

# **CP Capstone Project Final Report**

## **ข้อเสนอโครงการนวนิยายอดวิศวกรรมคอมพิวเตอร์ 2**

**เรื่อง**

**การรู้จำใบหน้าที่ต้านการปลอมแปลง**

**Anti-Spoofing Face Recognition**

**โดย**

นายธีรชัย อินทรคง รหัสนิสิต 6130268621

นายนัสรี อิสลาม รหัสนิสิต 6130287521

นายบวรวิช ลิมปวิทยากุล รหัสนิสิต 6131022721

นายสิริธิษณ์ คักดีพิบูลย์จิตร์ รหัสนิสิต 6131048021

**อาจารย์ที่ปรึกษา**

ดร.ดร.ธนารัตน์ ชลิตาพงศ์

ดร.ธนาพ กอบชัยสวัสดิ์

**รายงานนี้เป็นส่วนหนึ่งของวิชา 2110488 โครงการนวนิยายอดวิศวกรรมคอมพิวเตอร์**

## บทคัดย่อ

ระบบตรวจจับและรู้จำใบหน้าได้รับการพัฒนาอย่างต่อเนื่อง ในปัจจุบันเริ่มมีการใช้งานอย่างแพร่หลาย เช่น การใช้ระบบรู้จำใบหน้าในการยืนยันตัวตนในการทำธุกรรมทางการเงิน หรือการสแกนหน้าเพื่อควบคุมการเปิดปิดประตู เป็นต้น อย่างไรก็ต้องการยืนยันตัวตนด้วยใบหน้ายังมีจุดอ่อนที่เกิดขึ้นได้ เช่น การปลอมแปลงใบหน้า โดยการแสดงรูปภาพใบหน้าบุคคลอื่นทางหน้าจอของโทรศัพท์มือถือ หรือการใส่หน้ากากปลอมแปลงเป็นบุคคลอื่น ซึ่งอาจส่งผลให้เกิดความเสียหายต่าง ๆ ได้ จึงได้มีการศึกษาเครื่องมือที่จะใช้ตรวจจับการปลอมแปลงนี้ แต่ว่าเครื่องมือที่มีอยู่ในปัจจุบัน ยังมีการใช้ทรัพยากรการคำนวณในการแยกแยะใบหน้าปลอมแปลงมากเกินไป ในโครงการนี้จึงได้เสนอเครื่องมือที่จะตรวจจับการปลอมแปลงใบหน้า โดยลดปริมาณทรัพยากรที่ใช้ให้ได้น้อยที่สุดโดยที่ยังรักษาประสิทธิภาพการทำงานเอาไว้ได้

โครงการนี้ได้ทำการแบ่งเป็น 2 ปัจจัยอย ได้แก่ การป้องกันการโจมตีแบบ 2 มิติ (2D Attack) และการป้องกันการโจมตีแบบ 3 มิติ (3D Attack) โดยการป้องกันการโจมตีแบบ 2 มิติ ได้ใช้วิธีการสร้างแผนที่ความลึก (depth maps) จากภาพใบหน้าเพื่อทำการแยกแยะระหว่างใบหน้าจริงและการปลอมใบหน้าบุคคลอื่น โดยผลลัพธ์ของการทดลองยังอาจทำได้ไม่ดีในสภาพแสงและสภาพแวดล้อมบางรูปแบบ ส่วนการป้องกันการโจมตีแบบ 3 มิติ ได้ใช้สัญญาณซีพาระเบี่ยนสีของผิวหน้าคนในการแยกแยะระหว่างใบหน้าจริงและการโจมตีในรูปแบบการใส่หน้ากากเข้าหากันกล้อง โดยผลการทดสอบพบว่าการโจมตีที่สามารถป้องกันได้ดีนั้นจะเป็นการโจมตีที่ผู้โจมตีทำการใส่หน้ากากแบบเต็มบริเวณใบหน้า ในพื้นที่ที่แสงนั้นไม่มากจนเกินไป

## **Abstract**

Nowaday, face detection and recognition systems have been developed and widely used, for example, face recognition in financial transactions, face recognition for door access control. However, identity verification using a face recognition system has its weaknesses, such as showing another person's picture on a smartphone screen or wearing a mask. There have been many researches aiming to detect these kinds of spoofing but many methods require too much computation resource to accomplish this task. This project aims to detect face spoofing which requires the least computational cost as possible while preserving its performance.

This project is divided into two sub projects which are 2D attack detection and 3D attack detection. The 2D attack detection generates the depth map to distinguish between real face and attack. Experimental results show that our method does not provide good results in some light conditions and environment. The 3D attack detection uses the vital sign of the face that changes overtime to distinguish between real face and mask attack. Our method provides good results on full mask attacks with not too much light.

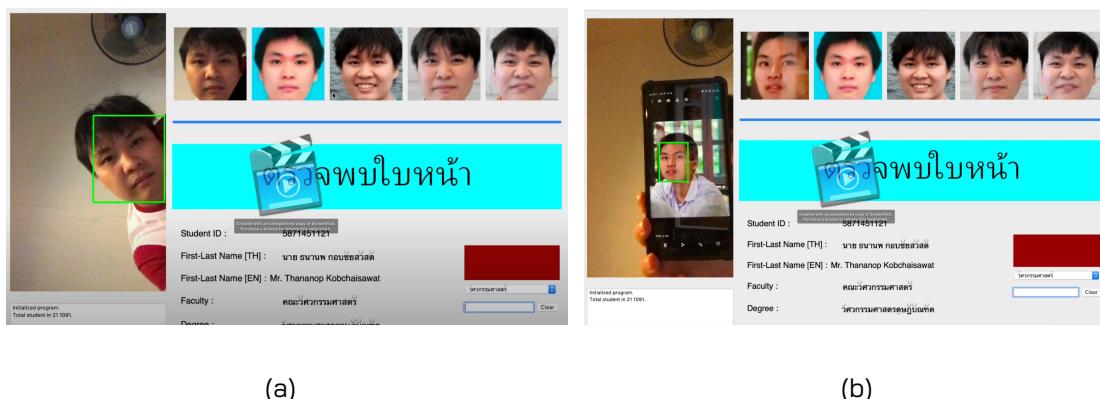
## บทที่ 1

### บทนำ

#### 1.1 ที่มาและความสำคัญ

การรู้จำใบหน้า (Face recognition) คือ กระบวนการระบุหรือยืนยันตัวตนของบุคคลโดยใช้ใบหน้า โดยนำข้อมูลลักษณะจำเพาะของใบหน้ามาเปรียบเทียบกับข้อมูลบนฐานข้อมูล เพื่อระบุว่าใบหน้าที่ตรวจจับนั้นตรงกับบุคคลใด โดยใช้อัลгорิทึมต่าง ๆ ในการวิเคราะห์องค์ประกอบต่าง ๆ บนใบหน้า เช่น คิ้ว ตา ปาก และริมฝีปาก เป็นต้น ปัจจุบันการรู้จำใบหน้านี้เป็นหนึ่งในเทคโนโลยีที่ได้รับความนิยมสูงที่สุดในระบบบรักษาความปลอดภัย และถูกใช้อย่างแพร่หลาย เช่น ระบบชำระเงินผ่านโทรศัพท์มือถือ และการยืนยันตัวตนเพื่อเข้าอาคารสำนักงาน เป็นต้น เนื่องจากความสะดวกในการใช้งาน ความมีเอกลักษณ์เฉพาะของบุคคล และความแม่นยำสูง

ระบบรู้จำใบหน้า (Face recognition system) ถูกนำมาใช้เป็นระบบตรวจสอบนักศึกษาที่เข้ารับปริญญาภายในงานรับปริญญาของจุฬาลงกรณ์มหาวิทยาลัย ระบบดังกล่าวถูกจัดทำในชื่อ โปรเจ็คไร้จัน (Project right one) ซึ่งเป็นระบบรู้จำใบหน้าที่มีประสิทธิภาพสูง โดยให้ความแม่นยำ (accuracy) ที่ 98% อย่างไรก็ตาม แม้ว่าระบบรู้จำใบหน้าของโปรเจ็คไร้จันสามารถใช้งานได้ถูกต้องและมีประสิทธิภาพในงานรับปริญญา แต่อาจเกิดปัญหาหากต้องการพัฒนาเพื่อนำไปใช้ในสถานการณ์อื่น ๆ เนื่องจากระบบรู้จำใบหน้าดังกล่าวไม่มีความสามารถในการแยกแยะว่า ข้อมูลใบหน้าที่เข้ามานั้นเป็นใบหน้าจริงของมนุษย์หรือการปลอมแปลงใบหน้า ทำให้เกิดช่องโหว่ต่อการปลอมแปลงชีวมิติ (Presentation attack) เช่น การโจมตีด้วยรูปภาพบุคคล (Photos attack), การโจมตีด้วยการย้อนเล่นวิดีโอบุคคล (Replay attack) และการโจมตีด้วยหน้ากากบุคคล (Mask attack) เป็นต้น ดังนั้นการพัฒนาเทคโนโลยีการตรวจจับและการต่อต้านการปลอมแปลงใบหน้าจึงเป็นส่วนสำคัญต่อการพัฒนาระบบตรวจสอบใบหน้า เพื่อให้สามารถนำไปใช้งานได้ในหลายสถานการณ์ โดยเฉพาะสถานการณ์ที่สูมเสียงต่อการโจมตีด้วยวิธีการปลอมแปลงใบหน้า



รูปที่ 1 (a) แสดงผลลัพธ์จากการรู้จำใบหน้าของใบหนามนุษย์จริงของโปรเจ็คไร้จัน (b) แสดงผลลัพธ์จากการรู้จำใบหน้าของการปลอมแปลงใบหน้าของโปรเจ็คไร้จัน

การต่อต้านการปลอมแปลงใบหน้า (Face anti-spoofing) คือ กระบวนการในการตรวจจับ การป้องกัน และการต่อต้านใบหน้าเท็จ (False facial verification) ซึ่งเกิดจากความพยายามในการสร้าง การตัดแปลง หรือการปลอมแปลงลีสิ่งต่าง ๆ ให้มีลักษณะคล้ายกับใบหน้าของบุคคลอื่น รวมถึงการปกปิดหรือแก้ไขบางส่วนของใบหน้าตนเองเพื่อไม่ให้ระบบตรวจจับใบหน้าสามารถระบุตัวตนได้ เช่น การลักยันต์ หรือแต่งหน้า เป็นต้น

อย่างไรก็ตามอัลกอริทึมสำหรับการต่อต้านการปลอมแปลงใบหน้าทั่วไปนั้นใช้เวลาและทรัพยากรการคำนวณที่สูง ซึ่งเป็นอุปสรรคต่อการนำไปประยุกต์ใช้ในสถานการณ์ทั่วไป เช่น การประยุกต์ใช้กับการยืนยันตัวตนสำหรับนักศึกษาเพื่อเข้าห้องเรียน รับปริญญา ซึ่งต้องการความรวดเร็ว เพื่อให้การเคลื่อนตัวของนักศึกษาที่เข้าร่วมงานเป็นไปอย่างรวดเร็ว การประยุกต์กับการยืนยันตัวตนเพื่อเข้าอาคารสำนักงาน (Access control system) ซึ่งต้องการความรวดเร็ว และบางกรณีอาจมีข้อจำกัดด้านทรัพยากรการคำนวณและพลังงาน เป็นต้น ไม่เพียงเท่านี้อัลกอริทึมสำหรับการต่อต้านการปลอมแปลงใบหน้าทั่วไปอาจต้องการเซนเซอร์เฉพาะทาง เช่น เซนเซอร์วัดความร้อน เซนเซอร์วัดความลึก และเซนเซอร์โพลาไรเซชัน เป็นต้น ซึ่งอาจเป็นอุปสรรค ต่อการประยุกต์ใช้กับอุปกรณ์ที่มีอยู่ทั่วไปซึ่งไม่ได้ผังเซนเซอร์ระดับสูงเหล่านั้น



รูปที่ 2 การปลอมแปลงชีวมิติ (Presentation attack)  
ที่มา [1]

จากเหตุผลดังกล่าว การพัฒนาเทคโนโลยีต่อต้านการปลอมแปลงที่ใช้เวลาน้อย ต้องการทรัพยากรต่ำ แต่ให้ประสิทธิภาพสูง และสามารถใช้กับอุปกรณ์ที่ผังเซนเซอร์กล้องอาจจีบีทั่วไปได้ จึงมีความสำคัญในการนำมาประยุกต์ใช้กับโลกปัจจุบัน

## 1.2 วัตถุประสงค์ของโครงงาน

เพื่อสร้างเครื่องมือสำหรับตรวจจับและต่อต้านการปลอมแปลงใบหน้า (Face Anti-Spoofing หรือ FAS) ทั้ง 2 ประเภท (หัวข้อ 1.4) ที่สามารถประยุกต์ใช้กับสถาปัตยกรรมของระบบการรู้จำใบหน้าของโปรเจ็คไพร์ทวัน (Project right-one) โดยเน้นให้ระบบใช้ทรัพยากรต่ำ แต่ยังคงมีประสิทธิภาพสูง และสามารถทำงานได้อย่างรวดเร็ว

### **1.3 ความต้องการของระบบ**

- ระบบจะต้องสามารถแยกแยะระหว่างใบหน้าคนจริงและการโจรตีได้ภายใน 5 วินาที
- ระบบจะต้องสามารถใช้ร่วมกับระบบรักษาใบหน้าของໂປຣເຈັດໄຮ່ວັນໄດ້
- ระบบจะต้องสามารถทำงานได้ด้วยกล้องถ่ายวิดีโอที่มีความสามารถปกติได้ (กล้องวิดีโอອาร์ຈິບ)

### **1.4 ขอบเขตของโครงงาน**

โครงงานนี้มุ่งเน้นการพัฒนาเทคโนโลยีการตรวจจับการปลอมแปลงใบหน้าที่มีลักษณะดังต่อไปนี้

1. การโจรตีทางกายภาพ (Physical presentation attacks) แบบทั้งใบหน้า (Whole region) ซึ่งมีวัตถุประสงค์เพื่อให้ระบุเป็นบุคคลอื่น (Impersonation) โดยเป็นการโจรตีแบบ 2 มิติ (two-dimension)
2. การโจรตีทางกายภาพ (Physical presentation attacks) แบบทั้งใบหน้า (Whole region) ซึ่งมีวัตถุประสงค์เพื่อให้ระบุเป็นบุคคลอื่น (Impersonation) โดยเป็นการโจรตีแบบ 3 มิติ (three-dimension)

### **1.5 ขั้นตอนและแผนการดำเนินโครงงาน**

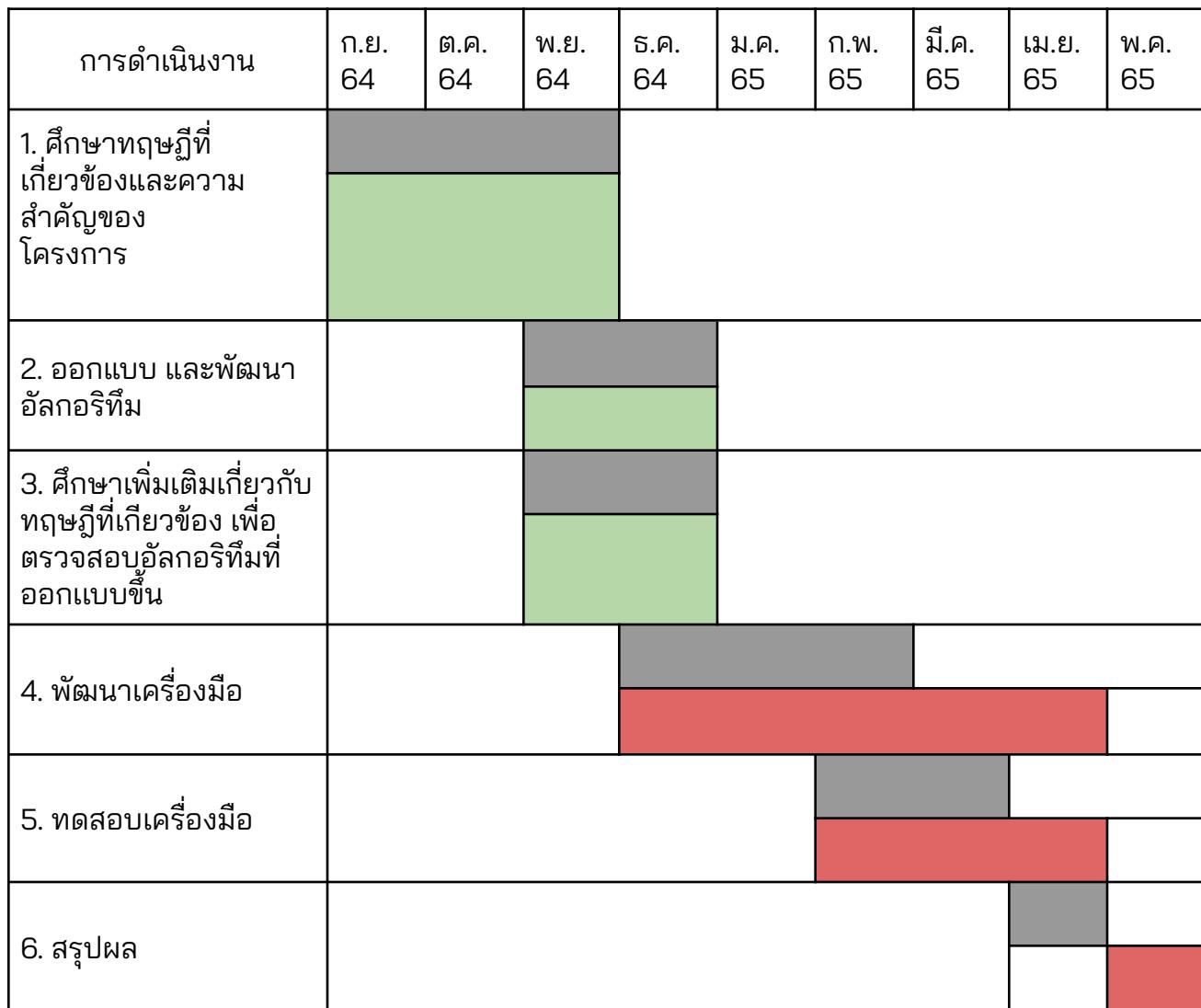
1.4.1 ศึกษาองค์ความรู้ และทฤษฎีที่เกี่ยวข้องกับโครงงาน

1.4.2 ออกแบบและพัฒนา

1.4.3 ทดสอบและวิเคราะห์ประสิทธิภาพของการทดลอง

1.4.4 สรุปผลการทำโครงงาน

ตารางที่ 1 ตารางแผนการดำเนินงาน



- แสดงกำหนดการที่วางไว้
- แสดงสิ่งที่เกิดขึ้นจริง ซึ่งเป็นไปตามกำหนดการที่วางไว้
- แสดงสิ่งที่เกิดขึ้นจริง ซึ่งล่าช้ากว่ากำหนดการที่วางไว้

## **1.6 ทีมงาน**

### **การตรวจจับการปลอมแปลงใบหน้าแบบ 2 มิติ**

1. นายธีรัช อินทร์คง รหัสนิสิต 6130268621

2. นายนัสรี อิสลาม รหัสนิสิต 6130287521

### **การตรวจจับการปลอมแปลงใบหน้าแบบ 3 มิติ**

1. นายบวรวิช ลิมป์วิทยากร รหัสนิสิต 6131022721

2. นายสิริธิษณ์ ดักดิพัญญ์จิตต์ รหัสนิสิต 6131048021

## **1.7 ประโยชน์ที่คาดว่าจะได้รับ**

ระบบฐานข้อมูลของໂປຣເຈັກໄຣຈົວນສາມາດตรวจจับและต่อต้านการปลอมแปลงใบหน้าได้อย่างมีประสิทธิภาพ ซึ่งจะส่งผลให้ระบบมีความปลอดภัยมากขึ้น และสามารถนำไปใช้งานได้หลากหลายสถานการณ์มากขึ้น โดยเฉพาะสถานการณ์ที่มีความสุ่มเสี่ยงต่อการโจรใต้ด้วยการปลอมแปลงใบหน้า

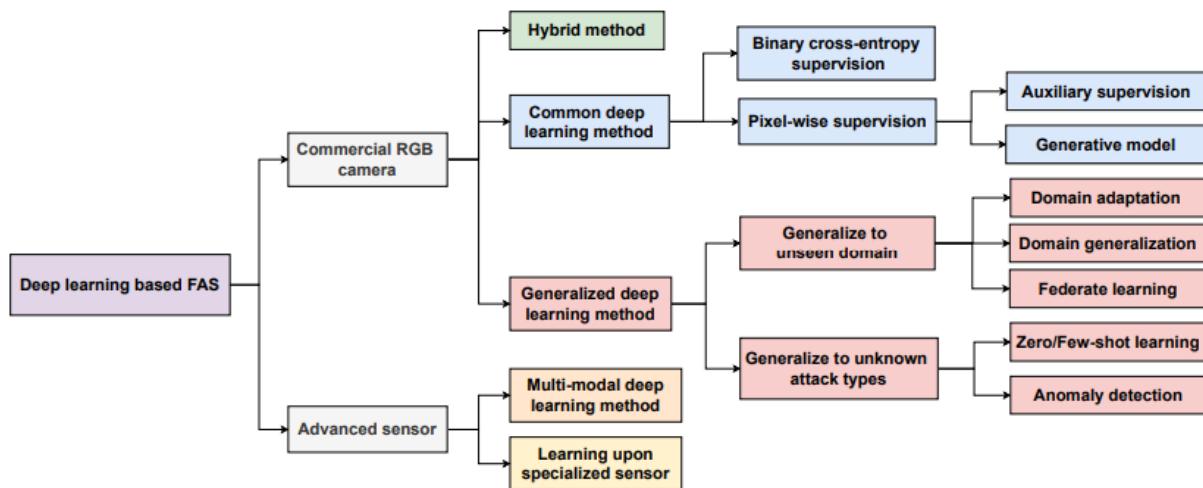
## บทที่ 2

### ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

#### 2.1 ทฤษฎีที่เกี่ยวข้อง

##### 2.1.1 Deep learning based FAS methods

เนื่องด้วยความสมจริงของการปลอมแปลงใบหน้าในปัจจุบัน มีการพัฒนาอย่างรวดเร็วทำให้วิธีการตั้งเดิมในการตรวจจับการปลอมแปลงอย่างการใช้ตัวสกัดคุณลักษณะที่สร้างด้วยมนุษย์ (Handcrafted feature) อาจไม่เพียงพอ ทำให้วิธีการใหม่อย่างการเรียนรู้เชิงลึก (Deep learning) เป็นที่นิยมในงานวิจัยเกี่ยวกับการตรวจจับการปลอมแปลงใบหน้า จากการสืบค้นงานวิจัยใน [1] พบว่า สามารถจำแนกวิธีการตรวจจับการปลอมแปลงใบหน้าได้ดังรูปที่ 2



รูปที่ 3 แผนผังจำแนก Deep learning based FAS methods  
(ที่มา : [1])

###### 1) Commercial RGB camera

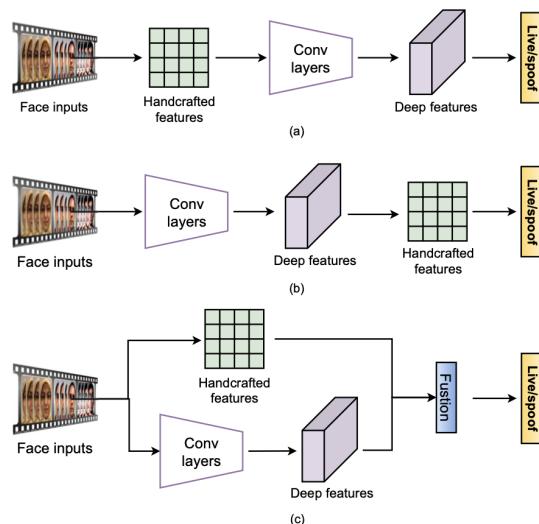
กล้องอาร์จีบีถูกใช้งานอย่างแพร่หลายในระบบตรวจจับใบหน้าในปัจจุบัน การตรวจจับการปลอมแปลงใบหน้าโดยใช้กล้องอาร์จีบี สามารถแบ่งได้เป็น 3 วิธีหลักคือ

###### 1.1) Hybrid method

การใช้โครงข่ายประสาทแบบconvoluted neural network (CNN) หรือการเรียนรู้เชิงลึก (Deep learning) ได้รับความนิยมใช้กันอย่างแพร่หลายในหลายปัญหาทางด้านคอมพิวเตอร์วิทัคโน (Computer vision) แต่อย่างไรก็ตาม อัลกอริทึมการซุดข้อเรียนรู้เชิงลึกนั้น สามารถเกิดปัญหาโอเวอร์ฟิตติ้ง (overfitting problem) หรือ ปัญหาที่ทำให้ตัวแบบจำลอง (model) มีความแม่นยำต่ำลงเมื่อนำไปใช้กับชุดข้อมูล (dataset) ที่ไม่เคยเจอ

เนื่องจากปริมาณและความหลากหลายของข้อมูลสอน (training data) ที่จำกัด ดังนั้น Hybrid method จึงเสนอวิธีการใช้ CNN หรือ Deep learning ร่วมกับ Handcrafted features เช่น Local Binary Pattern (LBP) และ Weber local descriptor ซึ่งจากการสืบค้นงานวิจัยที่เกี่ยวข้องว่า ตัวสกัดคุณลักษณะเหล่านี้มีความสามารถในการแยกแยะความแตกต่าง ของภาพใบหน้า คนจริงและภาพใบหน้าปลอมได้ซึ่งจะเพิ่มประสิทธิภาพในการแก้ปัญหาการปลอมแปลงใบหน้าในกรณีดังกล่าวมากขึ้น

แต่อย่างไรก็ตามเนื่องจาก Hybrid method ยังคงมีจุดอ่อนอันเนื่องมาจากการใช้ Deep learning ร่วมกับ Handcrafted feature ซึ่งปราศจากการพารามิเตอร์ที่เรียนรู้ได้ (learnable parameter) ทำให้การใช้ Handcrafted feature นั้นยากที่จะปรับพารามิเตอร์ให้ได้ผลลัพธ์สูงที่สุด เช่น การเลือกช่วงค่าลีต่าง ๆ มาใช้งาน



**รูปที่ 4** Hybrid method (a) Deep features from handcrafted features (b) Handcrafted features from deep features (c) Fused handcrafted and deep features  
(ที่มา : [1])

### 1.1.1) Deep features from handcrafted features

คือ การสกัดคุณลักษณะสำคัญของข้อมูลด้วย Handcrafted feature จากข้อมูลขาเข้าก่อนจะตามด้วยการใช้ CNN ดังรูปที่ 3 (a) เช่น [2] แยกคุณสมบัติของชิลโทแกรมการกระจายความแตกต่างของความเข้มจากภาพสะท้อน (intensity difference distribution histograms from reflectance images) จากนั้นจึงใช้ 1D CNN กับชิลโทแกรมเพื่อดูการเปลี่ยนแปลงการส่องสว่าง

### 1.1.2) Handcrafted features from deep features

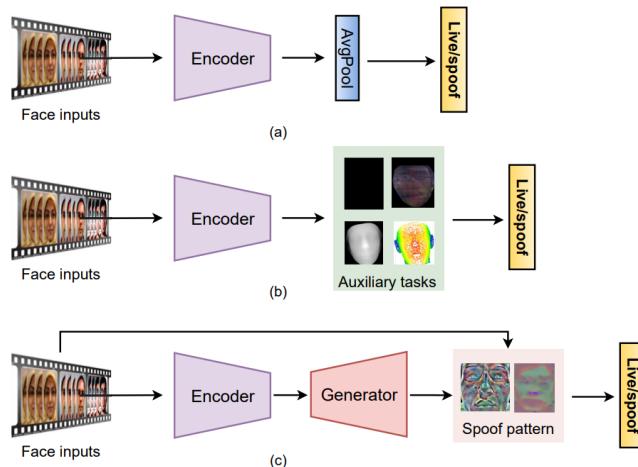
คือ การสกัดคุณลักษณะสำคัญของข้อมูลด้วย Handcrafted feature ออกมาจาก deep convolutional features ดังรูปที่ 3 (b) เช่น [3] สกัดข้อมูล Local Binary Pattern (LBP) จาก Convolution features ก่อนจะใช้ linear support vector machine (SVM) เพื่อจำแนกว่าเป็นการโจมตีหรือไม่

### 1.1.3) Fused handcrafted and deep features

คือ การทำทั้ง Handcrafted features และ deep convolutional features แยกกัน และนำผลของทั้งสองอย่างมารวมเข้าด้วยกัน ดังรูปที่ 3(c) เช่น [4] ดึงคุณสมบัติความแปรผันของความเข้มด้วย 1D CNN จากข้อมูลเข้าและแยกคุณสมบัติความว้างของภาพเบลอจากการเคลื่อนไหวด้วยอัลกอริทึม Similar Local Pattern (LSP) จากวิดีโอใบหน้าที่ขยายด้วยการเคลื่อนไหว ซึ่งจะถูกรวมเข้าด้วยกันก่อนจะใช้อัลกอริทึม SVM เพื่อแยกว่าเป็นการโจมตีหรือไม่

## 1.2) Common Deep Learning Method

เป็นการทำนายจากข้อมูลอินพุตไปเป็นคำตอบว่าเป็นการโจมตีหรือไม่ในลักษณะของปัญหาการจำแนก (Classification Problem) โดยไม่ต้องมีการนำ handcrafted features เข้ามาใช้



รูปที่ 5 Common Deep Learning Method (a) Direct Supervision with Binary Cross Entropy Loss (b)  
Pixel-wise supervision with Auxiliary Task (c) Pixel-wise Supervision with Generative Model  
(ที่มา : [1])

### **1.2.1) Direct Supervision with Binary Cross Entropy Loss**

เป็นการแบ่งรูปภาพออกเป็นกลุ่ม เช่น กลุ่มที่เป็นภาพจริง และ กลุ่มที่เป็นการโจมตี หรืออาจจะแบ่งการโจมตีออกเป็นหลายประเภท แล้วใช้ Loss Function ที่มีคุณสมบัติในการจำแนกข้อมูลนำเข้าว่า มีความใกล้เคียงกับข้อมูลกลุ่มใดมากที่สุด

### **1.2.2) Pixel-wise Supervision**

เป็นการนำข้อมูลระดับพิกเซล (Pixel) ที่สร้างขึ้นมาช่วยให้ CNN สามารถเรียนรู้ได้ดีขึ้น ซึ่งแบ่งออกเป็น 2 ประเภทคือ

#### **a. Pixel-wise supervision with**

[5] สร้างข้อมูลความลึก (depth maps) มาจากภาพตัวอย่าง เพื่อนำมาช่วยในการเรียนรู้ของ CNN

นอกจากความลึกแล้ว ยังมีข้อมูลประเภทอื่นอีก เช่น แสง สะท้อน (reflection map) [6] หรือ 3D point cloud map [7]

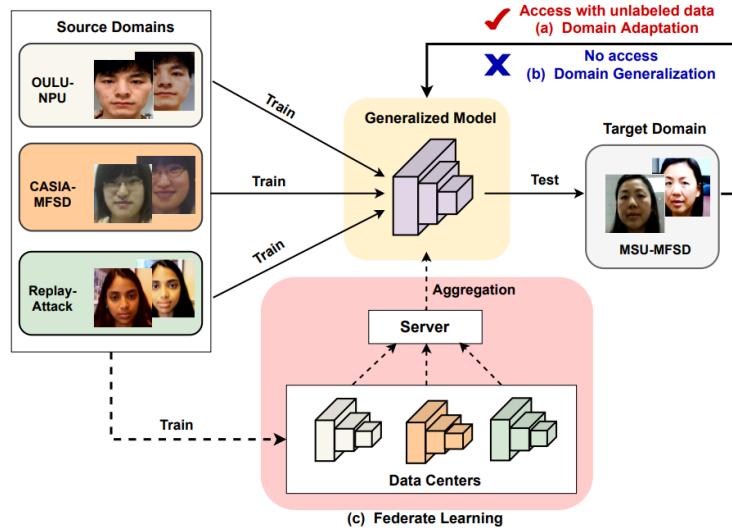
#### **b. Pixel-wise Supervision with Generative Model**

[8] ตีความปัญหาการตรวจสอบการปลอมแปลงใบหน้าใหม่ เป็นการแยกสัญญาณรบกวน (noise) ที่มองเห็นได้ในภาพที่เป็นการโจมตีออกมาจากใบหน้า

### **1.3) Generalized deep learning method**

เนื่องจาก deep learning method แบบปกติ มีแนวโน้มที่จะ overfit กับชุดข้อมูล (dataset) จึงอาจจะมีผลลัพธ์ที่ไม่ดี หากนำไปใช้ในสถานการณ์ที่ไม่เคยเจอ จึงมีการพัฒนา Generalized deep learning method ขึ้น โดยเน้นให้การเรียนรู้มีความเป็นทั่วไป เพื่อให้โมเดลยังคงมีประสิทธิภาพในกรณีที่การโจมตีเป็นการโจมตีในรูปแบบที่ไม่เคยเจอ เช่น การใส่หน้ากากรที่ทำด้วยวัสดุเฉพาะ หรืออยู่ในสภาพแวดล้อมที่ตัวโมเดลไม่เคยเจอ เช่น สภาพแสงเงา คุณภาพของกล้องและมุมกล้องที่แตกต่าง หรืออื่นๆ จากงานวิจัย [1] นั้นได้มีการจำแนกขั้นตอนวิธี ประเภทนี้ออกเป็น 2 ประเภทอยู่ ได้แก่

### 1.3.1) Generalization to Unseen Domain



รูปที่ 6 Generalization to Unseen Domain  
(ที่มา : [1])

เนื่องจากสภาพแวดล้อมที่แตกต่างจากเดิม เช่น สภาพแสงเงาที่ต่างกัน จะส่งผลให้โมเดลไม่มีความแม่นยำที่ Yao ลงได้ จึงมีการพยายามทำให้มีความเป็นทั่วไปในเชิงของสภาพแวดล้อม ซึ่งจะแบ่งออกมาได้ 3 วิธีคือ

#### a. Domain adaptation

เป็นการนำ ชุดข้อมูล (dataset) ที่ใช้ทดสอบโดยเดลมาช่วยในการ train ด้วย เพื่อให้โมเดลสามารถเรียนรู้โดเมนใหม่ ๆ ที่ไม่เคยเจอได้ เช่น ใน [9] ได้ทำการ train โมเดลกับชุดข้อมูลที่มีขนาดใหญ่มาก ๆ และจึงนำโมเดลที่ได้มาเป็นแกนหลักในการ train อีกโมเดลที่ชุดข้อมูลที่เหมือนกับสภาพแวดล้อมของ Domain เป้าหมาย

#### b. Domain generalization

เป็นการ train โมเดลจากชุดข้อมูลที่หลากหลาย เพื่อที่จะสามารถถกัด feature ที่มีความเป็นทั่วไปออกมาก ซึ่งมีความเป็นไปได้สูงที่ feature นี้จะมีในเป้าหมายที่เป็นโดเมนที่โมเดลไม่เคยเห็นเช่นเดียวกัน

#### c. Federate learning

เป็นการ train โมเดลจากชุดข้อมูลที่หลากหลาย เช่นเดียวกับ Domain generalization แต่แทนที่จะนำชุดข้อมูลมารวมกันก่อนนำเข้าโมเดลให้นำชุดข้อมูลของแต่ละแหล่งไปเข้าโมเดลแยกกัน จากนั้นจึงนำแต่ละโมเดลรวมอีกครั้งหนึ่ง เพื่อป้องกัน

ปัญหาความเป็นส่วนตัว (privacy) ของข้อมูล

### 1.3.2) Generalization to Unknown Attack Types

เป็นวิธีที่พยายามทำให้มีความเป็นทั่วไปในเชิงของการโจมตีซึ่งจะแบ่งออกมาได้ 2 วิธีคือ

#### a. Zero/Few-Shot Learning

เป็นวิธีการ train โมเดลโดยทำการ train โมเดลกับ ชุดข้อมูล (dataset) ที่มีความเป็นทั่วไปในระดับหนึ่งมาก่อนแล้ว มาใช้ในการทำงานการโจมตีที่ไม่เคยเห็นมาก่อน (Zero-shot) และนำตัวอย่างการโจมตีรูปแบบใหม่มาใช้ในการ train โมเดลในการทำงานต่อไป (Few-shot) เช่นใน [10] ได้ทำให้ตัวโมเดลมีการเรียนรู้จากช้อมูลที่เข้ามาใหม่เรื่อยๆ และตรวจจับใบหน้าที่ปลอมแปลงมาถึงแม้ว่าจะเป็นรูปแบบใหม่

ซึ่งข้อจำกัดของวิธีการนี้คือการที่ต้องมีตัวอย่างของการโจมตีแบบใหม่ เพื่อใช้ในการ train โมเดลเพิ่มก่อน ไม่เช่นนั้นประสิทธิภาพจะด้อยลงมาก และเมื่อทำการ train เพิ่มมาก ขึ้น อาจส่งผลทำให้ประสิทธิภาพของโมเดลต่ำลงได้

#### b. Anomaly Detection

เป็นวิธีการในการจำแนกการโจมตีด้วยวิธีการให้ตัวโมเดลเรียนรู้สิ่งที่ปกติให้ถูกต้องและนำযารว่าสิ่งที่แต่ต่างจากแนวทางนั้นเป็นการโจมตีทั้งหมด หรือคือเป็นการมองเป็นปัญหาการจำแนกประเภทคลาสเดียว (one-class classification) ที่จะทำการตัดสินว่าสิ่งนั้นเป็นค่าผิดปกติ (outlier) หรือไม่ เช่น ใน [11] ได้มีการสร้างโมเดลแบบ one-class classification ขึ้นมาในการตรวจจับใบหน้าที่ปลอมแปลงมา

ซึ่งวิธีการนี้ยังมีข้อจำกัดในการทำ FAS อยู่นั่นคือใบหน้าของคนนั้นมีความหลากหลายที่มาก เช่น สีผิวของใบหน้า ซึ่งถ้าหากว่าข้อมูลของใบหน้าปกติไม่ดีพอ หรือไม่มีความหลากหลายมากพอ ก็อาจจะทำให้โมเดลนี้มีประสิทธิภาพที่ไม่ดีพอไปด้วย

## 2) Deep FAS with advance sensor

เนื่องจากการใช้เพียงกล้องอาร์จีบีเพียงอย่างเดียวอาจจะมีข้อจำกัดในเชิงของข้อมูลที่ได้ ซึ่งอาจจะไม่เพียงพอต่อการระบุว่าเป็นการโจมตีหรือไม่ จึงมีการสร้าง Deep FAS Model โดยการนำข้อมูลจากมุมอื่น นอกจตามุมมองที่ได้จากการเซ็นเซอร์ในการรับแสงอาร์จีบีที่พับได้ในกล้องโดยทั่วไป ซึ่งเป็นรูปภาพแบบเดียวกับที่ตาของมนุษย์มองเห็น เช่น การใช้ข้อมูลที่ได้จากการเซ็นเซอร์ SWIR [12] ที่สามารถที่จะมองเห็นแสงที่มีความยาวคลื่นระหว่าง 1100-3000 นาโนเมตรได้ จะทำให้สามารถจับปริมาณของน้ำที่อยู่บนพื้นผิวนั้นได้ เป็นต้น ซึ่งข้อมูลในมุมมองอื่นนี้ ไม่ได้จำกัดเพียงแต่การใช้เซ็นเซอร์รับแสงชนิดอื่นเท่านั้น แต่รวมไปถึงการใช้ข้อมูลอื่นๆ เช่นร่วมได้ด้วย

เช่น การใช้แสงที่สะท้อนจากหน้าจอโทรศัพท์ไปสู่ใบหน้าของคน [13], [14] มาทำเป็นข้อมูลแสดงการสะท้อนและกระจายตัวของแสง เป็นต้น ซึ่งจะแบ่งได้เป็น 2 ประเภทคือ

### 2.1) Uni-Modal Deep Learning upon Specialized Sensor

เป็นการนำข้อมูลในมุมมองอื่น เพียงมุมมองเดียวมาใช้ในการสร้าง Deep FAS Model โดยไม่มีการนำข้อมูลอื่น ๆ เข้าร่วมด้วย

### 2.2) Multimodal Deep Learning

เป็นการนำข้อมูลจากหลายมุมมองมาสร้าง Deep FAS Model ซึ่งมี 3 วิธีการที่ใช้ในการสร้างได้คือ

**a) Feature-level** คือการนำข้อมูลในแต่ละมุมมองมา train โมเดลแยกกันแล้วจึงนำ Feature สำคัญที่ได้จากแต่ละโมเดลมารวมกัน

**b) Input-level** คือการนำข้อมูลในแต่ละมุมมองมารวมกันก่อนที่จะนำข้อมูลนั้นไป Train โมเดล เช่นข้อมูลรูปภาพอาร์จีบี ขนาด  $256 * 256$  จะได้เมทริกซ์ขนาด  $256*256*3$  รวมกับข้อมูลจากเซ็นเซอร์ SWIR ขนาด  $256*256$  จะได้เมทริกซ์ที่จะใช้ในการส่งให้ train โมเดลคือ  $256*256*4$  เช่นใน [15] ได้ทำการรวมข้อมูลนำเข้าจากรูปสีขาว-ดำ ความลึก และรังสีอินฟราเรด รวมกันส่งให้โมเดลประมวลผล

**c) Decision-level** คือการนำข้อมูลในแต่ละมุมมอง มา train โมเดลแยกกัน จากนั้นใช้โมเดลที่ได้นั้นในการทำนายผลลัพธ์แยกกันในแต่ละมุมมอง แล้วจึงนำผลการทำนายนั้นมาตัดสินอีกครั้งหนึ่งเป็นผลการทำนายสุดท้าย เช่นใน [16] ได้นำผลลัพธ์แยกกันโมเดลที่รับข้อมูลนำเข้าแต่ละอย่างแยกกับคือ รูปภาพอาร์จีบี ความลึกและรังสีอินฟราเรด มาประมวลผลต่ออีกทีหนึ่งเพื่อให้ได้เป็นผลลัพธ์สุดท้าย

ตารางที่ 2 ตารางเปรียบเทียบข้อดี - ข้อเสียของวิธีการต่าง ๆ

วิธีการตรวจจับการปลอมแปลงใบหน้า	ข้อดี	ข้อเสีย
Hybrid Method	<ul style="list-style-type: none"> <li>- สามารถเลือกคุณลักษณะ (feature) ที่ใช้ได้ตามต้องการ</li> <li>- โมเดลที่ได้สามารถทำความเข้าใจได้ง่ายกว่า</li> <li>- ใช้ทรัพยากรการคำนวนที่น้อยกว่า</li> </ul>	<ul style="list-style-type: none"> <li>- ต้องออกแบบคุณลักษณะ (feature) ด้วยตนเอง</li> <li>- คุณลักษณะไม่ได้เหมาะสมกับงานเสมอไป</li> </ul>
Common Deep Learning Method	<ul style="list-style-type: none"> <li>- สามารถทำความเข้าใจรูปแบบ (Pattern) ของการปลอมแปลงได้ง่ายกว่าวิธีการที่ใช้ Deep learning base วิธีการอื่น ๆ</li> </ul>	<ul style="list-style-type: none"> <li>- เกิดปัญหา overfitting หาก data set น้อยเกินไป</li> <li>- หากใช้ Auxiliary ช่วย ข้อมูล Auxiliary บางอย่างอาจไม่มีอยู่ในชุดข้อมูลเดิม</li> </ul>

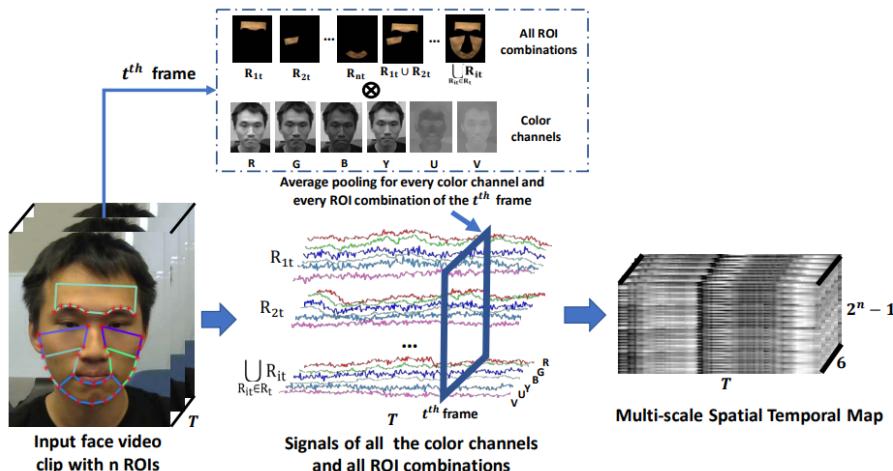
Generalized deep learning method	- ป้องกันการเกิด overfitting problem ได้ - เพิ่มโอกาสที่จะสามารถป้องกันการโจมตีในรูปแบบที่ไม่เดลไม่เดย์เห็นได้	- ใช้ทรัพยากรสูง - ความสามารถในการตรวจจับการโจมตีในรูปแบบปกติอาจายลลง
Deep FAS with advance sensor	มีความแม่นยำสูง เมื่อเทียบปริมาณทรัพยากรที่ใช้งานกับวิธีการอื่น ๆ	- ต้องการ sensor เฉพาะชุดข้อมูลมีปริมาณน้อย

## 2.2 งานวิจัยที่เกี่ยวข้อง

หลังจากที่ได้ทำการศึกษาแนวทางและข้อจำกัดต่าง ๆ ของ FAS method ในปัจจุบัน และได้เลือกที่จะแก้ปัญหาของ FAS ในเรื่องของทรัพยากรที่ใช้ในการประมวลผลที่สูง เราจึงได้ทำการค้นคว้างานวิจัยในเรื่องของ FAS ที่ฟอกสีหรือมีความเกี่ยวข้องกับประเด็นดังกล่าวข้างต้น เพื่อเป็นแนวทางในทางดำเนินงานต่อไป

### 2.2.1 Multi-scale spatial-temporal maps (MSTmap)

จาก [17] สร้างได้โดยทำการกำหนดพื้นที่จำนวน  $n$  พื้นที่ที่สนใจขึ้นมา จากนั้นจึงทำการสร้างการจัดกลุ่มขึ้นมาได้ทั้งหมด  $2^n - 1$  กลุ่ม จากนั้นนำค่าสีในแต่ละพื้นที่มาหาค่าเฉลี่ยกันทั้งหมด  $t$  ภาพในวิดีโอนั้น จะได้ MSTmap ที่มีขนาด  $2^n - 1 * T * C$  เมื่อ  $C$  เป็นจำนวนพื้นที่สีที่ใช้หลังจากนั้นจะทำการทำให้เป็นบรรทัดฐาน (normalized) ด้วยค่ามาก-น้อยที่สุดของค่าใน MSTmap นั้น ทำให้ได้ MSTmap ที่สามารถนำไปใช้งานได้ต่อไป

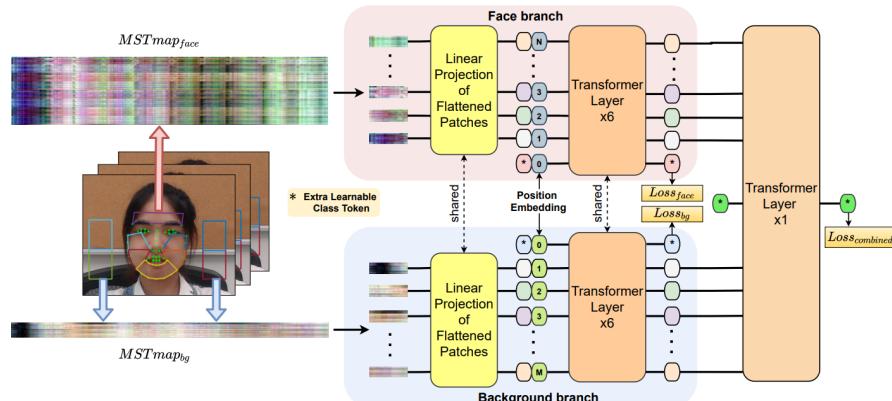


รูปที่ 7: การสร้าง MSTmap

(ที่มา [17])

## 2.2.2 TransRPPG

TransRPPG [18] เป็นเครื่องมือสำหรับตรวจจับการปลอมแปลงใบหน้าใน mask attack โดยนำเทคนิคของ Vision Transformer (ViT) มาใช้ร่วมกับ Remote Photoplethysmography (rPPG) ซึ่งเป็นเทคนิคในการตรวจจับความมีชีวิตชีวาของใบหน้าโดยวัดความเปลี่ยนแปลงของสีผิวนบนใบหน้า



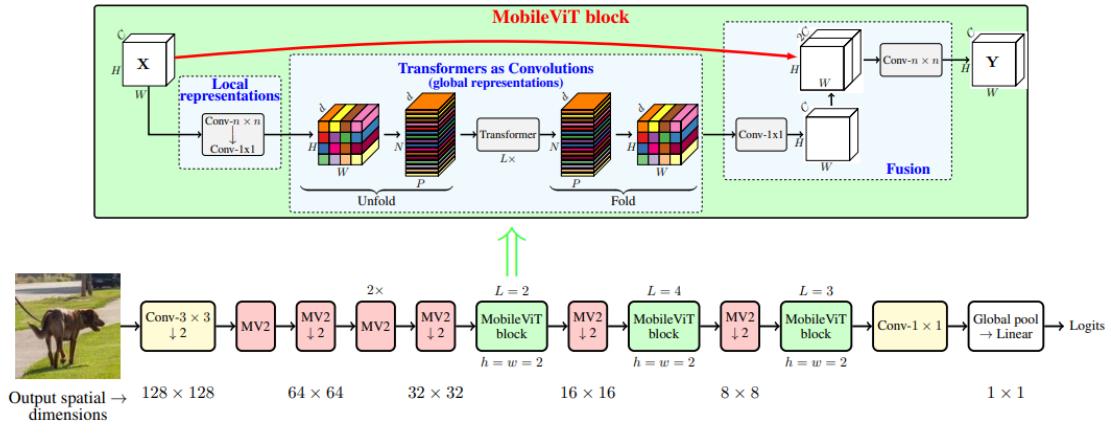
รูปที่ 8 ภาพรวมของ TransRPPG  
(ที่มา : [18])

ขั้นตอนของ TransRPPG เริ่มจากการนำวิดีโอใบหน้ามาสร้าง 2 multi-scale spatial-temporal maps (MSTMap) แบบเดียวกับ [17] จากพื้นที่ใบหน้าและพื้นหลัง จากนั้น MSTMap ทั้งสองจะถูกนำเข้า Vision Transformer เพื่อสกัด feature ต่าง ๆ ออกมารวมทั้ง rPPG feature และสุดท้าย feature จากทั้งสองจะถูกนำมารวมกันใน Transformer ชั้นสุดท้ายสำหรับ binary prediction (bonafide หรือ mask attack)

ซึ่งผลการทดลองพบว่าการใช้วิธีนี้ทำให้สามารถลดจำนวนพารามิเตอร์เหลือประมาณ 547K พารามิเตอร์และจำนวนการคำนวนเหลือ 763M FLOPs และมีความแม่นยำ AUC อยู่ที่ 98.3% เมื่อทำการ train ที่ชุดข้อมูล MARsV2 และ Test ที่ 3DMAD ซึ่งเป็นชุดข้อมูล 3 มิติ

## 2.2.3 MobileViT

MobileViT [19] เป็นโมเดลที่นำ CNN ไมเดลมาควบรวมกับ Vision Transformer โดยที่ใช้ข้อดีของทั้ง 2 ไมเดลมารวมกัน โดยที่ใช้ convolution layer ในการเข้ารหัส (encode) ข้อมูลในพื้นที่นั้น ๆ (local representation) และใช้ transformer block ในการเรียนรู้ข้อมูลในภาพรวม (Global representation) ทำให้ตัว MobileViT นั้นสามารถรักษาคุณสมบัติในพื้นที่เล็ก ๆ ไว้ได้ในขณะที่เรียนรู้ข้อมูลในภาพรวมไปด้วยกันได้



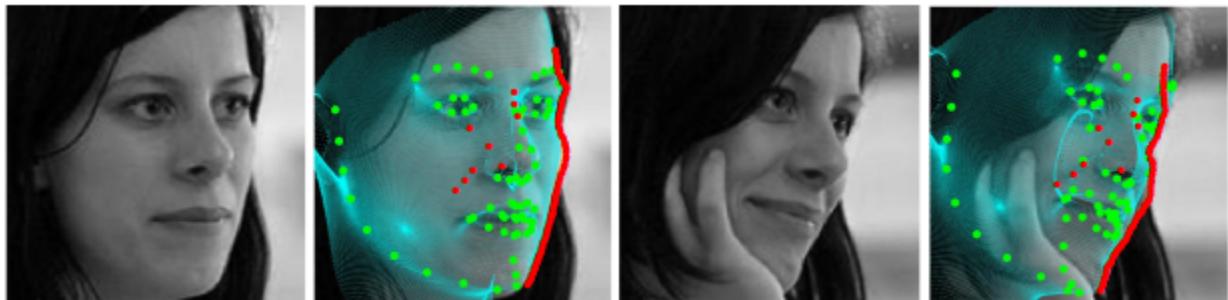
รูปที่ 9 MobileViT Architecture  
(ที่มา : [19])

โดยตัว MobileViT เมื่อทำการทดสอบแล้วพบว่าโมเดลนี้ มีลักษณะที่สามารถใช้ได้กับงานทั่วไปได้ดี อีกทั้งยังมีจำนวนของพารามิเตอร์ที่น้อย เมื่อเทียบกับโมเดลที่ได้ผลลัพธ์ที่ใกล้เคียงกันได้อีกด้วย เช่น เมื่อเปรียบเทียบกับ CNN โมเดลแล้ว MobileViT มีความแม่นยำมากกว่า EfficientNet [20] อยู่ 2.1% ในขณะที่มีจำนวนพารามิเตอร์ที่พอ ๆ กัน

## 2.2.4 Dense Face Alignment (DeFA)

Dense Face Alignment [21], [22] เป็นการหาตำแหน่งต่างๆ บนใบหน้าคน เช่น ตา จมูก ปาก จากรูปภาพ ดังรูปที่ 9 โดยสามารถสร้างรูปทรง 3 มิติ และแสดงข้อมูลความลึกของตำแหน่งต่างๆบนใบหน้า ซึ่งข้อมูลเหล่านี้ถูกนำมาใช้ประโยชน์ในงานต่างๆ ที่เกี่ยวกับใบหน้าคน รวมถึงการตรวจสอบการปลอมแปลงใบหน้า

ซึ่งจากการทดสอบบนชุดข้อมูล AFLW [23] พบว่าได้ความแม่นยำที่ค่อนข้างดี



รูปที่ 10 A pair of images with their dense 3D shapes obtained by imposing landmark fitting constraint, contour fitting constraint and sift pair constraint.  
(ที่มา : [21])

## 2.3 ชุดข้อมูล (Dataset)

### CelebA-Spoof

[24] เป็น ชุดข้อมูล ขนาดใหญ่ที่ถูกสร้างขึ้นมาจากการนำรูปภาพผู้มีชื่อเสียงมาสร้างรูปที่เป็นการโจมตีประเภทต่าง ๆ โดยมีรูปภาพคนจริงจำนวน 156,384 รูป และรูปที่เป็นการโจมตี 469,153 รูป ซึ่งในชุดข้อมูลนี้ประกอบด้วยการโจมตีหลายประเภท (ทั้งหมด 10 ประเภท) ซึ่งมีทั้งการโจมตีแบบ 2 มิติ (2D Attack) และการโจมตี 3 มิติ (3D Attack) และมีสภาพแวดล้อมที่หลากหลายโดยมีรูปภาพในสภาพแสงที่แตกต่างกัน และมีทั้งรูปภาพที่ถ่ายในอาคารและนอกอาคาร

ชุดข้อมูลนี้ออกแบบจากมีการระบุว่ารูปภาพแต่ละรูปเป็นรูปภาพคนจริงหรือไม่ แล้ว ยังมีการระบุคุณลักษณะอื่น ๆ อีกทั้งหมด 43 คุณลักษณะ ซึ่งเป็นคุณลักษณะของใบหน้าจำนวน 40 คุณลักษณะ เช่น ไส้เดือนตา ไส้หมาก มีหนวด เป็นต้น และคุณลักษณะที่เกี่ยวข้องกับ FAS อีก 3 คุณลักษณะ ได้แก่ ประเภทการโจมตี สภาพแวดล้อม และสภาพแสง

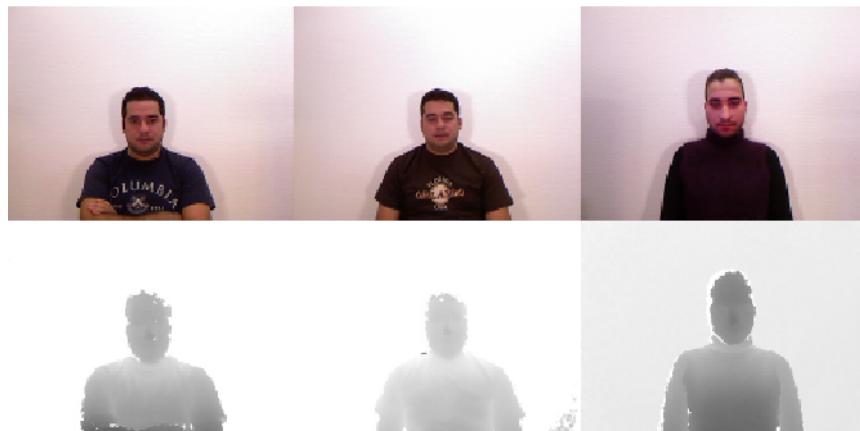
Spoof Type								Illumination Condition and Environment							
Print			Paper Cut			Replay		3D		Normal		Strong		Back	
Photo	Poster	A4	Face Mask	Upper Body Mask	Region Mask	PC	Pad	Phone	Mask	Indoor	Outdoor	Indoor	Outdoor	Indoor	Outdoor

รูปที่ 11 CelebA-Spoof  
(ที่มา : [24])

### **3DMAD**

[25] เป็น ชุดข้อมูล ขนาดเล็กที่ถูกสร้างขึ้นมาจากการวิดีโอของบุคคลจำนวน 17 คน ประกอบด้วยภาพวิดีโອ์ของคนจริงจำนวน 170 วิดีโอ และการโจมตี 3 มิติ (3D Attack) 85 วิดีโอ ซึ่งเป็นการโจมตีโดยใช้หน้ากากเรซิ่น โดยชุดข้อมูลนี้มีสภาพแวดล้อมที่เหมือนกันในทุกวิดีโอ

ชุดข้อมูลนี้นักวิจัยได้ใช้ภาพวิดีโอีนรูปแบบสีอาร์จีบีแล้ว ยังมีภาพวิดีโอีนรูปแบบของ Depth map มาให้ด้วย รวมทั้งมีการระบุตำแหน่งของดวงตาในทุก ๆ 60 ภาพ (frame) ของวิดีโอ

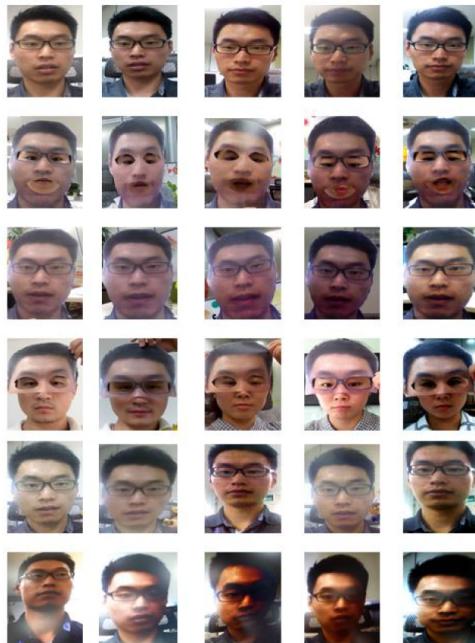


**รูปที่ 12 3DMAD**  
(ที่มา : [25])

### **Rose-Youtu**

[26] เป็น ชุดข้อมูล ที่ถูกสร้างขึ้นมาจากการวิดีโอของบุคคลจำนวน 20 คน ประกอบด้วยภาพวิดีโອ์ของคนจริงจำนวน 500 วิดีโอ และการโจมตี 2,850 วิดีโอ ซึ่งในชุดข้อมูลนี้ประกอบด้วยการโจมตีทั้งหมด 5 ประเภท ซึ่งเป็นการโจมตีแบบ 2 มิติ (2D Attack) 2 ประเภท ได้แก่ การโจมตีด้วยภาพบริ Rinne และการโจมตีด้วยการย้อนเล่นวิดีโอ และการโจมตี 3 มิติ (3D Attack) 3 ประเภท ซึ่งเป็นการโจมตีโดยใช้หน้ากากกระดาษที่มีแนวทางในการตัดแต่งต่างกัน

ชุดข้อมูลนี้นักวิจัยได้ใช้ภาพวิดีโอีนรูปแบบสีและวิดีโอีนรูปแบบ Depth map ไว้ในการฝึกอบรมเครื่องเรียน หรือไม่แล้ว ยังมีการระบุคุณลักษณะอื่น ๆ อีกทั้งหมด 6 คุณลักษณะ ได้แก่ ประเภทการโจมตี การพูดของบุคคล อุปกรณ์ที่ใช้บันทึกวิดีโอ การใส่แวกนตา สภาพแวดล้อม และไอเดียของบุคคล

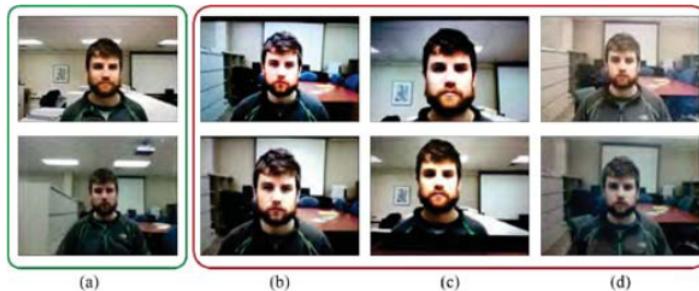


**รูปที่ 13 Rose-YouTu**  
(ที่มา : [26])

## MSU-MFSD

[27] เป็น ชุดข้อมูล ที่ถูกสร้างขึ้นมาจากการวิดีโอของบุคคลจำนวน 35 คน ประกอบด้วยภาพวิดีโอของคนจริงจำนวน 70 วิดีโอ และการโจมตีแบบ 2 มิติ (2D Attack) 210 วิดีโอ ซึ่งเป็นการโจมตีด้วยภาพปริ้น และการโจมตีด้วยการย้อนเล่นวิดีโอ โดยชุดข้อมูลนี้มีสภาพแวดล้อมที่เหมือนกันในทุกวิดีโอ

ชุดข้อมูลนี้ออกแบบจากมีการระบุว่าภาพวิดีโอแต่ละวิดีโอเป็นภาพวิดีโอุคนจริง หรือไม่แล้ว ยังมีการระบุคุณลักษณะอื่น ๆ อีกทั้งหมด 4 คุณลักษณะ ได้แก่ ประเภทการโจมตี อุปกรณ์ที่ใช้บันทึกวิดีโอ ความละเอียดของวิดีโอ และไอดีของบุคคล นอกจากนี้ยังมีการระบุตำแหน่งของใบหน้าในทุก ๆ ภาพ (frame) ของวิดีโอ



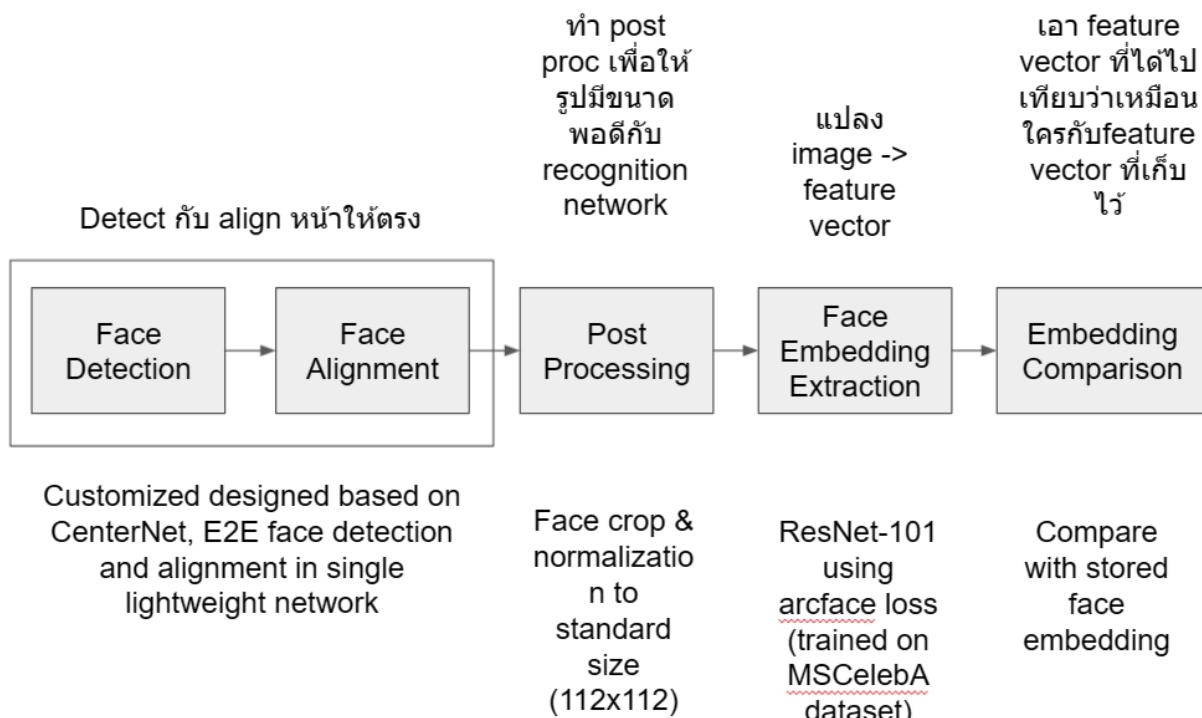
**รูปที่ 14 MSU-MFSD**  
(ที่มา [27])

## บทที่ 3

### แนวทางที่ทำ

#### 3.1 ระบบที่มีอยู่เดิม

รูปที่ 15 ระบบที่มีอยู่นั้นจะทำการตรวจสอบใบหน้าที่กล้องวิดีโอทำการถ่ายเอาไว้ หลังจากนั้นระบบจะทำการตัดพื้นที่บริเวณใบหน้ามาประมวลผลแล้วทำการแยกแยะใบหน้านั้นว่าเป็นใบหน้าของบุคคลไหน

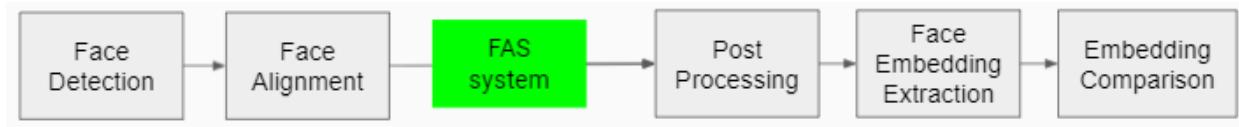


รูปที่ 15 The overview of the prior work

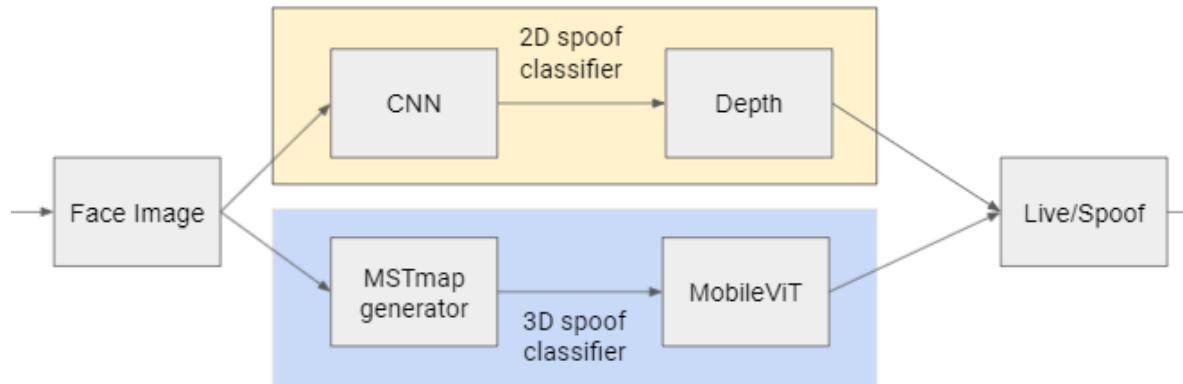
#### 3.2 ระบบที่เพิ่มขึ้นใหม่

วัตถุประสงค์ของโปรเจคนี้เพื่อเพิ่มเครื่องมือสำหรับตรวจสอบและต่อต้านการปลอมแปลงใบหน้าเข้าไปในระบบเดิม โดยเพิ่มระบบตรวจสอบการปลอมแปลงใบหน้าเข้าไปก่อนที่จะรู้จำใบหน้า ดังรูปที่ 15

รูปที่ 16 แสดงขั้นตอนต่างๆ ของระบบตรวจสอบการปลอมแปลงใบหน้า โดยหลังจากได้รับรูปภาพใบหน้ามาแล้ว จะนำไปใช้ในสองส่วน คือ การป้องกันการโจมตีแบบ 2 มิติ และการป้องกันการโจมตีแบบ 3 มิติ และนำผลจากทั้งสองส่วนมาใช้ตัดลินว่าเป็นการโจมตีหรือไม่



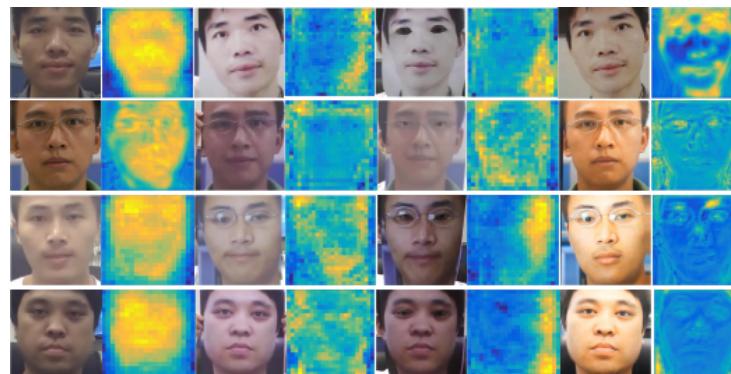
รูปที่ 16 The overview of the proposed method.



รูปที่ 17 The overview of face anti-spoofing system

### 3.2.1. การป้องกันการโจมตีแบบ 2 มิติ

จากการได้ศึกษางานวิจัยที่เกี่ยวข้องพบว่า การนำ Depth map เข้ามา ช่วยในการจำแนกการโจมตีได้อย่างมีนัยสำคัญ ดังนั้นทางผู้พัฒนาจึงคาดว่าจะนำ Depth map มาช่วยในการตรวจสอบการโจมตีแบบ 2 มิติ โดยใช้ CNN ที่รับข้อมูลภาพใบหน้าตัวอย่างและค่าความลึก (Ground truth depth maps) มาเรียนรู้เพื่อที่จะทำนายความลึกของภาพ



รูปที่ 18 The depth estimation. The first two columns are the live images, the rest six columns are three different types of spoof attacks (print, cut print and video attacks)  
(ที่มา : [5])

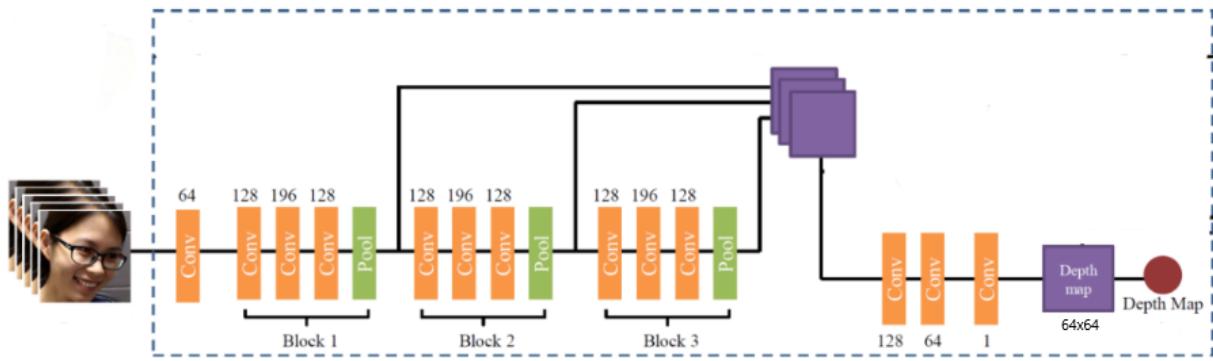
## 1) การเตรียมข้อมูล

สำหรับ ground truth ของภาพที่เป็นการจมตีแบบ 2 มิติจะมีค่าเป็นคูนย์ทั้งหมด ส่วน ground truth ของภาพจริงนั้น เราใช้ Dense face alignment (DeFA) methods [21], [22] ในการหารูป 3 มิติของใบหน้าอุอกมา และนำมิติความลึกของทรง 3 มิติที่ได้มาทำให้เป็น บรรทัดฐาน (normalized) ให้อยู่ระหว่าง 0 ถึง 1 และนำไปใช้เป็น ground truth ใน CNN

## 2) โมเดล

### Convolutional Neural Network

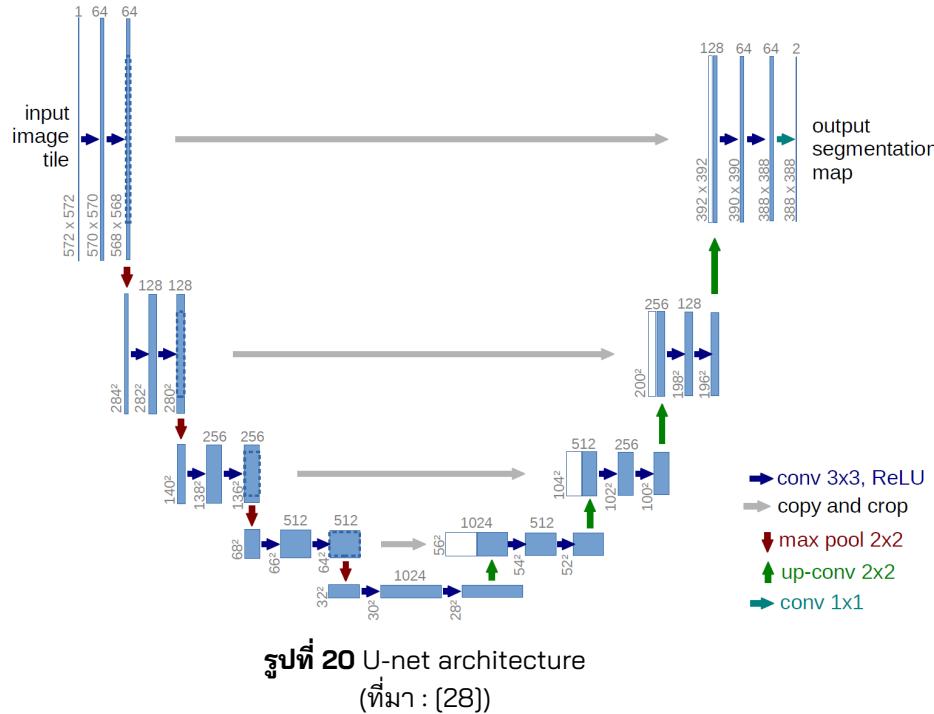
เราใช้ CNN ดังรูปที่ 19 ที่แต่ละชั้นประกอบด้วย convolutional layer 3 อัน ตามด้วย pooling และ resizing layer ที่เปลี่ยนให้เป็นขนาด  $64 \times 64$  และชั้นสุดท้ายจะได้ Depth map ออกมานั่นจะถูกนำมาเพียงกับ Ground truth depth map ที่เตรียมไว้



รูปที่ 19 Propose CNN architecture  
(ที่มา : [5])

### U-Net

เนื่องจากวิธีการที่นำในรูปที่ 19 เป็นวิธีที่มือญก่อและ เราจึงได้ลองนำ U-Net [28] มาใช้แทน โดย U-Net ประกอบด้วย contracting path (ด้านซ้าย) และ expansive path (ด้านขวา) ดังรูปที่ 20



รูปที่ 20 U-net architecture  
(ที่มา : [28])

### 3) วิธีการที่ทำ

เรา train โมเดลโดยรับข้อมูลเข้าเป็นภาพใบหน้าคนที่ถูกจัดตำแหน่งแล้ว และ Ground truth depth map สำหรับการหา classification score เราป้อนรูปภาพใบหน้าสำหรับทดสอบเข้าโมเดลเพื่อสร้าง Depth map และคำนวนคะแนนโดย

$$\text{Score} = \sqrt{\sum_{k=1}^n |x_k|^2}$$

โดย  $x$  คือค่าใน Depth map ซึ่งมีค่าอยู่ในช่วง 0 ถึง 1

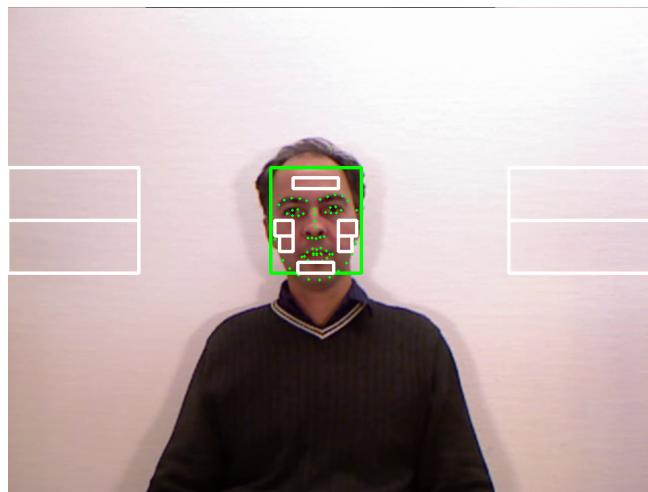
### 4) การวัดผล

เพื่อให้สามารถเปรียบเทียบผลลัพธ์กับงานที่มีอยู่ก่อนแล้วได้ การป้องกันการโจมตีแบบ 2 มิติ จะวัดผลโดยการคำนวณ Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER) และ Average Classification Error Rate (ACER) โดยที่  $\text{ACER} = (\text{APCER} + \text{BPCER})/2$

### 3.2.2. การป้องกันการโจมตีแบบ 3 มิติ

#### 1) การเตรียมข้อมูล

ก่อนนำข้อมูลเข้าโมเดลจะทำการแปลงข้อมูลนำเข้าจากวิดีโอของใบหน้าเปลี่ยนเป็นข้อมูลรูปภาพ โดยการสร้าง multi-scale spatial-temporal maps (MSTmap) ที่พื้นที่สี่เหลี่ยมจีบีจากภาพวิดีโอด้วยจำนวน 300 ภาพ (frame) (10 วินาทีสำหรับวิดีโอด้วยจำนวนภาพต่อวินาที (frame per second (fps)) เป็น 30 ภาพต่อวินาที หรือ 5 วินาทีสำหรับวิดีโอด้วยจำนวนภาพต่อวินาที (frame per second (fps)) เป็น 60 ภาพต่อวินาที) และกำหนดพื้นที่บนใบหน้าทั้งหมด 6 พื้นที่และพื้นหลัง 4 พื้นที่เช่นเดียวกับ [18] จะได้ MSTmap ขนาด  $63 \times 300 \times 3$  สำหรับพื้นที่บนใบหน้าและ  $15 \times 300 \times 3$  สำหรับพื้นหลัง



รูปที่ 21 การเลือกพื้นที่บนใบหน้าและพื้นหลังสำหรับการสร้าง MSTmap

#### 2) โมเดล

หลังจากที่ทำการเตรียมข้อมูลแล้วจะได้ multi-scale spatial-temporal maps (MSTmap) ของทั้งส่วนของใบหน้าและของพื้นหลังของกล้อง ซึ่งจะใช้เป็นข้อมูลนำเข้าส่งให้กับโมเดลต่อไป

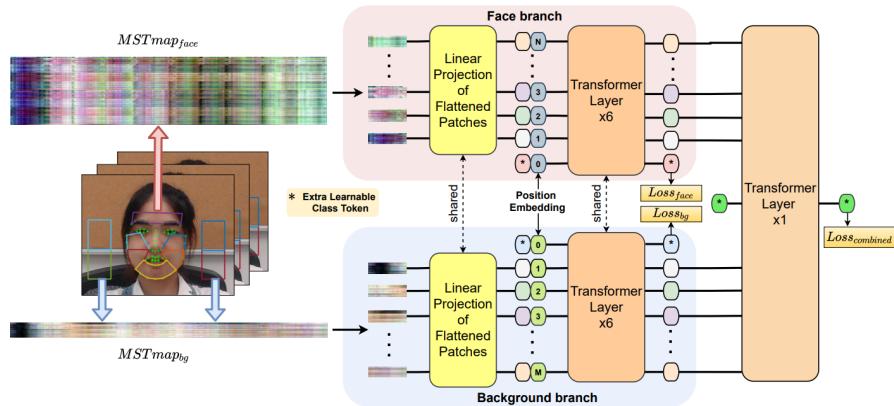
ในส่วนของโมเดลที่ใช้นั้น จะมีทั้งลีน 2 โมเดลด้วยกัน คือ โมเดล TransRPPG และ โมเดล MobileViT โดยมีวิธีการดำเนินการดังนี้

#### โมเดล Original TransRPPG และ Adapted TransRPPG

ในส่วนของโมเดลนี้เราจะดำเนินการตาม [18] จากรูป 22 โดยจะทำการดำเนินการทั้งหมด 2 รูปแบบ รูปแบบแรกนั้นดำเนินการโดยการตั้งค่า stride และ stride pad ในโมดูลของ Linear Projection ของโมเดลคือ (1,15) และ (3,30) ตามลำดับ ส่วนโมเดลในรูปแบบที่ 2 นั้นจะใช้

ค่าดั้งก่อตัวคือ (1,30) และ (3,60) ตามลำดับ และจะทำการเรียกโมเดล TransRPPG ในรูปแบบที่ 1 และ รูปแบบที่ 2 เป็น Original TransRPPG และ Adapted TransRPPG ตามลำดับ

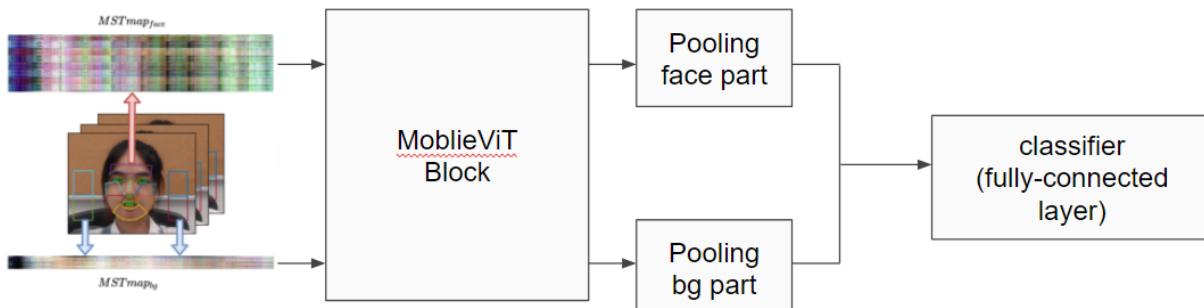
จากรูป 22 จะเห็นว่าโมดูลของ Linear Projection และ Transformer Layer นั้นจะมีการใช้พารามิเตอร์ร่วมกัน ในขั้นตอนการดำเนินการจะกระทำโดยที่ทำการส่งข้อมูลนำเข้าเป็น MSTmap ของใบหน้าและพื้นหลังเข้าไปพร้อมกัน เพื่อให้ตัวโมเดล TransRPPG นั้นทำการทำงานอย่างมาว่า MSTmap ของใบหน้าที่ได้รับนั้นเป็นใบหน้าของคน หรือเป็นคนที่ใส่หน้ากาก



รูปที่ 22 ภาพรวมของ TransRPPG  
(ที่มา : [18])

### โมเดล MobileViTRPPG

ในส่วนของโมเดลนี้นั้น ก็จะรับข้อมูลนำเข้าเป็น MSTmap ของใบหน้าและพื้นหลัง เช่นเดียวกันกับโมเดล TransRppg และจะทำการดำเนินการตาม [19] แต่ว่าจะมีการเปลี่ยนแปลง ในบางส่วนคือ การนำข้อมูลทั้ง MSTmap ของใบหน้าและพื้นหลังมาทำการเข้ารหัสด้วยตัวโมเดล MobileViT ก่อนหลังจากนั้นจะนำข้อมูลหลังจากที่ทำการเข้ารหัสแล้วมาต่อ กัน จากนั้นก็จะนำ ข้อมูลที่เข้ารหสนิลแล้วส่งเข้าสู่ชั้นเชื่อมโยงแบบสมบูรณ์ (Fully-Connected Layer) ดังแสดงในภาพ 23 ในการจำแนกระหว่างใบหน้าคนกับคนที่ใส่หน้ากาก โดยจะทำการเรียกโมเดลนี้ว่า MobileViTRPPG



รูปที่ 23 ภาพรวมของโมเดล MobileViT

### **3) การวัดผล**

การป้องกันการโจมตีแบบ 3 มิติ จะแบ่งการวัดผลออกเป็น 2 ประเด็น คือ การวัดประสิทธิภาพของโมเดล และขนาดของโมเดล

#### **ประสิทธิภาพของโมเดล**

วัดโดยการคำนวณ accuracy และสร้างเส้นโค้งของ Receiver Operating Characteristic (ROC Curve) เพื่อใช้ในการคำนวณ Equal Error Rate (EER), Area under the ROC Curve (AUC), และ False Fake Rate เมื่อ False Liveness Rate มีค่าเท่ากับ 0.01

#### **ขนาดของโมเดล**

วัดโดยใช้ปริมาณพารามิเตอร์ (Params) และปริมาณการคำนวณของโมเดล Giga Multiply-Accumulate Operations (GMACs)

## บทที่ 4

### การทดลองและผลลัพธ์

#### 4.1 การป้องกันการโจมตีแบบ 2 มิติ

การป้องกันการโจมตีแบบ 2 มิติ จะแบ่งการทดลองและผลลัพธ์ออกเป็น 2 ส่วน คือการดำเนินการก่อนการปรับปรุง และการดำเนินการหลังการปรับปรุง

##### การดำเนินการก่อนการปรับปรุง

ในขั้นตอนนี้จะมีการพัฒนาโมเดลทั้งหมด 2 โมเดล ได้แก่ CNN ดังรูปที่ 19 และ U-Net ดังรูปที่ 20 เราได้ทำการให้ข้อมูลนำเข้าเป็นรูปภาพจากชุดทดสอบของ CelebA-Spoof [24]

##### ผลการทดลองของการดำเนินการก่อนปรับปรุง

จากรายงานที่ 3 นี้ที่แสดงถึงประสิทธิภาพของโมเดลของการดำเนินการก่อนการปรับปรุง โดยวัดผลการคำนวน Attack Presentation Classification Error Rate (APCER) ซึ่งคำนวนจาก  $(APCER + BPCER)/2$  จะเห็นว่าประสิทธิภาพของโมเดลทั้ง 2 โมเดลของเรานั้นทำได้ไม่ดีมากนัก เมื่อเปรียบเทียบกับงานที่มีอยู่ก่อนแล้ว [5] ทั้งนี้การทดลองของเราใช้ชุดข้อมูลที่ไม่เหมือนกันงานที่ถูกนำมาเปรียบเทียบด้วย

ตารางที่ 3 ผลการทดลองประสิทธิภาพของโมเดลของการดำเนินการก่อนการปรับปรุง

Dataset	Method	Average Classification Error Rate (ACER)
Oulu	Y. Atoum, Y. Liu, A. [5]	0.041
CelebA-Spoof	Y. Atoum, Y. Liu, A. (our reimplementation)	0.220
	U-Net	0.295

##### สรุปผลการดำเนินการก่อนการปรับปรุง

จากการทดลองของแต่ละโมเดลนั้น จะเห็นว่าตัวของโมเดล CNN ตามรูปที่ 19 นั้นให้ผลลัพธ์ที่ดีกว่า U-Net ตามรูปที่ 20 ทางผู้ทำการทดลองจึงจะทำการดำเนินการปรับปรุงในส่วนของโมเดล CNN จากรูปที่ 19 ต่อไป

##### การปรับปรุงการทดลอง

จากรูปที่ 19 โมเดลของเรานั้นให้ผลลัพธ์ออกมาเป็น Depth map แต่เราไม่สามารถจำแนกประเภทของภาพใบหน้าจาก Depth map โดยตรงได้ เราจึงต้องมีการคิด classification score จาก Depth map อีกที

เราจึงตัดแปลงโมเดลจากรุปที่ 18 ให้กลายเป็นโมเดลจำแนกประเภท (classification model) โดยเพิ่ม dense layer เข้าไปหลังจากชั้นสุดท้ายของโมเดล ทำให้ผลลัพธ์สุดท้ายเป็นโอกาสในการเป็นภาพจริงซึ่งมีค่าอยู่ในช่วง 0 ถึง 1

ในการ train โมเดล เราได้ minimizes ทั้ง Depth map loss และ binary cross entropy loss ไปพร้อมกัน โดยที่ loss จาก Depth map ช่วยปรับค่าน้ำหนักแต่ส่วนที่เป็น CNN อย่างเดียว แต่ binary cross entropy loss จะปรับค่าน้ำหนักทั้งส่วน CNN และ dense layer ที่เพิ่มเข้ามาไปพร้อมกัน

### ผลการทดลองของการดำเนินการหลังปรับปรุง

จากตารางที่ 4 นี้ที่แสดงถึงประสิทธิภาพของโมเดลของการดำเนินการหลังการปรับปรุง จะเห็นว่าประสิทธิภาพของโมเดล classification model ที่ใช้ binary cross entropy loss เพียงอย่างเดียวมีประสิทธิภาพที่แย่ลง ส่วนโมเดลที่ minimizes ทั้ง Depth map loss และ binary cross entropy loss พร้อมกันนั้นมีประสิทธิภาพที่ดีขึ้นเมื่อเปรียบเทียบกับการทดลองก่อนปรับปรุง แต่ก็ยังมีประสิทธิภาพที่ด้อยกว่างานที่มีอยู่เดิม [5]

ตารางที่ 4 ผลการทดลองประสิทธิภาพของโมเดลของการดำเนินการก่อนการปรับปรุง

Dataset	Method	Average Classification Error Rate (ACER)
Oulu	Y. Atoum, Y. Liu, A. [5]	0.041
CelebA-Spoof	Our old method	0.220
	Predict class only	0.319
	Predict both depth and class	0.189

### สรุปผลการดำเนินการหลังการปรับปรุง

จากการทดลองนั้นจากเห็นว่าหลังจากการปรับปรุงตัวโมเดล มีประสิทธิภาพที่เพิ่มขึ้น ถึงแม้ว่าจะยังมีประสิทธิภาพที่ด้อยกว่างานที่มีอยู่เดิม [5] ทั้งนี้เนื่องจาก มีข้อจำกัดต่างๆ ทำให้การทดลองของเราทดสอบใน ชุดข้อมูล ที่ต่างออกไป และการทดลองของเรายังไม่ได้ใช้ pre-train model ทางผู้ทดลองได้เลือกให้ตัวโมเดลหลังการปรับปรุงซึ่งมีผลลัพธ์ดีที่สุดเป็นโมเดลที่จะใช้งานเป็นหลักในระบบการป้องกันการโจมตีแบบ 2 มิติ

### ข้อจำกัดของโมเดลจากการทดลอง

จากการทดลองในชุดข้อมูล CelebA-Spoof [24] นั้น เนื่องจากว่าลักษณะของแสง และสภาพแวดล้อมมีความหลากหลาย ซึ่งแสงและสภาพแวดล้อมมีผลต่อการทำงานของโมเดลค่อนข้างมาก ทำให้ผลของการจำแนกในบางสภาพแวดล้อมทำออกมาได้ไม่ดีมากนัก โดยเฉพาะการโจมตีประเภทภาพปริ้นท์ที่ทำนายอุปกรณามิดบอยกว่าภาพประเภทอื่น ดังรูปที่ 25



รูปที่ 24 successful anti-spoofing examples and their estimated depth maps



รูปที่ 25 failure examples and their estimated depth maps

## 4.2 การป้องกันการโจมตีแบบ 3 มิติ

การป้องกันการโจมตีแบบ 3 มิติ จะแบ่งการทดลองและผลลัพธ์ออกเป็น 2 ส่วน คือการดำเนินการก่อนการปรับปรุง และการดำเนินการหลังการปรับปรุง

### 4.2.1 การดำเนินการก่อนการปรับปรุง

ในขั้นตอนนี้จะมีการพัฒนาโมเดลทั้งหมด 3 โมเดล ได้แก่โมเดล Original TransRPPG, Adapted TransRPPG และ MoblieViT ในการดำเนินการทดลอง

#### โมเดล Original TransRPPG

สำหรับการทดลองในโมเดล Original TransRPPG นี้ เราได้ทำการให้ข้อมูลนำเข้าเป็น MSTmap ของคลิปวิดีโอที่มีความยาว 10 วินาที โดยที่มีจำนวนภาพต่อวินาที (frame per second (fps)) เป็น 30 ภาพต่อวินาที ทำให้ได้ MSTmap ที่เกิดจากภาพจำนวน 300 ภาพ

### **โมเดล Adapted TransRPPG**

สำหรับการทดลองในโมเดล Adapted TransRPPG นี้ เราได้ทำการให้ข้อมูลนำเข้าเป็น MSTmap ของคลิปวิดีโอที่มีความยาว 5 วินาที โดยที่มีจำนวนภาพต่อวินาที (frame per second (fps)) เป็น 60 ภาพต่อวินาที ทำให้ได้ MSTmap ที่เกิดจากภาพจำนวน 300 ภาพ

### **โมเดล MobileViTRPPG**

สำหรับการทดลองในโมเดล MobileViTRPPG นี้ เราได้ทำการให้ข้อมูลนำเข้าเป็น MSTmap ของคลิปวิดีโอที่มีความยาว 5 วินาที โดยที่มีจำนวนภาพต่อวินาที (frame per second (fps)) เป็น 60 ภาพต่อวินาที ทำให้ได้ MSTmap ที่เกิดจากภาพจำนวน 300 ภาพ เช่นเดียวกันกับการทดลองกับโมเดล Adapted TransRPPG

### **ผลการทดลองของการดำเนินการก่อนปรับปรุง**

#### **ประสิทธิภาพของโมเดล**

จากการ 5 นี้ที่แสดงถึงประสิทธิภาพของโมเดลของการดำเนินการก่อนการปรับปรุง จะเห็นว่าประสิทธิภาพของโมเดลทั้ง 3 โมเดลนั้นทำได้ไม่ดีมากนัก ซึ่งหากสังเกตุที่ค่า EER และ AUC นั้นจะเห็นว่าประสิทธิภาพของโมเดล MobileViTRPPG นั้นทำได้ไม่ดีมากนักเมื่อทำการเปรียบเทียบกับโมเดล Adapted TransRPPG ที่ใช้ข้อมูลนำเข้าแบบเดียวกัน

และจากผลลัพธ์จะเห็นว่าการใช้ข้อมูลนำเข้าเป็น MSTmap จากวิดีโอด้วยความยาว 5 วินาทีที่ 60 ภาพต่อวินาทีนั้น ได้ผลลัพธ์ที่แย่กว่า การใช้ข้อมูลนำเข้าเป็น MSTmap จากวิดีโอด้วยความยาว 10 วินาทีที่ 30 ภาพต่อวินาที อย่างมีนัยสำคัญ ซึ่งสาเหตุที่มีผลลัพธ์ที่แย่กว่าดังกล่าวนั้น ทางผู้ทดลองได้คาดว่ามาจากการที่ข้อมูลตั้งตันเป็นวิดีโอบนแบบความยาว 10 วินาทีที่ 30 ภาพต่อวินาที และผู้ทำการทดลองได้ทำการตัดวิดีโอีเป็น 5 วินาทีและทำการเพิ่มจำนวนรูปภาพเป็น 60 ภาพต่อวินาที ซึ่งไม่ได้เป็นการเพิ่มปริมาณของข้อมูลใหม่ให้กับวิดีโอด้วยความยาว 5 วินาทีที่ 60 ภาพต่อวินาทีมากนัก ส่งผลให้ผลลัพธ์ที่ได้ออกมาทำได้ไม่ดีเท่ากับการใช้วิดีโอด้วยความยาว 10 วินาทีที่ 30 ภาพต่อวินาที

**ตารางที่ 5** ผลการทดลองประสิทธิภาพของโมเดลของการดำเนินการก่อนการปรับปรุง

Data	Model	accuracy	EER	AUC	FFR@FLR=0.01
10 sec Video	Original TransRPPG	0.79	0.36	0.7696	1.0
5 sec Video	Adapted TransRPPG	0.69	0.4	0.6168	1.0

	MobileViTRPPG	0.69	0.52	0.5280	1.0
--	---------------	------	------	--------	-----

### ปริมาณพารามิเตอร์และการคำนวณของโมเดล

จากตาราง 6 นี้ที่แสดงถึงขนาดและปริมาณการคำนวณของโมเดลนี้ จะเห็นว่า ปริมาณพารามิเตอร์ของโมเดลทั้ง 3 โมเดลนั้นไม่ต่างกันมากนัก โดยที่โมเดล Adapted TransRPPG นั้นมีปริมาณพารามิเตอร์ที่น้อยที่สุดและ MobileViTRPPG นั้นมีปริมาณพารามิเตอร์ที่สูง

แต่สิ่งที่ต่างกันอย่างมากคือปริมาณการคำนวณของโมเดล ซึ่งโมเดล MobileViTRPPG นั้นมีมากที่สุด ซึ่งสาเหตุที่โมเดล MobileViTRPPG นั้นมีทั้งปริมาณพารามิเตอร์ที่มากและใช้ปริมาณการคำนวณมากที่สุดซึ่งต่างจากผลที่ [19] ได้กล่าวเอาไว้นั้น ผู้ทำการทดลองคาดว่าเป็นสาเหตุมาจากการที่ขนาดของโมเดล TransRPPG นั้นมีปริมาณพารามิเตอร์ที่น้อยอยู่แล้ว ซึ่งต่างจากโมเดล Vision Transformer ที่มีขนาดใหญ่(ปริมาณพารามิเตอร์ที่เยอะมาก)ที่ทางผู้เขียน [19] นั้นนำมาเปรียบเทียบ โดยการที่ทำการข้อรหัสข้อมูลเชิงพื้นที่ในโมเดล Vision Transformer ที่ขนาดใหญ่นั้น จะช่วยลดทั้งปริมาณพารามิเตอร์และปริมาณการคำนวณของส่วนที่เป็น Transformer และมาเพิ่มในส่วนของ convolution layer แทน ซึ่งน้อยเมื่อเทียบกับปริมาณพารามิเตอร์และการคำนวณที่ลดได้ในส่วนของ Transformer ซึ่งต่างจากตัวโมเดล TransRPPG ที่มีปริมาณพารามิเตอร์และการคำนวณที่น้อยอยู่แล้ว ดังนั้นการใช้ convolution layer มาช่วยนั้น จึงช่วยลดปริมาณพารามิเตอร์และการคำนวณในส่วน Transformer ได้น้อย กลับกันกลับเป็นการเพิ่มทั้งปริมาณพารามิเตอร์และการคำนวณในส่วนของ convolution layer มากกว่า จึงทำให้ทั้งปริมาณพารามิเตอร์และการคำนวณของโมเดล MobileViTRPPG นั้นมีมากกว่าตัวโมเดล TransRPPG ทั้ง 2 โมเดล

ตารางที่ 6 ปริมาณพารามิเตอร์และการคำนวณของโมเดลของการดำเนินการก่อนการปรับปรุง

	Params (Parameters)	GMACs (Flop)
Original TransRPPG	944.74 k	1.13
Adapted TransRPPG	899.62 k	0.55

MobileViTRPPG	1.01 M	3.97
---------------	--------	------

## สรุปผลการดำเนินการก่อนการปรับปรุง

จากการทดลองทั้งในด้านของประสิทธิภาพ ปริมาณพารามิเตอร์และการคำนวณของโมเดลนั้น จะเห็นว่าตัวของโมเดลที่ทำการดำเนินการตาม [18] นั้นให้ผลลัพธ์ที่ดีกว่า [19] ทางผู้ทำการทดลองจึงจะทำการดำเนินการปรับปรุงในส่วนของโมเดลที่ดำเนินการตาม [18] ต่อไป

### 4.2.2 การดำเนินการหลังการปรับปรุง

การปรับปรุงการทดลองในส่วนนี้นั้น ผู้ทำการทดลองได้กระทำการดำเนินการโดยแบ่งได้เป็น 2 ส่วน ได้แก่

#### การปรับปรุงโมเดล

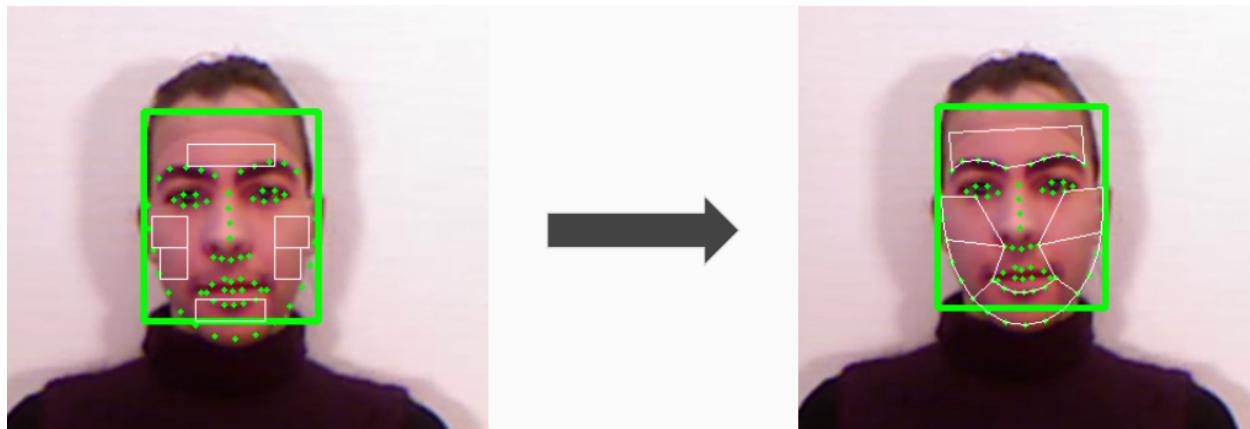
การปรับปรุงโมเดลนี้ทำโดยการทำ Hyper parameter tuning คือการปรับไฮเปอร์พารามิเตอร์ต่าง ๆ ของโมเดลใหม่ เช่นการปรับปริมาณ hidden layer ของโมดูล Transformer ใหม่เพื่อให้ผลลัพธ์ของโมเดลดีมากยิ่งขึ้น

#### การทำการเตรียมข้อมูลใหม่

ในขั้นตอนการเตรียมข้อมูลใหม่นี้ได้กระทำทั้งหมด 2 รูปแบบได้แก่

#### การปรับปรุงพื้นที่ที่สันใจบนใบหน้าใหม่

ในการสร้าง MSTmap นั้นจำเป็นที่จะต้องเลือกพื้นที่ที่สันใจในการนำมาร้านวน และสร้าง MSTmap ซึ่งในขั้นตอนนี้ทางผู้ทำการทดลองได้ทำการปรับพื้นที่ในส่วนนี้ใหม่ให้มีความละเอียดมากยิ่งขึ้นดังที่แสดงในรูป 26



รูปที่ 26 แสดงการปรับปรุงการเลือกพื้นที่บนใบหน้าสำหรับการสร้าง MSTmap

#### การทำ Data Augmentation บนข้อมูลสำหรับสร้าง MSTmap

ในขั้นตอนนี้ผู้ทดลองได้ทำการปรับปรุงการสร้าง MSTmap ใหม่สำหรับวิดีโอที่มีความยาว 5 วินาทีที่ 60 ภาพต่อวินาที เนื่องด้วยตัววิดีโอ 5 วินาทีนั้นเกิดจากการตัดวิดีโอ

ตั้งต้นที่มีความยาว 10 วินาทีมาใช้ เแล้วจากนั้นค่อยทำการเพิ่มปริมาณภาพจาก 30 ภาพ ต่อวินาทีเป็น 60 วินาที ซึ่งในขั้นตอนนี้ทำให้ผู้ทดลองต้องทิ้งข้อมูลอีก 5 วินาทีนั้นทิ้งไป ผู้ทดลองจึงได้ทำการปรับเปลี่ยนวิธีการตัดวิดีโอนี้ใหม่เป็นการทำ sliding window วิดีโอด้วยความยาว 10 วินาทีดังรูป 27 ซึ่งทำให้ในวิดีโอี 10 วินาที 1 วิดีโอนี้สามารถตัดวิดีโอด้วยความยาว 5 วินาทีได้มากขึ้น ในที่นี้ผู้ทดลองได้ทำการเลือกตัว sliding window นี้ทุก ๆ 1 วินาที ทำให้สามารถสร้างวิดีโอด้วยความยาว 5 วินาทีได้เพิ่มมากขึ้นจากเดิมเป็น 6 เท่าตัว ส่งผลให้ปริมาณ MSTmap ที่สร้างได้นั้นก็จะมากขึ้นจากเดิมเป็น 6 เท่าตัวตามไปด้วย



**รูปที่ 27** จำลองการทำ sliding window ตัดวิดีโอด้วยความยาว 5 วินาที (กรอบสีแดง เขียวและฟ้า) จากวิดีโอด้วยความยาว 10 วินาที

### ผลการทดลองของการดำเนินการหลังปรับปรุง

#### ประสิทธิภาพของโมเดล

จากตาราง 7 นี้ แสดงผลการทดลองประสิทธิภาพของโมเดลของการดำเนินการหลังการปรับปรุงเบรียบเทียบกับประสิทธิภาพของโมเดลของการดำเนินการก่อนการปรับปรุง จะเห็นว่าโมเดลหลังจากที่ทำการปรับปรุงแล้วนั้นมีประสิทธิภาพที่ดีมากขึ้นกว่า ก่อนปรับปรุงอย่างมีนัยสำคัญในทุกตัวชี้วัด และให้เห็นว่าการทำ Hyper parameter tuning และการปรับปรุงพื้นที่ที่สันใจบนใบหน้าใหม่ นั้นส่งผลต่อตัวโมเดลอย่างมาก และโมเดลแบบ Original TransRPPG นั้นสามารถทำผลลัพธ์ได้ดีกว่าแบบ Adapted TransRPPG ถึงแม้ว่าจะทำการทดลองกับข้อมูลวิดีโอด้วย 5 วินาทีก็ตาม และการที่ผลลัพธ์ของการทดลองกับข้อมูลของวิดีโอบนแบบ 5 วินาทีและแบบ 10 วินาทนั้น ไม่แตกต่างกันมากในโมเดล Original TransRPPG นั้นแสดงให้เห็นว่าการทำ Data Augmentation บนข้อมูลสำหรับสร้าง MSTmap นั้นก็ส่งผลเช่นเดียวกัน

**ตารางที่ 7** ผลการทดลองประสิทธิภาพของโมเดลของการดำเนินการหลังการปรับปรุงเปรียบเทียบกับ  
ประสิทธิภาพของโมเดลของการดำเนินการก่อนการปรับปรุง

Data	Model	Old				New			
		accuracy	EER	AUC	FFR@FLR=0.01	accuracy	EER	AUC	FFR@FLR=0.01
10 sec Video	Original TransRppg	0.79	0.36	0.7696	1.0	0.86	0.2705	0.8280	0.9647
5 sec Video	Original TransRppg	-	-	-	-	0.85	0.1960	0.8694	0.8
	Adapted TransRppg	0.69	0.4	0.6168	1.0	0.80	0.2568	0.8185	0.9078

### ปริมาณพารามิเตอร์และการคำนวณของโมเดล

จากตาราง 8 ที่แสดงถึงปริมาณพารามิเตอร์และการคำนวณของโมเดลของการดำเนินการหลังการปรับปรุงเปรียบเทียบกับปริมาณพารามิเตอร์และการคำนวณของโมเดลของการดำเนินการก่อนการปรับปรุง จะเห็นว่าการทำ Hyper parameter tuning นั้นส่งผลต่อทั้งปริมาณของพารามิเตอร์และการคำนวณ ทำให้โมเดลหลังจากที่ทำการปรับปรุงแล้วมีทั้งปริมาณของพารามิเตอร์และการคำนวณที่เพิ่มขึ้นเล็กน้อย

**ตารางที่ 8** ปริมาณพารามิเตอร์และการคำนวณของโมเดลของการดำเนินการหลังการปรับปรุงเปรียบเทียบกับปริมาณพารามิเตอร์และการคำนวณของโมเดลของการดำเนินการก่อนการปรับปรุง

	Before		After	
	Params (Parameters)	GMACs (Flop)	Params (Parameters)	GMACs (Flop)
Original TransRppg	944.74 k	1.13	1.07 M	1.31
Adapted TransRppg	899.62 k	0.55	1.03 M	0.64

## สรุปผลการดำเนินการหลังการปรับปรุง

จากการทดลองในด้านของประสิทธิภาพนั้นจากเห็นว่าหลังจากการปรับปรุงตัวโมเดลมีประสิทธิภาพที่เพิ่มขึ้นเป็นอย่างมาก แม้ว่าปริมาณพารามิเตอร์และการคำนวณจะเพิ่มขึ้นเล็กน้อย แต่ว่าเมื่อทำการเปรียบเทียบกับประสิทธิภาพของโมเดลที่เพิ่มขึ้นมาแล้วนั้น ถือว่าเป็นผลลัพธ์ที่สามารถยอมรับได้ จากผลลัพธ์ทั้งหมดทางผู้ทดลองจึงได้เลือกให้ตัวโมเดล Original TransRPPG ที่ทำการทดลองกับวิดีโocommunity 5 วินาทีที่ 60 ภาพต่อวินาทีเป็นโมเดลที่จะใช้งานเป็นหลักในระบบการป้องกันการโจมตีแบบ 3 มิติ

### ข้อจำกัดของโมเดลจากการทดลอง

จากการทดลองในชุดข้อมูล Rose-Youtu [26] นั้น เนื่องจากว่าตัวชุดข้อมูลนี้นั้นมีตัวอย่างข้อมูลที่มีการปิดบังใบหน้าบางส่วนดังรูป 28 ซึ่งการสร้างตัว MSTmap นั้นจะใช้การสร้างจากพื้นที่ที่สนใจบนใบหน้าแต่ละส่วน การที่ใบหน้าถูกปิดบังเพียงบางส่วนจะทำให้ตัว MSTmap จะยังสามารถสังเกตุสิ่งของใบหน้าที่เปลี่ยนไปตามช่วงเวลาของส่วนที่ยังไม่ถูกปิดได้อยู่ ทำให้ผลของการจำแนกระหว่างใบหน้าการโจมตีและใบหน้าที่ใส่หน้ากากนั้น ทำออกมาได้ไม่ดีมากนัก



รูปที่ 28 แสดงรูปการโจมตีที่มีการปิดบังใบหน้าบางส่วนภายในชุดข้อมูล Rose-Youtu

ในส่วนของคุณภาพของแสงก็ส่งผลต่อการสร้าง MSTmap เป็นอย่างมากอีกด้วย เนื่องจากการสร้าง MSTmap นั้นจะมีการทำให้เป็นบริหัดฐาน (normalized) ด้วยค่ามาก-น้อย ที่สุดอยู่ซึ่งถ้าหากว่าคุณภาพของแสงนั้นไม่ดี หรือว่ามีการเปลี่ยนความเข้มแสงโดยฉับพลัน จะทำให้ MSTmap ที่ได้ไม่มีความถูกต้องสำหรับใช้ในการจำแนกใบหน้าคนต่อไป ดังเห็นได้ในรูป 29



รูปที่ 29 แสดง MSTmap ของคลิปวีดิโอที่เกิดแสงจ้าขึ้นฉับพลันในขณะถ่าย

## บทที่ 5 ผลกระทบทางสังคมของโครงการ

โครงการนี้มุ่งพัฒนาเทคโนโลยีตรวจจับและการต่อต้านการปลอมแปลงใบหน้า (Face anti-spoofing) ซึ่งเป็นส่วนสำคัญที่เข้ามาปิดช่องโหว่ของเทคโนโลยีการรู้จำใบหน้า (Face recognition) จากการโจมตีของผู้ไม่หวังดี ซึ่งมุ่งใช้เทคนิคและเทคโนโลยีต่าง ๆ ในการปลอมแปลงหรือหลอกเลียนการตรวจจับใบหน้า นอกจากนี้แล้วโครงการนี้ยังมุ่งพัฒนาเทคโนโลยีตรวจจับและการต่อต้านการปลอมแปลงใบหน้าให้ใช้เวลาและทรัพยากรการคำนวณที่ต่ำ แต่มีประสิทธิภาพสูง ทำให้สามารถนำไปประยุกต์ใช้กับอุปกรณ์พกพาซึ่งมีข้อจำกัดด้านทรัพยากรของเครื่องได้

## บทที่ 6 บทสรุป

การทดลองของเรา มีวัตถุประสงค์เพื่อสร้างเครื่องมือสำหรับตรวจจับและการต่อต้านการปลอมแปลงใบหน้าโดยเน้นให้ระบบใช้ทรัพยากรต่ำ แต่ยังคงมีประสิทธิภาพสูง และสามารถทำงานได้อย่างรวดเร็ว โดยแบ่งเป็นส่วนการป้องกันการโจมตีแบบ 2 มิติ และการป้องกันการโจมตีแบบ 3 มิติ

สำหรับการป้องกันการโจมตีแบบ 2 มิติ ได้มีการทำนาย Depth map เพื่อช่วยในการตรวจจับการโจมตี จากผลการทดลอง การป้องกันการโจมตีแบบ 2 มิติในบางสภาพแวดล้อมทำออกมาได้ไม่ดีมากนัก โดยเฉพาะการโจมตีประเภทภาพปรินต์ที่ทำนายออกมายังไบท์ภาพประเภทอื่น

ส่วนการป้องกันการโจมตีแบบ 3 มิติ ได้ใช้ MSTmap ของส่วนของใบหน้าและของพื้นหลังของกล้องเพื่อช่วยในการตรวจจับการโจมตี ซึ่งจากผลการทดลอง การป้องกันการโจมตีแบบ 3 มิติ นั้น มีข้อจำกัดในเรื่องของแสงสว่างที่ไม่สามารถที่จะตรวจจับได้อย่างแม่นยำในพื้นที่ที่มีแสงสว่างมากจนเกินไปหรือมีการเปลี่ยนแปลงในช่วงเวลา นั้น ๆ หากเกินไป และอีกข้อจำกัดหนึ่งคือลักษณะของหน้าหาก ที่จำเป็นต้องเป็นหน้าหากที่ครอบคลุมใบหน้าด้านหน้าทั้งส่วน ซึ่งสามารถของการป้องกันการโจมตีแบบ 3 มิติจะลดต่ำลงหากว่าการโจมตีนั้นใช้หน้าหากปกปิดใบหน้าเฉพาะบางส่วนเท่านั้น

## บทที่ 7

### เอกสารอ้างอิง

- [1] Z. Ming, M. Visani, M. M. Luqman, and J.-C. Burie, "A Survey On Anti-Spoofing Methods For Face Recognition with RGB Cameras of Generic Consumer Devices," *ArXiv201004145 Cs*, Oct. 2020, Accessed: Oct. 29, 2021. [Online]. Available: <http://arxiv.org/abs/2010.04145>
- [2] L. Li, Z. Xia, X. Jiang, Y. Ma, F. Roli, and X. Feng, "3D face mask presentation attack detection based on intrinsic image analysis," *IET Biom.*, vol. 9, no. 3, pp. 100–108, 2020, doi: 10.1049/iet-bmt.2019.0155.
- [3] A. Agarwal, M. Vatsa, and R. Singh, "CHIF: Convoluted Histogram Image Features for Detecting Silicone Mask based Face Presentation Attack," in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Tampa, FL, USA, Sep. 2019, pp. 1–5. doi: 10.1109/BTAS46853.2019.9186000.
- [4] L. Li, Z. Xia, A. Hadid, X. Jiang, H. Zhang, and X. Feng, "Replayed Video Attack Detection Based on Motion Blur Analysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 9, pp. 2246–2261, Sep. 2019, doi: 10.1109/TIFS.2019.2895212.
- [5] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, Denver, CO, Oct. 2017, pp. 319–328. doi: 10.1109/BTAS.2017.8272713.
- [6] T. Kim, Y. Kim, I. Kim, and D. Kim, "BASN: Enriching Feature Representation Using Bipartite Auxiliary Supervisions for Face Anti-Spoofing," Oct. 2019, pp. 494–503. doi: 10.1109/ICCVW.2019.00062.
- [7] X. Li, J. Wan, Y. Jin, A. Liu, G. Guo, and S. Z. Li, "3DPC-Net: 3D Point Cloud Network for Face Anti-spoofing," p. 8.
- [8] A. Jourabloo, Y. Liu, and X. Liu, "Face De-Spoofing: Anti-Spoofing via Noise Modeling," *ArXiv180709968 Cs*, Jul. 2018, Accessed: Nov. 04, 2021. [Online]. Available: <http://arxiv.org/abs/1807.09968>
- [9] H. Li, S. Wang, P. He, and A. Rocha, "Face Anti-Spoofing With Deep Neural Network Distillation," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 5, pp. 933–946, Aug. 2020, doi: 10.1109/JSTSP.2020.3001719.
- [10] D. Pérez-Cabo, D. Jiménez-Cabello, A. Costa-Pazo, and R. J. López-Sastre, "Learning to Learn Face-PAD: a lifelong learning approach," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, Sep. 2020, pp. 1–9. doi: 10.1109/IJCB48548.2020.9304920.
- [11] "[2007.05856] Anomaly Detection-Based Unknown Face Presentation Attack Detection." <https://arxiv.org/abs/2007.05856> (accessed Nov. 06, 2021).
- [12] G. Heusch, A. George, D. Geissbuhler, Z. Mostaani, and S. Marcel, "Deep Models and Shortwave Infrared Information to Detect Face Presentation Attacks," *ArXiv200711469 Cs*, Jul. 2020, Accessed: Oct. 31, 2021. [Online]. Available: <http://arxiv.org/abs/2007.11469>
- [13] Y. Liu et al., "Aurora Guard: Real-Time Face Anti-Spoofing via Light Reflection," *ArXiv190210311 Cs*, Feb. 2019, Accessed: Oct. 31, 2021. [Online]. Available: <http://arxiv.org/abs/1902.10311>
- [14] A. F. Ebihara, K. Sakurai, and H. Imaoka, "Specular- and Diffuse-reflection-based Face Spoofing Detection for Mobile Devices," *ArXiv190712400 Cs*, Dec. 2020, Accessed: Oct. 31, 2021. [Online]. Available: <http://arxiv.org/abs/1907.12400>
- [15] O. Nikisins, A. George, and S. Marcel, "Domain Adaptation in Multi-Channel Autoencoder based Features for Robust Face Anti-Spoofing," *ArXiv190704048 Cs*, Jul. 2019, Accessed: Nov. 06, 2021. [Online]. Available: <http://arxiv.org/abs/1907.04048>
- [16] Z. Yu et al., "Multi-Modal Face Anti-Spoofing Based on Central Difference Networks," *ArXiv200408388 Cs*, Apr. 2020, Accessed: Nov. 06, 2021. [Online]. Available: <http://arxiv.org/abs/2004.08388>
- [17] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, "Video-based Remote Physiological

- Measurement via Cross-verified Feature Disentangling,” *ArXiv200708213* Cs, Jul. 2020, Accessed: Oct. 31, 2021. [Online]. Available: <http://arxiv.org/abs/2007.08213>
- [18] Z. Yu, X. Li, P. Wang, and G. Zhao, “TransRPPG: Remote Photoplethysmography Transformer for 3D Mask Face Presentation Attack Detection,” *ArXiv210407419* Cs, Apr. 2021, Accessed: Oct. 29, 2021. [Online]. Available: <http://arxiv.org/abs/2104.07419>
- [19] S. Mehta and M. Rastegari, “MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer,” *ArXiv211002178* Cs, Oct. 2021, Accessed: Oct. 29, 2021. [Online]. Available: <http://arxiv.org/abs/2110.02178>
- [20] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *Proceedings of the 36th International Conference on Machine Learning*, May 2019, pp. 6105–6114. Accessed: Nov. 05, 2021. [Online]. Available: <https://proceedings.mlr.press/v97/tan19a.html>
- [21] Y. Liu, A. Jourabloo, W. Ren, and X. Liu, “Dense Face Alignment,” *ArXiv170901442* Cs, Sep. 2017, Accessed: Oct. 31, 2021. [Online]. Available: <http://arxiv.org/abs/1709.01442>
- [22] A. Jourabloo and X. Liu, “Pose-Invariant Face Alignment via CNN-Based Dense 3D Model Fitting,” *Int. J. Comput. Vis.*, vol. 124, no. 2, pp. 187–203, Sep. 2017, doi: 10.1007/s11263-017-1012-z.
- [23] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, “Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization,” in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Barcelona, Spain, Nov. 2011, pp. 2144–2151. doi: 10.1109/ICCVW.2011.6130513.
- [25] N. Erdoganmus and S. Marcel, “Spoofing Face Recognition With 3D Masks,” *Inf. Forensics Secur. IEEE Trans. On*, vol. 9, pp. 1084–1097, Jul. 2014, doi: 10.1109/TIFS.2014.2322255.
- [26] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, “Unsupervised Domain Adaptation for Face Anti-Spoofing,” *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 7, pp. 1794–1809, Jul. 2018, doi: 10.1109/TIFS.2018.2801312.
- [27] Di Wen, Hu Han, and A. K. Jain, “Face Spoof Detection With Image Distortion Analysis,” *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 4, pp. 746–761, Apr. 2015, doi: 10.1109/TIFS.2015.2400395.
- [28] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *ArXiv150504597* Cs, May 2015, Accessed: May 10, 2022. [Online]. Available: <http://arxiv.org/abs/1505.04597>