Hindawi
Journal of Electrical and Computer Engineering
Volume 2023, Article ID 6645356, 13 pages
https://doi.org/10.1155/2023/6645356

Research Article

Attention-Based Image-to-Video Translation for Synthesizing Facial Expression Using GAN

Kidist Alemayehu,[1] **Worku Jifara**,[1] **and Demissie Jobir**[2]

[1]*School of Electrical Engineering and Computing, Department of Computer Science and Engineering, Adama Science and Technology University, Adama, Ethiopia*
[2]*School of Electrical Engineering and Computing, Department of Electronics and Communications Engineering, Adama Science and Technology University, Adama 1888, Ethiopia*

Correspondence should be addressed to Worku Jifara; worku.jifara@astu.edu.et

Received 13 June 2023; Revised 23 September 2023; Accepted 14 October 2023; Published 14 November 2023

Academic Editor: Nihal F. F. Areed

The fundamental challenge in video generation is not only generating high-quality image sequences but also generating consistent frames with no abrupt shifts. With the development of generative adversarial networks (GANs), great progress has been made in image generation tasks which can be used for facial expression synthesis. Most previous works focused on synthesizing frontal and near frontal faces and manual annotation. However, considering only the frontal and near frontal area is not sufficient for many real-world applications, and manual annotation fails when the video is incomplete. AffineGAN, a recent study, uses affine transformation in latent space to automatically infer the expression intensity value; however, this work requires extraction of the feature of the target ground truth image, and the generated sequence of images is also not sufficient. To address these issues, this study is proposed to infer the expression of intensity value automatically without the need to extract the feature of the ground truth images. The local dataset is prepared with frontal and with two different face positions (the left and right sides). Average content distance metrics of the proposed solution along with different experiments have been measured, and the proposed solution has shown improvements. The proposed method has improved the ACD-I of affine GAN from 1.606 ± 0.018 to 1.584 ± 0.00, ACD-C of affine GAN from 1.452 ± 0.008 to 1.430 ± 0.009, and ACD-G of affine GAN from 1.769 ± 0.007 to 1.744 ± 0.01, which is far better than AffineGAN. This work concludes that integrating self-attention into the generator network improves a quality of the generated images sequences. In addition, evenly distributing values based on frame size to assign expression intensity value improves the consistency of image sequences being generated. It also enables the generator to generate different frame size videos while remaining within the range [0, 1].

1. Introduction

Computer vision seeks to enable machines to perceive the world in the same manner that humans do and apply that knowledge to a variety of tasks and processes (such as image analysis, image recognition, and classification). Following the advancement of deep generative networks in computer vision, photo-realistic image generation [1–4], which focuses on generating realistic-looking images from random vector, and image-to-image translation [5, 6] which is concerned with translating images from one domain to another, are the two well-known examples that have been achieved. This study focuses on image-to-video translation specifically on generating image sequences of facial expression from a single neutral image.

Humans can imagine various scenes of what would be the next move based on a given still image, therefore allowing machines to learn such activity would be advantageous in a range of application sectors such as cinematography [7–10], movie production, photograph technology, and e-commerce. Image-to-video translation seeks to generate video sequences from the single static image; unlike image-to-image translation, it adds another temporal dimension to deep models; and, unlike video prediction, its

input is a static image with no temporal clues [6, 11–13]. Facial expression is defined as the movement of facial muscles on the skin that can convey an individual's emotional state. Facial expression synthesis focuses on creating new face shapes from the given input face without changing the particular characteristics of the given face; if implemented as an add-on, it can improve facial recognition accuracy and gender classification. Facial expression synthesis has several applications, including robot and avatars animation, video game development, and graphics interchange format (GIF) generation. Earlier methods synthesize facial expression using traditional approaches including 2D/3D image warping [14], image reordering [15], or flow mapping methods [16], the majority of which are example-based or morph-based. Recent methods rely on generative models to deal with facial expression synthesis. One of the most notable is GAN (generative adversarial network) [1], which has made significant advances in image generation. [3, 10, 17–21] have addressed the issue of facial expression but limited to static facial expression generation. Producing a facial expression video from one image is a one-to-many mapping problem in which the output has many more unknowns to solve than the input, which lacks temporal information. Recently, researchers addressed the challenge of image-to-video translation for generating facial emotions from a single picture, including facial geometry-based [22], face parsing-based [23], action unit-based [24], and 3D blend shape model-based [25]. The authors in [22, 23, 26] produce facial expressions images without background face (ear, hair, and half of the forehead). Other works [11] proposed a user-controllable framework to synthesize facial expressions, and this approach assigns expression intensity value manually to make the generated image sequences consistent; however, the manual annotation fails when the video is incomplete. To solve the manual annotation problem AffineGAN, the authors in [15] employed an affine transformation to give an expression intensity to each facial image in the latent space, assuming that the ground truth (real) image features are equal to the produced image features; if they are not equal, inconsistency occurs. This effort also necessitates extracting features from the target ground truth images to compute the expression intensity value. In addition to the abovestated problems, the majority of existing works only focus on synthesizing frontal and near-frontal face expressions. But the nonfrontal face images also need to be synthesized to apply the model for most real-world applications. Some works [20, 27] synthesize face expression and pose simultaneously, but they disregard expression intensity variations.

This work aims to generate good-quality and consistent image sequences of facial expression and varying frame size videos inspired by the AffineGAN [15, 28] network. In this regard, this study proposes to infer expression intensity value by evenly distributing values in the range [0, 1] according to the number of frames in a given video to produce consistent image sequences and varying frame size videos, as well as adding attention mechanisms, i.e., self-attention mechanism and constraints to the network architecture for better results.

2. Methodology

2.1. Model Architecture. The proposed model has been adapted from facial image-to-video translation by a hidden affine transformation called AffineGAN, for learning image-to-video translation which was introduced by the authors in [15]. In this work instead of three discriminators such as AffineGAN, the proposed model uses one discriminator network. In addition, attention mechanisms were also introduced to the generator network and express intensity value inferred by evenly distributing values according to the frame size in a given video. Hence, unlike that of the AffineGAN, the proposed model is built with one generator and one discriminator. The generator network was built as U-Net [7] architecture. The discriminator network is a patchGAN discriminator following the work [5] that penalizes structure only at the scale of nearby image patches.

The generator, as illustrated in Figure 1, is made up of two encoders, basic encoder Enc_b, and residual encoder Enc_r, as well as a decoder Dec. Both encoders are similar in structure and take the neutral image Io as input. The basic encoder is used to retain the feature of the neutral picture, while the residual encoder is utilized to capture the expression shift from neutral to peak expression. fb and fr are the features from the basic and residual encoder, respectively, as shown in equations (1) and (2), respectively.

The expression intensity e is used to control expression change and is calculated by equally distributing values within the range [0, 1] based on the number of frames in the training dataset. In essence, you provide the beginning and endpoints of an interval, as well as the number of total breakpoints you want inside that interval. The interval is the number of frames in a specific video, and the starting and finishing points are set to 0 and 1, respectively. For example, if the frame size (interval) is five, the intensity value will be [0, 0.25, 0.5, 0.75, 1], and the expression intensity helps the network learn that the expression should go from neutral to the peak. The target feature (ft) is obtained from basic features (fb), residual features (fr), and expression intensity (e) as shown in equation (3).

$$\text{Enc\_}b(\text{Io}) = \text{fb}, \tag{1}$$

$$\text{Enc\_}r(\text{Io}) = \text{fr}, \tag{2}$$

$$\text{ft} = \text{fb} + \text{fr} * e. \tag{3}$$

Finally, the target image $\widetilde{\text{It}}$ is generated by feeding target feature ft to the decoder Dec as shown in Figure 2.

$$\widetilde{\text{It}} = \text{Dec}(\text{ft}), \tag{4}$$

When $e = 0$, ft = fb, and it means that the generator has to generate a neutral face image, and when $e = 1$, the generated face image should be at the peak expression state. The discriminator network is utilized to contrast the produced expression image from that of the real expression image.

2.2. Generator Architecture. The generator network is utilized to create realistic-looking facial expressions with varying intensities. Both the encoders and the decoder are constructed as seven-layer neural networks with skip

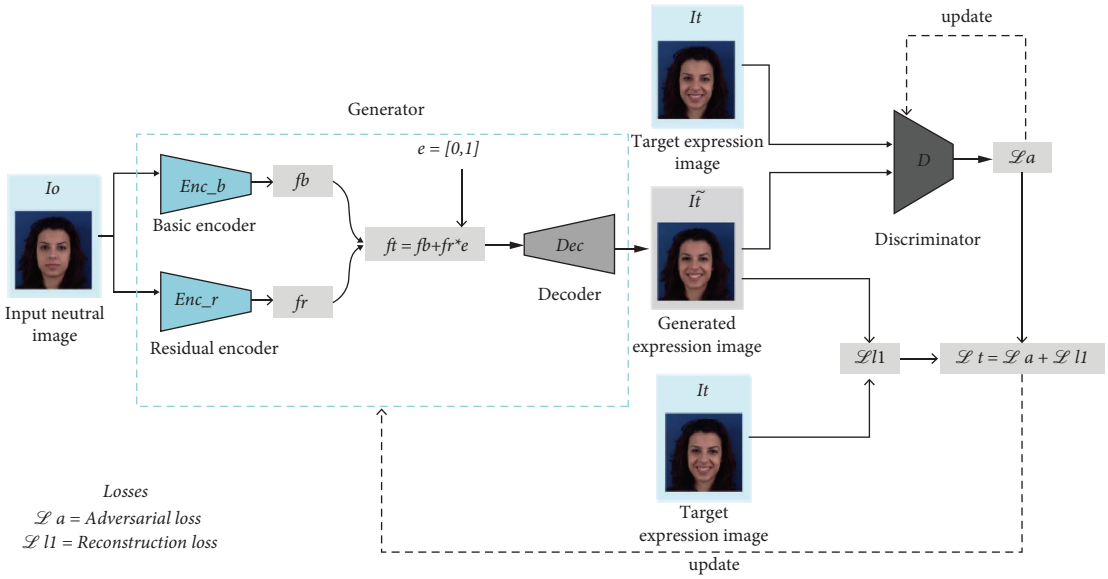

Figure 1: Proposed architecture.

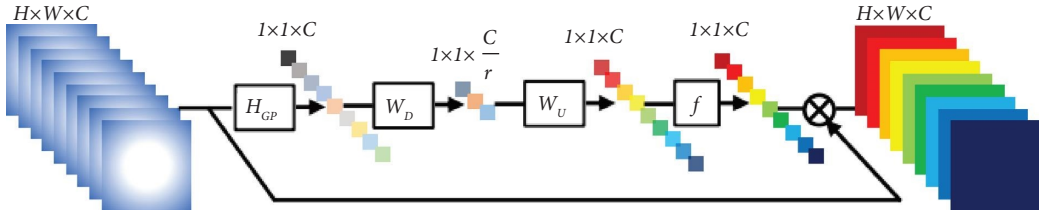

Figure 2: Channel attention architecture, adopted from [29].

connections based on U-Net architecture [7] as shown in Figure 3. U-Net, whose architecture is shaped like the letter "U," was originally designed for medical image segmentation. It produced such good results that it was later used in other applications. In U-Net architecture, there are two pathways. The first path is the contraction path called encoder, which is utilized to record the context of the image and the path is an expanding path also called decoder is utilized to achieve exact localization via transposed convolutions. To extend a feature to a target picture as revealed in Figure 3, U-Net employs the same feature maps that were utilized for contraction. This would preserve the image's structural integrity while drastically reducing distortion. Both the encoders contain convolution, leaky ReLu activation function, and instance normalization except for the first and the last layers, and it contains only the convolution at the first layer and convolution and leaky ReLu activation function at the last layer. The decoder contains transposed convolution, ReLu activation function, and instance normalization except for the last layer, and it only contains transposed convolution and ReLu activation function. The self-attention layer is incorporated at the first layer of the decoder to aid the generator to create images with fine details in every location that are precisely coordinated with fine features in distant parts of the image. Except on the first and last layer of the decoder, channel attention was planned to be added to help the network focus on crucial features while filtering out irrelevant features. However, for the sake of reducing experiment complexity, we have included self-attention and disregarded channel attention in the proposed method. We can consider the attention mechanism in Figure 3 as an empty set and will be considered in our future work.

2.3. Self-Attention Mechanism. SAGAN [29] presents a self-attention method into convolutional GANs. With the self-attention method, the generator can produce images with fine details in every location that are precisely coordinated with fine features in distant parts of the image.

The initial step toward self-attention is to break down each input feature into three separate vectors, $f(x)$, $g(x)$, and $h(x)$ as shown in Figure 4. To limit the number of channels, $1x1$ convolution is employed. Instead of inspecting every pixel, the self-attention component is concerned with the input activation's local regions. So, $h(x)$ is a representation of input features with a smaller number of channels and activation maps. The significance of the features of the self-attention map is determined using $f(x)$ and $g(x)$. The vector $g(x)$ at location x is calculated and compared with the vector $f(x)$ at all locations to compute an output feature at location x. The attention map is calculated with matrix multiplication between $f(x)$ and $g(x)$. Finally, the matrix multiplication between $h(x)$ and the

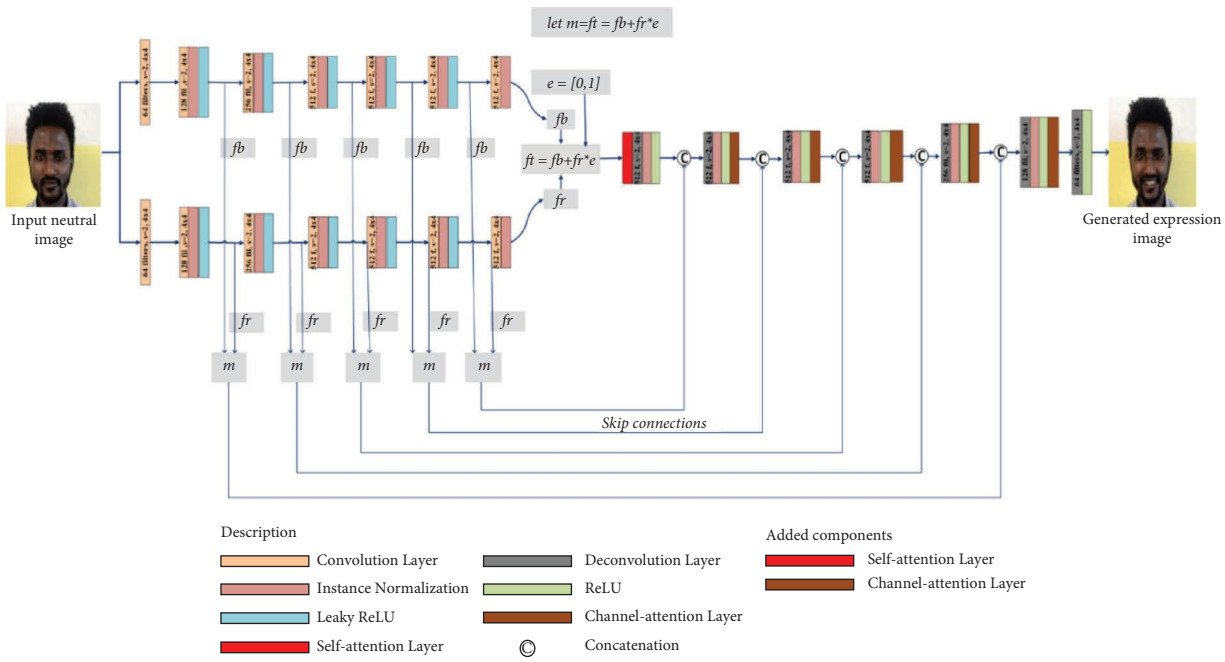

Figure 3: The proposed generator architecture.

attention map gives a self-attention feature map. The self-attention layer enables the generator to generate images with fine details in every location that are accurately coordinated with fine characteristics in distant regions of the image. To solve the challenge of limited convolution kernel size, the authors in [30] added a self-attention mechanism into the generator's upsampling block to generate long-range dependence of the picture. As a result of this discovery, in this work, we have incorporated self-attention mechanism in the generator's decoder block to construct the image's long-term dependency, as shown in Figure 3.

2.4. Channel-Attention Mechanism. Since each filter in the Conv layer works with a nearby receptive field, the output after convolution cannot leverage contextual information outside of the adjacent region. To make the generator network focus on crucial features, the authors in [2] proposed channel attention (CA) mechanism that makes use of feature channel interdependence. Global average pooling is utilized to convert channel-wise global location information into a channel descriptor.

Let $\mathrm{Img} = [\mathrm{img}1, \ldots, \mathrm{img}c, \ldots, \mathrm{img}C]$ be an input containing C feature maps of size HXW.

It has C feature maps with the size of HXW. Img can be shrunk through spatial dimensions HXW to acquire the channel-wise statistics $z \mathscr{E} \mathbb{R}^c$.

Then, the c^{th} element of z is determined by the following expression:

$$z_c = H_{\mathrm{GP}}(x_c) = \frac{1}{\mathrm{HXW}} \sum_{i=1}^{H} \sum_{J=1}^{W} x_c(i, j), \tag{5}$$

where $x_c(i, j)$ is the value at position (i, j) of c^{th} feature x_c. The global pooling function is denoted by $H_{\mathrm{GP}}(.)$. Such channel statistics can be considered as a collection of the neighborhood descriptors, and it contributed to expressing the entire image as shown in Figure 2.

A gating technique was also implemented to properly capture channel-wise dependencies from aggregated information via global average pooling. The gating mechanism should fulfill two conditions, as mentioned in [31]: first and foremost, it should be capable of learning nonlinear interactions between channels. Second, because several channel-wise properties might be stressed rather than a solitary one-hot activation, it must learn a nonmutually exclusive connection. A basic gating system with a sigmoid function is used here following the work of [31] as in the following equation:

$$s = f(\mathrm{WU}\delta(\mathrm{WD}z)), \tag{6}$$

where sigmoid gating and the ReLu activation function are denoted by $f(.)$ and $\delta(.)$, respectively. WD, is the weight of a Conv layer, that is used as a channel-downscaling with a reduction r ratio. The low-aspect signal is then expanded with ratio r by the channel-upscaling layer after ReLU activated it, whose weight is WU. The last channel statistics s are then obtained and employed to rescale the input x_c as follows:

$$\hat{x_c} = s_c.x_c, \tag{7}$$

where s_c and x_c are the scaling factors and feature map in the c^{th} channel.

The authors in [32] utilized channel-wise attention at the decoder part in an MRI reconstruction problem to focus on crucial features related to the ultimate objective while

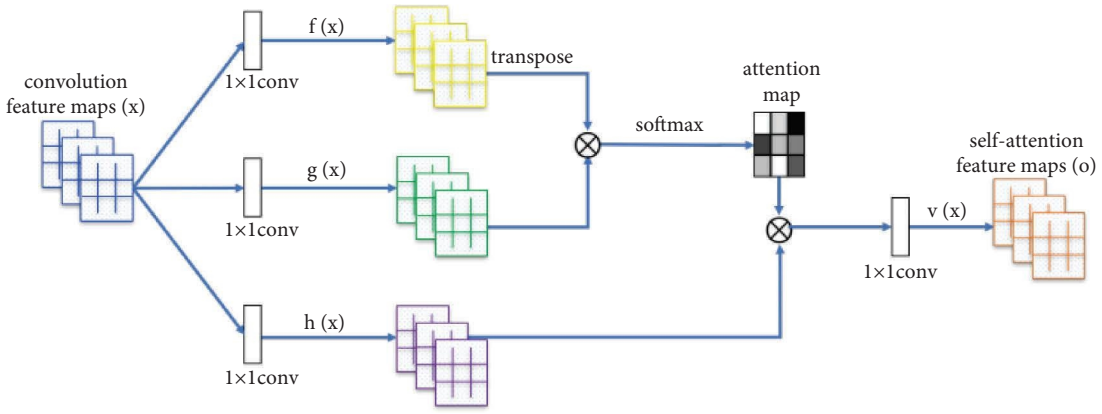

Figure 4: Self-attention mechanism introduced in the proposed GAN architecture.

filtering out irrelevant and noisy features and achieves very promising outcomes in terms of common MRI reconstruction metrics which are SSIM, PSNR, and NMSE. As a result of this discovery, in our future work, channel attention will be employed in the generator's second, third, fourth, fifth, and sixth decoder layers to focus on significant features while filtering out irrelevant features. In this work, channel attention in Figure 3 is empty and no any computation is available in this structure.

2.5. Discriminator Network. The discriminator network depicted in Figure 5 is a patchGAN discriminator following the work [5] that penalizes structure only at the scale of nearby image patches. It uses deep CNN to classify images as real or produced. Rather than learning to penalize structure across the entire image, as the traditional GAN does, the patchGAN penalizes structure only at the scale of nearby image patches. It decreases the discriminator load greatly since distinguishing local patches requires far less model capacity than discriminating complete images. The discriminator network, as seen in Figure 5, includes convolutional layers, instance normalization, and leaky ReLU except for the first and last layers, the discriminator's first layer comprises convolution and instance normalization, while the last layer simply contains convolution. It contrasts the generated video frames to the ground truth frames.

2.6. Model Learning Functions. The main aim of this work is to generate good-quality image sequences that are consistent in time. To achieve this objective, different learning functions such as loss functions and attention mechanisms were introduced to the network. The loss functions used in this work include adversarial loss $\mathscr{L}_a$ and reconstruction error $\mathscr{L}_{l1}$ between the produced and real images, as shown in the following equation:

$$\mathscr{L}_t = \mathscr{L}_a + \mathscr{L}_{l1}. \tag{8}$$

2.6.1. Adversarial Loss. In this study, mean absolute error is utilized to train the generator and the discriminator rather than vanilla GAN. It is determined by averaging the absolute difference between the actual and produced images, as it is formulated in the following equation:

$$\mathscr{L}_a = \frac{1}{2}\left|D\left(G\left(\mathrm{Io}, e\right)\right) - D\left(I_t\right)\right|, \tag{9}$$

where D is the discriminator network that contrasts the generated image by the generator G and the ground truth image I_t.

2.6.2. Reconstruction Loss. L_1 is a typical loss function that minimizes the absolute disparities between the estimated values and the existing target values. The L_1 loss function is more resilient and, in general, is unaffected by outliers. As shown in equation (10), L_1 loss was utilized between the produced and the real images following the work of [5].

$$\mathscr{L}_{L_1} = \left|I_t - \tilde{I}_t\right|, \tag{10}$$

where I_t the ground truth is (real) image and Ĩt is the generated image

3. Experiments

3.1. Datasets and Implementation Details. The proposed model was trained and tested using the MUG (multimedia understanding group) facial expression dataset and a local dataset that was collected by the researchers. The MUG dataset is appropriate for this study because it contains high-quality, well-organized video frames from neutral to peak expression. Also, the local dataset was created using the MUG dataset as a reference. The dataset is organized with different expressions. These are expressions that are happy, anger, surprise, and disgust. The number of frames varies from person to person. To increase the size of the dataset, the following image augmentation techniques were used: horizontal flip, random brightness, random saturation, adjusted brightness, and random contrast to train the model with more data. Random brightness and adjusted brightness techniques were used to change the image's brightness based on the brightness factor, brightness factor 32/255, and 0.3 utilized, respectively. Random saturation and random contrast are used to randomly change the saturation and

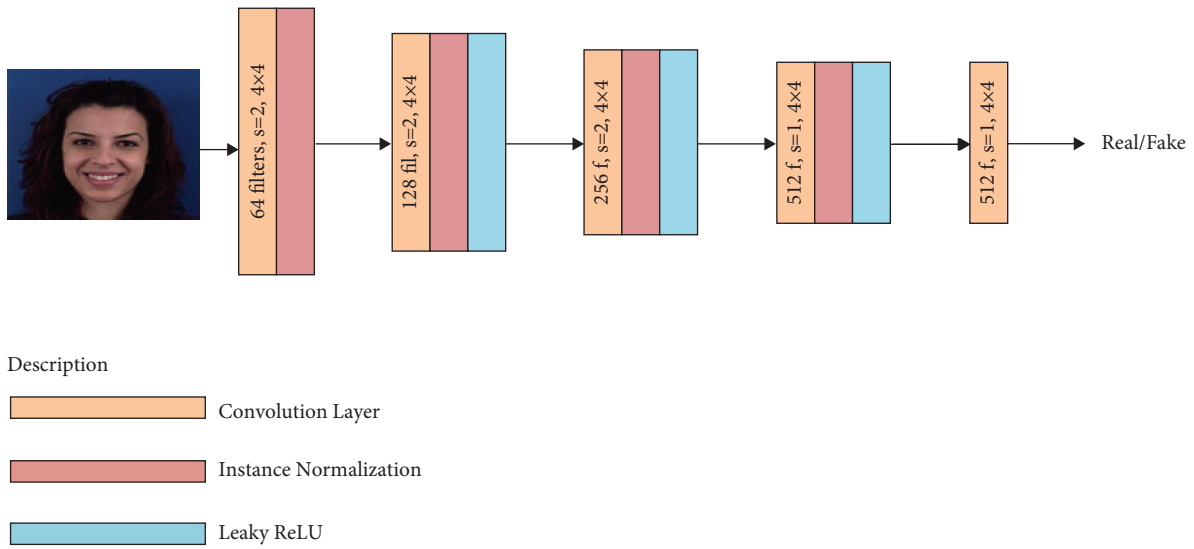

Figure 5: The proposed discriminator architecture.

contrast of the image based on the saturation and contrast range, respectively. Both the saturation and contrast ranges are set to the range of [0.5, 1.5]. In addition to the MUG dataset, we have collected few datasets for testing purpose.

The implementation of our work is mainly based on the AffineGAN using PyTorch deep learning framework. Adam optimizer with learning-rate = 0.002, beta_1 = 0.5, and beta_2 = 0.999 was utilized to update the network weight. Adam is described as integrating the benefits of two other extensions of stochastic gradient descent, specifically, AdaGrad and RMSProp [33], and it is straightforward to implement, computationally efficient, and requires little memory. The step size, often known as the "learning rate," is the amount by which the weights are changed during training. The learning rate and the weight loss for L_1 used for this work are similar to [15]. This work is implemented on Windows desktop with processor Intel®Intel™ i7-8700 CPU @ 20 GHz 3.19 GHz and graphics NVIDIA Corporation TU104 (GeForce RTX 2070 SUPER)/GeForce RTX 2070 SUPER/PCIe/SSE2 with a RAM capacity of 8 GB.

3.2. Experimental Classes. The propose method that is aimed to improve image-to-video translation for synthesizing facial expressions should be experimentally proved. To this end, we have conducted different experiments as shown in Table 1. The first experiment is the implementation of the baseline work, Affine GAN. AffineGAN inferred expression intensity value by applying affine transformation in the latent space, and it used alpha discriminator to control the value from deviating away from the range [0, 1]. BCE loss is used to train the generator and the discriminators.

The second experiment is the initial work of the proposed model. The expression intensity value is inferred as stated in the above section. Since the NumPy linespace function bound the values to the range [0, 1], the alpha discriminator is not necessary like AffineGAN. For the local discriminator, the boundary image is used as an informative region to calculate the loss between the multiplication of the informative region and the ground truth image with the multiplication of the informative region with a generated image. Instead of using the mouth region only as an informative region like [11, 15], the boundary image is used as an informative region since the boundary image can hold the structure of the face image and is also better at describing the facial pose than only the mouth area. Mean absolute error (MAE) loss is used to train the generator and the discriminators. The boundary is detected using an open-source face alignment library [34]. Lastly, in the third experiment, the self-attention mechanism was added at the first decoder layer of the generator for good-quality image generation. The result of each experiment is depicted in Section 4.

4. Experimental Results

In this study, three experiments were carried out. All the experiments were carried out on the same dataset, hardware, and software configuration for comparison purposes. The results are shown for the two datasets that are the MUG dataset and the local dataset.

4.1. AffineGAN. AffineGAN contains one generator and three discriminators. The first discriminator is used to contrast the generated image with the ground truth (real) image. The second discriminator is used to contrast the multiplication of the informative region (mouth region) and the produced image with the multiplication of the informative region and ground truth image. Mouth region mask is used as the informative region for the local discriminator. The third discriminator is used to keep the expression intensity value within the range [0, 1]. BCE loss is used to train the generator and the discriminators.

The first row is the expression intensity value of eight frames (i.e., the e values depicted above the facial image), and the second and the third rows show the results of AffineGAN

Table 1: List of experimental classes conducted.

Notation	Experiment	Dataset used
AffineGAN	Baseline work	MUG and local facial expression dataset
MAEGAN	Expression intensity inferred as stated, boundary image used as an informative region for the local discriminator and MAE loss used to train the generator and the discriminators	MUG and local facial expression dataset
MAEGAN + SA	With the addition of the self-attention layer and with local discriminator	MUG and local facial expression dataset

on the MUG dataset and local dataset, respectively. As shown in Figure 6, in the second row (i.e., the MUG dataset), there is a sudden change at expression intensity value of 0.142 which might affect the consistency of the video. This shows that there is a probability that inconsistency might occur in the AffineGAN network. There is also a quality problem at expression intensity values 0.857 and 1 on the MUG dataset. On the local dataset, (i.e., as shown in the third row) there is an issue with the mouth and teeth part.

4.2. MAEGAN. To tackle the problem with AffineGAN, the experiment called MAEGAN was carried out. MAEGAN is a mean absolute error loss function introduced in the generator network. The MAEGAN experiment bases expression intensity inferred as stated earlier, employee's boundary image used as an informative region for the local discriminator and mean absolute error (MAE) loss was used to train the generator and the discriminators. In this experiment, the expression intensity value is inferred as stated in the model architecture section, to make the generator generate consistent image sequences as well as different size videos while remaining in the range [0, 1]. Unlike the AffineGAN, the alpha discriminator is not necessary for this case since the NumPy linesapce function bound the value to the range [0, 1]. For the local discriminator, the boundary image is used as an informative region to calculate the loss between the multiplication of the informative region and the real image with the multiplication of the informative region with a produced image. Instead of using the mouth region only as an informative region like [11, 15], the boundary image is used as an informative region since the boundary image can hold the structure of the face image and is also better at describing the facial pose than only the mouth area. Mean absolute error loss is utilized to train the generator and the discriminators. With this experiment as shown in Figure 7, we have noticed some blurry noise which we believe it is as a result of detail abstract features were missed in the intermediate layers. To this end, we have introduced attention mechanism called, MAEGAN + SA, as described in Section 4.3. Each three rows are as described as in the previous section.

4.3. MAEGAN + SA. To solve the blurry problem in the previous experiment, a self-attention layer (SA) was added at the decoder of the generator. The first row is the expression intensity value of eight frames, and the second and the third rows show the results of MAEGAN on the MUG dataset and local dataset, respectively. As shown in Figure 8, there are no sudden changes in performing the expression, and the quality of the image sequence also gets better compared to the previous one. But generated images are somehow blurry, especially at peak expression intensity value in which further intervention is required. However, the generated images are far better than those of the images generated by AffineGAN (e.g., see the image generated at $e = 0.142$), as shown in Figure 8. In general, compared to the previously introduced AffineGAN and MAEGAN, MAEGAN + SA has better visual quality as shown in Figure 8 and Table 2, as shown in Table 2.

The experiment conducted with AffineGAN, MAEGAN, and MAEGAN + SA on one expression, i.e., only on the happy expression. In addition to those experiments conducted happy expression, we have conducted an experiment with both MUG and LOCAL dataset with the four facial expressions presented earlier in the dataset section. Figure 9 shows an experiment conducted on the MUG dataset with the four expressions. The first row is the expression intensity value of eight frames, the second row is the produced frames with happy expression, the third row shows the produced frames with anger expression, the fourth row shows the produced frames with surprise expression, and the last row shows the produced frames with disgust expression with MAEGAN + SA.

Similarly, the two figures shown in Figures 10 and 11 show the result of MAEGAN + SA on the local dataset with the four expressions presented earlier. The first row is the expression intensity value of eight frames, the second row is the produced frames with happy expression with near 45-degree left position, the third row shows the produced frames with anger expression with near 90-degree left position, the fourth row shows the produced frames with surprise expression with near 45-degree right position, and the last row shows the produced frames with disgust expression with near 90-degree right position with the MAEGAN + SA model on the local dataset. Figures 10 and 11 show the generated image sequences for the anger expression with frontal and two sides of the left and right side on the local dataset.

To further visualize our experiment, we have experimented with only three and five frames which are different from the eight-frame experimented as shown in Figures 12 and 13. In Figure 12, the first row is the expression intensity value of five frames and the second row is the produced frames of the MAEGAN + SA model with a happy expression on the local facial expression dataset. Similarly, in Figure 13, the first row is the expression intensity value of three frames and the second row is the produced frames of the MAEGAN + SA model with a happy expression on the local facial expression dataset.

Apart from the subjective evaluation, we have evaluated the proposed model objectively. To this end, we have used the metrics called average content distance (ACD), ACD-I, ACD-C, and ACD-G. ACD-I is utilized to evaluate the quality of face identities, and it computes the average distance between the original image and the output frames. ACD-C is used to assess content consistency, and it computes the average distance between all possible pairs of frames in a video. ACD-G is used to evaluate expression changes it computes the average frame-to-frame distance between the generated frames and the corresponding ground-truth ones. Accordingly, MAEGAN achieved 1.58 ± 0.016 ACD-I, 1.414 ± 0.007 ACD-C, and 1.738 ± 0.011 ACD-G, and MAEGAN + SA has achieved $\mathbf{1.584 \pm 0.001}$ ACD-I, $\mathbf{1.430 \pm 0.009}$ ACD-C, and $\mathbf{1.744 \pm 0.01}$ ACD-G. If we see these results, it is far better than the base line AffineGAN. The bold value indicates the best result in the experiments shown in Table 2.

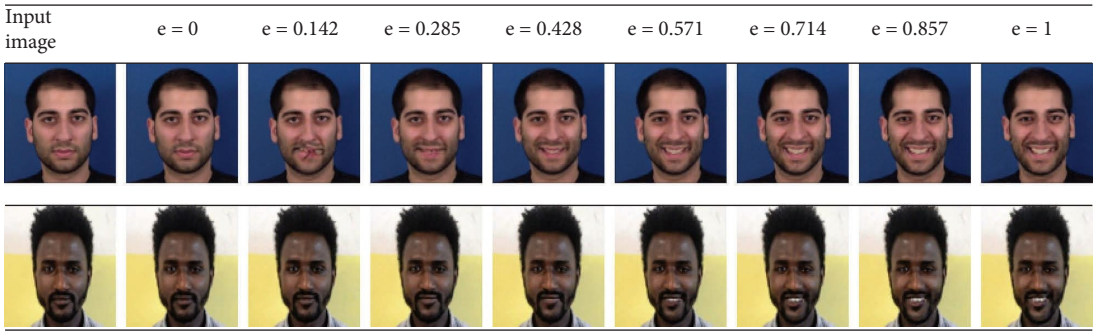

Figure 6: Facial image generated as a result of AffineGAN experimentation.

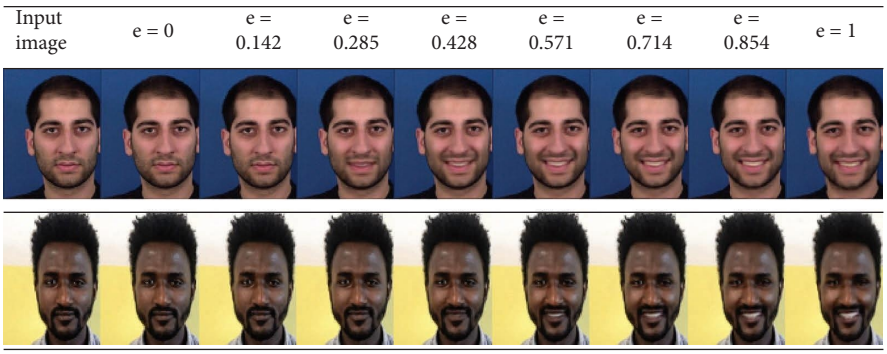

Figure 7: Facial image generated as a result of MAEGAN experimentation.

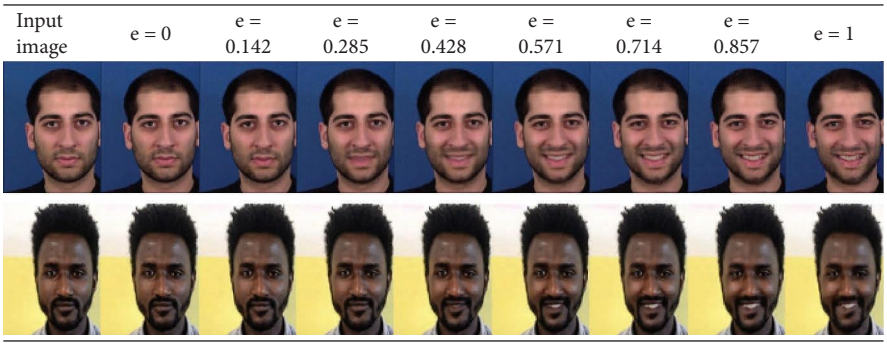

Figure 8: Facial image generated as a result of MAEGAN + SA experimentation.

Table 2: Average content distance, ACD (ACD-I, ACD-C, and ACD-G) scores result of the AffineGAN, MAEGAN, MAEGAN + SA, and the ground truth.

Numbers	Experiment	ACD-I	ACD-C	ACD-G
1	AffineGAN	1.606 ± 0.018	1.452 ± 0.008	1.769 ± 0.007
2	MAEGAN	1.58 ± 0.016	1.414 ± 0.007	1.738 ± 0.011
3	MAEGAN + SA	**1.584 ± 0.001**	**1.430 ± 0.009**	**1.744 ± 0.01**
4	Ground truth	1.38 ± 0.02	1.400 ± 0.01	0

All the ACD scores were also computed for the real (ground truth) for reference, and the last row indicates the result of the ACD scores for the ground truth. The ACD scores are lower, the better the model performance is.

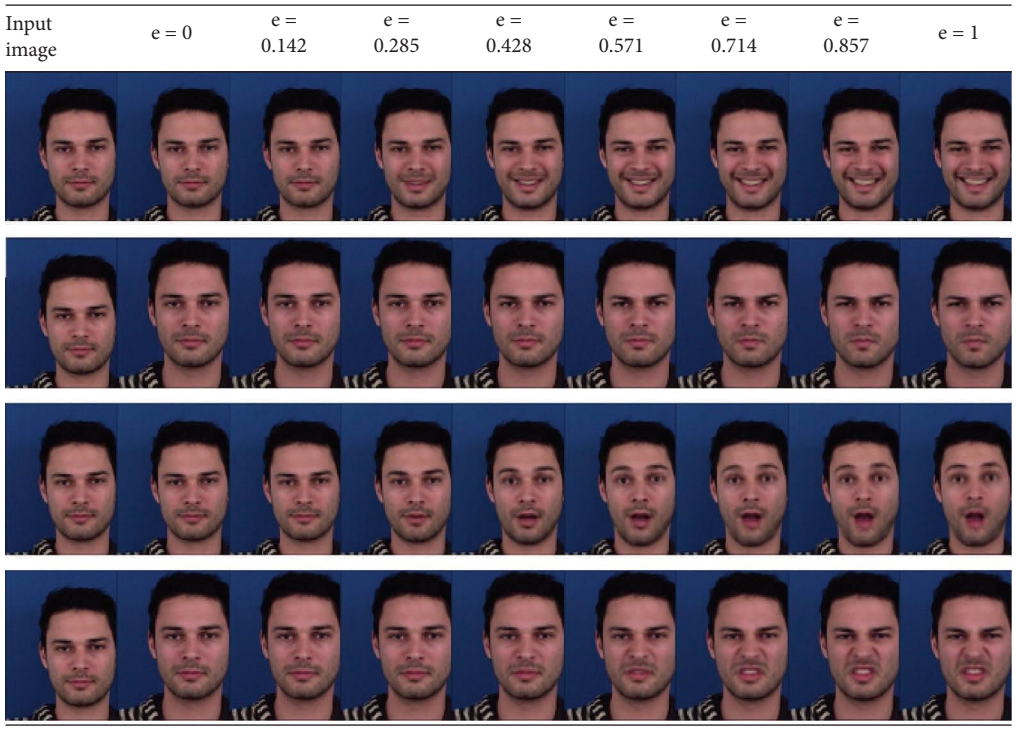

Figure 9: The generated frames with MAEGAN + SA for the four categories of expression on the MUG dataset.

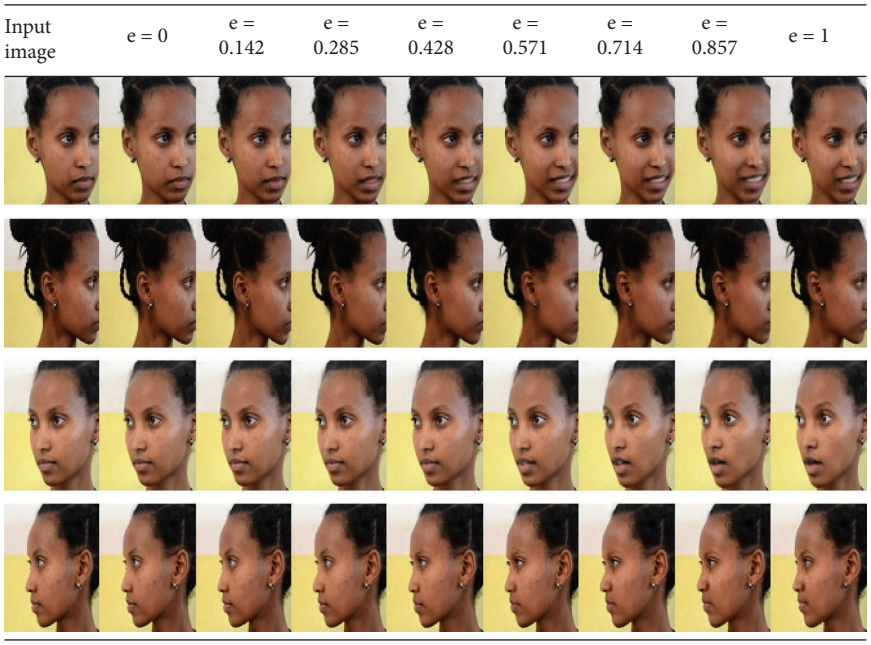

Figure 10: The generated frames with MAEGAN + SA for the four categories of expression on the local dataset (sample 1).

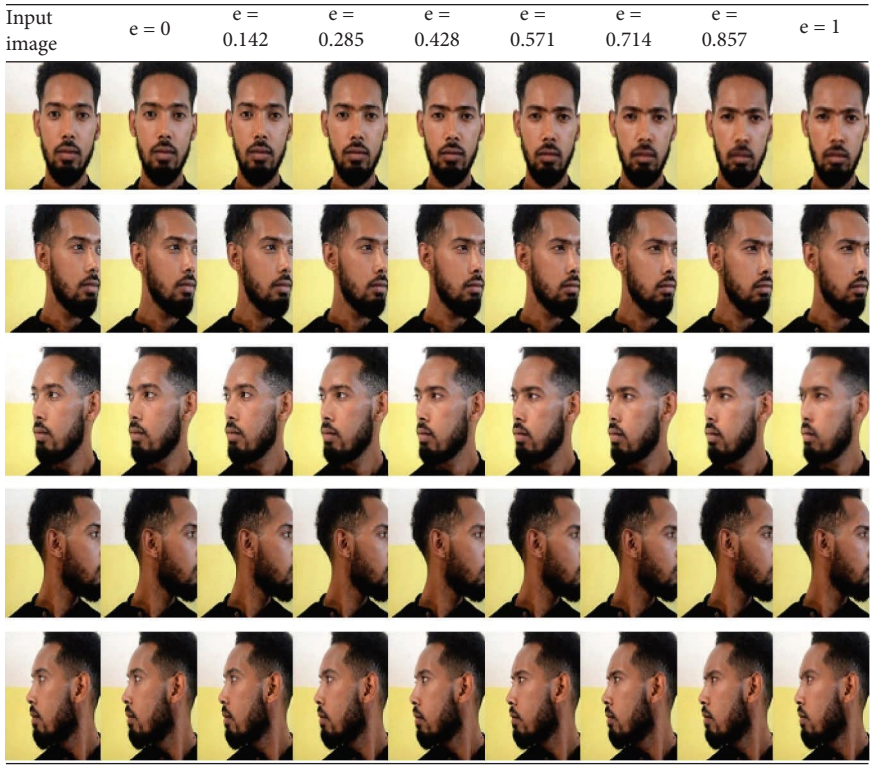

Figure 11: The generated frames with MAEGAN + SA for the four categories of expression on the local dataset (sample 2).

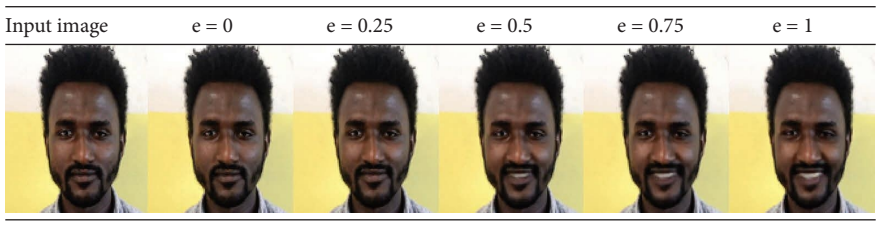

Figure 12: Generated image sequences with five frames with MAEGAN (local dataset).

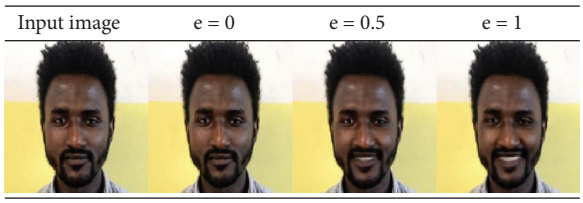

Figure 13: Generated image sequences with three frames with MAEGAN (local dataset).

5. Conclusion

Generally, this study introduced a method to infer the expression intensity value to make the generator able to generate consistent and varying frame size videos. This work also introduced self-attention disregarding the channel attention mechanisms to the network architecture and trains the model with mean absolute error loss. The local dataset was also prepared with frontal and with two different face positions on the left and right sides.

Finally, this work concludes that adding self-attention to the generator network, and utilizing MAE loss as adversarial loss improved the quality of the image sequences being generated. Evenly distributing values according to the frame number to assign expression intensity value improves the consistency of image sequences being generated and also enables the generator to generate different frame size videos while remaining in the range [0, 1].

Data Availability

The data used to support the study are included within the article.

Disclosure

This paper is prepared from student thesis Attention-based Image-to-Video Translation for Synthesizing Facial Expression using Generative Adversarial Network, in partial fulfillment of the requirement for the degree of master's in computer science and engineering.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Kidist Alemayehu has contributed in the designing, experimenting, presenting, and writing of the draft, and Worku Jifara (PhD) has contributed in formulating the problem, editing the entire document, organizing the entire paper, and improving the entire paper quality.

Acknowledgments

This research project was funded by Adama Science and Technology University under grant no. ASTU/SM-R/239/21.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the Lecture Notes in Computer Science, 11211 LNCS*, pp. 294–310, Munich, Germany, September 2018.

[3] Y. Yan, Y. Huang, S. Chen, C. Shen, and H. Wang, "Joint deep learning of facial expression synthesis and recognition," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2792–2807, 2020.

[4] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of the International Conference on Machine Learning*, pp. 7354–7363, PMLR, Vienna, Austria, July 2019.

[5] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings-30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, Hawaii, January 2017.

[6] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, Cambridge, MA, USA, June 2017.

[7] W. Weng and X. Zhu, "INet: convolutional networks for biomedical image segmentation," *IEEE Access*, vol. 9, pp. 16591–16603, 2021.

[8] Y. Zhou, Y. Song, and T. L. Berg, "Image2GIF: generating cinemagraphs using recurrent deep Q-networks," in *Proceedings-2018 IEEE Winter Conference on Applications of Computer Vision, WACV*, Lake Tahoe, NV, USA, March 2018.

[9] M. Liu, T. Wang, J. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, Video-to-Video synthesis," 2018, https://arxiv.org/abs/1808.06601.

[10] X. Wang, W. Li, G. Mu, and D. Huang, "Facial expression synthesis by u-net conditional generative adversarial networks," in *Proceedings of the 2018 ACM International Conference on Multimedia Retrieval*, pp. 283–290, Yokohama Japan, June 2018.

[11] L. Fan, W. Huang, C. Gan, J. Huang, and B. Gong, "Controllable image-to-video translation: a case study on facial expression generation," in *Proceedings of the 33rd aaai conference on artificial intelligence, aaai 2019, 31st innovative applications of artificial intelligence conference, iaai 2019 and the 9th aaai symposium on educational advances in artificial intelligence*, Honolulu, Hawaii, USA, February 2019.

[12] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. Metaxas, "Learning to forecast and refine residual motion for image-to-video generation," *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, Berlin, Germany, 2018.

[13] X. Zheng, Y. Guo, H. Huang, Y. Li, and R. He, "A survey of deep facial attribute analysis," *International Journal of Computer Vision*, vol. 128, no. 8–9, pp. 2002–2034, 2020.

[14] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[15] G. Shen, W. Huang, C. Gan et al., "Facial image-to-video translation by a hidden affine transformation," in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 2505–2513, Nice France, October 2019.

[16] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[17] B. Bozorgtabar, D. Mahapatra, and J. P. Thiran, "ExprADA: adversarial domain adaptation for facial expression analysis," *Pattern Recognition*, vol. 100, 2020.

[18] M. Chen, C. Li, K. Li, H. Zhang, and X. He, "Double encoder conditional GAN for facial expression synthesis," in *Proceedings of the Chinese Control Conference*, CCC, Wuhan, China, June 2018.

[19] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, "StarGAN: unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings-IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, PR, USA, June 2018.

[20] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He, "High-Fidelity face manipulation with extreme poses and expressions," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2218–2231, 2021.

[21] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, "Geometry-Contrastive GAN for Facial Expression Transfer," 2018, https://arxiv.org/abs/1802.01822.

[22] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *Proceedings of the 2018 ACM Multimedia Conference*, Seoul Republic of Korea, October 2018.

[23] Z. Lu, T. Hu, L. Song, Z. Zhang, and R. He, "Conditional expression synthesis with face parsing transformation," in *Proceedings of the 2018 ACM Multimedia Conference*, Seoul Republic of Korea, October 2018.

[24] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: anatomically-aware facial animation from a single image," *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, Berlin, Germany, 2018.

[25] E. Ververas and S. Zafeiriou, "Slidergan: synthesizing expressive face images by sliding 3d blendshape parameters," *International Journal of Computer Vision*, vol. 128, no. 10-11, pp. 2629–2650, 2020.

[26] H. Ding, K. Sricharan, and R. Chellappa, "ExprGAN: facial expression editing with controllable expression intensity," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI*, February 2018.

[27] S. Qian, K. Y. Lin, W. Wu et al., "Make a face: towards arbitrary high fidelity face manipulation," in *Proceedings of the IEEE International Conference on Computer Vision*, Cambridge, MA, USA, October 2019.

[28] A. Kidist and J. Worku, "An Attention-Based Image-To-Video Translation for Synthesizing Facial Expression Using GAN," Student Thesis, Adama Science and Technology University, Adama, Ethiopia, 2022.

[29] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of the 36th International Conference on Machine Learning*, ICML, Long Beach, CA, USA, June 2019.

[30] Z. Yuan, M. Jiang, Y. Wang et al., "SARA-GAN: self-attention and relative average discriminator based generative adversarial networks for fast compressed sensing MRI reconstruction," *Frontiers in Neuroinformatics*, vol. 14, pp. 611666–666, 2020.

[31] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[32] Q. Huang, D. Yang, P. Wu, H. Qu, J. Yi, and D. Metaxas, "MRI reconstruction via cascaded channel-wise attention network," in *Proceedings of the International Symposium on Biomedical Imaging*, pp. 1622–1626, Venice, Italy, April 2019.

[33] D. P. Kingma, J. Ba, and L. Adam, "A method for stochastic optimization," in *Proceedings of the 3rd international conference on learning representations, ICLR 2015-Conference Track Proceedings*, vol. 1–15, Guilin China, May 2015.

[34] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D Face Alignment problem?(and a dataset of 230, 000 3D facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1021–1030, Cambridge, MA, USA, June 2017.