“***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

December 2019

“***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

December 2019

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 (PHD) | \_ |

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).

**Yun Koo Chung (PHD) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**(Advisor) Signature Date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
(Chairperson) Signature Date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
(Internal Examiner) Signature Date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
(External Examiner) Signature Date

**Acknowledgment**

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

***This thesis work is the results of one-year research and hard work, so that it is very difficult part to acknowledge because so many people have helped me in so many different ways. Nonetheless, there are those that should be recognized for the critical help they have provided during the course of research.***

***I would like to express my special deepest gratitude and heartfelt thanks to my advisor and computer vision Special interest Group Leader Prof. Yun Koo Chung (Ph.D.) for his patiently guidance, valuable support and supervision starting form proposal until to the final completion of this thesis.***

***I also thanks computer vision program staff and postgraduate students, friends and parents for their every support and advice in my life. May God always Bless and keep them safe.***

Table of Contents

[CHAPTER ONE 2](#_Toc48646355)

[Introduction 2](#_Toc48646356)

[1.1. Backgrounds 5](#_Toc48646357)

[1.1.1. Generative Adversarial Networks 5](#_Toc48646358)

[1.2. Motivation 7](#_Toc48646359)

[1.3. Statement of the Problem 7](#_Toc48646360)

[1.4. The Objective of the Thesis 8](#_Toc48646361)

[1.4.1. General Objective 8](#_Toc48646365)

[1.4.2. Specific Objective 8](#_Toc48646366)

[1.5. Research Methodology 9](#_Toc48646367)

[1.5.1. Data Collection 9](#_Toc48646369)

[1.5.2. Literature Review 9](#_Toc48646370)

[1.5.3. Evaluation 9](#_Toc48646371)

[1.5.4. Implementation Tools 9](#_Toc48646372)

[1.6. Scope and Limitation 10](#_Toc48646373)

[1.6.1. Scope 10](#_Toc48646375)

[1.6.2. Limitations 10](#_Toc48646376)

[1.7. Organization of the thesis 10](#_Toc48646377)

[CHAPTER TWO 12](#_Toc48646378)

[Literature review 12](#_Toc48646379)

[2.1. Introduction 12](#_Toc48646382)

[2.2. Inside GAN 13](#_Toc48646383)

[2.2.1. GAN training 14](#_Toc48646386)

[2.2.2. Conditional GAN 15](#_Toc48646387)

[2.3. Image to Image translation 16](#_Toc48646388)

[2.4. Video to video translation 18](#_Toc48646389)

[2.4.1. Temporal information 19](#_Toc48646392)

[2.4.2. Spatio-temporal information 21](#_Toc48646393)

[2.5. Recent work summary 23](#_Toc48646394)

[CHAPTER THREE 27](#_Toc48646395)

[Materials and Methods 27](#_Toc48646396)

[3.1. Overview 27](#_Toc48646399)

[3.2. Dataset 27](#_Toc48646400)

[3.3. Development tools 28](#_Toc48646401)

[3.4. Design tools 28](#_Toc48646402)

[3.5. Prototype development framework 28](#_Toc48646403)

[3.5.1. TensorFlow 28](#_Toc48646408)

[3.5.2. Keras 29](#_Toc48646409)

[3.5.3. OpenCV 29](#_Toc48646410)

[3.5.4. MATLAB Deep Network Designer 30](#_Toc48646411)

[3.6. Evaluation methods 30](#_Toc48646412)

[3.6.1. User study 30](#_Toc48646413)

[3.6.2. Inception score 30](#_Toc48646414)

[CHAPTER FOUR 31](#_Toc48646415)

[The Proposed work & it’s Implementation 31](#_Toc48646416)

[4.1. Overview 31](#_Toc48646418)

[4.2. Introduction 31](#_Toc48646419)

[4.3. Image to image transfer 32](#_Toc48646420)

[4.1. Cycle-GAN 33](#_Toc48646421)

[4.1.1. Cycle Loss 33](#_Toc48646422)

[4.1.2. Identity Loss 33](#_Toc48646423)

[4.2. Feature preserving loss 34](#_Toc48646424)

[4.3. Temporal information extractor. 34](#_Toc48646425)

[4.4. Spatial-temporal information fusion 35](#_Toc48646426)

[4.5. Bring all together (Proposed Work) 36](#_Toc48646427)

[4.6. Working Environment. 42](#_Toc48646428)

[4.7. Set upping the environments 42](#_Toc48646429)

[4.8. CycleGAN specification 43](#_Toc48646430)

[4.9. Implement cycle GAN 44](#_Toc48646431)

[4.10. Temporal information extraction Implementation 45](#_Toc48646432)

[4.11. Experiment Class 45](#_Toc48646433)

[CHAPTER FIVE 46](#_Toc48646434)

[Evaluation, Results, and Discussion 46](#_Toc48646435)

[5.1. Overview 46](#_Toc48646437)

[5.2. Video to video Translation 46](#_Toc48646438)

[CHAPTER SIX 49](#_Toc48646439)

[Conclusion and Future work 49](#_Toc48646440)

[References 50](#_Toc48646441)

[Appendix 54](#_Toc48646442)

List of Figure

**Abbreviations**

GAN Generative Adversarial Network

Cycle GAN Cycle consistent Generative Adversarial Network

Recycle GAN Recurrent Cycle Consistency Generative Adversarial Network

RNN Recurrent Neural Network

Generator network transfer to target domain Y

Generator network transfer to target domain X

Dx Discriminator network for Generator

Dx Discriminator network for Generator

IoU Intersection-Over-Union

MP Mean Pixel Accuracy

AC Average Class Accuracy

GTA v Grand Thief Auto five(v)

3D three-dimensional.

Pix2pix pixel to pixel

FCN Fully connected network

# CHAPTER ONE

# Introduction

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.

In order to solve computer vision problems, Artiﬁcial Intelligence (AI) is an active ﬁeld that concerns this topic. It started when the nascent ﬁeld of computer science started to ask if a computer could become intelligent or mimic cognitive abilities that lead to knowledge such as learning, problem-solving, and reasoning. At the beginning of the development of AI, the software was hard-coded with knowledge about the world with a list of formal, mathematical rules. This approach never led to a major victory due to the struggle of describing the complexity of the world with sophisticated mathematical rules and formals. Instead of relying on hard-coded knowledge, AI systems needed a capability to extract their own knowledge. Systems started to extract patterns from raw data, this capability comes to be known as Machine Learning (ML) [3].

ML is a ﬁeld with many diﬀerent learning capabilities and it is still expanding. There are diﬀerent types of learning problems, (they are may not be the only types) the ﬁrst type is called supervised learning, that is when for every input variable , the output variable is known so an algorithm learns to map the input to the output and since the output (correct answer) is known for every input, the algorithm is said to be supervised. Another ML problem type is when only the input data is known, this is referred to unsupervised learning. The task here is to organize the data or to discover the structure or distribution of the data in order to learn more about it. Since there are no correct answers .

The last type is called semi-supervised machine learning and refers to problems where one part of the dataset is labeled and one part is unlabeled. This is very common because it is very expensive and time-consuming to label big datasets. Suppose a classiﬁcation problem where the data set is not fully labeled. Then unsupervised learning techniques can be used to discover the structure in the input variables. Or supervised learning techniques can be used to predict labels to every unlabeled . Even ML plays very tremendous work but it still fails to process complex data like image and video. So as to work with complex data problems Deep Learning (DL) an option.

DL is a subﬁeld of ML and has a special style for learning representations from data. Instead of learning one representation, DL algorithms learn successive layers of increasingly meaningful representations of the data. In other words, representations are expressed in terms of other, simpler representations. With this approach, a hierarchy of features is built and it is, therefore, possible to extract high-level features from raw data. This hierarchy of layers creates a deep Graph named Deep Learning. The quintessential example of a DL model is an artiﬁcial neural network (ANN)[1]. The research around DL exploded in 2012 when Alex Krizhevsky achieved remarkable results in the ImageNet competition (ILSVRC2012) using a convolutional neural network (CNN) [2]. But the pioneer of CNNs goes to Yann LeCun [3] when he in 1989 used a CNN to recognize handwritten digits. At that time, DL algorithms were outperformed by other ML algorithms due to two factors: the ﬁrst was because of the lack of available data and the second due to bad performance in hardware. So researchers did not see the potential of DL until a few years ago when the amount of data and the hardware performance increased. Today, DL is used in facial recognition, robotics, object detection, speech recognition, and translation.

One interesting outlet of unsupervised learning techniques is Generative Models. Usually, it is tough to analyze and understand data but generative models can do so. They are trained to discover the essence of data in some domain in command to generate similar data. This technique can be used in many tasks, for instance for image denoising, inpainting, super-resolution, structured prediction, exploration in reinforcement learning, etc. In the long run, the idea is to let computers automatically learn the natural features of data and to get a better understanding of the world.

A generative model that has recently achieved major success is called **Generative Adversarial Network (GAN)** [4] and it was introduced in 2014. GAN has been a hot research topic among computer vision researchers nowadays it learns a given data distribution in order to generate realistic-looking fake distribution. Basically, GAN contains two networks Neural networks that play zero-sum game namely called generator and discriminator- where generator generates fake data while discriminator tries to classify if the data generated is tangible or forged. This work tackle domain transfer for video.

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.*

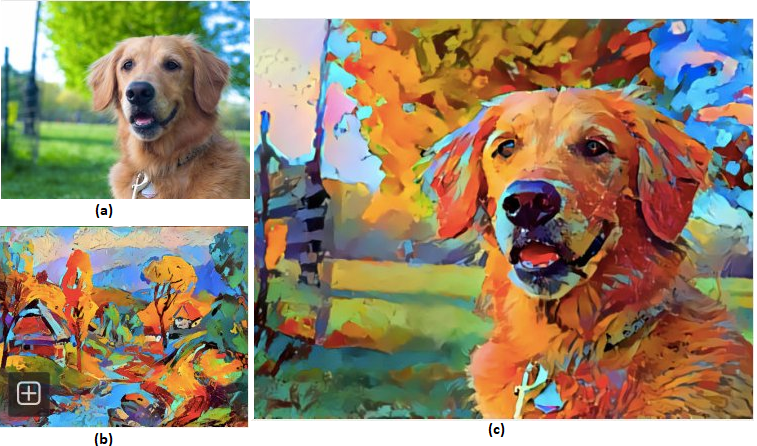


Figure 1 (a) input image. (b) style image. (c) output image

Last year (2018) all Artificial intelligence news [5] headlines were screaming about a paint drawn 100% by AI sold $432,500 (fascinating, isn't it?). (Because of style transfer data scientist doesn’t 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)

Perhaps the first successful neural style transfer paper was published in 2015 by [6]**.** After this work, many researchers came with a more realistic synthetic image. pix2pix [7] 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. [8] 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.

## Backgrounds

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.

### Generative Adversarial Networks

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.**,** in another word 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**and it aims to generate new data similar to the samples from real datasets. The Generator could act as a human art forger, which creates fake works of art.

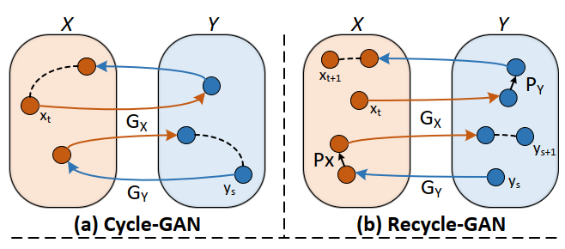
The second Network is entitled the **Discriminator**. This model’s goal is to recognize if an input data is ‘real’ — goes to the original dataset — or if it is ‘fake’ — generated by a falsifier Generator Network. In this scenario, a Discriminator is corresponding to the law enforcement agent (or an art 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.

Figure 2 Cycle Gan vs Recycle GAN

#### Conditional GAN

The conditional GAN[9] is an extension of the [4] 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 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 as shown Figure 3.

#### 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 human face [10] they lack generalization and doesn’t 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. [7] use conditional GAN [9], learn a mapping between paired input to the output image. The third category is unsupervised and unpaired data domain transfer like cycle GAN [11] which works enforcing cycle consistency for the unpaired image.

The recent state of artwork work ReCycle GAN by Bansal [11] motivated by [8] propose video retargeting via spatiotemporal constraint though directly synthesize future frame via temporal predictor to preserve temporal continuity But 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 **cycle** **motion transfer constraints also proposes to introduce Spatiotemporal video to** **video** translation for better realistic results.

## Motivation

Recent deep learning achievement has done because of the huge amount of data available nowadays, but still a there is a big problem to collect data especially when we need paired data set (day and night) since capturing datasets in two (or more) completely different environments is dreadful. This thesis work plan to improve video to video transfer which is one of the mechanisms to overcome such problems and it could have a great impact on computer vision and deep learning society. Even though this work focused on unpair dataset there is one major addition this thesis work could back specifically in data augmentation to improve data insufficiency in deep learning so as to improve convergence experience. since there is still no enough data in many computer vision problems. this issue remains one of the extremely challenging problems in computer vision when the real-life scenario is considered. This study tries to solve video temporal discontinuity problem by extending solutions presented in previous works [11], [12] explicitly for a video to video translation.

## 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 and respectively, cycle constraint enables an image to image translation in mutually frontward and backward mapping. There are two mapping functions mapping from domain and correspondingly form target domain to source and vice versa. where is input video frame at time and is a synthetic frame in domain same is true in domain. Cycle consistency constraint s.t as well as s.t.

Besides the preservation of cycle consistency in each frame this work-study mapping cycle motion consistency between consecutive frames in both domains. Meaning let optical flow between and is and optical flow between is , then, temporal cycle consistency need to enforce motion consistency via minimizing the difference between . Recycle-GAN [11] claims “*video to video translation is under constraint*” this work proposes toward add temporal cycle consistency to the extended video to video translation to see more constraints in its result.

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

* What are effective temporal cycle consistency constraints for a video to video translation?
* What is the effect of temporal cycle consistency for video translation impact on the unsupervised video to video transfer?

## The Objective of the Thesis



### General Objective

The general objective of the study is to design and implement **Temporal Cycle consistency** for **video to video translation**. This work motivated by [11] ReCycle GAN.

### Specific Objective

The following specific objectives will be addressed to achieve the general objective.

* Reviewing related works to understand the area and the works that are done by others.
* Gather dataset for training and testing.
* To preprocess the dataset in order to enhance its quality.
* To effectively extract temporal information from the video.
* To design a deep learning video translation model using Keras and TensorFlow framework.
* To blend spatial Information and Temporal Information.
* To train the model using a proper dataset.
* To test the trained model with a test set.
* To assess the performance of the model.

## Research Methodology

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



### 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.

### 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.

### 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.

### 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.

## Scope and Limitation



### Scope

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

* Translate a given domain video (sequence of image) to another domain.
* Blend spatial information to temporal information to improve the consistency of video to video translation.

### Limitations

This paper 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).
* The video does not zoom in or out throughout the whole process.

## 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 networks and temporal information blending with the spatial model. Flow chart and pseudocode for implementation, training, and testing with mathematical correspondent description have been discussed.

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

# CHAPTER TWO

# Literature review



## Introduction

The most impressive success in Deep Learning has, so far, involved discriminative models, i.e. models that map the dependence of unobserved target variables (y) on observed variables (x) – Classification problem. In simple terms, discriminative models suppose outputs based on inputs without considerate about how the input was generated. In another sense Generative models opposed to discriminative models, which maps how the input data was generated.

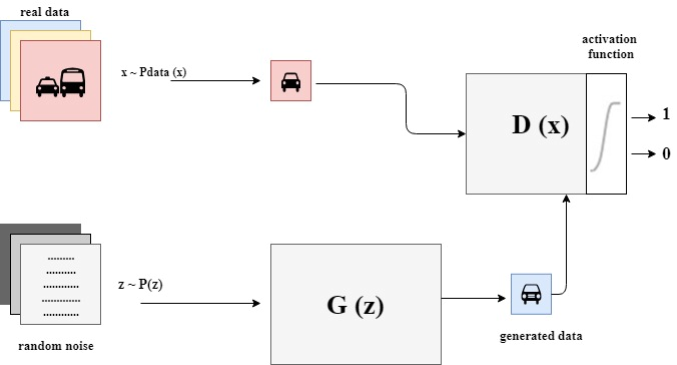
GANs (Generative Adversarial Networks) [4] is a fit into the generative type of network. GANs are taught to generate synthetic data alike to known input data. A GAN model consists of two types of neural networks inside, a generator model and a discriminator model. The two Deep Neural networks have an adversarial relationship where they fight against each other[[1]](#footnote-1). The generative model learns to mimic data while the discriminative model learns to determine whether a sample is from the model distribution or the data distribution .

Figure 4 GAN framework structure GAN framework consists of two networks: Discriminator (𝐷) and Generator (𝐺)

During training, both models improve their methods until the artiﬁcially generated data are indistinguishable from real data. In this chapter, the paper briefly describes the technologies, methods, and frameworks mentioned throughout the thesis.

## Inside GAN

In this section let see the detail inside of GAN. As we discuss GAN in the previous section, GAN consists of two independent networks Generator and Discriminator as shown in figure 4, which are represented by diﬀerentiable functions concerning each network’s input and parameters. The discriminator is deﬁned by a function where (observed variable) is the input which is a real dataset. gives the likelihood that came from (real distribution) rather than (fake distribution). It is a binary classiﬁer with two classes, when is real the probability is 1 and when 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 where (latent variable) is the input and generates an image as its output . 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 [4]:



Equation 1 Adversarial loss function

The generator can be seen as a kind of reverse CNN. It takes an -dimensional vector of noise and upsamples it to an image 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.



### 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 different for GAN training is all about process learning to mimic the real dataset samples. Unlike many deep learning models training is a bit tricky so let’s dive into it. But before that let sees a conflict between discriminator and generator.

The Discriminator’s goal is to be as precise possible (binary classification). For the real examples seeks to be as close as possible to 1 (label for the positive class). Meaning attempts to be 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 that are alike from the real data in the training dataset. Mathematically, the Generator strives to produce fake examples such that is as close to 1 as possible.

Table 1generator goal vs discriminator goal

|  |  |
| --- | --- |
| Generator | such that is as close to 1 as possible. |
| Discriminator | tries to be as close as possible to 0. |

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

1. Train the Discriminator:
   1. Take a random mini-batch of real examples: .
   2. Take a mini-batch of random noise vectors z and generate a mini-batch of fake examples:
   3. Compute the classification losses for and , and backpropagate the total error to update to minimize the classification loss.
2. Train the Generator:
3. Take a mini-batch of random noise vectors z and generate a mini-batch of fake examples:
4. Compute the classification loss for and backpropagate the loss to update 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 GAN model converges when the discriminator and the generator reach a Nash equilibrium. This is the optimal point for the minimax GAN loss function (equation 1). Regarding GAN at Nash equilibrium discriminator no longer able to distinguish between real and fake samples so it randomly classifies

### Conditional GAN

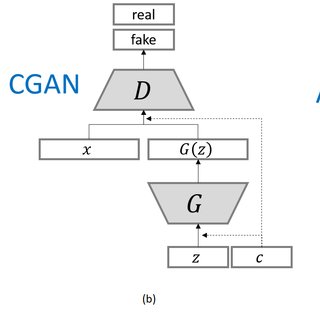
Even if GAN models are capable of generating new random plausible examples for a given dataset, there is no way to control the types of images that are generated. But the network tries to figure out the composite association between the latent space input to the generator in order to mimic the real dataset and the generated images[4], [13].

Figure 5 cGAN Architecture

Mirza et al propose The conditional generative adversarial network, or cGAN [9] for short, which is a type of GAN that involves the conditional generation of images by a generator model. Image generation can be conditional on a class label [[2]](#footnote-2), it forces the Generator network to generate only targeted class of images of a given type by conditional variable. Conditional variable is fade to the generator and discriminator networks as shown in figure 5 above. This work unlocks opportunities for many fascinating research topic like image to image translation, style transfer and video retargeting [11], [14]. The next section will discourse about Image to image translation.

## Image to Image translation

Let me 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****”* [15]I was wondering too winter to summer Absolutely fascinating.

Recent advancements in GANs [4] empowers style transfer models to create realistically looking [6]–[8], [16] adapted image (figure 6 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 to mean the dataset should contain images that are one to one correspondence as shown 6A. The major 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[7] 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 learn to map pixel from the first image to the second one.

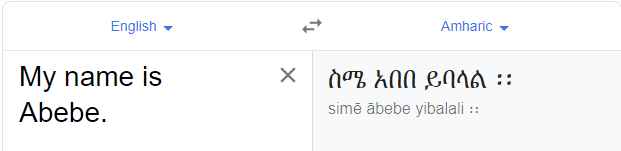
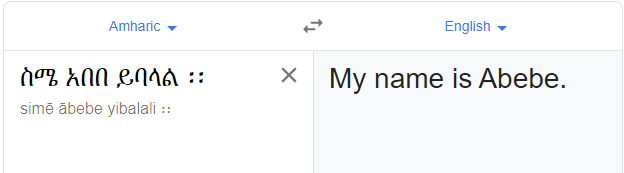
Because in reality pair datasets are very rear and expensive Zhu et al. came up with Cycle-GAN [8] which invented to learn bidirectional mapping in absence of paired training data via Cycle consistency loss. *Cycle Consistency loss* utilizes to learn transformation between two domains in a frontward 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 English to Amharic in both directions. When the user input “My name is Abebe” the model should generate “ስም አበበ ይባላል፡፡” perhaps if the user translates “ስም አበበ ይባላል፡፡” to English back again it should generate the original text “My name is Abebe”. 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 above.

Figure 6 language translation using google translator available on line.

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’s see it in bit detail using table

Table 2 Cycle GAN generator and discriminator operation.

|  |  |  |
| --- | --- | --- |
|  | Translate from to |  |
|  | Translate from to |  |
|  | Classify real and fake | 1 for , 0 for |
|  | Classify real and fake | 1 for , 0 for |

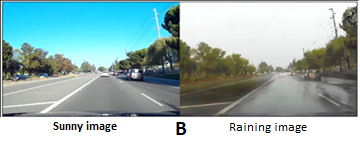
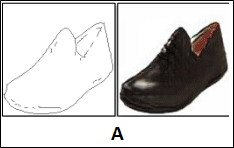
The loss function of the network could be formulated as:

Figure 7. (A) pair shoe dataset sample from Pix2pix (B) Sunny to Rainy translation from input and output image

Meaning translate a given image are and reconstructed image the difference should be the minimum . input image translated to another domain and retranslated back to its original

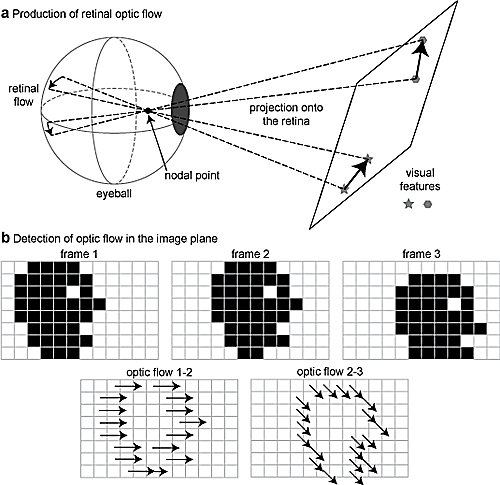
domain. Ahead of image transformation across domain video to video translation is an additional extension. Now let’s jump to video to video translation.

Figure 8 General Optical flow Computation Mechanism.

## 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 relation between them. This approach is non-trivial since the key issues which underlie flickering [17], [18] effect in the output video. To overcome the flickering effect Chen et al [18] consider temporal information along with spatial information. Specifically, they exploit previous frame optical flow to warp the current frame towards impose temporal constraints. Let see what is temporal information and different mechanism to extract.



### Temporal information

Temporal refers to time domain where in our case it can be seen as relation between sequence of while Spatial refers to RGB space frames. Spatiotemporal, or spatial-temporal, is used in information analysis when data is collected across both time and space. Straight forward approaches basically fail because it can’t consider both domains. Temporal information for video can describes a phenomenon in a certain pixel location with position change in time. For a video to video translation, we have various [[3]](#footnote-3)options to represent motion information. The next section would discuss those topics.

#### Optical flow

Optic flow is the change of structured pixels with specific intensity in successive images, or in another word, Optical flow is the motion of objects among successive frames, caused by the comparative movement among the object and camera. This make optical flow an ideal for encoding temporal information[11], [12], [19].

Figure 8 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) as shown in figure 8, which may be expressed as: where  is the motion in the  direction,  is the motion in the direction, and  is a neural network that takes in two consecutive frames  (frame at time  ) and  (frame at time as input.

Computing optical flow with deep neural networks [20], [21] requires huge amounts of training data which is principally hard to obtain. This is because tagging video footage for optical flow requires precisely figuring the exact motion of every point of an image to subpixel accuracy. To address the issue of labeling training data many research works, [20], [22], [23] 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 [24], an open-source CGI movie with optical flow labeling rendered for numerous sequences, and Flying Chairs [22], a dataset of numerous chairs hovering around random backgrounds both generated from the virtual world using Computer Graphics.

#### Pose estimation

Human pose estimation can be framed as the problem on localization of key points like eye, nose, elbows, wrists, etc. in images or videos sometimes called human joints. It is also defined as the exploration of an explicit pose in space of all articulated poses. Basically, pose estimation translate used in transferring motion from a deriving video to derived object in a video. Particularly human pose estimation used in transferring motion from one person to other as used [25], to transfer motion between different domain videos specifically for animating static image by driving motion [26] and facial expression transfer [27] between source and target persons.

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 one has universally adopted ConvNets as their main building block. They largely replace hand-crafted features & graphical models perhaps this approach has returned drastic advances on standard benchmarks. Basically,

#### 3D convolutional tensor

Figure 10 pose extraction

The 3D convolutional tensor mechanism is one of the orthodox methods which basically doesn’t consider temporal information explicitly. Since it considers presenting video scene [28] 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 basic reasons requires a high-efficiency machine, and the network becomes a fully black box. Meaning hard to tune parameters basically done in training Deep Learning models.

#### Recurrent temporal

Recurrent neural networks or RNNs are a type of neural networks inherently ideal for analyzing data from time series and other sequential figures make it ideal for video analysis. conceivably it’s overcame black box nature of 3D conv nets by adding additional parameter to tune the network. Recent works consider using LSTM (Long Short Time Memory.) which consider all previous frames as an input to minimize temporal residual error [19] other works use RNN for temporal predictor[11].

### Spatio-temporal information

In the previous section, we discussed temporal information (motion information) extraction mechanism. But since video consists of both temporal and spatial information, we need to discuss mechanisms to get the advantage over an early approach (spatial only). So 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. This approach is non-trivial since the key issues which motivate flickering effect in the results output video[17], [18].

To overcome the flickering effect Chen et al [18] 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 doesn’t consider newly introduced pixels in given frame scene). To overcome the occlusion and fast illumination change other fine work by Chen et al [29] MoCycle GAN introduce temporal motion translation to transfer estimated motion from source to target video while preserving temporal consistency. this work also relives temporal cycle constraint for motion reconstruction.

The current state of artwork [11] ReCycle GAN further extend 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 work by [19] proposes an optical flow residual error between ground truth and warped frame mechanism to guarantee the local and global consistency to overcome the temporal flickering and motion inconsistency between frames*.*

## Recent work summary

the following table illustrates a summary of previous works on the video to video translation[[4]](#footnote-4).

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Dataset (Data collection) | Architecture | Temporal information modeling | Result refinement | Evaluation matrix used | Limitation |
| Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks | Cityscapes, Horse to Zebra and Apple to Orange, Summer to Winter Yosemite, etc.. | Cycle Consistency Constraint. | No temporal Information considered. | - | FCN, AMT perceptual studies | Framewise image to image translation. |
| Recycle-GAN: Unsupervised Video Retargeting  Recycle GAN[11] | Viper, face, and flower datasets (more than 10,000 images) | CycleGAN with recurrent temporal predictor | Recurrent temporal predictor(pix2pix) | - | IoU, pixel accuracy, Average class accuracy | Temporal predictor basically fail to correctly predict, and no cycle consistency temporal cycle considered |
| Mocycle-GAN: Unpaired Video-to-Video Translation Mocycle GAN[29] | Flower video and viper dataset | CycleGAN with motion translator-based motion cycle consistency | Flownet2.0 with motion translator network | - | Human evaluation, IoU, pixel accuracy, Average class accuracy | Explicit motion translator |
| Animating Arbitrary Objects via Deep Motion Transfer Monkey-Net[27] | UvA-Nemo, Tai-Chi, and BAIR robot pushing datasets | RNN based Dense motion predictor and motion translation network | Keypoint detector with motion transfer network based on motion heatmap | - | L1, AKD, MKR, AED , and FID | No random input size. |
| Unsupervised Video-to-Video Translation Dina [28] | Volumetric MNIST,GTA segment to video and MRI-to-CT | 3D CycleGAN | Network implicitly learns to form input video | - | Human evaluation, pixel accuracy and L2 error between original and retranslated image | 3D tensor fails for temporal learning consistency between frames. |
| Preserving Semantic and Temporal Consistency for Unpaired Video-to-Video Translation [12] | Viper dataset | RNN based CycleGAN with flow estimator network and consistency warping network | Flownet2.0 base temporal fuse with spatial for improving occlusions problem | Use [30] to further reduce the Temporal warping error. | mIoU, fwIoU, and pixel accuracy | Only consider local temporal consistency |
| Video-to-Video Translation with Global Temporal Consistency[19] | DAVIS 2017 | RNN based CycleGAN, and RNN based Discriminator for global temporal consistency | Flownet2.0, temporal residual error minimizer | - | Peak Signal to Noise Ratio, Region Similarity, and Contour Accuracy | Complex architecture hard to train doesn’t consider temporal cycle continuity |

As shown in the above table, researchers design complex architectures used in previous works so as to learn a mapping from a domain to domain in an unsupervised manner.

# CHAPTER THREE

# Materials and Methods



## 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.

In order to achieve the objectives of this study, the following methods and techniques are employed.

## 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 face (Obama-trump),viper and, flowers are used for both training and testing stages as used in [11] .

* **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 in a different time and completely talk about different things). Each frame is 256 x 256 and about 8617 images 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, day ↔ night and sunrise ↔ sunset. For this study the frame resolution is Demote to (resolution: 256 × 256).

## 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 tools that are relevant to the research. The following sections give a brief detail about these development tools.

## Design tools

Design tools are media that are used for the production, expression, and perception of design ideas. Edraw Max [31] is used to design in the proposed system. It is a lightweight and powerful graphic design tool for creating professional-looking flowcharts, network diagrams, UML diagrams, and others. This tool is selected because [31].

* It has lots of high-quality shapes, example, and template,
* It easily visualizes complex information with a wide range of diagrams.
* Works with MS Office well and others.

## Prototype development framework



### TensorFlow

TensorFlow is an open-source software library, which is developed for high-performance numerical calculation and computation. Its flexible architecture that can be deployed easily on a variety of platforms like Central Processing Units (CPUs), Graphical Processing Units (GPUs), Tensor Processing Unit (TPUs). It also can be installed on desktops computers, clusters, mobile, and edge devices. Google Brain team (a group of engineers and scientists) is the one that developed the TensorFlow library. It supports machine learning, deep learning, and flexible numerical computation [32]. The following diagram shows the power score of deep learning framework based on usage, popularity, and interest [33].

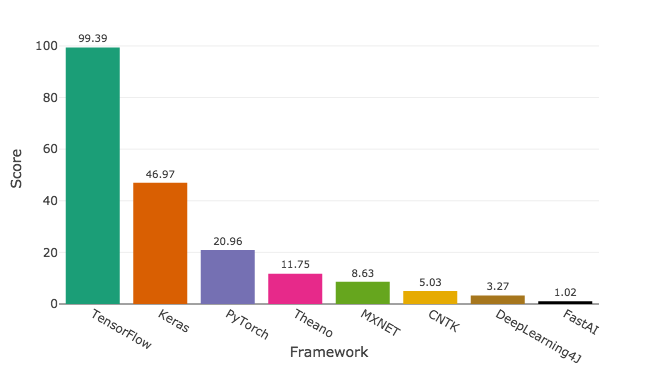


Figure 11 Deep Learning Framework comparision..

As shown in the above diagram, Tensorflow is by far the most used and popular deep learning framework.

### Keras

Keras is a high-level API built on TensorFlow (and can be used on top of Theano too). Compared to TensorFlow it is more user-friendly and easier to use. Keras is user-friendly, supports **Modularity, Easy extensibility, and works with Python [34]**. Keras has the following properties.

* It allows easy and fast prototyping
* Supports both convolutional networks and recurrent networks, as well as combinations of the two.
* It is user-friendly, modular, and extensible.
* It can run runs seamlessly on CPU and GPU.

Keras is the 2nd most popular and used deep learning framework according to the above diagram.

### OpenCV

OpenCV is an open-source computer vision and machine learning software library built to provide a common infrastructure for computer vision applications [35]. It has Python, Java, C++, and MATLAB interfaces and also supports almost every operating system. OpenCV was made for image processing, so each function and data structure were designed with the image processing engineers in mind.

### MATLAB Deep Network Designer

MATLAB deep network designer [36] 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 specially for complex network deep architectures and GAN networks. This even further helps Developers to track and debug errors on the premature design stage.

## Baselines

To verify the effectiveness of our model, we compare it with models that focus on video translation with GANs. Since our model architecture is based on Recycle-GAN and take unpaired video data as input, we choose CycleGAN [8] and Recycle-GAN [11] as the baselines of our experiments.

– **CycleGAN** [8] uses two generators to translate images, with the idea of cycle consistency. We employ it to translate the video frames and make comparisons, with the intent to ﬁgure out the effect of the spatio–temporal constraint.

– **Recycle-GAN** [11] uses two generators and two predictors to translate videos. It puts forward a recycle loss to work with cycle loss and recurrent loss for content translation and style preservation, which takes the temporal information into consideration. The contrasts with Recycle-GAN aim to demonstrate the signiﬁcant improvements that our method makes about the spatial–temporal information

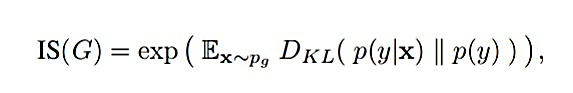
### Evaluation methods

The result will be analyzed to describe the performance of video to video translation model 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 succeeding is qualitative analysis metric to evaluate this work.

#### User study

This evaluation method uses volunteers to assess whether the given video is real or fake after he/she sees random real and fake videos to determine whether or not the generated data is any good. The normal score value is evaluated as per the figure of entities.

#### 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, or IS for short, measure 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 about IS discussed in the appendix.

# CHAPTER FOUR

# The Proposed work & it’s Implementation



## 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, so with the purpose of achieve this objective image to image translation, temporal information extraction and Spatio-temporal information fusion are main building stones. This chapter can be ideally portioned to three major sections, the first introduces image to image translation (spatial domain transfer) to translate a given domain image into another domain. the second deals with video to video translation problems and constraints to solve. The last section explains how temporal information extracted and a fusion between spatial-temporal fusion.

## Introduction

We set out to study the question if it is even possible to embed additional constraint into the video to video translation latent space the results coherent video. let’s recoup some recent approaches a bit about how *constraints* Influence temporal inconsistency problems in a video to video translation. Generally, we can classify those attempts into two various classes. the first-class use conditional GAN [9] to generate a new temporally consistent frame using frame prediction. one of the recent works based on this mechanism [11] use temporal predictor based on pix2pix [7] architecture and cycle-gan neural network. the idea is temporal predictor network is trained to learn the relation between streamed previous frames, so the network could predict how the objects in a frame flow to new generated synthetic frame, however, this approach fails when the ***new*** object introduced in the frame. Since, the temporal predictor uses previous frame it cannot estimate how the new introduced object in the frame should flow in the current frame. The second approach extends the previous one by introducing optical flow-based frame warping [12], [19], [29], [37]. The intuition of this approach is a frame (generated frame) is warped one optical flow between consecutive frame (e.g. and ). Recent work [37] Generator network transfer input frames to output frames . Alongside they compute the optical flow between the neighboring frames (e.g. ). Then is warped using to produce to align . Implicitly the network transfers spatial information from domain to domain and using optical flow it learns how to transfer motion from domain to domain. [19] extend by further introducing global temporal constraint.

Inspired by [11],[12] and [37] this thesis work proposes to add an additional temporal transfer from domain to a domain by introducing motion cycle consistency constraint. In the succeeding section, let us discuss the proposed architecture.

Generally, this work proposes to add the following things

## Unpaired Image to image transfer

For generating a fake video from translating one domain video to another domain. Spatial image Translator network is important. This network takes a sequence of images as an input. To do so the input video should change to the sequence of the frame using frame extractor network.

One thing to keep in mind is that the translated image should preserve **contain information** but possibly the **style information** would be changed and it should be closed to as real image in another domain. The translator network should consider this constraint while learning in training. In this research work CycleGAN [8] is used as an image to image translation network, which gives an advantage of unpaired training manner.

## Cycle-GAN

After the genius insight of original GAN work by Goodfellow et.al [4] various creative work immerges. Cycle-GAN is one of them perhaps this work is an ice breaker that immerses pair dataset for the domain to domain translation. To achieve an unpair image to image translation cycle-GAN introduce cycle consistency constrain or cycle loss in short with vanilla GAN loss or GAN loss.

### Cycle Loss

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. This statement could be formulated as:

### Identity Loss

Identity loss is introduced to diminish translation of the images already in domain A to the Generator from , because the CycleGAN should understand that they are already in the correct domain. This means translating Amharic text to Amharic using English to Amharic translator since the input is Amharic the network should make no change.

So, the hole loss would be:

***where***: are generators,,and are discriminators respectively both domain and 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 also extended by for video translation.

## Problems in Translation Networks

Highly expressive parametric models have achieved great success in machine learning [38], where learning goals and measurement measures are typically well defined and easily calculated. 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 are the most common problems for video translation.

Let say we have two generators and to translate from one domain to another domain and two discriminatorsand , where trained to translate from to and from to . and discriminators and to classify between real and fake in both domains. Video and are sample videos from respective domain and . where are ith frame of video . Each frame may contain various number of objects.

1. **Object disappearance**: is a problem object in a given video frame in domain shall also appear in translated appear in another domain image. meaning if a car appears in is should also appear in Mathematically,
2. **Object dislocation**[[5]](#footnote-5): happen when an object in frame from domain changes its position, when translated in domain . Object dislocation also, can be seen as abrupt object movement. Mathematically,
3. **Artifacts**: An image frame artifact is any element that occurs in the picture that is not present in the initial picture set.
4. **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.

The problems described above are appropriate for the problem of translation, where only spatial transformation is considered. For an approach of mixing spatial and temporal learning functions, this thesis work gets a better result.

## Solutions

## 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., = 0) only if the inverse mapping recovers the original contents, regardless of how incorrect the forward mapping was.

By adding the above loss, we inspire the video frames to have the consistent semantics earlier and afterward the translations. This thesis work uses MobileNetV2 [39] as a feature extractor which enforces the content information appears in original image also should appeared on translate. as an example, if a person and a dog appear in image A so does in translated image A2B albeit the style modified.

## Temporal information extractor.

In Section 2.4.1. of this paper we talk about how video to video translation extends image to image translation by imposing temporal information to spatial domain to create more temporal coherence video. we also discuss using Flownet 2.0 by E. Ilg et.al. [20] temporal extractor which belongs to optical flow-based temporal information extractor. Even we discussed various options for temporal information extraction in section 2.4, optical flow based is appropriate for this work because both flower datasets and the Obama trump dataset contain less dynamic scene video[[6]](#footnote-6) and very object-centric. Figure 10 shows a clearer view of the input and output of temporal information extractor. Detail of flownet 2.0 architecture specification is described in detail in the appendix.

The next section would discuss how to blend those spatial and temporal pieces of information to test the proposed hypothesis.

Figure 13 Flownet 2.0 optical flow extraction

input image

flownet2



Optical flow between frames

## Spatial-temporal information fusion

In the previous session, we talk about how valuable temporal information is for a video to video translation, but the question lay how we Blend temporal information to spatial one to generate better consistent video. To impose temporal information two approaches have been implemented in this work. The first one is warping the translated and the later use temporal consistency discriminator network. Where the first implemented inside Generator network the later one is discriminators job.

Warping: Image warping algorithms are in a widespread class of nonlinear image transforms, correlated to the spatial component of image[40]. They are applied to the correction of image distortions, morphing, and special 2D eﬀects. A formal way of definition could be modifying pixel location based on the calculation of the pixel’s coordinate in the destination image using the source pixel coordinates.in our case, it can mathematically be described as:

Let is an optical flow between successive frames , translated image in y domain which equals to and is a function which alters translated image base on optical So

Where warped translated image and it is temporally consistent.

We use simple binary interpolation warping. even if the warping is a good way of fusing temporal information but It basically fails in two common cases

1. when a new object is introduced in a frame and,
2. when an object in the previous frame occluded by an object in a frame.

## Bring all together (Proposed Work)

In the previous sections, we see the various part of this thesis work in discrete means. Now in this section let’s put them all together so that we can see the network architecture as one big network.

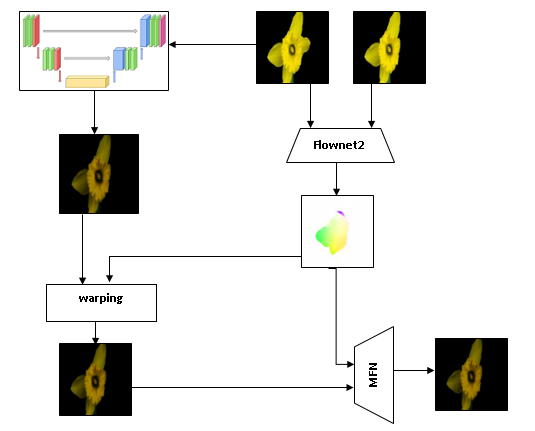


Figure 15 Generator Network Architecture.

So, the loss function of the network would be:

***Where***:

To improve visual quality further, motivated by [41] we implement discriminator which takes consecutive images to decide its real or fake. The architecture and the output stay the same with Patch GAN instead the differences are just the input and the number of channels meaning rather that of differentiating between single frames, discriminator design it in a way that it observes two constitutive of synthesized frames and relates them with two constitutive of the real frames which make it ideal since the discriminator takes into account a temporal aspect of the video generation problem.

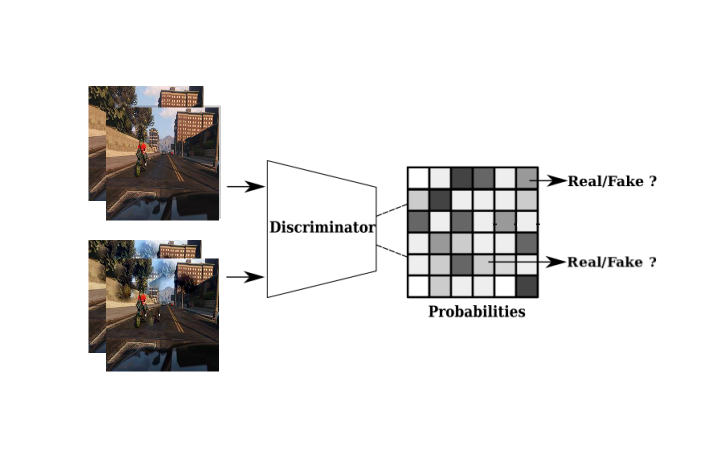


Figure 16 Discriminator Architecture

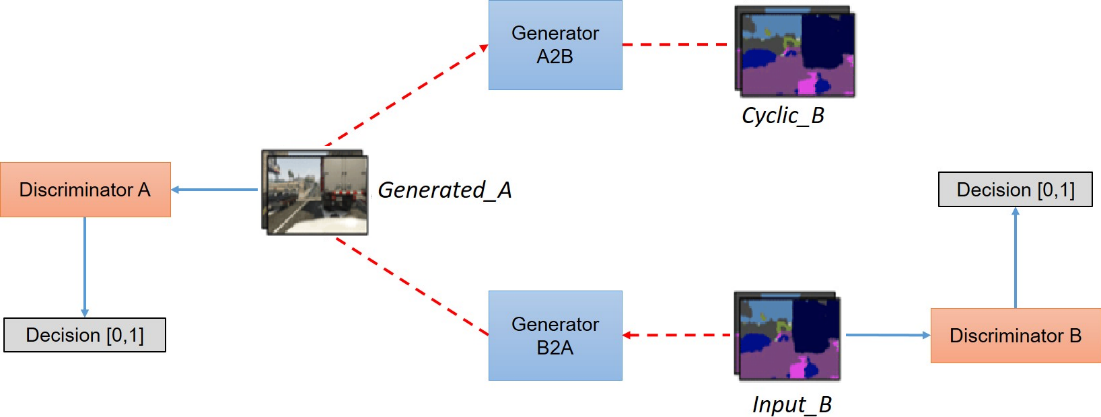
Even the discriminator network has minor differences with cycle-GAN architecture the rest is quite the same. This thesis work whole architecture is shown in the figure below.

Figure 17 Cycle-GAN Architecture

pseudo code for training would look like

while epoch

for each image in Dataset

A <= read pair of images from domain A

B <= read pair of images from domain B

f <= compute optical\_flow A and A-1

A2B <= warp A2B using f

A2B <= Translate A to B Domain // GAN loss

B2A <= Translate B to A Domain

A2A <= Translate A to A Domain

B2B <= Translate B to B Domain

A2B2A <= Translate A2B to A Domain

B2A2B <= Translate B2A to B Domain

evaluate A2B and B2A are real or fake

compute loss between A2B2A and A //cycle loss

compute loss between B2A2B and B

update loss

epoch = epoch + 1

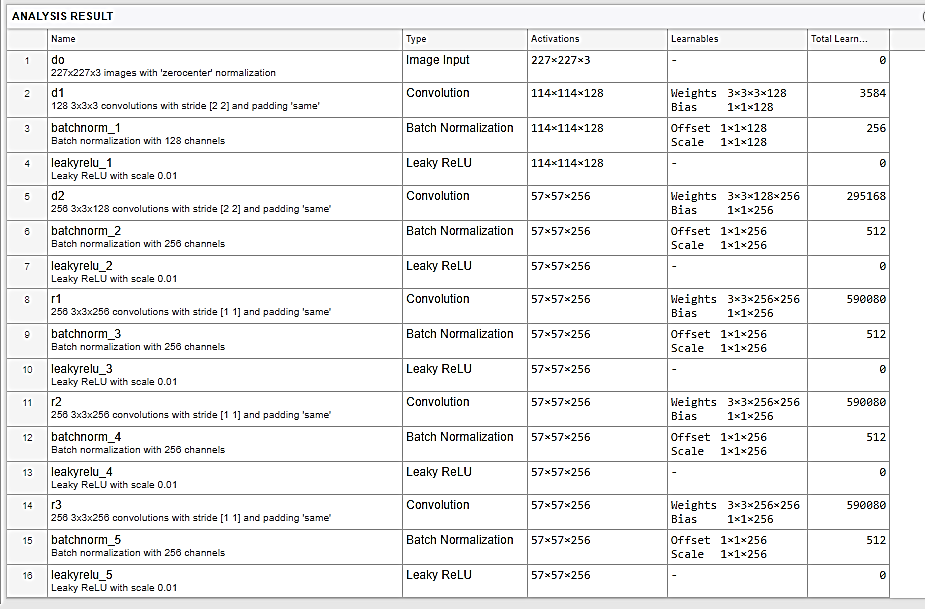
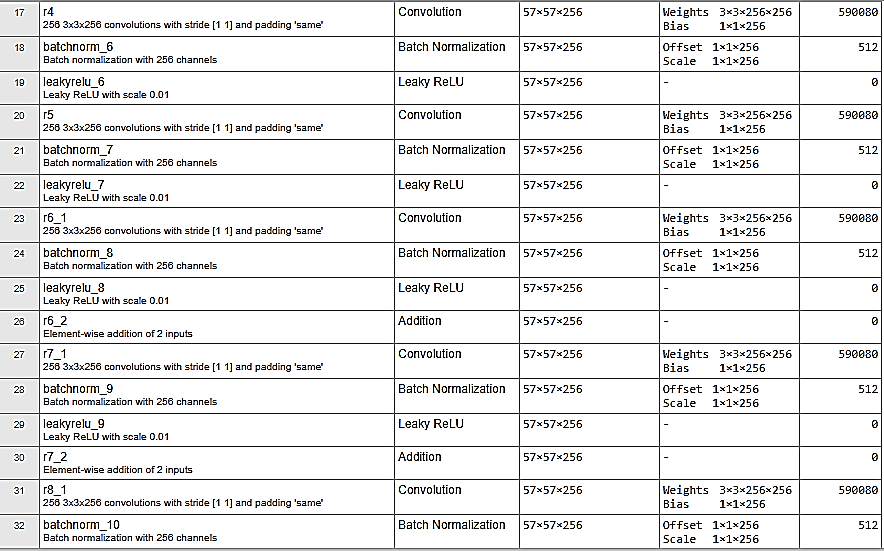
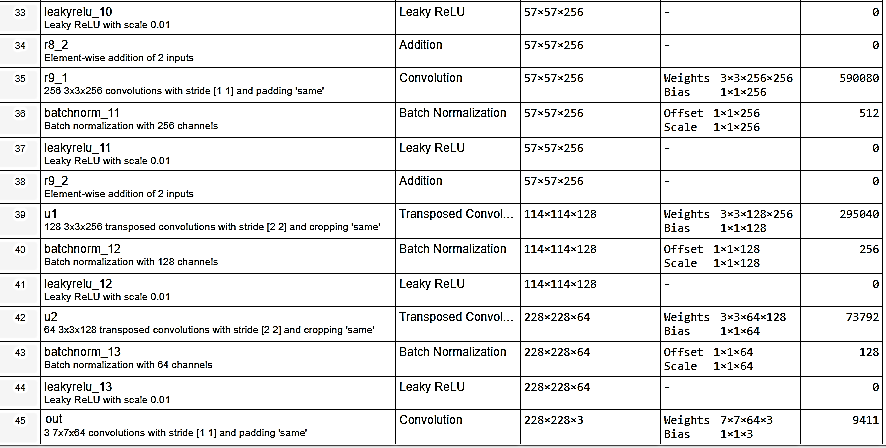


Table 4 Cycle GAN architecture





## Working Environment.

In this section explain 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.
  + Operating system: Windows 10
  + Processor: Intel ® Core™ i7-2300QM CPU @ 2.00GHz
  + Graphics: Intel ® Graphics 3000
  + Primary Memory (RAM): 8.00 GB
  + System Type: 64-bit Operating System, x64-based Processor
* Desktop: The desktop computer is used for developing a video for video translation.
  + Operating system: windows 10
  + Processor: Intel ® Core™ i5-4580 CPU @ 3.29GHz x 4
  + Graphics: Intel ® HD Graphics 4600
  + GPU: GeForce RTX 2070 Super 6 GB RAM
  + Primary Memory (RAM): 14.00 GB
  + 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. and TensorFlow-GPU 2.1 used. In the next section list of experiments class conducted for evaluating the proposed hypothesis are discussed.

## Set upping the environments

Various applications and IDEs are used for this work.

* + 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.

## CycleGAN specification

CycleGAN[[7]](#footnote-7) implemented through TensorFlow Keras, was used during the experimentation, there are two different implementations available ﬁrst was a CycleGAN used on 9 residual blocks which used in this work, second, consist 6 residual blocks. Those two networks differ on the number of residual blocks in their architecture.

CycleGAN comprises of two identical Generator and two identical Discriminator networks since we have two domains. 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 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. For further detail check table 4.3.1.

The discriminator had 70 x 70 PatchGAN. this network composed of 5 convolutional layers denotes a 4 × 4 Convolution-Instance Normalization with LeakyReLU layer and stride 2. After the last layer, apply a convolution to produce a 1-dimensional output. We do not use Batch Normalization for the ﬁrst layer. The slope of leaky in leakyReLU was 0.2. see table 4.4.3 for more detail.

Weights in convolutional layers were initialized with a truncated normal distribution initializer with a standard deviation of 0.02, all other layers used a random normal initializer with a standard deviation of 0.02. All biases were initialized to 0. The decay for the moving average for the batch instance normalization was set to 0.9, the epsilon was set to 10−5. Every network used ADAM optimizer with the momentum term . Additionally, the loss functions were,

In the LeakyReLU activation, the gradient of the leak was set to 0.2. Lastly, the training process was balanced by making two training steps for the generator for each training step made for the discriminator. Most of the conﬁgurations were adopted from the CycleGAN paper found and from implementations of CycleGAN by its authors in GitHub.

## Implement cycle GAN

A CycleGAN is made up overall of two architectures: a generator and a discriminator. The generator architecture is used to create two models, Generator AB and Generator BA. The discriminator architecture is used to create additional two architectural models, Discriminator A, and Discriminator B.

The generator network is an encoder-decoder category of network. It takes an image as an input and outputs another image. It has two parts: an encoder and a decoder.

G\_A2B = module.ResnetGenerator(input\_shape=(crop\_size,crop\_size, 3))

G\_B2A = module.ResnetGenerator(input\_shape=(crop\_size,crop\_size, 3))

The discriminator network is equivalent to the architecture of the discriminator in a PatchGAN network[7]. Basically, it takes an image of the shape of (256, 256, 3) and predicts whether the image is real or fake.

D\_A = module.ConvDiscriminator(input\_shape=(num\_img, crop\_size,crop\_size, 3))

D\_B = module.ConvDiscriminator(input\_shape=(num\_img, crop\_size,crop\_size, 3))

The general architecture composed of the above four independent networks. The objective of Generator is to diminish the adversarial loss function against an adversary Discriminator, which constantly tries to maximize it. Differently from the other network it takes consideration of two images (num\_img = 2) generated from generator to decide fake or real. As discussed in section 4.6 above.

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])

similar to other network types GAN is no different. Learning function has to explicitly defined in order to the network learns to translate image.

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

## Temporal information extraction Implementation

Section 4.4 Discuss about extraction mechanism of temporal information used in this paper briefly. Eventually this section present python implementation. As discussed this work use flownet2 as temporal extractor [20].

A2B = G\_A2B(A, training=True) #translated image of current frame.

f = flownet2(A,prev)# compute optical flow between successive frames.

prev = A #set A as previous frame for next iteration.

A2B = insertTemporalInformation(A2B,f) #warp next frame using f

A2Bprev = A2B # for temporal aware discriminator network

This work also uses additional temporal aware discriminator network.

A2B\_d\_logits = D\_B(A2B, A2Bprev, training=True)

And cycle loss of temporal aware network would be

A2B2A\_cycle\_loss = cycle\_loss\_fn(A, A2B2A)#A2B2A is retranslate image.

## Experiment Class

To evaluate the essence of temporal information for video translation testing an initial hypothesis is mandatory, to do so three different classes of experiments are conducted as shown below on the table for each dataset group.

The first class is all about classical Cycle GAN image translation on a given sequence of images Which taking into consideration spatial domain only. The second regarding consider using temporal consistency aware discriminator network.

Table 5 lists of experimental classes.

|  |  |  |
| --- | --- | --- |
| class | Experiment | Model used |
| 1 | Frame wise video translation (spatial domain) | Cycle GAN. |
| 2 | Temporal information Consider. | Cycle GAN and flownet2. |
| 3 | Optical flow temporal information Blend With NOAFMN | Cycle GAN, flownet2, & temporal aware Discriminator. |

.

# CHAPTER FIVE

# Evaluation, Results, and Discussion



## Overview

This chapter presents the evaluation of the video to video translation and the integration of temporal information to improve flickering output by the previous approach. It also discusses the result by comparing it with other related works.

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 in relation to the principal research question posed in this thesis: *How to preserve temporal consistency for a video to video translation? And this thesis work proposes a hypothesis that “adding temporal consistency constrain would improve temporal consistency between successive frames”*

## Video to video Translation

The video to video translation takes an image from the scene as an input a generate an equivalent image in other domain with the consideration preserving temporal information. The image dataset contains 180,030 images from the 88,132 datasets. The full set is torn apart into training and testing dataset for training and testing the deep learning architecture. we randomly sampled images from this dataset as training data and testing data Table below show sample of dataset. The dataset contains images of one object or multiple objects. This work conducts three training classes as discussed in section 4.8. The first model is trained on cycle-GAN architecture with considering spatial information only to translate to other domain so that it can specialize to translate images from domain to domain.

In this research work we use inception score (IS) and human study to evaluate the experimental outcome.

|  |  |  |
| --- | --- | --- |
| Environment | Sample1 | Sample2 |
| Rain |  |  |
| Snow |  |  |
| Sunset |  |  |
| Night |  |  |
| Day |  |  |

# CHAPTER SIX

# Conclusion and Future work

# References

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

[2] G. E. H. Alex Krizhevsky, Ilya Sutskever, “ImageNet Classification with Deep Convolutional Neural Networks,” *ILSVRC2012*, pp. 1–1432, 2007.

[3] 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.

[4] I. Goodfellow *et al.*, “Generative Adversarial Nets (NIPS version),” *Adv. Neural Inf. Process. Syst. 27*, 2014.

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

[6] 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.

[7] 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*, 2017, vol. 2017-Janua, pp. 5967–5976.

[8] 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*, 2017, vol. 2017-Octob, pp. 2242–2251.

[9] M. Mirza and S. Osindero, “Conditional Generative Adversarial Nets,” Nov. 2014.

[10] 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.

[11] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, “Recycle-GAN: Unsupervised Video Retargeting,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11209 LNCS, pp. 122–138.

[12] 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.

[13] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016.

[14] 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*, 2017, vol. 2017-January, pp. 5967–5976.

[15] “Image-to-Image Translation: Machine Learning Magic that Converts Winter Photos Into Summer - Abto Software, Lviv, Ukraine.” [Online]. Available: https://www.abtosoftware.com/blog/image-to-image-translation. [Accessed: 03-Mar-2020].

[16] 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.

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

[18] 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*, 2017, vol. 2017-Octob, pp. 1114–1123.

[19] 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.

[20] 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*, 2017, vol. 2017-Janua, pp. 1647–1655.

[21] 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*, 2018, pp. 8934–8943.

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

[23] 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*, 2018, pp. 8934–8943.

[24] 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.

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

[26] 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.

[27] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “Animating Arbitrary Objects via Deep Motion Transfer,” 2018.

[28] D. Bashkirova, B. Usman, and K. Saenko, “Unsupervised Video-to-Video Translation,” no. Nips, 2018.

[29] 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.

[30] 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.

[31] “Edraw Max - Excellent Flowchart Software & Diagramming Tool.” [Online]. Available: https://www.edrawsoft.com/edraw-max/. [Accessed: 02-Jun-2020].

[32] “TensorFlow.” [Online]. Available: https://www.tensorflow.org/. [Accessed: 02-Jun-2020].

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

[34] “Keras API reference.” [Online]. Available: https://keras.io/api/. [Accessed: 02-Jun-2020].

[35] “OpenCV.” [Online]. Available: https://opencv.org/. [Accessed: 02-Jun-2020].

[36] “Design, visualize, and train deep learning networks - MATLAB.” [Online]. Available: https://www.mathworks.com/help/deeplearning/ref/deepnetworkdesigner-app.html. [Accessed: 05-Jun-2020].

[37] S. Jiang, H. Liu, Y. Wu, and Y. Fu, “Spatially Constrained Generative Adversarial Networks for Conditional Image Generation,” pp. 1–20, 2019.

[38] T. R. Shaham, T. Dekel, and T. Michaeli, “SinGAN: Learning a Generative Model from a Single Natural Image,” May 2019.

[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 4510–4520, Jan. 2018.

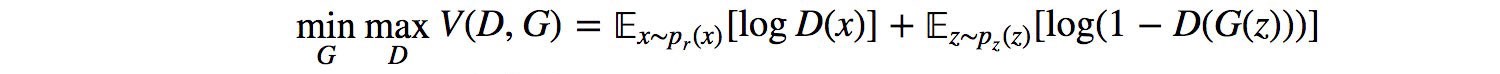
[40] P. Mazurek, “Optimization of memory accesses for CUDA architecture and image warping algorithms,” in *Advances in Intelligent Systems and Computing*, 2013.

[41] P. Zablotskaia, “Generative adversarial networks for pose-guided human video generation,” 2020.

# Appendix

Appendix A: GAN loss function proof

We have mentioned optimizing GAN is optimizing JS-divergence. This is not obvious from the cost function:



**p = q**,

Appendix B: Flownet 2.0 specification

Appendix C: GAN loss function proof

Appendix D: GAN loss function proof

Appendix E: GAN loss function proof

1. Some authors see GAN in other perspective rather adversarial: collaboration of two networks to Mimic a give real data distribution. [↑](#footnote-ref-1)
2. conditioning variable *C* could be any type of information. Like Image, tabular information or…. [↑](#footnote-ref-2)
3. only significant related approach for this kind of problem discussed (regarding to temporal information extraction). [↑](#footnote-ref-3)
4. These present papers are only substantial papers that directly relate to thesis work, and all are from 2017 onwards. [↑](#footnote-ref-4)
5. Object disappearance and object dislocation in a situation like face to face translation wouldn't be a question. [↑](#footnote-ref-5)
6. 3D convolution tensor and RNN based approaches need very high efficiency machines, where pose estimation works fine when salience object is human. [↑](#footnote-ref-6)
7. In this paper when we say CycleGAN it includes ***identity loss*** even though it doesn’t include in original work. [↑](#footnote-ref-7)