The Transactions of the Korea Information Processing Society
TKIPS Vol.14, No.3, pp.135~142
eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2025.14.3.135

Diffusion-based Audio-to-Visual Generation for High-Quality Bird Images

Adel Toleubekova† · Joo Yong Shim†† · XinYu Piao††† · Jong-Kook Kim††††

ABSTRACT

Accurately identifying bird species from their vocalizations and generating corresponding bird images is still a challenging task due to limited training data and environmental noise in audio data. To address this limitation, this paper introduces a diffusion-based audio-to-image generation approach that satisfies both the need to accurately identify bird sounds and generate bird images. The main idea is to use a conditional diffusion model to handle the complexities of bird audio data, such as pitch variations and environmental noise while establishing a robust connection between the auditory and visual domains. This enables the model to generate high-quality bird images based on the given bird audio input. Plus, the proposed approach is integrated with deep audio processing to enhance its capabilities by meticulously aligning audio features with visual information and learning to map intricate acoustic patterns to corresponding visual representations. Experimental results demonstrate the effectiveness of the proposed approach in generating better images for bird classes compared to previous methods.

Keywords : Audio-to-visual generation, Diffusion models, Image generation, Audio features, Multi-modal generation

확산 모델 기반 오디오-비주얼 고품질 새 이미지 생성

Adel Toleubekova† · 심 주 용†† · 박 흠 우††† · 김 종 국††††

요 약

새의 소리로부터 종을 정확히 식별하고 해당 새의 이미지를 생성하는 것은 제한된 훈련 데이터와 오디오 데이터의 환경적 잡음으로 인해 어려운 과제이다. 이를 해결하기 위해 본 논문은 새의 소리를 정확히 식별하고 해당하는 이미지를 생성하는 확산 모델 기반 오디오-이미지 생성 기법을 제안한다. 본 연구의 핵심 아이디어는 조건부 확산 모델을 활용하여 새 소리 데이터의 피치 변동과 환경 잡음과 같은 복잡성을 처리하고, 청각적 정보와 시각적 정보 간의 견고한 연결을 구축하여 주어진 오디오 입력에 대해 고품질의 새 이미지를 생성해 내는 것이다. 제안된 방법에서는 오디오 특징을 시각적 정보와 결합해 복잡한 음향 패턴에 대응되는 시각적 표현으로 연결되도록 학습하여, 오디오 처리 능력을 향상함으로써 이미지 생성 성능을 향상했다. 실험 결과로 제안된 접근 방식이 기존 방법에 비해 종별로 더 잘 구분되는 고품질의 이미지를 생성할 수 있음을 보여준다.

키워드 : 오디오-비주얼 생성, Diffusion 모델, 이미지 생성, 오디오 특징, 멀티모달 생성

1. Introduction

Multi-modal generation has gained significant attention in recent years. In particular, the recent advances in diffusion models [1], cross-modal generation tasks, such as image-to-text, and text-to-image generation, have shown remarkable improvement [2-4]. While text-guided synthesis in visual data has been widely studied and implemented, the progress in audio-to-image generation has fallen behind.

Audio data contains abundant information that could be used to generate meaningful visual representations. A particularly promising area is wildlife monitoring, where converting the sound data into images provides valuable visual data for observing and analyzing various species. Advancements in acoustic sensors has improved wildlife detection by capturing sound data, such as bird calls in dense forests where direct observation is challenging. Converting such audio data into visual representations of wildlife species could significantly enhance monitoring efforts by offering more information for accurate species identification.

※ 이 논문은 교육부가 지원하는 한국연구재단의 기초과학연구사업(5년)의 지원을 받아 수행되었음.
† 비 회 원 : 고려대학교 전기전자공학부 학부과정
†† 준 회 원 : 고려대학교 정보통신기술연구소 박사후연구원
††† 준 회 원 : 고려대학교 전기전자공학과 석박사통합과정
†††† 종신회원 : 고려대학교 전기전자공학부 교수
Manuscript Received : January 8, 2025
Accepted : February 17, 2025
* Corresponding Author : Jong-Kook Kim(jongkook@korea.ac.kr)

Generative models that translate one data modality (*i.e.*, sound) into another (*i.e.*, images) assist in completing recognition tasks for the wildlife species.

Despite the advances in wildlife observation, however, it still remains challenges to effectively and accurately identify species in remote or dense habitats. Deep learning has emerged as a tool that offers an automated, efficient, and accurate method to identify species even in challenging environments. Traditional methods, such as sound files manual analysis, are time-consuming and often error-prone [5]. Plus, image-based recognition models depend on visual inputs, which may not be available in certain environments due to varying climatic and habitat conditions, thereby limiting monitoring capabilities. Another significant hurdle is the limited training data for diverse bird species, leading to inaccuracies in model outputs and reduced quality. Moreover, current deep learning models are challenged by environmental noise, limited diverse data, and overlapping natural sounds, which highlights the need for more effective methods to translate audio data into visual representations.

To overcome these problems, this work closely studies the conditional diffusion model [2] for handling complexities such as pitch variations and environmental noise in bird sound data and generating high-quality bird images. Based on our observations, this conditional diffusion model has the potential to produce detailed and high-quality images compared to other models such as GANs [6] and autoencoders [7]. This is because conditional diffusion models are designed to handle noisy and can capture fine-grained features and subtle details in complex sound data through an iterative denoising process. Based on a thorough review of the literature, importantly, previous research on bird species image generation has not explored the application of diffusion models in the audio-to-image domain. This lack of research provides an opportunity to use diffusion models that condition audio features.

This paper proposes the diffusion-based audio-to- image generation approach to address limitations in current wildlife monitoring techniques. The proposed approach first converts raw audio data into visual features (*i.e.*, mel-spectrograms) via feature extraction on the sound encoder. Then, the conditional diffusion model refines these features and generates the final bird images. In this paper, the conditional diffusion model introduced reveals the intrinsic connections between audio features and visual representations within wildlife monitoring. The proposed method enhances the audio-visual research framework and demonstrates its effectiveness by improving image generation metrics. The contributions of this paper can be summarized as follows:

- This paper proposes a novel approach that applies diffusion models for the first time to audio-to-image generation for bird species identification.
- The proposed approach also reveals the intrinsic connections between audio features and visual representations within wildlife monitoring.
- Experimental results show that the proposed approach achieves 98.4 FID score, demonstrating improved performance over the previous work.

This paper is organized as follows. Related work is described in Section 2. Section 3 introduces the proposed approach in more detail. Then, Section 4 shows the experiment results for the proposed approach compared to the existing methods. Finally, Section 5 concludes the paper.

2. Related Work

In cross-modal generation tasks, most of the prominent works focus on text-to-image generation tasks [4, 8]. A well-known example is Imagen [9], which generates images with fine-grained details and realistic textures by hierarchical sampling using diffusion models. Imagen uses a large-scale T5 model for text encoding to create a representation of the prompt from the Conceptual Captions dataset [10]. This dataset contains image URLs and caption pairs designed for the training and evaluation of machine-learned image captioning systems. Captions after being encoded are then used by an image generator, which begins with adding a Gaussian noise [11] and progressively refines it to create a small image depicting the scene described in the caption. Additionally, Imagen incorporates Classifier-Free Guidance [12], which reduces the dependency on noisy or less informative data during generation. Imagen achieved a new state-of-the-art in text-to-mage generation, specifically performing better in terms of photo-realism, detail and language understanding [13].

With significant advancements in deep learning, the audio-to-visual generation is also gaining traction and showing remarkable growth. By using deep neural networks (DNNs), models capture the spatial and semantic information embedded in sound to construct detailed visual scenes. Kim et al. [14] introduced Sound2Scene, which employs generative adversarial networks (GANs) to generate

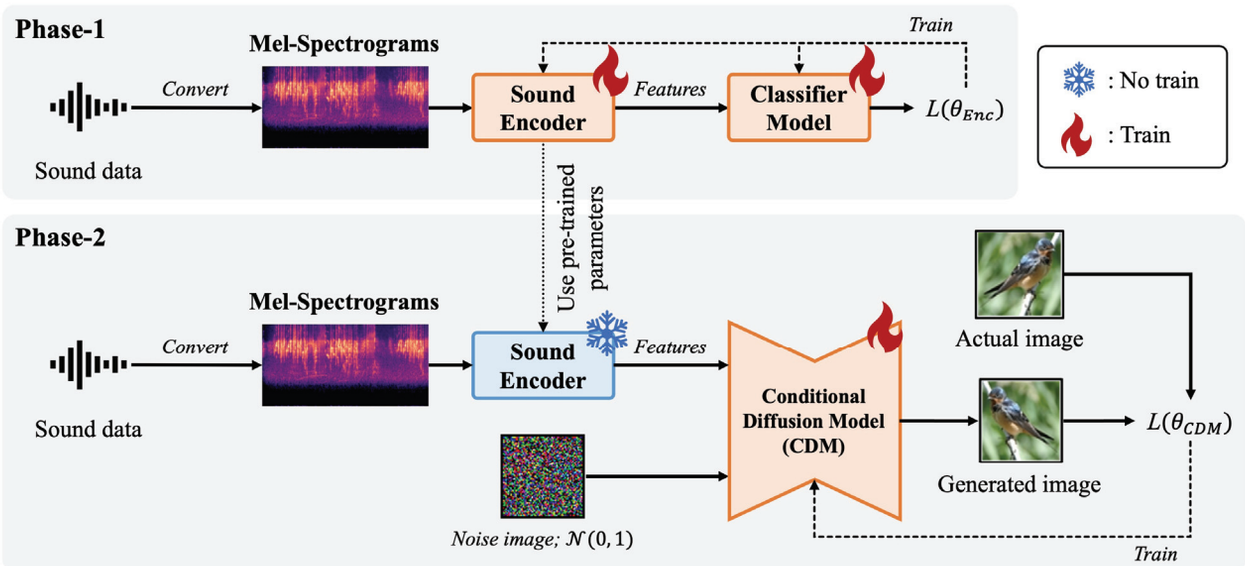

Fig. 1. Overall process of the proposed approach.

diverse images by first translating sounds into visual embeddings and then aligning them with visual space to finally synthesize images. Another important feature of Sound2Scene is its controllability, where visual representations change by applying simple manipulations on inputs in the waveform space or learned latent space. Shim et al. [6] introduced an audio-to-visual cross-modal generation using ACGAN model designed to generate realistic bird images based on the corresponding audio recordings. This approach uses a streamlined sound encoder that simplifies the sound feature extraction process compared to other models, while still delivering powerful performance results.

However, previous works showed poor performance in generating high-quality images of birds from audio recordings, making it difficult to accurately identify birds as shown in Fig. 3. In terms of the wildlife monitoring, audio-to-visual generation is also important for detecting habitat changes or patterns by quickly generating images of birds in low visibility environments like jungles. Thus, this paper aims to enhance the previous work [6] by exploring the capabilities of diffusion models in-depth and achieving significant improvements in generating high-fidelity bird images.

3. Proposed Method

3.1 Overview

This paper presents a diffusion-based audio-to-visual generation framework. The overall process is shown in Fig. 1. The proposed framework consists of two phases: (i) Sound Feature Extraction (Phase-1) and (ii) Image Generation (Phase-2). Phase 1 focuses on extracting audio features, where bird sounds are first converted into Mel-Spectrograms and then processed by a Sound Encoder to extract relevant features. The Sound Encoder is trained using additional classifier layers to classify sounds based on class labels. In Phase 2, the final target outputs are generated using a Conditional Diffusion Model (CDM). The audio features extracted from Phase-1 are used as conditional inputs to generate corresponding bird images. In the forward diffusion process, Gaussian noise [11] is gradually added to the input images, and in the denoising process, starting from the noisy

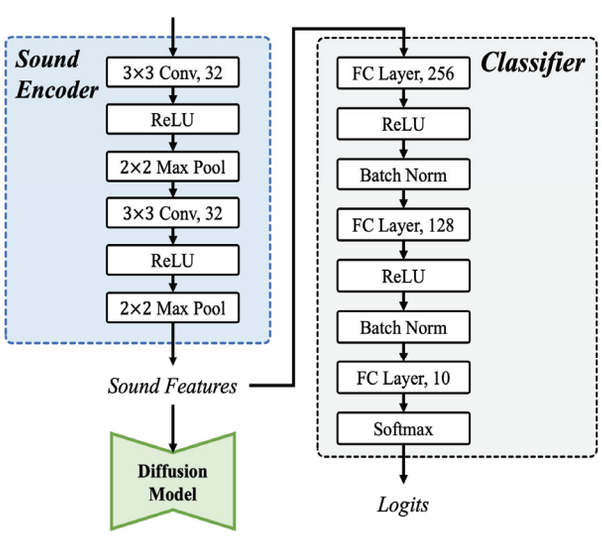

Fig. 2. Sound feature extraction model.

version of the image, the U-Net denoises it step by step, guided by the audio features. This entire process is designed to produce high-quality images that accurately reflect the characteristics of the audio input.

3.2 Sound Feature Extraction

The classifier-based model is implemented and trained to meaningfully extract sound features. The detailed architecture is depicted in Fig. 2. A classifier-based model uses a convolutional neural network (CNN) to extract features from the spectrograms. The extracted features are then input to a diffusion model for further processing. The CNN used comprises convolutional layer, ReLU activation functions, max-pooling layers, and batch normalization layers. The final layer of the CNN is a fully connected layer with a softmax activation function, which outputs the predicted class probabilities.

Before the whole process, input raw sound files are first preprocessed and converted to mel-spectrograms. Mel-spectrograms are often preferred over simple waveforms for training audio features, since they provide a two-dimensional representation of sound (time and frequency), which is richer in information compared to the one-dimensional waveform (time only). Using just raw materials is not preferred as it makes it hard to process data, thus raw files are transformed to mel-spectrograms. In detail, the audio files are converted to their corresponding mel-spectrogram at 22050 Hz, with a window length of 2048, a hop length of 512, and 128 mel filters. Mel-spectrograms then proceed to a classifier- based feature extraction model where it extracts the features received from the last convolutional layer [15]. The pooled features are formatted into a single vector of embedding values.

After receiving the visual representation of the audio files, the feature extraction model is trained further on those new representations. The model has 3 convolution layers, each followed by a max pooling with Relu activations. After each constitutional layer, 2x2 max-pooling layers are performed. The above-mentioned constitutional layers are flattened and thus fully connected layers of sizes 256, 128, and 10 are added to the end of the network. Adding other 3 fully connected layers makes this model to be called "classifier-based" as in Fig. 2. To prevent overwriting, there were applied batch normalization and dropouts with a value of 0.5. Features are extracted from the last convolution layer of the classifier model. The classification results obtained on different datasets are shown in the Results and Analysis part of this work.

3.3 Image Generation

The Conditional Diffusion Model (CDM) is used in this work for image generation. Unlike previous image generation approaches, where diffusion models are usually conditioned on visual features such as image embeddings, this method uses sound features extracted from audio data to guide the diffusion process. By using audio instead of visual features, the proposed approach demonstrates a novel application of diffusion models, bridging the gap between audio and visual modalities for cross-domain generation tasks. This method enables the generation of images that semantically correspond to input sounds, offering a unique perspective on conditional image synthesis.

Specifically, this proposed model uses frozen encoded data received from a classifier-based feature extraction subnetwork and the pre-loaded images from the AVC-B dataset [6]. Audio features after being concatenated are paired with the corresponding images and fed into the CDM. The generation model is trained using, (1) the Gaussian Diffusion model, which defines the added Gaussian noise into the input files, (2) the Denoising model, which is trained to predict noise induced in the diffusion process and generate images using the U-Net architecture.

1) Diffusion Process

Gaussian diffusion model is used to define diffusion process operations. Taking only one argument - timesteps - creates the variance schedule. The variance schedule identifies the variance of Gaussian noise which is attached to the visual data. This schedule is before introduced and implemented in DDPM work [1]. Apart from that, constants and buffers were also used in the DDPM work. The q-sample method then performs noising of the images to a given timestep in the diffusion process, while Q-posterior and predict-start-from-noise functions calculate the parameters that are used for the denoising, that is, the reverse diffusion process.

2) Denoising Process

Denoising model is the next part of the model to be completed after adding the noise to the image. The classifier-free guidance (CFG) model [12] is employed for the CDM to improve the fidelity and diversity of the generated images by effectively amplifying the influence of the conditioning input. Unlike traditional approaches requiring

an external classifier, CFG integrates the conditioning directly into the noise estimation process. The use of conditional U-Net model [16] makes a conditional generation possible. As the U-Net is compiled for all timesteps, it needs to have time conditioning to know where and what amount of noise should be removed. Time conditioning generates a new tensor t, which has timestep conditioning information. The proposed model is also conditioned on the audio features received from the mel-spectrograms using the sound feature extraction model. The diffusion model operates by gradually transforming a distribution of random Gaussian noise into a coherent image over several timesteps, guided by the learned data distribution and audio features.

Given the noise z_t at time step t, the noise prediction conditioned on the sound features s (extracted from the sound encoder) is denoted as $\epsilon_\theta(z_t, s)$, while the unconditional prediction is $\epsilon_\theta(z_t)$. The guided noise estimate is calculated as:

$$\tilde{\epsilon_\theta} = (1-w)\epsilon_\theta(z_t) + w\epsilon_\theta(z_t, s) \quad (1)$$

where w is the guidance scale, a hyperparameter controlling the trade-off between fidelity and diversity. A higher w strengthens the influence of the conditioning sound input s, ensuring the generated image aligns more closely with the semantic content of the audio, while lower w increases diversity by reducing the conditioning impact. This approach simplifies the architecture by eliminating the need for an external classifier and enhances the semantic consistency of the generated images with the input sound, striking a balance between diversity and precision.

4. Result

4.1 Experimental Setup

As described in Section III, the proposed model is trained in two training phases. In Phase-1, the sound encoder is trained to extract meaningful features using 3,740 preprocessed audio mel-spectrograms with 30 epochs and 40 mini-batch sizes. Plus, the sound encoder is trained using cross-entropy (CE) loss and Adam optimizer [17] with 0.001 learning rate. The sound feature for each audio input is extracted in 12×12×128 shape from the final layer of the model. These extracted audio features are then paired with corresponding 64 × 64 resolution bird images from the dataset. In Phase-2, the conditional diffusion model is trained to generate images using the audio features obtained in Phase-1 as conditional input. The noise process in Phase-2 involved adding random Gaussian noise to the input images. Audio feature dimensions were set to 512, and the noising was applied over 1,000 time steps. The whole model was trained with 800 epochs and 512 batch sizes.

For evaluation, this paper uses the AVC-B dataset introduced in previous work [6]. This dataset consists of images and audio recordings of 10 different bird species, such as Barn Swallow, Blue Jay, Common Tern, Common Yellow throat, Hooded Oriole, Mallard, Red-winged Blackbird, Sayornis, Song Sparrow, Tropical King bird. The AVC-B dataset has 3,740 audio files and 3,600 image files with 64 × 64 image size. To address the imbalanced number of data between audio and image files, the images were repeated to match the audio samples circularly. This repetitive use ensured that each audio sample was paired with an image, effectively tackling the mismatch between the two data modalities. In addition, by aligning the dataset sizes, consistent processing across the entire dataset was achieved, facilitating more accurate outcomes. The experiments are tested on the computer system consisting of the NVIDIA A100 GPU with 80GB memory. The proposed model is implemented by using PyTorch 2.3.0 version, CUDA 12.1 version, and MinImagen [18].

4.2 Evaluation

1) Quantitative Evaluation

Table 1 shows the quantitative evaluation results for the proposed approach and four alternative methods referenced in the prior study [6]. The Frechet Inception Distance (FID) [13] score are used as a metric to assess the performance of the generative models. This score measures the distance between the pixel distributions of generated and real images, where lower values indicate superior performance. Note that, the FID scores in Table 1 are from the previous work [6].

The results in Table 1 indicate that the proposed model achieved the lowest FID score. Specifically, while the A2V method [6] achieved an FID score of 203.86, the proposed model demonstrated a significantly better performance by achieving 98.43. Additionally, the proposed model surpassed the LMS method, which achieved the next best FID score of

Table 1. Experimental results

Methods	Baseline	Label	LMS	A2V	Proposed
FID (↓)	0.00	254.10	166.22	203.86	**98.43**

166.22 among the compared methods. In addition to FID performance, the proposed classifier used for the sound extraction showed better accuracy in bird species classification, demonstrating 74.7% accuracy, compared to 72.5% in the prior work. Interestingly, despite utilizing fewer layers than the previous method[6], the proposed sound encoder showed improved efficiency and accuracy in feature extraction.

This result presents that diffusion-based models yield better results in image generation compared to GAN-based model for the same task. The diffusion-based model generates images that more closely resemble actual bird images in the dataset, reflecting improved quality and realism in the audio-to-image generation process. There are several factors that contribute to the performance improvements of diffusion models over GANs [2]. Firstly, diffusion models mitigate issues related to mode collapse, which is a common problem in GANs where the model fails to capture the diversity of the data distribution. These diffusion models benefit from a more stable training process because they do not involve adversarial training, which is sensitive to hyper-parameter tuning and prone to instability. Moreover, the iterative nature of the diffusion model allows for more refined data generation, leading to progressively improved quality of generated outputs. Collectively, these factors and the above-mentioned quantitative results explain why diffusion models are better suited to this task than GANs.

2) Qualitative Evaluation

The Abstract should concisely state what was done, how it was done, principal results, and their significance. It should

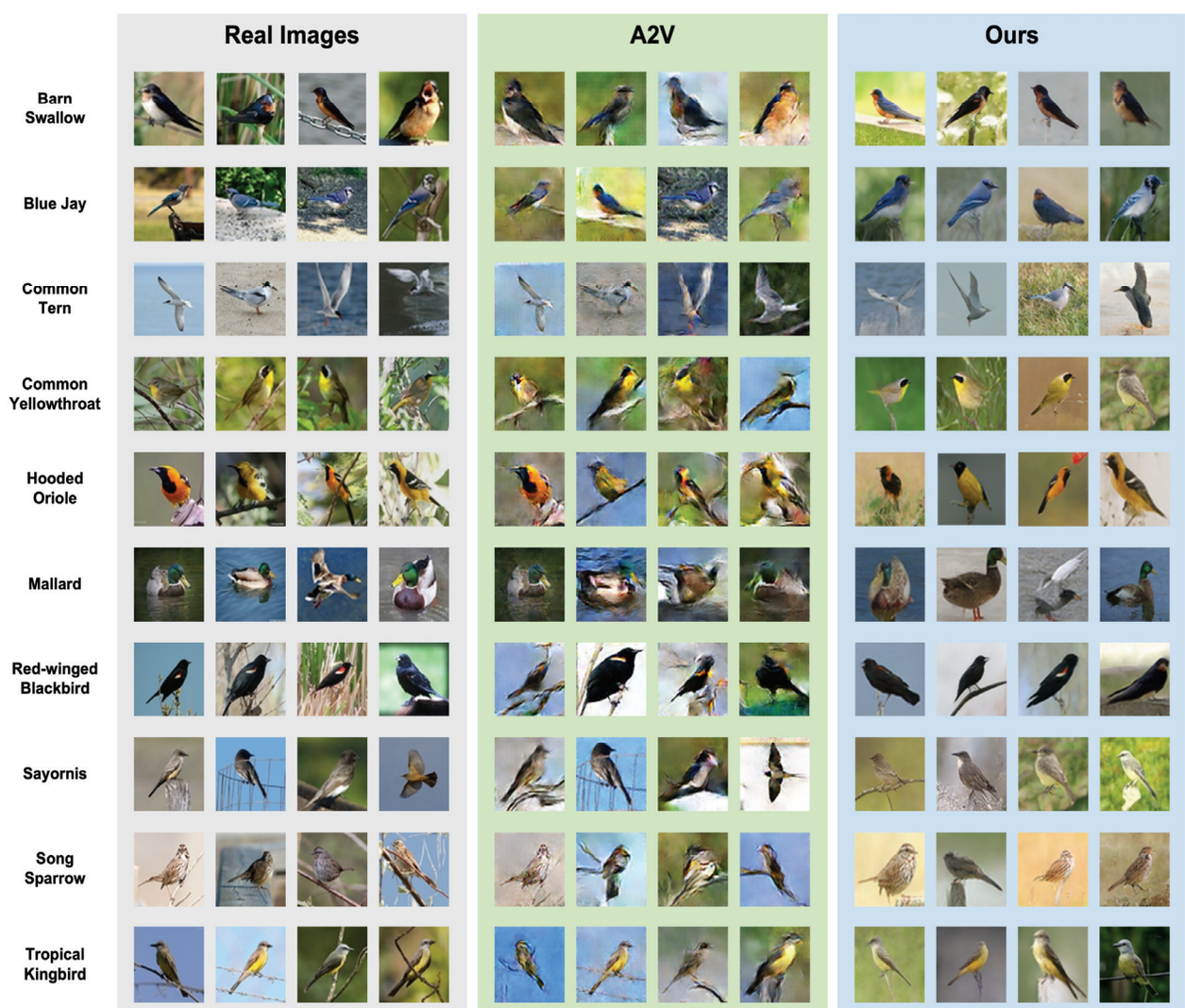

Fig. 3. Generated images.

be less than 300 words for all forms of publication. The abstract should be written as one paragraph and should not contain tabular material or numbered references. At the end of abstract, keywords should be given in 3 to 5 words or phrases. For qualitative evaluation, the generated images for 10 different bird species are presented in Fig. 3. The most representative images generated by the proposed approach compared to those produced using using the A2V model[6], which employed a simple sound encoder and ACGAN.

As shown in Fig. 3, bird species such as the Mallard, Sayornis, and Song Sparrow, which performed well in the feature extraction model, produced clearer and more detailed images due to their prominent and distinctive features. For instance, species like the Blue Jay, Common Tern, and Mallard were easily distinguishable thanks to their highly recognizable characteristics.

However, some bird species displayed overlapping features, such as similar colors or body shapes, which could lead to confusion. For example, the Barn Swallow and Red-winged Blackbird share similar coloring, with the red spot on the Blackbird being a key distinguishing feature. This red spot may appear orange from certain angles, occasionally resembling the Barn Swallow. Similarly, species pairs like the Common Yellowthroat and Hooded Oriole, as well as the Sayornis and Song Sparrow, share notable similarities. The Yellowthroat is identified by its yellow patch under the mouth, whereas the Oriole features a distinct colored hood. The Song Sparrow is characterized by its feather pattern, which the Sayornis lacks.

Overall, the proposed model consistently produced higher-quality images for 10 distinct bird species compared to the A2V model. These qualitative results further confirm the superiority of the proposed approach in the audio-to-visual generation task, demonstrating enhanced clarity and detail in the generated images.

5. Conclusion

This paper introduces a novel approach to the audio-to-visual generation task by using a conditional diffusion model. The proposed method has two main parts: (i) Sound Feature Extraction and (ii) Image Generation. By integrating these two parts, the proposed method effectively extracts meaningful features from audio files and uses to generate realistic visual representations. The proposed approach achieved an FID metric of 98.426, significantly outperforming previous methods, in generating high-quality images. Notably, the proposed model produced clearer and more detailed images for 10 distinct bird species. This work demonstrates the advantages of diffusion models over GANs, especially in generating greater realism and quality outputs. This study introduces the application of diffusion models to the audio-to-visual generation task, providing a practical and scalable solution for future developments.

References

[1] S. Y. Hea and E. G. Kim, "Design and implementation of the differential contents organization system based on each learner's level," *The KIPS Transactions: Part A*, Vol.18, No.6, pp.19-31, 2011.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, Vol.33, pp.6840-6851, 2020.

[3] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in Neural Information Processing Systems*, Vol.34, pp.8780-8794, 2021.

[4] X. Xu, Z. Wang, G. Zhang, K. Wang, and H. Shi, "Versatile diffusion: Text, images and variations all in one diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.7754-7765, 2023.

[5] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, Vol.35, pp.36479-36494, 2022.

[6] F. Rong, "Audio classification method based on machine learning," in *2016 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, IEEE, pp.81-84, 2016.

[7] J. Y. Shim, J. Kim, and J.-K. Kim, "Audio-to-visual cross-modal generation of birds," *IEEE Access*, Vol.11, pp.27719-27729, 2023.

[8] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," in *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, pp.353-374, 2023.

[9] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.5729-5739, 2023.

[10] R. Gandikota and N. Brown, "Pro-DDPM: Progressive growing of variable denoising diffusion probabilistic models for faster convergence," in *Proceedings of the British Machine Vision Conference (BMVC)*, p.121, 2022.

[11] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image

alt-text dataset for automatic image captioning," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2018.
[12] F. Russo, "A method for estimation and filtering of Gaussian noise in images," *IEEE Transactions on Instrumentation and Measurement*, Vol.52, No.4, pp.1148-1154, 2003.
[13] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint*, arXiv:2207.12598, 2022.
[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in Neural Information Processing Systems*, Vol.30, 2017.
[15] K. Sung-Bin, A. Senocak, H. Ha, A. Owens, and T.-H. Oh, "Sound to visual scene generation by audio-to-visual latent alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6430-6440, June 2023.
[16] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," *arXiv preprint*, arXiv:1706.09559, 2017.
[17] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, Vol.9, pp.82031-82057, 2021.
[18] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint*, arXiv:1412.6980, 2014.
[19] AssemblyAI, MinImagen: A minimal implementation of the Imagen text-to-image model [Internet], https://github.com/AssemblyAICommunity/MinImagen.

Adel Toleubekova
https://orcid.org/0009-0004-7999-5107
e-mail : aaaddellle@korea.ac.kr
She has been undergraduate scholar at the School of Electronic Engineering, Korea University, Seoul, Republic of Korea, from March 2021, and currently pursuing final year of study. Her current research interests include multi-modal generation, image generative modals.

Joo Yong Shim
https://orcid.org/0000-0003-2828-9232
e-mail : shimjoo@korea.ac.kr
She has been a postdoctoral scholar at the Department of Electrical and Computer Engineering, Korea University, Seoul, Republic of Korea, since March 2024, where she received her Ph.D. in electrical and computer engineering, in February 2024. She also received her B.S. in electrical engineering from Korea University, Seoul, Republic of Korea, in February 2019.
Her current research interests include image generation, cross-modal generative models, and their applications to mobility systems.

XinYu Piao
https://orcid.org/0009-0000-7502-4080
e-mail : xypiao97@korea.ac.kr
He received the B.S. degree from the School of Electrical Engineering, Korea University, in 2020 and is currently pursuing the Ph.D. degree from the Department of Electrical and Computer Engineering, Korea University, Seoul, Republic of Korea. His research interests include systems for artificial intelligence (AI), algorithms for deep learning, hardware resource management, and cloud computing.

Jong-Kook Kim
https://orcid.org/0000-0003-1828-7807
e-mail : jongkook@korea.ac.kr
He (Senior Member, IEEE and ACM) received the B.S. degree from the Department of Electronic Engineering, Korea University, Seoul, South Korea, in 1998, and the M.S. and Ph.D. degrees from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, USA, in 2000 and 2004, respectively. He was at the Samsung SDS's IT Research and Development Center, from 2005 to 2007. He is currently a Professor with the School of Electrical Engineering, Korea University, where he joined in 2007. His research interests include heterogeneous distributed computing, energy-aware computing, resource management, evolutionary heuristics, distributed mobile computing, artificial neural networks, efficient deep learning, systems for AI and distributed robot systems.