Science and Engineering Connect, ISSN 3027-7914 (Online)
Vol. 47 No. 2 pp. 99-117 (2024)

บทความวิจัย (Research Article)

การสร้างภาพใบหน้าคนร้ายจากคำบรรยายรูปพรรณสัณฐาน
Text-to-face Synthesis for Criminal Sketch

เมธี ประเสริฐกิจพันธุ์, พีรพัทธ ตั้งไพบูลย์, ไตรทิพย์ ศุภศิริวัฒนา, ดวงธิดา แซ่แต้,
วิชิต ชำนาญนาวา, ณัฐภณ อัศวเหม, ฐิติรัตน์ ศิริบวรรัตนกุล*
Mathee Prasertkijaphan, Peerapat Tungpaiboon, Taitip Suphasiriwattana, Duangthida Sae-Tae,
Wichit Chamnannawa, Nattaphon Assavahem, Thitirat Siriborvornratanakul*
คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์ กรุงเทพฯ ประเทศไทย
Graduate School of Applied Statistics, National Institute of Development Administration,
Bangkok, Thailand
*Corresponding author E-mail: thitirat@as.nida.ac.th

Received 3 August 2022; Revised 6 April 2024; Accepted 19 April 2024

บทคัดย่อ

ความเป็นมาและวัตถุประสงค์ : การแปลงข้อความเป็นรูปภาพได้รับความนิยมแพร่หลายขึ้น เนื่องจากสามารถสร้างประโยชน์ให้กับหลายวงการ เช่น ในด้านการสืบสวนสอบสวนที่ต้องการร่างภาพคนร้าย หรือแม้กระทั่งการตามหาบุคคลสาบสูญ อย่างไรก็ตาม งานวิจัยด้านการแปลงข้อความเป็นรูปภาพส่วนใหญ่มักเป็นการศึกษาบนวัตถุที่ไม่ซับซ้อน เช่น วัตถุทั่วไป ดอกไม้ หรือนก แต่การศึกษาบนใบหน้าคนยังมีไม่มากนัก ส่วนหนึ่งเป็นเพราะฐานข้อมูลใบหน้าคนยังไม่สมบูรณ์ เช่น ฐานข้อมูลของ Labeled Faces in the Wild และ MegaFace ที่มีเพียงรูปภาพแต่ไม่มีคำบรรยาย งานวิจัยนี้เป็นการพัฒนาระบบร่างภาพใบหน้าคนร้ายอัตโนมัติ โดยอาศัยเพียงอินพุตข้อความที่เป็นคำบรรยายลักษณะรูปพรรณสัณฐาน

วิธีดำเนินการวิจัย : ในงานวิจัยนี้ ผู้วิจัยใช้ฐานข้อมูลของ CelebA HQ ซึ่งเป็นฐานข้อมูลที่มีทั้งรูปภาพและคำบรรยาย จากนั้นใช้วิธีเข้ารหัสข้อความด้วย BERT Sentence Encoder และ CLIP Text Encoder เพื่อถอดรหัสคำบรรยายออกเป็นคำศัพท์และแปลงเป็นตัวเลข แล้วนำตัวเลขดังกล่าวไปทดสอบในแบบจำลอง 4 แบบ ได้แก่ StyleGAN3+CLIP, StyleGAN3+CLIP (Fine-Tuning), DC+CLIP (Fine-Tuning) และ DF+CLIP (Fine-Tuning)

ผลการวิจัย : จากผลการทดลองเปรียบเทียบแบบจำลองทั้ง 4 แบบด้วยตัวชี้วัดเชิงปริมาณอย่าง FID พบว่าแบบจำลองที่ใช้ StyleGAN3 เป็น Generator มีประสิทธิภาพสูงสุดในการสร้างภาพที่มีคุณภาพดีและตรงตามคำบรรยาย ทั้งนี้ ผลลัพธ์ที่ดีที่สุดคือเมื่อใช้ StyleGAN3 คู่กันกับแบบจำลอง CLIP ที่ผ่านการ Fine-tuning เพิ่มเติมด้วยชุดข้อมูล CelebA HQ

สรุป : เพื่ออำนวยความสะดวกและเพิ่มประสิทธิภาพของกระบวนการร่างภาพใบหน้าคนร้ายในงานสืบสวนสอบสวน งานวิจัยนี้ทำการพัฒนาระบบร่างภาพใบหน้าคนร้ายโดยอาศัยอินพุตข้อความที่เป็นคำบรรยายรูปพรรณสัณฐานของคนร้ายแต่เพียงอย่างเดียว จากการทดลองผสมผสาน Generative Adversarial Network (GAN) เข้ากับแบบจำลอง CLIP พบว่า การผสมผสาน StyleGAN3 และ CLIP ให้ผลลัพธ์การสร้างภาพใบหน้าได้มีคุณภาพดีและตรงตามคำบรรยายมากที่สุด

การนำไปใช้ประโยชน์ในเชิงปฏิบัติ : งานวิจัยนี้สามารถนำไปพัฒนาต่อยอดเพื่อสร้างระบบร่างภาพใบหน้าคนร้ายจากคำบรรยายรูปพรรณสัณฐานสำหรับช่วยเหลือในการสืบสวนสอบสวน หรือใช้ในบริบทอื่นที่คล้ายกัน เช่น การออกแบบใบหน้าของนางแบบหรือนายแบบใหม่เพื่อประโยชน์ในการสร้างสื่อโฆษณา

คำสำคัญ : การเรียนรู้เชิงลึก, การสร้างภาพใบหน้าจากข้อความ, StyleGAN3, Contrastive Language-Image Pre-training

Abstract

Background and Objectives: Transformation of text into images is gaining widespread popularity because such transformation offers benefits in various fields, including in investigations requiring sketching of suspects or even in the search for missing persons. However, most researches on text-to-image transformation tend to focus on such simple objects or common items as flowers or birds. Studies on human faces are still relatively rare, partly because facial databases are incomplete. Such databases as Labeled Faces in the Wild and MegaFace contain only images without descriptive text. The present research aimed to develop an automatic suspect face sketching system using only textual descriptions of physical characteristics as input.

Methodology: The present research used CelebA HQ database, which contains both images and descriptions. BERT Sentence Encoder and CLIP Text Encoder were employed to encode the text, converting the descriptions into vocabulary and eventually transforming them into numerical values. These numerical values were tested on four models: StyleGAN3+CLIP, StyleGAN3+CLIP (Fine-Tuning), DC+CLIP (Fine-Tuning), and DF+CLIP (Fine-Tuning).

Main Results: Based on the experimental comparison of the four models using the quantitative indicator FID, we found that the model using StyleGAN3 as the Generator exhibited the highest performance in generating high-quality images that matched the descriptions. The best results were

obtained when StyleGAN3 was paired with the CLIP model, which had been further fine-tuned with the CelebA HQ dataset.

Conclusions: To facilitate and enhance the efficiency of the process of sketching suspect faces in investigations, the present research developed a system for generating suspect face sketches based solely on textual descriptions of the suspect's physical features. The experimental results, which were obtained through the combined use of Generative Adversarial Networks (GAN) and CLIP model, revealed that the combination of StyleGAN3 and CLIP produced the highest quality and most accurate face sketches based on the descriptions.

Practical Application: The present research can be further developed to create a system for generating suspect face sketches from textual descriptions to aid in investigations. It can also be used in similar contexts, such as designing new model faces for advertising media creation.

Keywords: Deep Learning, Text-to-face Synthesis, StyleGAN3, Contrastive Language-Image Pre-training

Introduction

การร่างภาพคนร้ายเป็นงานที่ท้าทายและสำคัญต่อกระบวนการสืบสวนสอบสวนของตำรวจ โดยต้องอาศัยทักษะการวาดภาพและความรู้ความเชี่ยวชาญในเรื่องสรีระวิทยาประกอบกันจึงจะได้รูปภาพผลลัพธ์ที่ดี ในปัจจุบันผู้เชี่ยวชาญงานนี้มีจำนวนไม่มากนัก การร่างภาพคนร้ายจะกระทำโดยอาศัยการวาดภาพของคนร้ายตามคำบอกเล่าของพยาน ในบางครั้งภาพที่ได้อาจจะไม่ถูกต้องซึ่งอาจเกิดจากความผิดพลาดทั้งจากผู้เล่าและนักร่างภาพ จากการศึกษาของ Jalan และคณะ [1] พบว่าการร่างภาพคนร้ายจะมีภาพเพียง 13 ภาพจาก 160 ภาพหรือประมาณ 8 เปอร์เซ็นต์เท่านั้นที่ทำได้ใกล้เคียงกับภาพของคนร้ายตัวจริง ซึ่งถือว่ามีความแม่นยำค่อนข้างน้อย

ในปัจจุบันแม้ว่าการร่างภาพคนร้ายจะเปลี่ยนจากการใช้มือวาดเพียงอย่างเดียวมาเป็นการใช้ซอฟต์แวร์ช่วยเหลือ แต่ก็พบว่ายังมีข้อจำกัดอยู่ เช่น EvoFit ซอฟต์แวร์ตัวใหม่ล่าสุดสำหรับช่วยสร้างภาพร่างคนร้ายนั้น มีการเปลี่ยนกระบวนการร่างภาพคนร้ายเป็นการให้พยานบอกลักษณะคนร้ายโดยคร่าว ๆ เช่น เพศ อายุ และเชื้อชาติ จากนั้นโปรแกรมจะสุ่มภาพใบหน้าขึ้นมาให้พยานเลือกภาพที่คิดว่าคล้ายคนร้ายมากที่สุด แล้วโปรแกรมก็จะสร้างภาพใบหน้าถัดไปที่ใกล้เคียงกับใบหน้าที่เลือก ทำซ้ำเช่นนี้ไปเรื่อย ๆ จนได้ภาพที่พยานมั่นใจที่สุด แล้วจึงค่อยนำภาพใบหน้ามาปรับแต่งทรงผม โครงหน้าเพิ่มเติม อย่างไรก็ตามวิธีนี้ใช้เวลาในการสร้างภาพนาน เนื่องจากต้องอาศัยการทำซ้ำหลาย ๆ ครั้งและโปรแกรมยังไม่สามารถสร้างภาพที่มีคุณภาพสูงได้

การบอกลักษณะของคนร้ายในปัจจุบันมีอยู่ด้วยกันหลายวิธี Kotian และคณะ [2] สรุปการบอกลักษณะของคนร้ายหรือคนที่น่าสงสัยออกเป็น 2 วิธี วิธีแรกคือการบอกรายละเอียดของคนร้ายจากลักษณะเด่น โดยให้หานักร่างภาพที่มีฝีมือและบอกเฉพาะลักษณะเด่นของคนร้าย เพื่อให้นักร่างภาพร่างรูปภาพคนร้ายขึ้นมา ซึ่งขั้นตอนนี้อาจใช้เวลาหลายชั่วโมงกว่าจะได้รูปที่สมบูรณ์ วิธีที่สองคือให้เก็บข้อมูลภาพของเหตุการณ์ที่เกิดขึ้นเท่าที่สามารถถ่ายเก็บไว้ได้ จากนั้นนำภาพเหล่านี้มอบให้กับทุกสถานีตำรวจเพื่อกระจายข่าวและช่วยในการตามหาคนร้าย เมื่อพิจารณาจากความเป็นจริงแล้วความแม่นยำในการระบุตัวตนของคนร้ายของวิธีที่ 2 นั้นค่อนข้างน้อยกว่าวิธีที่ 1 จึงทำให้มีความพยายามในการพัฒนากระบวนการร่างภาพคนร้ายและการระบุตัวคนร้ายจากรูปภาพให้มีประสิทธิภาพสูงขึ้น เช่น การนำเทคนิคปัญญาประดิษฐ์ (Artificial intelligence) ประเภทการเรียนรู้เชิงลึก (Deep learning) มาช่วยร่างภาพคนร้าย ดังตัวอย่างงานของ Nair และคณะ [3] ที่อาศัยสถาปัตยกรรมการเรียนรู้เชิงลึกแบบ Deep-Convolutional Neural Network (D-CNN) หรืองานของ Xu และคณะ [4] ที่จับคู่ภาพคนร้ายในฐานข้อมูลจากภาพร่าง โดยใช้แบบจำลองการเรียนรู้เชิงลึกที่ชื่อ VGG อย่างไรก็ตามงานทั้งสองนี้ยังต้องอาศัยทักษะและความเชี่ยวชาญของนักร่างภาพเพื่อสร้างภาพต้นแบบที่มีคุณภาพดี

ในงานวิจัยชิ้นนี้ ผู้วิจัยทดลองนำการระบุตัวคนร้ายวิธีที่ 1 มาพัฒนาต่อยอด โดยสร้างระบบปัญญาประดิษฐ์ต้นแบบที่ผสมผสานความสามารถในการเข้าใจภาษาข้อความ (Text understanding) และความสามารถในการวาดภาพใบหน้าคน (Face image synthesis) ของปัญญาประดิษฐ์เข้าด้วยกันในระบบเดียว ผู้วิจัยหวังว่าผลงานวิจัยชิ้นนี้จะสามารถถูกนำไปใช้ต่อยอดและช่วยอำนวยความสะดวกให้ผู้พิทักษ์สันติราษฎร์สามารถตามจับตัวคนร้ายหรือผู้ต้องสงสัยได้อย่างรวดเร็วและมีประสิทธิภาพมากขึ้น

Related Works

ในปัจจุบันหนึ่งในแบบจำลองการเรียนรู้เชิงลึกที่นิยมใช้ในการสร้างภาพใบหน้าบุคคลจากคำบรรยายลักษณะคือ Generative Adversarial Networks (GAN) โดยจะต้องให้ข้อมูลที่เป็นชุดตัวเลขกับแบบจำลองเพื่อนำไปประมวลผลและสร้างรูปภาพที่มีความเกี่ยวข้องกับตัวเลขชุดดังกล่าว โดยแบบจำลองจะพยายามเรียนรู้ที่จะสร้างภาพให้ตรงกับข้อมูลที่ได้รับมามากที่สุด ซึ่งการจะทำให้แบบจำลองสร้างรูปภาพใบหน้าคนได้ตามคำบรรยายลักษณะที่กำหนดนั้นประกอบด้วยการทำงาน 3 ส่วนสำคัญดังต่อไปนี้

ส่วนแปลงการคำบรรยายลักษณะใบหน้า

เป็นการแปลงข้อมูลตัวอักษรให้อยู่ในรูปชุดตัวเลข (Latent space) เพื่อให้แบบจำลองนำชุดตัวเลขนี้ไปสร้างรูปภาพใบหน้าบุคคลที่มีลักษณะตรงตามคำบรรยาย การแปลงข้อมูลคำบรรยายลักษณะใบหน้ามนุษย์นั้นเป็นงานที่ยากและท้าทายต้องอาศัยเครื่องมือที่สามารถทำความเข้าใจกับความซับซ้อนของภาษาที่ใช้บรรยายได้ ตัวอย่างเช่น Wadhawan และคณะ [5] ได้นำ Embedding GLoVe ที่ถูกสร้างมาจากฐานข้อมูลคำศัพท์ขนาดใหญ่มาใช้ เพื่อช่วยให้แบบจำลองเข้าใจคำบรรยายและสร้างชุดตัวเลขดังกล่าวได้ดีขึ้น หรือในงานของ Kotian และคณะ [2] มีการใช้เทคนิค Embedding รวมกับแบบจำลองการเรียนรู้เชิงลึกแบบ Long Short-Term Memory (LSTM) เพื่อแปลงข้อมูลคำบรรยายเป็นชุดตัวเลขเพื่อเป็นข้อมูลตั้งต้น

ในการสร้างรูปภาพใบหน้าบุคคล

ในส่วนของ Sun และคณะ [6] ได้ใช้แบบจำลองการเรียนรู้เชิงลึกแบบ Bidirectional LSTM (Bi-LSTM) รวมกับการทำ Normalization ข้อมูลก่อนส่งให้ขั้นตอนถัดไป ทั้งนี้เพื่อจัดการคำบรรยายในจำนวนมากได้ดียิ่งขึ้น นอกจากนี้ Oza และคณะ [7] รวมถึง Wang และคณะ [8] มีการใช้แบบจำลองการเรียนรู้เชิงลึกที่ซับซ้อนขึ้นอย่าง Bidirectional Encoder Representations from Transformers (BERT) เข้ามาช่วยในการสกัดข้อมูลจากคำบรรยายใบหน้า เพื่อให้ได้ชุดข้อมูลตัวเลขที่ดี เช่นเดียวกับ Sun และคณะ [9] รวมถึง Li และคณะ [10] ที่มีการประยุกต์ใช้แบบจำลองการเรียนรู้เชิงลึกขั้นสูงอย่าง Contrastive Language–Image Pre-training (CLIP) มาช่วยในการแปลงข้อมูลคำบรรยายให้กลายเป็นชุดตัวเลข

ส่วนการสร้างภาพใบหน้า

การสร้างรูปภาพใบหน้าของแบบจำลองประเภท GAN นั้นโดยมากจะเริ่มต้นจากการสุ่ม Noise vector และลักษณะใบหน้า (Feature) เพื่อสร้างภาพใบหน้าขึ้นมา ขั้นตอนการสุ่ม Noise vector นั้นมักใช้การสุ่มจากการกระจายแบบ Gaussian โดยที่ Noise vector และ Feature เหล่านั้นจะโดนป้อนผ่านแบบจำลองย่อยภายในที่เรียกว่า Generator อันเป็นโครงข่ายประสาท (Neural network) ที่มีเลเยอร์ประเภท Convolution หลายชั้นที่ทำหน้าที่สร้างรูปใบหน้าขึ้นมา หลังจากนั้นภาพใบหน้าที่สร้างขึ้นจาก Generator จะถูกป้อนเข้าไปให้กับแบบจำลองย่อยภายในอีกตัวที่เรียกว่า Discriminator เพื่อให้ Discriminator ตรวจสอบว่าภาพที่ป้อนเข้าไปนั้นเป็นภาพปลอม (Fake image หรือ Artificial image) ที่ถูกสร้างขึ้นจาก Generator หรือเป็นภาพจริง (Real image) ของใบหน้าคนที่มีอยู่จริง ทั้งนี้การเรียนรู้ของ Generator จะค่อย ๆ เกิดขึ้นจากข้อมูลการตรวจสอบของ Discriminator ที่ป้อนกลับ (Backpropagate) ไปสอน Generator อีกทีทำให้ Generator สามารถพัฒนาตัวเองและสร้างภาพใบหน้าที่ใกล้เคียงกับภาพใบหน้าจริงได้มากขึ้นเรื่อย ๆ นอกจากนี้ Discriminator เองก็ต้องมีการถูกทำ Pre-train ด้วยข้อมูลลักษณะคำบรรยายพร้อมกับรูปภาพก่อนจะตรวจสอบภาพที่ได้รับจาก Generator

แบบจำลอง GAN ในปัจจุบันก้าวหน้าไปมากจนสามารถควบคุมภาพของสิ่งที่ต้องการสังเคราะห์ขึ้นมาใหม่ได้ อย่างไรก็ตาม Hermosilla และคณะ [11] พบว่าแบบจำลอง GAN ระดับชั้นนำของโลกแต่ละตัวนั้นมีวัตถุประสงค์การใช้งานที่แตกต่างกันไป เช่น แบบจำลอง Deep Convolutional GAN (DCGAN) มีวัตถุประสงค์ในการปรับปรุงความละเอียดและผลลัพธ์คุณภาพของภาพที่ได้จาก GAN ดั้งเดิม (Vanilla GAN) ทว่าความละเอียด (Resolution) ของภาพที่สังเคราะห์ได้ก็ยังจำกัดที่ 64x64 พิกเซล ในขณะที่แบบจำลอง The CoupledGAN (CoGAN) นั้นแตกต่างจาก GAN ทั่วไปเพราะถูกเพิ่มให้มี Generator ถึง 2 ตัวแต่ยังคงมี Discriminator เพียง 1 ตัวเท่าเดิม ด้วยวิธีนี้ Generator จึงสามารถสร้างภาพที่มีลักษณะแตกต่างกันมากได้ เช่น ใบหน้า, สีผม, สีตา, ผิว แต่ทั้งนี้ความละเอียดที่ได้ก็ยังมีข้อจำกัด

ในขณะที่ Sun และคณะ [9] กล่าวถึงแบบจำลอง Deep Fusion GAN (DFGAN) [12] ว่าถูกสร้างมาเพื่อช่วยแก้ปัญหาด้านทรัพยากรการคำนวณที่เพิ่มสูงขึ้นจากการที่ GAN มี Generator อยู่ภายในหลายตัว

วิธีที่นำเสนอคือการปรับ Noise (Affine transformation) ให้เหมาะสมและใช้เพียง Generator ตัวเดียว เพื่อสร้างรูปภาพ อย่างไรก็ตามวิธีนี้ยังมีข้อจำกัดในเรื่องของคุณภาพของรูปภาพที่มีความละเอียดเพียง 256 x256 พิกเซล ในส่วนของ Wang และคณะ [8] พบว่าความละเอียดของภาพที่ถูกสร้างโดยแบบจำลอง GAN นั้นถูกปรับปรุงให้ดีขึ้นอย่างมีนัยสำคัญจากหลักการของแบบจำลอง GAN ที่ชื่อ StyleGAN ของบริษัท NVIDIA โดยหลักการนี้ทำให้สามารถสร้างรูปภาพที่มีความละเอียดสูงถึง 1024 x1024 พิกเซล และด้วยความละเอียดขนาดนี้ก็ทำให้ภาพที่ได้มีความเป็นธรรมชาติมากขึ้น

ส่วนการสร้างภาพใบหน้าจากข้อความ

ในอดีตแบบจำลองสำหรับประมวลผลข้อความและประมวลผลรูปภาพนั้นมักเป็นแบบจำลองสองตัวแยกกัน แต่ละตัวถูกฝึกสอนมาบนข้อความหรือรูปภาพเพียงอย่างใดอย่างหนึ่ง การจะนำแบบจำลองสองส่วนที่ฝึกสอนมาคนละแบบและด้วยคนละชนิดข้อมูลมาใช้ร่วมกันให้ได้ผลดีนั้น จำเป็นที่จะต้องศึกษาเทคนิคการปรับจูนหรือเทคนิคอื่นที่สามารถเชื่อมต่อความเข้าใจของข้อมูลทั้งสองประเภทให้ไปในทิศทางเดียวกันได้ ในงานของ Patashnik และคณะ [13] พบว่าแบบจำลองชื่อ StyleGAN มีความสามารถในการแยกคุณสมบัติ (Disentanglement properties) ของภาพเป็นอย่างดี เหมาะกับการนำไปต่อยอดใช้จัดการรูปภาพ (Image manipulation) หลายแบบไม่ว่าจะเป็นภาพที่สร้างขึ้นโดยจำลอง (Synthetic) หรือภาพถ่ายก็ตาม โดยงานนี้ได้นำแบบจำลองชื่อ Contrastive Language-Image Pre-training Model (CLIP) มาช่วยทำการจัดการกับรูปภาพผ่านข้อความ (Text-based semantic image manipulation) แบบที่ไม่มีข้อจำกัดอยู่กับ preset สำเร็จรูปที่กำหนดไว้ อีกทั้งใช้ CLIP เพื่อลดภาระงานการสร้างคำอธิบายใหม่

ในงานของ Deorukhkar และคณะ [14] มีการทดลองใช้ชุดข้อมูล CelebA มาเข้ารหัส (Encode) คู่ของ Image-Text แต่ละคู่โดยใช้แบบจำลองชื่อ Sentence BERT จากนั้นนำเวกเตอร์ความหมาย (Semantic vectors) ที่ได้มาเป็นอินพุตส่งต่อไปยัง Generator และ Discriminator อีกทีหนึ่ง งานนี้ใช้ตัวชี้วัดในการวัดผลได้แก่ FID, Clean FID และ Inception score ผลการทดลองพบว่า Self-attention GAN และ DFGAN สามารถสร้างรูปคุณภาพสูงได้แต่ก็ใช้เวลาฝึกสอนค่อนข้างนานเนื่องจากความซับซ้อนของตัวแบบจำลองเอง ในขณะที่ DCGAN ให้ผลลัพธ์ที่ความแม่นยำใกล้เคียงกันแต่ใช้เวลาในการคำนวณน้อยกว่า ในส่วนของ DFGAN และ SAGAN นั้นสร้างรูปภาพได้หลากหลายกว่าวิธีการก่อนหน้าที่ได้กล่าวไป

ที่ผ่านมามีความพยายามนำแบบจำลองชั้นนำของโลกสำหรับศาสตร์ด้านการประมวลผลภาษาธรรมชาติ (Natural language processing หรือ NLP) และศาสตร์ด้านคอมพิวเตอร์วิทัศน์ (Computer vision) มาผนวกรวมหรือใช้งานร่วมกัน เช่น การนำ Word embedding จากศาสตร์ NLP ทั้ง Glove, ELMo หรือ BERT มาใช้ร่วมกับ Generator ของแบบจำลอง GAN ในปี ค.ศ. 2021 มีการเกิดขึ้นของแบบจำลองใหม่ความสามารถสูงอย่าง CLIP ทำให้เกิดเป็นกระแสความนิยมของการสร้างงานที่ใช้แบบจำลอง CLIP เป็นสะพานเชื่อมระหว่างทั้งสองศาสตร์มากขึ้น ซึ่งรวมถึงการใช้แบบจำลอง CLIP ในการสร้างรูปภาพจากข้อความด้วย ซึ่งหนึ่งในเทคนิคที่ได้รับความนิยมคือการนำแบบจำลอง CLIP มาใช้ร่วมกับแบบจำลอง StyleGAN เนื่องจาก

StyleGAN มีความสามารถในการสร้างภาพที่ความละเอียดสูงและสมจริงมากกว่า GAN ทั่วไป อย่างไรก็ตาม ผู้วิจัยมีความเห็นว่าการรวมตัวกันของ 2 แบบจำลองความสามารถสูงนี้ยังมีจุดที่ยังสามารถพัฒนาต่อยอดได้อีก คือ การค้นคว้าและพัฒนาว่าจะทำอย่างไรให้ความหลากหลายของคำบรรยายเชิงเปรียบเทียบ (Slang words) นั้นสามารถถูกสร้างเป็นภาพใบหน้าได้

โดยสรุปจะเห็นว่าปัจจุบันมีงานวิจัยเรื่องการแปลงข้อความเป็นภาพอยู่จำนวนมาก จากการทบทวนวรรณกรรมผู้วิจัยพบว่างานที่ใช้แบบจำลอง GAN นั้นส่วนใหญ่นิยมใช้ StyleGAN ซึ่งมีประสิทธิภาพดีและให้ผลลัพธ์การสร้างรูปภาพเป็นที่น่าพอใจ แต่สำหรับงานวิจัยที่เกี่ยวกับ Text-to-feature นั้น ยังไม่พบว่ามีวิธีใดที่ได้ผลลัพธ์ดีอย่างเป็นเอกฉันท์โดยเฉพาะในบริบทของคำบรรยายลักษณะใบหน้าคน จำเป็นต้องมีการวิจัยเพิ่มเติมเพื่อให้ได้ประสิทธิภาพและความชัดเจนในส่วนนี้มากขึ้น ดังนั้นในงานวิจัยชิ้นนี้ผู้วิจัยจึงจะเน้นศึกษาในส่วนของ Text-to-feature หรือ Feature extractor ต่อยอดจากงานวิจัยในอดีต โดยผู้วิจัยเลือกนำ StyleGAN มาใช้และทำการพัฒนาต่อยอดเฉพาะในส่วนการแปลงข้อมูลคำบรรยายให้เป็นข้อมูลชุดตัวเลขที่จะถูกนำไปใช้เป็นเงื่อนไขในการสร้างภาพใบหน้าโดย StyleGAN ทั้งนี้โดยจะมีการทดลองนำเทคนิคต่าง ๆ มาปรับใช้กับ StyleGAN เช่น CLIP, ELMo และ BERT เพื่อหาส่วนผสมของผลลัพธ์ที่ดีที่สุดซึ่งใช้ทรัพยากรการคำนวณที่สมเหตุสมผล

Research Methodology

งานวิจัยชิ้นนี้ใช้แพลตฟอร์ม Google Colab Pro+ ในการพัฒนาแบบจำลองโดย Graphics Processing Unit (GPU) ที่ใช้คือ NVIDIA V100 แพ็กเกจการเขียนโปรแกรมที่เกี่ยวข้องกับการพัฒนา ได้แก่ Torch 1.9.1, Torchvision 0.10.1 และ Cuda11 ซึ่งรองรับการทำงานทั้งกับแบบจำลอง StyleGAN3 [15] และ CLIP [16] ทั้งนี้ผู้วิจัยเลือกใช้แบบจำลอง StyleGAN3-ffhqu-256x256 ซึ่งดาวน์โหลดจาก Github-NVlabs/stylegan3 ในส่วนของแบบจำลอง CLIP นั้นผู้วิจัยใช้เวอร์ชัน ViT-B/32 ซึ่งดาวน์โหลดจาก Github-openai/CLIP นอกจากนี้ผู้วิจัยยังติดตั้งแพ็กเกจ Sentence-transformers 1.1.0 เพื่อใช้ในการตัดคำ (Tokenization) และเพื่อเรียกใช้งาน Bert-base-nli-mean-tokens สำหรับแปลงข้อมูลคำที่ตัดแล้วให้กลายเป็นเวกเตอร์ด้วย แบบจำลองการเรียนรู้เชิงลึกที่ทำการทดลองในงานวิจัยชิ้นนี้มี 4 แบบจำลองดังรายละเอียดในหัวข้อย่อย Model 1 ถึง Model 4 ต่อไปนี้

ทั้งนี้ การตั้งค่า (Hyperparameter tuning) ต่าง ๆ ในผลงานวิจัยนี้เกิดจากการสุ่มทดลองของผู้วิจัยเองภายใต้ทรัพยากรการฝึกสอนแบบจำลองที่มีจำกัด ทั้งนี้เมื่อพิจารณาจากจำนวน Hyperparameter ที่มีอยู่มาก ทั้งในตัวของแบบจำลองการเรียนรู้เชิงลึกขนาดใหญ่แต่ละตัวที่ใช้ในงานวิจัยนี้ ทั้งในขั้นของการนำแบบจำลองการเรียนรู้เชิงลึกขนาดใหญ่หลายตัวมาเชื่อมต่อกัน และทั้งในขั้นของการฝึกสอนแบบจำลองการเรียนรู้เชิงลึกที่เชื่อมต่อกันแล้ว กล่าวได้ว่าความเป็นไปได้ในการทดลอง Hyperparameter tuning ในงานวิจัยนี้มีจำนวนมากนับอนันต์ วิธีการอย่าง Grid search จึงไม่อาจการันตีว่าผู้วิจัยจะสามารถสำรวจความเป็นไปได้นับอนันต์นี้ได้อย่างถ้วนทั่ว ด้วยเหตุนี้การทดลองปรับค่าแบบสุ่มจึงเป็นวิธีการที่ผู้วิจัยเลือกใช้

Model 1: StyleGAN3 + CLIP

ในขณะที่แบบจำลอง StyleGAN3 [15] สามารถสร้างรูปภาพใบหน้าบุคคลได้อย่างคมชัดและสมจริง และแบบจำลอง CLIP [16] สามารถหาตัวเลขความคล้ายคลึงระหว่างข้อความและรูปภาพได้ ในการทดลองแรกนี้ผู้วิจัยจะลองนำทั้งสองแบบจำลองมาผสานกันภายใต้บริบทของการสร้างภาพใบหน้าบุคคลจากข้อความบรรยายลักษณะ โดยแบบจำลอง StyleGAN3 ที่ผู้วิจัยเลือกมาใช้คือ StyleGAN3 ที่ถูกฝึกสอนมาบนชุดข้อมูล Flickr-Faces-HQ Dataset (FFHQ) ซึ่งจะสร้างรูปภาพใบหน้าบุคคลขนาด 256 x 256 พิกเซล

ดังโครงสร้างแสดงใน Figure 1 คำบรรยายลักษณะใบหน้าบุคคลจะถูกเปลี่ยนให้กลายเป็นเวกเตอร์ตัวแทนขนาด 768 มิติโดยแบบจำลอง BERT [17] แต่เนื่องจากแบบจำลอง StyleGAN3 นั้นถูกออกแบบมาให้สร้างรูปจากเวกเตอร์อินพุตขนาด 512 มิติ เพื่อปรับขนาดของเวกเตอร์ให้สอดคล้องกันผู้วิจัยจึงทำการเพิ่ม Linear layer จำนวน 2 ชั้นแทรกลงไปเพื่อแปลงจำนวนมิติของเวกเตอร์ดังกล่าวจาก 768 ให้กลายเป็น 512 เมื่อได้เวกเตอร์ในขนาดที่ต้องการแล้ว การสร้างรูปภาพจะถูกดำเนินการต่อโดยใช้เพียงแบบจำลองย่อยส่วน Generator ที่อยู่ภายในแบบจำลองใหญ่ StyleGAN3 เมื่อได้ภาพที่สร้างจาก StyleGAN3 แล้ว ภาพที่ถูกสร้างขึ้นดังกล่าว (Fake image) และภาพจริงจากชุดข้อมูล (ภาพเป้าหมายหรือ Real image) จะถูกนำไปป้อนให้กับส่วน Image encoder ของแบบจำลอง CLIP ต่อไป

ฟังก์ชันสูญเสีย (Loss function) ที่ใช้ในการฝึกสอนแบบจำลองนี้ประกอบด้วยการคำนวณ 2 ส่วน ซึ่งผลลัพธ์จะถูกนำมาบวกรวมกัน การคำนวณส่วนแรกคือ Reconstruction loss (สมการที่ 1) เพื่อเทียบภาพเป้าหมายกับภาพที่สร้างขึ้นตรง ๆ ในแบบพิกเซลต่อพิกเซล และการคำนวณส่วนที่สองคือ Spherical Distance Loss (SD loss) [18] (สมการที่ 2) ซึ่งเปรียบเทียบเวกเตอร์ของภาพทั้งสองที่ได้จาก CLIP image encoder ในงานวิจัยชิ้นนี้แบบจำลองจะถูกฝึกสอนซ้ำเป็นจำนวน 100 epochs โดยใช้ Batch size ขนาด 16 ใช้ Optimizer แบบ AdamW ซึ่งมี Learning Rate เท่ากับ 0.03 และ beta2 เท่ากับ 0.999

$$\text{Reconstruction Loss} = \frac{\sqrt{\sum_w \sum_h (x_{wh} - y_{wh})^2}}{w \times h} \quad (1)$$

เมื่อ w และ h คือความกว้างและความสูงของภาพ ในขณะที่ x_{wh} และ y_{wh} หมายถึงข้อมูลของรูปภาพที่ GAN สร้างขึ้นมาและภาพเป้าหมายตามลำดับ ณ ตำแหน่งพิกเซลที่ (w, h)

$$\text{SD Loss} = \left(\arcsin\left(\frac{\|x' - y'\|_F}{2}\right)\right)^2 \times 2 \quad (2)$$

เมื่อ x' และ y' คือ เวกเตอร์จาก CLIP image encoder ที่ได้จากภาพที่สร้างขึ้นมาโดย GAN และภาพเป้าหมายตามลำดับ

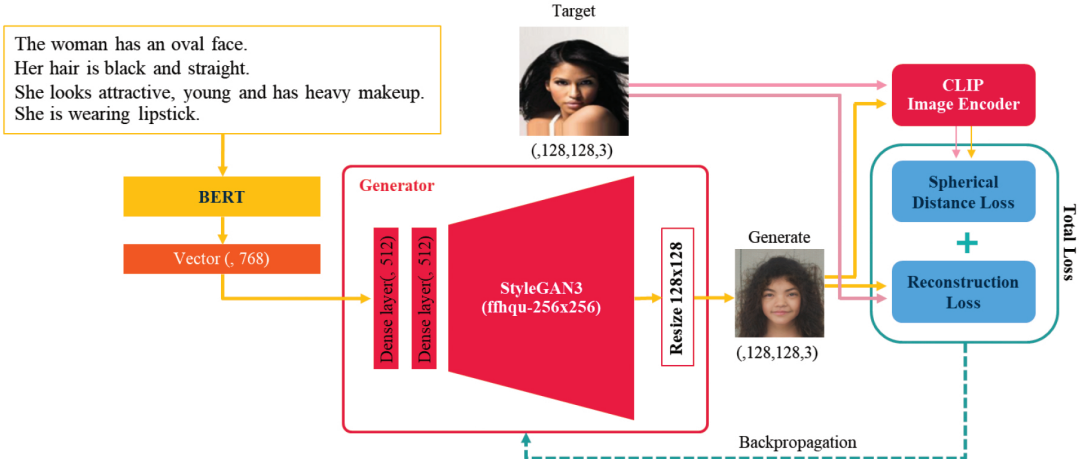

Figure 1 [Model 1] Architecture of StyleGAN3 + CLIP

Model 2: StyleGAN3 + CLIP (Fine-Tuning)

ในหัวข้อนี้ผู้วิจัยได้เพิ่มการฝึกสอนแบบ Fine-tuning ให้กับ CLIP โดยใช้ชุดข้อมูลรูปภาพใบหน้าคู่กับข้อความบรรยายจำนวน 50,000 คู่ สมมติฐานของผู้วิจัยคือการ Fine-tuning เพิ่มเติมนี้น่าจะช่วยให้ CLIP มีความสามารถที่เฉพาะเจาะจงกับงานการสร้างภาพใบหน้าคนจากข้อความบรรยายลักษณะมากขึ้น เมื่อเทียบกับการใช้ CLIP แบบดั้งเดิมที่ถูกฝึกสอนมากับรูปภาพและคำบรรยายที่ไม่ได้เจาะจงเฉพาะกับใบหน้าคน การตั้งค่าสำหรับการ Fine-tuning ส่วนของ CLIP นี้ผู้วิจัยใช้ Optimizer คือ Adam ที่มี Learning Rate เท่ากับ 0.00005 และตั้งค่า beta1 เท่ากับ 0.9 และ beta2 เท่ากับ 0.98 และ eps เท่ากับ 0.000001 ส่วน weight decay เท่ากับ 0.2 ในส่วนของฟังก์ชันสูญเสียใช้สมการ Cross entropy และทำการฝึกสอนแบบจำลองด้วย Batch size ขนาด 64 ซ้ำทั้งหมด 30 epochs ทั้งนี้ผู้วิจัยเลือกบันทึกเก็บชุดของ Weights เฉพาะใน Epoch ที่มีค่าสูญเสีย (Loss) ต่ำที่สุดเท่านั้น และนำ Weight ของ CLIP ชุดที่ได้มาใช้สำหรับฝึกสอนส่วน Generator ต่อไป

ในส่วนของการสร้าง Generator นั้น จาก Figure 2 จะเห็นว่า Generator ของการทดลองนี้จะรับอินพุตซึ่งเป็นเวกเตอร์ตัวเลขขนาด 512 มิติ โดยอินพุตดังกล่าวจะผ่านเข้าสู่ Linear layer จำนวน 1 ชั้น ที่สร้างเอาต์พุตขนาด 512 มิติออกไปเพื่อผ่านชั้นของ Batch Normalization และผ่าน Leaky ReLU ที่เป็นฟังก์ชันกระตุ้น (Activation function) อีกทีหนึ่ง ผลลัพธ์จากขั้น Leaky ReLU จะเป็นเวกเตอร์ตัวเลขขนาด 512 มิติที่ถูกนำไปบวกเข้ากับเวกเตอร์ Noise ที่มีขนาดเท่ากันซึ่งเกิดจากการสุ่ม เวกเตอร์ผลลัพธ์การบวกที่มีขนาดเท่าเดิมคือ 512 มิตินั้นจะถูกนำไป Mapping เข้ากับ Latent space ของ StyleGAN3 (ในการทดลองนี้ผู้วิจัยเลือกใช้ StyleGAN3 ซึ่งถูก Pre-train มาบนชุดข้อมูล FFHQ) ผลลัพธ์การสร้างภาพของ StyleGAN3 จะได้ภาพขนาด 256 x 256 พิกเซลออกมา โดยภาพนี้จะถูกนำไปย่อขนาด (Resize) ให้เหลือขนาด 128 x 128 พิกเซล ทั้งนี้ในส่วนของ Generator ผู้วิจัยได้ตั้งค่าที่เกี่ยวกับการฝึกสอนไว้ คือ ใช้ Optimizer เป็น AdamW มี Learning Rate เท่ากับ 0.025 ตั้งค่า beta1 เท่ากับ 0.9 และ beta2 เท่ากับ 0.999 และมี Weight Decay เท่ากับ 0.01

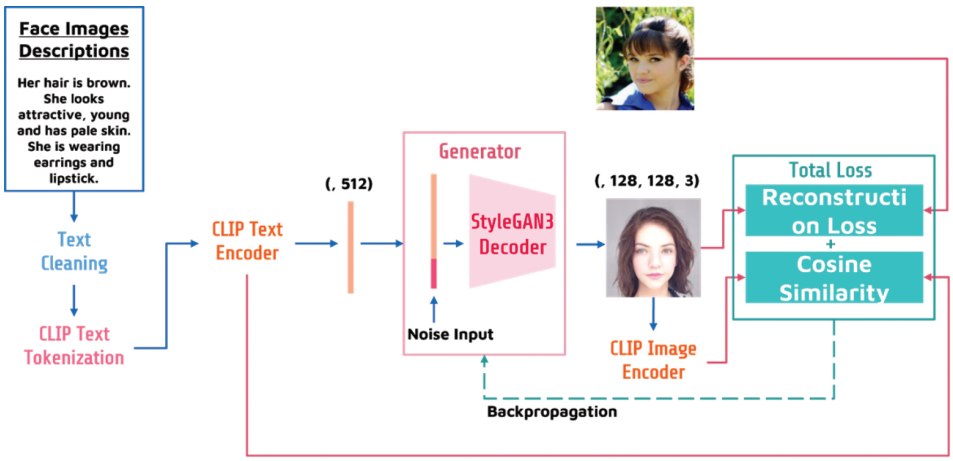

Figure 2 [Model 2] Architecture of StyleGAN3 + CLIP (Fine-Tuning)

การฝึกสอน Generator ดังแสดงใน Figure 2 นั้น ผู้วิจัยนำ CLIP ที่ผ่านการ Fine-tuning มาใช้ โดยนำอินพุตซึ่งเป็นข้อความบรรยายลักษณะมาตัดคำด้วย CLIP text tokenizer และนำผลลัพธ์ที่ได้เข้าสู่ CLIP text encoder เพื่อเปลี่ยนข้อมูลให้กลายเป็นเวกเตอร์ขนาด 512 มิติซึ่งพร้อมต่อการส่งเข้า Generator ในลำดับถัดไป ผลลัพธ์ภาพขนาด 128 x 128 พิกเซลที่สร้างจาก Generator ถูกนำมาใช้คำนวณฟังก์ชันสูญเสีย ซึ่งเป็นค่าเฉลี่ยเท่า ๆ กันระหว่างการคำนวณ 2 ส่วน ได้แก่ ส่วนการคำนวณ Reconstruction loss เทียบกับภาพเป้าหมาย (เหมือนกับหัวข้อ Model 1) และส่วนการคำนวณ Cosine similarity โดยในการคำนวณส่วนที่ 2 นั้นแทนที่จะเป็นการเปรียบเทียบเวกเตอร์ระหว่างภาพ 2 ภาพเหมือนกับในหัวข้อ Model 1 ผู้วิจัยเปลี่ยนมาเป็นการเปรียบเทียบระหว่างเวกเตอร์ของภาพและเวกเตอร์ของข้อความแทน โดยภาพที่ Generator สร้างขึ้นจะถูกนำไปผ่าน CLIP image encoder แปลงให้กลายเป็นเวกเตอร์และถูกนำไปเทียบกับเวกเตอร์ของข้อความบรรยายลักษณะที่ได้จาก CLIP text encoder ฟังก์ชันสูญเสียแบบใหม่นี้ถูกใช้เพื่อส่งกลับ (Backpropagate) ค่าความสูญเสียที่คำนวณได้ไปปรับปรุง Generator ให้สร้างภาพคุณภาพดีขึ้นที่ตรงกับข้อความบรรยายลักษณะมากขึ้น ทั้งนี้การฝึกสอนนี้ผู้วิจัยใช้ Batch size เท่ากับ 16 และทำการฝึกสอนซ้ำทั้งหมด 50 epochs

Model 3: Deep Convolution (DC) + CLIP (Fine-Tuning)

การทดลองนี้ผู้วิจัยได้แรงบันดาลใจมาจาก Heusel และคณะ [19] โดยผู้วิจัยได้เปลี่ยนส่วน Discriminator มาเป็นการใช้ CLIP (Fine-tuning) แทน นอกจากนี้ผู้วิจัยยังนำ Weight ของ CLIP (Fine-tuning) มาใช้สำหรับฝึกสอน Generator ด้วย ทั้งนี้โครงสร้างภายในของ Generator จะยังคงรูปแบบดั้งเดิมของแบบจำลองที่ชื่อ DCGAN [20]

ส่วนของ Generator นั้นจะมีโครงสร้างดังแสดงใน Table 1 และถูกนำไปต่อเชื่อมกับส่วนอื่น ๆ เพื่อทำการฝึกสอนดังแสดงใน Figure 3 จากภาพจะเห็นว่าคำบรรยายลักษณะใบหน้าบุคคลจะถูกแบบจำลองชื่อ

BERT แปลงให้เป็นเวกเตอร์ที่มีขนาด 768 มิติเช่นเดียวกับหัวข้อ Model 1 จากนั้นเวกเตอร์นี้จะถูกส่งเข้าชั้น Linear layer เพื่อที่จะแปลงมิติของเวกเตอร์จาก 768 เป็น 256 มิติตามที่ Generator ในการทดลองนี้ต้องการ โดยที่จะมีการทำ Batch normalization และใช้ Leaky ReLU เป็นฟังก์ชันกระตุ้นด้วย เมื่อนำเวกเตอร์ที่ได้นี้มารวมกับเวกเตอร์ Noise ขนาด 100 มิติที่สร้างขึ้นก็จะได้อินพุตที่พร้อมส่งให้กับ Generator สำหรับรูปภาพที่ถูกสร้างขึ้นโดย Generator นั้นจะถูกนำไปคำนวณฟังก์ชันสูญเสียเปรียบเทียบกับภาพเป้าหมายโดยใช้การคำนวณ 2 ส่วนเช่นเดียวกับในหัวข้อ Model 1

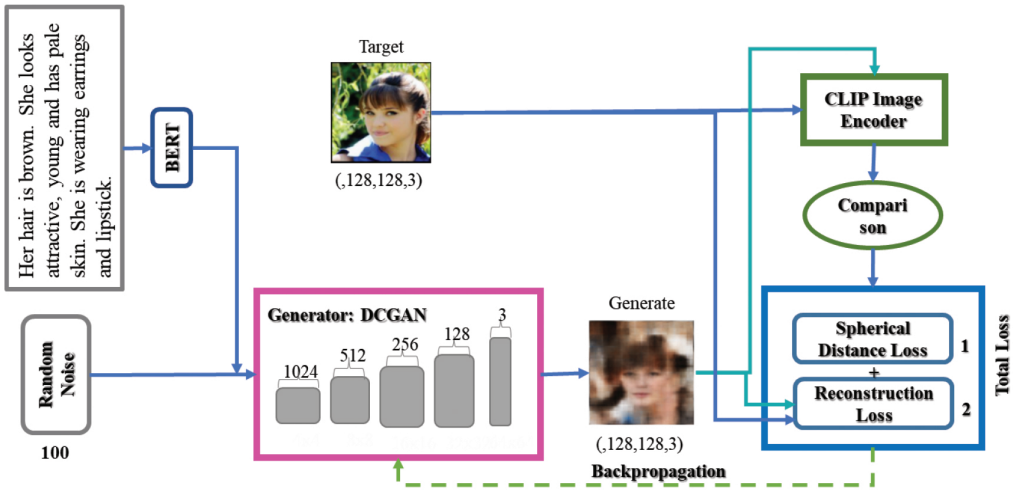

Figure 3 [Model 3] Architecture of DC + CLIP (Fine-tuning)

Table 1 Architecture and parameters of DC's generator

Layers	Input	Output	Kernel size	Stride	Padding	Activation function
Convolution1	356	1024	4*4	1	0	LeakyReLU
Batch normalization1	1024	N/A	N/A	N/A	N/A	N/A
Convolution2	1024	512	4*4	2	1	ReLU
Batch normalization2	512	N/A	N/A	N/A	N/A	N/A
Convolution3	512	256	4*4	2	1	ReLU
Batch normalization3	256	N/A	N/A	N/A	N/A	N/A
Convolution4	256	128	4*4	2	1	ReLU

Table 1 Architecture and parameters of DC's generator (Continued)

Layers	Input	Output	Kernel size	Stride	Padding	Activation function
Batch normalization4	128	N/A	N/A	N/A	N/A	N/A
Convolution5	128	128	4*4	2	1	ReLU
Batch normalization5	128	N/A	N/A	N/A	N/A	N/A
Convolution6	128	3	4*4	2	1	Tanh
Parameters	Learning rate: 0.0002 Batch size = 64, Epochs = 100 Betas = (0.5, 0.5)			Loss function: Reconstruction Loss and Spherical Loss Optimizer: Adam		

Model 4: Deep Fusion + CLIP (Fine-Tuning)

การทดลองนี้ได้รับแรงบันดาลใจของตัวโครงสร้างจาก Deep Fusion Generative Adversarial ของ Tao และคณะ [12] โดยผู้วิจัยใช้โครงสร้าง Deep Fusion ดั้งเดิมมาปรับปรุงเฉพาะในส่วน Discriminator โดยผู้วิจัยนำ CLIP (Fine-Tuning) ไปแทนที่ Discriminator ตัวดั้งเดิมได้ออกมาเป็นโครงสร้าง Deep Fusion + CLIP (Fine-Tuning) ดังแสดงใน Figure 4

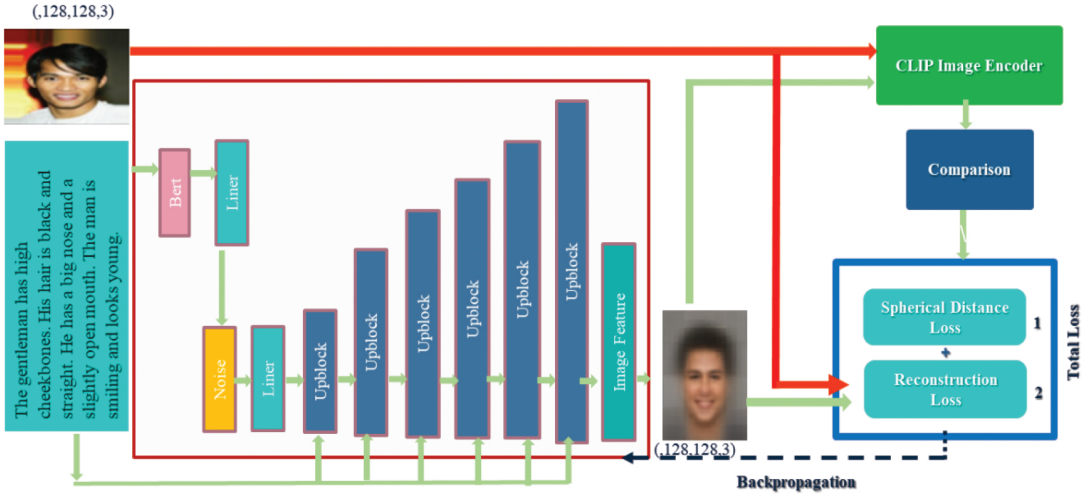

Figure 4 [Model 4] Architecture of DF + CLIP (Fine-Tuning)

โครงสร้างของ Generator แสดงดัง Table 2 โดยจาก Figure 4 จะเห็นว่าหลักการทำงานเริ่มต้นจาก BERT ที่แปลงคำบรรยายลักษณะใบหน้าให้กลายเป็นเวกเตอร์ความหมาย (Semantic vector) ซึ่งประกอบด้วยตัวเลข

ทั้งหมด จากนั้น Semantic vector นี้จะถูกนำไปเข้า Linear layer เพื่อลดขนาดจาก 768 มิติให้เหลือ 256 มิติ ผลที่ได้ถูกนำไปรวมกับเวกเตอร์ Noise ขนาด 100 มิติ แล้วจึงนำเข้า Linear layer อีกครั้งก่อนที่จะถูกส่งเข้าไปในส่วนของ Upblock0 จนถึง Upblock5 ตามลำดับ ซึ่งการทำงานของ Upblock0 ถึง Upblock5 นี้จะมีค่าของ Semantic vector ขนาด 768 มิติที่ได้จาก BERT เป็นเงื่อนไขประกอบในการทำงานด้วย ผลลัพธ์จาก Upblock ที่ 5 จะถูกส่งไปสร้างเป็นข้อมูลของ Image feature สำหรับใช้สร้างภาพใบหน้าออกมา ภาพที่สร้างได้จะมีขนาด 128 x128 x 3 พิกเซล ทั้งนี้ภายใต้กระบวนการแต่ละ Upblock จะใช้ฟังก์ชันกระตุ้นคือ ReLU และใช้ Adam optimizer ที่กำหนด Beta1 = 0, Beta 2 = 0.9 โดยมีจำนวนรอบการฝึกสอนอยู่ที่ 10 epochs แต่ละ Epoch มีการฝึกสอน 5,000 iterations แต่ละครั้งใช้จำนวนภาพฝึกสอน 4 ภาพ ค่า Learning Rate คือ 0.0001

Table 2 Architecture of DF's generator

Layers	Input	Output	Kernel size	Stride	Padding	Activation function	Remark
Convolution1	512	512	3	1	1	N/A	Upblock (N=0,1,2,3,4,5)
Convolution2	512	512	3	1	1	N/A	
Linear 1	256	256	N/A	N/A	N/A	Relu	affine0 – gamma
Linear 2	256	512	N/A	N/A	N/A	N/A	
Linear 3	256	256	N/A	N/A	N/A	Relu	affine0 – beta
Linear 4	256	512	N/A	N/A	N/A	N/A	
Linear 5	256	256	N/A	N/A	N/A	Relu	affine1 – gamma
Linear 6	256	512	N/A	N/A	N/A	N/A	
Linear 7	256	256	N/A	N/A	N/A	Relu	affine1 – beta
Linear 8	256	512	N/A	N/A	N/A	N/A	
Linear 9	256	256	N/A	N/A	N/A	Relu	affine2 – gamma
Linear 10	256	512	N/A	N/A	N/A	N/A	
Linear 11	256	256	N/A	N/A	N/A	Relu	affine2 – beta
Linear 12	256	512	N/A	N/A	N/A	N/A	
Linear 13	256	256	N/A	N/A	N/A	Relu	affine3 – gamma
Linear 14	256	512	N/A	N/A	N/A	N/A	
Linear 15	256	256	N/A	N/A	N/A	Relu	affine3 – beta
Linear 16	256	512	N/A	N/A	N/A	N/A	

ในส่วนของการดัดแปลง Discriminator เดิมมาเป็น CLIP (Fine-tuning) โดยผู้วิจัยนำ Weight ของ CLIP (Fine-tuning) มาใช้นั้น ในการทดลองนี้หน้าที่ของ CLIP (Fine-tuning) คล้ายกันกับในหัวข้อ Model 1 คือผู้วิจัยใช้เพียง CLIP image encode เพื่อแปลงภาพที่สร้างจาก Generator และภาพเป้าหมายให้กลายเป็นเวกเตอร์ จากนั้นจึงนำภาพทั้งสองมาคำนวณฟังก์ชันสูญเสียเทียบกันด้วย Spherical Distance Loss และ Reconstruction Loss เช่นเดียวกับหัวข้อ Model 1

Experimental Results

ในการวัดคุณภาพของภาพที่สร้างจากแต่ละแบบจำลองนั้น Fréchet Inception Distance (FID) ถือเป็นหนึ่งในตัวประเมินประสิทธิภาพที่นิยมใช้กันอย่างกว้างขวางในงานวิจัย สำหรับงานวิจัยชิ้นนี้ผู้วิจัยทำการวัดผลเชิงปริมาณด้วยค่า FID โดยนำภาพตัวอย่างที่สร้างจากแต่ละแบบจำลองจำนวนแบบจำลองละ 16 ภาพ ไปเปรียบเทียบกับภาพเป้าหมายที่เป็นภาพจริงและคำนวณค่า FID ได้ผลสรุปดัง Table 3 โดยจะเห็นว่าแบบจำลองที่ทำค่า FID ได้ต่ำที่สุด (ดีที่สุด) อันดับที่ 1 และ 2 คือ StyleGAN3+CLIP (Fine-Tuning) และ StyleGAN3+CLIP ตามลำดับ ซึ่งมีความแตกต่างอย่างมากจากค่า FID ของ Deep Fusion+CLIP (Fine-Tuning) และ Deep Convolution (DC)+CLIP (Fine-Tuning) เมื่อพิจารณาจากตัวอย่างใน Figure 5 จะเห็นถึงความแตกต่างของภาพที่สร้างจากแบบจำลองที่ใช้ StyleGAN3 เป็น Generator Network กับอีก 2 แบบจำลองที่เหลืออย่างชัดเจน

Table 3 FID scores of the four experimental models

Model	FID
StyleGAN3 + CLIP	242.45
StyleGAN3 + CLIP (Fine -Tuning)	**193.70**
Deep Convolution (DC) + CLIP (Fine -Tuning)	328.94
Deep Fusion + CLIP (Fine -Tuning)	309.15

Bold font represents the best score (the minimum FID score)

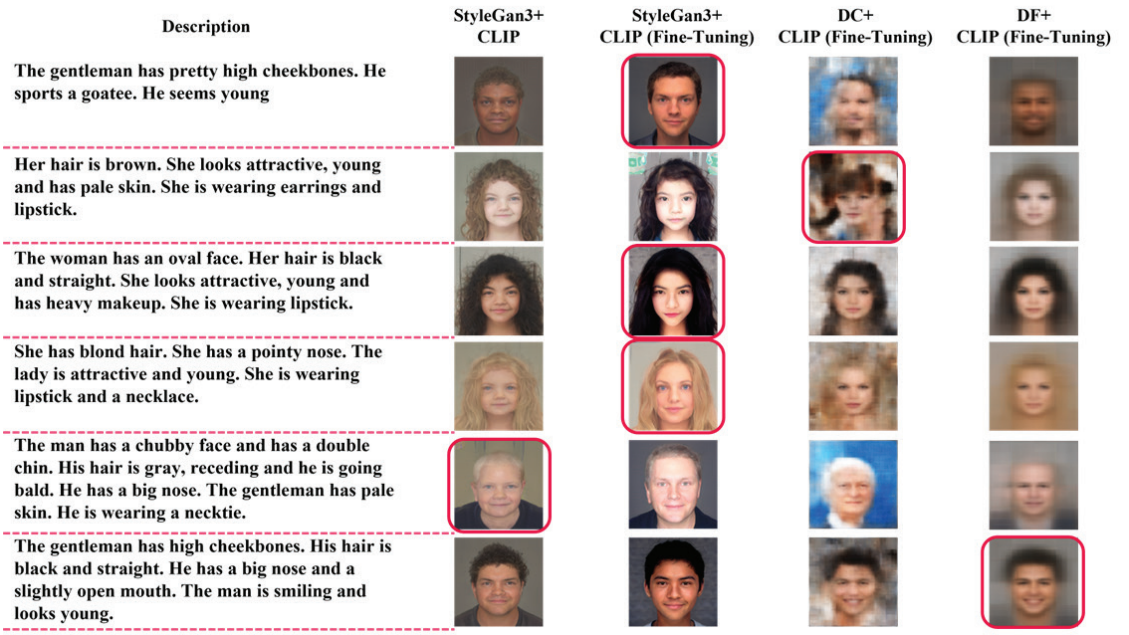

Figure 5 Comparison among facial images of individuals from each model. The solid pink frame in each row indicates the image selected as the best match to the description, based on the summary of an online opinion survey of 36 volunteers

หากพิจารณาจากตัวอย่างภาพที่แต่ละแบบจำลองสร้างขึ้นมาได้ใน Figure 5 แบบจำลอง StyleGAN3+CLIP สามารถสร้างภาพใบหน้าบุคคลที่มีความสมจริง มีเพศสีผิวและสีผมเป็นไปตามคำบรรยาย แต่ภาพใบหน้าบุคคลที่สร้างมาได้นั้นยังมีลักษณะที่คล้ายคลึงกัน ดังใน Figure 5 คำบรรยายที่ 2 และ 3 จะเห็นได้ว่ารูปภาพผู้หญิงที่สร้างขึ้นมาใบหน้าไม่แตกต่างกันมากนัก ไม่มีความหลากหลาย นอกจากนี้ยังไม่สามารถสร้างรายละเอียดเล็ก ๆ บนภาพได้ เช่น หนวด หรือเครื่องประดับ ดังใน Figure 5 คำบรรยายที่ 2 ที่บรรยายว่าผู้หญิงใส่ต่างหู แต่แบบจำลองก็ยังไม่สามารถสร้างต่างหูขึ้นมาได้

ในส่วนของ StyleGAN3+CLIP (Fine-Tuning) ที่มีการนำ CLIP แบบจำลองไปปรับปรุงเพิ่มนั้นสามารถสร้างภาพใบหน้าบุคคลที่มีความสมจริงตรงกับคำบรรยายและคมชัดมากขึ้น อีกทั้งยังสามารถสร้างใบหน้าที่หลากหลายได้ ดังจะเห็นว่าภาพที่สร้างจากแบบจำลอง StyleGAN3+CLIP (Fine-Tuning) มีความหลากหลายของใบหน้าอย่างชัดเจนและยังมีลักษณะเด่นต่าง ๆ เช่น สีตา ทรงผม สีผม การยิ้ม การแต่งหน้าละเอียดขึ้น แต่ก็ยังคงไม่สามารถสร้างรายละเอียดเล็ก ๆ เช่น หนวดเครา หรือเครื่องประดับได้เช่นเดิม ดังใน Figure 5 คำบรรยายที่ 3 ที่บอกว่าผู้หญิงทาลิปสติก แบบจำลองก็สามารถสร้างรูปภาพที่คล้ายกับผู้หญิงทาลิปสติกขึ้นมาได้ แต่ในคำบรรยายที่ 2 แบบจำลองนี้ก็ยังไม่สามารถสร้างต่างหูขึ้นมาได้

ในขณะที่อีก 2 แบบจำลองที่เหลือที่ไม่ได้ใช้ StyleGAN3 เป็น Generator นั้น ไม่สามารถสร้างภาพใบหน้าบุคคลที่มีความคมชัดในรายละเอียดของภาพได้ดีเท่ากับ 2 แบบจำลองข้างต้น แม้ว่าทั้ง DC+CLIP

(Fine-Tuning) และ DF+CLIP (Fine-Tuning) จะสามารถสร้างภาพใบหน้าบุคคลที่มีรายละเอียดสำคัญตรงตามคำบรรยาย เช่น เพศ ตา จมูก ปาก สีผม ได้ แต่ด้วยคุณภาพของภาพที่ไม่คมชัด ทำให้ส่วนใบหน้าและพื้นหลังมีความเบลอมองเห็นได้ไม่ชัดเจนยากที่จะระบุถึงตัวตนของบุคคล

Conclusion

ผู้วิจัยได้นำความสามารถของ StyleGAN3 และ CLIP เข้ามาช่วยในการสร้างรูปภาพจากคำบรรยาย (Text2Face Generation) โดยได้ทดลองสร้างแบบจำลองทั้งหมด 4 แบบจำลอง โดยมี 2 แบบจำลองที่ใช้ StyleGAN3 เป็น Generator คือ แบบจำลอง StyleGAN3+CLIP และแบบจำลอง StyleGAN3+CLIP (Fining-Tuning) ส่วนอีก 2 แบบจำลองที่ไม่ได้ใช้ StyleGAN3 เป็น Generator คือ แบบจำลอง DC+CLIP (Fine-Tuning) และแบบจำลอง DF+CLIP (Fine-Tuning) การวัดคุณภาพของภาพที่แบบจำลองสร้างขึ้นด้วยการใช้ตัวชี้วัดค่า FID ได้ข้อสรุปว่าแบบจำลองที่ใช้ StyleGAN3 เป็น Generator สามารถสร้างภาพที่มีคุณภาพดีกว่าและตรงตามคำบรรยายมากกว่า โดยคุณภาพของภาพจะออกมาดีที่สุดเมื่อใช้ StyleGAN3 คู่กับแบบจำลอง CLIP แบบจำลองที่ผ่านการ Fine-tuning เพิ่มเติมด้วยชุดข้อมูล CelebA HQ

ข้อสังเกตที่สำคัญของงานวิจัยชิ้นนี้อีกส่วนหนึ่งคือแบบจำลอง CLIP ที่ใช้จำเป็นจะต้องผ่านการ Fine-tuning บนชุดข้อมูลรูปภาพใบหน้าคู่กับคำบรรยายเสียก่อนถึงนำมาใช้งานได้ดี และแบบจำลอง CLIP ก็มีข้อดีที่สามารถใช้งานได้ง่ายสามารถนำมาใช้คู่กับ Generator ใด ๆ แทนที่ Discriminator ได้ ช่วยลดขั้นตอนในการฝึกสอนแบบจำลองและลดปัญหาที่เกิดจากการสอนทั้ง Generator และ Discriminator ไปพร้อมกัน อีกทั้ง CLIP ยังสามารถทำ Text encoder ได้ด้วยตัวเองทำให้ไม่จำเป็นต้องพึ่งพาแบบจำลองด้าน NLP (เช่น BERT) เพื่อมาทำหน้าที่ Text encoder แยกอีกตัวหนึ่ง นอกจากนี้การเลือกฟังก์ชันสูญเสียก็มีส่วนสำคัญที่ทำให้การฝึกสอนแบบจำลองเป็นไปอย่างมีประสิทธิภาพ ซึ่งจากงานวิจัยชิ้นนี้พบว่า Reconstruction loss สามารถสร้างผลลัพธ์จากการฝึกสอนได้ดีที่สุด อย่างไรก็ตาม แม้แบบจำลองของงานวิจัยนี้จะสามารถสร้างรูปภาพที่ตรงกับข้อความบรรยายภาษาอังกฤษได้ โดยสามารถเห็นลักษณะเด่น เช่น เพศ สีตา สีผม ทรงผม การยิ้ม การแต่งหน้าได้ชัดเจน อีกทั้งภาพยังมีความคมชัดใกล้เคียงกับภาพใบหน้าของบุคคลจริง แต่ก็ยังไม่สามารถสร้างรายละเอียดเล็ก ๆ เช่น หนวดเครา หรือเครื่องประดับได้ จำเป็นจะต้องมีการศึกษาพัฒนาในจุดนี้ต่อไป

งานวิจัยชิ้นนี้ถือเป็นรูปแบบหนึ่งของปัญญาประดิษฐ์แบบรู้สร้าง (Generative AI) ประเภทแบบจำลองซึ่งทำหน้าที่สร้างรูปภาพจากข้อความ (Text-to-image model) หากเปรียบเทียบกับแบบจำลองลักษณะนี้ซึ่งเป็นที่รู้จักแพร่หลายอย่าง DALL-E, Midjourney และ Stable Diffusion ซึ่งนิยมใช้สถาปัตยกรรมภายในเป็น Diffusion model แบบจำลองของงานวิจัยชิ้นนี้มีความแตกต่างตรงที่ใช้สถาปัตยกรรมของ StyleGAN คู่กับ CLIP อีกทั้งงานวิจัยชิ้นนี้ยังเพิ่มการฝึกสอนแบบจำลองในขอบเขตที่เฉพาะเจาะจงกับคำบรรยายลักษณะใบหน้าและการสร้างภาพใบหน้าบุคคลเท่านั้น ทั้งนี้เพื่อให้แบบจำลองมีความเชี่ยวชาญเฉพาะทางและจัดการกับความซับซ้อนของใบหน้ามนุษย์ได้ดีกว่าแบบจำลองที่ถูกฝึกสอนมาอย่างกว้าง ๆ สำหรับสร้างภาพใด ๆ

ผลลัพธ์ที่ได้จากงานวิจัยชิ้นนี้นอกจากจะมีประโยชน์สำหรับผู้พิทักษ์สันติราษฎร์ใช้สร้างภาพร่างใบหน้าคนร้ายจากคำบรรยายลักษณะแล้ว ยังสามารถต่อยอดไปยังการใช้งานในบริบทอื่นที่อนุญาตให้ผู้ใช้ทั่วไปสามารถดีไซน์ออกแบบใบหน้าของบุคคลใหม่ ๆ และควบคุมลักษณะของใบหน้าที่ต้องการผ่านการบรรยายข้อความ ซึ่งถือเป็นการควบคุมที่ทำได้ง่ายและผู้ใช้ไม่จำเป็นต้องมีความเชี่ยวชาญในการใช้เครื่องมือออกแบบกราฟิกส์แต่อย่างใด

References

1. Jalan, H.J., Maurya, G., Corda, C., Dsouza, S. and Panchal, D., 2020, "Suspect Face Generation," *Proceedings of the 2020 3rd International Conference on Communication System*, *Computing and IT Applications (CSCITA)*, 3-4 April 2020, Mumbai, India, pp. 73-78.
2. Kotian, C., Lokhande, S., Jain, M. and Pavate, A., 2020, "D2F: Description to Face Synthesis Using GAN," *Proceedings of the International Conference on Recent Advances in Computational Techniques (IC-RACT)*, 27-28 March 2020, Mumbai, India, 8 p.
3. Nair, K.R., Sam, S.S., Praveena, K.P., Juju, K. and Cherian, S., 2021, "Transfer Learning with Deep Convolutional Neural Networks in Forensic Face Sketch Recognition," *Proceedings of the International Conference on IoT Based Control Networks & Intelligent Systems (ICICNIS 2021)*, 28-29 June 2021, Kerala, India, pp. 1-5.
4. Xu, J., Xue, X., Wu, Y. and Mao, X., 2020, "Matching a Composite Sketch to a Photographed Face using Fused HOG and Deep Feature Models," *The Visual Computer*, 37 (4), pp. 765-776.
5. Wadhawan, R., Drall, T., Singh, S. and Chakraverty, S., 2020, "Multi-Attributed and Structured Text-to-Face Synthesis," *IEEE International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET)*, 10-10 December 2020, Bengaluru, India, pp. 1-7.
6. Sun, J., Li, Q., Wang, W., Zhao, J. and Sun, Z., 2021, "Multi-caption Text-to-Face Synthesis: Dataset and Algorithm," *Proceedings of the 29th ACM International Conference on Multimedia, Association for Computing Machinery*, 20-24 October 2021, New York, USA, pp. 2290–2298.
7. Oza, M., Chanda, S. and Doerman, D., 2022, "Semantic Text-to-Face GAN-ST2F," *arXiv,* preprint arXiv:2107.10756. https://doi.org/10.48550/arXiv.2107.10756

8. Wang, T., Zhang, T. and Lovell, B., 2021, "Faces la Carte: Text-to-Face Generation via Attribute Disentanglement," *Winter Conference of Applications on Computer Vision (WACV)*, 5-9 January 2021, Virtually, pp. 3380-3388.
9. Sun, J., Deng, Q., Li, Q., Sun, M., Ren, M. and Sun, Z., 2022, "AnyFace: Free-style Text-to-Face Synthesis and Manipulation," *arXiv,* preprint arXiv:2203.15334v1. https://doi.org/10.48550/arXiv.2203.15334
10. Li, Z., Min, M.R., Li, K. and Xu, C., 2022, "StyleT2I: Toward Compositional and High-Fidelity Text-to-Image Synthesis," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 19-24 June 2022, New Orleans, Louisiana, USA, pp. 18197-18207.
11. Hermosilla, G., Tapia, D.H., Allende-Cid, H., Castro, G.F. and Vera, E., 2021, "Thermal Face Generation Using StyleGAN," *IEEE Access*, 9, pp. 80511-80523.
12. Tao, M., Tang, H., Wu, F., Jing, X., Bao, B. and Xu, C., 2022, "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 19-24 June 2022, New Orleans, Louisiana, USA, pp. 16515-16525.
13. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D. and Lischinski, D., 2021, "StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery," *Proceedings of the International Conference on Computer Vision (ICCV)*, Virtually, 11-17 October 2021, Virtually, pp. 2085-2094.
14. Deorukhkar, K., Kadamala, K. and Menezes, E., 2022, "FGTD: Face Generation from Textual Description," *Inventive Communication and Computational Technologies, Lecture Notes in Networks and Systems*, 311, pp. 547-562.
15. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J. and Aila, T., 2021, "Alias-Free Generative Adversarial Networks," *Conference on Neural Information Processing Systems (NeurIPS 2021)*, 6-14 December 2021, Virtually, 12 p.
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I., 2021, "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of the 38th International Conference on Machine Learning (PMLR)*, 18-24 July 2021, Virtually, 16 p.
17. Devlin, J., Chang, M., Lee, K. and Toutanova, K., 2019, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2-7 June 2019, Minneapolis, Minnesota, USA, pp. 4171-4186.

18. Cui, J., Jin, L., Kuang, H., Xu, Q. and Schwertfeger, S., 2021, "Underwater Depth Estimation for Spherical Images," *Journal of Robotics*, 2021, 12 p.
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S., 2017, "GANs Trained by a Two Time-scale Update Rule Converge to a Local Nash Equilibrium," *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 4-9 December 2017, Long Beach, CA, USA, pp. 6629–6640.
20. Radford, A., Metz, L. and Chintala, S., 2016, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *International Conference on Learning Representations (ICLR)*, 2-4 May 2016, San Juan, Puerto Rico.