masters in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

October 2020

"***Temporal Cycle consistency:*** for a video-to-video translation."

Kirubel Abebe Senbeto

Advisor Prof Yun Koo Chung

A Thesis Submitted to the Department of Computing School of Electrical Engineering and Computing.

Presented in Partial Fulfilment of the Requirement for the Degree of Masters in Computer Science and Engineering

School of Electrical Engineering and Computing

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

October 2020

Declaration

I hereby declare that this MSc thesis is my original work and has not been presented for a degree in any other university, and all sources of material used for this thesis have been duly acknowledged.

Name	Signature
Kirubel Abebe Senbeto	_____

This MSc thesis has been submitted for examination with my approval as a thesis by

Advisor. Name	Signature
Yun Koo Chung (Ph.D.)	_____

APPROVAL OF THE BOARD OF EXAMINERS

We, the undersigned, members of the Board of Examiners of the final open defense by **Kirubel Abebe Senbeto,** have read and evaluated his thesis entitled "**Temporal Cycle consistency: for a video-to-video translation**" and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirement of the degree of Masters in Computer Science and Engineering (CSE).

Name	Signature	Date
Kirubel Abebe Senebto		
Name of the Student		
Prof. Yun Koo Chung		
Advisor		
External Examiner		
Internal Examiner		
Chair Person		
Head of Department		
School Dean		
Post Graduate Dean		

ACKNOWLEDGMENT

First and foremost, I would like to thank the HOLY TRINITY FATHER, SON, and HOLY SPRITE GHOST the Almighty GOD who created everything seen and unseen. Also, I love to Praise the Virgin Mary the Holy Mother of JESUS CHRIST by the song of St. Yared the Ethiopian.

This thesis research is the outcome of a one-year study and hard work, but it's very hard to remember when so many people have helped me in far too many different ways. However, some should be honored for the essential assistance they have offered in the process of study.

I would like to thank Prof. Yun Koo Chung (Ph.D.) Special Interest Group Head for his patient guidance on any issue and his recommendation for a start-up process. Dr. Mesfin Abebe (Ph.D.) for his suggestions and support on this paperwork. Another person worth mentioning is Anteneh Tlaye for his support in evaluating the outputs produced. I am also grateful for all the support and advice I have received from computer postgraduate students, and friends.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2D	2-Dimensional
3D	2-Dimensional
AED	Annotation Edit Distance
AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
CC	Cycle Constraint
CC+CP	Cycle Constraint Plus Feature Preserving
CC+CP+TD	Cycle Constraint Plus Feature Preserving Plus Temporal Aware Discriminator
CGAN	Conditional Generative Adversarial Network
CGI	Computer-Generated Imagery
CNN	Convolutional Neural Network
Conv-nets	Convolutional Neural
CPU	Central Processing Unit
CYCLEGAN	Unpaired Image-To-Image Translation Using Cycle-Consistent
DL	Deep Learning
DX	Discriminator X
DY	Discriminator Y
FCN	Fully Convolutional Networks
FID	Fréchet Inception Distance
GAN	Generative Adversarial Networks
GPU	Graphics Processing Unit
GX	Generator X
GY	Generator X
HFR	High Frame Rate
IDE	Integrated Development Environment
IoU	Intersect Over Union

IS	Inception Score
L1	Manhattan Distance Or L1 Norm
L2	Euclidean Distance
LSTM	Long Short-Term Memory
mIoU	Mean Intersect Over Union
ML	Machine Learning
MoCycle-GAN	Mocycle-GAN: Unpaired Video-To-Video Translation
MPI	Max Planck Institute
P1	Human Evaluation Study Protocol One
P2	Human Evaluation Study Protocol Two
P(z)	Noise Data Distribution
PIX2PIX	Image-To-Image Translation With Conditional Adversarial Nets
RC	Recycle-GAN
RC+TD	Recycle-GAN Plus Temporal Discriminator
ReCycle-GAN	Recycle-Gan: Unsupervised Video Retargeting
RGB	Red Green Blue
RNN	Recurrent Neural Network
TPU	Tensor Processing Unit
VAE	Variational Autoencoder
$\boldsymbol{X \rightarrow Y}$	X To Y
$\boldsymbol{Y \rightarrow X}$	Y To X

ABSTRACT

Generative Adversarial Networks (GANs) is a deep learning method that is developed for synthesizing data. Area of applications for which it can be used are image-to-image translations, Video-to-video translation, and video retargeting. However, to train those models there is a need for large amounts of complex paired data which is hard to find. To collect datasets, especially when we need a paired dataset is time-consuming and expensive. One way to overcame this problem is to collect datasets in one domain and translate it to another domain using image translation techniques to make it a paired dataset. Various research has leveraged enormous in image translation by the use of GANs on an unpaired dataset. As far as video translation is concerned, current GAN-based approaches do not entirely leverage space-time knowledge in videos.

This research examines the idea of using GANs for the utilization of spatial-temporal information in a video by extending the unpaired video-to-video translations model (ReCycle-GAN) to enhance spatial-temporal video translation. In particular, previous methods suffer from Object disappearance, Object dislocation, and flickering Artifacts. To Mitigate these issues, this work proposes to adds feature preserving loss and temporal aware discriminator to the Cycle GAN and ReCycle GAN to generate more temporal consistent videos. Extensive qualitative and quantitative assessments demonstrate the notable success of the proposed system against existing methods. Average human evaluation study *has shown that this research excels at 60% compared to CycelGAN and 35% on ReCycle GAN. This paper concludes that Adding feature preserving constraints and temporal aware discriminator does improve temporal coherency of output video.*

Keywords: Cycle GAN, ReCycle GAN, Spatial-temporal information, Unsupervised Video-to-video translation

CHAPTER ONE

1. INTRODUCTION

1.1. Background

Computer Vision is measured among the most fascinating fields in computer engineering and artificial intelligence. The chase of providing machines with a sense of sight that is even better than that of humans is keeping researchers busy and motivated. There is an extensive range of problems with active research within the field of computer vision, such as facial recognition, object classification, scene recognition, and Domain transfer. In this thesis, the focus is on Domain Transfer.

Unsupervised Video retargeting is transferring of sequential content information form one domain to another while preserving the style of the target domain. Such domain transfer could be served in numerous areas including motion translation from one person to another person and video colorization – monochrome video to color and day ↔ night. Recent works use the generative adversarial network for retargeting and style transfer image-to-image translation problems. This proposal purposes to extend ReCycle GAN by Bansal et al [1] to improve frame to frame continuity (motion consistency) by introducing additional constraints to the network.

Style transfer is a subproblem of domain transfer that aims to translate or map domain to domain. Such domain transfer could be served in numerous areas, including classical language translation to motion translation from one person to another person and video colorization. Since this work uses Images and video as input data, we can say that *style transfer is a process of repainting a given image by style image while preserving it contain.*

$$Input\ Image\ +\ Style\ Image\ ->\ Output\ Image\ (Styled\ Input)$$

Last year (2018), all Artificial intelligence news [2] headlines were screaming about a paint drawn 100% by AI sold $432,500 (fascinating, isn't it?). (Because of style transfer data scientist does not need to buy a hundred-thousand-dollar painting for decorating his living room while he can have one when he is home siting in Infront of his laptop, marvelous)

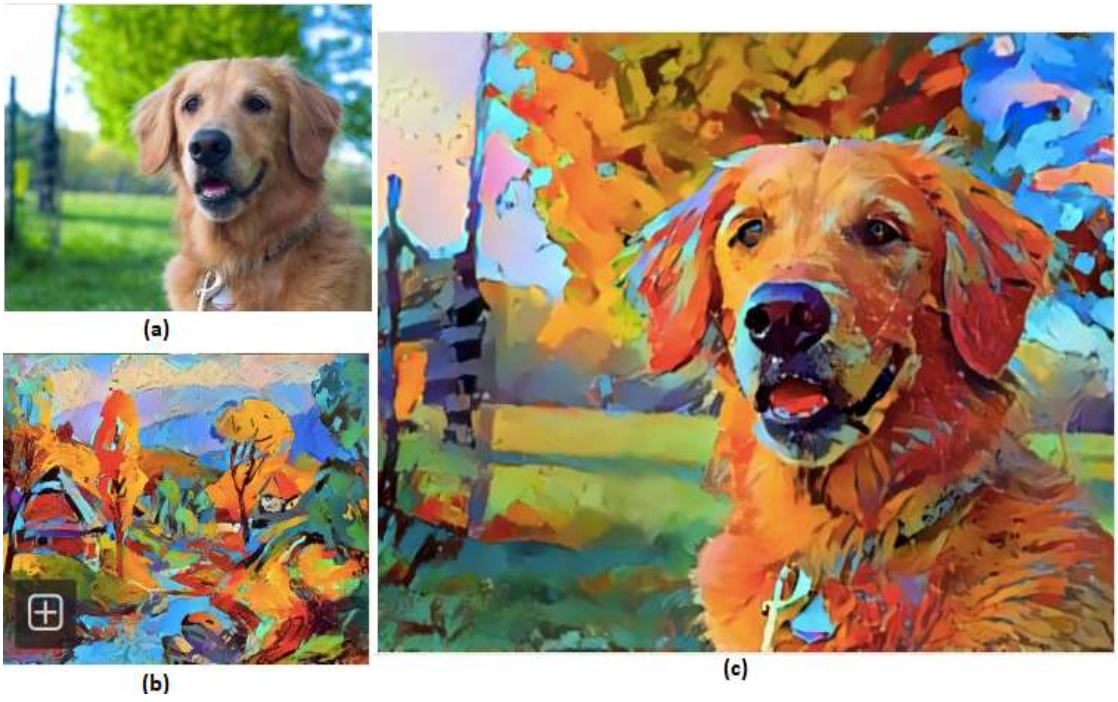


Figure 1-1(a) input image. (b) style image. (c) output image

Source: Adapted from [3]

Perhaps the first successful neural style transfer paper was published in 2015 by [4]**.** After this work, many researchers came with a more realistic synthetic image. pix2pix [5] introduces with a supervised image-to-image translation, but pix2pix needs paired data for training which is expensive and unlikely -needs paired data examples from both domains to learn. Other finest GAN paper Cycle-GAN by Zhu et al. [6] present unsupervised style transfer to overcome pix2pix problem due ***cycle consistency*** *–If I take an input image of horse feed it to the network it generates zebra image then take the output image as an input again run the second transformation I expect to get the same horse image I started with.* Cycle-GAN place foundation for unsupervised image transfer problem in computer vision.

Video-to-video translation is a natural extension of an image-to-image translation (since the video is a sequence of images). Recent works use the generative adversarial network for retargeting and style transfer images to image translation problems. This work aims to extend video-to-video translation to the improved frame to frame continuity (motion consistency) by introducing additional constraints to the network.

In order to clearly understand this thesis research question, we need to have a clear and brief introduction to the following topics. A more detailed discussion will be held in the proceeding section.

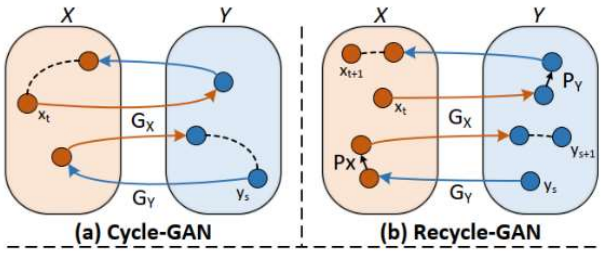


Figure 1-2 Cycle Gan vs Recycle GAN

Source: Adapted from[1]

1.1.1. Generative Adversarial Networks

GAN (Generative Adversarial Networks) fit into the conventional algorithms called Generative models. The term 'generative' refers to the fact that these networks can learn how to produce data samples that are similar to real ones in the training dataset. It is a sub-set of ML which aims to study algorithms that learn the data distribution of the given data, deprived of specifying a target value. This method builds upon the success of using deep neural networks in content generation.

Generative Adversarial Networks are collected of two Networks work against each other in a zero-sum game framework. The first network is called a Generator. The goal is to produce new data close to that of samples from real datasets. The Generator could act as a human art forger, which creates fake works of art.

The second Network is entitled the Discriminator. This model aims to recognize if an input data is ***'real'*** — goes to the original dataset — or if it is ***'fake'*** — generated by a falsifier Generator Network. In this scenario, a Discriminator is corresponding to the law enforcement agent (or an art expert), which tries to spot artworks as truthful or fraud. Successful training of a GAN requires reaching an equilibrium state between two opposing objectives, unlike CNN or Long Short-Term Memory (LSTM) where the training objective is to minimize or maximize the value of a single loss function.

Conditional GAN

The conditional GAN [10] is an extension of the [5] original vanilla GAN, by introducing a conditioning variable into both generator and discriminator network. So instead of generating random data, the newly introduced condition variable would allow generating a particular data distribution specified by the conditioning variable. Mainly, the random noise input to the generator will be concatenated with a variable specifying the condition to generate the fake data, meaning to generate the fake data cGAN use random noise and newly introduced conditional variable.

Video-to-video transfer

Video-to-video transfer is a domain transfer problem that aims to transfer sequential content information form one domain to another while preserving the style of the target domain. Current approaches for domain transfer categories broadly into three classes. Early techniques use classical computer vision mechanism work specifically designed for particular body parts such as the human face [7] they lack generalization and does not work well if there is occlusion. The second approach use paired image-to-image translation such as pix2pix -in an image it takes a pixel, then converts to another pixel. [5] use conditional GAN [8], learn a mapping between paired input to the output image. The third category is unsupervised and unpaired data domain transfer like Cycle-GAN [1], which enforces cycle consistency for the unpaired image.

The recent state of artwork work ReCycle-GAN by Bansal [1] motivated by [6] proposes video retargeting via spatiotemporal constraint though directly synthesizing future frames via temporal predictor to preserve temporal continuity. Bansal et al., clams video-to-video translation are **still under constraint** since their work result shows of video-to-video transfer has very flickering output. This proposal proposes to extend Bansal et al., work to improve temporal continuity between adjacent consecutive frames by introducing additional **temporal cycle consistency constraints also proposes to introduce Spatiotemporal video-to-video** translation for better realistic results.

1.2. Motivation

Recent deep learning achievement has been done because of the enormous amount of data available nowadays. However, still, there is a big problem to collect data especially when there is a need for paired data set (such as day and night) since capturing datasets in two (or more) completely different environments is dreadful. After Goodfellow paper [9], GAN has been used for various applications in a wide range of areas for numerous applications. Image-to-image translation is maybe of the significant one. Image translation could be the mechanism to overcome such problems. Advancement in GAN enables scientists and researchers to create fake images indistinguishable from real ones[10][11].

By extending the image translation idea, various researchers propose a number of approaches for the video-to-video translation network to learn both spatial and temporal domains but failed [12] to Achieve the potential found by image translation networks. Generally, currently used video-to-video translation networks, prone to object disappearance problems, and arbitrary strange motion on the generated videos make translated videos more unrealistic.

1.3. Statement of the Problem

Problem formulation: Inspired by recent work Recycle-GAN in the unpaired video-to-video translation, The notion of a research problem. Let we have two videos archives in source and target domain $X = \{x\}_{t=1}^{T}$ and $Y = \{y\}_{x=1}^{T}$ respectively, cycle constraint enables an image-to-image translation in mutually frontward and backward mapping. There are two mapping functions $G_{AB}\ and\ \ G_{BA}$ mapping from domain $X \rightarrow Y$ and $Y \rightarrow X$ correspondingly form target domain to source and vice versa. $G_{AB}(x_t) = \widetilde{x_t}$ where x_t is input video frame at time t and $\widetilde{x_t}$ is a synthetic frame in X domain same is true in Y domain. Cycle consistency constraint $G_{BA}(\widetilde{x_t}) = \widetilde{\widetilde{x}_t}$ so then $\widetilde{\widetilde{x}_t} \approx x_t$ as well as $G_{BA}(\widetilde{y_t}) = \widetilde{\widetilde{y}_t}$ so then $\widetilde{\widetilde{y}_t} \approx y_s$.

Besides the preservation of cycle consistency in each frame this work-study mapping temporal consistency between consecutive frames in both domains. Meaning let optical flow between x_t and x_{t+1} is f_t and optical flow between $\widetilde{\widetilde{x}_t}\ and\ \widetilde{\widetilde{x_{t+1}}}$ is $\widetilde{\widetilde{f}_t}$, then, temporal cycle consistency need to enforce motion consistency via minimizing the difference between $f_t\ and\ \widetilde{\widetilde{f}_t}$. Recycle-GAN [1] claims **"*video-to-video translation is under constraint*"** this

work proposes toward adding temporal cycle consistency to the extended video-to-video translation to see more constraints in its result.

To do so, an extensive experimental attempt was made with the purpose of answering the following research question.

- » Can Adding additional constraints improve temporal coherency for video translation?
- » What is the effect of temporal discriminator on the unsupervised video-to-video transfer?

1.4. The Objective of the Thesis

1.4.1. General Objective

The general objective of the study is to add and implement Temporal Cycle Consistency constraint for the Video-to-video Translation. This work is motivated by [1] ReCycle-GAN.

1.4.2. Specific Objectives

The following specific objectives are addressed to achieve the general objective.

- » Reviewing related works to understand the area and the works that are done by others.
- » Gathering dataset for training and testing.
- » Preprocessing the dataset in order to enhance.
- » Embedding feature preserving loss to Cycle-GAN
- » Modifying the discriminator network to make it aware about temporal information.
- » Changing the discriminator network to supply it with temporal knowledge.
- » Adding learning constraints to the Cycle-GAN and recycle-GAN networks.
- » Designing a GAN model for video translation model using Keras and TensorFlow framework.
- » Training the model using a proper dataset.
- » Testing the trained model with a test set.
- » Evaluating the performance of the model.

1.5. Research Methodology

The following methods and techniques are applied in order to meet the objectives of this study.

1.5.1. Data Collection

This study uses a machine learning approach to solve the problem, so data is an essential part of the study. Videos (sequence of Images) are collected for both training and testing. Those data (Datasets) are collected directly from the internet (available popular unpaired dataset) for the purpose of the study. Besides the popular datasets available on the internet, this work plan to **collect local video dataset to inference the study**. Most of these videos were long and made up of several frames, each shot being a different scene.

1.5.2. Literature Review

This study uses a literature review to enhance the research. Recent related literature is reviewed to get an insight into current trends and methods to solve the problem at hand. Necessary documents and tools are also reviewed for the development of the prototype.

1.5.3. Evaluation

The result will be analyzed to describe the performance of the proposed architecture on a test data set. The performance of this work will be analyzed in real-world scenarios videos from the dataset.

1.5.4. Implementation Tools

For the development of the deep learning network architecture in addition to reporting this thesis work finding the following tools and software will be used.

- » OpenCV, TensorFlow, and Keras API, MATLAB will be used for modeling networks, coding the as well as training and testing.
- » Microsoft Word, PowerPoint, and Grammarly are software plain to use for editing, Presentation, and check Plagiarism checking.
- » GPU to train the network more efficiently.

1.6. Scope and Limitation of the Study

1.6.1. Scope of the Study

The scope of the thesis work within a given time and resource includes: -

- » Translate a given domain video (sequence of image) to another domain.
- » Add learning constraints to Cycle-GAN and the ReCycle-GAN network.
- » Blend spatial information to temporal information to improve the consistency of video-to-video translation using introduced constraints.

1.6.2. Limitations of the Study

This thesis work does not cover the following due to time and resource limitations.

- » One to many video-to-video translation is **not** a part of this work. The network will be trained to translate from one domain image to another domain, which is one to one correspondence (Doesn't consider multi-domain translation such like [13]).
- » The video does not zoom in or out throughout the whole process.

1.7. Organization of the Thesis

The remainder of this thesis is organized as follows:

Chapter Two: discusses the background literature and related works regarding the image-to-image translation, video retargeting, and video-to-video translation. This chapter also elasticities the theoretical framework of Deep Learning and Generative Adversarial Network.

Chapter Three: features the research methodology, including different methods and techniques used to develop the solution and select the appropriate one. Data collection method, design tools, prototype development framework and platforms, and evaluation methods are also discussed

Chapter Four: will cover points about the proposed solution in detail and the working environment setup. Discuss the specification of an image-to-image translation network and temporal information blending with the spatial model. Flow chart and pseudocode for

implementation, training, and testing with mathematical correspondent descriptions have been discussed.

Chapter Five: Explains how the desired proposed solution is implemented. The working environment, cycle-GAN implementation, with training pseudocode implementation described using snip code.

Chapter Six: The obtained testing result from Temporal Cycle consistency for a video-to-video translation model is presented and Compare with the other related work in order to have the best judgment.

Chapter Seven: concludes the research and provides directions for possible future work.

CHAPTER TWO

2. LITERATURE REVIEW AND RELATED WORK

2.1. Introduction

In the early days, the software was hard-coded with commands and rules represented in a mathematical formula Those methods short-handed as they attempt to simulate sophisticated real-world situations and data. To solve this scarcity another software development paradigm emerges which asks whether the machine can understand as humans by having cognitive taught called Artificial intelligence or AI. AI has made a lot of improvements in recent years, allowing AI applications completely capable of gathering and extracting data to learn from the knowledge sequence called machine learning [14].

Machine learning (ML) is a vast research area with diverse learning capacities and it keeps to grow. From different learning capabilities of machine learning, unsupervised learning is one of them. This learning type is the task of clustering unorganized data to organize them based on the information they composed. Another kind of machine learning is supervised learning, unlike the unsupervised approach, for every data input there is a corresponding output label (x, y), the network task is to learn how to map from input x to it label y in testing time (R.B. data is paired dataset). Semi-supervised learning is another kind of machine learning approach when some section data is labeled and the rest section is unlabeled. Even ML plays very tremendous work, in recent days, but it still fails to process complex data like image and video. So as to work with such complex data types Deep Learning is an alternative which subfield of machine learning.

Deep learning (DL) is another mechanism of learning from data in sequential filter layers while the previous approaches learn from data representation. DL filter present data information one in terms of another which build a hierarchical ordered of features. These features enable us to extract high-level features. The first emerged deep learning was the artificial neural network ANN [15]. Convolutional Neural Network (CNN) was introduced by Yann LeCun [17] in 1989 to recognize handwritten digits but the lack of a large dataset and low computing capability at a time limit its popularity. CNN bloom after Alex-Net [16] archive a significant win in the ImageNet contest in 2012 using the CNN image classifier

network. Alex et al work opens many scientists and researchers eye to the power of deep learning. Today high-performance models and networks are designed for face detection, object classification, and recognition.

CNN models score a high success on classification models, classification models are tasked to predict label y for given input x. In other sense, Generative models are the Multiplicative reciprocal of classification networks.

The most impressive new emerging deep learning approaches are Generative models. these models are trained to learn the essence of the data distribution to generate fake samples that are similar to real ones. GANs application is a very broad and image-to-image translation, video-generation can be some examples of its application area.

GAN introduced by Goodfellow et.al in 2014 [9] the model consists of two stand-alone DL networks named Generator and Discriminator model which have fight one another in an adversarial relationship. (footnote 1). Hence the discriminator plans to distinguish fake samples $p(z)$ from the real input data $pdata(x)$. Oppositely the generator network is all about generative fake samples as good as the real ones. During training, both models update their loss to improve until Nash-equilibrium is reached. The next section discusses the detail inside of GAN.

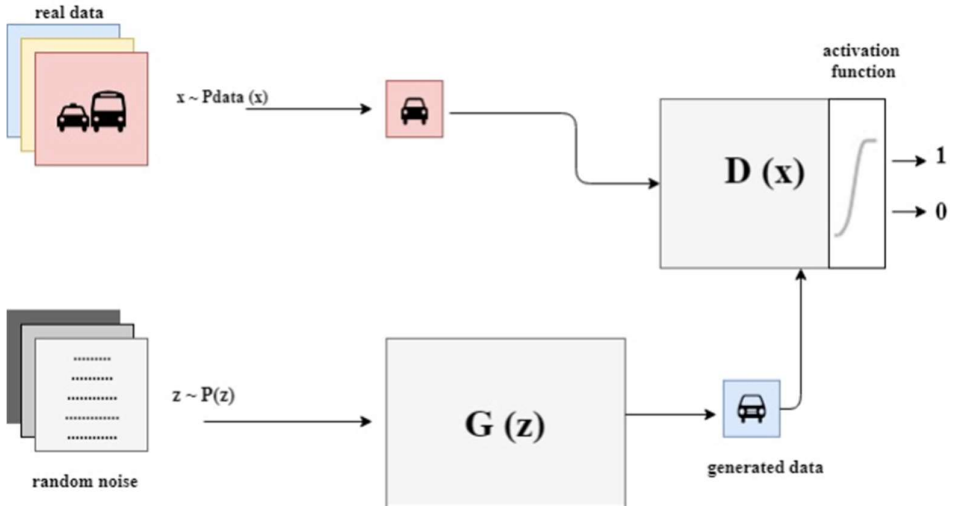


Figure 2-1 GAN framework structure GAN: Discriminator D(x) and Generator G(z)

In this chapter, the paper briefly describes the technologies, methods, and frameworks mentioned throughout the thesis.

2.2. Inside GAN

Lets see the detail inside of GAN. As discuss GAN in the previous section, GAN consists of two independent networks Generator and Discriminator as shown in figure 2_1, which are represented by differentiable functions concerning each network's input and parameters. The discriminator is defined by a function $D(x)$ where x (observed variable) is the input which is a real dataset. $D(x)$ gives the likelihood that x came from $pdata$ (real distribution) rather than $p(z)$ (fake distribution). It is a binary classifier with two classes, when x is real the probability is 1 and when x is synthetic the probability is 0. The discriminator can be seen as a typical CNN that transforms a 2- or 3 (grayscale or RGB) dimensional matrix of pixels into probabilities.

The generator $G(z)$ accepts input from a random noise distribution $p(z)$ where z (latent variable) is the input and generates an image as its output x_{fake}. The generated image is fed into the discriminator network $D(x)$, which attempts to classify the image as real or generated by G. The result of the classification is backpropagated to the generator to help it learn how to produce images with a closer representation of the input data.

The loss function used in training the networks is formulated as [9]:

$$\mathrm{L_{adv}} = \min_{G} \max_{D} E_{x\epsilon X}(\log D(x)) + E_{x\epsilon Z}(\log(1 - D(G(z)))) \qquad \textit{eq.(2.1)}$$

Equation 1 Adversarial loss function

The generator can be seen as a kind of reverse CNN. It takes an z-dimensional vector of noise and upsamples it to an image using transposed convolution(transconv) to be specific transconv can be seen as a convolutional upsampling. Conceptually, the discriminator in GAN provides guidance to the generator on what images to create implicitly in the training process. Now we can discuss how to training GAN.

2.2.1. GAN Training

Machine learning is all about Generalization in which the model learns from real-world examples so that it can predict the test set accurately. No difference for GAN training is all about the process of learning to mimic the real dataset samples. Unlike many deep learning

models, training is a bit tricky so let us dive into it. However, before that, let sees an adversarial conflict between discriminator and generator.

The Discriminator's goal is to be as precise possible (binary classification). For the real examples $x, D(x)$ seeks to get as real as possible to 1 (label for the positive class). Meaning $x_{fake}, D(x_{fake})$ attempts to converge 0 as possible (label for the negative class).

The Generator's goal is the reverse. It tries to find a way to fool the Discriminator by producing fake example x_{fake} that are alike from the real data in the training dataset. Mathematically, the Generator strives to produce fake examples x_{fake} such that $D(x_{fake})$ is as close to 1 as possible.

Table 1 generator goal vs discriminator goal

Generator	x_{fake} such that $D(x_{fake})$ is as close to 1 as possible.
Discriminator	x_{fake}, such that $D(x_{fake})$ tries to be as close as possible to 0.

Now let's back to GAN and see pseudocode for training GAN (*R.B* its iterative process)

I. Train the Discriminator:
   a. Take a random mini-batch of real examples: x.
   b. Take a mini-batch of random noise vectors z and generate a mini-batch of fake examples: $G(z) = x_{fake}$.
   c. Compute the classification losses for $D(x)$ and $D(x_{fake})$, and backpropagate the total error to update $\theta^{(D)}$ to minimize the classification loss.

II. Train the Generator:
   a. Take a mini-batch of random noise vectors z and generate a mini-batch of fake examples: $G(z) = x_{fake}$.
   b. Compute the classification loss for $D(x_{fake})$, and backpropagate the loss to update $\theta^{(G)}$ to maximize the classification loss.

Unlike other deep learning training Notice that in step 1, the Generator's parameters are not updated intact while training the Discriminator. Similarly, the Discriminator's parameters intact while in the Generator session. The reason GAN allows updates only to the biases and weights of the network being trained is to isolate all deviations to only the constraints that

are under the network's control. This guarantees that separately generator and discriminator get relevant signals about the updates to make, without interacting from the other's updates meaning each two players taking turns to update their weights. This process continues until the Nash equilibrium.

GAN is based on the adversarial game between two networks. In short, if the Generator wins the Discriminator loses and vice versa of the other wins. In-game theory, the Generative network converges when the generator and the discriminator hit the Nash equilibrium. This is the optimum point for the GAN loss minimax function (equation 1). Regarding GAN at Nash equilibrium discriminator no longer able to distinguish between real and fake samples so it randomly classifies ($accuracy = 50\%$).

2.2.2. Conditional GAN

Even though GAN models are able to produce new random possible examples for sample data, There is no way to identify the types of images generated. However, in order to imitate the original data set and images, the network seeks to define the composite relation between the latent space inputs in the generator.[9], [18].

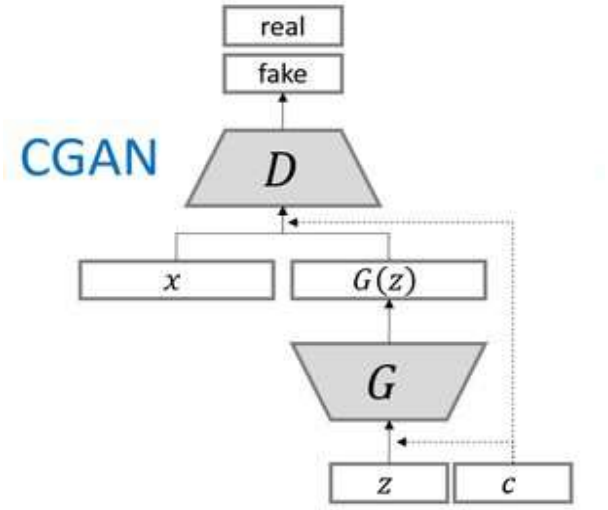


Figure 2-2 cGAN Architecture

Mirza et al. propose The conditional generative adversarial network, or cGAN [8] for short, which is a type of GAN that involves the conditional generation of images by a generator

model. Image generation can be based on the label of the class [1]. It requires the Generator network to produce only the target class of frames of a given form by a conditional variable. The conditional variable C is fade to the generator and discriminator networks as shown in figure 2-2 above. This work unlocks opportunities for many fascinating research topic like image-to-image translation, style transfer and video retargeting [1], [5]. The next section will discourse about Image-to-image translation.

2.3. Image-to-Image Translation

Let start by Abto software AI software company from Europe say about style transfer when they announce their research product *"you may hear A magician can make his trick with just a wave of a magic wand, but its old news. Here in our lab, our engineers can make their magic with just one click! Interested* ***how the same winter landscape would look in summer****"* [19] I was wondering too winter to summer Absolutely fascinating.

Recent advancements in GANs [9] empowers style transfer models to create realistically looking [4]–[6], [20] adapted image (2-4 B show image-to-image transfer from sunny to rainy). Image-to-image translation aims to learn a mapping function between the input image and out image in different domains. When we talk about Image-to-Image basically learning involves the precise modification of an image while preserving contain information and it requires large datasets of paired images that are complex to prepare, meaning the dataset should contain images that are one to one correspondence as shown 2-4. The primary difficulty in the image-to-image translation is they need paired data set for training, but in reality, doing so is very expensive and not scalable, but some work achieves good results. pix2pix[5] is one of them, which is a conditional Generative model by Isola et al. train in a supervised manner using a paired dataset that fits into a supervised image-to-image translation. Pix2pix as the name indicates it learn to map pixel from the first image to the second one.

Because in reality pair datasets are very rear and expensive Zhu et al. came up with Cycle-GAN [6] which was invented to learn bidirectional mapping in the absence of paired training data via Cycle consistency loss. *Cycle Consistency loss* utilizes to learn transformation

[1] conditioning variable *C* could be any type of information. Like Image, tabular information or….

between two domains in a frontward and backward fashion. Cycle consistency constraint is not a new idea; in fact, very old news in natural language processing. The following example gives a simple illustration. Assume using language translation from Amharic (አማርኛ) to English in both directions. When the user input "ስም አበበ ይባላል።" the model should generate "My name is Abebe" perhaps if the user translates "My name is Abebe" to Amharic back again it should generate the original text "ስም አበበ ይባላል።". Meaning the difference between the original text and regenerated text should be minimum. I use google translator to demonstrate this example, as shown in figure 2-3.

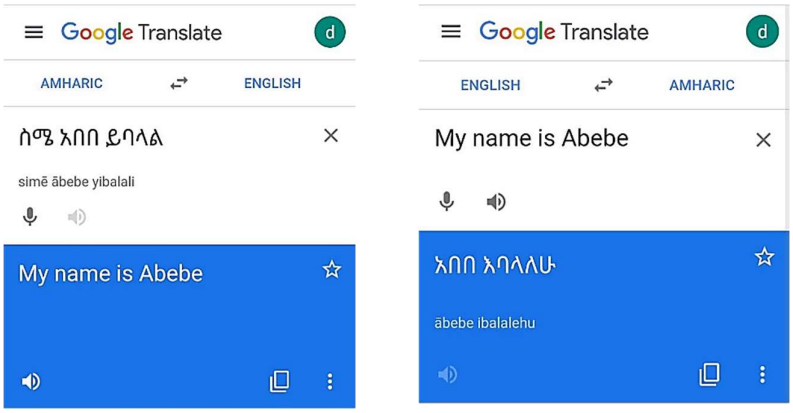


Figure 2-3 Amharic to English language translation using google translator (Example).

The general architecture of Cycle-GAN contains two generators and discriminators for each domain. Where one generator translates from domain A to B while the others do the reverse. Let us see it in bit detail using a table 2.

Table 2 Cycle-GAN generator and discriminator operation.

G_{AB}	Translate from A to B	$A -> \dot{A}$
G_{BA}	Translate from A to B	$B -> \dot{B}$
D_A	Classify real A and fake $\dot{B}$	1 for A, 0 for $\dot{A}$
D_B	Classify real B and fake $\dot{A}$	1 for B, 0 for $\dot{B}$

$A\ and\ B\ are\ real\ image\ from\ domain\ A\ and\ B\ respectively.$

$While\ \dot{A}\ and\ \dot{B}are\ generated\ images\ from\ G_{A_B}and\ G_{B_A}respectively.$

$R.B\ \dot{A}\ is\ in\ domain\ B\ whwere\ as\ \dot{B}\ in\ domain\ A$

The loss function of the network could be formulated as: $min \sum \left|\left|x - G_{A_B}\left(G_{B_A}(x)\right)\right|\right|$

Meaning translate a given image are x and reconstructed image x_{rec} the difference should be the minimum $(x \approx x_{rec})$. x_{rec} Input image x translated to another domain and retranslated back to its original domain. Ahead of image transformation across domain video-to-video translation is an additional extension.

Figure 2-4 (A) pair shoe dataset sample from Pix2pix, (B) Sunny to Rainy translation from input and output image

Source: Adapted from [5]

2.4. Video-to-video translation

Video-to-video translation is a natural extension of an image-to-image translation. Translating video points toward learning the **appearance of objects in a scene** and **realistic motion movement between successive frames**. A straightforward way to video-to-video translation carry out the image-to-image translation in each frame of input videos without considering those frames has a relation between them. This approach is non-trivial since this is key issues that underlie the flickering [12], [21] effect in the output video. To overcome the flickering effect, Chen et al. [21] consider temporal information along with spatial information. Specifically, they exploit previous frame optical flow to warp the current frame towards imposing temporal constraints. Let see what temporal information and different mechanism to extract is. But before that it worth a brief discussion about the problem in current approaches.

2.5. Problems in Translation Networks

As discussed in previous sections, video-to-video translation is an immediate extension of image translation, so every limitation of image translation is extended correspondingly. Furthermore, Object disappearance, Object dislocates, and Artifacts [22]–[27] are the most common problems for video translation.

Let say we have two generators G_{AB} and $G_{BA}$to translate from one domain to another domain and two discriminators D_A and D_B, where G_{AB} trained to translate from A to B and $G_{AB}$from B to A. and discriminators $D_A$and D_B to classify between real and fake in both domains. Video X and Y are sample videos from respective domain A and B. $X = \{x_1, x_2, x_3 \dots, x_n\}$ where x_i are the i[th] frame of video X. Each frame may contain various objects. $\{O_{x_1}^1, O_{x_1}^2, O_{x_1}^3 \dots O_{x_1}^n\} \epsilon\, x_1$ Objects in a frame can be seen as a group of connected pixels.

- » Object disappearance: is a problem object O_i in a given video frame x_t in domain A shall also appear in translated appear $\widetilde{x_1}$ in another domain image. meaning if a car appears in x_t is should also appear in $\widetilde{x_1}$. Mathematically,
- » $if\ \{O_{x_1}^i\} \epsilon\, x_1, then\ \{O_{\widetilde{x_1}}^i\} \epsilon\, \widetilde{x_1} where: \widetilde{x_t} = G_{AB}(x_t)$
- » Object dislocation[2]: happen when an object O_i in frame x_t from a domain A changes its position when translated in $\widetilde{x_1}$domain B. Object dislocation also can be seen as an abrupt object movement. Mathematically,
- » $if\ \{O_{x_1}^i\} \epsilon\, x_1 \& locate\ [(a1, b1) - (a2: b2)]\ then, \{O_{\widetilde{x_1}}^i\} \epsilon\, \widetilde{x_1}\ should\ locate\ in\ [(a1, b1) - (a2: b2)]$
- » $where: \widetilde{x_t} = G_{AB}(x_t), a\ \&\ b\ are\ spatial\ location\ in\ x\ and\ y\ direction$
- » Artifacts: An image frame artifact is any element that occurs in the picture that is not present in the initial picture set.
- » Tide Spatially to the input: The optimizer is required to learn a solution that is strongly similar to the input due to the reconstruction loss on the input itself.

[2] Object disappearance and object dislocation in a situation like face to face translation wouldn't be a question.

The problems described above are appropriate for the problem of translation, where only spatial transformation is considered.

2.6. Temporal Information

Temporal refers to time-domain, wherein our case, it can be seen as a relation between the sequence of frames in the video, while Spatial refers to RGB space frames. Spatiotemporal or Spatio-temporal is used in the study of information as data is gathered over time and space. Straight forward approaches Generally fail because they cannot consider both domains. Temporal information for video can describe a phenomenon in a particular pixel location with position change in time. For a video-to-video translation, we have various options to represent motion information. The next section would discuss those topics. For illustration, the extraction of time knowledge can be split into two separate groups. One explicit temporal information extraction: this kind of network operates in such a way that the model extracts temporal information directly, such as optical flow and pose estimation. Then the model imposes temporal information on the generated frame. The other tacit does

not explicitly collect temporal knowledge but aims to learn temporal dimension through specially modeled learning layers of the model. Examples could be 3D Conv-nets, RNNs, and temporal constraint models. Indeed, some of the works have been done to blend the above techniques, such as Park et al. [28].

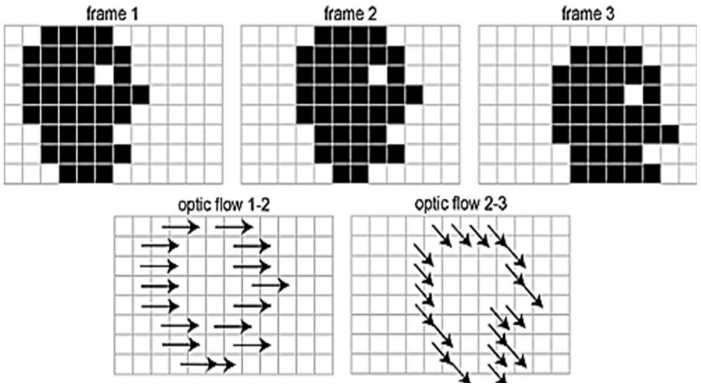


Figure 2-5 Detection of the optical flow in 3 consecutive images.

Source: Adapted from [29]

2.6.1. Optical flow

Optic flow is the change of structured pixels with specific intensity in successive images, or in other words, Optical flow is the motion of objects among successive frames, caused by the relative movement among the object and camera. This make optical flow an ideal for encoding temporal information[1], [28], [30].

Figure 2-5 shows three sequence images, and in the next row shows the Optic flow between the modification in these images over a vector field. The research underlines the precise, pixel-wise estimation of optic flow, which is a computationally challenging task.

Nowadays, better computational resources and Recent advancements in Deep learning enable researchers to estimate optical flow. Generally, such approaches take two video frames as input to output the optical flow (color-coded image), which may be expressed as $(u, v) = f(x_{t-1}, x_t)$ where u is the motion in the x direction, v is the motion in the y direction, and f is a neural network that takes in two consecutive frames x_{t-1} (frame at time $= t - 1$) and x_t (frame at time $= t$) as input.

Computing optical flow with deep neural networks [31], [32] requires vast amounts of training data which is principally hard to obtain. This is because marking optical flow video footage requires a detailed finding of the precise motion of each point of a frame to the precision of the subpixels. To address the issue of labeling training data, many research works, [31], [33], [34] used computer graphics to simulate massive realistic worlds. Since virtual worlds are produced by complex computer instruction, the motion of each and every point of an image in a video sequence is known. Some examples of such include MPI-Sintel [35], an open-source CGI movie with optical flow labeling rendered for numerous sequences, and Flying Chairs [33], a dataset of numerous chairs hovering around random backgrounds both generated from the virtual world using Computer Graphics.

2.6.2. Pose Estimation

Human pose estimation can be framed as the problem on the localization of key points like eye, nose, elbows, wrists, etc. in images or videos frequently referred to as human joints. It is also known as the exploration of the overt position of all articulated poses in space. Basically, pose estimation translate used in transferring motion from a deriving video to derived object in a video. Mainly human pose estimation is used in transferring motion from

one person to another as used [36][37], to transfer motion between different domain videos specifically for animating static image by driving motion as shown in figure 2-6 [38] and facial expression transfer [39] between source and the target person.

We have two types of pose estimation classical and deep learning; the former is all about represents an object by a group of "parts" organized in a deformable configuration, and Later, ConvNets was commonly embraced as their core building block. They largely replace hand-crafted features & graphical models perhaps this approach has returned drastic advances on standard benchmarks.

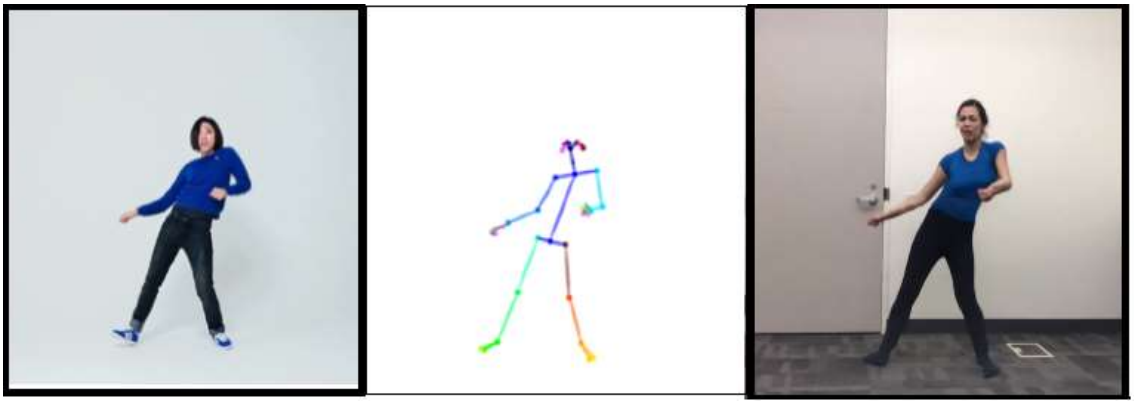


Figure 2-6 pose extraction to transfer pose.
Source: Adapted from[36]

2.6.3. 3D convolutional tensor

The 3D convolutional tensor mechanism is one of the orthodox methods, which basically does not consider temporal information explicitly. Since it considers presenting video scene [22] as a 3-dimensional tensor meaning it takes the whole video as input, and the network eventually learns the relation between consecutive frames to preserve temporal consistency implicitly. In due course, this approach is not used frequently because of two fundamental reasons requires a high-efficiency machine, and the network becomes an entirely black box, which means hard to tune parameters basically done in training Deep Learning models.

2.6.4. Recurrent temporal

Recurrent neural networks or RNNs are a type of neural network inherently ideal for analyzing data from time-series, and other sequential figures make it ideal for video analysis. Possibly it overcomes the black-box nature of 3D Conv nets by adding an additional

parameter to tune the network. Recent works consider using LSTM (Long Short Time Memory.) which takes into account all previous frames as an input to minimize temporal residual error [30].

2.7. Related Works

In the previous section, we discussed the temporal information (motion information) extraction mechanism. However, since video consists of both temporal and spatial information, we need to discuss mechanisms to get the advantage over an early approach (spatial only). Hence instead of applying Spatio information only (meaning split a video as a sequence of images and apply for domain transfer on each, then stitch them back) by assuming frame constraint has no relation[6]. This approach is non-trivial since the key issues which motivate the problem listed in section 2.6. the results output video[12], [21].

Another work, Unsupervised Video-to-Video Translation model temporal information using a 3D convolutional layer embedded on Cycle-GAN, but the model black-box nature makes it hard to train. Nevertheless, the result shows a lack of a robust nest to model the temporal information.

To overcome the flickering effect, Chen et al. [21] consider temporal information along with spatial information. Specifically, they exploit previous frame optical flow to warp the current frame to impose temporal constraints, but this paradigm prone to occlusion and fast illumination change (since optical flow does not consider newly introduced pixels in given frame scene).

Video-to-Video Translation with Global Temporal Consistency [30] by further extend the optical flow frame warping network, Wei et al. present a mechanism focusing on the video-level consistency by residual error based on discriminator to minimize the total L1 distance between the optical flow map of consecutive frames eventually this approach failed on longer video and result failed in fast motion videos since the network constrained to minimize temporal difference along the whole video.

Another fine work by Chen et al. [40] MoCycle-GAN introduces temporal motion translation to transfer estimated motion from source to target video while preserving temporal consistency. This work also relives the temporal cycle constraint for motion

reconstruction. Even if explicit motion translation network is a blessing, model parameter increased enormously.

The current state of artwork [1] ReCycle-GAN further extends cycle consistency constraint by intercorporate it with temporal predictor network to predict over spatiotemporal predictor though directly synthesize future frame via temporal predictor to preserve temporal continuity.

Another recent quality work by [28] proposes an optical flow warping ground truth and content loss on frame mechanism to guarantee the consistency to overcome the temporal flickering and motion inconsistency between frames temporal flow consistency is another addition of this work, which basically excellent if the two domains are similar in nature, but has no much impact on slight different motion videos such as on flower dataset.

2.8. Related work summary

the following table illustrates a summary of previous works on the video-to-video translation[3].

Table 3 previous works summary on the video-to-video translation.

Related papers	***Dataset (Data collection)***	***Architecture***	***Temporal information modeling.***	***Temporal constraint applied***	***Evaluation matrix used***	***Limitation***
Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks[6]	Cityscapes, Horse to Zebra, Apple to Orange, Summer to Winter Yosemite.	Cycle Consistency Constraint.	No temporal Information is considered.	-	FCN, IS	Framewise image-to-image translation.

[3] These present papers are only substantial papers that directly relate to thesis work, and all are from 2017 onwards.

Unsupervised Video-to-Video Translation [22]	Volumetric MNIST, GTA segment to video and MRI-to-CT	3D Cycle-GAN	The network implicitly learns to form input video (3D-Conv- net)	-	Human evaluation, pixel accuracy, and L2 error between original and retranslated image	3D tensor fails for temporal learning consistency between frames.
Video-to-Video Translation with Global Temporal Consistency[30]	DAVIS 2017	RNN based Cycle-GAN, and RNN based Discriminator for global temporal consistency	Flownet2.0, temporal residual error minimizer	Generator + Discriminator Network	Peak Signal to Noise Ratio, Region Similarity, and Contour Accuracy	Complex architecture hard to train. Inappropriate for in videos contain fast object motion. Not work for long videos.
MoCycle-GAN: Unpaired Video-to-Video Translation MoCycle-GAN [40]	Flower video and viper dataset	Cycle-GAN with motion translator-based motion cycle consistency	Flownet2.0 with motion translator network	Generator Network	Human evaluation, IoU, pixel accuracy, Average class accuracy	Explicit motion translator

Recycle-GAN: Unsupervised Video Retargeting: ReCycle-GAN [1]	Viper, face, and flower datasets (more than 10,000 images)	Cycle-GAN with recurrent temporal predictor	Recurrent temporal predictor(pix2pix)	Generator Network	Human evaluation, IoU, pixel accuracy, Average class accuracy, IS	Temporal predictor basically fails to correctly predict, and no cycle consistency temporal cycle considered
Preserving Semantic and Temporal Consistency for Unpaired Video-to-Video Translation [28]	Viper dataset	Cycle-GAN with flow estimator network and consistency warping network	Flownet2.0 base temporal fuse with spatial for improving occlusions problem	Generator Network, Use [41] to further reduce the Temporal warping error.	mIoU, fwIoU, and pixel accuracy	Input domain videos shall have very similar content.

As shown in the above table, researchers design complex architectures in previous work to learn a mapping from domain to domain transfer in an unsupervised manner. However, those works do not consider the content difference between a real and translated video, which is a significant problem that causes object dislocation and disappearance. they does not consider the power of the temporal aware discriminator network. This thesis work was done to add the loss for content preserving and modify learning losses to the baseline network to improve the generated video quality even further.

CHAPTER THREE

3. MATERIALS AND METHODS

3.1. Overview

The thesis research questions were outlined in Chapter one, along with a mathematical formulation and an overview of the method used to investigate the associated plans. This chapter provides further details of the methodology, dataset, and experimental metrics to answer the research questions.

The following approaches and procedures are used to accomplish the goals of this study.

3.2. Dataset

This study uses a machine learning approach to solve video-to-video translation problems in an unsupervised manner, so data is an essential part of the study. Images of a face (Obama-trump), Viper and, flowers are used for both training and testing stages as used in [1]. In addition to inference the result of this work, I collect a local dataset called **አዲስ** (Adiss).

- ***Obama-trump***: is a recently released dataset for style transfer and video retargeting. This dataset contains a sequence of images of Obama and Trump making an interview (though at a different time and completely talk about different things). Each frame is 256 x 256 and about 8617 images are included.
- ***Flower Video Dataset***: is a recently released dataset for video translation. This dataset contains the time-lapse videos which depict the flourishing or fading of several flowers but lacking any sync. The resolution of the respective videos is 256 × 256—this work use flower-to-flower for domain transfer between dissimilar types of flowers.
- ***Viper***: is a prevalent visual perception benchmark to facilitate both low-level and high-level vision tasks -semantic segmentation and optical flow. It comprises videos from a realistic virtual world game (i.e. GTA V), which are composed while driving, riding, and walking in various ambient circumstances (day, sunset, snow, rain, and night). Each frame (resolution: 1920 × 1080) is annotated with pix-level labels, for video-to-labels and labels-to-video, viper could be a benchmark for evaluating the

translations between videos and segmentation label maps, and day ↔ sunset. For this study, the frame resolution is Demote to (resolution: 256 × 256).

» ***አዲስ (Adiss)***: is a local dataset for video retargeting containing two very popular politicians Prime minister Abiy Ahmed and Debretsion Gebremicahel making an interview press conference. Frame size is 256 x 256, and around 5000 images are included.

Table 4 Training Dataset Sample (from Obama - Trump and Flower Datasets)

Flower1	
Flower2	
Obama	
Trump	
Abiy	
Debretsion	

Table 5 Viper Dataset Sample Examples

Rain	***Snow***	***Sunset***	***Day***
			

3.3. Development tools

For this research, numerous types of development tools are used to design and implement the proposed thesis work. The development tools section gives a description and justification of these development tools. These tools include prototype development tools and platforms, UML Modeling tools, and other relevant tools to the research. The following sections give a brief detail about these development tools.

3.4. Design tools

Design tools are mediums that are used for the creation, presentation, and interpretation of design concepts. Edraw Max [42] is used to design in the proposed system. It is a lightweight and powerful graphic design tool for creating professional-looking flowcharts, network diagrams, flowchart diagrams, and others. This tool is selected because [42].

- » It has lots of high-quality shapes, example, and template,
- » Easily visualizes complicated details through a broad range of graphics.
- » It works well with MS Office.

3.5. Prototype development framework

3.5.1. TensorFlow

TensorFlow is an open-source software library optimized for maximum-performance numerical modelling and processing. Its modular architecture can be easily implemented on a range of platforms such as Central Processing Units (CPUs), Graphical Processing Units (GPUs), Tensor Processing Units (TPUs). It can also be mounted on personal computers, clusters, handheld devices, and edge devices. Tensorflow Supports artificial learning, deep learning, and versatile numerical computing [43] The following diagram displays the power score of the deep learning system based on application, popularity, and interest [44].

The following diagram demonstrates the power score of the deep learning system based on application and popularity. As shown in the below diagram, TensorFlow is by far the most used and popular deep learning framework.

- Makes fast and rapid prototyping;
- Embraces all Convolution networks and recurrent networks, as well as variations of each.
- User-friendly, modular, and extensible.
- It can run efficiently on GPU or CPU.

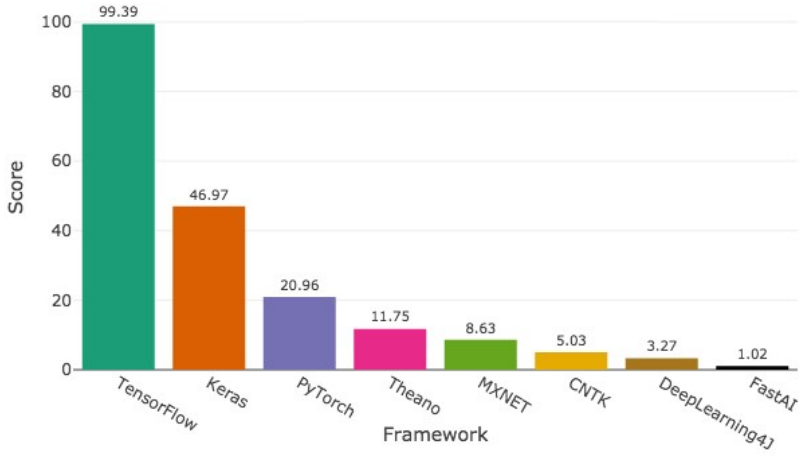


Figure 3-1 Deep Learning Framework comparison.

Source: Adapted from [45]

3.5.2. OpenCV

OpenCV is an open-source computer vision software library intended to provide a shared infrastructure for image processing and computer vision applications [46]. It has Python, Java, C++, and MATLAB interfaces and supports nearly any operating system as well. OpenCV was developed for image processing, meaning that and feature and data structure was developed with the image processing engineers in mind.

3.5.3. MATLAB Deep Network Designer

MATLAB deep network designer [47] is an application developed by MATLAB which developed for easy Design, Visualize, and train deep learning networks using drag & drop simple user interactive mechanism. This tool is a relief for AI developers, especially for complex network deep architectures and GAN networks. This even further helps Developers to track and debug errors on the early design stage.

3.6. Baseline Works

To validate our model's effectiveness, we equate it with models that dwell on translating video with GANs. Since our model architecture is based on Recycle-GAN and takes as input unpaired video data, we chose Cycle-GAN [8] and Recycle-GAN [11] as the baselines for our experiments.

» Cycle-GAN [8] converts images using two generators, with the assumption of cycle consistency. This work uses it to translate the video frames and make comparisons in order to understand the Spatio-temporal constraint effect better.
» ReCycle-GAN [11] uses two generators and two predictors for video translation. It puts forward a recycle loss to work with cycle loss and recurrent loss for content conversion and style preservation, taking into account the temporal detail.

The purpose of contrasts Cycle-GAN and ReCycle-GAN is to show the substantial improvements achieved by our model in terms of spatial-temporal knowledge.

3.7. Feature Extraction Network

Several states of art Deep CNN based architectures have been proposed over the last decades for the classification of images. These different state of art CNN based feature extraction architecture include ResNet[48], Inception[49], Xception [50], EfficientNet[51], and others.

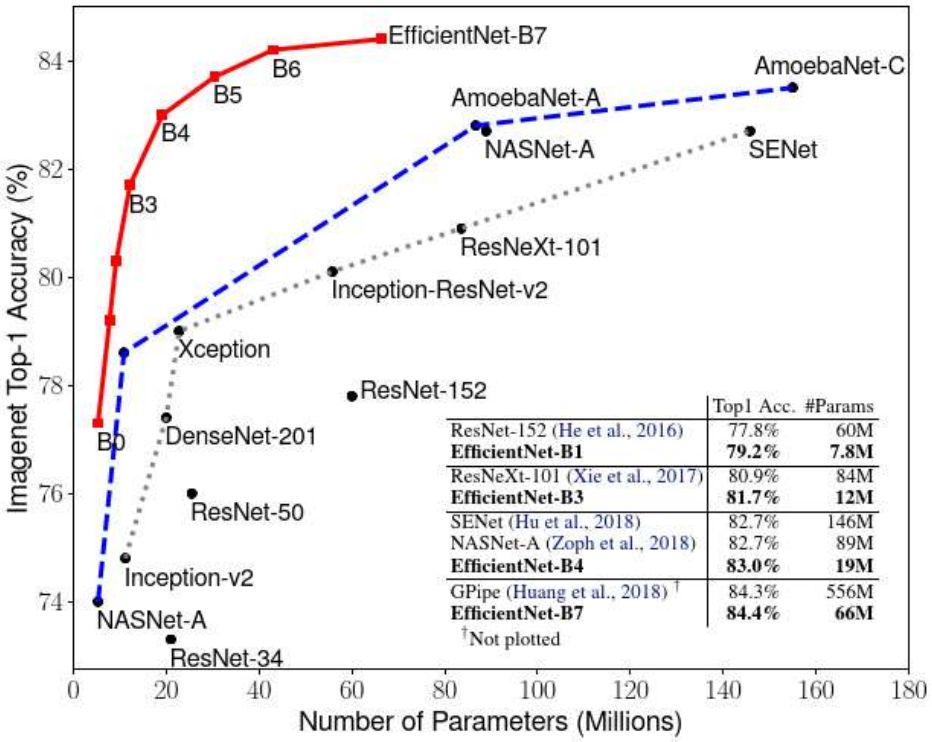


Figure 3-2 Benchmark Analysis on EfficientNet-B7

Source: Adapted from[51]

The above figure shows the top1 accuracy vs the number of parameters. The x-axis represents the number of trainable parameters. EfficientNet-B7. another architectures like VGG have more than 155 parameters million and ResNet has round 60 million. EfficientNet-B7 has less number of parameters and operations compared to most architectures, which makes it able to run fast on different devices that have less computing. Perhaps, Other architectures like ResNet-50 have a fewer number of parameters and operations as shown in the figure above, but their accuracy fails as well. Based on the above observation, EfficientNet-B7[51] is used in this research for feature extraction in the task of computing feature map of a given image.

3.8. Temporal Discriminator Network

Temporal information is a motion relation between the sequence of frames as discussed in chapter two. Recent works do not consider the power of discriminator networks regarding temporal information. An ablation study has been conducted to decide the number of frames the discriminator should consider for classifying between fake and real. Detail results are found in Appendix F.

The study performed on temporal discriminator with one frame, with two frames, with three frames, with four frames, and with five frames but the training fails with five frames because of memory inefficiency problem. From the result, Temporal Discriminator with three input frames performs well than comparative models. So in this study temporal discriminator with five frames has been used.

3.9. Evaluation Methods

The result will be analyzed to describe the video-to-video translation model's performance on a test data set. The dataset is split into different training and testing set using different test sizes. The algorithm is evaluated using the test set. One big problem with GANs is that there is no robust way to beyond visual inspection. The next subsection present is a qualitative analysis and a quantitative metric to evaluate this work.

3.9.1. Human Evaluation Study

This evaluation method uses 15 volunteers to assess whether the given video is real or fake after he/she sees real and fake videos to determine whether or not the generated data is any good. The average score value is evaluated as per the figure of entities. Motivated by the ReCycle-GAN Human evaluation study, this thesis work uses two protocols. First, the input video, Synthesized videos of other approaches, and this work result are seen simultaneously for the participants. T**hey are asked which one has higher consistency, better smoothness, and better continuity between video frame sequences**. Second, only Synthesized fake videos are seen simultaneously for the participants, and they are asked which one has higher consistency and **looks more natural Translation**. The higher the Human evaluation score means, the better the performance of the network.

3.9.2. Inception Score

The inception score is a commonly used evaluating algorithm for GANs. It uses a pre-trained inception V3 network (trained on ImageNet) to extract the features of both generated and real images. The inception score [52] for short IS, **measures the variety and the quality of the created images**. The superiority of the model is good if it has a high inception score. Further detail and limitations of IS are found in Appendix A.

$$D_{KL}(P\|Q) = \sum_{i} P(i)\, log \frac{P(i)}{Q(i)} \qquad \text{eq.(3.1)}$$

CHAPTER FOUR

4. PROPOSED LOSS FUNCTION.

4.1. Overview

This chapter presents the proposed solution to video-to-video translation problems for improving temporal consistency. The Generated video should be able to have better consistency between a succession of frames. This chapter can be ideally portioned into three major sections; the first introduces model Architecture to translate a given domain image into another domain—the second deals with Network optimizing loss functions. The last explains training Pseudocode to train our model.

4.2. Model Architecture.

The model architecture has directly Adopted from the architecture defined in *"Unpaired Image-to Image Translation using Cycle-Consistent Adversarial Networks"* [6] and *"Recycle-GAN: Unsupervised Video Retargeting"* [1] for Learning Domain Translation.

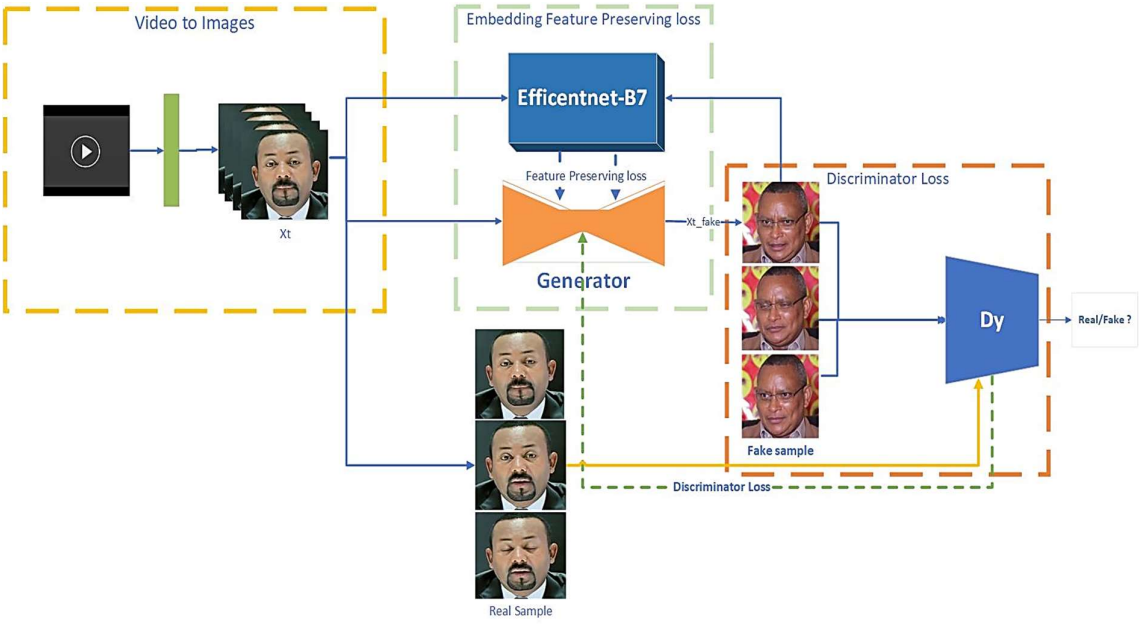


Figure 4-1 Model Architecture.

As shown in figure 4-1, Adjustments have been made to the discriminator network, and additional losses have been applied to the Generative Network; section 5.4 addressed the depth Implementation detail of the model architecture.

4.3. Model Learning Functions

The key objective of this thesis work is to optimize the use of space-time knowledge. In order to address our research query, this work adds loss functions, and change the discriminator network so that it can address temporal coherency to the Cycle-GAN and ReCycle-GAN. As the network model is based on Cycle-GAN and Recycle-GAN.

As discussed on the problem statement, the network seeks to transform a series in time domain images from the source domain, $X = \{x_1, x_2, x_3 \dots, x_n\}$, to a sequence of domain changed images, $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \tilde{x}_3 \dots, \tilde{x}_n\}$, With the exclusion of problems listed in section 2.5. The function is then to acquire the mapping of $G_{AB} : \mathrm{X} \rightarrow \mathrm{Y}$. Note that our model uses sequential unpaired image data as input during training.

4.3.1. Proposed Network Learning Function.

Because this work follow the GAN architecture, the vanilla adversarial loss is also used in this thesis work, called $\boldsymbol{\ell}_{\mathrm{GAN}}$. And the cycle consistency loss $\boldsymbol{\ell}_{cycle}$ in Cycle-GAN [6] is adopted. Besides, the recurrent loss $\boldsymbol{\ell}_{recurrent}$ and the recycle loss $\ell_{recycle}$ in Recycle-GAN [1] are also leveraged. Meanwhile, this work introduces constrain $\boldsymbol{\ell}_{\boldsymbol{featurePreserving}}$ to impel the model and improve the whole translation. Further more the GAN loss changed accordingley to the temporal discriminatro used. The full loss function of this work is as follows:

$$\min_{G,P}\max_{D} \boldsymbol{\ell} all(G,P,D) = \ell_{GAN}(G_x, \boldsymbol{D_x^{t:3}}) + \ell_{GAN}(G_y, \boldsymbol{D_y^{t:3}}) + \alpha\ell_{cycle}(G_x, G_y) + \alpha\ell_{cycle}(G_y, G_x) + \beta\ell_{recurrent}(P_x) + \beta\ell_{recurrent}(P_y) + \gamma\ell_{recycle}(G_x, G_y, P_x) + \gamma\ell_{recycle}(G_x, G_y, P_y) + \theta\boldsymbol{\ell}_{\boldsymbol{featurePreserving}}(\boldsymbol{G_x}, \boldsymbol{G_y}) + \theta\boldsymbol{\ell}_{\boldsymbol{featurePreserving}}(\boldsymbol{G_x}, \boldsymbol{G_y})$$

Where α, β, γ and ɵ are used parameter of learning. Indeed, the network needs more learning constraints, the aim of which is to demonstrate a significant consistency. Let's look in detail at all loss constraints.

Keeping in mind that the translated image should preserve **contain information** but perhaps not the **style.** It should be close to the real image in another domain. The translator network should consider this constraint while learning in the training process.

Cycle Loss

Only unpaired samples are used independently in the respective videos during learning, without the need for paired input results. To fix this, the consistency of cycle continuity is necessary and leveraged by our process, which can be written as[6]:

$$L_{cyc}(G_{AB}, G_{BA}) = E_{x\sim pdata}(x)\ [\|G_{BA}(G_{AB}(x) - x)\|_1] + E_{y\sim pdata}(y)\ \left[\left\|G_{AB}(G_{BA}(y)) - y\right\|_1\right] \quad \textit{eq.(4.1)}$$

Cycle consistency is a loss function asks a question to answer “the original image and the twice-translated (reconstructed image.) image are the same”? If this fails, we may not have a coherent mapping A-B-A. Meaning the original image A and the retranslated image A2B2A mean square distance should be minimum.

Identity Loss

Perhaps the most straightforward loss, Identity loss ensures that the network retains the overall color structure of the image. So, adding a regularization concept lets us keep the tint of the photo in line with the original shot. Imagine that as a way to guarantee that the network can still recreate the original image even after adding several filters[6].

Identity loss is introduced to diminish translation of the images already in domain A to the Generator from G_{BA}, because the Cycle-GAN should understand that they are already in the correct domain. Meaning translating Amharic text to Amharic using English to Amharic translator; since the input is Amharic, the network should make no change.

$$L_{identity}(G_{AB}, G_{BA}) = E_{x\sim pdata}(x)\ [\|G_{AB}(x) - x\|_1] + E_{y\sim pdata}(y)\ [\|G_{BA}(y) - y\|_1] \quad \textit{eq.(4.2)}$$

So, the full loss would be: CycleGAN $= GAN\ loss + cycle\ loss + Identity\ Loss$

$$L(G_{AB}, G_{BA}, D_x, D_y)$$

$$= l_{GAN}(G_{AB}, D_y, X, Y) + l_{GAN}(G_{BA}, Dx, Y, X) + \alpha l_{cyc}(G_{AB}, G_{BA}) + \beta l_{identity}(G_{AB}, G_{BA})$$

where: $G_{AB}\ and\ G_{BA}$ are generators, D_x,and $D_y$are discriminators, respectively both domain X and Y are samples from both domain datasets.

The cycle-loss and identity-loss were extended to various temporal domains. However, these works consider only the spatial information in 2D images and completely disregard the temporal information for modeling, which is also extended by video translation.

Feature Preserving Loss

Indeed, classic cycle-consistency does not essentially assurance the transformation to be semantically consistent. This is, as a result, it does not consider any semantic correspondence during the translation, and thus the system can accomplish textbook cycle-consistency (i.e., L_{cyc}= 0) only if the inverse mapping recovers the original contents, regardless of how incorrect the forward mapping was. This assumption causes object disappearance problem hence, the network doesn't consider content translation to another domain.

$$L_{Fpreserving}(G_{AB}, G_{BA}) = E_{x\sim pdata}(x)\ \left[\left\|mNET(G_{AB}(x)) - mNET(x)\right\|_1\right] + E_{y\sim pdata}(y)\ \left[\left\|mNET(G_{BA}(y)) - mNET(y)\right\|_1\right] \quad \textit{eq.(4.3)}$$

where: *mNET stand for pre-trained EfficientNet-B7*

By adding the above loss, we inspire the network to minimize the **Object Disappearance** problem list in section 2.5 to have consistent semantics earlier and afterward the translation. This thesis work uses EfficientNet-B7[51] as a feature extractor that enforces the content

information that appears in the original image also should appear on translate. For example, if a person and a dog appear in image A so does in translated image A2B, albeit the style modified. (i.e. EfficientNet-B7 current state of classification algorithm tested on Image-net Dataset)

Recurrent Loss

To handle video data, the temporal ordering of the sequential frames must be taken advantage of. In Recycle-GAN [1], we adopt a recurrent temporal P_x predictor to predict frames in the future based on the past frame details. The repeated deficit is as follows:

$$\min_{P_x} l_{recurrent}(Px) = \sum_t \|x_{t+1} - Px(x_{t-1}^t)\|_1 \qquad \textit{eq.(4.4)}$$

Where $Px(x_{t-1}^t)$ is a prediction of Px given x_{t-1} and x_t as concatenated input.

Recycle Loss

Merging image generator [6] and temporal prediction network. The recycle loss[1] across domains and time can be described as:

$$\min_{P_x G_y P_y,} l_{recycle}(G_x, G_y, P_x) = \sum_t \left\|x_{t+1} - G_x(P_y(G_y(x_{t-1}^t)))\right\|_1 \qquad \textit{eq.(4.5)}$$

4.4. Temporal aware Discriminator

To improve visual quality further, a discriminator takes three consecutive generated images to decide whether it is real or fake. The Discriminator architecture and the output stay the same with Patch GAN [5]; instead, the differences are just the input and the number of channels. Rather than differentiating between single frames, the discriminator network is designed in a way to observes three constitutive of synthesized frames and three constitutive of the real frames. This approach makes the discriminator network optimal because it takes into account the temporal nature of the video generation issue, such as **Object Dislocation**.

$$L_{adv} = \min_{G} \max_{D} E_{x \epsilon X}(\log(\boldsymbol{D}(\boldsymbol{x_t}, \boldsymbol{x_{t-1}}, \boldsymbol{x_{t-2}})) + E_{x \epsilon Z}(\log(1 - \boldsymbol{D}(\boldsymbol{G}(\boldsymbol{z_t}), \boldsymbol{G}(\boldsymbol{z_{t-1}}), \boldsymbol{G}(\boldsymbol{z_{t-2}}))))) \quad \textit{eq.(4.6)}$$

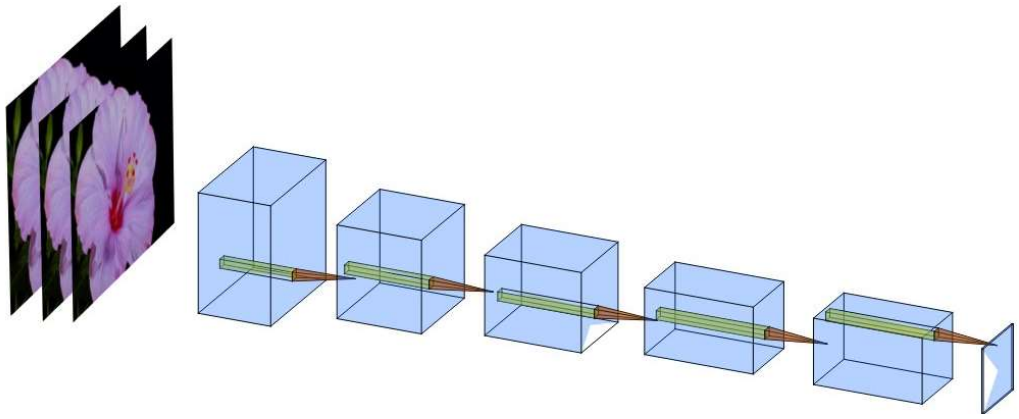


Figure 4-2 Temporal Discriminator Network

4.5. Training Pseudocode

Training algorithms for this thesis work have been introduced in this section as this study compares earlier research, Cycle-GAN, and ReCycle-GAN. Their training algorithms have present in Appendix D and E. As discussed in the preceding section, the previous approach does not consider content translation, which leads to Object Dislocation and Object Disappearance problem. However, this work emphasizes minimizing the content difference between fake and real images, as shown in training pseudocode **line-12.** On the other hand, this work modified the patch-GAN discriminator to make it a temporal aware network, **line-4,** and **line-13,** which care about the relation among consecutive three frames.

Table 6 Training pseudocode CycleGAN with feature preserving loss and temporal discriminator

	This thesis work Training pseudocode:
1	$Take\ a\ sample\ mini-batch: x, x_{t-1}, x_{t-2}, y, y_{t-1}, y_{t-2}$
2	Train D:
3	$Translate\ A, B: \tilde{x} = G_{AB}(x),\ \tilde{y} = G_{BA}(y)$
4	$Compute: \boldsymbol{D_A(x, x_{t-1}, x_{t-2}), D_B(y, y_{t-1}, y_{t-2}), D_A(y, \widetilde{y_{t-1}, y_{t-2}}), D_B(x, \widetilde{x_{t-1}, x_{t-2}})}\ then,$
5	$dloss = \frac{1}{4} * \sum((D_A(x), D_B(y)), (D_A(\tilde{x}), D_B(\tilde{y})))$
6	$update\ \Theta^{(dloss)}\ to\ minimize\ classification\ loss.$
7	Train G:
8	$Compute: \tilde{\tilde{x}} = G_{AB}(\tilde{x}),\ \tilde{\tilde{y}} = G_{AB}(\tilde{x}), xI = G_{BA}(x), yI = G_{AB}(y),$ $\widetilde{\widetilde{x\_pred}} = G_{BA}(Py([G_{AB}(x_{t-1}), G_{AB}(x_{t-2}))),$ $\widetilde{\widetilde{y\_pred}} = G_{AB}(Px([G_{BA}(y_{t-1}), G_{BA}(y_{t-2})))$
9	$cycle\_loss = \frac{1}{2}(mae((x - \tilde{\tilde{x}}), (y - \tilde{\tilde{y}})))$
10	$Recycle\_loss = \frac{1}{2}(mae((x - \widetilde{\widetilde{x\_pred}}), (y - \widetilde{\widetilde{y\_pred}})))$
11	$identity\_loss = \frac{1}{2}(mae((x - xI), (y - yI)))$
12	$\boldsymbol{feature\_preserving\_loss = \frac{1}{2}(mae((mNet(x) - mNet(\tilde{x})), (mNet(y) - mNet(\tilde{y}))))}$
13	$\boldsymbol{D_A(\tilde{y}), D_B(\tilde{x}): d\_loss = \frac{1}{2}\sum(DB(x, \widetilde{x_{t-1}, x_{t-2}}), DA(y, \widetilde{y_{t-1}, y_{t-2}}))}$
	$update\ \Theta^{(\text{d\_loss,cycle\_loss,identity\_loss,}Recycle\_loss, feature\_preserving\_loss)}\ min\ loss.$

Bold indicate the added constraints.

CHAPTER FIVE

5. IMPLEMENTATION OF THE PROPOSED WORK

5.1. Overview

In this chapter, the implementation of the proposed solution is described. The working environment, cycle-GAN implementation, and experimental class conducted are discussed.

5.2. Working Environment.

This section explains the hardware stack that we used to implement our experiments in addition to describing the hardware stack.

- » Laptop: The Laptop computer is used for developing a Network Architecture.
  - o Operating system: Windows 10
  - o Processor: Intel ® Core™ i7-2300QM CPU @ 2.00GHz
  - o Graphics: Intel ® Graphics 3000
  - o Primary Memory (RAM): 8.00 GB
  - o System Type: 64-bit Operating System, x64-based Processor
- » Desktop: The desktop computer is used for developing a video for video translation.
  - o Operating system: windows 10
  - o Processor: Intel ® Core™ i5-4580 CPU @ 3.29GHz x 4
  - o Graphics: Intel ® HD Graphics 4600
  - o GPU: GeForce RTX 2070 Super 6 GB RAM
  - o Primary Memory (RAM): 14.00 GB
  - o System Type: 64-bit Operating System, x64-based Processor

Visual Studio Code and Jupiter notebook are used as a development IDE, with python interpreter 3.6 on a laptop computer. For implementing the proposed domain transfer problem OpenCV 3.7. Furthermore, TensorFlow-GPU 2.2 used. In the next section list of experiments class conducted for evaluating the proposed hypothesis are discussed.

5.3. Environmental Setup

In this thesis work, different software and IDEs have been used.

Anaconda: in an application used to install the up-to-date version of python with its different module and IDEs, for implementing the proposed solution an anaconda application version 1.9.7 with 64-bit support used.

Jupyter Notebook: is the most popular and handy IDE among AI and deep learning researchers to work with python. This thesis work uses Jupiter note-book 6.0.0.

5.4. Implement Cycle-GAN

A Cycle-GAN is made up overall of two GAN architectures: a generator and a discriminator. The generator architecture contains two separate models, Generator AB and Generator BA. In the same manner, the discriminator architecture contains an additional two architectural models, Discriminator A, and Discriminator B. table 7 shows convolutional layers of Cycle-GAN architecture.

The generator network is an encoder-decoder category network. It takes an image as an input with the shape (256,256,3), and outputs generated image with the same shape. Based on base work Cycle GAN two generator networks are defined. The Generators had consisted of 15 layers. Four convolution layers followed by nine residual blocks and two deconvolutional layers - deconvolution means transposed 2-D convolution. The LeakyReLU activation was on all layers except the last layers output layer in the same manner Instance normalization was used in every layer beside the last one.

```
G_A2B = module.ResnetGenerator(input_shape=(256,256, 3))
G_B2A = module.ResnetGenerator(input_shape=(256,256, 3))
```

The discriminator network is equivalent to the discriminator's architecture in a Patch GAN network[5]. Basically, it takes an image of the shape of (256, 256, 3) and predicts whether the image is real or fake. This network composed of 5 convolutional layers denotes a 4 × 4 filter Convolution-Instance Normalization with LeakyReLU layer and stride 2. After the last layer, apply a convolution to produce a 1-dimensional output. The slope of leaky in leakyReLU was 0.2.

```
D_A = module.ConvDiscriminator(input_shape=(256,256, 3))
D_B = module.ConvDiscriminator(input_shape=(256,256, 3))
```

Table 7 Cycle GAN architecture convolutional layers

Layer	***Generators***
1	Convolutional-(Filters-32, Kernel size-7, Stride-1), Instance normalization, LeakyReLU
2	Convolutional-(Filters-64, Kernel size-3, Stride-2), Instance normalization, LeakyReLU
3	Convolutional-(Filters-128, Kernel size-3, Stride-2), Instance normalization, LeakyReLU
4-12	Residual block-(Filters-128, Kernel size-3, Stride-1), Instance normalization, LeakyReLU
13	Convolutional-(Filters-64, Kernel size-3, Stride-0.5), Fractionally strided, Instance normalization, LeakyReLU
14	Convolutional-(Filters-32, Kernel size-3, Stride-0.5), Fractionally strided, Instance normalization, LeakyReLU
15	Convolutional-(Filters-3, Kernel size-7, Stride-1), Instance normalization, Tanh
Layer	***Discriminators***
1	Convolutional-(Filters-64, Kernel size-4, Stride-2),LeakyReLU with slope 0.2
2	Convolutional-(Filters-128, Kernel size-4, Stride-2), Instance normalization, LeakyReLU with slope 0.2
3	Convolutional-(Filters-256, Kernel size-4, Stride-2), Instance normalization, LeakyReLU with slope 0.2
4	Convolutional-(Filters-512, Kernel size-4, Stride-2), Instance normalization, LeakyReLU with slope 0.2

The Generator's objective is to diminish the adversarial loss function against an adversary Discriminator, which constantly tries to maximize it. Similar to other network types of GAN is no different. The learning function has to be explicitly defined in order for the network to learn to translate the image.

```
self.combined = tf.keras.Model(inputs=[img_A, img_B], outputs=[valid_A,
valid_B,reconstr_A, reconstr_B,img_A_id, img_B_id,img_A_id, img_B_id])

#define loss function
d_loss_fn, g_loss_fn =
gan.get_adversarial_losses_fn(adversarial_loss_mode)
cycle_loss_fn = tf.losses.MeanAbsoluteError()
identity_loss_fn = tf.losses.MeanAbsoluteError()
G_loss = (A2B_g_loss + B2A_g_loss) + (A2B2A_cycle_loss +
B2A2B_cycle_loss) * cycle_loss_weight + (A2A_id_loss + B2B_id_loss) *
identity_loss_weight
```

Almost all of the configurations were taken from the Cycle-GAN paper and the implementations of its authors on GitHub.

5.5. Temporal Predictor Network Implementation

This thesis work uses Recycle-GAN temporal predictor network Px and Py for video retargeting, which is identical to the pix2pix [5] generator network. However, the input layer has been modified to receive two successive previous images.

```
inputs1 = tf.keras.layers.Input(shape=[256,256,3])
inputs2 = tf.keras.layers.Input(shape=[256,256,3])
#(bs, 256, 256, channels*2)
inputs = tf.keras.layers.concatenate([inputs1, inputs2])
```

The temporal predictor network predicts the next frame based on two previous frames taken as input. Like every neural network, the temporal predictor network is similar and has been explicitly defined in the network.

As shown in the code snip, *train_p* function takes six argument variables. A, A_1, and A_2 are in domain A and the rest in domain B. Since pix2pix needs paired dataset A_1 and A_2 concatenated as an input, the network predicts A_p, which is predicted frame-based given inputs. The loss is the L1 distance between A_P and A, which is used to update the gradient weights.

```
#input temporal predictor network
from temporal_predictor import Generator
```

```
Px = Generator(inputs)
Py = Generator(inputs)

P_optimizer = keras.optimizers.Adam(learning_rate = 2e-4, beta_1 = 0.5)
#A_1, A_2, B_1 and B_2 are the previous two frames in Domain A and Domain
B
@tf.function
def train_P(A, A_1, A_2, B, B_1, B_2):
    with tf.GradientTape() as pt:
        A_p = Px([A_1, A_2],training = True)
        B_p = Py([B_1, B_2],training = True)
        xl1_loss = P_loss_fn(A, A_p)
        Px_loss = xl1_loss * LAMBDA
        yl1_loss = P_loss_fn(B, B_p)
        Py_loss = yl1_loss * LAMBDA
        P_loss = (Px_loss + Py_loss)* args.cycle_loss_weight
#update gradient weight
   P_grad = pt.gradient(P_loss, Px.trainable_variables
+Py.trainable_variables)
   P_optimizer.apply_gradients(zip(P_grad, Px.trainable_variables +
Py.trainable_variables))
    return A_p, B_p, {'Px_loss': Px_loss,  'Py_loss': Py_loss}
```

5.6. Feature Preserving Loss Implementation

As discussed in the previous sections, feature preserving loss aims to minimize content information deference between real and the translated fake images. To do so Efficientnet-b7 pre-trained model is imported. Since the aim is to extract the feature map of the input image, the last four layers are removed, as shown in code snip. Using a pre-trained EfficientNetB7 model, the new Keras model is created.

Another function *get_content_feature* is define was then defined to measure the content of two pictures. in order to compute feature map new, which returns a feature map of the input images. Then compute feature preserving loss between the real image and fake image pair sets has been computed then the loss has been used to update network weight. Meaning the content loss would be the L1 distance between *M_B, M_B2A, and M_A, M_A2B*.

```
import efficientnet.tfkeras as eff #import pretrained EfficientNet-B7
#remove the last four layers
```

```
base_model =
eff.EfficientNetB7(input_shape=(256,256,3),include_top=False)
x = base_model.layers[-4].output
mNet = tf.keras.Model(inputs = base_model.input, outputs=x)

def get_content_features(a,b):
      return mNet(a), mNet(b)
M_A, M_A2B = get_content_features(A, A2B)
M_A_A2B = identity_loss_fn(M_A, M_A2B)
M_B, M_B2A = get_content_features(B, B2A)
M_B_B2A = identity_loss_fn(M_B, M_B2A)
```

5.7. Temporal aware Discriminator Network Implementation

This work also uses additional temporal aware discriminator network. As discussed in the presiding section, it takes three images to discriminate whether the images are real or fake.

```
def build_discriminator(n):
      inputA = tf.keras.Input(shape = (256,256,3))
      inputB = tf.keras.Input(shape = (256,256,3))
      inputC = tf.keras.Input(shape = (256,256,3))
      h = tf.keras.layers.Concatenate()([x, y, z])
      d1 = conv2d(h, 64, 4, 2)
      d2 = conv2d(d1, 128, 4, 2)
      d3 = conv2d(d2, 256, 4, 2)
      d4 = conv2d(d3, 512, 4, 2)
      d5 = conv2d(d4, 1, 4, 1)
      x = tf.keras.Model(img,d5,name=n)
      return x
```

The Discriminator architecture and the output stay the same with Patch GAN [5]; instead, the differences are just the concatenated input and the number of channels. This enforces the network to strictly focus on the relation among generated images and its relation with two previous images.

```
      A_d_logits = D_A((A,A_1,A_2), training=True)
      B2A_d_logits = D_A((B2A,B2A_1,B2A_2), training=True)
      B_d_logits = D_B((B,B_1,B_2), training=True)
      A2B_d_logits = D_B((A2B,A2B_1,A2B_2), training=True)
```

5.8. Experiment Class

To evaluate the essence of temporal information for video translation testing the initial hypothesis is mandatory. Five different classes of experiments are conducted, as shown below on the table for each dataset group. The first three experiments focus on video translation on flower and viper datasets, while the rest two are basically for video retargeting on Obama-Trump and (አዲስ) Adiss Datasets. The first class is all about vanilla Cycle-GAN image translation on a given sequence of images, considering the spatial domain only. The second is regarding consider using feature preserving loss. The third one includes temporal Discriminator build up on the second experiment. The fourth experimental class uses vanilla ReCycle-GAN aiming video retargeting, and the last one merges ReCycle-GAN with temporal discriminator which become a total of ten experiments

Table 8 lists of experimental classes.

Notation	***Experiment***	***Training Epochs***	***Dataset***	***Model used***
CC	Baseline CycleGAN	200 epochs, 20 epochs	Flower dataset and Viper dataset	Cycle-GAN
CC+CP	CycleGAN baseline generator trained with additional feature preserving loss	200 epochs, 20 epochs	Flower dataset and Viper dataset	Cycle-GAN and EfficientNet-B7
CC+CP+TD	CycleGAN baseline generator trained with additional feature preserving loss and temporal Discriminator Network	200 epochs, 20 epochs	Flower dataset and Viper dataset	Cycle-GAN, flownet2, and Temporal aware discriminator.
RC	Baseline ReCycle-GAN	30 epochs	Obama trump dataset and Adiss Dataset	ReCycle-GAN
RC+TD	ReCycle-GAN with temporal Discriminator	30 epochs	Obama trump dataset and Adiss Dataset	ReCycle-GAN & temporal Discriminator

CHAPTER SIX

6. EVALUATION, RESULTS, AND DISCUSSION

6.1. Overview

Previous Chapters identified the methodologies that were selected to experimentally investigate the research propositions—this section reports on the outcomes of the Experimental stage. The data collected and information are analyzed concerning the principal research goal posed in this thesis: How to preserve temporal consistency for a video-to-video translation? Moreover, this thesis work proposes a hypothesis that *"adding temporal consistency constrain would improve temporal consistency between successive frames."*.

6.2. Video-to-video Translation

The video-to-video translation takes a video from the scene as an input to generate an equivalent video in other domains with the consideration of preserving temporal information. This work conducts different training experiments to explain the qualitative and quantitative outcomes of comparing the baselines on which the study is based tested on different datasets. This research work uses the inception score (IS) and a Human evaluation study to evaluate the experimental outcome. Using the training algorithms mentioned in the segment. 4.2.

The models compared in the evaluation are shown in Table 7. As discussed in the evaluation metrics, the Human evaluation study follows two protocols. The first asks which video looks more real showing generated videos only which are labeled P1 in the next section. The second questions which one looks more realistic and natural translation by showing real input video and the fake generated videos side by side and the result labeled P2. The next section confers results and discussion on the experiment output found on each dataset.

6.2.1. Flower to Flower

Figure 6-1 demonstrates this method's synthesized frames on the Flower dataset.[4] The videos in this dataset show the blooming of different flowers, which is a relatively slow process, meaning the shifts between adjacent frames are relatively small. Our algorithm can generally preserve the consistency of a sense and content information based on the given input video. The translated flower in each target field retains a continuity for much of the time, with input flower at a different domain. Table 8 shows the inception score of experimental runs of the network.

Table 9 IS score and Human evaluation study Result on flower Dataset

Methods	Flower			
	IS		Average Human evaluation study per domain[5]	
Real dataset	1.165 0.030	1.248±0.055		
	Domain A	Domain B	Domain A	Domain B
CC	1.022±0.002	1.102±0.031	25%	3.1%
CC + CP	1.023±0.009	1.122±0.184	9.4%	0%
CC + CP + TD	**1.138±0.041**	**1.162±0.025**	**65.6%**	**96.9%**

Bold values indicate the best results in the experiments.

Figure 6-2 shows that even if Temporal Continuity can be preserved by model CC+CP+TD, it contains many artifacts because of the vanishing gradient problem. In training, the model does not really have much weight changes after some epochs (meaning the gradient becomes very low-approximate to zero, multiplication of a small number further minimizes the loss) so any gradient update does change almost nothing in backpropagation. As a result, the discriminator network of CC+CP+TD becomes too complicated to be tricked by the generator network. As seen in figure 6-2 below, the discriminator loss becomes slightly similar to zero, and the generator loss will escalate to one. All the generator generated images are known as false or, in other words, the discriminator network quickly bits the generator, which is not what we want. However, we were searching for the Nash equilibrium of the two networks (Generator and Discriminator) to balance each other. On the other hand, even

[4] Each experiment takes around 2.13 day for training.(200 epochs)

[5]Human Evaluation User study for flower translation found at: https://forms.gle/dG3jo9iVskvXxLWLA

though the Inception score of CC+CP outperforms vanilla Cycle-GAN, the Human evaluation study shows CC excel CC+CP. Indeed, The CC+CP network even amplifies the flickering effect.

To mitigate the vanishing gradient and gradient explosion problem, applying gradient penalty to the discriminator network had been applied. the CC+CP+TD improves the generated output video quality, as shown in figure 6-3 below. (however, for a fair comparison gradient penalty result has not included in the Human evaluation study and Inception score report.).

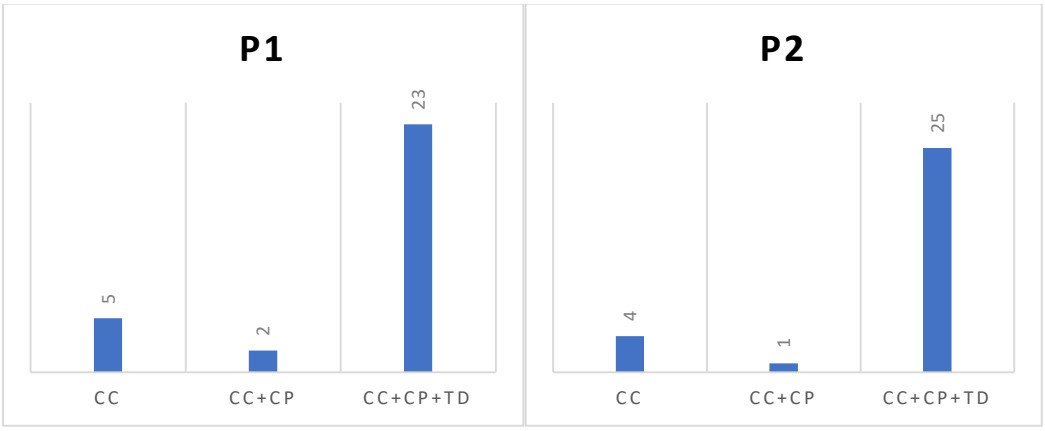


Figure 6-1 Human Evaluation Study on flower dataset

(left) Human evaluation study found after showing fake videos only to participants,
(right) Human evaluation study based on generated videos with the corresponding real video input.
Higher values indicate the best results in the experiments.

From the above observation, this work performs advantages over Cycle-GAN(CC) and Cycle-GAN with Feature preserving loss (*CC+CP*), due to the improvements of video continuity and stability brought by the spatial-temporal constraint.

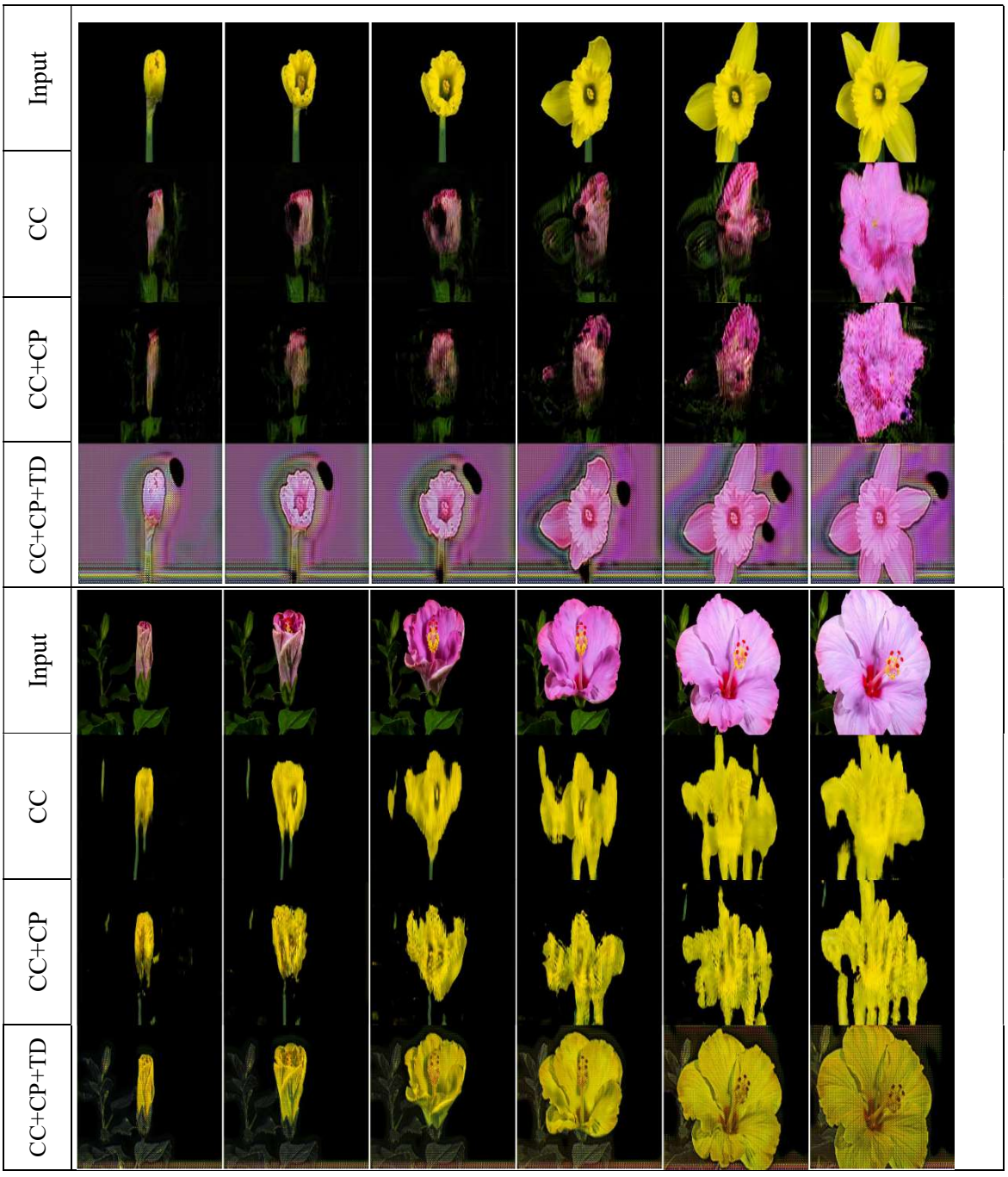


Figure 6-2 flower to flower translation result

The real images labeled Input that the synthetic images are based on. The second-row is the result of Cycle-GAN, Third-row shows Cycle-GAN with Feature preserving loss, the last include Temporal Aware discriminator and Temporal warping.

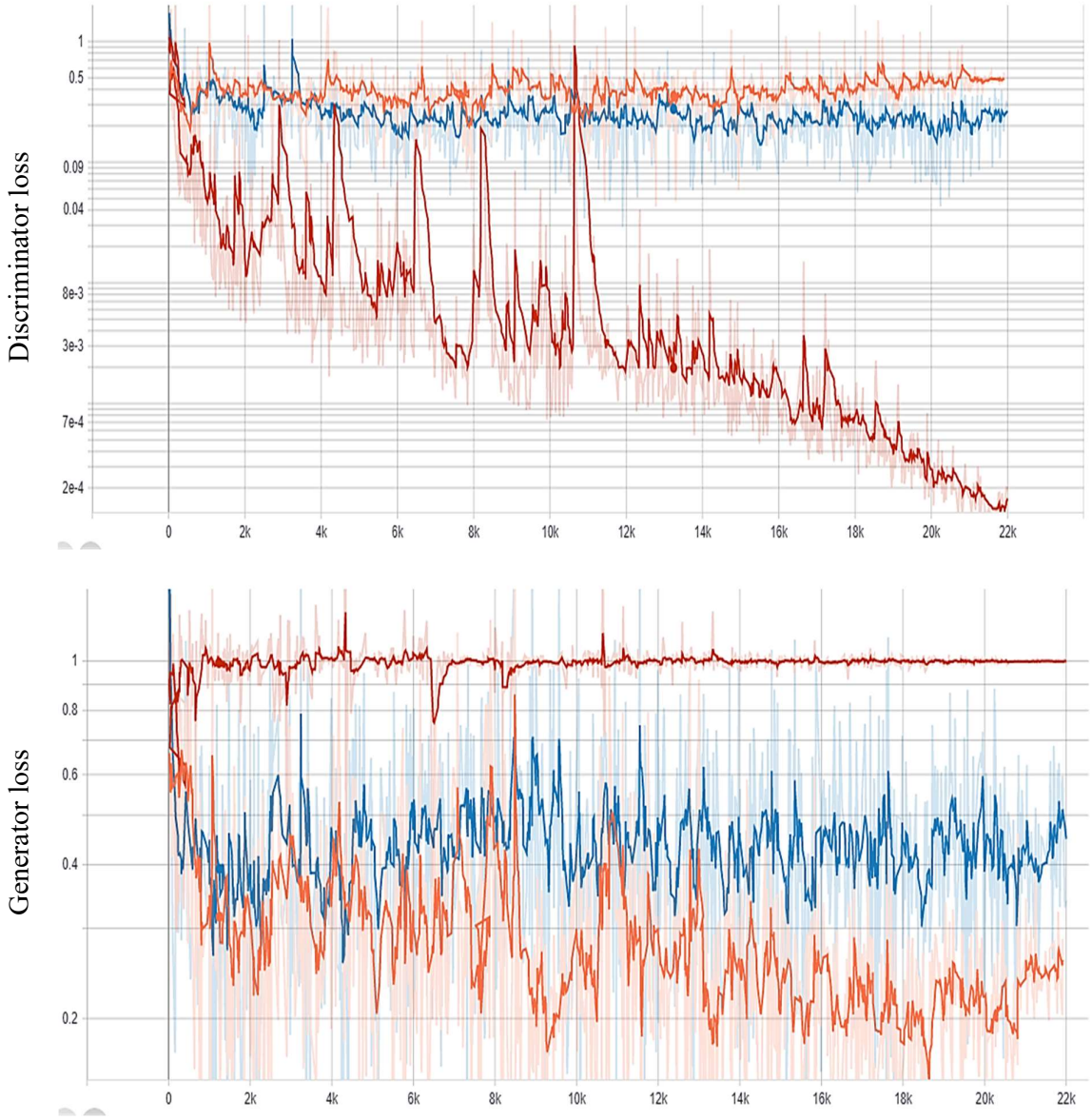


Figure 6-3 weight Vanishing problem on CC+CP+TD

Top Discriminator network, and bottom Generator network,

Red: CC+CP+TD, Blue: CC+CP, Orange: CC. The CC+CP+TD discriminator loss quickly fail to zero result the generator loss to explode to one.

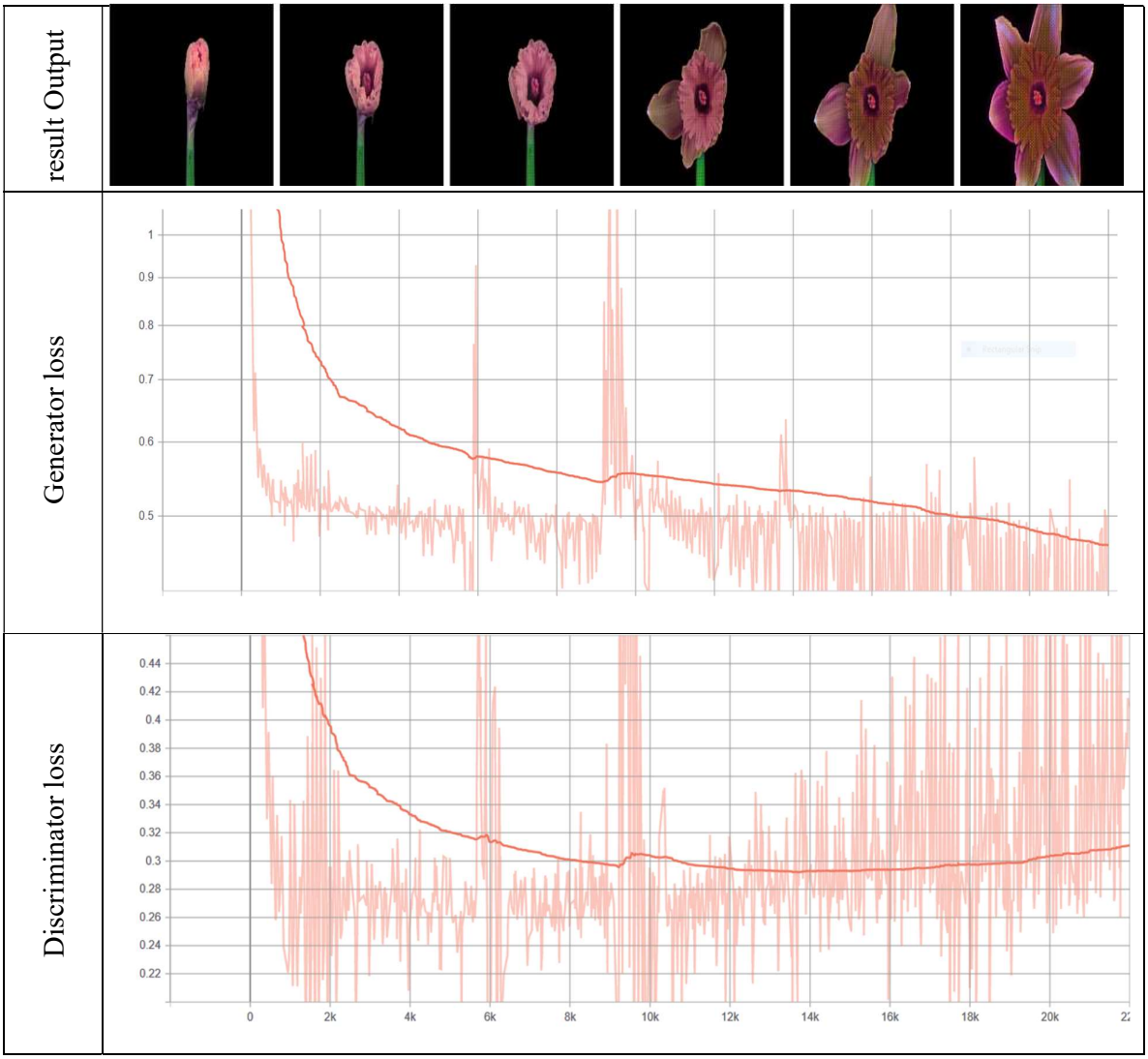


Figure 6-4 CC+CP+TD with gradient penalty

Number of examples (top) Output sample (middle) Generator loss,
and (bottom) Discriminator loss

6.2.1. Sunset-to-Day

Similar to flower translation, this experiment uses the same training setup, except uses a subset of viper dataset target to translate Day time video to Sunset video and vice versa. It takes around 1.16 sec per image and a total of 3.75 days, train for 20 epochs. The task of Sunset to Day is

shown to explain the influence of our exploitation of the proposed solution of this thesis; therefore, more focus on the video quality improvements shown in Figure. 6-4 and figure 6-5,

Figure 6-5 Sunset to Day translation Output Result

Eventually, this work positively improves visual quality, as confirmed in IS and Human evaluation study results (Table 11). This experiment may tell us a great deal about our method because the network can convert complex datasets successfully while compared to the flower dataset.

Table 10 IS score and Human evaluation study Result on Viper Dataset

Methods	Day to Sunset			
	IS		Average Human evaluation study per domain [6]	
Real data	*3.56s±0.21*	*3.81±0.44*		
	Day	Sunset	Day	Sunset
CC	2.50±0.17	2.71±0.19	0%	5%
CC + CP	3.09±0.07	**3.64±0.26**	40%	40%
CC + TD	**3.23±0.13**	3.61±0.11	**60%**	**55%**

Bold values indicate the best results in the experiments.

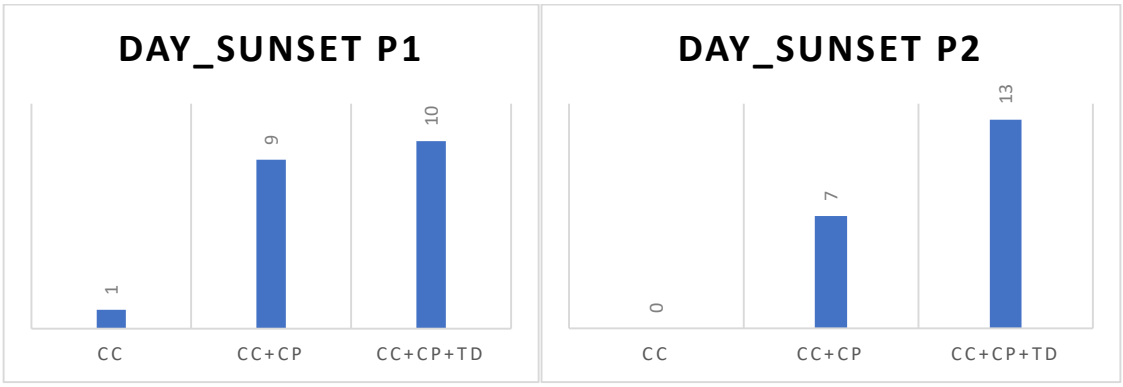


Figure 6-6 Human Evaluation Study on Viper dataset

((left) A human evaluation study found after showing fake videos only to participants, (right) A human evaluation study found after showing fake videos with the corresponding real video input. Higher values indicate the best results in the experiments CC+CP almost catch CC+CP+TD.

As shown in the above result report CC+CP performs much better in Sunset to Day datasets than the flower translation, as shown in table 11 above. Even more CC+CP performs as good as this thesis work as figure 6-6 shows (P1 result show this work exceed CC+CP by one). I suppose it is because the Efficientnet-B7 feature extracting network trains on ImageNet. I did not think there were substantial flower blooming (in fact it only contains 1197 images of flowers, which is around 0.0084% of the entire dataset.) instances in training so the network might not be able to extract adequate features in flower video. Hence, the network performed

[6] Human Evaluation User study for Day to Sunset found at: https://forms.gle/xbkt9aFxA4YmNFnH6

badly due to this cause. CC+CP+TD was impacted by the vanishing of the gradient problem in the tiny dataset as seen in the flower to flower translation, and maybe the viper dataset is very big as the pixelated and the artifacts problem in the flower translation has diminished even better.The Human evaluation study scores tell that a majority of the participants prefer our synthesized videos than those comparative models. 17.5% over Cycle-GAN with feature preserving loss and 55% on Cycle-GAN.

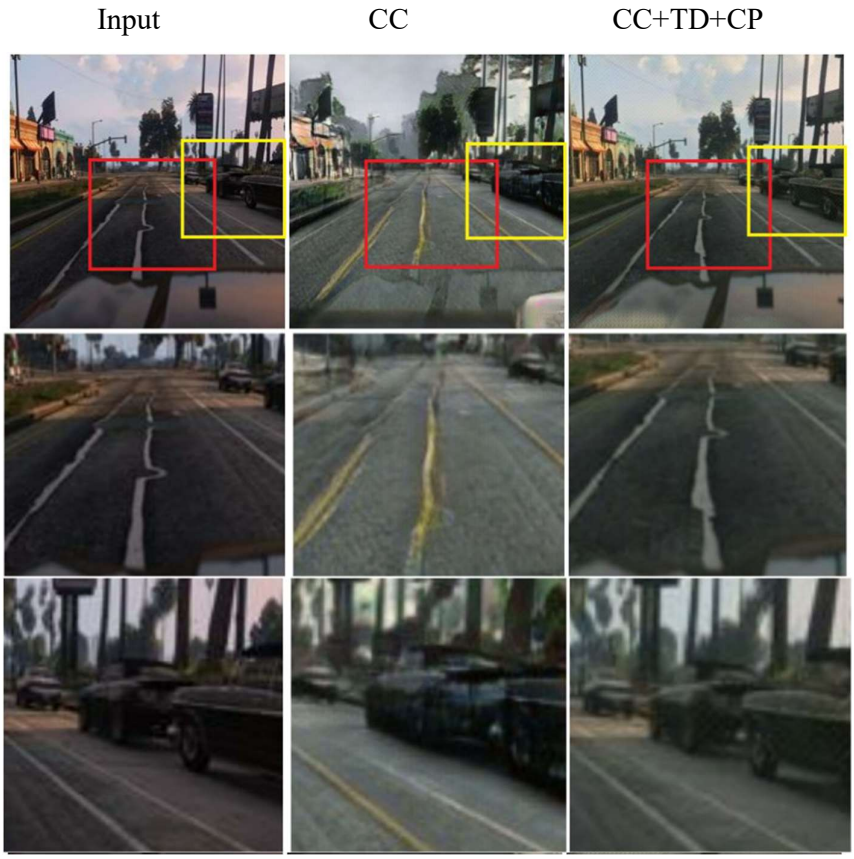


Figure 6-7 Comparison between Cycle-GAN with this thesis work on Sunset to Day

(left) input images, (center) Cycle-GAN, (right) this thesis work, CC+CP+TD can preserve the detailed content and color information than CC.

Figure 6-7 shows that CC does not retain detailed information, but this model generates a decent result positively toward content translation compared with CC. As shown in the above figure, the car shape altered a bit and the color of the road line change as the result of Cycle-GAN. But the proposed model can preserve the color and shape of the input frame better than comparative work.

6.2.2. Face to Face

In this experiment, we evaluate Obama to Trump translation using the Recycle-GAN and ReCycle-GAN with temporal discriminator. Result shows both approaches are capable of accessing the stylistic facial gestures of Donald Trump and Barack Obama[7] ***(Please note that the photos are very minimal in their representation)***. Nevertheless, mouth motion slightly Differ, as shown in the above figure 6-7. For example, in Trump to Obama translation, RC+TD model fetches trump mouth movement more reasonably than the comparative model Recycle GAN. Training takes around 4.12 days for 30 epochs.

The IS result shows this thesis work ReCycle-GAN with Temporal Discriminator (RC+TD) advantages over ReCycle-GAN(RC). (r.b. Content preserving network cannot be implemented since this work focuses on video Retargeting). In comparison with that of the ReCycle-GAN, our network increases the IS with a slight advantage as shown below, and it also outperforms human evaluation study by 40% of base work.

Table 11 Obama to Trump Inception Score and Human evaluation Study.

Methods	Obama to Trump			
	IS		Average Human evaluation study per domain [8]	
Real data	1.283±0.102	1.069±0.274		
	Obama	Trump	Obama	Trump
RC	1.035±0.120	1.048±0.010	40%	20%
RC + TD	**1.041±0.013**	**1.068±0.011**	**60%**	**80%**

Bold values indicate the best results in the experiments.

RC model has been suffered from image flipping problem as shown below in figure 6-10, which truly inherited from the Convolutional neural network. (but for fair comparison in figure 6-8, RC output images have been aligning according to the input.) however, RC+TD could

[7] Please check the website for better comparison: https://sites.google.com/astu.edu.et/kirubelabebe/temporal-cycle-consistency-constraint-for-video-to-video-translation-result

[8] Human Evaluation User study for Day to Sunset found at: https://forms.gle/ydauVZbixebUJVYJA

maintain input video alignment which indicates the temporal discriminator network has better orientation awareness in its model weight.

On the other hand, the temporal discriminator network positively impacts video retargeting based on qualitative and quantitative evaluations discussed above; however, it also forces the model to make the L1 distance between consecutive frames to become very small and similar, doing so limit motion change which sometimes make the result too static and unreal.

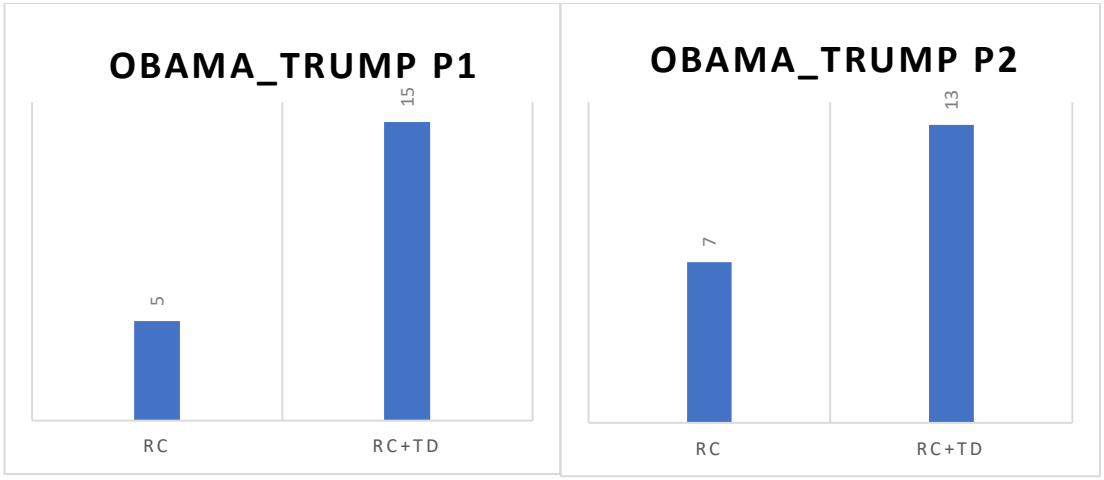


Figure 6-8 Human Evaluation Study on Obama Trump dataset

((left) A human evaluation study found after showing fake videos only to participants, (right) A human evaluation study found after showing fake videos with the corresponding real video input. Higher values indicate the best results in the experiments

Figure 6-9 RC Trump Generated image sequences

Generated video sequence from the RC model result fliped trump out.

Figure 6-10 Obama to Trump Translation Result

Row label as six sequential inputs are the inputs to the network, and the rests are the corresponding output of the network. The top three are Obama to Trump, and the bottom ones are the reverse translation.

6.2.1. Adiss (አዲስ)

This experiment extends face to face to implicate performance of the model on the local dataset which relatively is complex datasets contain full face including hair, eyeglass, colorful

background, and unaligned face direction. As shown in the generated output, both models almost perfectly capture the **head** movement of the input video.

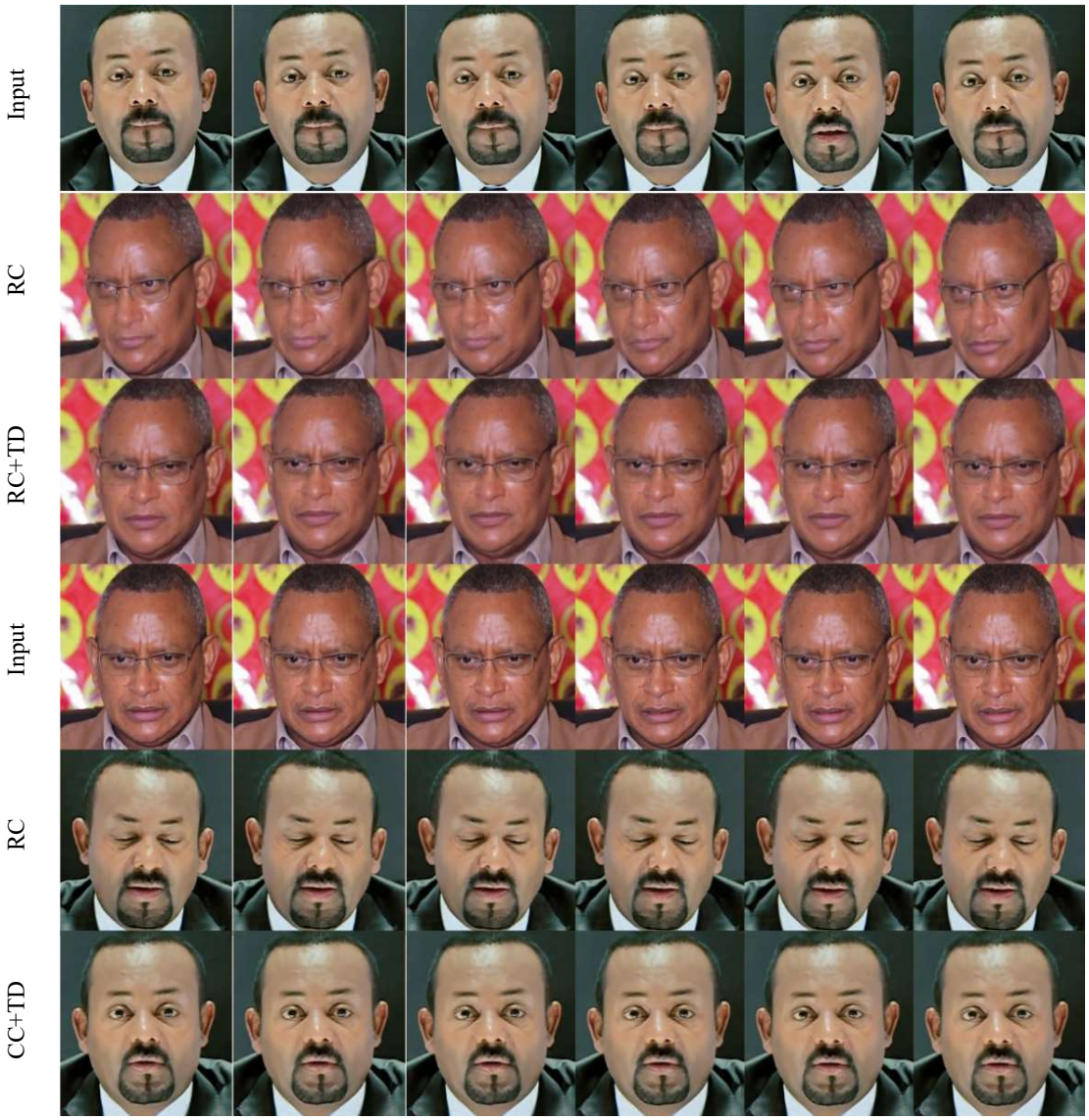


Figure 6-11 Abiy to Debretsion Translation Result.

Row label as six sequential inputs are the inputs to the network, and the rests are the corresponding output of the network. The top three are Abiy to Debretsion, and the bottom ones are the reverse translation.

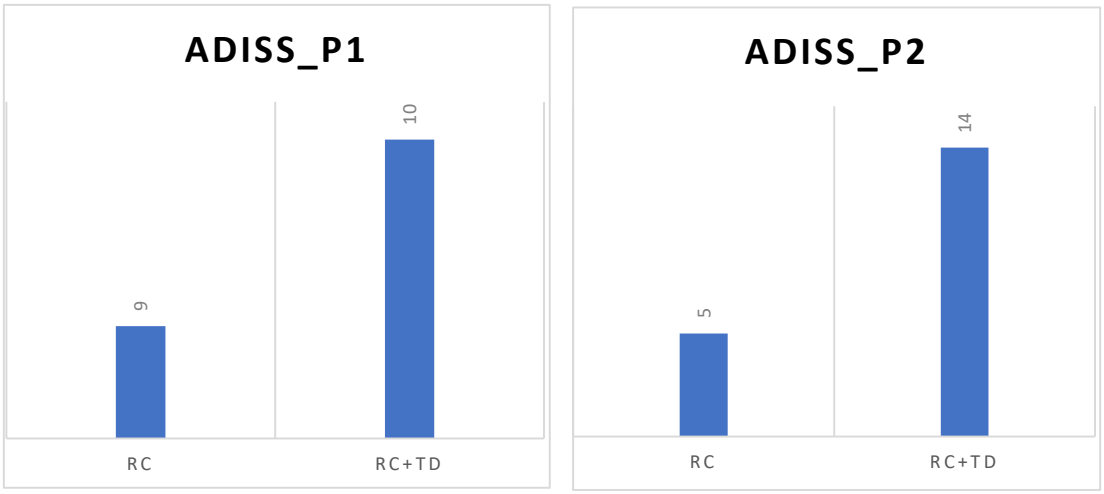
The quantitative result of RC+TD improves the inception score on Debretsion's fake video with a significant amount. on the other side, the thesis work lost by base work with a slight margin fake Abiy video.

The human evaluation study claims, participants ware confused to choose a better one among the result found from RC and RC+TD. This indicates the poorer quality of the Adiss(አዲስ) dataset in several aspects such as it doesn't consider face keypoint alignment and face size normalization and network convergence. Additionally, from the result in figure 6-11 above its clear, the network basically emphases on translating head motion rather than lip movement on the input video.

Table 12 Abiy-to-Debretsion translation result

Methods	Abiy-to-Dibretsion			
	IS		Average Human evaluation study per domain [9]	
Real data	*1.250±0.055*	*1.475±0.047*		
	Debretsion	Abiy	Debretsion	Abiy
RC	1.171±0.054	**1.314±0.033**	35%	35%
RC + TD	**1.218±0.030**	1.307±0.031	**65%**	**65%**

Bold values indicate the best results in the experiments



((left) A human evaluation study found after showing fake videos only to participants, (right) A human evaluation study found after showing fake videos with the corresponding real video input. Higher values indicate the best results in the experiments

[9] Human Evaluation User study for Day to Sunset found at: https://forms.gle/u2qi7DwBnyJCeW2d9

Even if both networks typically produces an unsatisfactory result the user study indicates this thesis model has a slight edge in both P1, P2 score. Furthermore, the RC+TD Average Human evaluation study per domain surpasses ReCycle-GAN by 30%.

6.3. Video-Translation Summary

Based on Quantitative and Qualitative results, on Flower-to-Flower translation and Sunset-to-Day, the simplest model *CC*, which only trained only on the Cycle-loss, has the lowest score among the models, indicating that the (with respect) cycle GAN architecture is not complex enough for the video-to-video translation. Since CC only considers the spatial domain, the translation lacks knowledge of the temporal domain. The CC+CP model shows a non-significant improvement in the flower dataset but performs relatively well on Viper Dataset. CC+CP+TD outperforms the baseline work Cycle-GAN by almost 61% and 49.95% on Cycle-GAN with feature preserving loss.

6.4. Video-Retargeting Summary

Based on ReCycle-GAN RC model Generated result, it lacks shape and orientaino knowlage as shown on Obama-to-trump translation in which case the model flip trump face horizontally. Inception Score on Rc and RC+TD indicate optimistic progress of the model applied, which illustrate the robustness of this thesis work in order to learn better spatial-temporal information from the video and to produce better content consistency. In reality, both models are capable of learning the style of the input video and speaker action, but the outcomes are far from flawless. Evidently, ReCycle-GAN with temporal discriminator output in the Human Evaluation analysis reveals an average 30 percent increase in the model performance.

CHAPTER SEVEN

7. CONCLUSION AND FUTURE WORK

7.1. Conclusion

Video-to-video translation is a natural extension of an image-to-image translation. Translating video points toward learning objects' appearance **in a scene** and **realistic motion movement between successive frames**. A straightforward way to video-to-video translation carry out the image-to-image translation in each frame of input videos without considering those frames that have a relation between them. This approach is non-trivial since the underlying flickering problem effect is in the output video.

The purpose of this study was to improve temporal coherence for the video-to-video translation by adding constraints to the GAN network learning function trained on the unpaired dataset, which starts on the ReCycle-GAN claim. Among the investigation, the goal was to generate as visually realistic video as possible. To do so, this thesis adds Feature preserving loss, and Temporal aware discriminator to the baseline works. Indeed, these changes make our model very aware of the perpetual spatial-temporal information changes in the video.

Different from early approaches, which focus only on the Generated image, look real or fake based on Spatial information only. Our approach enforces the discriminator network to emphasize not only on the spatial domain to judge real or fake but also check temporal coherency between the Generated image and its preceding two frames. Object Disappearing appears to be another issue in recent works, so this thesis introduces a loss-preserving constraint to minimize the distance between the extracted Efficientnet-B7 features on the generated fake image and the original input. Our model Combines the above two losses to preserve temporal information.

In fact, this work has been impacted by the Efficientnet-B7 model weight which makes it dependent, furthermore, it impacted by dataset type and size such as seen on the flower to flower translation. The model becomes unstable and collapses by vanishing gradient problems. Even more, it produces artifacts in the generated fake images. However, Applying

a gradient penalty to the discriminator network improves the vanishing gradient problem, and a better result is generated.

Compared with baseline works Cycle-GAN[6], and ReCycle-GAN[1], qualitative and quantitative experimental findings indicate. The Cycle-GAN model, trained only on Cycle Loss, has the lowest evaluation score among this thesis work. Hence, it suggests that Cycle-GAN architecture is not complex enough for the video-to-video translation. Since it considers the spatial domain only, the translation lacks information on the temporal domain. Another version Cycle-GAN with content preserving loss also was very dependent on the feature extraction model used.

In the case of video retargeting experiments, the RC model can capture the style and content of a video as well as our model. However, lousy alignment result has been seen in the generated trump video result as well in the fake Abiy video, in which the eye was closed in the entire period. RC+TD shows better shape aware retaining tendency from the experiments, which helps to better quality generated videos.

Experiments show more significant variation among the result, both qualitatively and quantitatively. This thesis's achievement is that it excels in the human assessment analysis by xxx percent and xxx percent in the IS score of Cycle-GAN. This Research work concludes that Adding constraints to video-to-video translation does improve temporal coherency.

7.2. Limitation and Future work

The thesis method does not come without limitations. As observed in the experiment, the model is strongly dependent on the EfficientNET-B7 outputs, and since the feature preserving loss is not designed to be consistent across frames. Generated output depends essentially on the feature extraction network's performance on a specific training dataset, as discussed on 6.2.1. This naturally leads to inconsistency in the results produced. One approach to resolving this issue is a retune feature extraction network on the training dataset.

Furthermore, rather than a simple concatenation of output images in a manner to make temporal ware, the discriminator network could be modeled in a much efficient approach using the LSTM network so further research could be work to extend in a better approach.

Although the number of participants in human evaluation is a very small range from 10 to 15 persons, further evaluation of human study would improve for better judgment.

Finally, this thesis's work might take us in a very different direction: letting it learn using the discriminator network with considering a substantially long sequence of frames, focus on how to construct synthetic intermediate frames between successive frames, could increase the video's frame rate. HFR or (High frame rate) videos will increase the movement representation and consequently provide better pictures to increase the audience's accuracy. Perhaps it could need considerably more than that of two frames as used in this thesis work.

REFERENCES

[1] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, “Recycle-GAN: Unsupervised Video Retargeting,” *CoRR*, vol. abs/1808.0, 2018, [Online]. Available: http://arxiv.org/abs/1808.05174.

[2] “What Happens Now That An AI-Generated Painting Sold For $432,500?” https://www.forbes.com/sites/williamfalcon/2018/10/25/what-happens-now-that-an-ai-generated-painting-sold-for-432500/#f7702aca41ca (accessed Dec. 12, 2019).

[3] “Attempts on Real Time Style Transfer – mc.ai.” https://mc.ai/attempts-on-real-time-style-transfer/ (accessed Sep. 22, 2020).

[4] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, “Temporal cycle-consistency learning,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, no. 12, pp. 1801–1810, Apr. 2019, doi: 10.1109/CVPR.2019.00190.

[5] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, vol. 2017-Janua, pp. 5967–5976, doi: 10.1109/CVPR.2017.632.

[6] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, Dec. 2017, vol. 2017-Octob, pp. 2242–2251, doi: 10.1109/ICCV.2017.244.

[7] C. Cao, Q. Hou, and K. Zhou, “Displaced dynamic expression regression for real-time facial tracking and animation,” in *ACM Transactions on Graphics*, 2014, vol. 33, no. 4, doi: 10.1145/2601097.2601204.

[8] M. Mirza and S. Osindero, “Conditional Generative Adversarial Nets,” Nov. 2014, Accessed: Oct. 10, 2019. [Online]. Available: http://arxiv.org/abs/1411.1784.

[9] I. Goodfellow *et al.*, “Generative Adversarial Nets (NIPS version),” *Adv. Neural Inf.*

Process. Syst. 27, 2014, doi: 10.1001/jamainternmed.2016.8245.

[10] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2019, vol. 2019-June, pp. 4396–4405, doi: 10.1109/CVPR.2019.00453.

[11] T. C. Wang, M. Y. Liu, J. Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 8798–8807, doi: 10.1109/CVPR.2018.00917.

[12] H. Huang *et al.*, “Real-time neural style transfer for videos,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, vol. 2017-Janua, pp. 7044–7052, doi: 10.1109/CVPR.2017.745.

[13] L. Hui, X. Li, J. Chen, H. He, and J. Yang, “Unsupervised Multi-Domain Image Translation with Domain-Specific Encoders/Decoders,” in *Proceedings - International Conference on Pattern Recognition*, Nov. 2018, vol. 2018-August, pp. 2044–2049, doi: 10.1109/ICPR.2018.8545169.

[14] Jake VanderPlas, *Python Data Science Handbook*. O’Reilly Media, Inc.

[15] R. Rojas, “Neural Networks: A Systematic Introduction. ,” *Springer New York, NY, USA - Verlag New York, Inc.*, 1996.

[16] G. E. H. Alex Krizhevsky, Ilya Sutskever, “ImageNet Classification with Deep Convolutional Neural Networks,” *ILSVRC2012*, pp. 1–1432, 2007, doi: 10.1201/9781420010749.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998, doi: 10.1109/5.726791.

[18] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep

convolutional generative adversarial networks," Nov. 2016.

[19] "Image-to-Image Translation: Machine Learning Magic that Converts Winter Photos Into Summer - Abto Software, Lviv, Ukraine." https://www.abtosoftware.com/blog/image-to-image-translation (accessed Mar. 03, 2020).

[20] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 8789–8797, Nov. 2018, doi: 10.1109/CVPR.2018.00916.

[21] D. Chen, J. Liao, L. Yuan, N. Yu, and G. Hua, "Coherent Online Video Style Transfer," in *Proceedings of the IEEE International Conference on Computer Vision*, Dec. 2017, vol. 2017-Octob, pp. 1114–1123, doi: 10.1109/ICCV.2017.126.

[22] D. Bashkirova, B. Usman, and K. Saenko, "Unsupervised Video-to-Video Translation," no. Nips, 2018, [Online]. Available: http://arxiv.org/abs/1806.03698.

[23] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic Speech-Driven Facial Animation with GANs," *Int. J. Comput. Vis.*, 2019, doi: 10.1007/s11263-019-01251-8.

[24] A. R. Kosiorek, H. Kim, I. Posner, and Y. W. Teh, "Sequential attend, infer, repeat: Generative modelling of moving objects," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. NeurIPS, pp. 8606–8616, 2018.

[25] S. Tulyakov, M.-Y. Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing Motion and Content for Video Generation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 1526–1535, doi: 10.1109/CVPR.2018.00165.

[26] B. Kratzwald, Z. Huang, D. P. Paudel, A. Dinesh, and L. Van Gool, "Improving Video Generation for Multi-functional Applications," 2017, [Online]. Available: http://arxiv.org/abs/1711.11453.

[27] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixé, and N. Thuerey, “Learning Temporal Coherence via Self-Supervision for GAN-based Video Generation,” Nov. 2018, Accessed: Jan. 25, 2020. [Online]. Available: http://arxiv.org/abs/1811.09393.

[28] K. Park, S. Woo, D. Kim, D. Cho, and I. S. Kweon, “Preserving semantic and temporal consistency for unpaired video-to-video translation,” *MM 2019 - Proc. 27th ACM Int. Conf. Multimed.*, pp. 1248–1257, Aug. 2019, doi: 10.1145/3343031.3350864.

[29] C. Militello, L. Rundo, and M. C. Gilardi, “Applications of imaging processing to MRgFUS treatment for fibroids: a review,” *Transl. Cancer Res.*, vol. 3, no. 5, pp. 472–482, 2014, doi: 10.21037/3200.

[30] X. Wei, S. Feng, J. Zhu, and H. Su, “Video-to-video translation with global temporal consistency,” *MM 2018 - Proc. 2018 ACM Multimed. Conf.*, pp. 18–25, 2018, doi: 10.1145/3240508.3240708.

[31] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “FlowNet 2.0: Evolution of optical flow estimation with deep networks,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, vol. 2017-Janua, pp. 1647–1655, doi: 10.1109/CVPR.2017.179.

[32] D. Sun, X. Yang, M. Y. Liu, and J. Kautz, “PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 8934–8943, doi: 10.1109/CVPR.2018.00931.

[33] A. Dosovitskiy *et al.*, “FlowNet: Learning optical flow with convolutional networks,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 2758–2766, 2015, doi: 10.1109/ICCV.2015.316.

[34] D. Sun, X. Yang, M. Y. Liu, and J. Kautz, “PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 8934–8943, doi: 10.1109/CVPR.2018.00931.

[35] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, “A naturalistic open source movie for optical flow evaluation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7577 LNCS, no. PART 6, pp. 611–625, doi: 10.1007/978-3-642-33783-3_44.

[36] J. P. Bennett, “Everybody Dance Now!,” *J. Phys. Educ. Recreat. Danc.*, vol. 77, no. 1, pp. 6–7, Jan. 2019, doi: 10.1080/07303084.2006.10597803.

[37] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, “GANimation: One-Shot Anatomically Consistent Facial Animation,” *Int. J. Comput. Vis.*, no. January, 2019, doi: 10.1007/s11263-019-01210-3.

[38] S. Webber, M. Harrop, J. Downs, T. Cox, N. Wouters, and A. Vande Moere, “Everybody Dance Now: Tensions between participation and performance in interactive public installations,” *OzCHI 2015 Being Hum. - Conf. Proc.*, pp. 284–288, 2015, doi: 10.1145/2838739.2838801.

[39] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “Animating Arbitrary Objects via Deep Motion Transfer,” 2018. Accessed: Oct. 08, 2019. [Online]. Available: http://arxiv.org/abs/1812.08861.

[40] Y. Chen, Y. Pan, T. Yao, X. Tian, and T. Mei, “Mocycle-GAN: Unpaired video-to-video translation,” *MM 2019 - Proc. 27th ACM Int. Conf. Multimed.*, pp. 647–655, Aug. 2019, doi: 10.1145/3343031.3350937.

[41] W. S. Lai, J. Bin Huang, O. Wang, E. Shechtman, E. Yumer, and M. H. Yang, “Learning blind video temporal consistency,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11219 LNCS, pp. 179–195, doi: 10.1007/978-3-030-01267-0_11.

[42] “Edraw Max - Excellent Flowchart Software & Diagramming Tool.” https://www.edrawsoft.com/edraw-max/ (accessed Jun. 02, 2020).

[43] “TensorFlow.” https://www.tensorflow.org/ (accessed Jun. 02, 2020).

[44] J. Hale, “Deep Learning Framework Power Scores,” 2018. .

[45] “AI deep learning frameworks ranking 2018 | Statista.” https://www.statista.com/statistics/943038/ai-deep-learning-frameworks-ranking/ (accessed Sep. 22, 2020).

[46] “OpenCV.” https://opencv.org/ (accessed Jun. 02, 2020).

[47] “Design, visualize, and train deep learning networks - MATLAB.” https://www.mathworks.com/help/deeplearning/ref/deepnetworkdesigner-app.html (accessed Jun. 05, 2020).

[48] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[49] C. Szegedy *et al.*, “Going deeper with convolutions.”

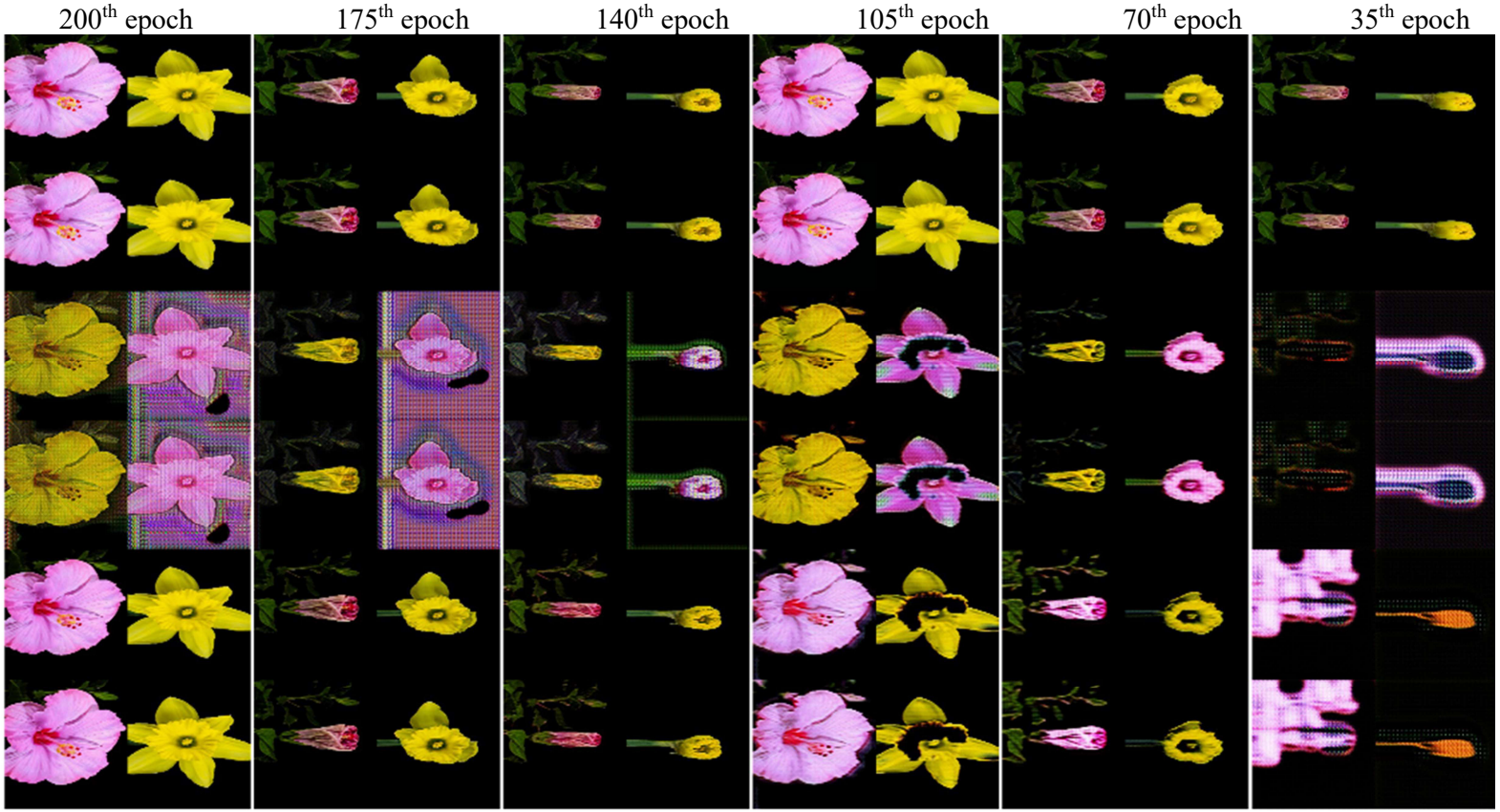
[50] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, vol. 2017-January, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.

[51] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, May 2019, Accessed: Aug. 29, 2020. [Online]. Available: http://arxiv.org/abs/1905.11946.

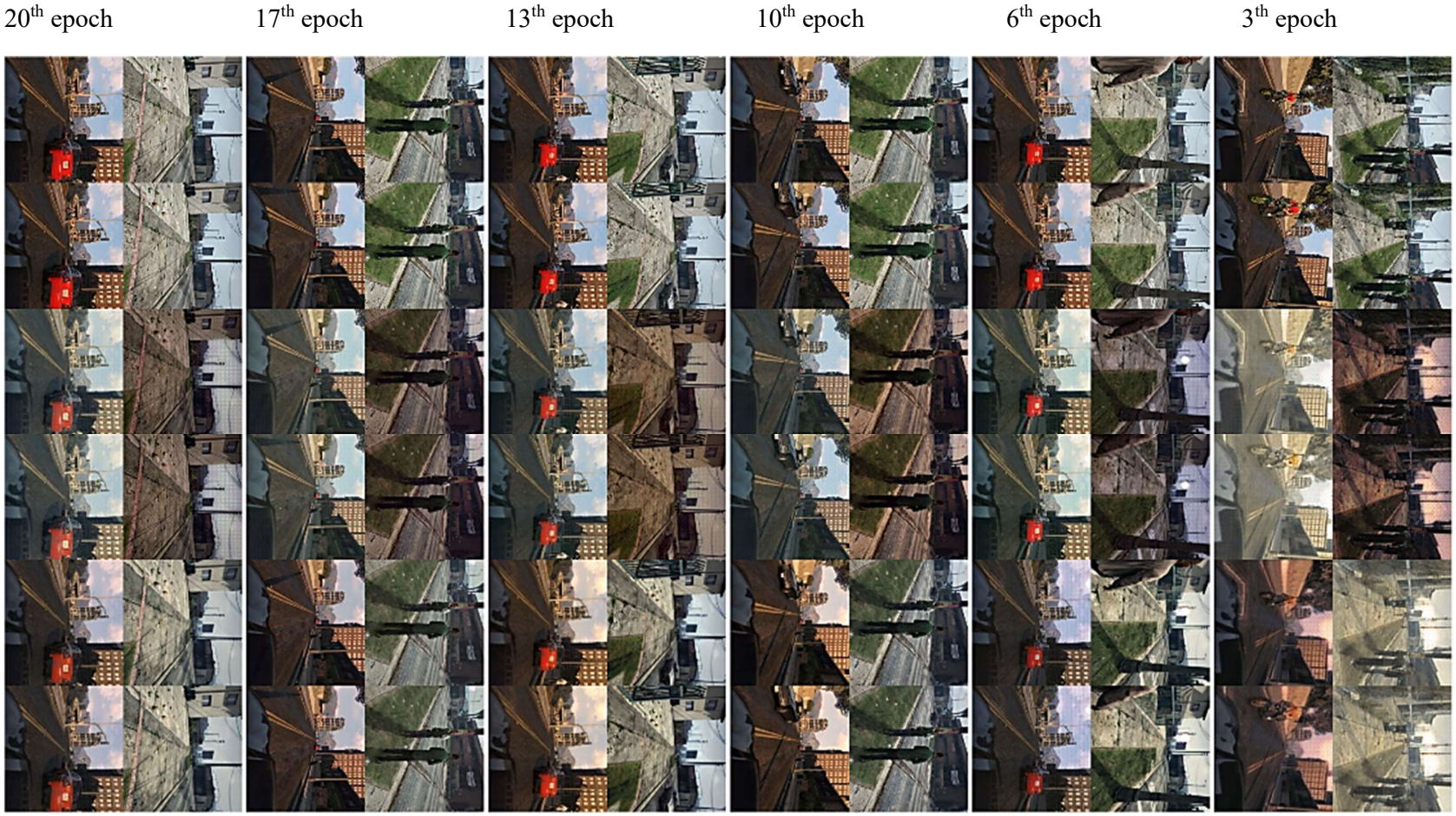
[52] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved Techniques for Training GANs,” *Adv. Neural Inf. Process. Syst.*, pp. 2234–2242, Jun. 2016, Accessed: Sep. 25, 2020. [Online]. Available: http://arxiv.org/abs/1606.03498.

APPENDIX

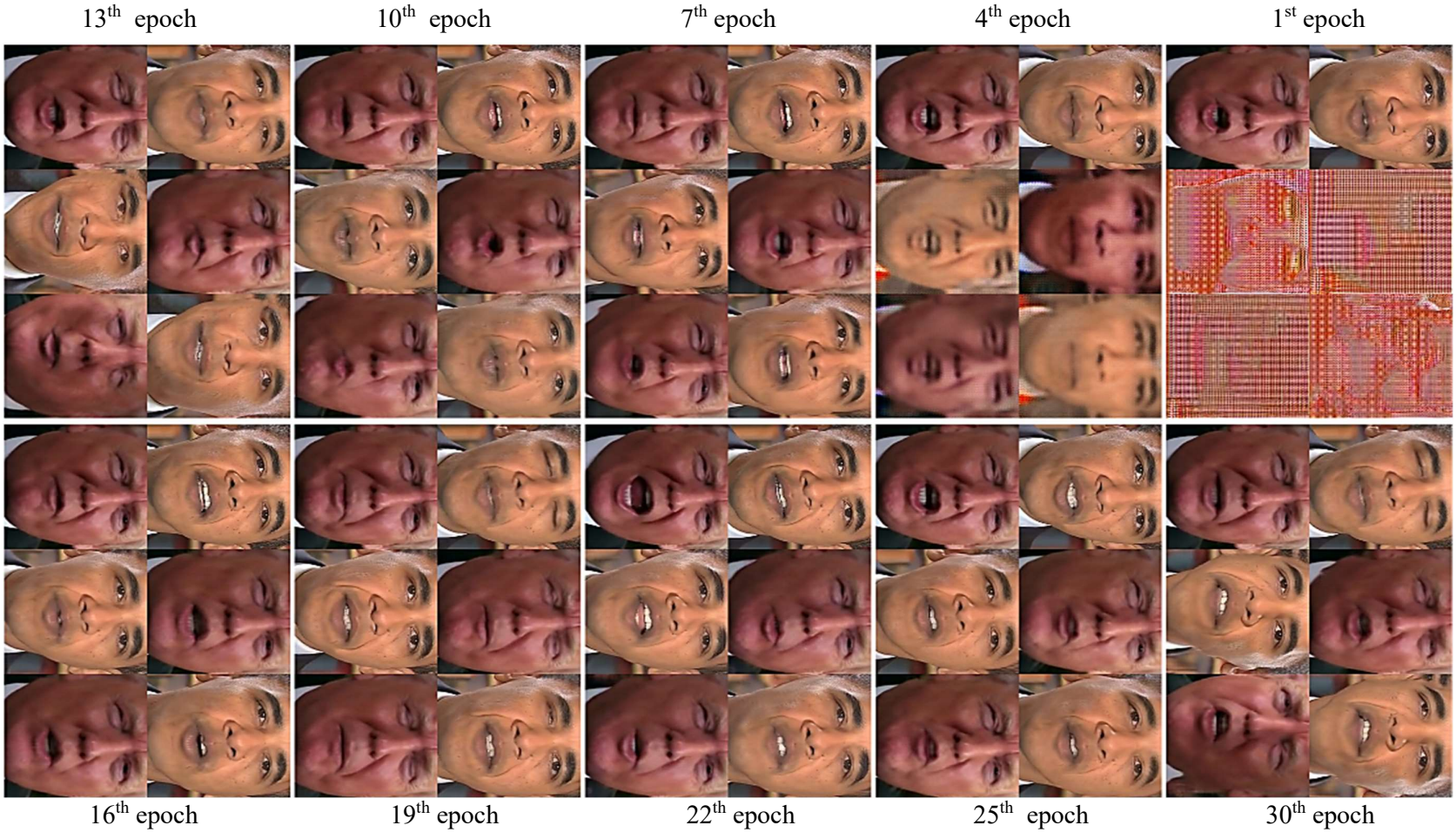
Appendix A: Result on Different epochs



The result on different epoch on the flower dataset, row label shows the corresponding epoch

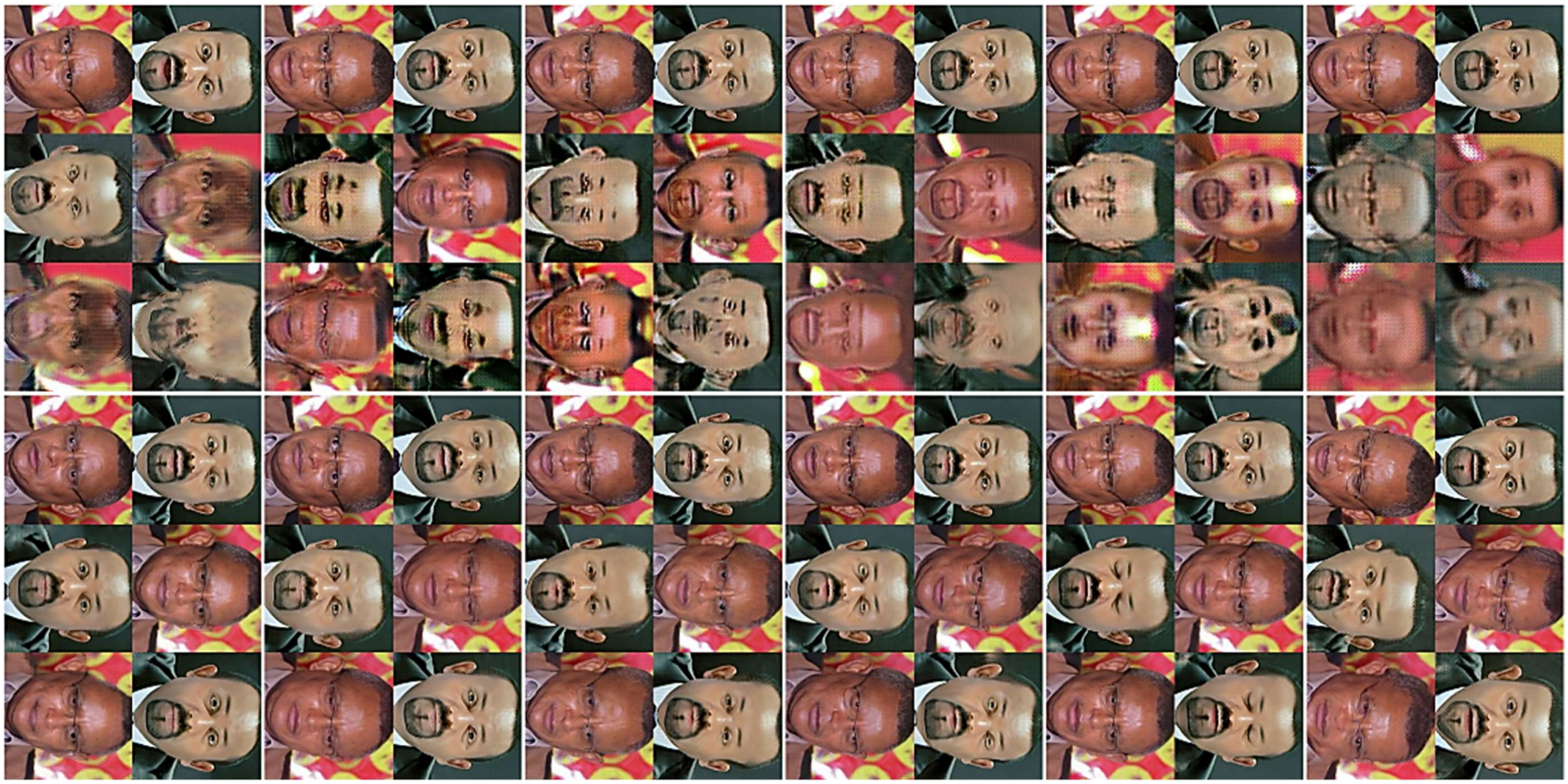


The result on different epoch left on the viper dataset, the label shows the corresponding epoch



Result on different epoch on Obama Trump dataset; row label shows the corresponding epoch

17th epoch | 14th epoch | 10th epoch | 7th epoch | 3th epoch | 1st epoch



30th epoch | 29th epoch | 27th epoch | 25th epoch | 23th epoch | 20th epoch

Result on different epoch on Adiss(አዲስ) dataset; row label shows the corresponding epoch

Appendix B: Training pseudocode:

Cycle-GAN Training pseudocode:

$take\ a\ sample\ mini-batch: x, y$

Train D:

$Translate\ A, B: \tilde{x} = G_{AB}(x),\ \tilde{y} = G_{BA}(y)$

$Compute: D_A(x), D_B(y), D_A(\tilde{y}), D_B(\tilde{x})\ then,$

$$dloss = \frac{1}{4} * \sum((D_A(x), D_B(y)), (D_A(\tilde{x}), D_B(\tilde{y})))$$

$update\ \Theta^{(dloss)}\ to\ minimize\ classification\ loss.$

Train G:

$Compute: \tilde{\tilde{x}} = G_{AB}(\tilde{x}),\ \tilde{\tilde{y}} = G_{AB}(\tilde{x}), xI = G_{BA}(x), yI = G_{AB}(y)$

$cycle\_loss = \frac{1}{2}(mae((x - \tilde{\tilde{x}}), (y - \tilde{\tilde{y}})))$

$identity\_loss = \frac{1}{2}(mae((x - xI), (y - yI)))$

$D_A(\tilde{y}), D_B(\tilde{x}): d\_loss = \frac{1}{2}\sum(DA(\tilde{y}), DB(\tilde{x}))$

$update\ \Theta^{(\text{d\_loss,cycle\_loss,identity\_loss})}\ to\ maximizeclassification\ loss.$

Cycle-GAN with Feature Preserving Training pseudocode:

$Take\ a\ sample\ mini-batch: x, y$

Train D:

$Translate\ A, B: \tilde{x} = G_{AB}(x),\ \tilde{y} = G_{BA}(y)$

$Compute: D_A(x), D_B(y), D_A(\tilde{y}), D_B(\tilde{x})\ then,$

$$dloss = \frac{1}{4} * \sum((D_A(x), D_B(y)), (D_A(\tilde{x}), D_B(\tilde{y})))$$

$update\ \Theta^{(dloss)}\ to\ minimize\ classification\ loss.$

Train G:

$Compute: \tilde{\tilde{x}} = G_{AB}(\tilde{x}),\ \tilde{\tilde{y}} = G_{AB}(\tilde{x}), xI = G_{BA}(x), yI = G_{AB}(y),$

$cycle\_loss = \frac{1}{2}(mae((x - \tilde{\tilde{x}}), (y - \tilde{\tilde{y}})))$

$identity\_loss = \frac{1}{2}(mae((x - xI), (y - yI)))$

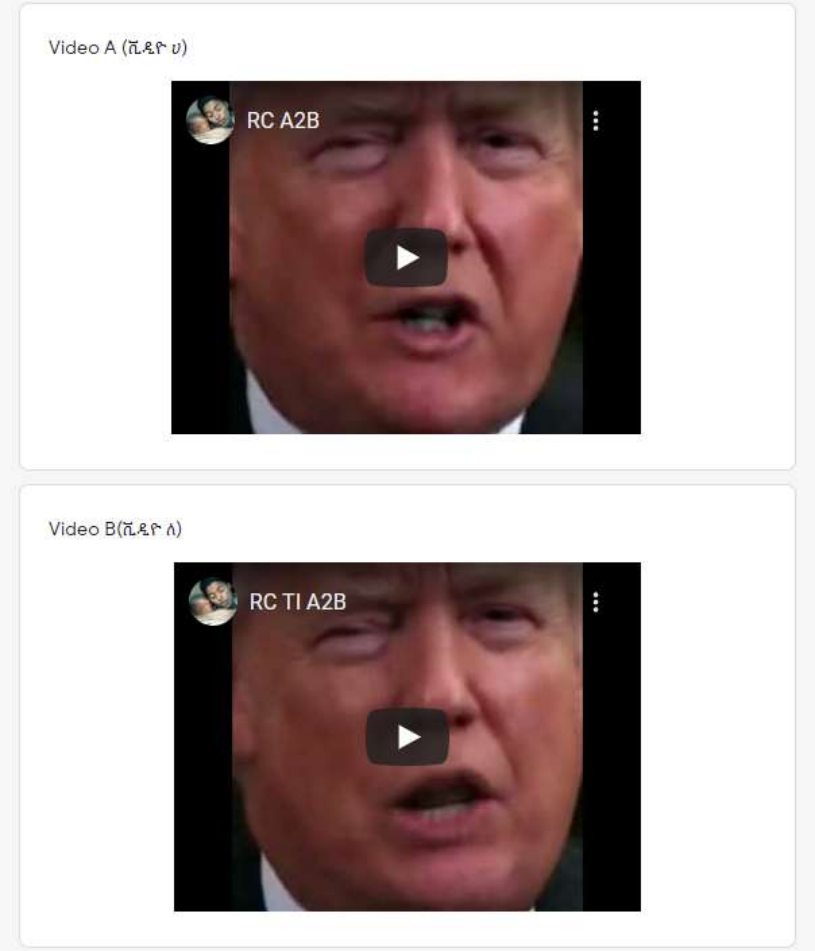
$$\boldsymbol{feature\_preserving\_loss = \frac{1}{2}(mae((mNet(x) - mNet(\tilde{x})), (mNet(y) - mNet(\tilde{y}))))}$$

$D_A(\tilde{y}), D_B(\tilde{x}): d\_loss = \frac{1}{2}\sum(DA(\tilde{y}), DB(\tilde{x}))$

$update\ \Theta^{(\text{d\_loss,cycle\_loss,identity\_loss,}feature\_preserving\_loss)}\ to\ minimize\ classification$
.

Appendix C: User Study Evaluation Form Questions Sample

Q1, Please carefully watch the next two videos, and answer the question accordingly
እባክዎን የሚቀጥሉትን ሶስት ቪዲዮዎች በጥንቃቄ ይመልከቱ ፣ እናም ለጥያቄው በዚሁ መሠረት ይመልሱ?



From the Above 2 Videos which one looks more real Trump Video? (ከላይ ከ 2 ቪዲዮዎች የትኛው ይበልጥ እውነተኛ የትራምፕ ቪዲዮን ይመስላል?)

- Video A (ቪዲዮ ሀ)
- Video B (ቪዲዮ ለ)

Q2, Please carefully watch the next two videos, and answer the question accordingly
እባክዎን የሚቀጥሉትን ሶስት ቪዲዮዎች በጥንቃቄ ይመልከቱ ፣ እናም ለጥያቄው በዚሁ መሠረት ይመልሱ?

Video A (ቪዲዮ ሀ)

Video B (ቪዲዮ ለ)

From the Above 2 Videos which one looks more real Obama Video? (ከላይ 2 ቪዲዮዎች የትኛው ይበልጥ እውነተኛ የአባማ ቪዲዮን ይመስላል?)

- Video A (ቪዲዮ ሀ)
- Video B (ቪዲዮ ለ)

Trump to Obama (ከትራምፕ ወደ አባማ)

Q3, Please carefully watch next two videos, and answer the question accordingly

እባክዎን የሚቀጥሉትን ሁለት ቪዲዮዎች በጥንቃቄ ይመልከቱ ፣ እናም ለጥያቄው በዚሁ መሠረት ይመልሱ?

in the next videos, you will see Obama to Trump video translation

በቀጣዮቹ ቪዲዮዎች ውስጥ አባማን ወደ ትራምፕ የቪዲዮ ትርጉም ያያሉ

You will see two concatenated videos, the one on the left is original, and the right is fake.

ሁለት ተጣምረው የሚንቀሳቀሱ ቪዲዮዎችን ያያሉ ፣ በስተግራ ያለው የመጀመሪያው ኦሪጅናል ሲሆን በቀኝ በኩል ደግሞ ሃሰተኛ ነው::

Video A (ቪዲዮ ሀ)

Video B (ቪዲዮ ለ)

From the Above 3 which looks like a more realistic and natural translation? ይበልጥ ተጨባጭ እና ተፈጥሮአዊ የምስል ትርጉም ያለው የትኛው እንደሆነ ይጠይቁ? *

- Video A (ቪዲዮ ሀ)
- Video B (ቪዲዮ ለ)

Obama to Trump(ከአባማ ወደ ትራምፕ)

Q4, Please carefully watch the next two videos, and answer the question accordingly

እባክዎን የሚቀጥሉትን ሁለት ቪዲዮዎች በጥንቃቄ ይመልከቱ ፣ እናም ለጥያቄው በዚሁ መሠረት ይመልሱ?

in the next videos, you will see Trump to Obama video translation

በቀጣዮቹ ቪዲዮዎች ውስጥ ከትራምፕ ወደ አባማን የቪዲዮ ትርጉም ያያሉ

You will see two concatenated videos, the one on the left is original, and the right is fake.

ሁለት ተጣምረው የሚንቀሳቀሱ ቪዲዮዎችን ያያሉ ፣ በስተግራ ያለው የመጀመሪያው ኦሪጅናል ሲሆን በቀኝ በኩል ደግሞ ሃሰተኛ ነው::

Video A (ቪዲዮ ሀ)

Video B (ቪዲዮ ለ)

From the Above 2, which looks like a more realistic and natural translation? ይበልጥ ተጨባጭ እና ተፈጥሮአዊ የምስል ትርጉም ያለው የትኛው እንደሆነ ይመስላል?

- Video A (ቪዲዮ ሀ)
- Video B (ቪዲዮ ለ)

Appendix F: Ablation study

This thesis work proposed to implement a temporal discriminator network, which is a modification on the patchGAN network by concatenated previously generated output images. This Ablation study has been done to come up with the number of images the discriminator network shall consider to justify output is whether real or fake. The figure below shows a comparison among different discriminator models and it indicates the GAN network performs well when three frames are considered by the discriminator[10].

The optical flow between consecutive frames of (input) original image, (one) vanilla patchGAN, (two) discriminator consider two frames, (three) discriminator consider three frames, (four) discriminator considers four frames.

[10] For better comparision please watch: https://youtu.be/OWh-ffKROfQ