applied sciences

MDPI

Article

Sound-to-Imagination: An Exploratory Study on Cross-Modal Translation Using Diverse Audiovisual Data

Leonardo A. Fanzeres * and Climent Nadeu

Signal Theory and Communications Department, Polytechnic University of Catalonia (UPC), C. Jordi Girona 1–3, 08034 Barcelona, Spain; climent.nadeu@upc.edu
* Correspondence: leonardo.areias@upc.edu

Abstract: The motivation of our research is to explore the possibilities of automatic sound-to-image (S2I) translation for enabling a human receiver to visually infer occurrences of sound-related events. We expect the computer to 'imagine' scenes from captured sounds, generating original images that depict the sound-emitting sources. Previous studies on similar topics opted for simplified approaches using data with low content diversity and/or supervision/self-supervision for training. In contrast, our approach involves performing S2I translation using thousands of distinct and unknown scenes, using sound class annotations solely for data preparation, just enough to ensure aural–visual semantic coherence. To model the translator, we employ an audio encoder and a conditional generative adversarial network (GAN) with a deep densely connected generator. Furthermore, we present a solution using informativity classifiers for quantitatively evaluating the generated images. This allows us to analyze the influence of network-bottleneck variation on the translation process, highlighting a potential trade-off between informativity and pixel space convergence. Despite the complexity of the specified S2I translation task, we were able to generalize the model enough to obtain more than 14%, on average, of interpretable and semantically coherent images translated from unknown sounds.

Keywords: computational imagination; cross-modal learning; deep audiovisual learning; generative adversarial networks (GANs); information bottleneck; sound-to-image translation

Citation: Fanzeres, L.A.; Nadeu, C. Sound-to-Imagination: An Exploratory Study on Cross-Modal Translation Using Diverse Audiovisual Data. *Appl. Sci.* **2023**, *13*, 10833. https://doi.org/10.3390/app131910833

Academic Editors: Min Yang, Hao Liu, Shanxiong Chen and Yinong Chen

Received: 30 June 2023
Revised: 18 September 2023
Accepted: 21 September 2023
Published: 29 September 2023

1. Introduction

In the last few decades, acoustic event detection (AED) has evolved from adopting techniques initially developed for automatic speech recognition [1] to the use of deep learning (DL) models [2]. However, most current AED systems still rely on classification processes, and from a human receiver's standpoint, the model output has not changed much. In such systems, inference results are limited to discrete labels that represent sound concepts. A class-based output might suit an automatic audio monitoring system within a restricted acoustic context. Nonetheless, considering the intricate nature of environmental sounds, which encompass a vast and diverse spectrum of concepts, such output can easily mislead the human receiver, inducing a poor interpretation of the actual sound scene. In a previous study [3], deaf participants tested a mobile sound-recognition system and expressed their preference for images rather than text to represent sounds in the application. In the present work, as an alternative to sound classification, we explore the possibilities of automatic sound-to-image (S2I) translation for visually conveying the occurrence of acoustic events. We propose a system (Figure 1) that, given an audio input, is able to 'imagine' the scene with the sound-emitting source, generating an original image based on knowledge acquired through audiovisual learning. Furthermore, the system is expected to generate images that are interpretable and semantically coherent with the corresponding acoustic event of the captured audio. Throughout this text, we will occasionally refer to these images as informative. Here, the use of this term is aligned with the concept of 'informativity' in the context of text translation, as described by Neubert and Shreve [4]:

"Informativity in the translation process is a measure of the information a translation provides to an L2 reader about L1 events, states, processes, objects, individuals, places and institutions. The original information source was an L1 text intended for L1 audience. Translation opens an information channel between senders and receivers who could not normally inform one another about their respective states of affairs." L1 and L2 mean, respectively, source language and target language. Analogously, within the context of S2I translation, we define 'informativity' as the measure of information transmitted from the aural to the visual modality about sound-related events and sound-emitting sources, as well as any picturable elements in the surroundings that can be inferred from the sound. This might include landscapes, environments, people, backgrounds, surfaces, and objects.

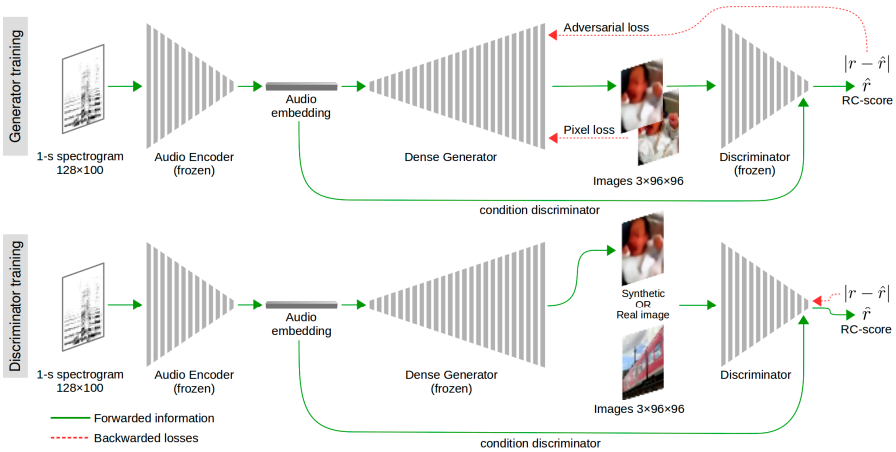

Figure 1. S2I translator training scheme.

From pioneering works that employed data mining techniques to match words and image parts [5], to recent approaches based on DL models, cross-modal initiatives have gained the perspectives of new horizons. These proposals all share a fundamental strategy: the creation of a bridge that connects different modalities. DL models enable efficient data processing, since they can achieve higher abstraction levels through automatically learned features. Additionally, in the case of convolutional neural networks (CNNs) employed for computer vision, interpretable features may be generated in their inner layers. According to Zhou et al. [6], semantic parts and objects emerge spontaneously with CNNs trained for scene classification. Gonzalez-Garcia et al. [7] verified that roughly 10 to 20% of these inner features can represent interpretable concepts tied to textures, materials, semantic parts, and objects. Furthermore, Liang et al. [8] observed that images and sounds have complementary information about the occurrence of common events on a video stream. Those findings reveal a potential strategy for building the cross-modal bridge, one which is based on the assumption that both visual and aural modalities share extractable semantic information about acoustic events. For example, a video of a beach scene might contain both the imagery and sounds of waves crashing on the shore. If we can capture the aural–visual correspondence with a tractable and meaningful representation, then we will be able to trace the path towards the aimed direction. Due to the complexity involved, previous studies adopted simplified approaches using data with low content diversity and/or supervision/self-supervision. In contrast, we propose to perform cross-modal S2I translation addressing diverse audiovisual content. Although the addressed sonic universe is restricted to five sound classes, namely, Baby cry, Dog, Rail transport, Fireworks and Water flowing, our models were trained on over eight thousand distinct scenes from a dataset characterized by both inter- and intraclass diversity. We also highlight that our translator model does not require supervision or self-supervision for training. Sound class annotations were used solely for data preparation, just enough to ensure that the acoustic related element/event was present in both aural and visual modalities. The use of supervision or self-supervision for training the translator would restrict the aural and/or

visual representations to pre-defined concepts. Additionally, in these approaches, it is hard to know if the model performed well due to a connection between modalities or if the generated image merely fitted the class features.

The system we present (Figure 1) constitutes an end-to-end solution obtained after the training of an autoencoder to define the audio latent space and a generative adversarial network (GAN) [9] equipped with a deep densely connected generator to perform cross-modal translation and synthesize the images. Additionally, we introduce an approach that employs informativity classifiers as a way to quantitatively evaluate the performed S2I translation. This enables us to analyze the influence of network-bottleneck variation on the translation process. The results subtly indicate a trade-off between pixel space convergence and informativity, with better results observed, respectively, for higher and lower dimensionalities of the audio embedding space. Though the specified S2I translation task is quite challenging, we were able to obtain models that effectively translated over 14%, on average, of unknown sounds into informative images. To the best of our knowledge, this is the first study to tackle S2I translation with such diversity of audiovisual content, all achieved without resorting to any type of supervision. Furthermore, we present the techniques that we have developed to address issues like latent space continuity, model generalization, and GAN training stabilization.

This text is organized as follows: In the next section, we provide a review of prior research on cross-modal processes. Section 3 explains the challenges inherent in performing S2I translation. In Section 4, the details of the S2I translator are outlined. Section 5 presents the results we have obtained with our translator. Lastly, we conclude the study by sharing our final considerations and offering suggestions for future works.

2. Related Work

The present study proposes a S2I translation system employing DL models [10] to generate content that is both perceptually meaningful and semantically coherent for human receivers. In this section, we provide a concise literature review of studies that have employed DL methods for aural–visual cross-modal processes.

Chen et al. [11] conducted a study on S2I translation using conditional GANs [12]. While adopting a translation structure similar to the one used by Isola et al. [13] and Zhu et al. [14], they addressed a different problem: cross-modal content generation in both directions, image-to-sound (I2S) and S2I. Their aim was to translate audio tracks of musical solo performances into images of a person playing the corresponding instrument, and vice versa. Following the strategy of previously mentioned cross-modal approaches, and drawing from the DCGAN architecture [15], their system consisted of an encoder and a conditioned GAN. In the context of S2I translation, their model produced good results when tested on the URMP audiovisual dataset [16] comprised of studio-quality video tracks featuring uniformly framed individuals playing instruments against a blue background. However, when tested on a more diverse dataset, the quality of the synthesized images dropped considerably. Hao et al. [17] presented a framework called CMCGAN to perform cross-modal aural–visual mutual generation. While also capable of performing I2S and S2I translation using the URMP dataset, their framework demonstrated an enhancement in quality for cross-modal reverse translation from synthetic images/sounds compared to the same task using ground-truth image/sound pairs. This progress was attributed to their improved handling of dimension and structural asymmetry across different modalities, which was achieved through noise injection. Similar to the work by Chen et al. [11], the weak points of this approach are its low diversity and the uniformity of the audiovisual content. Duan et al. [18] utilized the same URMP dataset to carry out a cascade coarse-to-fine S2I translation. Instead of feature embeddings, they opted for a supervised approach to keep the cross-modal translation consistent with high-level semantics. Employing attention mechanism on generators, in addition to class-based loss across all learning stages, and a residual class label to guide finer image generation, they were able to improve significantly the results obtained by Chen et al. [11] and Hao et al. [17]. Different from these

approaches, our work proposes addressing S2I translation without employing any type of supervision for training, i.e., avoiding class-based losses, and utilizing larger and more diverse audiovisual datasets.

Wan et al. [19] present the outcomes of their S2I translation proposal, one which uses conditional GANs trained on video data. Unlike ours, their approach is entirely supervised. Apart from extracting the audio feature vector from SoundNet [20], their generator and discriminator are both trained with an auxiliary classifier to enhance the semantic coherence between the generated image and the corresponding input sound. Also, they propose a sound–image similarity score to improve discriminator training. Similar to the strategy employed by Chen et al. [11], their S2I translator generates relatively informative images when applied to audiovisual data characterized by low content diversity and uniform backgrounds. The sound classes used are: Baseball, Dam, Plane, Soccer, and Speedboat. Utilizing a subset of the same dataset, Yang et al. [21] present an S2I translator based on a stacked GAN architecture. Like the work of Wan et al. [19], their approach is entirely supervised. They also use SoundNet-extracted audio features for the generator input and an auxiliary classifier for GANs training. Moreover, they reverse the translation via an I2S net to validate the audio-content consistency between the original audio embedding and the reversed one. The authors showcase the outcomes of their S2I translation for two sound classes, namely, Baseball and Soccer, obtaining informative images. However, the dataset utilized for this experiment consists of only 2065 sound–image training pairs. Furthermore, a notable limitation present in both studies is the data-splitting method, which does not prevent sound–image pairs of the same video from appearing in both training and test sets. This procedure compromises the evaluation of the system, since the model may be tested with sounds from known scenes.

Another S2I task that has already been explored involves generating images of faces from speech audio. Duarte et al. [22] present an end-to-end solution, named Wav2Pix, aimed at addressing this cross-modal challenge. They employ a GAN conditioned on an audio embedding extracted from speech. The model is capable of generating realistic and diverse facial images. However, their method requires the use of a clean dataset with precisely framed faces and high-quality audiovisual content. Additionally, to obtain acceptable results, their generator needs to be conditioned on known voices. Oh et al. [23] present a model called Speech2Face which tackles a similar problem. They train an encoder to align visual embeddings with those generated by a pre-trained face-recognition network [24]. The subsequent decoding process is performed using a separately trained reconstruction model [25], generating images of faces in a canonical form—precisely framed, frontally positioned, and displaying a neutral expression. Unlike our approach, these works address a specific domain, which significantly reduces audiovisual content diversity.

Chatterjee and Cherian [26] present the Sound2Sight framework, designed for generating video frames conditioned on the audio track and preceding frames. Briefly described, their framework follows an encoder–decoder auto-regressive generator architecture, one which produces one video frame at a time using two long short-term memory (LSTM) networks. Different from our approach, Sound2Sight does not process pure S2I translation, as the visual modality is also present in the input. This aspect enhances the generation of plausible images, but characterizes the process as a multimodal task rather than a cross-modal one. Moreover, the conducted experiment utilizes three datasets separately, each consisting of specific audiovisual content, resulting in limited diversity.

Also addressing a S2I task, Shim et al. [27] propose an end-to-end solution for generating images of birds conditioned on call sounds of correspondent species. After training a sound classifier, they obtain the audio embedding, which then serves as input to a conditional GAN. Unlike our proposal, they employ a class-based encoder. Additionally, their adversarial training is also supervised, as their discriminator is trained not only to assess the realness of generated images but also to predict the species label. In another study on bird sounds, Hao et al. [28] present AECMCGAN, a framework based on their previous work [17], and designed to perform both I2S and S2I translation with the addition of at-

tention modules to capture intra- and inter-modality global dependencies. They obtained improved results compared to prior studies by using their own dataset for cross-modal translation of bird sounds and a subset of the previously mentioned URMP dataset. However, both datasets are domain-specific, focusing on a limited range of audiovisual diversity. Another drawback of these two studies is that the split of the data does not prevent the model from being tested with sounds from known audio streams. As previously mentioned, this procedure compromises the evaluation of translation quality.

Sanguineti et al. [29] conducted a study on multi-modal image generation for audio-visual inpainting. Yet, their model is also capable of performing cross-modal generation conditioned on sound only. Their pipeline consists of a coarse-to-fine two-staged process. For the first stage, they employ self-supervision for the audio data, generating a low-resolution image using a model adapted from PixelCNN [30]. Subsequently, a finer image is generated through a GAN conditioned on the low-resolution image and the audio features. For cross-model tasks, they achieved over 25% and 21% accuracy using diverse data from AudioSet [31] and VGGSound [32] datasets, respectively. Notably, unlike our approach, their method is partially self-supervised. Additionally, for the second stage, the GAN is also trained employing a perceptual loss using high-level feature maps from a pre-trained VGG network [33] which is a supervised model.

Sung-Bin et al. [34] present a S2I solution based on visual discrete representation learning. They employ self-supervision to train both a visual encoder and an image generator which is part of a GAN conditioned on the representations obtained from the visual encoder. They then train an audio encoder using contrastive loss to align the audio embedding to the anchored visual latent space. Their model's training and evaluation utilize the VGGSound [32] and VEGAS [35] datasets. Using the audio embeddings from the aligned encoder and a frozen generator, their system returns more than 83% of images with the correct sound-emitting source depicted. In contrast to our approach, their training is self-supervised, which, as discussed earlier, is prone to producing class-biased results.

Furthermore, the studies conducted by Zhu et al. [36] and Vilaça et al. [37] provide comprehensive surveys on state-of-the-art audiovisual-correlation learning methodologies.

3. Inherent Challenges in S2I Translation Processes

S2I translation shares a common challenge with other cross-modal tasks, which is finding the semantic correlation between the two modalities. In our case, the system must connect the acoustic events present in the audio stream to the semantically correlated elements in the visual modality. While it is an easy and intuitive process for humans to learn semantic correlation between images and sounds, it becomes a challenging task for machines, largely due to the disparity between the audio waveform domain and the image RGB color domain [36]. For instance, there exists a common conceptual or semantic entity between the sound of a foreground dog bark in a background acoustic environment and an image featuring a prominently positioned dog. However, the heterogeneity of the representation of the dog entity across these two domains hinders the successful execution of S2I translation.

3.1. Computational Imagination

Due to its characteristics, the cross-modal generation involved in S2I translation belongs to the broader field of artificial intelligence (AI) known as computational imagination [38,39]. The complexity of this task is also a consequence of the fact that such artificial systems aim to assimilate an ability that can be considered exclusive to humans. Based on Stevenson's work [40], Beaney [41] discusses an alignment between philosophers regarding a possible definition of 'imagining', which can be conceived as 'thinking of something that is not present to the senses'. In the context of our study, the missing part is the entire visual modality. When instructing the translator to generate an image based on the input sound, such as a baby crying, we expect the output to be a complete and particular scene with a baby crying, catching all possible information from the audio signal to picture an

informative image. And we aim to avoid producing stereotypical content, such as an image of a baby standardly positioned on a generic background. Given that the input sound will likely be different from any sound known by the translator, we assume that the generated image will probably not resemble the visual surroundings corresponding to the captured sound. Besides, many elements of the original scene may leave no trace in the audio signal, such as the color of the baby's clothes, for instance. Thus, as in a sound-to-imagination mental process, we expect the computer to use its audiovisual knowledge to 'imagine' an approximate scene featuring the sound-emitting source, as well as related elements that can contribute to the picturing of an informative image.

3.2. Computational Creativity and Divergence/Convergence Methods

Another inherent characteristic of cross-modal generation processes is computational creativity, which is fundamental to providing original outputs. To 'imagine' the surrounding scene corresponding to the sound, the system may need to blend known images, gathering visual 'memories' from the imagery acquired during the training phase in order to create an original image. To reach this goal, the system must analyze the input sound, involving the encoded patterns learned from the training sounds, to generate a representation of it that effectively drives the 'creation' of the output image. The outlined process follows ideas frequently exposed in philosophy, psychology, and cognitive science about the interrelations between imagination, perception, memory, knowledge, and creativity [41]. Pereira and Cardoso [42] emphasize the importance of extending the established AI techniques to improve the divergence/convergence abilities [43,44] of algorithms in order to achieve 'creativity'. Among other AI approaches, genetic algorithms are probably the first to employ divergence/convergence methods to obtain original results, but are mostly limited to exploring narrower knowledge spaces [42]. More recent AI techniques, such as GANs, are capable of exploring wider spaces. The adaptation of this architecture with a conditioned generator [12] is a common approach in current studies on cross-modal tasks. However, despite their effectiveness in producing realistic results, GANs are known to be unstable due to adversarial training [45,46]. A solution presented by Radford et al. [15], named Deep Convolutional GANs (DCGAN), helps overcome this behavior. Their approach comprises a set of architectural constraints that have demonstrated stabilization of GAN output in most settings, although other forms of instability, like filter collapse, persist [15]. These issues lead to a fundamental problem in training generative models, namely, the necessity to increase the diversity of synthesized content [47,48]. Addressing this becomes especially crucial when pursuing creativity, as in our case. Greater diversity in training data can indeed enhance the model's ability to generate diverse output [49]. On the other hand, excessive diversity may hamper model generalization. We are acutely aware of the challenge posed by data diversity, both in training and in data generation. In our work, we utilize diverse audiovisual content to train the translator and aim to obtain a model with enough generalizability to 'imagine' informative images from unknown sounds. Also, the model must be sufficiently 'creative' to generate diverse and original images. This task is undoubtedly challenging, and despite the low percentage of informative images obtained (around 14%), we demonstrate that it is possible to take steps towards the S2I translation objective.

3.3. Addressing the Problem

We train our translator without any type of supervision and using 48,945 sound–image pairs extracted from more than 8000 different scenes with diverse audiovisual content. These scenes are representative of the type of visual output we expect the translator to generate. Additionally, we test the translator exclusively with scenes unknown to the model. These criteria set our study apart from previous works that adopted more simplified approaches, without addressing content diversity and/or employing supervision/self-supervision for training, as explained in the previous section. In some of the studies mentioned, the task being performed is similar to an image retrieval process, in which

the system fetches from a database the image that best fits the query. The exploration we have conducted, along with the corresponding solutions and the S2I translator we provide, represent a modest attempt to tackle this significant challenge. Potential alternative approaches include attention-based methods [50,51], causal reasoning [52,53], explicit incorporation of prior knowledge through rules and constraints [54], bags of acoustic events [55] and/or visual elements [56], sound separation [57], image segmentation [58], and different levels of supervision. These techniques may impose restrictions that could even make the translator produce more realistic results. Yet, such data structuring may lead the system to become in conflict with the objectives of the translation we propose, a situation which demands a commitment to diversity and imagination. Instead of relying on explicit human-generated knowledge, we explore the power of deep neural networks (DNN), including GANs, to model the aimed S2I translator. Given the task's complexity, designing and training these DNN models has proven challenging. Without minimizing the difficulties, except for constraining the number of sound classes used, we adopt a clear, exploratory approach, presenting a solution that gives clues on how to tackle the problem. Nonetheless, we also provide brief descriptions and suggest possible explanations for the observed phenomena. Furthermore, we opted to train the model without employing class-based losses. This choice aimed to prevent class-biased outcomes, allowing the translator to freely use the acquired multimodal knowledge without the potential restrictions imposed by a supervised approach. As discussed in the Introduction, the use of supervision for training the translator would restrict the aural and/or visual representations to pre-defined concepts, whether they are acquired by humans (for supervision) or extracted from unsupervised discrete representation learning (for self-supervision). A class-based loss function, for instance, may help the model to generate informative images. But it is hard to know if the model performed well due to a connection between modalities or if it is simply an image that fitted the class features. With our proposed translator, an informative image is necessarily generated through a successful connection between aural and visual modalities, because there is no clue at any stage of the process about the class of the input sound or the target image.

4. Sound-to-Image Translator

We developed an end-to-end S2I translator that includes a convolutional audio encoder and a conditioned deep densely connected cross-modal generator trained with an (also conditioned) discriminator. As discussed in the previous section, S2I translation is a complex task, and, since the beginning of this research, we have been aware that dealing with content diversity would be a major challenge. This understanding has guided decisions regarding neural network architecture, model regularization and training algorithm, which are all detailed in this section.

4.1. Overview

The training process of the S2I translator is depicted in Figure 1. Initially, an audio autoencoder is trained using log-mel spectrograms computed from 1 s audio segments. Then, the frozen encoder is used to extract the audio embeddings that will be forwarded to the generator. During the training phase, the generator will try to fool the discriminator, which is trained once for every five updates of the generator. At this stage, the discriminator will receive balanced batches of real and synthetic images, along with their corresponding target scores, for modeling visual feature extraction. An aural–visual coherence check is performed by concatenating the source-audio embedding with the input of the discriminator's last layer, effectively merging aural and visual modalities. This enables the discriminator to jointly assess both the realness of the generated images and their semantic coherence with the corresponding audio, and output what we call a realness-and-coherence score (RC-score).

4.2. Network Architecture

As depicted in Figure 2, the audio autoencoder consists of a mirrored architecture of 26 convolutional layers, each one followed by batch normalization (BN). The inner layers of both the encoder and decoder are activated by rectified linear units (ReLU), while the last layer employs a hyperbolic tangent function (TanH). As for the decoder part, the network applies dropout regularization to each inner layer to prevent overfitting.

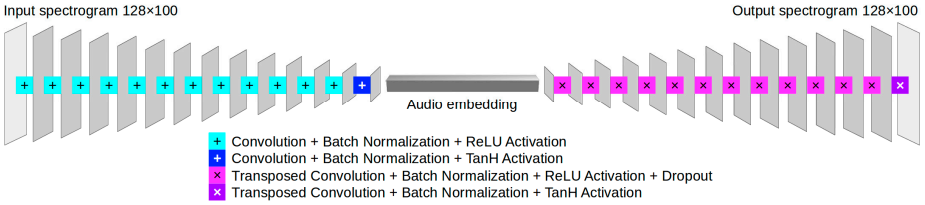

Figure 2. Audio autoencoder architecture.

Regarding the generator architecture, a notable improvement over the baseline S2I translator was achieved by employing a deeper 25-layer architecture, especially when applying "skip connections" to make a dense generator. This final architecture draws inspiration from DenseNet [59] (Figure 3). Compared to the initial 13-layer sequential generator, which resembles the audio decoder in structure, the new deeper and denser architecture allows for a substantial improvement in the quality of the generated images. The convolutional layers of the generator are followed by BN and ReLU activation, except for the output layer, which employs TanH. Additionally, we include dropout regularization between each dense block to improve model generalization, as well as to prevent deterministic inference through its application also at test time, as utilized by Gal and Ghahramani [60] and Isola et al. [13].

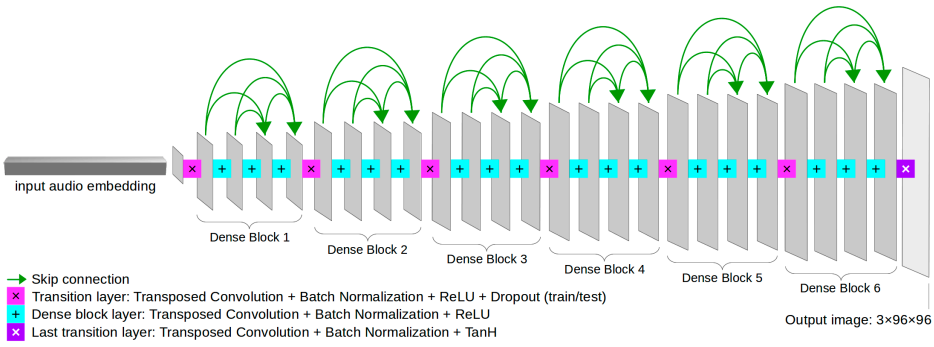

Figure 3. Generator architecture with skip connections.

As for the discriminator architecture (Figure 4), it closely resembles the structure of the audio encoder, differing primarily in the input and the shape of the feature maps. Additionally, the last convolutional layer of the discriminator is conditioned with the embedding extracted from the audio encoder. Moreover, dropout regularization is applied after each inner layer. The last convolutional layer directly outputs a scalar corresponding to the RC-score.

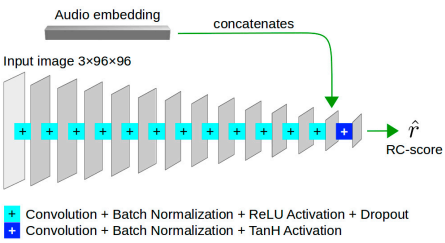

Figure 4. Discriminator architecture.

4.3. Training Methodology

Here we define a set of spectrogram–embedding–image data triples $\{S_i,\ x_i,\ Y_i\}$, consisting of spectrograms $S_i \in \mathbb{R}^{h\times w}$, audio feature vectors $x_i \in \mathbb{R}^f$, and real-color images $Y_i \in \mathbb{R}^{h\times w\times c}$, in which each element of the triple corresponds to the same i^{th} acoustic event. All the real numbers are limited to the interval $[-1,\ 1]$, since pixel values are normalized to the mentioned range before entering the network, and audio embeddings, as well as the realness-and-coherence score (RC-score) $r \in \mathbb{R}$, both activated by TanH, fit the same interval. Regarding dimensions: h and w are, respectively, the height and width of spectrograms or images, c is the number of channels of color images, and f is the dimensionality of the audio feature space. The audio encoder A_E and the audio decoder A_D are defined, respectively, as the transformations $A_E(S) : \mathbb{R}^{h\times w} \to \mathbb{R}^f$ and $A_D(x) : \mathbb{R}^f \to \mathbb{R}^{h\times w}$. From the side of the adversarial networks, the generator G and the discriminator D are defined, respectively, as $G(x) : \mathbb{R}^f \to \mathbb{R}^{h\times w\times c}$ and $D(Y \vee G(x),\ x) : \mathbb{R}^{h\times w\times c+f} \to \mathbb{R}$, where Y denotes the real image that is entered into the discriminator, alternating with the synthetic image from $G(x)$.

With respect to the computation of loss functions, we chose to use the mean squared error (MSE) in the pixel space for measuring the distance between the target and the generated spectrogram/image during the training of both the audio autoencoder and the generator. This strategy provided an unbiased and scalable solution, enabling us to perform extensive testing, so that, by observing the behavior of the translator, we were able to improve the architecture and tune the networks for enhanced performance.

The optimization of the autoencoder networks A_E and A_D is performed, minimizing the pixel loss $L_A\left(S, \hat{S}\right)$ defined in Equation (1), where b is the batch size, and, as stated earlier, h and w are the height and width of spectrograms. The loss is computed as the MSE between the target spectrogram S and the generated one $\hat{S} \leftarrow A_D(A_E(S))$. We include the batch iteration in the equations, since this is how the losses are effectively computed, being averaged among all instances of the batch.

$$L_A(S,\ \hat{S}) = \frac{1}{bhw}\sum_{i=1}^{b}\sum_{j=1}^{h}\sum_{k=1}^{w}\left(S_{ijk} - \hat{S}_{ijk}\right)^2 \tag{1}$$

The discriminator D is optimized through the minimization of the score loss $L_D(r,\ \hat{r})$ defined in Equation (2), which is calculated from the batch-averaged MSE between the output RC-score $\hat{r} \leftarrow D(Y \vee G(x), x)$ and the target RC-score r, which can be the maximum (1) or the minimum (-1) score value, depending on whether the input image is real or synthetic, respectively.

$$L_D(r,\ \hat{r}) = \frac{1}{b}\sum_{i=1}^{b}(r - \hat{r}_i)^2 \tag{2}$$

The optimization of the generator G is guided by two objectives. The first aims to minimize the pixel loss $L_G\left(Y, \hat{Y}\right)$ defined in Equation (3), where c, as stated earlier, is the number of channels of color images. The loss is computed as the MSE between the target image Y and the generated one $\hat{Y} \leftarrow G(x)$.

$$L_G(Y,\ \hat{Y}) = \frac{1}{bchw}\sum_{i=1}^{b}\sum_{j=1}^{c}\sum_{k=1}^{h}\sum_{l=1}^{w}\left(Y_{ijkl} - \hat{Y}_{ijkl}\right)^2 \tag{3}$$

The second objective aims to minimize the adversarial loss based on the RC-score obtained from the discriminator D. We implemented the moving-average adversarial loss $L_G^{ma}(r_{max},\ \hat{r})$ defined in Equations (4) and (5), where r_{max} is the maximum RC-score value, t is the current epoch number, $\overline{L}_{Gi}$ is the average adversarial loss for epoch i, and k is the number of averaged data-points. When employing the moving-average loss instead of the current batch loss $L_G(r_{max}, \hat{r})$ (Equation (4)) to train the generator, adversarial training instability was significantly attenuated. In our case, this was especially important because

the ratio of generator/discriminator training update, which was 5, was higher than typically applied, causing even more instability than usual during GAN training.

$$L_G(r_{max},\, \hat{r}) = \frac{1}{b}\sum_{i=1}^{b}(r_{max} - \hat{r}_i)^2 \tag{4}$$

$$L_G^{ma}(r_{max}, \hat{r}) = \frac{L_G(r_{max}, \hat{r}) + \sum_{i=t-k+1}^{t-1} \overline{L}_{Gi}}{k} \tag{5}$$

The final generator loss is expressed in Equation (6). The adversarial loss is scaled by a factor λ to balance the mean amplitude of the two terms.

$$L_G = L_G\left(Y, \hat{Y}\right) + \lambda L_G^{ma}(r_{max}, \hat{r}) \tag{6}$$

Leaving apart the audio encoder training, we present in Algorithm 1 a pseudo-code describing the main steps of the training of our translator.

Algorithm 1 Pseudo-code for the training of the translator's GAN.

Input b (b is the batch size)
Input n_g (n_g is the number of training iterations of the generator)
Input n_{gd} (n_{gd} is the number of training iterations of the generator per discriminator training)
Input r_{min} (r_{min} is the minimum RC-score value)
Input r_{max} (r_{max} is the maximum RC-score value)
Input λ (λ is the adversarial loss scale factor)

for n_g iterations **do**
  Get b spectrograms S from stored data:
    $S_{batch}\ \{S_1,\, S_2,\, \ldots,\, S_b\} \leftarrow data$
  with A_E frozen
    Get b audio embeddings x from the audio encoder:
      $x_{batch}\ \{x_1,\, x_2,\, \ldots,\, x_b\} \leftarrow A_E\ (S_{batch})$
  end with
  Get b real images Y from stored data:
    $Y_{batch}\ \{Y_1,\, Y_2,\, \ldots,\, Y_b\} \leftarrow data$
  if current iteration number is multiple of n_{gd} **then**
    with generator G frozen
      Update the discriminator D to minimize:
        $(L_D(r_{max},\, D(Y_{batch},\, x_{batch})) + L_D(r_{min},\, D(G(x_{batch}),\, x_{batch})))$
    end with
  end if
  with discriminator D frozen
    Update the generator G to minimize:
      $(L_G(Y_{batch},\, G(x_{batch})) + \lambda L_G^{ma}(r_{max},\, D(G(x_{batch}),\, x_{batch})))$
  end with
end for

5. Experiments

In this section, we explain the heuristics behind our approach, detailing the training strategies employed and the datasets used for the experiments. Also, we present our solution for evaluating the translated images using informativity classifiers. We complete the section providing both quantitative and qualitative evaluations of the S2I translation results. For further information regarding the experiments and the code implemented for the networks training, please refer to https://purl.org/s2i (accessed on 27 September 2023).

5.1. Data Used

The AudioSet, as described by Gemmeke et al. [31], consists of a large-scale audiovisual dataset of manually annotated acoustic events. Starting from the literature and manual

curation, the authors defined a structured hierarchical ontology of 632 audio classes to collect data from human labelers. The goal of their task was to probe the presence of specific audio classes in 10 s segments of YouTube videos. The complete dataset contains over 2 million videos and the labeled segments employ part of the AudioSet ontology. The provided data is characterized by its highly diverse audiovisual content. Dealing with such variety makes our study distinct from most previous works. Despite the fact that we use only five sound classes, the sound–image pairs are extracted from over eight thousand distinct scenes, resulting in substantial inter- and intraclass diversity. For training and testing our S2I translator, we employed a subset of AudioSet named VEGAS, made available by Zhou et al. [35] for their study on cross-modal image-to-sound translation. This dataset provides cleaner data, in which the starts and ends of addressed acoustic events are precisely annotated. Additionally, the tracks have been inspected to verify whether the elements/events related to the sound were present in both visual and aural modalities, and non-matching segments were removed. The complete VEGAS dataset consists of 28,109 videos of a maximum duration of 10 s, distributed across ten sound classes, among which we use five: Baby cry, Dog, Rail transport, Fireworks, and Water flowing. The original VEGAS dataset is unbalanced; thus, to prevent class biasing, an equal number of segments is used for all classes. Table 1 presents the number of original video tracks for each sound class, along with their corresponding 1 s segments designated for training, validation, and test sets.

Table 1. Sound classes and their respective number of video tracks and 1 s segments for training, validation and test sets.

Sound Classes	No. of Original Video Tracks (max. 10 s)	No. of 1 s Video Segments		
		Training	Validation	Test
Baby cry	2059	9789	1115	1365
Dog	2785	9789	1115	1365
Fireworks	3115	9789	1115	1365
Rail transport	3259	9789	1115	1365
Water flowing	2924	9789	1115	1365
Total	14,142	48,945	5575	6825

Regarding audio data, log-mel spectrograms were computed according to the following procedure: the signal was split into 25 ms frames, with a 15 ms overlap; a Hamming window was applied to the frames and the short-time Fourier-transform (STFT) was computed; its squared magnitude was integrated in 128 sub-bands using triangular weights according to a non-linear mel-scale, and the logarithm of those sub-band energies was computed. For a 1 s audio segment, a matrix of 100 $\times$ 128 was obtained. The segmentation of one of these spectrograms and the respective frames from the original video are illustrated in Figure 5. As for visual data, images were extracted from the central frame of the corresponding 1 s video segment. Before being loaded into the neural network, these extracted images were square cropped at the center and then resized to 96 $\times$ 96 pixels.

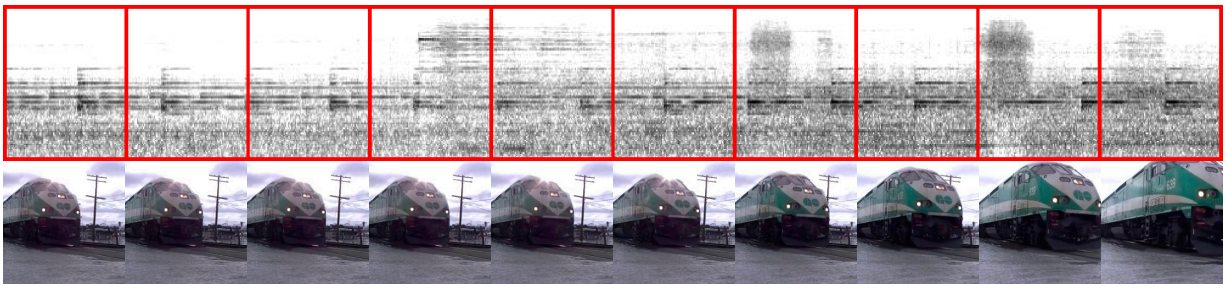

Figure 5. Images of 1 s mel spectrogram segments and respective central frames extracted from a video track of the class 'Rail transport'.

5.2. Preliminary Exploration and Training Details

5.2.1. Generator's Architecture Evolution

As mentioned in Section 4.2, a significant enhancement in the quality of translated images was due to the evolution of the generator architecture from a 13-layer sequential design to a 25-layer densely connected structure. The deeper architecture, composed of dense blocks inspired by DenseNet [59], enabled the translator to picture shapes, colors, and sharp edges much closer to the original ground-truth training images. In this way, it was possible to achieve a reasonable quality of translation without overfitting the model. Given the difficulty in making a quality comparison using test data due to visual decoupling between target and translated images, we show in Figure 6, as illustration, images generated using training data, i.e., known sound–image pairs. In this figure, we display results obtained from all three generator architectures: sequential 13-layer, sequential 25-layer, and dense 25-layer.

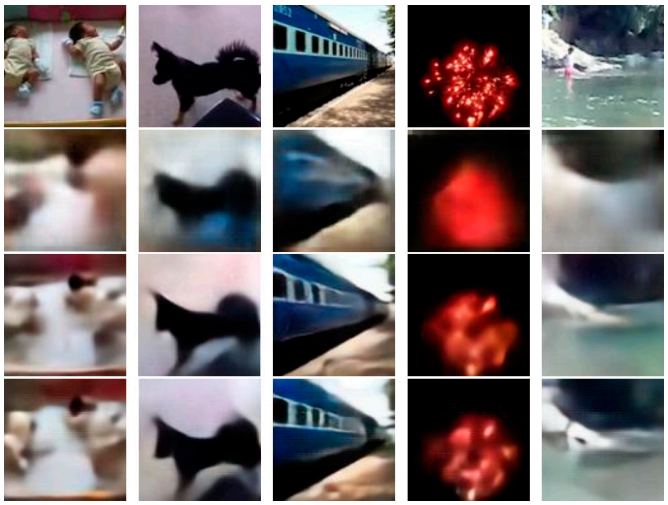

Figure 6. Illustrative comparison between ground truth (1st row on top), and images generated by the sequential 13-layer (2nd row), sequential 25-layer (3rd row), and the dense 25-layer generator (4th row) using known sound–image pairs.

5.2.2. Models' Generalization

With respect to the training of the audio autoencoder, we successfully softened the latent space through dropout regularization applied to the audio decoder, which improved generalization significantly. For the generator, as mentioned in Section 4.2, we included test time dropout regularization between each dense block. This strategy, in addition to helping to generalize the model, induces stochasticity within the generator, enabling greater visual variety in translations. In practice, this technique noticeably enhanced the diversity of the generated images. However, it appears that test time dropout only contributed to improving model generalization up to a certain extent, as translated images were often non-informative. Actually, evaluating the generalizability of the generator is not an easy task, since the synthesized images most likely do not share the visual structure of the images corresponding to the input sound. This visual decoupling implies that even if generated images are semantically coherent with the original sound, visual elements will not necessarily appear on the same position as in the target images. Consequently, pixel loss obtained from the test set becomes useless. In fact, visual matching is rare, and even when it occurs, it is an approximate match. Aware of those limitations, we decided to focus our analysis on two key aspects that we consider suitable for assessing the quality of the generated images, namely, interpretability and semantic coherence, which can both be summarized in the term 'informativity.'

5.2.3. Networks Activation

Throughout our exploration, we noticed that the activation functions of both the generator and the discriminator were playing prominent roles in the process. This phenomenon aligns with findings reported by Glorot et al. in tasks such as image classification and text sentiment analysis tasks [61], as well as domain adaptation [62], where network activation sparsity improved generalization. We observed the same effect in our translator during the network's tuning phase. Based on these insights, we chose to employ ReLU activation for all inner layers of both the autoencoder and the GAN. The network sparsity provoked by ReLU layers helped to regularize the model, allowing it to capture the essential semantic information from the input sounds and keep it until the lower levels of abstraction of the generator model, close to the effective synthesis of the image at the network's end. This explains the better generalization observed, along with a noticeable increase of the informativity of the generated images. On the other hand, training the translator with Leaky ReLU or exponential linear units (ELU) activation resulted in less stability, and model output frequently transitioned from blurry to sharp (yet rather abstract) images. Although producing details of increased validity for the training data, models activated by these functions were prone to overfitting, generating more non-interpretable images full of visual artifacts when translating unknown sounds. In fact, some models activated by Leaky ReLU were not able to generate one single informative image.

5.2.4. Data Balancing and Networks Initialization

Furthermore, we ensured an equal number of real and synthetic images, finding no need to vary their ratio, as indicated by Lucas et al. [63], since we did not experience noticeable mode collapse issues, except for the initial stages of GAN tuning. Across all networks, we applied the Xavier initialization method, referred to by Glorot and Bengio [64] as 'normalized initialization'. Therefore, weights W for each network layer were sampled from the random uniform distribution defined in Equation (7), where n_i is the number of incoming network connections and n_{i+1} is the number of outgoing network connections for the layer i.

$$W \sim U\left(-\frac{\sqrt{6}}{\sqrt{n_i + n_{i+1}}}, \frac{\sqrt{6}}{\sqrt{n_i + n_{i+1}}}\right) \tag{7}$$

5.2.5. Algorithm Hyperparameters and Technology Stack

All autoencoders of the five different embedding dimensions were trained with an initial learning rate of 0.05 and momentum 0.9. For both the generator and discriminator networks across all embedding dimensions, an initial learning rate of 0.1 and a momentum of 0.5 were employed. The scale factor λ of the generator adversarial loss (Equation (6)) was set to 0.1, and the discriminator was trained once for every five updates of the generator. We utilized mini-batch gradient descent optimization for all training, employing a batch size of 64 for each network update. The entire system was implemented in Python version 3.5.2. Machine-learning-related code was implemented using PyTorch library version 1.1.0. The training processing of all models was run on a Supermicro™ SYS-7048GR-TR server allocating 160 gigabytes of RAM and two Intel™ Xeon E5-2670 processors at 2.30GHz. The execution was accelerated using a GeForce™ Titan Xp GPU accessed via a CUDA™ platform, version 9.0.176.

5.3. Informativity Classifiers

As an additional contribution, we introduce a solution using classifiers to infer whether the translated images are interpretable and semantically coherent, or, briefly, if they are informative. According to the criteria set for the present study, an image is considered informative when it surpasses a minimum threshold of informativity, ensuring that the depiction of the primary sound-emitting source is recognizable.

Assessing the true quality of S2I translation posed an additional difficulty during the experiment. Since class information was available, we initially attempted to use

ordinary image classifiers to evaluate our translator. The idea was to verify whether the generated image would be classified as the input sound corresponding class. However, this alternative proved ineffective, as the reported classification scores were unrealistically high. As discussed in Section 3, S2I translation is a challenging task, and generated images are mostly non-informative. During the tests, over 50% of generated images were reported to match the original sound class, but upon visual inspection, it became evident that this did not represent the translator's actual performance, even when using the best generator models. Two facts made the evaluation of the S2I translation more difficult, producing biased results as to classification. Firstly, the low percentage of informative images complicated the task, which was to be expected, due to the inherent difficulty of the process. This turned out to be a problem, as the large number of non-informative images had the potential to distort the classification outcomes significantly. For instance, if these non-informative images were randomly classified among the five sound classes, they would considerably inflate the overall effective accuracy. Secondly, inaccuracy resulted from the fact that even informative images were essentially different from real images. They often exhibited less sharpness and less visual diversity compared to the actual images used for training, causing the classifier to perform inaccurately when testing synthetic images. We even tried to analyze the output vector of the softmax function from the visual classifier in an attempt to find a correlation between the class distribution and classification reliability, but no improvement was observed through this approach.

After the aforementioned unsuccessful attempts, we proposed to classify images as informative or non-informative in order to provide a quantitative evaluation of the S2I translator's performance. It was necessary to train one informativity classifier for each sound class, since general classifiers performed poorly. Besides, such general classification would not benefit the experiment in any way, given that our goal was uniquely to report the translators' performance as accurately as possible. Thus, we trained five informativity classifiers using both informative and non-informative synthetic images translated from validation sounds. It is important to note that two different visual data sectioning processes were employed here. One groups images into five classes of sound, regardless of whether they are real or synthetic images, while the other categorizes them as informative or non-informative images, serving solely to assess the translator's performance. Considering these data perspectives, we built five balanced datasets comprising a total of 5000 synthetic images selected among the outputs of 17 previously trained S2I translator models. Each dataset for a specific sound class contained 1000 images (800 for training and 200 for testing), evenly distributed between informative and non-informative classes. Since, at that moment, we had not yet trained informativity classifiers, the screening of generator models relied on the reported pixel loss and class matching of translated images, while the final image selection was based on subjective evaluation.

The S2I translators performed differently for each sound class. For instance, when testing the translator with water flowing sounds, it was able to generate approximately 18% of the informative images, whereas with dog sounds, the best performance was about 6%. Due to these disparities, we had to employ a larger number of translator models for classes with lower performance levels until we completed the set of 500 informative images for each sound class. Regarding the non-informative class, although using just one model would have been sufficient to complete the 500-image set, we maintained the same number of selected images per model to prevent bias in the informativity classifier. This was necessary because image generation models may leave a fingerprint on the output. By ensuring uniformity in the number of selected images per model, we mitigated the risk that artifact patterns [65], or even the level of blur, provided hints about the input image's original class.

We want to emphasize that the informativity classifiers were exclusively employed to assess the translator's performance, and thus, no supervisory information was utilized during the translator's training. The architecture of these classifiers consists of a CNN with five convolutional layers for visual features extraction, and two fully connected layers for

classification. Batch normalization, ReLU activation, and dropout (for training) are applied after each layer, except for the last fully connected layer, where the log probabilities are computed from the output of a softmax function. All five classifiers were trained with an initial learning rate of 0.001, momentum 0.9, and weight decay of 5×10^{-5}, obtaining models with the following accuracies: Baby cry—80%, Dog—80%, Rail transport—84%, Fireworks—82%, and Water flowing—81.5%.

5.4. S2I Translation Results

For quantitative evaluation, we present the performance of different S2I translator models regarding the variation of audio embedding space dimensionality. For qualitative evaluation, we also offer an analysis of a carefully chosen selection of images that accurately represent the quality of S2I translation achieved.

5.4.1. Quantitative Evaluation

By using the set of five informativity classifiers, we were able to make an extensive comparison of the performance levels of different S2I translators. Employing the architecture presented in Section 4.2, the best performance of our translator was obtained with an audio-embedding dimension of 512, resulting in an average informativity of over 14% across all five sound classes. The history of translators' informativity during training is reported in Figure 7a–e. Due to the oscillation of informativity obtained during the model's adversarial training, we evaluated the translator's performance using two averaging metrics: one reports the overall average from epoch 500 to 4000, while the other presents the maximum 50-epochs moving average among the five sound classes (Figure 7f). Additionally, we omitted the informativity data from the two epochs immediately after the discriminator update. Images generated during these epochs often misled the classifiers, leading to erroneous reporting of higher informativity levels despite the presence of visual artifacts in the images. Instead, we relied on the average of the previous five epochs for accurate evaluation.

The pixel loss history during the generators' training is also reported using a 50-epochs moving average (Figure 8a). In general, models trained within a broader latent space showed faster convergence and achieved lower pixel losses (Figure 8a,b), although this tendency was not verified in extreme dimensions 128 and 2048. In spite of that, informativity decreased when the embedding dimension was greater than 512 (Figure 7f). Although in a subtle way, this outcome suggests a possible influence of audio embedding dimension variation on translation performance, and the results seem to point out a trade-off between convergence (in the pixel space) and informativity, with better results observed, respectively, for higher and lower feature space dimensionality. The increased flow of information across the network might have led to model overfitting, which is difficult to confirm, since the generators' testing losses are uninformative due to the visual decoupling mentioned in Section 5.2. On the other hand, models trained in a more constrained feature space could present poorer convergence, but they generally produced more informative images, especially for intermediate dimensions. We hypothesize that this occurs due to the semantic generalization induced by constraining the information flow between source and target spaces, as demonstrated in previous studies on supervised learning [66], generative adversarial learning [67,68], and domain adaptation [69,70]. In such processes, the reduction of the latent variable dimensionality forces the network to extract essential semantic information. In the case of cross-modal tasks, this helps bridge source and target modalities by exploiting high-level semantics that connect the two. In S2I translation, the concept of an acoustic event is what links aural and visual modalities, and the entirety of the AudioSet [31] structure is based on such concepts. It is also important to emphasize that the proposed S2I translation does not simply prioritize image realism, but also places significant emphasis on informativity.

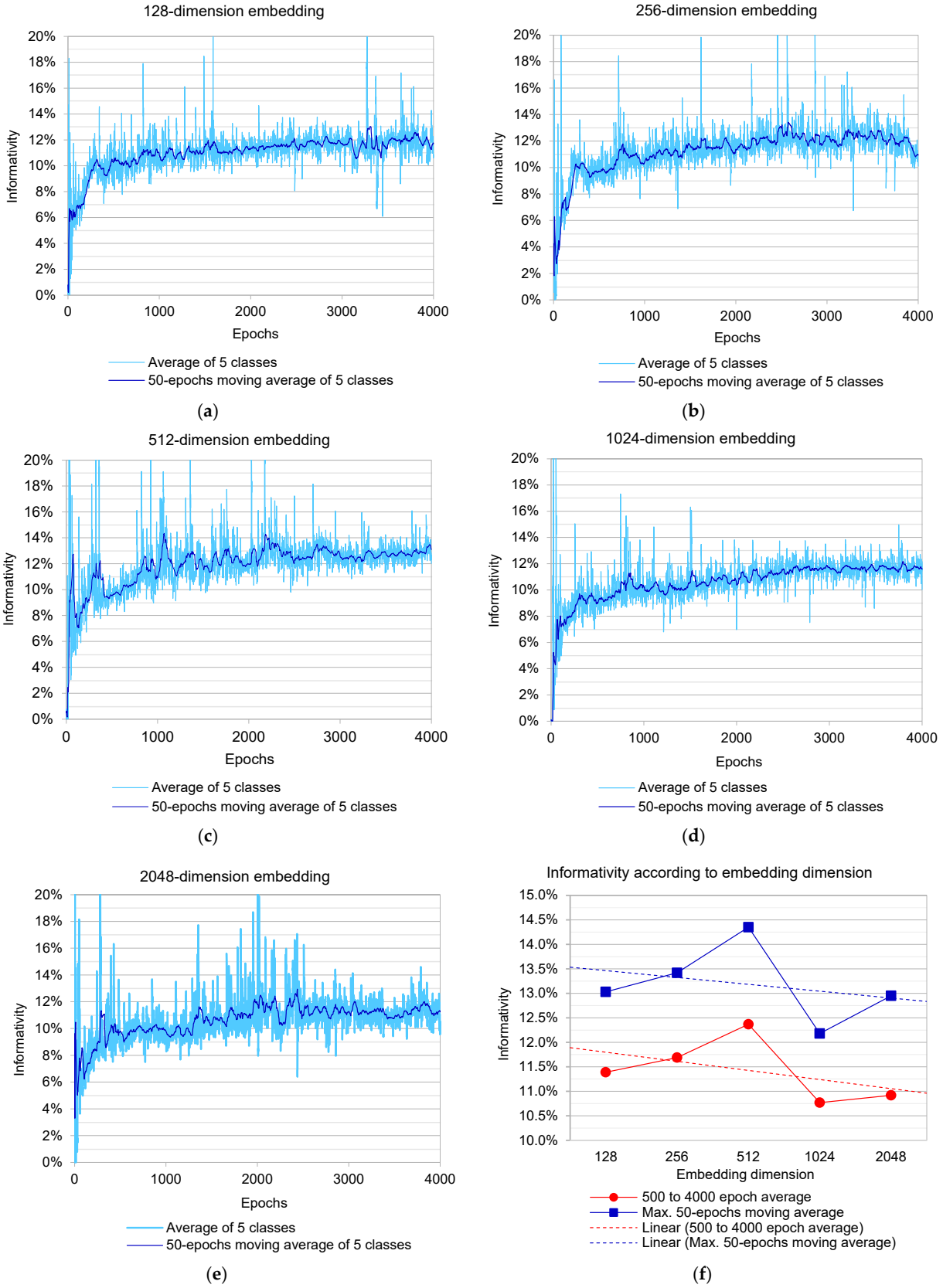

Figure 7. (**a**–**e**) Translator's informativity history for each embedding dimension. (**f**) Informativity variation according to embedding dimension.

Figure 8. (**a**) 50-epochs moving-average pixel loss from training for each embedding dimension. (**b**) Training pixel loss variation according to embedding dimension.

Luo et al. [69] improved GAN training stability for semantic segmentation through bottleneck constraint. We have managed to overcome major instability issues in long-term training using a moving-average discriminator loss, and we haven't identified any correlation between embedding dimension and training stabilization. Therefore, we assume that the improvement of the performance obtained through the bottleneck constraint is not related to GAN stability. We still cannot assert that the reduced bottleneck enforced semantic alignment between modalities, and further tests would be needed to confirm the observed tendency, especially quantifying the information propagation across the network. Nonetheless, we highlight the potential of controlling the information flow for solving problems related to semantic alignment involving different modalities and/or domains.

5.4.2. Qualitative Evaluation

All the generated images presented in this section are conditioned on sounds from unknown videos, and the results demonstrate the translator's achieved generalizability. In Figure 9, we showcase a selection of translated images spanning the five sound classes. Most of the images presented in this section were obtained from a single translator model conditioned on 512-dimensional audio embeddings. Despite the blurry aspect present in most areas of the images, which we discuss next, there are identifiable borders and interpretable shapes in all sound classes, and even, occasionally, sharp details. Color coherence is also noticeable, and the visual structure in most images is drawn in accordance with real scenes. Besides, in most images we can see well-pictured volumes, with correct light and shadow effects. As mentioned previously, we expected content diversity to be our main challenge. In fact, real fireworks images, which present lower diversity compared to other classes, showed the best results, not only in terms of informativity, but also in image sharpness. On the other hand, this could be attributed to the abstract visual nature intrinsic to fireworks scenes, which may make them easier to be rendered by the translator. On the other hand, translations of dog sounds presented the worst results, which we believe to be due to certain characteristics of this class. Since dog sounds are typically short or nearly instantaneous events, it seems that our sound-signal segmentation might not have been sufficiently adequate to effectively model the semantic alignment between the aural and visual modalities for this specific category.

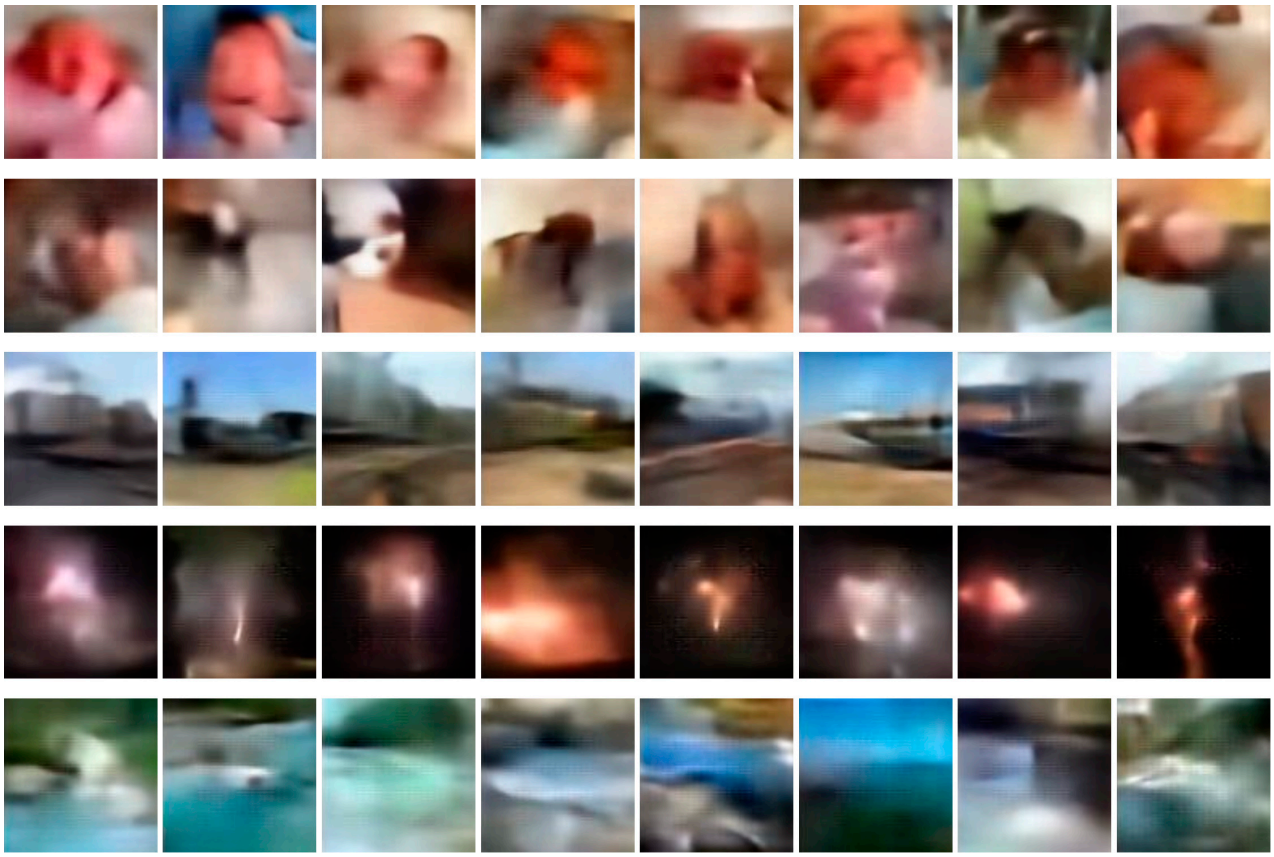

Figure 9. Qualitative demonstration of results among the five sound classes, from top to bottom row: Baby cry, Dog, Rail transport, Fireworks, and Water flowing. S2I translation was performed from unknown sounds using a single translator model conditioned on 512-dimension audio embeddings.

Blurry areas and uncertainty mapping. Apart from dealing with model generalization, we also had to face another important issue regarding the quality of translation. In most generated images, a lack of sharpness was observed, which can be a consequence of using averaged pixel-wise losses, which is known to produce blurry results. Alternatively, these outcomes could also have been triggered by the dropout regularization we used to improve the autoencoder generalization, since it softens the latent-feature space. On the other hand, the adversarial loss may have compensated this tendency to a certain extent [71], since blurry images are penalized due to their unrealistic appearance. Apart from that, S2I translation raises another concern related to image generation: the confidence level of the provided translation. Although we aim to generate well depicted sharpen images, since blurry areas are usually less informative, we hypothesize that the blurry areas may work as an uncertainty map [72,73]. In other words, when the model is not sure about what to draw somewhere in the image, it may produce fuzzy shapes, leaving explicit the inherent uncertainty of the inference. Although we cannot assume that the uncertainty mapping will automatically occur, it is worth investigating the translator's ability to provide such information, either merged with the translated image or in a separated uncertainty layer.

Visual structure and diversity. Another characteristic of the translator is that the model successfully produced diverse outputs, as can be seen in the synthetic images shown throughout this section. Yet, if these are compared to the original images which correspond to the audio input, there is a visible loss of diversity. Furthermore, in Figure 10, and in some examples more than in others, we can notice the aforementioned visual decoupling between the synthetic images and their corresponding ground-truth of the test set, which occurred in most translations. However, we could find exceptions to this rule, and even images translated from unknown sounds may occasionally resemble scenes corresponding

to the input sound. In Figure 11, we present a selection of translations that produced such a result. Using the currently evaluated model, in most cases presented, the translated images share the original visual structure and also the color, but sometimes the generated image has a mirrored structure compared to the original one (Figure 11, first row (a) on top and third row (d)). Also, the translated image may match the original visual structure, but colors and textures are different, like the aforementioned mirrored image (Figure 11, first row (a)) which pictures a baby crying and the dog image (fourth row (a)), in which, despite having completely different backgrounds, the blurry silhouette in the translated image shows the dog in a quite similar pose when compared to its original pair. Some of the remaining examples exhibit unexpected visual matching in Rail transport, Fireworks and Water flowing scenes, and there is even a Baby cry translation (Figure 11, third row (b)) that matches quite precisely both structure and color, except for the baby's clothes. In fact, sound signals carry clues about the surrounding space and scene elements, and this information may be modeled by the translator in order to help picture the output image. Therefore, what seems to happen by chance can be revealing learning processes occurring in multimodal intermediate layers of the translator.

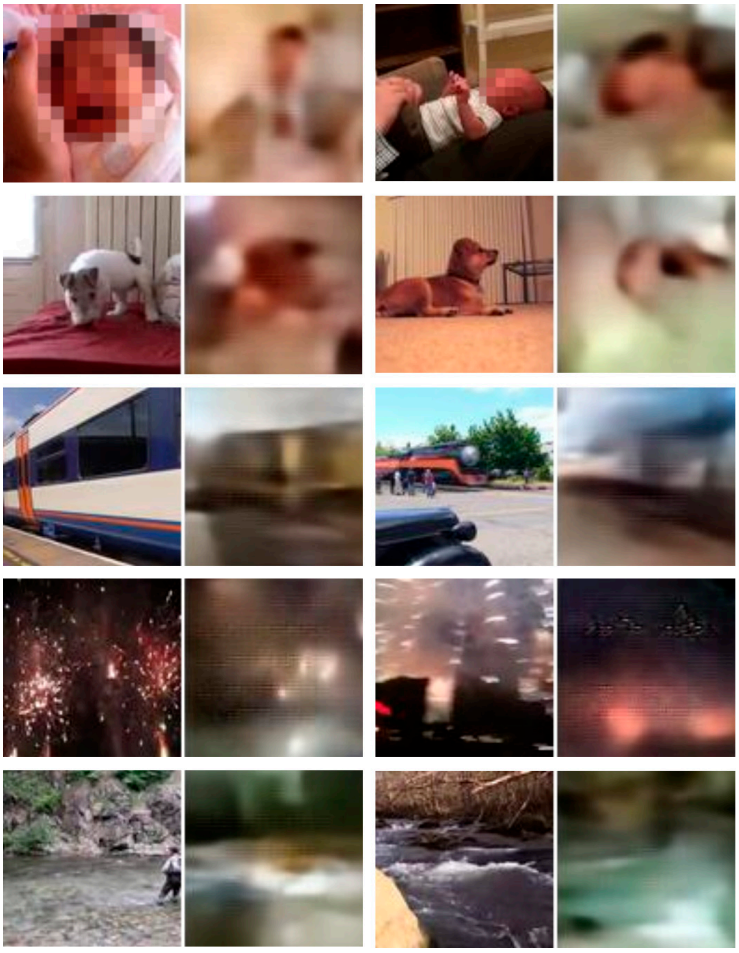

Figure 10. Comparison showing a visual decoupling between original and synthetic images among the five sound classes; from top to bottom row: Baby cry, Dog, Rail transport, Fireworks, and Water flowing. S2I translation was performed from unknown sounds using a single translator model conditioned on 512-dimension audio embeddings. The babies' faces in the images on the left were intentionally pixelated to preserve the children's identities.

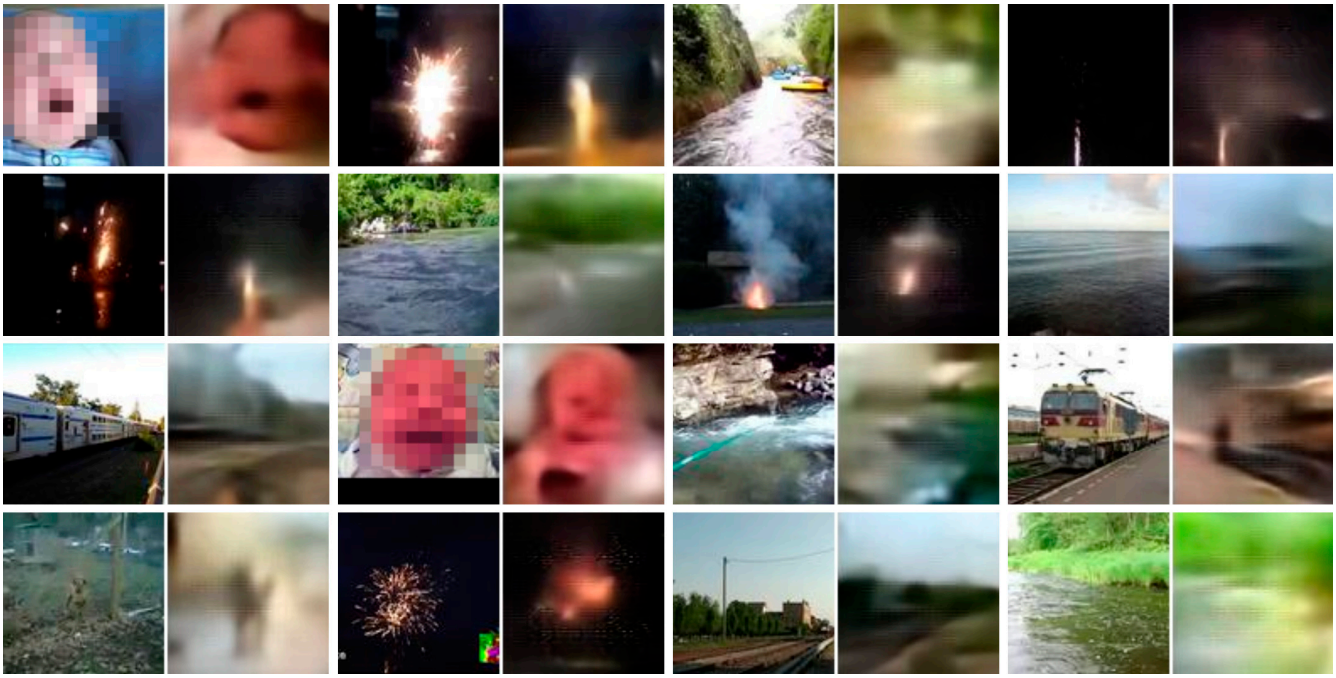

Figure 11. Sharing of similar visual structure (and also color in some cases) between synthesized images and input-sound-corresponding images. Showing examples from the five sound classes: Baby cry—1st row (**a**), 3rd row (**b**); Dog—4th row (**a**); Rail transport—3rd row (**a**,**d**), 4th row (**c**); Fireworks—1st row (**b**,**d**), 2nd row (**a**,**c**), 4th row (**b**); and Water flowing—1st row (**c**), 2nd row (**b**,**d**), 3rd row (**c**), 4th row (**d**). S2I translation was performed from unknown sounds using a single translator model conditioned on 512-dimension audio embeddings. The babies' faces in the original images were intentionally pixelated to preserve the children's identities.

Beyond informativity. Apart from the informative/non-informative images perspective, we identified the following types of translated images:

- Defective—Images that may even be informative, but inaccurately depict elements concerning color, luminosity, size, and/or position.
- Incomplete—Images that may be informative to a certain degree, but lack some essential part of the element of interest. Or, despite having a coherent surrounding scene, the sound-emitting source is omitted.
- Artifactual—Non-informative images that are rather abstract, consisting basically of unrecognizable forms.
- Implausible—Images that may occasionally be informative, but that contain awkward or unlikely elements.
- Surreal—Images that may present some degree of informativity but have a curious or fantastical appearance.
- Creepy—Images that may be partially informative, but that portray parts of living beings in a harrowing way, or that contain ghostly or alien-looking elements. These images could sometimes also be considered defective or surreal, depending on the case.
- Multi-informative—These images depict elements from two or more sound classes.

Bad translations. Regarding badly translated images, such results occur due to different reasons. For instance, certain acoustic events might not have been accurately modeled, leading to unusual or nearly abstract images, as occurred with most dog images. In other instances, the model just mistranslated the input sound, generating, for example, a waterfall scene for a passing train sound, or vice versa. Also, issues related to the dataset itself may impact translation quality. For instance, regarding fireworks sounds, frames from the original training videos frequently contained white subtitles in the lower area, which ended

up being modeled as part of the sound-emitting source. Consequently, these subtitles occasionally appear in generated images as horizontal luminous lines over a dark background (Figure 12, first row (d)). With respect to defective, incomplete, and artifactual images, such outputs occurred with different frequencies, occasionally significantly compromising the informativity of the translation. For instance, a typical defective image occurs when specific parts are inaccurately depicted, as seen in Figure 12 (first row (a)). Here, the baby's eye regions are brighter than the rest of the face, when usually the opposite is likely to happen. However, that which is clearly a bad output can also reveal the translator's ability to separately model different semantic components within the image. Another example of a defective image can be seen in Figure 12 (first row (e)), where the visual structure of the water scene seems to be upside down. And what could be the corner of a rocky beach with clear green water ends up looking like an abstract image. With respect to incomplete images, in the case of rail transport sounds, for instance, the translator may generate an interpretable image of a landscape which is coherent with the acoustic event in question. However, the absence of a train or railway makes the image incoherent (Figure 12, first row (c)). Another example of incomplete image is showcased in Figure 12 (first row (b)), where a potentially informative image of a dog is compromised due to the lack of elements indicating the location of the animal's head, resulting in a somewhat abstract image. Considering the entire dataset, abstract images occurred more frequently in the translation of dog sounds compared to other classes, but non-sense pictures may show up at any time. The most common outputs of this type are artifactual images, of which we show some examples in Figure 12 (second row (a to e)).

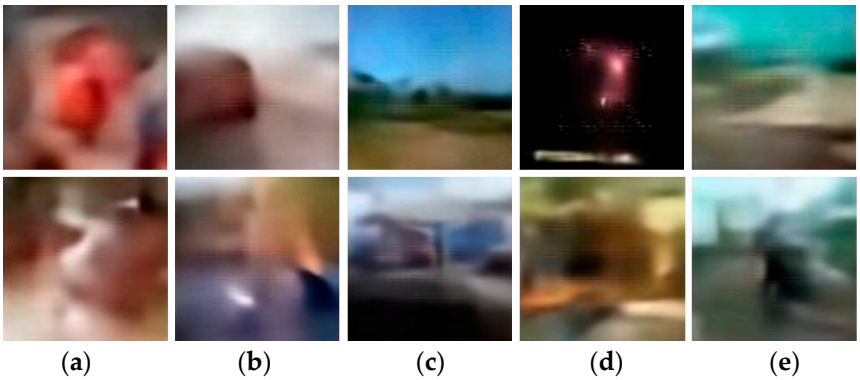

Figure 12. Typical examples of badly translated images from the five sound classes, from left to right column: (**a**) Baby cry, (**b**) Dog, (**c**) Rail transport, (**d**) Fireworks, and (**e**) Water flowing. S2I translation was performed from unknown sounds using a single translator model conditioned on 512-dimension audio embeddings.

Model 'creativity'. As mentioned in Section 3, since the translator must generate images based solely on an input sound, without any visual information about the original scene, we realized that the cross-modal generation performed in S2I translation implies addressing the problem of computational imagination. Despite the fact that our goal was to produce realistic images, there is an inherent 'creativity' required to enable the translator to 'imagine' complete scenes. Interestingly, we observed that the translator occasionally generated rather 'creative' results. While these outcomes were unintended, they were expected to happen, and they are an indicator of some level of 'creativity' achieved, since they demonstrate that the translator was capable of using learned visual features to generate original forms. Some of these 'creative' outputs included surreal, implausible, and creepy images, as illustrated in Figure 13. Sometimes the model's 'creativity' might have compromised the informativity of the image, while in other cases, it just added a bit of fantasy to the picture. With respect to implausible images, although they may occasionally be informative, the presence of awkward elements ends up diverting attention from relevant information. In Figure 13 (first row (a)), we show an example of this, an implausible picture of what seems to be a

baby reading a pink-covered book. Also from the Baby cry sounds, we can see in the image below (second row (a)) a somewhat creepy picture resembling an alien's face. Regarding surreal images, also in Figure 13 (first row (b)), we can see what was supposed to be a dog, since the image was translated from dog sounds. However, due to the shape of the head, it looks more like a fantastical creature with a furry face. Likewise, the image below it (second row (b)), also translated from dog sounds, shows a figure resembling a dark gray dummy sitting on the floor of some indoor environment. When translating train sounds, 'creative' outputs emerged as well. In the same figure (first row (c)), we can observe a depiction resembling a white spaceship landing on a grassy field under a blue sky, or, alternatively, it could be interpreted as a white face with black eyes. The image below it (second row (c)) is also intriguing, showing what seems to be an outdoor scene in daylight with some sort of machinery over a gray pavement. From fireworks sounds, we obtained some colorful compositions, with one that looks more like a sunset scene (first row (d)) and another that is rather abstract, with something on the top that could be a fireworks burst (second row (d)). Lastly, with respect to water sounds, the translator pictured a sort of water-made creature, or potentially a blue fish (first row (e)), while the image below it (2nd row (e)) appears to show a face surrealistically merged with the landscape.

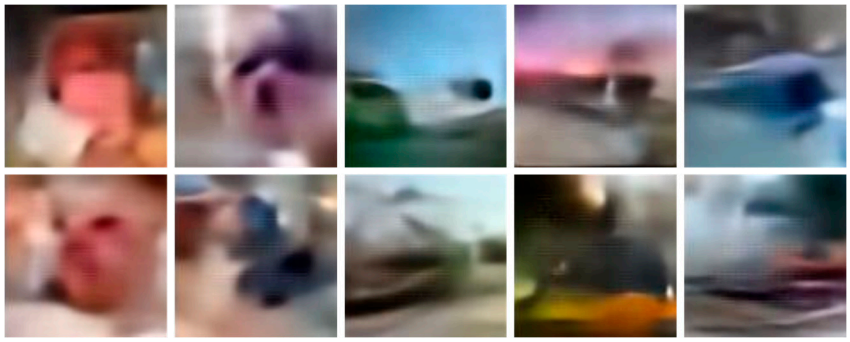

Figure 13. 'Creative' outputs from the five sound classes, from left to right column: (**a**) Baby cry, (**b**) Dog, (**c**) Rail transport, (**d**) Fireworks, and (**e**) Water flowing. S2I translation was performed from unknown sounds using a single translator model conditioned on 512-dimension audio embeddings.

Multi-informative images. Furthermore, the translator demonstrated an ability to produce multi-informative outputs, synthesizing more than one acoustic event within a single image. However, such inferences rarely occurred. Based on our subjective evaluation using 22 different models, the translator generated, on average, two multi-informative images from the entire test set of 6825 sounds. The easiest to spot are the ones that picture people inside the scene. In Figure 14, we show multi-informative images translated from sounds of people and fireworks (first row on top) and from sounds of people and water (second row). These images were translated from unknown sounds using seven different translator models. The scarcity of such images is likely attributable to the fact that the dataset was not originally built for inferring multiple sounds concurrently. As previously explained, the dataset used to train the translator was pre-cleaned to ensure that the sound-related element/event of interest was present both in aural and visual modalities. But, given the diversity of audiovisual content, video segments might contain other acoustic events occurring simultaneously. The problem is that there is no guarantee that these additional acoustic events are present in both modalities of the training data. To prevent any incoherence of this kind, it would be necessary to ensure that all acoustic events of the audio stream had their visual elements represented in the corresponding video frames. However, such an approach could filter the data excessively, making it impossible to train the translator. Nonetheless, the obtained multi-informative images serve as indicators that our strategy of not using class-based losses for training allowed the translator to freely share features and spontaneously generate images spotting the presence of people in the acoustic scenes. The

voice is the most common audible indicator of human presence, whether through speech (Figure 14, first row (a, c, d), second row (a, b, d)), shouting (first row (a), second row (a)), or singing (first row (b)). But in the case of the image in the second row (c), also in Figure 14, human presence was detected by a loud and fussy dog-paddle swim performed by a man in a river. In this instance, a single sound conveyed information about two events: the man swimming and the movement of water.

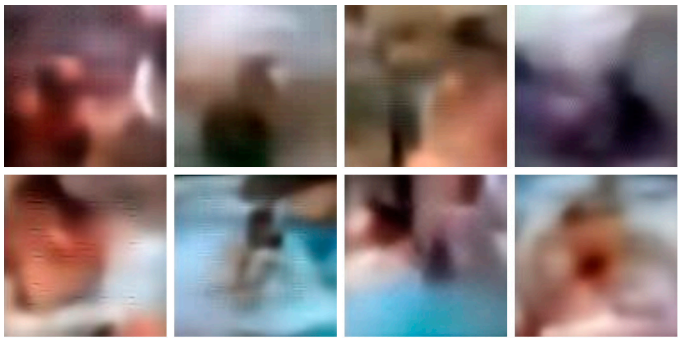

Figure 14. Multi-informative images translated from sounds of people and fireworks (1st row on top), and from sounds of people and water (2nd row). S2I translation was performed from unknown sounds using seven different translator models conditioned on the following audio embedding dimensions: 128 (1st row (**a**,**b**)), 512 (1st row (**c**,**d**) and 2nd row (**a**–**c**)), 1024 (2nd row (**d**)).

6. Conclusions and Future Work

In our exploratory study, we designed, trained, and tested an end-to-end S2I translator with a deep dense generator architecture. We provided detailed information about the model, the heuristics of the approach, and an evaluation of the results from a S2I translation experiment. Additionally, we introduced a solution using informativity classifiers to assess the translator's performance. To the best of our knowledge, this is the first work to address S2I translation without employing supervision or self-supervision for training, and using a dataset characterized by a high degree of audiovisual diversity. Despite the fact that the translator often produced non-informative outputs, our model was capable of generating an average of over 14% interpretable and semantically coherent images. Some of these even exhibited visual structures resembling the input-sound's corresponding images. As discussed throughout the text, the adopted strategy of not using any type of supervision to train the translator ensured that all informative images were necessarily generated through a successful connection between aural and visual modalities. In addition to achieving informativity, the translator was able to produce visually diverse results. We also have found that the translator sometimes produced 'creative' results, picturing original forms, and, less frequently, spontaneously generated multi-informative images. Furthermore, we conducted a performance comparison among five different S2I translator models, varying the dimensionality of the audio embedding space. The results subtly indicated a trade-off between pixel-space convergence and informativity, with better results observed, respectively, for higher and lower feature-space dimensions. We hypothesize that the increased informativity of generated images was due to semantic generalization induced by constraining the flow of information between source and target spaces. However, further studies focusing on quantifying information along the network would be needed to confirm our assumption. Additionally, other explanations for the influence of bottleneck variation on performance must be considered. Apart from the control of information flow, the use of a deeper and denser architecture was decisive for the improvement of the translator's generalization. These solutions, among others which we have presented, allowed us to overcome the problem to a certain extent. We highlight the necessity of further exploring the characteristics of the networks that can interfere in the generalization of GANs applied to cross-modal tasks. We also encourage approaches aiming to optimize

the model using perceptual losses jointly with the adversarial loss and the averaged pixel-wise loss. Moreover, finding feasible solutions to address a broader sonic universe is a key step for taking forward the research on S2I translation.

Author Contributions: Conceptualization, L.A.F. and C.N.; Data curation, L.A.F.; Investigation, L.A.F. and C.N.; Methodology, L.A.F. and C.N.; Software, L.A.F.; Supervision, C.N.; Writing—original draft, L.A.F.; Writing—review and editing, L.A.F. and C.N. All authors have read and agreed to the published version of the manuscript.

Funding: The present work was supported in part by the Brazilian National Council for Scientific and Technological Development (CNPq) under PhD grant 200884/2015-8. Also, the work was partly supported by the Spanish State Research Agency (AEI), project PID2019-107579RB-I00/AEI/10.13039/501100011033.

Data Availability Statement: The datasets analyzed during the current study are available in a repository accessible from https://purl.org/s2i_data (accessed on 27 September 2023). These datasets were derived from the AudioSet dataset, available at https://g.co/audioset (accessed on 27 September 2023).

Acknowledgments: The authors are thankful to Santiago Pascual for his advice on the implementation of GANs. We also thank Josep Pujal for his support in using the computational resources of the Signal Theory and Communications Department at the Polytechnic University of Catalonia (UPC).

Conflicts of Interest: The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ballas, J.A.; Howard, J.H. Interpreting the Language of Environmental Sounds. *Environ. Behav.* **1987**, *19*, 91–114. [CrossRef]
2. Gencoglu, O.; Virtanen, T.; Huttunen, H. Recognition of Acoustic Events Using Deep Neural Networks. In Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, 1–5 September 2014; pp. 506–510.
3. Fanzeres, L.A.; Vivacqua, A.S.; Biscainho, L.W.P. Mobile Sound Recognition for the Deaf and Hard of Hearing. *arXiv* **2018**. [CrossRef]
4. Neubert, A.; Shreve, G.M. *Translation as Text*; Kent State University Press: Kent, OH, USA, 1992; ISBN 978-0-87338-695-1.
5. Barnard, K.; Duygulu, P.; Forsyth, D.; de Freitas, N.; Blei, D.M.; Jordan, M.I. Matching Words and Pictures. *J. Mach. Learn. Res.* **2003**, *3*, 1107–1135.
6. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Object Detectors Emerge in Deep Scene CNNs. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
7. Gonzalez-Garcia, A.; Modolo, D.; Ferrari, V. Do Semantic Parts Emerge in Convolutional Neural Networks? *Int. J. Comput. Vis.* **2018**, *126*, 476–494. [CrossRef]
8. Liang, J.; Jin, Q.; He, X.; Yang, G.; Xu, J.; Li, X. Detecting Semantic Concepts in Consumer Videos Using Audio. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AL, Canada, 15–20 April 2015; pp. 2279–2283.
9. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 2672–2680.
10. Deng, L.; Yu, D. Deep Learning: Methods and Applications. *Found. Trends Signal Process.* **2014**, *7*, 197–387. [CrossRef]
11. Chen, L.; Srivastava, S.; Duan, Z.; Xu, C. Deep Cross-Modal Audio-Visual Generation. In Proceedings of the Thematic Workshops of ACM Multimedia 2017, Mountain View, CA, USA, 23–27 October 2017; Association for Computing Machinery: New York, NY, USA; pp. 349–357.
12. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**. [CrossRef]
13. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-To-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
14. 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 (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
15. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the 4th International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
16. Li, B.; Liu, X.; Dinesh, K.; Duan, Z.; Sharma, G. Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications. *IEEE Trans. Multimed.* **2019**, *21*, 522–535. [CrossRef]
17. Hao, W.; Zhang, Z.; Guan, H. CMCGAN: A Uniform Framework for Cross-Modal Visual-Audio Mutual Generation. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
18. Duan, B.; Wang, W.; Tang, H.; Latapie, H.; Yan, Y. Cascade Attention Guided Residue Learning GAN for Cross-Modal Translation. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milano, Italy, 10–15 January 2021; pp. 1336–1343.

19. Wan, C.; Chuang, S.; Lee, H. Towards Audio to Scene Image Synthesis Using Generative Adversarial Network. In Proceedings of the ICASSP 2019–IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 496–500.
20. Aytar, Y.; Vondrick, C.; Torralba, A. SoundNet: Learning Sound Representations from Unlabeled Video. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Curran Associates Inc.: Red Hook, NY, USA, 2016; pp. 892–900.
21. Yang, P.-T.; Su, F.-G.; Wang, Y.-C.F. Diverse Audio-to-Image Generation via Semantics and Feature Consistency. In Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 7–10 December 2020; pp. 1188–1192.
22. Duarte, A.; Roldan, F.; Tubau, M.; Escur, J.; Pascual, S.; Salvador, A.; Mohedano, E.; McGuinness, K.; Torres, J.; Giro-i-Nieto, X. Wav2Pix: Speech-Conditioned Face Generation Using Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; Volume 3.
23. Oh, T.-H.; Dekel, T.; Kim, C.; Mosseri, I.; Freeman, W.T.; Rubinstein, M.; Matusik, W. Speech2Face: Learning the Face Behind a Voice. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 5–20 June 2019; pp. 7539–7548.
24. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep Face Recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*; British Machine Vision Association: Durham, UK, 2015.
25. Cole, F.; Belanger, D.; Krishnan, D.; Sarna, A.; Mosseri, I.; Freeman, W.T. Synthesizing Normalized Faces from Facial Identity Features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3386–3395.
26. Chatterjee, M.; Cherian, A. Sound2Sight: Generating Visual Dynamics from Sound and Context. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 701–719.
27. Shim, J.Y.; Kim, J.; Kim, J.-K. S2I-Bird: Sound-to-Image Generation of Bird Species Using Generative Adversarial Networks. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 2226–2232.
28. Hao, W.; Han, M.; Li, S.; Li, F. An Attention Enhanced Cross-Modal Image–Sound Mutual Generation Model for Birds. *Comput. J.* **2021**, *65*, bxaa188. [CrossRef]
29. Sanguineti, V.; Thakur, S.; Morerio, P.; Del Bue, A.; Murino, V. Audio-Visual Inpainting: Reconstructing Missing Visual Information with Sound. In Proceedings of the ICASSP 2023–IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes, Greece, 4–10 June 2023; pp. 1–5.
30. Van den Oord, A.; Kalchbrenner, N.; Vinyals, O.; Espeholt, L.; Graves, A.; Kavukcuoglu, K. Conditional Image Generation with PixelCNN Decoders. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Curran Associates Inc.: Red Hook, NY, USA; pp. 4797–4805.
31. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780.
32. Chen, H.; Xie, W.; Vedaldi, A.; Zisserman, A. Vggsound: A Large-Scale Audio-Visual Dataset. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 721–725.
33. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**. [CrossRef]
34. Sung-Bin, K.; Senocak, A.; Ha, H.; Owens, A.; Oh, T.-H. Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment. *arXiv* **2023**. [CrossRef]
35. Zhou, Y.; Wang, Z.; Fang, C.; Bui, T.; Berg, T.L. Visual to Sound: Generating Natural Sound for Videos in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3550–3558.
36. Zhu, H.; Luo, M.-D.; Wang, R.; Zheng, A.-H.; He, R. Deep Audio-Visual Learning: A Survey. *Int. J. Autom. Comput.* **2021**, *18*, 351–376. [CrossRef]
37. Vilaça, L.; Yu, Y.; Viana, P. Recent Advances and Challenges in Deep Audio-Visual Correlation Learning. *arXiv* **2022**. [CrossRef]
38. Mahadevan, S. Imagination Machines: A New Challenge for Artificial Intelligence. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
39. Davies, J. Artificial Intelligence and Imagination. In *The Cambridge Handbook of the Imagination*; Abraham, A., Ed.; Cambridge University Press: Cambridge, UK, 2020; pp. 162–171. ISBN 978-1-108-65929-1.
40. Stevenson, L.F. Twelve Conceptions of Imagination. *Br. J. Aesthet.* **2003**, *43*, 238–259. [CrossRef]
41. Beaney, M. *Imagination and Creativity*; Open University Worldwide: Milton Keynes, UK, 2010; ISBN 978-0-7492-1735-8.
42. Pereira, F.C.; Cardoso, A. *Conceptual Blending and the Quest for the Holy Creative Process*; ResearchGate: Lyon, France, 2002.
43. Guilford, J.P. *The Nature of Human Intelligence*, 1st ed.; McGraw-Hill: New York, NY, USA, 1967; ISBN 978-0-07-025135-9.
44. Gabora, L. Reframing Convergent and Divergent Thought for the 21st Century. In Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci), Montreal, QC, Canada, 24–27 July 2019; pp. 1794–1800.
45. Mescheder, L.; Nowozin, S.; Geiger, A. The Numerics of GANs. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Long Beach, CA, USA; pp. 1823–1833.

46. Mescheder, L.; Geiger, A.; Nowozin, S. Which Training Methods for GANs Do Actually Converge? In Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–13 July 2018; Volume 80, pp. 3481–3490.
47. Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward Multimodal Image-to-Image Translation. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA; pp. 465–476.
48. Mao, Q.; Lee, H.-Y.; Tseng, H.-Y.; Ma, S.; Yang, M.-H. Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1429–1437.
49. Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; Abbeel, P. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 23–30.
50. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2018; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
51. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-Attention Generative Adversarial Networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 10–15 June 2019; pp. 7354–7363.
52. Pearl, J. *Causality: Models, Reasoning and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2009; ISBN 978-0-521-89560-6.
53. Walker, C.M.; Gopnik, A. Causality and Imagination. In *The Oxford Handbook of the Development of Imagination*; Taylor, M., Ed.; Oxford University Press: Oxford, UK, 2013; pp. 342–358. ISBN 978-0-19-998303-2.
54. Stewart, R.; Ermon, S. Label-Free Supervision of Neural Networks with Physics and Domain Knowledge. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 2576–2582.
55. Grzeszick, R.; Plinge, A.; Fink, G.A. Bag-of-Features Methods for Acoustic Event Detection and Classification. *IEEEACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1242–1252. [CrossRef]
56. Kato, H.; Harada, T. Image Reconstruction from Bag-of-Visual-Words. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 955–962.
57. Kavalerov, I.; Wisdom, S.; Erdogan, H.; Patton, B.; Wilson, K.; Le Roux, J.; Hershey, J.R. Universal Sound Separation. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; pp. 175–179.
58. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
59. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
60. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the International Conference on Machine Learning, PMLR, New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
61. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
62. Glorot, X.; Bordes, A.; Bengio, Y. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In Proceedings of the 28th International Conference on Machine Learning, Omnipress, Madison, WI, USA, 28 June–2 July 2011; pp. 513–520.
63. Lucas, T.; Tallec, C.; Ollivier, Y.; Verbeek, J. Mixed Batches and Symmetric Discriminators for GAN Training. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2844–2853.
64. Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
65. Odena, A.; Dumoulin, V.; Olah, C. Deconvolution and Checkerboard Artifacts. *Distill* **2016**, *1*, e3. [CrossRef]
66. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep Variational Information Bottleneck. *arXiv* **2017**, arXiv:1612.00410.
67. Jeon, I.; Lee, W.; Kim, G. IB-GAN: Disentangled Representation Learning with Information Bottleneck GAN. 2018. Available online: https://openreview.net/pdf?id=ryljV2A5KX (accessed on 27 September 2023).
68. Peng, X.B.; Kanazawa, A.; Toyer, S.; Abbeel, P.; Levine, S. Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow. *arXiv* **2020**, arXiv:1810.00821.
69. Luo, Y.; Liu, P.; Guan, T.; Yu, J.; Yang, Y. Significance-Aware Information Bottleneck for Domain Adaptive Semantic Segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6778–6787.
70. Song, Y.; Yu, L.; Cao, Z.; Zhou, Z.; Shen, J.; Shao, S.; Zhang, W.; Yu, Y. Improving Unsupervised Domain Adaptation with Variational Information Bottleneck. In Proceedings of the 24th European Conference on Artificial Intelligence, Santiago de Compostela, Spain, 29 August–8 September 2020; IOS Press: Amsterdam, The Netherlands, 2020; Volume 325, pp. 1499–1506.

71. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
72. Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA; pp. 5580–5590.
73. Sedai, S.; Antony, B.; Mahapatra, D.; Garnavi, R. Joint Segmentation and Uncertainty Visualization of Retinal Layers in Optical Coherence Tomography Images Using Bayesian Deep Learning. In *Computational Pathology and Ophthalmic Medical Image Analysis, Proceedings of the 1st International Workshop, COMPAY 2018, Granada, Spain, 16–20 September 2018*; Stoyanov, D., Taylor, Z., Ciompi, F., Xu, Y., Martel, A., Maier-Hein, L., Rajpoot, N., van der Laak, J., Veta, M., McKenna, S., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 219–227.