Creating a Video from Facial Image Using Conditional Generative Adversarial Network

Bui Thanh Hung Ho Vo Hoang Duy Vo Quoc Huy

Data Science Laboratory
Faculty of Information Technology
Industrial University of Ho Chi Minh city, Vietnam
buithanhhung@iuh.edu.vn h.hoangduy2002@gmail.com huyvo8500@gmail.com

Abstract. Create a video from facial images that hold significant meaning in generating natural-looking videos from a single image. This technique is widely used in various fields such as filmmaking or social media. Previous methods have had limitations, such as creating high-quality videos lacking naturalness in reproducing character movements. Some studies have focused solely on producing high-quality videos, resulting in a loss of diversity in the content and style of the image. In this study, we propose a method to create a short video with natural facial movements of the lips, eyes, and related facial parts using deep learning techniques, convolutional neural networks (CNN), Hiden Affine transformation combined with Conditional Generative Adversarial Network (cGAN), image processing technique and computer vision methods. We will evaluate this method on CK-Mixed datasets and compare it with other research methods. Based on the results, we will proceed to develop an application that can create a facial motion video from a single input image and test its practicality.

Keywords: Creating a Video, Facial Images, Deep learning, CNN, Hiden Affine transformation, cGAN.

1 Introduction

In recent years, computer vision has seen significant advances, allowing computers to perceive and understand visual information with increasing precision. One fascinating application of computer vision technology is its ability to convert still images into dynamic video.

Converting still images to dynamic video offers wide application potentials in the multimedia content creation, virtual reality, video editing and assistive technologies, enriching the viewer experience and opening up new creative possibilities. At the same time, it also

contributes to opening new doors in exploiting the potential of computer vision technology to create creative and engaging content.

In this study, we focused on creating a video from a single facial photo, including the movements of the lips, eyes, and relevant parts of the face. We propose an innovative and effective method for compositing dynamic video from a facial image, integrating lip, eye and related facial movements to create a natural video. Our contribution of this study includes the following:

- We develop a method based on deep learning technique that combines Hidden Affine Transformation and image processing techniques to be able to synthesize a dynamic video sequence from a single facial image, which integrates smooth movements of lips, eyes, and related parts of the face to create a realistically and vividly video.
- We use cGAN deep learning model to learn Affine transforms, these Affine parameters are calculated based on information from input face image and used to generate frames then synthesize into a video with lip, eye and related facial movements.
- We conduct experiments and complete evaluation on diverse datasets, in order to prove the effectiveness and stability of our proposed model.
- We evaluate the results according to image quality, practicality and coherence and compare our method with existing methods.

In addition to the introduction, the rest of the paper includes the following. Section 2 presents related works. Section 3 describes our proposed model in detail. Section 4 presents the experimental and comparative evaluation of our results with other methods. Section 5 presents the conclusions of our study and future direction research.

2 Related Works

Video making from images is the process of transforming a series of images into a sequence of consecutive frames to create a moving video of the objects in the image. Many approaches were proposed and most of the methods use deep learning because of its effectiveness [1-4].

Using cINN model and other methods such as Affine transform, nonlinear deep learning and autoregressive model to synthesize video from input images were presented in [5]. However, the above method has some limitations such as the limited capacity of the estimation model, and the difficulty in dealing with large and complex distributions.

Yang Zhou [6] introduced a method using deep learning to create dynamic video of speakers with smooth moving lips and natural portrait rendering. The model can create animations that show mouth synchronization, individual facial expressions, and head movements better than current advanced technology, but the model still has some limitations that cannot generate high resolution images of speakers.

Guangyao Shen [7] proposed a method to convert a face image to

video using a hidden affine transform. This method is implemented using a deep learning model called AffineGAN, with the architecture based on the GAN (Generative Adversarial Networks) model. This method can create new videos with various facial expressions, and produce smoother, more realistic videos. The limitation of this research is requirements of careful data preparation; limit of the affine transform; depending on the quality of the input data; generalizability and loss of information.

Haomiao Ni [8] proposed Latent Flow Diffusion Models (LFDMs) as a class of latent flow based generation models, used to simulate the distribution of complex data such as images and videos. The proposed model can generate video by transforming a given image with flow sequences generated in hidden space based on layer condition. The limitation of this model is that LFDM is limited to processing video containing a single moving object; LFDM is now conditionally based on class labels instead of natural written descriptions and at sampling with 1000-step DDPM, GAN is faster than LFDM.

Ming-Yu Liu [9] proposed a model for creating video from single image. This process may include the generation of intermediate frames between the available frames, to generate a continuous forecast of motion. The advantage of this method is that it provides the ability to create dynamic video from single image, extending the ability to create video from sparse image data; create realistic and continuous video frames in motion. However, this method has some limitations such as requiring large amount of training data and computational complexity; depending on fining tune parameters and training process to achieve good video quality and difficulty reproducing natural details and movements in the video.

Long Zhao [10] introduced a method to generate video from input image by predicting and improving redundant motion. The above model can predict excess motion, potentially producing high-quality video. The limitation of this method is that it depends on the original frame; high computational complexity, limited to handling complex cases such as fast moving subjects, complex motion, or rapid changes in light in a video.

Based on previous studies, in this study, we propose an effective model to synthesize dynamic video from a still image by integrating lip, eye and related parts on the face into composite video to create natural and realistic based on deep learning technique that combines Hidden Affine Transformation and image processing techniques. We use a convolutional neural network (CNN) for extracting facial features and a conditional Generative Adversarial Network (cGAN) for video synthesis from image features. The proposed model will be presented in the next section.

3 Methodology

3.1. The proposed model

From the initial raw data set, we use data preprocessing methods and then put it through a deep learning model to synthesize video from face images. Next will save the parameters of the model and optimize those parameters. To model for more realistic and vivid motion video. Fig 1 describes our proposed model.

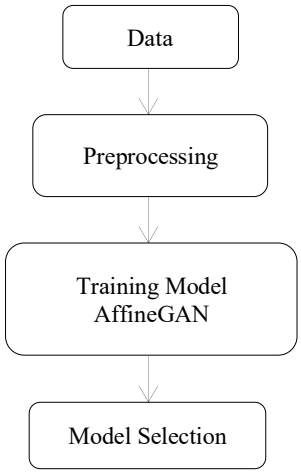

Fig 1. The proposed model

3.2. Conditional Generative Adversarial Network (cGAN)

GAN was proposed by Mirza and Osindero in 2014 [11]. cGAN is a type of GAN in the field of deep learning that is capable of generating new data based on a specific input condition. cGan architecture is presented in Fig. 2.

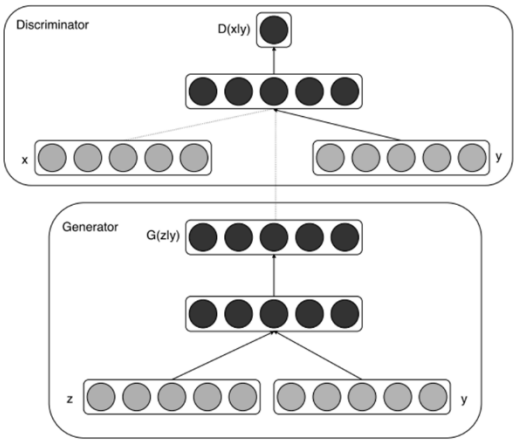

Fig. 2: cGAN architecture [11]

The generator network and the discriminator network are two main parts of the cGAN.

Generator:
The generator network takes as input a label and a face image. Labels

contain information about facial expressions or other attributes. The generative network uses a convolutional and deconvolutional layer-based architecture to map from face images to video frames. The convolutional layer helps to extract features from the face image and hidden affine transform, while the deconvolutional layer helps to create detailed video frames.

We use CNN in the generator network to analyze and extract important features from facial images, including features of facial shape and structure, details such as eyes, nose, mouth and expressive features. These features are then used to perform the Hidden Affine transform and generate a video from the input face image.

Discriminator:
Different from the traditional discriminant network, the discriminant network in cGAN takes an additional label as input, which helps to differentiate based on both the image content and the label. The discriminant network receives video frames and labels as input, and then goes through convolutional layers for feature extraction. We use cGAN to convert face images into video through the creation of subsequent frames in the video sequence. The cGAN model used is presented as follows:

Generative Network:
From an input face image, this component generates a video corresponding to that image. The cGAN model uses an Encoder-Decoder structured generator network, in which:
Encoder is responsible for extracting features from the input face image. Then, the generator network uses a Hidden Affine Transformation to map the features from the image space to the video space. Hidden Affine Transformation is a secret linear transform that maps features from face space to video space. This helps to create natural movement in the resulting video.
Decoder: Next, the features are mapped over the decoder network to produce the output video.

Discriminative Network:
The discriminant network is trained to distinguish between the videos generated by the generator network and the actual videos. The discriminant network uses a conventional classifier network architecture, trained to properly distinguish between the generated face video and the actual video. The task of the generator network is to produce videos that the discriminant network cannot distinguish from the actual video.

3.3. Hidden Affine Transformation

Hidden affine transformation is a method to apply affine transformations on deep learning data without explicitly defining affine parameters. Instead, the deep learning model automatically learns affine transformations based on the training data.

The hidden affine transform allows the model to learn how to generate

video frames from the original face image by scaling, rotating, and shifting the facial features. How to perform hidden affine transformations in deep learning models usually involves using neural networks to learn and predict the affine parameters corresponding to each video frame. These affine parameters are calculated based on input information, such as a face image, and are used to generate new video frames.

In this study, we use Hidden Affine transform to improve face image to video transformation by applying hidden linear transforms to reproduce natural facial expressions and movements.

4. Experiments

4.1. Dataset

In this research, we use the CK-Mixed dataset. CK-Mixed is a Cohn-Kanade (CK+) dataset [12]. This dataset includes 593 videos of 6 types of emotions, most of which are in gray scale. After processing video data, we extracted frames from the video to obtain an image data set of 8000 images. We proceed to divide the data set in the ratio of 8:2. Table 1 presents an overview of the CK-Mixed dataset and some images of the dataset are presented in Fig. 3.

Table 1. The dataset

Dataset	Number of images
Train	5400
Test	1600

Fig 3. Some pictures in the CK-Mixed dataset

4.2.Dlib

To locate the lip position on the face in the image we used Dlib [13] which is an open source library developed by Davis E. King.

After determining the lip position, we apply the OpenCV library [14] to cut that lip part from the original image and convert this lip part to a binary image to create a patch used to fix the problem in motion video. Fig. 4 details our preprocessed lip patch.

Fig. 4: Our preprocessed lip patch.

4.3.Evaluation

Metric

To evaluate the model in detail and quantitatively, we use the following metrics, PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index Measure) [15], ACD (Average Content Distance) [16], ACD-I (Average Content Distance - Identity) [17]. The higher the SSIM and PSNR scores, the better the video quality is produced, the lower ACD and ACD-I scores show the better results

MSE (Mean Square Error) [18]:

This loss function used in LSGAN (Least Squares GAN) calculates the average error between the Discriminator output and the target label. MSE is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

BCE (Binary Cross-Entropy) [19]:

This loss function is common used in binary classification problems. The BCE function is used to measure the difference between binary predictions and actual labels in a binary classification problem. Typically, the output of a binary classification model is mapped to the range [0, 1], using the Sigmoid activation function, for example. BCE calculates the cross-entropy between the binary prediction and the actual label as a numeric value. The formula of BCE is calculated as follows:

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot log(\hat{y}) + (1 - y_i) \cdot log(1 - \hat{y}_i)$$

4.4.Result

We used OpenFace [20], a deep model trained for face recognition and capable of surpassing human performance, to extract facial features.

We create a video from a single input facial image based on the above proposed method. This method is capable of creating smooth and realistic motion video from the input face image. The results were evaluated based on criteria such as similarity to the original image, authenticity and diversity of the video. Figure 5 shows an example of the

prediction results after model execution.

Fig. 5: The result of our proposed model

To evaluate our model with other models, we chose 5 advanced models, achieving good results built based on the GAN model for comparison, which are: VGAN[21], MoCoGAN[22] and ImaGINator[23].

Regarding the models used for comparison, we implemented on the publicly available source code of VGAN and ImaGINator, for MoCoGAN we used the results given in [22]

The results of the comparison between the 5 methods and our proposed model are presented in Table 2. The results of the loss function evaluating the difference between the generated image (fake_image) and the image (real_image) are presented in Fig. 6 and Fig. 7 shows Discriminator loss results.

Table 2: Comparison table between methods

Method	PSNR	SSIM	ACD	ACD-I
VGAN	16.32	0.41	0.14	1.55
MoCoGAN	18.16	0.58	0.15	0.90
ImaGINator	20.29	0.85	0.08	0.29
AffineGAN	**35.50**	**0.91**	**0.06**	**0.16**

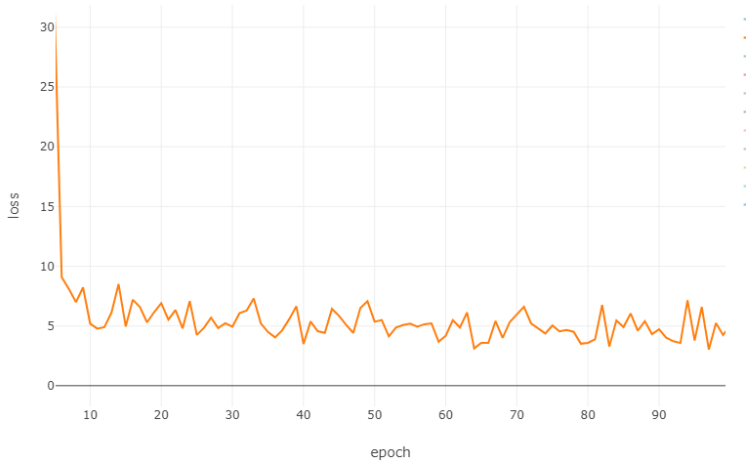

Fig. 6: The loss function evaluates the difference between fake_image and real_image

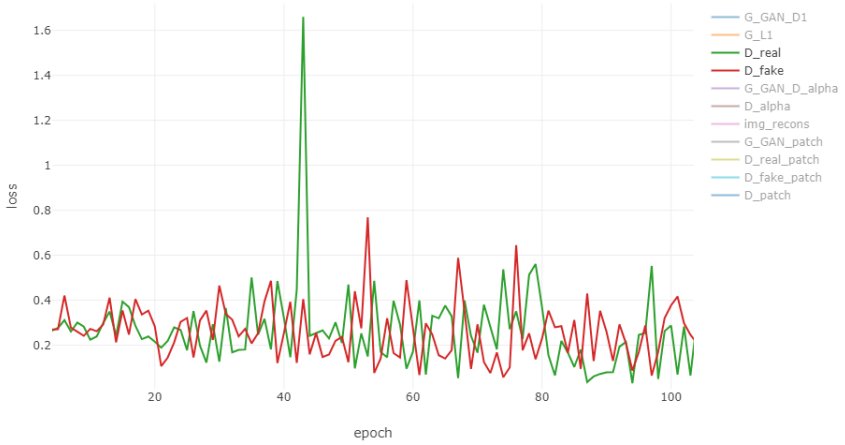

Fig. 7: Discriminator Loss

Looking at the results in Table 2, we can see that our proposed method has the best results compared to other comparison methods in all 4 measurements ACD, ACD-I, SSIM and PSNR,. This result shows that the cGAN deep learning method combined with Hidden Affine Transformation gives better results than other GAN methods.

We evaluate the difference between the generated image (`fake_image`) and the real image (`real_image`) based on L1Loss [24]. The goal is to minimize the distortion between the generated image and the real image. In Fig. 6, we see that the loss decreases over time as the Generator improves its ability to generate images. The ultimate goal is to achieve the lowest G_L1 value, as the Generator generates images with a high degree of similarity to the real image, which means producing high quality images that are difficult to distinguish from the real image.

In Fig. 7, D_real and D_fake are the discriminator outputs when real and fake data are input, respectively. D_real is the output of the discriminator when inputting real data. D_fake is the output of the discriminator when dummy data is introduced. The goal of the model is to generate fake data that is difficult for the discriminator to distinguish from real data. By optimizing the loss function, the model tries to make D_real close to 1 and D_fake close to 0, thereby generating high quality artificial data.

However, when analyzing in detail, we find that our proposed model also has limitations such as: there are still some special cases when video synthesis is difficult, for example when the original image low quality or indistinct, or when facial movements are complex and subtle. These restrictions can cause problems such as image noise, unnatural video, or inaccurate video.

5. Conclusion

The experimental results of the study have demonstrated the effectiveness and potential of the proposed method in creating dynamic video from a single image. By using CNN model for feature extraction and cGAN model for video synthesis, the method has achieved impressive results. Composite videos contain natural and realistic movements of lips, eyes, and other parts of the face. Experiments on

diverse datasets have demonstrated the stability and good reproducibility of the method in many different cases.

In order to improve and develop the method in the future, we will extend the model architecture to handle more difficult and special cases, including low image quality, motion diversity, and large face variability; focus on improving the realism of the composite video, ensuring that facial movements and details are accurately reflected and explore the applicability of the method in different fields, such as cinema, media and entertainment.

References

1. Bui Thanh Hung, Vijay Bhaskar Semwal, Neha Gaud, Vishwanth Bijalwan, Violent Video Detection by Pre-trained Model and CNN-LSTM Approach. Proceedings of Integrated Intelligence Enable Networks and Computing. Springer Series in Algorithms for Intelligent Systems, 2021.
2. Ankur Gupta, Sabyasachi Pramanik, Hung Thanh Bui and Nicholas M. Ibenu, Machine Learning and Deep Learning in Steganography and Steganalysis, Multidisciplinary Approach to Modern Digital Steganography Book, IGI Global, 2021.
3. Bui Thanh Hung, Content based Image Retrieval using Multi-Deep Learning Models, Next Generation of Internet of Things. Lecture Notes in Networks and Systems, pp 347-357, vol 445. Springer, Singapore, 2022.
4. Bui Thanh Hung, Using Deep Unsupervised Method for Stock Prediction, Lecture Notes in Networks and Systems book, LNNS, volume 288, 2022.
5. M. Dorkenwald, T. Milbich, A. Blattmann, R. Rombach, K. G. Derpanis and B. Ommer, Stochastic Image-to-Video Synthesis using cINNs, 2021.
6. Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis And D. Li, MakeItTalk: Speaker-Aware Talking-Head Animation, 2021.
7. G. Shen, W. Huang, C. Gan, M. Tan, J. Huang, W. Zhu, and B. Gong, Facial Image-to-Video Translation by a Hidden Aine Transformation, 2019.
8. H. Ni, C. Shi, K. Li, S. X. Huang and M. R. Min, Conditional Image-to-Video Generation with Latent Flow Diffusion Models, 2023.
9. M.-Y. Liu, X. Huang, J. Yu, T.-C. Wang and A. Mallya, Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications, 2020.
10. L. Zhao, X. Peng, Y. Tian, M. Kapadia and D. Metaxas, Learning to Forecast and Refine Residual Motion for Image-to-Video Generation, 2018.
11. M. Mirza and S. Osindero, Conditional Generative Adversarial Nets, 2014.
12. Lucey, Patrick, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, 2010 IEEE

computer society conference on computer vision and pattern recognition-workshops. IEEE, 2010.
13. DLIB: http://dlib.net/
14. OpenCV: https://opencv.org/
15. A. Horé and D. Ziou, Image Quality Metrics: PSNR vs. SSIM, 20th International Conference on Pattern Recognition, Istanbul, Turkey, 2010, pp. 2366-2369.
16. S. Tulyakov, M.-Y. Liu, X. Yang, J. Kautz, Mocogan: Decomposing motion ancontent for video generation, 2018.
17. L. Zhao, X. Peng, Y. Tian, M. Kapadia, D. Metaxas, Learning to forecast and refine residual motion for image-to-video generation, 2018.
18. Schluchter, Mark D, Mean square error, Encyclopedia of Biostatistics 5, 2005.
19. Li, Li, Miloš Doroslovački, and Murray H. Loew, Approximating the gradient of cross-entropy loss function, IEEE Access 8, 111626-111635, 2020.
20. B. Amos, B. Ludwiczuk, M. Satyanarayanan, OpenFace: A general-purpose face recognition library with mobile applications, 2016.
21. Vondrick, Carl, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics, Advances in neural information processing systems, 2016.
22. Tulyakov, Sergey, Ming-Yu Liu, Xiaodong Yang, Jan Kautz, Mocogan: Decomposing motion and content for video generation, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.
23. Wang Yaohui, Piotr Bilinski, Francois Bremond, Antitza Dantcheva, Imaginator: Conditional spatio-temporal gan for video generation, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020.
24. Chai, Tianfeng, and Roland R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE), Geoscientific model development discussions 7.1, pp 1525-1534, 2014.