information

MDPI

Article

Generative Adversarial Networks (GANs) for Audio-Visual Speech Recognition in Artificial Intelligence IoT

Yibo He [1], Kah Phooi Seng [1,2,3] and Li Minn Ang [3,*]

[1] School of AI and Advanced Computing, Xi'an Jiaotong Liverpool University, Suzhou 215000, China; yibo.he22@xjtlu.edu.cn (Y.H.); jasmine.seng@xjtlu.edu.cn (K.P.S.)
[2] School of Computer Science, Queensland University of Technology, Brisbane City, QLD 4000, Australia
[3] School of Science Technology and Engineering, University of the Sunshine Coast, Sippy Downs, QLD 4556, Australia
* Correspondence: lang@usc.edu.au

Abstract: This paper proposes a novel multimodal generative adversarial network AVSR (multimodal AVSR GAN) architecture, to improve both the energy efficiency and the AVSR classification accuracy of artificial intelligence Internet of things (IoT) applications. The audio-visual speech recognition (AVSR) modality is a classical multimodal modality, which is commonly used in IoT and embedded systems. Examples of suitable IoT applications include in-cabin speech recognition systems for driving systems, AVSR in augmented reality environments, and interactive applications such as virtual aquariums. The application of multimodal sensor data for IoT applications requires efficient information processing, to meet the hardware constraints of IoT devices. The proposed multimodal AVSR GAN architecture is composed of a discriminator and a generator, each of which is a two-stream network, corresponding to the audio stream information and the visual stream information, respectively. To validate this approach, we used augmented data from well-known datasets (LRS2-Lip Reading Sentences 2 and LRS3) in the training process, and testing was performed using the original data. The research and experimental results showed that the proposed multimodal AVSR GAN architecture improved the AVSR classification accuracy. Furthermore, in this study, we discuss the domain of GANs and provide a concise summary of the proposed GANs.

Keywords: Internet of things (IoT); generative adversarial networks (GANs); deep learning; audio-visual speech recognition

check for updates

Citation: He, Y.; Seng, K.P.; Ang, L.M. Generative Adversarial Networks (GANs) for Audio-Visual Speech Recognition in Artificial Intelligence IoT. *Information* **2023**, *14*, 575. https://doi.org/10.3390/info14100575

Academic Editors: Vasco N. G. J. Soares, João M. L. P. Caldeira, Bruno Bogaz Zarpelão and Jaime Galán-Jiménez

Received: 9 September 2023
Revised: 12 October 2023
Accepted: 17 October 2023
Published: 19 October 2023

1. Introduction

The Internet of things (IoT) refers to the connection of physical and virtual objects to the Internet for data collection, exchange, and automated operations, to provide a smarter and more convenient living and working experience. Initially introduced in 1999, the IoT has since evolved from its early stages to become a tangible reality, driven by the rapid advancement and extensive utilization of cloud computing [1] and wireless sensor networks [2]. It has been applied to many fields, which include smart homes, environment monitoring, and intelligent transportation. The wide application of IoT is due to a variety of key technologies, including radio-frequency identification (RFID) technology [3], sensor networks, computer vision technology, and intelligent computing techniques [4].

The IoT leverages the mentioned key technologies to collaboratively enhance environmental sensing capabilities, giving it a distinct advantage. Inspired by this property, we apply the IoT paradigm to multimodal audio-visual speech recognition. Traditional audio-visual speech recognition (AVSR) only recognizes video–audio in a single scene for classification. In contrast to them, IoT with heterogeneous sensors provides an opportunity for effective multimodal audio-visual speech recognition. Specifically, the IoT not only captures the visual data of the speaker through visual sensors, but also collects the speaker's environmental information, such as ambient noise and lighting conditions, in combination with audio sensors. Multiple environmental factors can improve the classification accuracy of AVSR.

However, IoT generates a large amount of multimodal data, and these big data have complex characteristics that pose significant challenges for data storage and identification. The motivation of this paper was to investigate a novel multimodal generative adversarial network AVSR architecture, which simultaneously guarantees energy efficiency for IoT and classification accuracy for multimodal audio-visual speech recognition. The architecture should fulfill the following requirements: (1) First, the processed data can be stored in a small storage space; and (2) second, the processed data can correctly reflect the feature information of the original data for further applications. Specifically, we utilize the processed data for audio-visual speech recognition classification.

In recent decades, great efforts have been made toward accurate audio-visual speech recognition classification. Audio-visual speech recognition was an early application of multimodal audio-visual sensing in the 1950s. This type of research was influenced by the McGurk effect [5], which states that vision and sound interact. The AVSR sensor architecture is constructed from three modules: (1) an audio/acoustic module, which extracts speech features and is robust to noise; (2) a visual/image module, which extracts contour and color information from the mouth area; and (3) a fusion module, which combines/fuses the extracted audio features and visual features. One example of a fusion module uses hidden Markov models (HMMs) [6]. The earliest work on AVSR can be dated back to around two decades ago, when using hand-crafted visual features to improve HMM-based ASR systems was proposed [7]. The first modern AVSR system [8] proposed using deep neural networks. The field has been rapidly developing since then. Most works are devoted to architectural improvements; for example, Zhang et al. [9] proposed a temporal focal block and spatiotemporal fusion. Another line of research focuses on a more diversified learning scheme to improve AVSR performance. Li et al. [10] used a cross-modal student–teacher training scheme.

In tandem with the advancement of precise audio-visual speech recognition classification, researchers have initiated the integration of AVSR into Internet of things (IoT) devices. In a prior investigation, Mehrabani et al. [11] introduced the incorporation of HMM-based ASR systems into smart connected homes, enhancing convenience and bolstering security. Additionally, Dabran et al. [12] developed tools for real-time caption generation, to assist those with hearing impairments. Nevertheless, the application of HMM-based speech recognition models in real-world scenarios has yielded unsatisfactory classification performance. Furthermore, Ma et al. [13] identified that the speech recognition framework for smart IoT devices transmitted speech data in plain text, posing a potential risk to user privacy. Consequently, an outsourced privacy-preserving speech recognition framework was proposed, utilizing long short-term memory (LSTM) neural networks and edge computing, thereby enhancing the classification performance, while preserving data security. With the increasing adoption of speech recognition technology in IoT devices, Bäckström [14] investigated the establishment of unified and standardized communication protocols among voice-operated devices, emphasizing the paramount importance of preserving privacy in IoT devices. Subsequently, privacy protection has emerged as a foremost consideration in the realm of IoT application technologies.

Recently, generative adversarial networks (GANs) have been extensively researched and have made significant progress in optimizing many areas. The flexibility of GANs means they can be used in machine learning to solve different problems, from generative tasks such as image synthesis [15], style transfer [16], super-resolution [17], and image completion [18], to decision-making tasks such as classification [19] and segmentation [20]. In early research on the super-resolution of GANs, style-based generative adversarial networks (StyleGAN) [21] were the primary generative models. In 2017, the authors of the StyleGAN series proposed progressive growing of GANs (Progressive-GAN) [22], which is based on the idea of gradually increasing the resolution of the generator and discriminator during training. However, shifting to high-resolution produces a different level of effectiveness. Therefore, Tero Karras et al. proposed StyleGAN in 2018. StyleGAN is a high-resolution image generation model that includes facial expressions, face orientation, and hairstyle, as well as texture details such as skin tone and brightness. At this stage, StyleGAN has been proven to work reliably on various datasets. StyleGAN2 [23] optimized and improved the image quality of StyleGAN and addressed the issue of image artifacts. Based on the ability of StyleGAN2 to generate sufficiently realistic face images, generative facial prior generative adversarial networks (GFP-GAN) [24] were proposed for performing face restoration. The core of GFP-GAN utilizes the "knowledge" contained in the trained face generation model, termed a generative facial prior (GFP), such as in StyleGAN2. GFP contains rich details of features and face color and treats the face as a whole, dealing with hair, ears, and facial contours. Compared to other face super-resolution methods, GFP-GAN provides better detail in the recovery of features, a more natural overall appearance, and enhanced color. However, the existing generative adversarial architectures are not suitable for generating multimodal audio-visual speech recognition data.

The aim of this paper is to propose a novel multimodal generative adversarial network AVSR architecture for IoT. First, we discuss different variants of the GAN algorithmic architecture and briefly summarize the proposed GAN variants. Second, we propose a GAN architecture for multimodal audio-visual sensing. We apply a fusion of a GFP-GAN and an audio-visual speech recognition model to the super-resolution processing of low-quality images generated by IoT sensors, to enhance facial feature details and allow for more accurate feature extraction on the visual side to improve performance.

The contributions of this paper can be summarized as follows:

- Introduction of a novel multimodal generative adversarial network AVSR architecture for IoT: We propose an innovative AVSR architecture, leveraging a multimodal generative adversarial network, which combines GFP-GAN and audio-visual speech recognition models. This integration enhances facial feature details, resulting in improved classification performance for IoT applications.
- Exploration of model lightweighting techniques: we investigate various model lightweighting techniques, such as modular task design and the integration of cache modules. These optimizations effectively reduce the computational complexity of the model, while preserving a high performance. This adaptability makes the model well-suited for deployment in resource-constrained IoT devices.
- Focus on privacy and security: we delve into the privacy and security considerations associated with different data sources. Our research includes the establishment of rigorous privacy protocols, authentication mechanisms, and the incorporation of federated learning principles. These measures collectively enhance the privacy and security of model data, which is critical for IoT applications.
- Extensive experimental validation: through extensive experimental evaluations, we demonstrate the versatility of our AVSR architecture in various IoT scenarios, affirming its applicability across diverse real-world contexts.

The paper is organized as follows: Related works on different GAN variant architectures are discussed in Section 2. Section 3 describes the methodology used for the study, based on the GFP-GAN architecture for AVSR and the Wave2Lip-GAN architecture for AVSR. Section 4 discusses the results. We give some concluding remarks in Section 5.

2. Literature Review

Since the introduction of GANs, many researchers have used GANs to enhance the data processing, model optimization, and security of IoT systems, to improve IoT performance and intelligence. These IoT networks using GANs have been applied to various tasks and have shown impressive performance. In this section, we discuss several important GAN variant architectures applied to the tasks of picture conversion, image generation, and image super-resolution of IoT sensor data, starting with the basic GAN architecture.

2.1. Generative Adversarial Networks

The classical approach to GANs is an artificial neural network consisting of two models: (1) a generator network, and (2) a discriminator network. As shown in Figure 1, the basic theory behind GANs is that the generator learns to create new data that are similar to a given dataset. In contrast, the discriminator learns to differentiate between the accurate and the generated data.

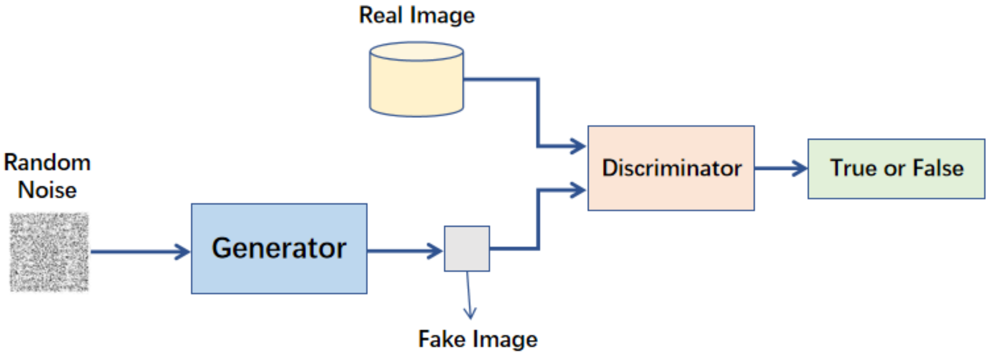

Figure 1. The basic structure of a GAN.

The generator takes random noise as input and attempts to generate an output that is similar to the training data. The discriminator takes both real and generated data and tries to distinguish between the real data and the fake data. The two models are trained simultaneously in a game-like process, where the generator tries to produce better and more convincing outputs, while the discriminator tries to become better at detecting fake data.

The ultimate goal of GANs is to have a generator that can produce an output that is indistinguishable from real data, and a discriminator that can no longer tell the difference between real and generated data. Goodfellow provided a mathematical description of the training loss function for the generator and discriminator $\min_G \max_D V(D, G)$, denoted as G and D, respectively. The following equation describes the aim of the GAN:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \tag{1}$$

The training of the GAN is iterated in two steps, to reach the optimization goal. First, the discriminator is trained to differentiate the real samples from the fake samples. The optimization objective is derived from the aforementioned formula, specifically the maxV part. Since it is a maximization term, during the optimization process using gradient descent, the loss function $\mathcal{L}_D$ for optimization is as follows:

$$\mathcal{L}_D = -\mathbb{E}_{x \sim p_{gt}(\mathbf{x})}[\log D(\mathbf{x})] - \mathbb{E}_{z \sim N(\mathbf{z}|0,\mathbf{I})}[\log(1 - D(G(z))] \tag{2}$$

The second step involves training the generator, denoted as G. The optimization objective is to minimize the value function V. It is important to note that the total loss function only includes the second term related to G. Therefore, the loss function $\mathcal{L}_G$ should be

$$\mathcal{L}_G = \mathbb{E}_{z \sim N(z|0,I)}[log(1 - D(G(z))] \tag{3}$$

2.2. Fully Connected Generative Adversarial Network

The original energy-based GAN utilized fully connected neural networks to construct generators and discriminators. This architectural variant is commonly employed for straightforward image datasets like MNIST [25] and CIFAR-10 [26]. However, in this GAN variant, the generator employs both ReLU and sigmoid activation functions. Unfortunately, this GAN did not exhibit satisfactory generalization performance as the complexity of the images was increased. Consequently, the GANs could only generate lower-resolution images of 32 × 32 pixels when applied to the MNIST dataset.

2.3. Deep Convolutional Generative Adversarial Network

Following the successful applications of convolutional neural networks (CNNs) [27] in computer vision tasks, researchers [28] recognized their potential and combined CNNs with GANs to introduce deep convolutional generative adversarial networks (DCGAN). In this innovation, they employed a deconvolutional neural network architecture [29] for the generator and replaced the original multi-layer perceptron (MLP) structure with fully convolutional networks [30]. Additionally, DCGAN incorporated batch normalization [31] and ReLU activation [32]. Consequently, DCGAN gained popularity, and the utilization of deconvolution became a widely adopted architectural approach to GAN generators. Figure 2 illustrates the architecture of the generator in DCGAN. However, DCGAN exhibits better performance only on images with lower resolution, due to limitations in the model capacity and optimization challenges.

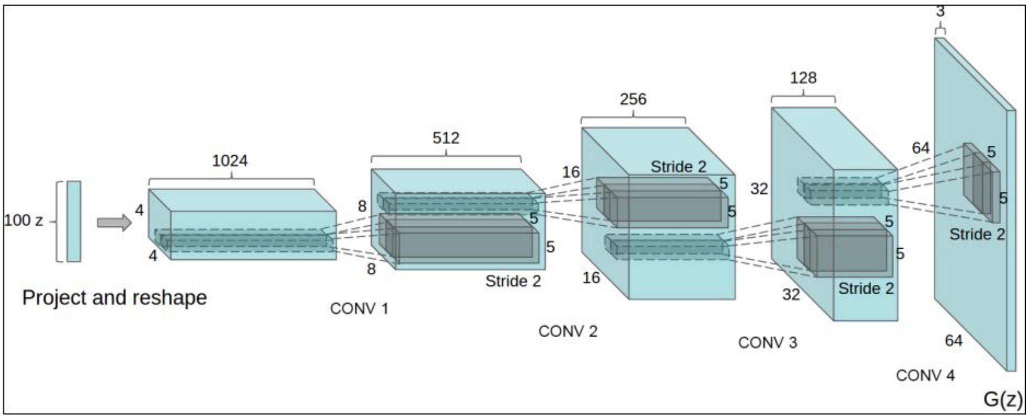

Figure 2. DCGAN generator architecture reproduced from Radford et al. [29].

2.4. Laplacian Pyramid Generative Adversarial Network

Prior to the development of DCGAN, Denton et al. [33] introduced the Laplacian pyramid GAN (LAPGAN). In this GAN architecture, the resolution of the synthetic samples is progressively increased throughout the generation process. LAPGAN utilizes a cascade of CNNs within the framework of the Laplacian pyramid to upsample the images, enabling the generation of synthetic images up to a resolution of 96 × 96 pixels. This cascade structure enhances the training stability and facilitates high-resolution modeling. Figure 3 illustrates the architecture of LAPGAN.

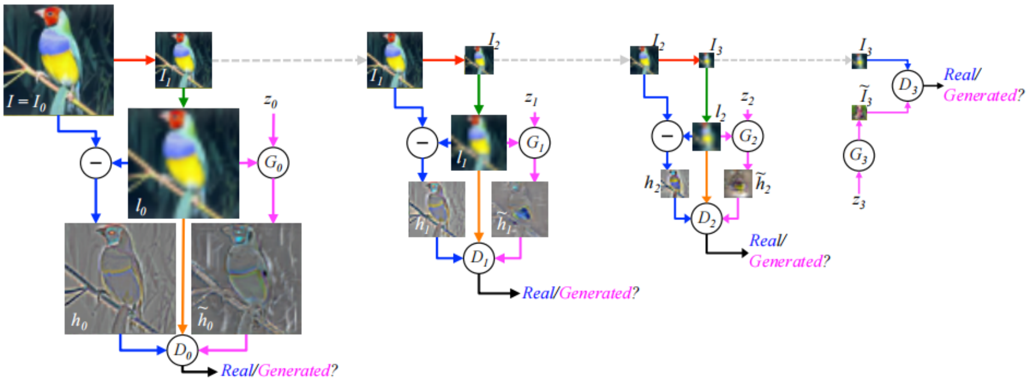

Figure 3. LAPGAN network architecture reproduced from Denton et al. [33].

2.5. Self-Attention Generative Adversarial Networks

Following the proposal of DCGAN, researchers observed that conventional CNNs were limited in their ability to capture global spatial information and effectively model image datasets with multiple classes, such as ImageNet [34]. This posed a challenge for

GAN networks to learn and generate such complex images. To address this, Han Zhang et al. [35] introduced the self-attention generative adversarial network (SAGAN). SAGAN incorporated a self-attention mechanism, allowing it to focus on relevant dependencies and capture long-range spatial relationships in the image generation process. The self-attention module architecture of SAGAN is depicted in Figure 4.

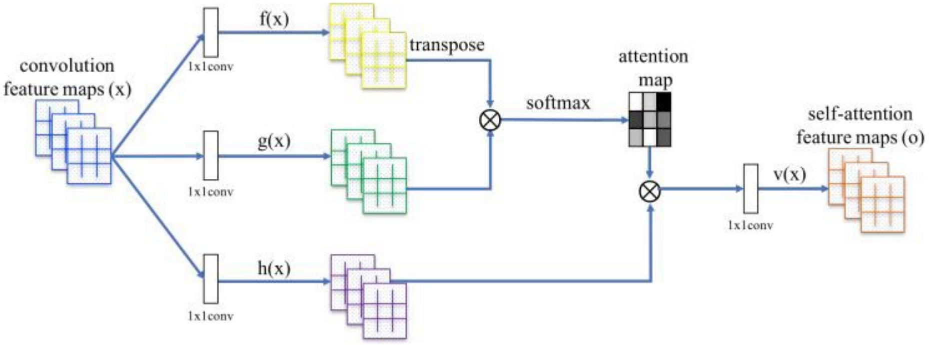

Figure 4. The self-attention module for SAGAN [35].

2.6. Cycle-Consistent Adversarial Networks

Traditional GANs can be used to generate realistic images, but they usually require paired data, which are not readily available. In this context, Junyan Zhu et al. [36] proposed cycle consistent adversarial network (CycleGAN). CycleGAN is an unsupervised image transformation GAN framework that contains two generators and two discriminators. Bidirectional transformation of two uncorrelated image domains is achieved through cyclic consistency loss, which eliminates the need for paired data.

2.7. Bicycle Generative Adversarial Networks

After the proposal of CycleGAN, Junyan Zhu et al. found that their CycleGAN could only satisfy the single-modal requirement. Therefore, they proposed the bicycle generative adversarial network (BicycleGAN) [37] architecture, to introduce a new consistency constraint for multimodal image translation. BicycleGAN allows image translation to expand from unimodal to multimodal. Figure 5 shows the architecture of BicycleGAN, which consists of two generators and two discriminators to realize the bidirectional mapping transformation.

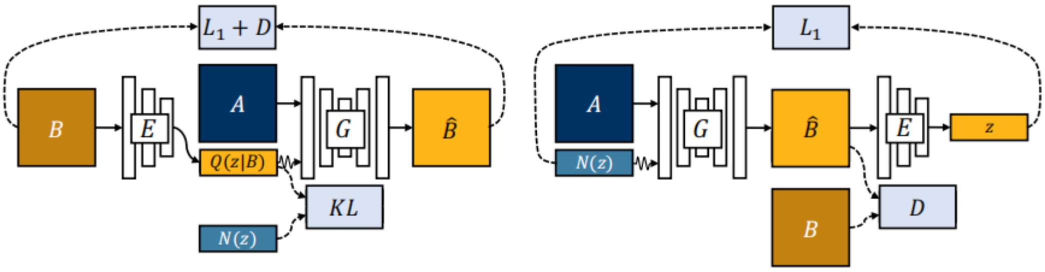

Figure 5. Bidirectional network architecture for BicycleGAN [37].

2.8. Star Generative Adversarial Networks

In the task of image translation, researchers need to train a separate model for each task, which leads to a dramatic increase in the number of models. To overcome this problem, Yunjey Choi et al. [38] proposed a unified GAN architecture (StarGAN) for multi-domain image translation. The overall structure of StarGAN is a conditional domain-based generative adversarial network consisting of a generator and a discriminator. This is different from bidirectional generative adversarial networks (GANs) because StarGAN is designed to achieve multi-domain image translation without the need to train separate generators and discriminators for each pair of transformations between domains. Figure 6 shows the structure of StarGAN, where the two generators shown are split by the same generator.

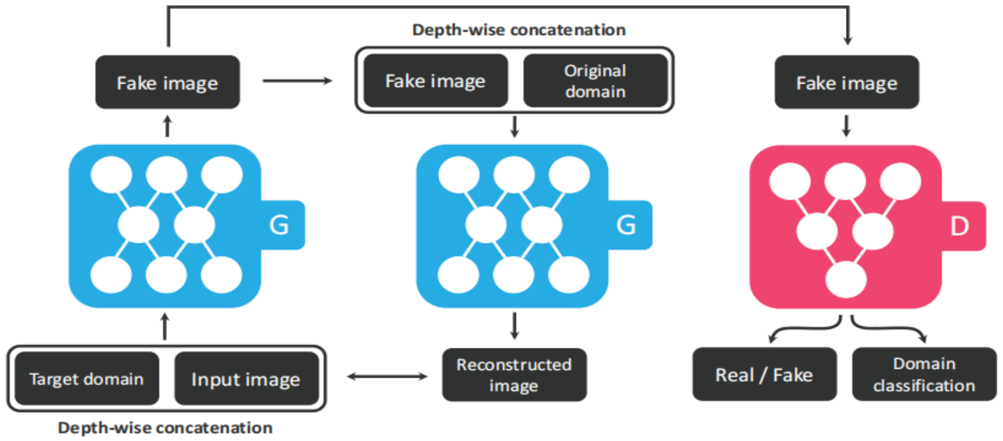

Figure 6. Overview of StarGAN [38], consisting of two modules: a discriminator D, and a generator G.

2.9. Super-Resolution Generative Adversarial Network

For the generation task, how to generate high-quality and high-resolution images is challenging. Traditional image super-resolution methods are mainly based on interpolation or filtering techniques, and this very easily leads to image distortion and blurring. In this context, Christian Ledig et al. [39] were pioneers in successfully applying GAN to image super-resolution with their proposed super-resolution generative adversarial network (SRGAN). SRGAN consists of a generator that introduces the residual dense blocks module and a discriminator. The residual dense blocks generator captures the features and details of an image more efficiently, which is then coupled with the perceptual loss, to generate a more realistic and detailed high-resolution image. Figure 7 shows the internal structure of the SRGAN generator.

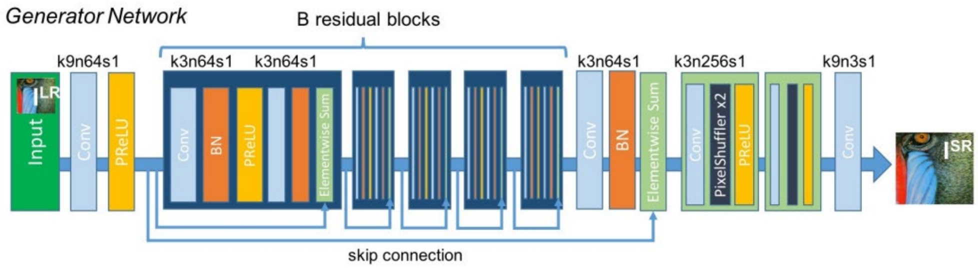

Figure 7. SRGAN architecture of generator [39].

2.10. Diverse Generative Adversarial Network

Although GAN has made significant developments in image super-resolution, it lacks high-resolution textures for the generated images. Thus, Masoumeh Zareapoor et al. [40] proposed diverse generative adversarial network (DGAN). DGAN is a diverse GAN architecture that contains multiple generators and a discriminator. Compared to a single generator, it recovers realistic textures using multiple generators to produce different samples. Figure 8 shows the multiple generator architecture of DGAN. Table 1 shows a summary of GAN architectures for image translation, image generation, and image super-resolution representation.

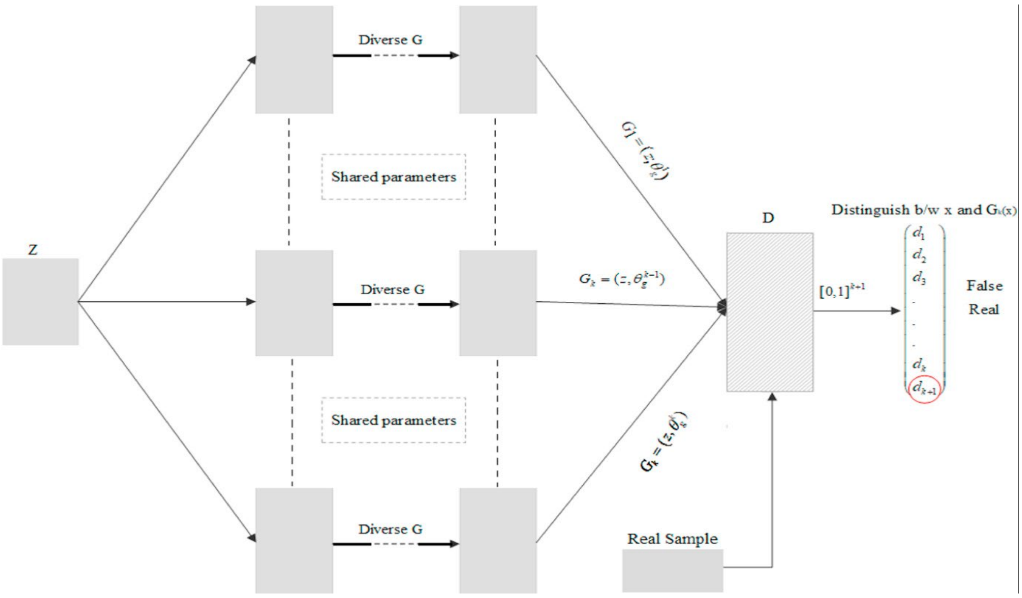

Figure 8. Multiple generator architectures of DGAN [40].

Table 1. Architecture of GANs for image translation, image generation, and image super-resolution representation.

Category/Domain Area	Year	Main Contributions	Datasets	Reference
Research work on GANs for image translation	2017	CycleGAN: Proposed architecture for image translation without pairing data	CycleGAN Pix2Pix	JunYan Zhu et al. [36]
	2018	BicycleGAN: Implementing multimodal image translation	CycleGAN Pix2Pix	JunYan Zhu et al. [37]
	2018	StarGAN: Proposed a Unified GAN architecture for multi-domain image translation	CelebA RaFD ImageNet	Yunjey Choi et al. [38]
Research work on GANs for image generation	2014	FCGAN: Early generative adversarial network models utilizing fully connected neural networks.	MNIST CIFAR10	Ian Goodfellow et al.
	2014	DCGAN: The first model to combine deep convolutional neural networks (CNNs) with generative adversarial networks.	MNIST CIFAR10	Xu et al. [29]
	2019	SAGAN: Proposed structure for GAN image generation with self-attention.	ImageNet	Han Zhang et al. [35]
Research work on GANs for image super-resolution	2015	LAPGAN: Step-by-step image super-resolution using the Laplace pyramid framework.	CIFAR10	Lai et al. [33]
	2017	SRGAN: A milestone in the successful introduction of GAN into the field of image super-resolution.	Set5	Christian Ledig et al. [39]
	2019	DGAN: Implementing a multi-sample image super-resolution architecture.	DIV2K	Masoumeh Zareapoor et al. [40]

3. Multimodal Audio-Visual Sensing

In IoT, multimodal audio-visual sensing is an important technology for multimodal sensors, to capture visual and audio information. This section discusses the proposed multimodal generative adversarial network AVSR architecture. The early part of this section will detail the architecture of multimodal audio-visual sensing. The latter part of the section will discuss how the GAN structure is incorporated into the AVSR architecture. Figures 9 and 10 show the proposed architecture of AVSR. The architecture utilizes deep learning, along with a transformer model (TM-CTC) trained on CTC loss, based on the self-attentive module.

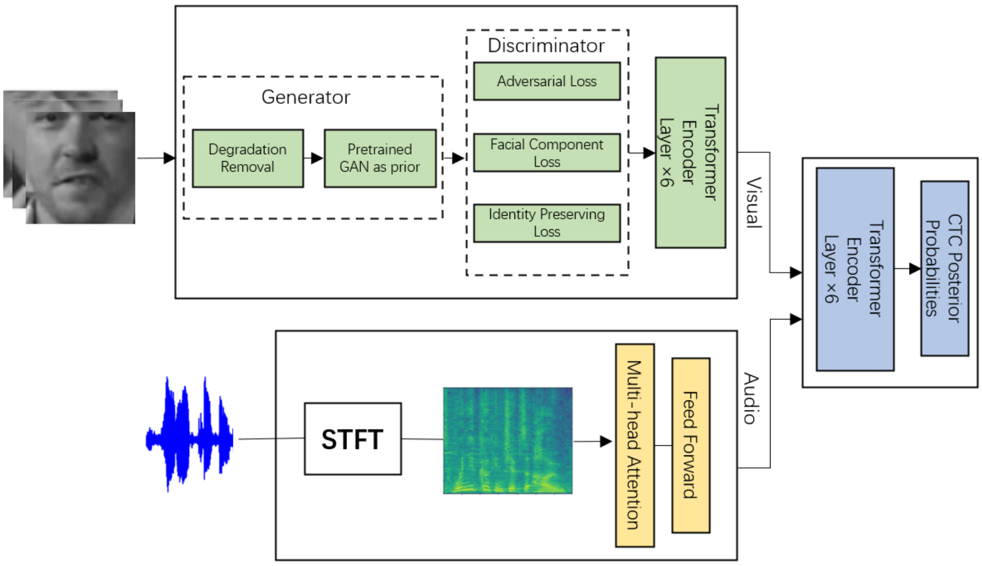

Figure 9. GFP-GAN multimodal AVSR architecture.

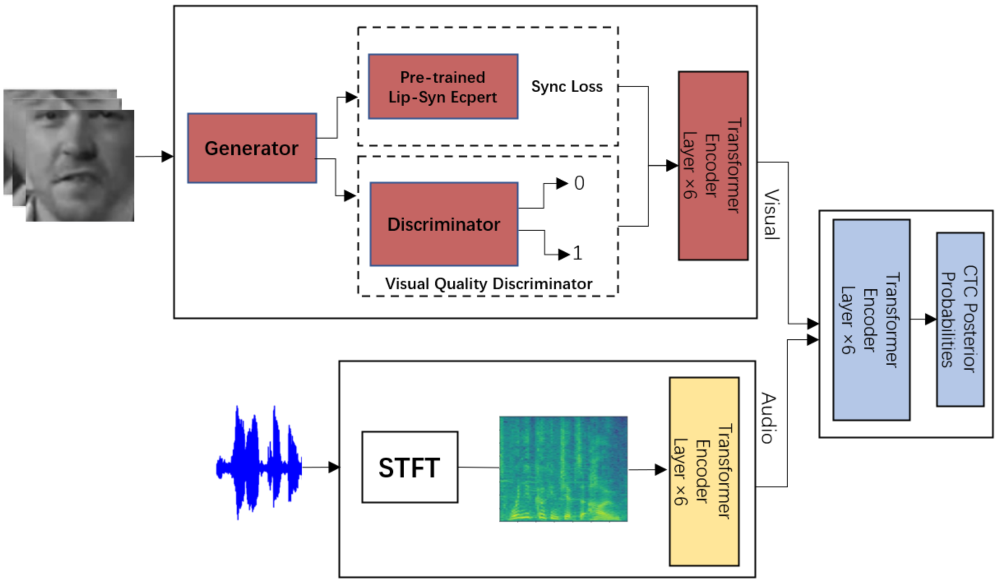
Pre-trained Lip-Syn Ecpert
Sync Loss
Generator
Discriminator
0
1
Visual Quality Discriminator
Transformer Encoder Layer ×6
Visual
Transformer Encoder Layer ×6
CTC Posterior Probabilities
STFT
Transformer Encoder Layer ×6
Audio

Figure 10. Wave2Lip-GAN multimodal AVSR architecture.

3.1. Multimodal Audio-Visual Sensing

Multimodal audio-visual sensing can enhance the ability of IoT systems to comprehensively sense and intelligently respond to the environment by combining multiple sensing modalities such as audio and video. The multimodal data obtained from multimodal sensors can contain more complex feature information such as environmental noise, lighting conditions, etc. This is of great significance for our subsequent classification task of audio-visual speech recognition.

3.2. AVSR with GFP-GAN

This section details the integration of the GAN structure into the AVSR architecture.

Figure 9 illustrates the model architecture for audio-visual speech recognition. The architecture comprises the following components: (1) audio-modal sensing input, (2) video-modal sensing input, and (3) audio-visual modal fusion. The top half of Figure 9 (left) presents the architecture of the video modal sensing input, where the visual images are processed using ResNet. Two-dimensional ResNet and three-dimensional convolution are applied to the images, to align with the audio frames.

Figure 9 (bottom left) showcases the architectural components of the audio-only modal sensing input. The audio signal is processed using a short-time Fourier transform (STFT) to extract audio features. These features convert the original 16 kHz audio waveform into vector representations across a 321-dimensional spectral amplitude. The 321-dimensional spectral amplitude is calculated within a 40 kHz window with a hop length of 10 ms. The video clip has a frame rate of 25 frames per second (40 ms per frame), and each video frame corresponds to four audio frames.

Figure 9 (top left) showcases the architectural components of the visual-only modal sensing input. Before the video sensing input enters the GFP-GAN architecture, the video data are processed into a 112 × 112 image sequence. Figure 11 presents the generator architecture of the AVSR using GFP-GAN. Within the network, the data first enter the generator, which consists of a degradation removal module and a pretrained facial GAN as a prior. The degradation removal module eliminates blurred noise and generates clean multiresolution features. The facial prior in the generator generates features containing rich facial details based on the multi-resolution features, ultimately producing a high-fidelity face image through spatial transformation. The newly generated image is then fed into a discriminator, responsible for determining the discriminative loss settlement for the entire face and local discriminative loss calculations for specific components, such as the left and right eyes and the mouth. The generator and discriminator are connected from coarse to fine by direct latent code mapping and several channel-segmentation spatial-feature-transformation (CS-SFT) layers. In the visual backend, we employ pretrained visual features to extract image characteristics. For the pretraining of these models, we leverage word excerpts from the extensive MVLRS [41] dataset to pretrain the visual front-end. In this pretraining phase, a two-layer temporal convolutional back-end is employed for segment classification. The use of pretrained visual features offers our model a generic, efficient, and high-performance image representation. By incorporating features pretrained on extensive datasets, our model effectively mitigates training complexity and resource demands, thereby expediting model training and enhancing classification accuracy.

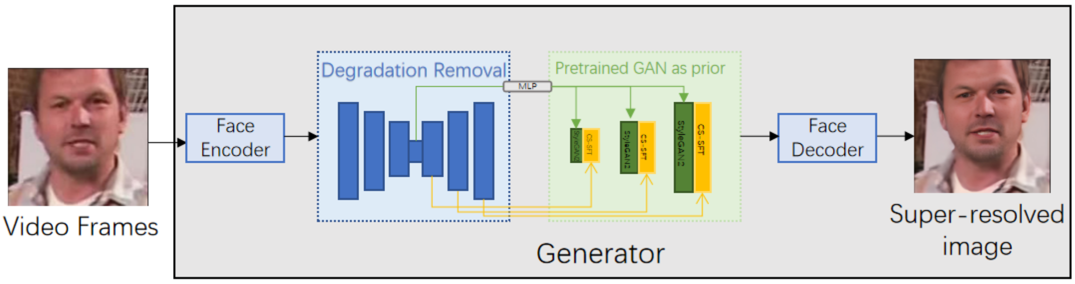

Figure 11. The generator architecture of the AVSR using the GFP-GAN.

Figure 9 (bottom right) demonstrates the architectural components of audio-visual modal fusion. In this stage, the fusion of modalities relies on the TM-CTC model. Two stacks of six transformer encoder layers form the encoder (one for each modality), and a single stack of six transformer encoder layers forms the joint decoder. Encoded feature vectors from these two modalities are connected and linearly transformed into 512-dim vectors. The network generates CTC posterior probabilities for each input frame and trains the entire stack structure to minimize CTC losses.

Figure 12 illustrates the algorithmic flowchart of GFP-GAN. The input image enters the degradation removal module using the U-Net structure. We obtain the latent features and spatial features of the input image through U-net sampling. The latent features undergo processing via a multilayer perceptron to create 16 latent codes, each having a dimension of 512, enhancing the preservation of semantic attributes. These latent codes, in turn, give rise to intermediate convolutional features within the StyleGan multilevel convolutional layers situated in the facial prior module. During the concluding phase of the generator, the intermediate convolutional features are seamlessly integrated with the initial spatial features through spatial modulation techniques, including translation scaling. This spatial modulation effectively aligns the realism and fidelity across distinct features. The output image from the generator, when assessed by the discriminator, computes the total loss function and gradient in conjunction with the original image. This computation encompasses adversarial loss, facial component loss, and identity preserving loss. Adversarial loss employs a logistic regression loss function to guarantee the generation of authentic textures. Facial component loss quantifies the discrepancy between the generated image and the original image, focusing on specific regions of interest. Identity preserving loss ensures that the features of the generated image remain minimally distant from those of the original image within the deep feature space.

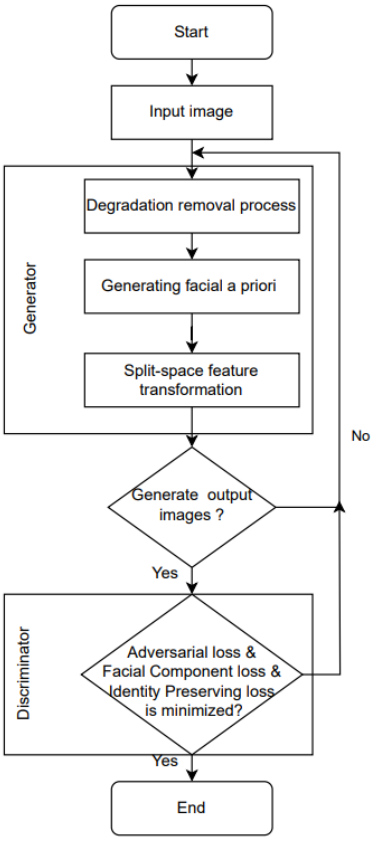
Start
Input image
Generator
Degradation removal process
Generating facial a priori
Split-space feature transformation
No
Generate output images ?
Yes
Discriminator
Adversarial loss & Facial Component loss & Identity Preserving loss is minimized?
Yes
End

Figure 12. Algorithm flowchart of GFP-GAN.

We propose multimodal audio-visual sensing using GFP-GAN, which achieves a better balance of realism and fidelity compared to conventional multimodal audio-visual sensing. In real-world environments, multimodal audio-visual sensing receives input data that may contain noise and blur. This often results in a visually realistic but low-fidelity output. The proposed multimodal audio-visual sensing architecture generates high-fidelity and realistic face images in realistic environments, as evident from the clear visibility of mouth movements, leading to an improved recognition accuracy.

3.3. AVSR with Wave2Lip-GAN

In this section, we introduce another multimodal audio-visual sensing architecture utilizing Wave2Lip-GAN [42]. While the previous section extensively described the AVSR architecture, this part will directly focus on how AVSR incorporates Wave2Lip-GAN.

Figure 13 illustrates the architecture of Wave2Lip-GAN for AVSR. Prior to entering the TM-CTC architecture, the audio and video sensing input undergoes processing through Wave2Lip-GAN. The generator network consists of two encoders and a decoder. The face encoder generates intermediate facial features for the video frames, while the audio encoder generates intermediate audio features from the audio signals. These generated video and audio features are combined and fed into the face decoder, which produces lip and audio synchronized output frames. The generated frames then undergo evaluation by the discriminator, to assess the quality of the lip video and synchronization. The outcome of this process is a new dataset to be trained in the TM-CTC.

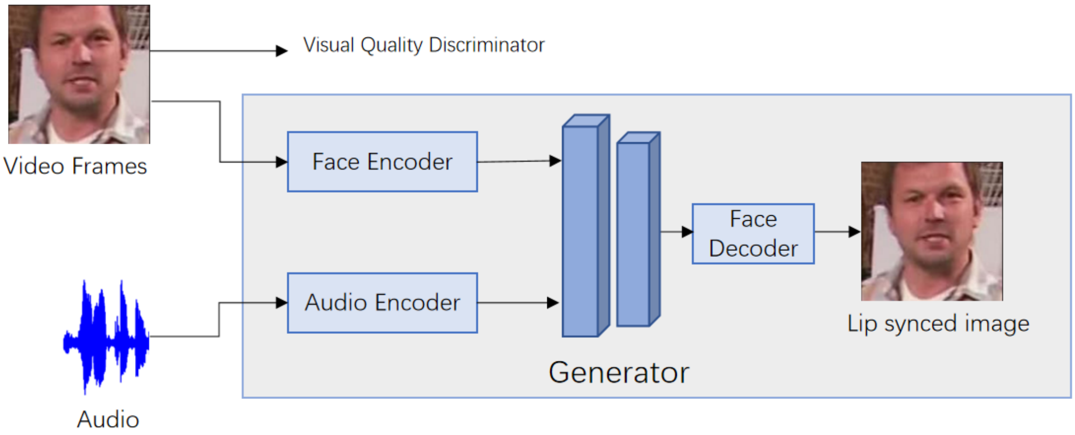

Figure 13. The proposed architecture uses Wave2Lip-GAN for AVSR.

We utilized the Wave2Lip-GAN model, which includes two discriminator modules. The first discriminator module assesses the synchronization quality between the generated lip image and the audio. The second discriminator module evaluates the quality of the lip image generated by the generator, comparing it with the synchronized lip image.

We propose multimodal audio-visual perception using Wave2Lip-GAN, which achieves a superior alignment between modalities compared to the average methods. Multimodal audio-visual sensing requires modal fusion, due to the independent nature of audio and video modalities. The challenge in fusing these modalities lies in achieving proper alignment. The architecture we propose for multimodal audio-visual sensing enables synchronization between the audio and video modalities, allowing for authentic alignment judgment and ultimately improving video quality.

4. Experiments

To validate the effectiveness of our architecture, we evaluated it on a multimodal audio-visual dataset. In this section, we first introduce the multimodal audio-visual dataset collected through IoT audio-visual sensors. Then, we elaborate on the implementation details of the AVSR architecture for multimodal generative adversarial networks for IoT. Finally, we analyze the classification results and the results of IoT energy efficiency.

4.1. Description of Datasets

This section provides a description of the datasets used for training and evaluation purposes. The Lip Reading Sentence 2 (LRS2) and Lip Reading Sentences 3 (LRS3) datasets were utilized for these tasks.

The LRS2 dataset [43] consists of a vast collection of audio and video data, totaling 224 h of content. It comprises 144,000 video sequences extracted from British Broadcasting

Corporation (BBC) recordings. In particular, there are 96,318 utterances for pretraining (195 h), 45,839 for training (28 h), 1082 for validation (0.6 h), and 1243 for testing (0.5 h). The dataset exhibits diversity in terms of various factors, including head pose, lighting conditions, video type, and the presence of multiple speakers. This diversity ensured a robust training and evaluation environment for the proposed multimodal audio-visual sensing architectures. In our study, LRS2 encompassed a diverse range of speech data collected from various scenarios, including news reports, cinematic dialogues, educational lectures, interviews, and more. These speech datasets exhibit notable variations, encompassing a wide spectrum of accents, speech tempos, and language styles characteristic of distinct speakers. Additionally, this dataset comprises multimodal data, encompassing not only speech but also relevant mouth movements, further enriching its informational content.

The LRS3 dataset [44] provides an extensive collection of audio and video data, amounting to over 438 h. It contains 151,819 video sequences extracted from TED and TEDx presentations. Specifically, there are 118,516 utterances in the pretraining set (408 h), 31,982 utterances in the training-validation set (30 h), and 1321 utterances in the test set (0.9 h). In addition to being used for training and evaluation, this dataset was also employed to train external language models. These language models were trained using a text corpus that incorporated subtitles, enabling them to enhance the language processing capabilities of the multimodal audio-visual sensing architectures. The LRS3 dataset captures video data from TED and TEDx events available on the YouTube channel. It encompasses a wide spectrum of individuals, not only actors in scripted films or theatrical productions. Additionally, the video footage exhibits reduced variability, resulting in a higher frequency of complete sentences accompanied by continuous facial trajectories.

Figure 14a,b showcase examples from the LRS2 and LRS3 datasets, respectively. These examples provide visual representations of the data present in the datasets, highlighting the variation in visual cues, lip movements, and speaker characteristics captured in the dataset samples. The LRS2 and LRS3 datasets are provided as mp4 files with a frame rate of 25 fps, encoded using the h264 codec. The audio data are presented in a format with a single channel, featuring a 16-bit bit depth and a sampling frequency of 16 kHz, while the corresponding text and the alignment boundaries of each word are included in the plain text files. The utilization of these datasets in the experiments ensured the evaluation of the proposed multimodal audio-visual sensing architectures on real-world, diverse, and challenging data.

(**a**)

(**b**)

Figure 14. Sample images from datasets. (**a**) Sample images from LRS2 [43]. (**b**) Sample images from LRS3 [44].

4.2. Experiment 1–GFP-GAN

In this subsection, we present the results obtained from the AVSR architecture using the GFP-GAN setup. As part of the experimental procedure, we performed pre-processing and data augmentation steps on the data. Specifically, we modularized the code for data preprocessing and data enhancement, to handle data from different IoT devices. In the preprocessing module, the data came in mp4 format, and we used the open-source tool FFmpeg (Fast Forward Moving Picture Experts Group) to extract the audio, while converting the video into a sequence of images with a size of 224 × 224 per frame. Then, a bounding box of 120 × 120 was used to crop the mouth ROIs. The cropped frames were further converted to gray-scale and normalized with respect to the overall mean and variance of the training set.

In the data augmentation module, we removed random frames and horizontal flipping with a probability of 0.5 after random cropping of size 112 × 112 of all frames of a given image sequence, in order to eliminate variations associated with facial rotation and scaling. Additionally, we introduced clutter to the audio data by adding background noise with a signal-to-noise ratio (SNR) of 5 dB and an audio stream with a probability of pn = 0.25. We generated babble noise samples by mixing a combination of 20 different audio samples from the LRS2 dataset. This was carried out to assess and enhance the model's ability to generalize in the presence of audio noise.

Following the random cropping of the 112 × 112 mouth image sequence, the data were fed into the GFP-GAN architecture. To preserve the original size of the image sequence, while improving clarity, we set the scaling factor to 1. This ensured that the image sequence was processed without any distortion in size.

To evaluate the performance of the AVSR architecture, we utilized word error rate (WER) as the evaluation metric. The WER is calculated using the following equation:

$$WER = \frac{S + D + I}{N} \tag{4}$$

where S, D, I, and N represent the number of substitutions, deletions, insertions, and words in the reference, respectively.

By employing WER as the evaluation metric, we were able to assess the accuracy and effectiveness of the AVSR architecture in transcribing spoken words from the input audio and visual data. The lower the WER, the higher the accuracy and alignment between the predicted transcription and the reference transcription.

In our experiments, we utilized the PyTorch library for implementation, leveraging the computational power of an NVIDIA V100 GPU with 40 GB of memory. The NVIDA V100 is a public version of the card manufactured by NVIDIA, and the device is sourced from the United States. During GPU training of the model, the minibatch size (default = 32) was reduced by half each time we encountered an out of memory error. We used an early stopping tactic and hyperparameter tuning to avoid the overfitting effect. Within each iterative cycle, our attention was directed towards scrutinizing the model's performance on the validation dataset, all while meticulously tracking the influence of the learning rate on the performance. Once the validation set WER had been flattened, training was forced to terminate. At the same time, we set an initial learning rate of 10^{-4}, which was reduced by a factor of 2 every time the validation error plateaued, down to a final learning rate of 10^{-6}. The network models were trained using the ADAM optimizer, which is a popular choice for deep learning tasks. Moreover, we undertook additional measures to enhance the robustness and overall generalization prowess of our model. To achieve this, we implemented a fusion of dropout and label smoothing techniques, employing a parameter value of p = 0.1. In the context of a dropout with p = 0.1, a stochastic process was introduced, whereby each neuron possessed a 10% probability of being randomly excluded during each training iteration. This strategic approach effectively mitigated the model complexity. Simultaneously, within the realm of label smoothing, the original single thermal-encoded labels underwent a transformation into a more diffuse distribution. This transformation served to instill a degree of uncertainty into the model's label predictions, thereby thwarting excessive confidence and, subsequently, combating overfitting.

During the model training process, we implemented a curriculum-based learning strategy. We initiated the training by exclusively using single-word examples and progressively increased the sequence length as the network continued to learn. These shorter sequences were originally part of longer sentences in the dataset. This approach had several noteworthy benefits: the convergence on the training set was significantly expedited, and the curriculum substantially mitigated overfitting. This improvement could be attributed to the natural way of introducing data into the training process. Additionally, we had the flexibility to fine-tune the model by adjusting the iterations at which we introduced curriculum learning. The number of words incorporated into each iteration of curriculum learning followed a sequential pattern: 1, 2, 3, 5, 7, 9, 13, 17, 21, 29, and 37, totaling 11 iterations in all.

4.3. Experiment 2—Wave2Lip-GAN

In this subsection, we first outline the experimental setup of the proposed architecture. Subsequently, we will present the results obtained from the AVSR system using the Wave2Lip-GAN setup.

In experiment 2, the data underwent a preprocessing step before being fed into the AVSR architecture. The audio and video inputs for multimodal audio-visual sensing were initially processed through the Wave2Lip-GAN block. To accurately detect the face region, the bounding box parameters were set with pads = (0 20 0 0), ensuring the inclusion of the entire face area, including the chin. The parameter nosmooth = true was utilized in the framework to prevent excessive smoothing during face detection.

After passing through the Wave2Lip-GAN block, the new data followed the same preprocessing steps as in experiment 1. This included adding noise, randomly cropping the image sequence to a size of 112 × 112, and horizontally flipping the images. To evaluate the performance of the AVSR system in experiment 2, the same evaluation metric, WER, was employed as in experiment 1. This facilitated a direct comparison between the two architectures.

4.4. Comparison Results for Classification and Discussion

In this subsection, we compare the two multimodal generative adversarial network AVSR architectures together with a multimodal AVSR architecture. Tables 2 and 3 present the experimental results, showcasing the performance in terms of word error rate (WER) on the LRS2 and LRS3 datasets, respectively. The experiments were conducted under two conditions: clean input (without noise), and added noise. The performance of the AVSR architecture and its individual components was compared in three scenarios: using GFP-GAN, using Wave2Lip-GAN, and using nothing. This comparison allowed us to evaluate the impact and effectiveness of incorporating the GFP-GAN for improving the performance of the AVSR system.

Table 2. Performance of AVSR on the LRS2 dataset.

AVSR Architecture		**Greedy Search**	**Beam Search (+LM)**
		Clean Input	
TM-CTC+ GFP-GAN	AV	10.20%	6.80%
TM-CTC+Wav2Lip GANs	AV	11.90%	8.40%
TM-CTC	AV	10.60%	7.00%
		Added Noise	
TM-CTC+GFP-GAN	AV	29.40%	22.40%
TM-CTC+Wav2Lip Gans	AV	35.70%	27.90%
TM-CTC	AV	30.30%	22.80%

Table 3. Performance of AVSR on the LRS3 dataset.

AVSR Architecture		Greedy Search	Beam Search (+LM)
		Clean Input	
TM-CTC+GFP-GAN	AV	11.60%	8.00%
TM-CTC+Wav2Lip GANs	AV	13.80%	12.60%
TM-CTC	AV	12.20%	10.80%
		Added Noise	
TM-CTC+GFP-GAN	AV	32.40%	25.50%
TM-CTC+Wav2Lip Gans	AV	39.20%	31.90%
TM-CTC	AV	34.70%	26.80%

In our experiments, we introduced noise into the audio input of both AVSR frameworks that utilized the GFP-GAN. The noise was generated by adding murmurs to the original audio. Recognition of multimodal audio-visual sensing in the presence of noise poses a challenging problem that needs to be addressed. In the AVSR framework, the addition of noise led to a decrease in word error rate (WER) by more than 20% compared to the AVSR performance without added noise. In addition, we can also clearly observe from the experimental results in Figures 15 and 16 that the results of the model with added noise were much higher than the results in clean environments, both on the LRS2 and LRS3 datasets. The results in the clean environment all remained below 15%, while the results in the noisy environment were all above 20%. Furthermore, the AVSR framework utilizing the GFP-GAN outperformed another framework in the presence of loud sounds in the environment.

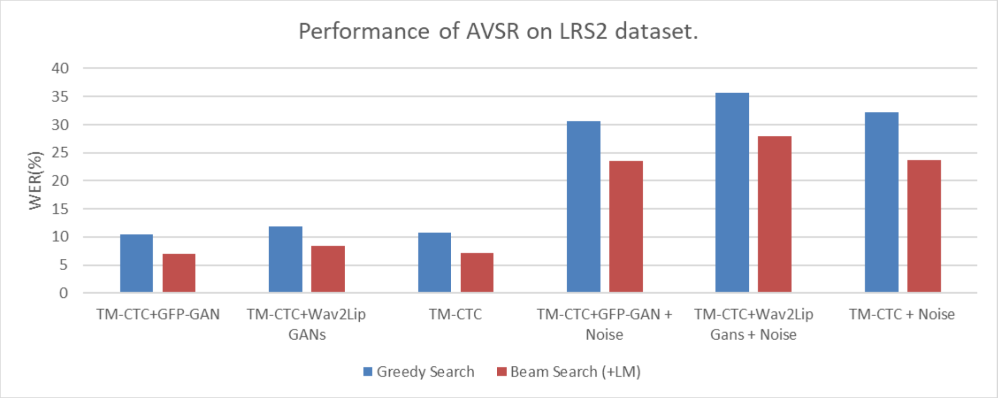

Figure 15. Bar chart of all experimental results on the LRS2 dataset.

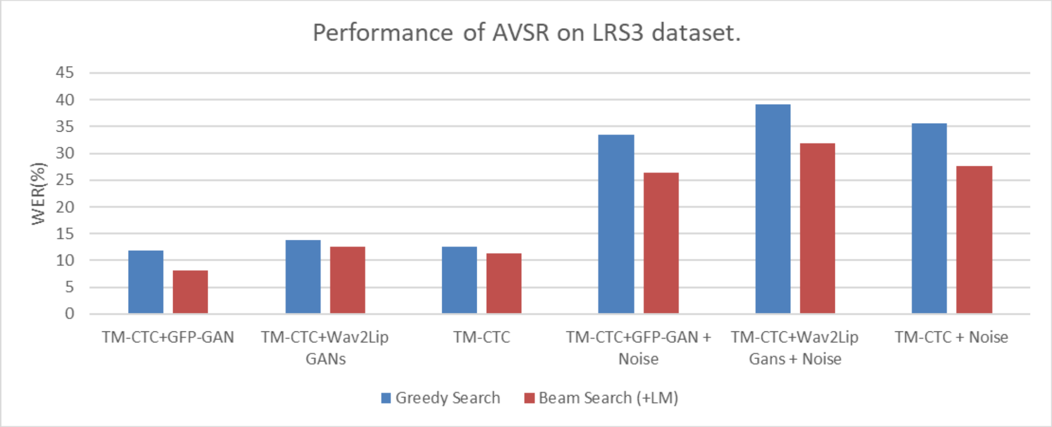

Figure 16. Bar chart of all experimental results on the LRS3 dataset.

The movement of the lips in the AVSR architecture provides valuable cues for speech recognition, particularly when the speech signal is heavily corrupted by noise. When the GFP-GAN was used in conjunction with the AVSR architecture, both architectures outperformed the AVSR architecture alone, in terms of performance. In the AVSR architecture with TM-CTC decoding, incorporating beam search and an external language model proved beneficial for improving performance. A notable characteristic observed in all three different AVSR architectures was that beam search yielded better results compared to greedy search. We also clearly observed that the results using beam search (red bars) were all lower than the results of greedy search (blue bars), as shown by the experimental results in Figures 15 and 16. This suggests that multimodal audio-visual sensing can achieve explicit linguistic consistency when integrated with external language models. Figure 17 shows some samples that reflect the AVSR results during our experiments. We can clearly observe that the texture appears to be significantly sharper in Figure 17a, performing well under the AVSR framework, compared to Figure 17b, performing poorly. In addition, using the GFP-GAN framework shown in Figure 17c clearly improved the facial texture clarity as well.

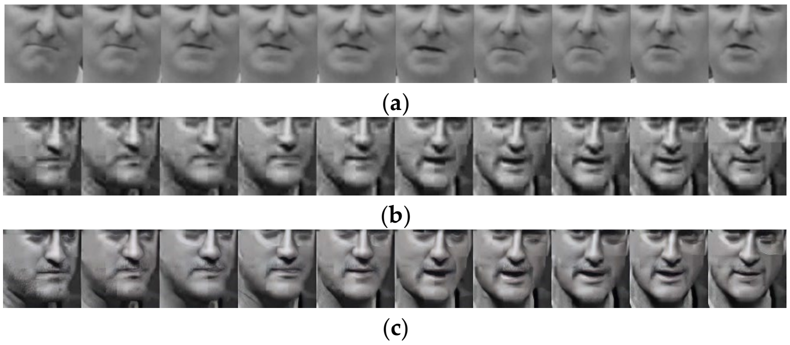
(**a**)
(**b**)
(**c**)

Figure 17. Example images after preprocessing. (**a**,**b**) show two different speakers in the same dataset and using the same preprocessing. (**b**,**c**) show the same speaker, the difference is (**c**) used GFP-GAN preprocessing. (**a**) Best examples of AVSR results after preprocessing from LRS2 [43]. (**b**) Worst examples of AVSR results after preprocessing from LRS2 [43]. (**c**) Worst examples of AVSR results after preprocessing and GFP-GAN from LRS2 [43].

Overall, the discussed findings highlight the effectiveness of the AVSR architectures, particularly when incorporating GFP-GAN, in addressing the challenges posed by noise in multimodal audio-visual sensing. The improvements in performance and linguistic consistency demonstrate the potential of these architectures for enhancing speech recognition in real-world environments.

4.5. Comparison Results for Energy Efficiency Generalizability and Discussion

In this subsection, we explore the energy efficiency of the proposed method. Tables 4 and 5 show the training time of the multimodal generative adversarial network AVSR architecture in the different cases. In IoT applications, assessing the energy efficiency of resources holds significant importance, especially considering hardware limitations. We gauged the resource energy efficiency by measuring the training time of the computational architecture model. By the time modal fusion was executed, the feature vector representation had expanded to 512 dimensions in our test architecture model. The required storage space varied depending on the number of video frames in the dataset. For instance, in the case of Figure 17a, our picture sequence necessitated 400 kilobytes of storage space, calculated as ((512 × 8)/1024) × 10. Compared to the video and audio space size of 160 KB that the IoT sensor collected initially, this reduced the storage space by 37%.

Table 4. Training efficiency of AVSR on the LRS2 dataset. The runtime means the time spent on the model training simulation, and this result is based on equipment of an NVIDIA V100 GPU with 40 GB of memory.

AVSR Architecture		**Greedy Search Run Time (h)**	**Beam Search (+LM) Run Time (h)**
		Clean Input	
TM-CTC+GFP-GAN	AV	72	90
TM-CTC+Wav2Lip GANs	AV	69	88
TM-CTC	AV	96	115
		Added Noise	
TM-CTC+GFP-GAN	AV	75	97
TM-CTC+Wav2Lip Gans	AV	72	92
TM-CTC	AV	100	122

Table 5. Training efficiency of AVSR on the LRS3 dataset.

AVSR Architecture		**Greedy Search Run Time (h)**	**Beam Search (+LM) Run Time (h)**
		Clean Input	
TM-CTC+GFP-GAN	AV	78	104
TM-CTC+Wav2Lip GANs	AV	80	98
TM-CTC	AV	100	118
		Added Noise	
TM-CTC+GFP-GAN	AV	79	109
TM-CTC+Wav2Lip Gans	AV	89	115
TM-CTC	AV	103	126

To ensure that the versatility of our AVSR architecture extended beyond the confines of the LRS2 and LRS3 datasets, we further diversified our validation process by randomly selecting datasets from various scenarios within Lip Reading in the Wild (LRW) [45]. LRW, as the largest audio-visual dataset, boasts an expansive collection of over 500 h of video clips, spanning a multitude of scenarios and contexts, closely resembling the audio-visual data source format prevalent in IoT devices. Our AVSR architecture possesses the capability of handling audio-visual datasets sourced from diverse origins. The preprocessing module played a pivotal role in this process, as it standardized audio-visual data into the requisite picture sequences and waveform files essential for seamless integration with the AVSR architecture, employing advanced FFMPEG technology. Table 6 clearly illustrates that our AVSR architecture remained adaptable and consistently demonstrated robust performance when applied to alternative datasets.

Table 6. Performance of AVSR on the LRW dataset.

AVSR Architecture		**Greedy Search**	**Beam Search (+LM)**
		Clean Input	
TM-CTC+GFP-GAN	AV	13.40%	10.50%
TM-CTC+Wav2Lip GANs	AV	17.30%	16.20%
TM-CTC	AV	15.30%	13.80%
		Added Noise	
TM-CTC+GFP-GAN	AV	34.80%	27.30%
TM-CTC+Wav2Lip Gans	AV	42.20%	34.50%
TM-CTC	AV	37.60%	29.20%

In the experiments on resource energy efficiency, both the GFP-GAN AVSR architecture and the Wave2Lip-GAN architecture showed a significant improvement in resource energy efficiency on training, with a maximum improvement of more than 20 h, which saved nearly one-fifth of the resource usage for the device. The experiments with both architectures emphasized the untapped potential of the multimodal generative adversarial network AVSR architecture for further applications for IoT devices and IoT energy efficiency. Furthermore, our AVSR architecture places a paramount emphasis on processing data from an array of IoT devices, underscoring its heightened potential for broader application and scalability across diverse IoT device ecosystems.

5. Audio-Visual Speech Recognition for IoT

In the realm of IoT applications, the significance of model accuracy is paramount. However, equally critical to us is the seamless integration of this model into IoT devices. In this section, we explore the technical aspects and considerations vital for the effective implementation of the multimodal generative adversarial network AVSR architecture into the IoT ecosystem. The first segment of this section discusses the privacy and security issues of IoT data sources. The latter part of this section focuses on the optimization and deployment strategies pertinent to the multimodal generative adversarial network AVSR architecture within the context of IoT.

5.1. Privacy and Security of Audio-Visual Data

In the context of real-world IoT scenarios, our foremost considerations revolve around the preservation of privacy and the security of audio-visual data sources. Notably, our data sources, LRS2 and LRS3, were meticulously curated from the BBC, necessitating a rigorous adherence to a privacy policy agreement with the BBC itself. This stringent agreement is in place to guarantee that sensitive data are exclusively accessible to duly authorized devices and users. As part of our security measures, we deployed stringent password protocols and integrated multi-factor authentication mechanisms, all working in concert to meticulously limit data access to authorized entities. Furthermore, it is imperative that data access be contingent upon the execution of a stringent privacy policy agreement. This agreement delineates crucial facets, including the data type, processing objectives, storage duration, and user authorization, with a notable emphasis on data security measures and user rights. Furthermore, it encompasses provisions for notifying and managing policy alterations and incorporates a dispute resolution mechanism, thus furnishing users with a transparent and dependable framework for data governance.

To fortify the security of audio-visual data storage, we have embraced the federated learning approach. Rather than centralizing data collection, our models can be downloaded and trained on distinct local devices. This approach serves as a robust guardian of user privacy, ensuring that data remain securely confined within the confines of the local device. Simultaneously, this decentralized data storage method possesses the capacity to effectively isolate different data sources, thus preempting any unintended mixing or leakage of data.

5.2. Extension and Application of AVSR

Optimizing and deploying our AVSR architecture is imperative for bolstering its scalability to IoT devices. We implemented lightweight optimizations to enhance the performance of our AVSR architecture. Initially, we introduced modularity to our AVSR architecture by segregating each task into distinct modules, facilitating diverse task scheduling. Simultaneously, a caching module was incorporated, to allow for direct storage of trained model parameters and weights, thus preventing unnecessary computational resource expenditure through repeated model runs. Throughout the model's inference process, we cached intermediate results for subsequent reuse. This proactive caching strategy effectively obviated the necessity for redundant retraining and furnished an uninterrupted approach for storing and retrieving model outputs in the presence of repetitive data inputs, thereby mitigating computational overheads linked to repetitive computations.

To broaden the applicability of our AVSR architecture, we encapsulated the model as an invocable service within docker containers. This enables neighboring IoT devices to locally retrieve the model through authenticated requests directed at the container's designated port. Given the diverse resource profiles of various IoT devices, we offer a versatile low-latency processing solution. Our multimodal AVSR architecture seamlessly accommodates IoT devices with video and audio data sources, facilitated by our preprocessing module's capacity to standardize these data sources into sequences of images and waveform files for model inference. Simultaneously, we adapt the batch size to match the hardware specifications of distinct IoT devices, leveraging batch processing to enhance inference speed. Furthermore, our AVSR model is available for local downloads, thus positioning model inference in closer proximity to the data source, effectively curtailing data transmission delays. Leveraging cached models for data reasoning results in a substantial boost in computational speed. For instance, processing 0.9 h of data requires a mere 13 s.

Leveraging extensive and diverse real-world datasets from LRS2 and LRS3, our AVSR architecture demonstrated its strength in solving real-world applications. Given both the LRS2 and LRS3 datasets are derived from authentic real-world settings, our AVSR architecture consistently demonstrated a remarkable classification performance across an array of real-world contexts, encompassing scenarios like subway environments, educational interviews, and news interviews. These outcomes underscore the significant potential of our AVSR architecture in real-world application scenarios. To further underscore the versatile applicability of our AVSR architecture across multiple scenarios, we additionally employed random selection of datasets from various scenarios within LRW for validation purposes. The commendable performance observed in these diverse scenarios reaffirmed the adaptability and efficacy of our AVSR architecture in tackling real-world IoT scenarios.

Beyond the scenarios within our dataset, the potential real-world applications of our AVSR architecture are manifold. In the realm of smart homes, audio-visual speech recognition empowers users to command home devices such as smart lighting, thermostats, and security systems through simple verbal directives. This not only enhances the quality of daily living but also fosters energy efficiency. Furthermore, in the healthcare sector, audio-visual speech recognition could revolutionize the interaction with medical devices, aiding healthcare professionals in recording and retrieving patient information, thereby elevating the efficiency and precision of medical services. Within the domain of intelligent transportation systems, audio-visual speech recognition technology is poised to elevate the driving experience, by enabling drivers to control vehicle functions, access navigation instructions, and acquire traffic information through voice commands, thereby enhancing road safety and driving efficiency. Moreover, audio-visual speech recognition plays a pivotal role in diverse sectors, including industrial automation, smart cities, and education. Its integration brings greater intelligence and interactivity to IoT devices, promising to transform and elevate the capabilities of these systems.

6. Conclusions

This paper focused on the application of a novel multimodal generative adversarial network AVSR architecture for artificial intelligence Internet of things (IoT) and its performance in AVSR classification accuracy. This research explored the use of traditional and GAN-based generative adversarial network (GAN) techniques to enhance the AVSR architecture. Experiments conducted on real datasets such as LRS2 and LRS3 demonstrated the effectiveness of the proposed AVSR architecture. The results validated that the GAN architecture was well suited for multimodal sensor inputs and improved on the performance of the AVSR framework. They also verified that our multimodal generative adversarial network AVSR architecture has efficient processing in terms of resource loss and energy efficiency. In our research, there were potential limitations. We focused on exploring and validating multimodal generative adversarial network AVSR architectures for artificially intelligent IoT. However, it is important to note that there is still considerable potential for further development in enhancing the robustness of multimodal generative adversarial networks and applying them to different IoT devices. Moreover, when considering the deployment of the AVSR architecture across various IoT devices, it becomes imperative to address critical concerns such as privacy security and the imperative need for low-latency processing of data sources. In the future, we aim to further explore the application and performance of the multimodal generative adversarial network AVSR architecture in AI IoT to meet the hardware constraints of IoT devices. This includes the scalability of the modal generation adversarial network AVSR architecture to different IoT devices and its applicability to different scenarios. This provides potential avenues to extend the research and applications in IoT to improve the generalizability of the multimodal generative adversarial network AVSR architecture on IoT hardware. Furthermore, the management of sensitive data in IoT devices constitutes a crucial concern for our prospective exploration of AVSR architectures for AI-driven IoT. Algorithms like data encryption storage present promising avenues for addressing this challenge.

Author Contributions: Conceptualization, K.P.S., Y.H. and L.M.A.; methodology, Y.H. and K.P.S.; resources, K.P.S.; data curation, Y.H., K.P.S. and L.M.A.; writing—original draft preparation, K.P.S., Y.H. and L.M.A.; writing—review and editing, K.P.S., Y.H. and L.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in refs [43–45].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Atzori, L.; Iera, A.; Morabito, G. The internet of things: A survey. *Comput. Netw.* **2010**, *54*, 2787–2805. [CrossRef]
2. Zhao, F.; Wang, W.; Chen, H.; Zhang, Q. Interference alignment and game-theoretic power allocation in MIMO heterogeneous sensor networks communications. *Signal Process.* **2016**, *126*, 173–179. [CrossRef]
3. Roberts, C.M. Radio frequency identification (RFID). *Comput. Secur.* **2006**, *25*, 18–26. [CrossRef]
4. Stergiou, C.; Psannis, K.E. Recent advances delivered by mobile cloud computing and internet of things for big data applications: A survey. *Int. J. Netw. Manag.* **2017**, *27*, e1930. [CrossRef]
5. Tiippana, K. What is the McGurk effect? *Front. Psychol.* **2014**, *5*, 725. [CrossRef] [PubMed]
6. Kinjo, T.; Funaki, K. On HMM speech recognition based on complex speech analysis. In Proceedings of the IECON 2006—32nd Annual Conference on IEEE Industrial Electronics, Paris, France, 6–10 November 2006; IEEE: Piscataway, NJ, USA, 2006; pp. 3477–3480.
7. Dupont, S.; Luettin, J. Audio-visual speech modeling for continuous speech recognition. *IEEE Trans. Multimed.* **2000**, *2*, 141–151. [CrossRef]
8. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *44*, 8717–8727. [CrossRef] [PubMed]
9. Zhang, X.; Cheng, F.; Wang, S. Spatio-temporal fusion based convolutional sequence learning for lip reading. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 713–722.
10. Li, W.; Wang, S.; Lei, M.; Siniscalchi, S.M.; Lee, C.H. Improving audio-visual speech recognition performance with cross-modal student-teacher training. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 6560–6564.
11. Mehrabani, M.; Bangalore, S.; Stern, B. Personalized speech recognition for Internet of Things. In Proceedings of the 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), Milan, Italy, 14–16 December 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 369–374.
12. Dabran, I.; Avny, T.; Singher, E.; Danan, H.B. Augmented reality speech recognition for the hearing impaired. In Proceedings of the 2017 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS), Tel-Aviv, Israel, 3–15 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
13. Ma, Z.; Liu, Y.; Liu, X.; Ma, J.; Li, F. Privacy-preserving outsourced speech recognition for smart IoT devices. *IEEE Internet Things J.* **2019**, *6*, 8406–8420. [CrossRef]
14. Bäckström, T. Speech coding, speech interfaces and IoT-opportunities and challenges. In Proceedings of the 2018 52nd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 28–31 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1931–1935.
15. Brock, A.; Donahue, J.; Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint* **2018**, arXiv:1809.11096.
16. Park, T.; Liu, M.; Wang, T.; Zhu, J. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

17. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Loy, C.C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
18. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with contextual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 5505–5514.
19. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved Techniques for Training Gans. In *Advances in Neural Information Processing Systems (NeurIPS)*; The MIT Press: Cambridge, MA, USA, 2016; pp. 2234–2242.
20. Xue, Y.; Xu, T.; Zhang, H.; Long, L.R.; Huang, X. Segan: Adversarial network with multi-scale L_1 loss for medical image segmentation. *Neuroinformatics* **2018**, *16*, 383–392. [CrossRef]
21. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
22. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growingof GANs for Improved Quality, Stability, and Variation. *arXiv preprint* **2017**, arXiv:1710.10196.
23. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
24. Wang, X.; Li, Y.; Zhang, H.; Shan, Y. Towards real-world blind face restoration with generative facial prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9168–9178.
25. Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Process. Mag.* **2012**, *29*, 141–142. [CrossRef]
26. Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features from Tiny Images*; MIT: Cambridge, MA, USA; NYU: New York, NY, USA, 2009.
27. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part I 13. Springer International Publishing: Cham, Switzerland, 2014; pp. 818–833.
28. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint* **2015**, arXiv:1511.06434.
29. Xu, L.; Ren, J.S.; Liu, C.; Jia, J. Deep convolutional neural network for image deconvolution. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 90–1798.
30. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint* **2014**, arXiv:1412.6806.
31. Santurkar, S.; Tsipras, D.; Ilyas, A.; Madry, A. How does batch normalization help optimization? In Proceedings of the 2018 Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Volume 31.
32. Li, Y.; Yuan, Y. Convergence analysis of two-layer neural networks with ReLU activation. In Proceedings of the 2017 Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
33. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep Laplacian pyramid networks for fast and accurate super-resolution. *arXiv preprint* **2017**, arXiv:1704.03915.
34. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
35. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. *arXiv preprint* **2019**, arXiv:1805.08318.
36. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
37. Zhu, J.Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward multimodal image-to-image translation. *arXiv preprint* **2017**, arXiv:1711.11586.
38. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint* **2018**, arXiv:1711.09020.
39. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint* **2017**, arXiv:1609.04802.
40. Zareapoor, M.; Celebi, M.E.; Yang, J. Diverse adversarial network for image super-resolution. *Signal Process. Image Commun.* **2019**, *74*, 191–200. [CrossRef]
41. Chung, J.; Zisserman, A. Lip reading in profile. In Proceedings of the Ritish Machine Vision Conference, London, UK, 4–7 September 2017; British Machine Vision Association and Society for Pattern Recognition: Durham, UK, 2017.
42. Prajwal, K.R.; Mukhopadhyay, R.; Namboodiri, V.P.; Jawahar, C.V. A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 484–492.
43. The Oxford-BBC Lip Reading Sentences 2 (LRS2) Dataset. Available online: https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html (accessed on 8 March 2023).

44. Lip Reading Sentences 3 (LRS3) Dataset. Available online: https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs3.html (accessed on 8 March 2023).
45. The Oxford-BBC Lip Reading in the Wild (LRW) Dataset. Available online: https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html (accessed on 8 July 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.