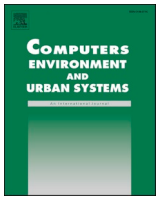
Computers, Environment and Urban Systems 110 (2024) 102122

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



From hearing to seeing: Linking auditory and visual place perceptions with soundscape-to-image generative artificial intelligence

Yonggai Zhuang [a], Yuhao Kang [b,c,f,*], Teng Fei [a,e,**], Meng Bian [d], Yunyan Du [e]

[a] School of Resource and Environmental Sciences, Wuhan University, Wuhan, China
[b] GISense Lab, Department of Geography, University of South Carolina, USA
[c] Center for GIScience and Geospatial Big Data, University of South Carolina, USA
[d] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
[e] State Key Laboratory of Resources and Environmental Information System, Beijing, China
[f] Department of Geography and the Environment, University of Texas at Austin, USA

ARTICLE INFO

Keywords:
Soundscape
Street view images
Sense of place
Stable diffusion
Generative AI
LLMs

ABSTRACT

People experience the world through multiple senses simultaneously, contributing to our sense of place. Prior quantitative geography studies have mostly emphasized human visual perceptions, neglecting human auditory perceptions at place due to the challenges in characterizing the acoustic environment vividly. Also, few studies have synthesized the two-dimensional (auditory and visual) perceptions in understanding human sense of place. To bridge these gaps, we propose a Soundscape-to-Image Diffusion model, a generative Artificial Intelligence (AI) model supported by Large Language Models (LLMs), aiming to visualize soundscapes through the generation of street view images. By creating audio-image pairs, acoustic environments are first represented as high-dimensional semantic audio vectors. Our proposed Soundscape-to-Image Diffusion model, which contains a Low-Resolution Diffusion Model and a Super-Resolution Diffusion Model, can then translate those semantic audio vectors into visual representations of place effectively. We evaluated our proposed model by using both machine-based and human-centered approaches. We proved that the generated street view images align with our common perceptions, and accurately create several key street elements of the original soundscapes. It also demonstrates that soundscapes provide sufficient visual information places. This study stands at the forefront of the intersection between generative AI and human geography, demonstrating how human multi-sensory experiences can be linked. We aim to enrich geospatial data science and AI studies with human experiences. It has the potential to inform multiple domains such as human geography, environmental psychology, and urban design and planning, as well as advancing our knowledge of human-environment relationships.

1. Introduction

Sense of place, a fundamental concept in geography, reflects complex interactions between humans and their environment (Agnew, 2011; Cresswell, 2014; Martin, 2017). Sense of place emphasizes a multidimensional relationship that extends beyond mere physical presence, to include human subjective experiences and emotional connections attached to place (Kang, Gao, & Roth, 2019; Kang, Jia, et al., 2019; Manzo, 2003). The various sensory stimuli from our surrounding environments play an important role in shaping our perceptions and evoking human emotions, contributing to the development of a personal and affective sense of place (Raymond et al., 2017; Schreuder et al., 2016; Spence, 2020). As demonstrated in Tuan's seminal work *Topophilia (*Tuan, 1974*)*, human sense of place is informed by multiple dimensions such as visual cues, auditory environment, tactile sensations, and smell. Notably, visual and auditory perceptions are key drivers in our interactions with the surrounding environment (Pocock et al., 1994; Tuan, 1975; Yildirim & Arefi, 2022).

Visual perception, in particular, acts as one of the primary ways we engage with our environments, allowing us to observe details of our environments. Our eyes become the windows to read the stories of the world. We gaze at landscapes, observe architectural details, and watch

* Corresponding author at: GISense Lab, Department of Geography, University of South Carolina, USA.
** Corresponding author at: School of Resource and Environmental Sciences, Wuhan University, Wuhan, China.
E-mail addresses: 2019301030116@whu.edu.cn (Y. Zhuang), yuhaokang@sc.edu (Y. Kang), feiteng@whu.edu.cn (T. Fei), bian@whu.edu.cn (M. Bian), duyy@lreis.ac.cn (Y. Du).

https://doi.org/10.1016/j.compenvurbsys.2024.102122
Received 4 December 2023; Received in revised form 14 March 2024; Accepted 23 April 2024
Available online 1 May 2024

life's myriad activities unfold in different places (Fischer & Whitney, 2014). Through this visual lens, we often form our initial impressions and long-lasting memories associated with a specific place (Cosgrove, 2003; Michael et al., 2019; Smith, 2015). Given its importance, researchers have made efforts to measure human visual perceptions. Several quantitative approaches and technologies have been developed, such as photographic analyses (Dunkel, 2015) and landscape evaluations (Clement et al., 2013; Hong et al., 2019; Jeon & In Jo, 2020). Recently, with the emergence of geospatial big data and advanced computer vision approaches, researchers have leveraged street view images (Kang et al., 2021; Kang et al., 2023), remote sensing images (Xu et al., 2021; Wu et al., 2023), and user-generated content (UGC) (Bhandari et al., 2019; Gao et al., 2021; Wilson & Soranzo, 2015) for mapping and analyzing visual elements from photos at places. Leveraging these approaches enables us to better understand how individuals visually perceive a space, which could further inform urban design decision-making to ensure that spaces are both functional and aesthetically pleasing (Åsa Ode Sang, Knez, Gunnarsson, & Hedblom, 2016).

Similarly, the sounding environment, or soundscape, offers an auditory experience, capturing the essence of a place through its unique blend of natural (e.g., bird calls or rustling leaves) and anthropogenic sounds (e.g., traffic or conversations) (Brooks et al., 2014; Jeon & Hong, 2015; Schafer, 1977). This auditory dimension, while sometimes subtle, may significantly influence our emotional responses, evoking a wide range of feelings from serenity to dissonance (Cain et al., 2013). Auditory experiences, such as rolling thunder and whistling wind, tend to touch us more and trigger stronger emotional reactions than visual perceptions (Tuan, 1974). These experiences influence our daily perceptions and interactions with places (Pijanowski et al., 2011). Beyond the immediate auditory experience, soundscapes are significantly associated with a variety of health outcomes. On the one hand, persistent noise pollution, for instance, has been linked to a range of health concerns, from sleep disorders to cardiovascular problems (In Jo & Jeon, 2020; Schulte-Fortkamp & Fiebig, 2006). On the other hand, the quality of a soundscape can influence mental health. Prior studies have demonstrated that serene natural sounds are often associated with relaxation and improved well-being, while chaotic, jarring noises can heighten stress and anxiety (Buxton et al., 2021; Watts et al., 2016). Recognizing these discoveries, scholars have employed various methods to measure soundscapes such as spectrograms and power spectrums. These techniques have been utilized to capture the frequency, intensity, energy, and temporal characteristics of the acoustic environment (Holmes et al., 1971; Oppenheim & Schafer, 2004b; Prestigiacomo, 1962).

Despite advancements in measuring and understanding human visual and auditory perceptions, we identify two significant research gaps in the existing literature. First, it remains an ongoing challenge to vividly characterize the acoustic environment to understand human auditory experiences at place. While researchers have leveraged several methods such as spectrograms (Marchegiani & Posner, 2017), power spectrums (Arnal et al., 2015), and acoustic signal attributes (Fuller et al., 2015; Hopkins et al., 2022) to depict soundscapes, their lack of intuitiveness often limits their effectiveness, especially when compared with those methods for visual perceptions. Such methods, though technically accurate, may not engage with our humans effectively to provide a comprehensive understanding of the inherent characteristics of the soundscape environment. The second gap refers to the lack of synergistic consideration of the multidimensionality of the sense of place. Echoing Tuan's insights about the holistic nature of human perceptions, we experience the world through all our senses simultaneously (Tuan, 1974). Multiple dimensions of the human senses contribute uniquely to our experience and enrich our human experience of place. While the significance of various sensory dimensions has been well-acknowledged, most existing GIS and geospatial data science studies have focused on a singular dimension of the human sense of place, such as vision (Landeschi et al., 2016; Minelli et al., 2014; Wróżyński et al., 2016), or hearing (Bilaşco et al., 2017; Keyel et al., 2017; Primeau & Witt, 2018). Few studies have made efforts to link the two dimensions together to provide a more comprehensive understanding of the human sense of place, such as measuring noise levels from visually perceived street view images (Huang et al., 2023; Zhao et al., 2023). This narrow focus overlooks the complexity and depth of human-environment interactions, which are inherently shaped by multi-sensory experiences. Therefore, exploring the interplay between the visual and auditory senses is crucial to offer insights into how we experience and connect with our surrounding environments.

To bridge the two research gaps, we propose a novel computational framework that synthesizes auditory and visual perceptions by translating soundscapes into visual representations of place. We aim to explore the potential of visualizing acoustic environments as place images. Specifically, can we create visual imagery from auditory soundscapes, i.e., can we visualize the place we hear? The advancement in geospatial big data and geospatial artificial intelligence (GeoAI) presents new opportunities to visualize soundscapes and associate the two dimensions of sensory experiences, which could significantly advance our knowledge of human-environmental relationships (Janowicz et al., 2020; Kang et al., 2024). Stable Diffusion, a cutting-edge deep learning technique, has stood out due to its widely recognized image generation capabilities (Moliner et al., 2023; Yang et al., 2023). Stable Diffusion could translate complex datasets into visual representations. It has been applied in various research areas, demonstrating its utility in generating and enhancing vision-based image data. Utilizing Stable Diffusion offers new opportunities to effectively translate auditory soundscapes into visual street view images, providing a more intuitive way to characterize soundscapes of places, thereby bridging the first research gap. Furthermore, by associating auditory soundscapes with visual street view images, our study could provide a holistic depiction of human multi-sensory experiences at place. By doing so, it not only addresses the second research gap but also provides new insights into the multifaceted nature of human-environment relationships.

To this end, we design a computational framework that converts street soundscapes into visual representations of street view images by developing an advanced Soundscape-to-Image Diffusion model. Our study is driven by the following two research questions:

(1) How to visualize soundscape with the Soundscape-to-Image Diffusion model to depict human multi-sensory experiences at place?
(2) To what extent do soundscapes provide visual cues sufficient for generating recognizable streetscape images that accurately depict places?

Our computational framework contains four steps: First, we collected videos taken on streets that contain both visual and auditory elements. Then, we processed and formatted the data to create a multi-sensory dataset. After that, we fed such a dataset to train a Soundscape-to-Image Diffusion model capable of associating visual and auditory information; Once trained, the model could generate visual images of places based solely on auditory inputs. Finally, to ensure the robustness and accuracy of our framework, we conducted four distinct validation tasks to evaluate the performance and reliability of the model including both machine-based and human-centered evaluations.

The major contributions and innovations of this study are threefolds: First, our study adds to the existing literature about the sense of place in environmental psychology and human geography by combining human auditory and visual interactions with the environment. Traditionally, the two dimensions of human experiences at place were often studied separately. Our study emphasizes the simultaneous impact of both dimensions in fostering a human sense of place. Second, we developed a novel computational framework and a cutting-edge Soundscape-to-Image Diffusion model that explicitly links the auditory

and visual perceptions. Using such a framework, we may understand how auditory cues contribute to our visual perceptions of places. Third, this study has broader implications for multiple domains. The results may enhance our knowledge of the impacts of visual and auditory perceptions on human mental health, may guide urban design practices for place-making, and may improve the overall quality of life in communities.

2. Literature review

2.1. Visualizing soundscape

Researchers have developed a variety of approaches to visualize soundscapes, aiming to translate acoustic information into visual formats. Approaches for soundscape visualization can be primarily classified into two categories: *comprehensive representation* and *component-based representation*. Comprehensive representation focuses on presenting audio files in an intuitive graphical form rather than specific sound-emitting objects or sound spaces in a soundscape. The comprehensive representation approaches aim to visualize certain aspects of environmental audio files, such as waveforms (Phillips et al., 2017), spectrograms (Prestigiacomo, 1962), power spectrums (Holmes et al., 1971), feature vector plots (Park et al., 2014), and frequency analyses (Oppenheim & Schafer, 2004a). These diagrams can exhibit certain features of the soundscape.

In contrast, component-based representation methodologies aim to map soundscapes as well as their semantic information with appropriate classifications and symbology (Aiello et al., 2016; Sacchelli & Favaro, 2019). A significant challenge for soundscape visualization refers to the effective extraction and clear presentation of semantic information, as well as the environmental characteristics of the soundscape. As environmental soundscapes may vary significantly due to their unique acoustic elements and characteristics at different places (Lillis et al., 2014). Given these obstacles, it is challenging to convert the acoustic environment to visual formats, making it complex to account for multisensory human experiences at places.

2.2. Stable diffusion

The emergence of Stable Diffusion offers new opportunities to link auditory and visual data. Stable Diffusion is a probabilistic generative model that generates structured data like images based on input prompts (e.g., texts, audio) using a noise injection method (Ho et al., 2020). It outperforms several traditional deep learning models such as the generative adversarial networks (GANs) (Goodfellow et al., 2014; Kang, Gao, & Roth, 2019; Ramesh et al., 2022). Stable Diffusion has been widely used for cross-modal data generation, especially in the Text-to-Image domain where it converts text descriptions to corresponding images (Nichol et al., 2021). Several notable large language models (LLMs) such as Imagen (Saharia et al., 2022a, b), Midjourney (Oppenlaender, 2022), and DALL-E2 (Ramesh et al., 2022) incorporate Stable Diffusion as a crucial component in their architecture for image generation from texts (Saharia et al., 2022a, b). Building upon the foundation of image generation, Stable Diffusion has further expanded its generative capabilities from images to videos (Chen et al., 2023), such as medical video generation and animated video generation (Reynaud et al., 2023). Furthermore, Stable Diffusion has extended its application for the Text-to-Music domain, such as MusicLM and AudioLDM (Agostinelli et al., 2023). All these advancements in Stable Diffusion highlighted its potential for multimodal data generation, and thereby we select it as our method for connecting audio information with visual information by generating street view images from auditory soundscapes. Inspired by the ImageBind model (Girdhar et al., 2023) developed by Meta, we recognize that advanced audio models can capture visual scene-related information. Therefore, we leverage Stable Diffusion to generate images that resemble real data from auditory soundscapes in this paper (Girdhar et al., 2023; Han et al., 2023).

It is worth mentioning that we are among the first to develop the Soundscape-to-Image Diffusion model. Our model does not directly rely on any existing open-source Sound-to-Image model. The rationale behind this is that most existing Sound-to-Image models focus on Music-to-Image, while paying limited attention to mapping soundscape to street view images (Williams et al., 2023; Wu et al., 2021).

3. Data and method

In this study, we propose a novel computational framework for generating street view images based on acoustic environment collected at streets. The proposed method comprises four key components: (1) data collection; (2) soundscape feature processing; (3) a Soundscape-to-Image Diffusion model, and (4) evaluation, as illustrated in Fig. 1. We first collected both street auditory soundscapes and street view images. Then, we processed audio files as high-dimensional vectors that encode their semantics so that they can be fed into the Stable Diffusion model. After that, we developed our Soundscape-to-Image Diffusion that links acoustic and visual perceptions together to generate street view images from soundscapes. After training this Soundsccape-to-Image Diffusion model, we evaluated its performance based on the validation data with both machine-based and human-centered metrics.

3.1. Dataset

The data utilized in this research were downloaded from publicly available street view videos on YouTube. We conducted searches using keywords such as “street walk,” “city walk,” and “village walk,” and manually screened the videos found based on these keywords. Only videos that have clear street views and clear street sounds (i.e., no background music) were selected in our study. We aim to build a diverse dataset around the theme of streetscapes as the style of the videos may exhibit regional and cultural differences. The collected videos capture a wide range of real-world sounds and scenes from diverse regions, including Asia, North America, and Europe, featuring metropolitan areas like Hong Kong, Oslo, London, Tokyo, as well as some smaller cities in the United States. Notably, the street view scenes exhibit complexity and diversity in terms of color and objects. By collecting videos from these varied regions with diverse objects and scenes, we aim to illustrate the potential generalizability of the proposed model and highlight the significance of the research findings. In total, the dataset contains 1,667 min of content from 102 videos, offering the data foundation for training the Soundscape-to-Image Diffusion model.

After that, we segmented all videos into multiple 10-s video clips to create audio-image pairings. Within each clip, we extracted a 10-s audio segment, saving it in a .wav format. Given that audio clips are streamed content, as opposed to static images, we strategically paired these audio segments with images at the midpoint – the 5th second – of each clip. The rationale behind the 10-s duration is that we aim to achieve a balance between content diversity and consistency. A 10-s audio clip provides sufficient auditory information about the street environment, and the visual scene within such a short timeframe remains relatively consistent. By doing so, all audio-image pairs could be fed in our Soundscape-to-Image Diffusion model with a standardized format. Upon pairing all video clips, a total of 10,000 audio-image pairs were generated, contributing to the dataset used for the study. After creating all audio-image pairings, we extracted 90% of all audio-image pairs as the training set, and the remaining 10% of audio-image pairs as the validation set for performance evaluation. Then, we input the 10-s audio .wav files to the audio encoder of the Soundscape-to-Image Diffusion model, and the corresponding street view images were used as the training label to guide the model to generate street view images.

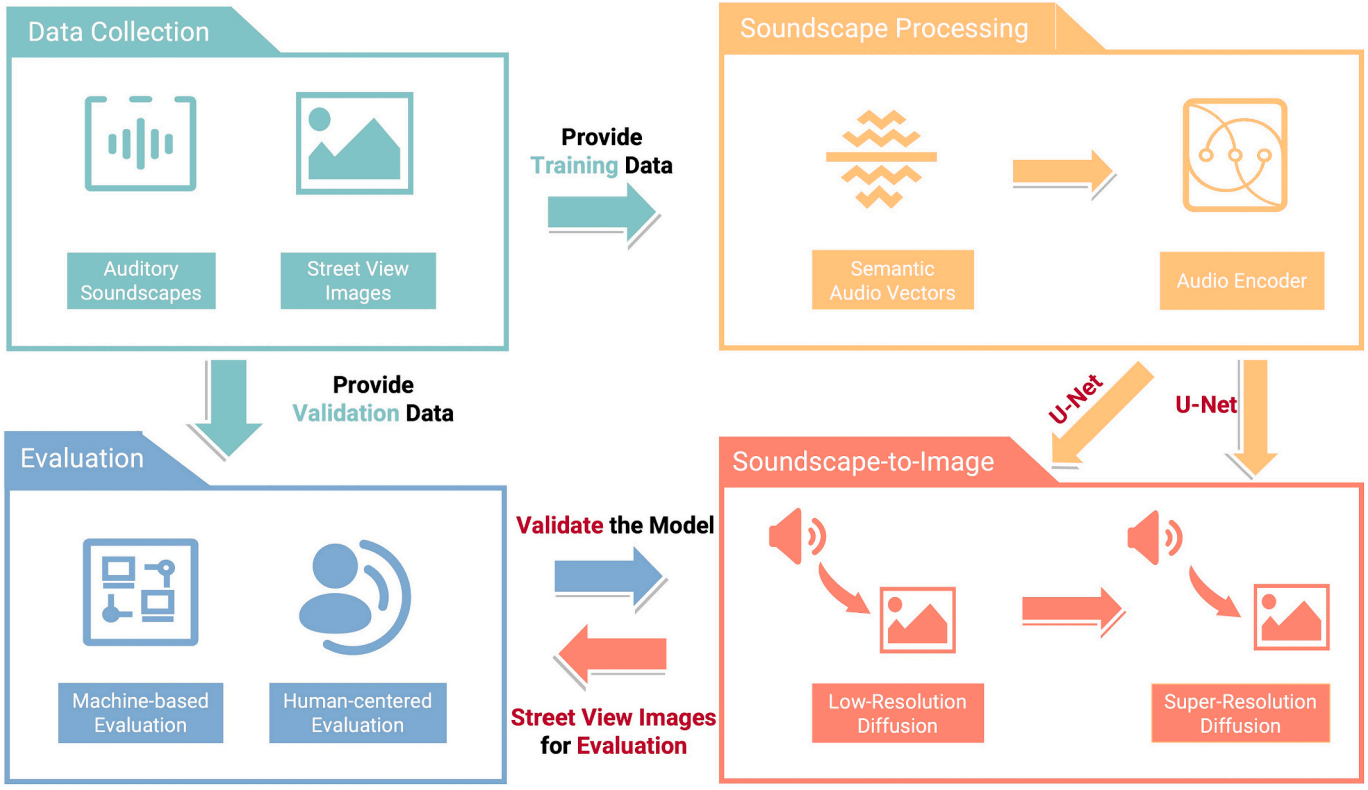


Fig. 1. The computational framework of this study.

3.2. Soundscape feature processing

After building the training and testing dataset of the audio-image pairs, we preprocessed and transformed the original street soundscape data into high-dimensional semantic audio vectors. Such audio vectors will be further utilized to guide the street view image generation in the next stage. To do so, we utilized an audio encoder designed to process each 10-s stereo audio .wav file at a 44.1 kHz sampling rate. It then employed a short-time Fourier transformation using a Hamming window (size 520, hop size 160 samples) to capture the frequency and phase of sine waves in localized temporal regions. Subsequently, a Mel-filter was applied to derive the Mel-spectrum in each Hamming window, enhancing the representation of relevant spectral features. Such processing steps have been widely employed in prior studies with their effectiveness thoroughly validated (Eronen et al., 2006).

Following preprocessing, the audio vector was input into a convolutional neural network (CNN) to transform the initial 64 frequency bands at each time step into a 768-dimensional feature representation (as illustrated in Fig. 1). Such high-dimensional audio vectors capture detailed street soundscapes from both semantic and locational aspects. These audio vectors then serve as the input for the Soundscape-to-Image Diffusion model to generate corresponding street view images.

3.3. Soundscape-to-image diffusion model

Based on the high-dimensional semantic audio vectors, we further trained a Soundscape-to-Image Diffusion model. The goal of such a model is to produce high-resolution street view images from high-dimensional audio vectors, associating human auditory and visual experiences. To achieve this, we developed our Soundscape-to-Image Diffusion model by adopting Imagen, a cutting-edge image generation model.

Stable Diffusion has been widely used for generating imagery data. It often requires a set of prompts (text, audio prompts, or other modalities) as inputs to perform supervised data generation. These prompts are then transformed into high-dimensional feature vectors by a text encoder to guide the noise-generation process (Lee et al., 2023). The model initially generates a set of random values with the same dimensions as the output data. Guided by the prompts, systematic noises are iteratively added to the initial random data through multiple iterations to generate real data. The training process is based on the inverse process of adding noise, utilizing training samples (such as images or audio clips) that include prompts and labels. The model learns the probability distribution of each noise step in reverse from the training data labels. These labels guide the model to identify and learn the noise patterns associated with the prompts (Rombach et al., 2021). Such models are commonly used to generate superior-quality images and text by mastering the reverse diffusion process.

Our Soundscape-to-Image Diffusion was developed building upon Imagen (Saharia et al., 2022a, b), which is an advanced Text-to-Image Stable Diffusion model developed by Google. As a diffusion model, it consists of: (1) a Text Encoder that maps text to a sequence of embeddings, represented as high-dimensional vectors, and (2) a cascading of conditional diffusion models including a Text-to-Image Diffusion Model and two Super-Resolution Diffusion Models that map image embeddings to progressively higher resolutions. The Text-to-Image diffusion model initially generates random noise images. It then utilizes a UNet network to transform the high-dimensional text vectors into structured noise vectors. Such noise vectors are continuously appended to the noise images, gradually generating low-resolution images after multiple iterations. Following this, the Super-Resolution Diffusion Model further appends structured noise derived from those high-dimensional vectors to the low-resolution images. It then progressively upgrades the low-resolution images into high-resolution images, demonstrating its ability to translate complex textual descriptions into vivid visual representations. More details on Imagen can be referred to Saharia et al. (2022a, 2022b).

In this study, we developed our Soundscape-to-Image Diffusion

model by extending the Imagen model. We first designed an audio encoder that processes the high-dimensional audio vectors mentioned in Section 3.2 to replace the original Text Encoder. Our Soundscape-to-Image Diffusion contains a cascading of conditional diffusion models including a Low-Resolution Diffusion Model and a Super-Resolution Diffusion Model. The Low-Resolution Diffusion Model first generates a random 64 × 64 noise image based on the 768-dimensional audio vector. Subsequently, it uses a UNet structure to transform the audio vector into a 64 × 64-dimensional noise vector. This noise vector is then appended to the noise image 256 times, ultimately transforming the random noise image into a 64 × 64 low-quality image. Then, the Super-Resolution Diffusion Model begins by generating a random 256 × 256 noise image. It combines the 64 × 64 low-quality image with the audio vector and feeds them into another UNet network, generating a 256 × 256-dimensional noise vector. This noise vector is then appended to the noise image 256 times (with a stride of 256), ultimately transforming the random noise image into a 256 × 256 high-resolution street view image. Notably, different from the original structure, we leveraged one Super-Resolution Diffusion Model to generate high-resolution images aiming for optimizing computational resources and efficiency. It is worth noting that the UNet network plays an important role in feature vector conversion. By utilizing the UNet network, semantic audio vectors can be projected from the semantic auditory space to the semantic visual space. It facilitates the alignment of audio segments with corresponding images in high-dimensional space, thereby guiding the stable diffusion network to generate street view images for further evaluation. In sum, once trained, the proposed Soundscape-to-Image Diffusion model could leverage the high-dimensional auditory soundscape data to generate corresponding street view images. The code of our model is openly available on GitHub at: https://github.com/GISense/Soundscape-to-Image.

3.4. Evaluation

After training the model, we evaluated the performance of our model by associating auditory and visual experiences at place based on the constructed dataset. Prior studies in image generation have faced challenges in evaluating the quality of generated samples, mainly due to the significant differences between the generated and original images. Similarly, this study encounters difficulties in evaluating the quality of the generated images as well. A key question here is to determine whether the generated images are visually similar to the ground truth data. To comprehensively evaluate the image synthesis performance of our Soundscape-to-Image Diffusion model, we conducted two types of assessments including both machine-based methods as well as human-centered assessments.

Regarding machine-based methods for evaluating the model performance, we devoted our efforts from two perspectives. First, we assessed and computed a loss function between the training and validation dataset. The Low-Resolution Stable Diffusion Model and Super-Resolution Diffusion Model in our Soundscape-to-Image Diffusion model were trained separately. Both diffusion models were trained using 90% of the audio-image pairs in the dataset mentioned earlier. During the training process of these two diffusion models, we computed the Mean Squared Error (MSE) loss between the predicted image and the real street view image. After a certain number of epochs, when the loss converged, we terminated the training process and saved the model weight parameters. We then used the remaining 10% of the audio-image pairs in the dataset as validation data for evaluation.

Second, we evaluated the image similarity at the street element level. The rationale behind this was to determine to what extent objects in the real-world images were captured from auditory clips and appear in the generated images as well. In particular, we focused on three key street elements including greenery, buildings, and sky, as they have been extensively studied in prior studies (Ren et al., 2023; D'Alessandro et al., 2018; Tan et al., 2022). To do so, we randomly selected 100 different video clips from the dataset, and compared the objects in both simulated and in-situ images from three aspects: greenery, buildings, and sky. We utilized a DeepLabV3 semantic segmentation model (Salih Can Yurtkulu, Yusuf Hüseyin Şahin, & Gozde Unal, 2019), pre-trained on the ADE20K dataset (Zhou et al., 2019) to perform image segmentation, and computed the proportions of greenery, buildings, and sky pixels in the images (See Fig. 2). After that, we calculated correlation coefficients between the original and generated images to measure their similarities. Despite that these two machine-based evaluating approaches could provide insights into quantifying the accuracy of the image generation process, they may not capture all aspects of image quality and coherence due to the subjectivity of human perceptions and experiences.

We equally value the importance of human evaluation in understanding human subjective auditory and visual perceptions. By keeping human-in-the-loop, we approach it from two perspectives including both qualitative and quantitative methods. First, we manually selected three groups of places with significantly different characteristics: urban vs. rural settings, areas with abundant greenery vs. sparse greenery, and places with clear vs. obstructed sky views. We believe that a variety of factors such as the degree of urbanization, the presence of trees, and the openness of the place are crucial elements in influencing the visual and auditory settings of a place. If our model can successfully generate distinct images that accurately reflect the auditory cues as well as place characteristics, it indicates that the developed Soundscape-to-Image Diffusion model could effectively link what we hear to what we see. Thus, we selected a small subset of street soundscape samples and generated five street view images for each audio clip. Through a qualitative comparison between the generated street view images and the actual street view images, we may estimate the model performance. It should be acknowledged that Stable Diffusion models typically produce varied outputs even with the same inputs. Thus, we generated five images for each audio clip to enhance the robustness of our results.

Second, we invited participants to provide quantitative feedback on the effectiveness of our method. Participants were presented with 10 audio segments, each randomly selected from the validation data. After listening to each audio segment, they were asked to choose the image that best reflected the audio from a set of three generated images. Among these three images, one was generated based on the given audio using our Soundscape-to-Image Diffusion model, while the other two were generated from audio in different categories of places. In total, we received responses from 22 individuals. By examining whether their selections align with the actual audio content, we could validate our model from a human-centered perspective.

4. Results

4.1. Street view image generation

Fig. 3 shows several example street view images generated based on different auditory clips using the proposed Soundscape-to-Image Diffusion model. We manually classified places into three categories based on their characteristics. Each group consists of two contrasting places with very different visual/soundscape characteristics: (a) urban vs. rural settings; (b) areas with high greenery vs. less greenery; (c) places with high vs. low percentages of sky views. Despite the fact that some areas and pixels in the generated images are blurred, we believe that these images capture similar visual characteristics to those of the original images. In the following sections, we perform multiple approaches to evaluate the model performances.

4.2. Machine-based evaluation with a loss function

During the training process of the Soundscape-to-Image Diffusion model, the Low-Resolution Diffusion Model and the Super-Resolution Diffusion Model were trained independently. Therefore, we present the model losses of the two models in Fig. 4. The loss functions for both

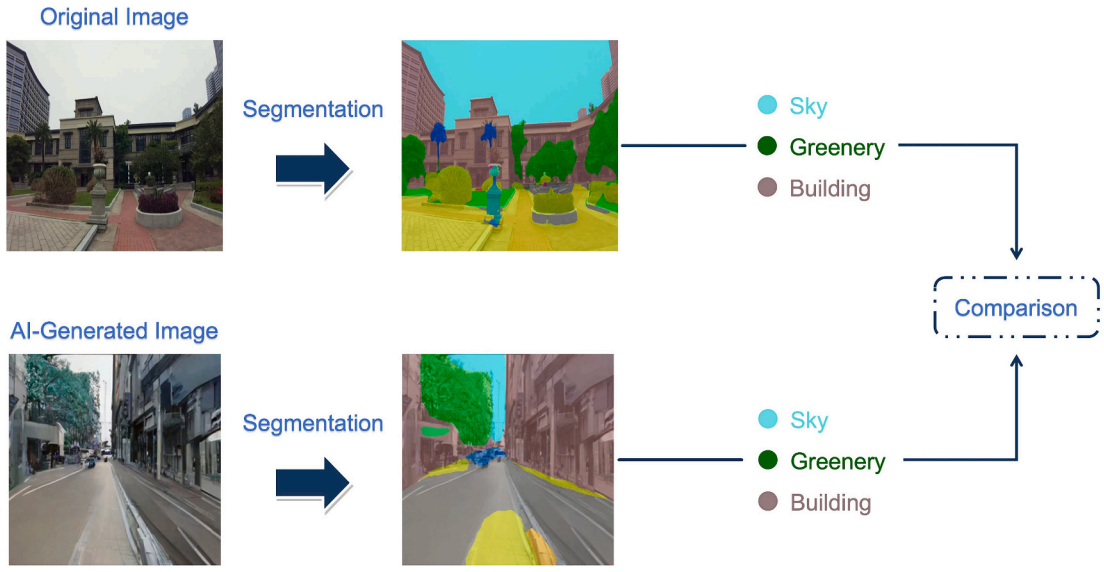


Fig. 2. Calculate the proportion of sky, greenery, and buildings in both the original and generated images separately through image segmentation.

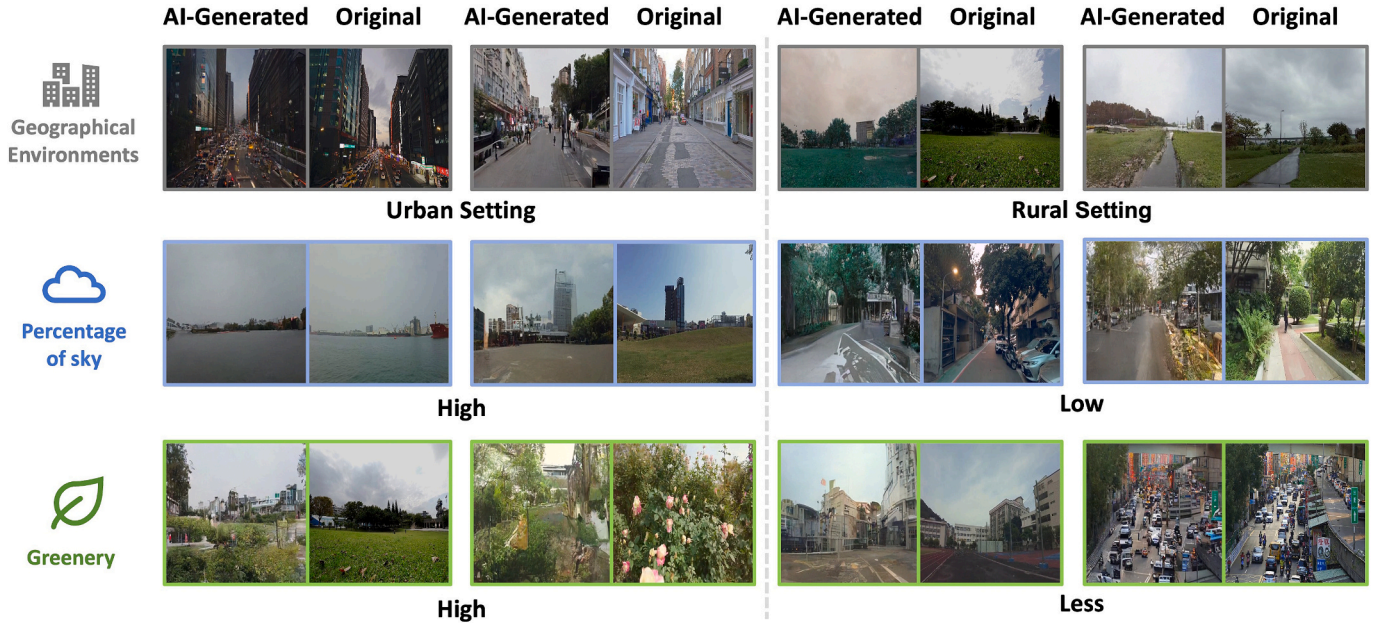


Fig. 3. Example street view images generated using the Soundscape-to-Image Diffusion model at different places.

models exhibited similar trends. In the first five epochs of training, there was a sharp decrease in the loss, indicating significant improvements. Then, the reduction rate in the loss function decelerated, suggesting the model was approaching convergence. After 30 epochs of training, both models showed considerable progress in reducing the loss function values, and thereby we terminated the training process. For the Low-Resolution Diffusion Model, the noise MSE loss function decreased from 0.261 to 0.022. Similarly, for the Super-Resolution Diffusion Model, the noise-loss function decreased from 0.109 to 0.008.

In parallel, we computed the loss functions based on the validation data. The results showed that the average loss function values computed by both models on the validation dataset (0.022 and 0.008) were very close to the loss function values computed based on the training set. These outcomes demonstrate the effectiveness of the training process, showing that our proposed Soundscape-to-Image Diffusion Model could generate images from audio clips accurately.

4.3. *Machine-based evaluation at the element level*

We then evaluated the image similarity at the street element level to determine whether the objects in the ground-truth images were similar to those in the generated images. We selected 100 different random soundscapes from the testing dataset and compared the images they generated through the proposed Soundscape-to-Image Diffusion model with the original images. For each soundscape, we generated five images, resulting in a total of 500 pairs for comparison. The correlation coefficients between the generated 500 images and the original images were calculated based on three urban elements including the proportion of greenery, building, and sky pixels in the image (Fig. 5). It is worth noting that, Stable Diffusion models often produce different outputs with the same inputs. Generating five images for each audio clip could enhance the reliability of our comparison results. To complete this task, we performed semantic segmentation on both original and generated images using the DeepLabV3 semantic segmentation model(Zhou et al.,

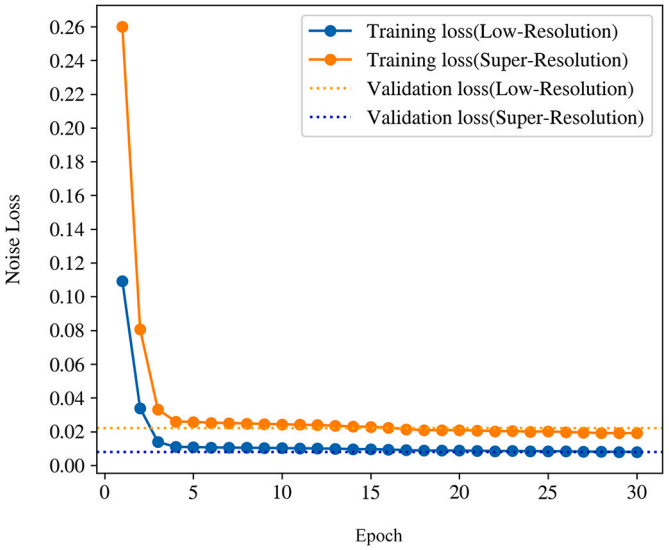


Fig. 4. The loss functions in the training process.

2019) (Fig. 2), and then the proportions of greening, building, and sky pixels in the images were quantified. Finally, the correlation coefficients of those objects between the original and generated images were calculated.

Fig. 5 shows the comparison results based on the proportions of different street elements. According to Pearson's correlation coefficient, we observed a strong positive association regarding the proportion of the sky between the generated images and the original images, with a coefficient of 0.80. Similarly, we found a significant correlation for the proportion of greenery pixels, with a coefficient of 0.69. It suggests that our generative model could successfully learn visual cues from street soundscapes and infer the corresponding visual information of the street such as the openness of the sky and the ratio of greenery. In comparison, the correlation was slightly lower for building pixels, with a coefficient of 0.59. It may indicate that buildings have more complex and diverse textures and styles, and it might be challenging to fully learn the features related to different buildings. In sum, by quantifying the three notable street elements, we demonstrate that our model could translate acoustic environments to visual information at places, indicating the potential of our approach in image generation from soundscapes.

4.4. Human-centered qualitative evaluation

After performing the above two machine-based evaluation metrics, here, we integrated human-centered evaluation approaches for a more comprehensive assessment of our model's performance. We first generated sample street view images for different places. We hypothesize that if the characteristics of different places can be reflected in the images generated from the audio with distinct features, then the effectiveness of our method could be verified to a certain extent.

We show several examples at different places in Fig. 6. In the first group, one soundscape was collected from a busy urban street characterized by traffic noises, while the other was collected from a quiet forest path in rural areas filled with natural sounds. For the generated image in an urban setting, we can identify the road and the tall gray buildings on both sides of the road, and very less greenery. In comparison, images generated from rural soundscapes contain roads, much greenery, and visible water bodies, as well as the reflection of the trees in the water.

In the second group, one soundscape was collected in a grassy open space, while the other one was collected from a quiet cul-de-sac surrounded by buildings in a city. The images generated from the soundscape with greenery and open space also contain large grass areas and trees, while for the images generated from the soundscape at places with less open space and greenery, we can barely see grass and trees. (Fig. 6).

In the last group, the input soundscape was recorded at places with different levels of sky visibility. Our proposed Soundscape-to-Image Diffusion model successfully captured this difference (Fig. 6). In a scene with high tree density and limited light, the generated images are also relatively gloomy and difficult to see the sky. In terms of a place with open spaces and high visibility, the generated images can well reflect clouds in sky and have a larger proportion of sky pixels.

To summarize, our human-centered qualitative evaluation based on the three groups of sample street view images all demonstrate that our proposed Soundscape-to-Image Diffusion model learns visual cues from soundscapes to generate street view images, and could effectively link the auditory and visual perceptions together.

4.5. Human-centered quantitative evaluations with volunteers

Moreover, we invited volunteers to quantitatively assess the quality of our generated outputs. In particular, we have human observations and performed visual assessments for the images generated using our proposed Soundscape-to-Image Diffusion model.

We randomly selected 10 audio segments from the validation dataset and used our model to generate 10 corresponding street view images. We then launched a survey and invited participants to match the audio clips and street view images. The main objective was to test whether human evaluators could correctly associate each audio clip with its generated street view image. In this evaluation process, participants were asked to listen to an audio segment and then select the street view image that most closely matched the audio from a set of three images. Among the three images, only one (the corresponding item) was generated by our model based on the given audio, while the other two were randomly selected from images in other place categories.

We show a screenshot of the user interface of our survey in Fig. 7. In total, we invited 22 people to participate in our survey. Results showed

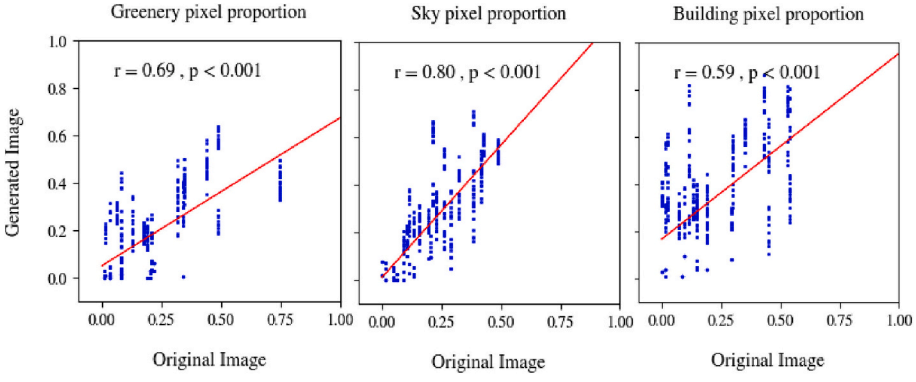


Fig. 5. Scatter plots of correlation coefficients regarding proportions of different street elements between original and geneated images.

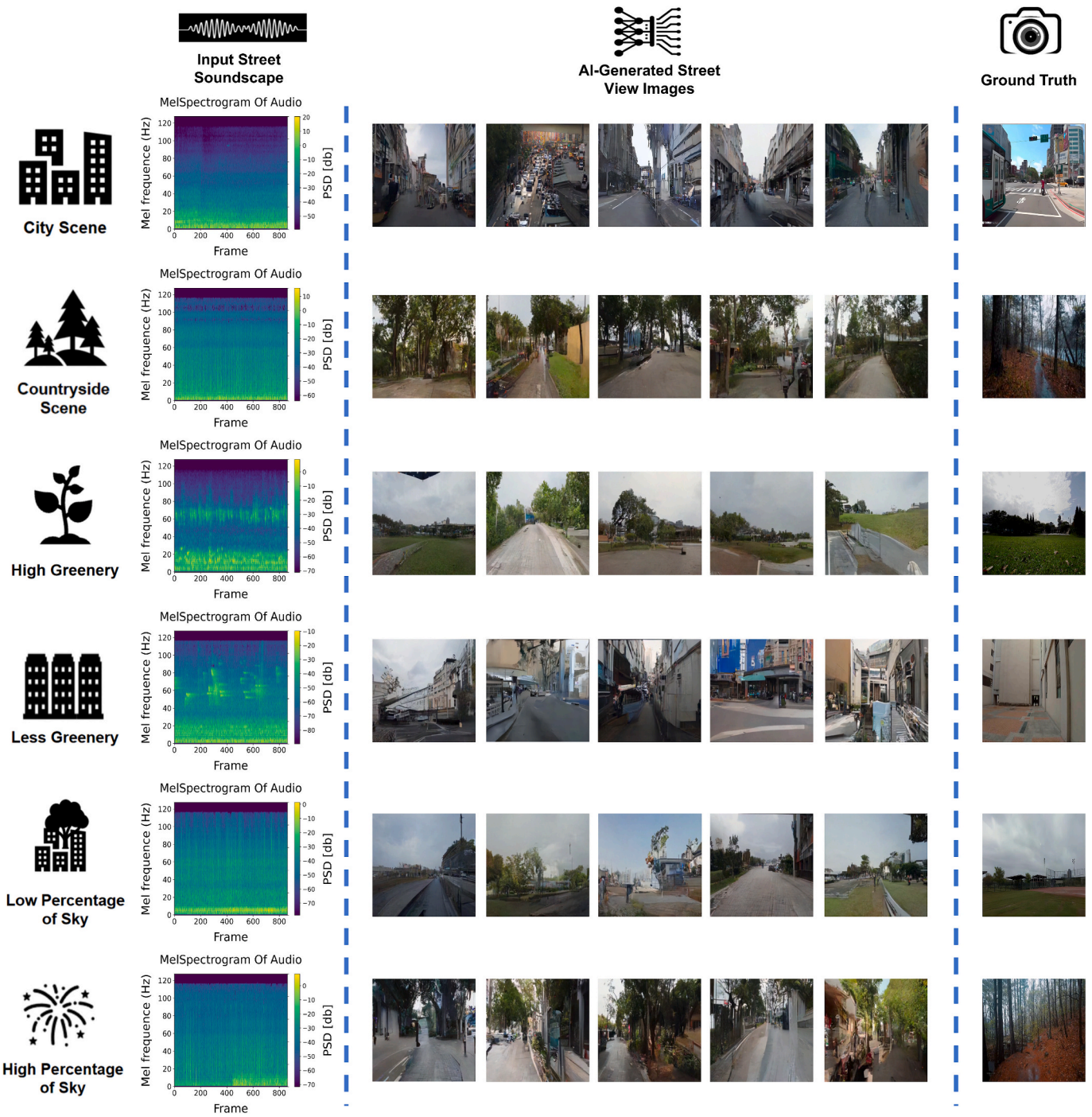


Fig. 6. Sample street view images in the three groups (urban vs. rural places; high and low greenery scene; high and low sky visibility places).

that the average matching rate of human raters was 80.455%, which to some extent verified the consistency and reliability of our model. Among 22 participants, the matching rate of correctly linking audio and corresponding street view images ranges from 50% (0.5) to 100% (1.0), with a variance of 0.1225. Our results demonstrate that the images generated from our proposed Soundscape-to-Image Diffusion model could well capture place characteristics of acoustic environments, shedding lights for advancing human multi-sensory experiences at place, and modeling our human-environment interactions.

5. Discussions

We list several takeaways of this study in this section.

5.1. Implications for understanding human experiences at place

In this study, we developed a novel computational framework by creating a Soundscape-to-Image Diffusion model that successfully associates human auditory and visual perceptions. The results obtained from the study demonstrate that the model could effectively visualize soundscape characteristics by generating corresponding street view images with the same environmental settings as the audio inputs. For instance, for an audio clip that contains vehicle noises and human voices, the model could generate an image that depicts commercial streets. Additionally, the model could learn visual cues from soundscapes, such as the density of tree canopy, architectural styles, appearances of roads and houses, and other environmental factors that represent soundscape

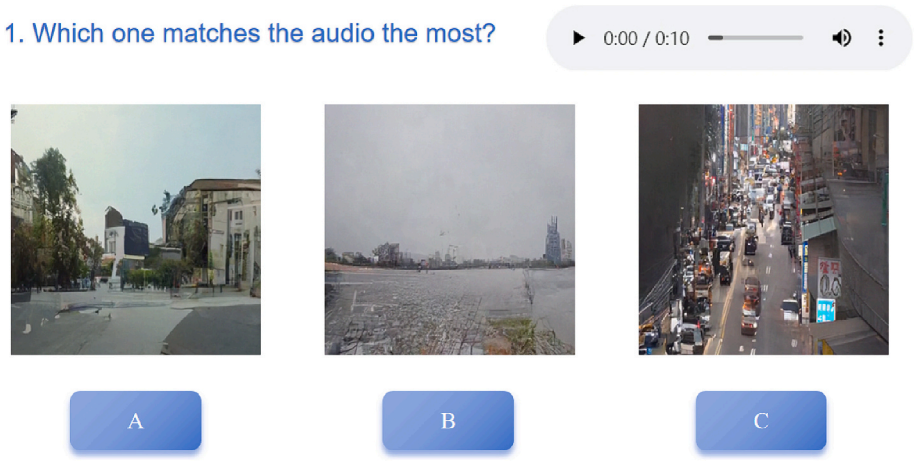


Fig. 7. A screenshot of our survey's user interface.

characteristics. Interestingly, the images maintain spatial relationships among objects within street view images to a certain degree, highlighting the adaptability and generalizability of our model.

The proposed methods could not only enrich our understanding of human multi-sensory experiences at place but also offer new insights into the complex nature of human-environment interactions. Our study reveals the connections between objects in street soundscapes and visual representations of streets. We discovered that soundscapes capture environment settings such as the presence of greenery and buildings, and the openness of the space. These characteristics could be visualized in street view images, highlighting the comprehensive and nuanced relationship between soundscape and visual elements at a specific place, contributing to a better understanding of how auditory and visual information are intertwined in the environment.

Furthermore, we unexpectedly discovered a potential connection between light and street soundscapes. The data spans various conditions, including both sunny and cloudy days, as well as different times of day. We tested our model under three different lighting conditions, namely, sunny day, cloudy day, and clear night. Surprisingly, the generated street view images exhibited a high level of accuracy in reflecting lighting conditions, closely matching real-world place images (Fig. 8). One possible explanation may refer to the variations in human and non-human activities throughout the day, which in turn influences the soundscape. For example, a commercial street on a weekday morning might be relatively quiet but becomes busy with human activities in the evening as people finish work, leading to noises of heavy traffic. Later at night, the areas quiet down again, with the soundscape shifting to the nocturnal chirping of insects. In addition to the findings we have discussed, there may exist more hidden connections between the street soundscape and the visual representation of the street. Our

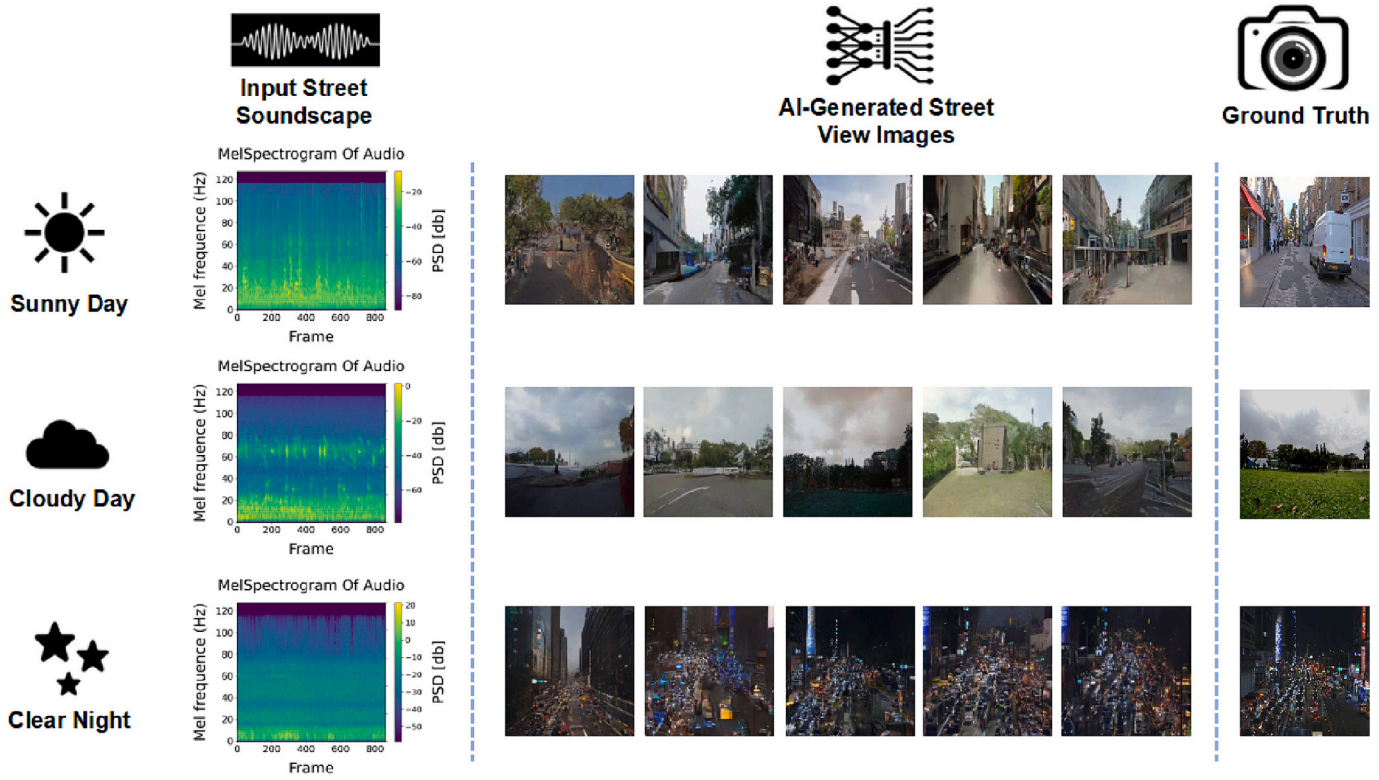


Fig. 8. Our model generated scenes under different environmental lighting based on audio.

soundscape visualization methodology could capture such diverse soundscape features, providing a comprehensive portrayal of the acoustic environment. This highlights the complex interplay among various environmental factors in shaping our overall perceptions of places, emphasizing the importance of analyzing multi-sensory experiences simultaneously.

5.2. Methodological and ethical implications

From a visualization perspective, our research offers novel insights into visualizing audio features. Previous research often involves segmenting audio into various dimensions based on sound elements like frequency and pitch or relies on sound separation and classification to extract relevant events from complex sounds. In contrast, our proposed Soundscape-to-Image method exhibits certain advantages. It eliminates the need for fixed event classification, resulting in greater universality and more intuitive results. In addition, street soundscapes are often complex and overlapping, as the soundscape is the result of a combination of street activities and spatial configurations of street buildings and facilities. Such complex interactions are challenging to be visualized through conventional methods, but our Soundscape-to-Image Diffusion approach could effectively and vividly represent them to human eyes.

Despite the advancements in our proposed methods, we observe several potential ethical questions that need to be considered before being used in real-world practices (Kang et al., 2024; Ntoutsi et al., 2020; Shaw & Sui, 2021). Two issues, namely, inaccuracy, and irreproducibility, were observed in the street view images generated by our Soundscape-to-Image Diffusion model.

We found that the details of AI-generated street view images may differ from reality, and may have the following inaccuracies, including blurred boundaries, strange shapes, and mixed color styles. We provide several examples in Fig. 9. As illustrated in the figure, the outlines of buildings and roads might not be clear. On the road surface, there seemed to be crosswalks, but their shapes and directions appeared strange and irregular. Additionally, the color styles of the buildings observed in images were not consistent, mixing different color styles, which is uncommon in real life.

Additionally, we observed that street view images generated by our Soundscape-to-Image Diffusion model cannot be reproduced even using the same auditory soundscape. Due to the randomness inherent in the stable diffusion generation process, it is impossible to generate two images with exactly the same content, objects, and layouts. Currently, street view images generated from the same audio could just exhibit similarities in certain features but not the overall scene. For example, in Fig. 10, several street view images were generated from the same auditory soundscape. They all have dense tree canopy but there are still differences in the style, orientation, and colors of trees and roads. The lack of reproducibility presents challenges in validating and replicating the images generated from the Soundscape-to-Image Diffusion model, thus raising crucial ethical questions (Kedron et al., 2021).

5.3. Limitations and future directions

We acknowledge several limitations of this study to guide improvements in the future. First, due to constraints in computing resources, our Soundscape-to-Image Diffusion model could be further improved based on larger datasets with more hyperparameters to generate higher-resolution images. We could better improve the model performance with deeper investigation of optimal parameters such as audio length or audio encoder for generating higher-quality images. Second, we are considering incorporating Reinforcement Learning with Human Feedback (RLHF) into our algorithm to ensure that the generated image details better align with human preferences. Other audio encoders might be tested and integrated such as ImageBind to further refine the model architecture. Finally, we only focused on three objects in streetscapes including greenery, sky view, and buildings. In the future, we aim to broaden the scope of our research by incorporating more street elements. This not only allows us to validate the reliability and generalizability of the model from different perspectives but also inform placemaking from a more diverse perspectives.

6. Conclusions

This study proposes a novel framework for visualizing street soundscapes by developing a Soundscape-to-Image Diffusion model. Our computational framework contains four key components: (1) data collection; (2) soundscape feature processing; (3) a Soundscape-to-Image Diffusion model; and (4) evaluation. We collected videos and created a multi-sensory dataset that contains audio-image pairs. Then, we trained a Soundscape-to-Image Diffusion model using this dataset to associate visual and auditory information. Validated through machine-based and human-centered evaluations, the Soundscape-to-Image Diffusion model, which includes Low- and Super-resolution Diffusion Models, effectively translates semantic audio vectors into visual representations of places.

Through the Soundscape-to-Image Diffusion model, our study provides new insights for professionals, researchers, and designers. From the perspective of human geography, our research breaks through the limitation of past studies where the two dimensions of human experiences at a place were often studied separately and innovatively synthesizes the two-dimensional (auditory and visual) perceptions to understand the human sense of place. From the viewpoint of environmental psychology, our research contributes to a further understanding of how auditory cues contribute to our visual perceptions of places,

Fig. 9. The generated image exhibits inconsistencies with reality in terms of blurred boundaries, strange shapes, and mixed color styles.

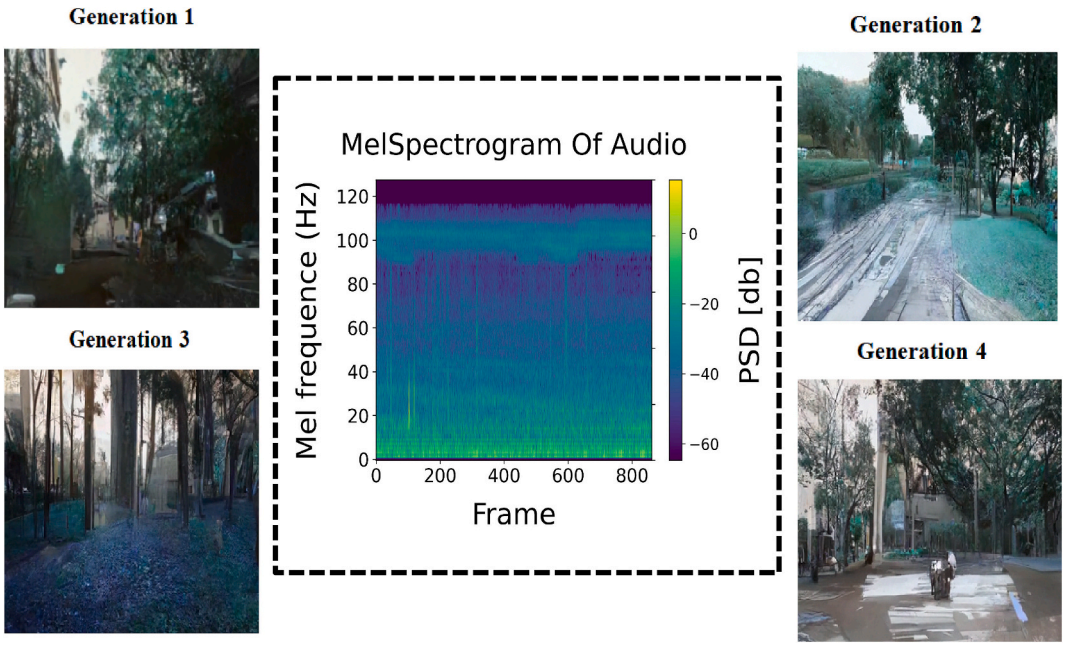


Fig. 10. Four images of street soundscape visualization were generated by the same soundscape (visualized in the middle using Melspectrogram). This soundscape was recorded under a tree in a park. The park was filled with bird songs and without human voices or traffic sounds. Although the generated images are not exactly the same, they all reflect the common features of trees, nature, and the absence of people and vehicles.

enhancing our knowledge of the impacts of visual and auditory perceptions on human mental health. For many urban designers, our findings can better help them understand the relationship between urban landscapes and noise, as well as advance their knowledge of human-environment relationships.

CRediT authorship contribution statement

Yonggai Zhuang: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Yuhao Kang:** Conceptualization, Visualization, Writing – original draft, Writing – review & editing. **Teng Fei:** Conceptualization, Funding acquisition, Visualization, Writing – original draft, Writing – review & editing. **Meng Bian:** Data curation, Writing – review & editing. **Yunyan Du:** Funding acquisition, Writing – review & editing.

Acknowledgment

We acknowledge the support by grants from the National Natural Science Foundation of China under grant number 42271476 (TF), the 351 Talent Project of Wuhan University (TF), and the Key Laboratory of Resources and Environmental Information System (TF & YD), all of which have been instrumental in facilitating our project's success. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders.

References

Agnew, J. A. (2011). *Space and place*.
Agostinelli, A., Denk, T., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., … Frank, C. (2023). *MusicLM: Generating music from text*.
Aiello, L., Schifanella, R., Quercia, D., & Aletta, F. (2016). Chatty maps: Constructing sound maps of urban areas from social media data. *Royal Society Open Science, 3*.
Arnal, L., Flinker, A., Kleinschmidt, A., Giraud, A.-L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, 25.
Åsa Ode Sang, Knez, Igor, Gunnarsson, Bengt, & Hedblom, Marcus (2016). The effects of naturalness, gender, and age on how urban green space is perceived and used. *Urban Forestry & Urban Greening, 18*, 268–276.
Bhandari, U., Chang, K., & Neben, T. (2019). Understanding the impact of perceived visual aesthetics on user evaluations: An emotional perspective. *Information & Management, 56*, 85–93.
Bilaşco, Ş., Govor, C., Roşca, S., Vescan, I., Filip, S., & Fodorean, I. (2017). GIS model for identifying urban areas vulnerable to noise pollution: Case study. *Frontiers of Earth Science, 11*, 214–228.
Brooks, B., Schulte-Fortkamp, B., Voigt, K., & Case, A. (2014). Exploring our sonic environment through soundscape research & theory. *Acoustics Today, 10*, 30–40.
Buxton, R., Pearson, A., Allou, C., Fristrup, K., & Wittemyer, G. (2021). A synthesis of health benefits of natural sounds and their distribution in national parks. *Proceedings of the National Academy of Sciences, 118*, Article e2013097118.
Cain, R., Jennings, P., & Poxon, J. (2013). The development and application of the emotional dimensions of a soundscape. *Applied Acoustics, 74*, 232–239.
Chen, W., Wu, J., Pan, X., Wu, H., Li, J., Xia, X., … Liang, L. (2023). *Control-A-video: Controllable text-to-video generation with diffusion models*.
Clement, J., Kristensen, T., & Grønhaug, K. (2013). Understanding consumers' in-store visual perception: The influence of package design features on visual attention. *Journal of Retailing and Consumer Services, 20*, 234–239.
Cosgrove, D. (2003). *Landscape and the european sense of sight - eyeing nature*.
Cresswell, T. (2014). *Place: An introduction*.
D'Alessandro, F., Evangelisti, L., Guattari, C., Grazieschi, G., & Orsini, F. (2018). Influence of visual aspects and other features on the soundscape assessment of a university external area. *Building Acoustics*, 25.
Dunkel, A. (2015). Visualizing the perceived environment using crowdsourced photo geodata. *Landscape and Urban Planning, 142*, 173–186.
Eronen, A. J., Peltonen, V. T., Tuomi, J. T., Klapuri, A. P., Fagerlund, S., Sorsa, T., … Huopaniemi, J. (2006). Audio-based context recognition. *IEEE Transactions on Audio, Speech and Language Processing, 14*, 321–329.
Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience, 17*, 738–743.
Fuller, S., Axel, A. C., Tucker, D., & Gage, S. H. (2015). Connecting soundscape to landscape: Which acoustic index best describes landscape configuration? *Ecological Indicators, 58*, 207–215.
Gao, S., Yu, L., Kang, Y., & Zhang, F. (2021). *User-generated content: A promising data source for urban informatics*.
Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K., Joulin, A., & Misra, I. (2023). *ImageBind: One embedding space to bind them all*.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial networks. *Advances in Neural Information Processing Systems, 3*.
Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., … Yu, Q. (2023). *ImageBind-LLM: Multi-modality instruction tuning*.
Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising diffusion probabilistic models*. ArXiv, abs/2006.11239.
Holmes, C., Brook, M., Krehbiel, P., & McCrory, R. (1971). On the power spectrum and mechanism of thunder. *Journal of Geophysical Research, 76*, 2106–2115.

Hong, J. Y., Lam, B., Ong, Z.-T., Ooi, K., Gan, W.-S., Kang, J., … Tan, S.-T. (2019). Quality assessment of acoustic environment reproduction methods for cinematic virtual reality in soundscape applications. *Building and Environment, 149*, 1–14.
Hopkins, J. M., Edwards, W., & Lin, S. (2022). Invading the soundscape: Exploring the effects of invasive species' calls on acoustic signals of native wildlife. *Biological Invasions, 24*, 3381–3393.
Huang, J., Fei, T., Kang, Y., Li, J., Liu, Z., & Guofeng, W. (2023). Estimating urban noise along road network from street view imagery. *International Journal of Geographical Information Science, 38*, 128–155.
In Jo, H., & Jeon, J. Y. (2020). Effect of the appropriateness of sound environment on urban soundscape assessment. *Building and Environment, 179*, 106975.
Janowicz, K., Gao, S., McKenzie, G., Hu, Y., & Bhaduri, L. (2020). GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science, 34*, 1–13.
Jeon, J. Y., & Hong, J. Y. (2015). Classification of urban park soundscapes through perceptions of the acoustical environments. *Landscape and Urban Planning, 141*, 100–111.
Jeon, J. Y., & In Jo, H. (2020). Effects of audio-visual interactions on soundscape and landscape perception and their influence on satisfaction with the urban environment. *Building and Environment, 169*, 106544.
Kang, Y., Abraham, J., Ceccato, V., Duarte, F., Gao, S., Ljungqvist, L., … Ratti, C. (2023). Assessing differences in safety perceptions using GeoAI and survey across neighbourhoods in Stockholm, Sweden. *Landscape and Planning, 236*, 1–11.
Kang, Y., Gao, S., & Roth, R. (2019). *Transferring multiscale map styles using generative adversarial networks*.
Kang, Y., Gao, S., & Roth, R. (2024). Artificial intelligence studies in cartography: A review and synthesis of methods, applications, and ethics. *Cartography and Geographic Information Science*, 1–32.
Kang, Y., Jia, Q., Gao, S., Zeng, X., Wang, Y., Angsuesser, S., … Fei, T. (2019). *Extracting human emotions at different places based on facial expressions and spatial clustering analysis*.
Kang, Y., Zhang, F., Gao, S., Peng, W., & Ratti, C. (2021). Human settlement value assessment from a place perspective: Considering human dynamics and perceptions in house price modeling. *Cities, 118*, 103333.
Kedron, Peter, Frazier, Amy, E, Trgovac, Andrew, B, … Stewart. (2021). Reproducibility and replicability in geographical analysis. *Geographical Analysis, 53*(1), 135–147.
Keyel, A. C., Reed, S. E., McKenna, M. F., & Wittemyer, G. (2017). Modeling anthropogenic noise propagation using the sound mapping tools ArcGIS toolbox. *Environmental Modelling & Software, 97*, 56–60.
Landeschi, G., Dell'Unto, N., Lundqvist, K., Ferdani, D., Campanaro, D. M., & Touati, A.-M. L. (2016). 3D-GIS as a platform for visual analysis: Investigating a Pompeian house. *Journal of Archaeological Science, 65*, 103–113.
Lee, S., Hoover, B., Strobelt, H., Wang, Z., Peng, S. Y., Wright, A., … Chau, P. (2023). *Diffusion explainer: Visual explanation for text-to-image stable diffusion*.
Lillis, A., Eggleston, D. B., & Bohnenstiehl, D. R. (2014). Estuarine soundscapes: Distinct acoustic characteristics of oyster reefs compared to soft-bottom habitats. *Marine Ecology Progress Series, 505*, 1–17.
Manzo, L. (2003). Beyond house and haven: Toward a revisioning of emotional relationships with places. *Journal of Environmental Psychology, 23*, 47–61.
Marchegiani, L., & Posner, I. (2017). Leveraging the urban soundscape: Auditory perception for smart vehicles. In *2017 IEEE international conference on robotics and automation (ICRA)* (pp. 6547–6554).
Martin, B. (2017). Soundscape composition: Enhancing our understanding of changing soundscapes. *Organised Sound, 23*, 1–9.
Michael, I., Ramsøy, T. Z., Balakrishnan, M. S., & Kotsi, F. (2019). A study of unconscious emotional and cognitive responses to tourism images using a neuroscience method. *Journal of Islamic Marketing, 10*.
Minelli, A., Marchesini, I., Taylor, F. E., De Rosa, P., Casagrande, L., & Cenci, M. (2014). An open source GIS tool to quantify the visual impact of wind turbines and photovoltaic panels. *Environmental Impact Assessment Review, 49*, 70–78.
Moliner, E., Lehtinen, J., & Välimäki, V. (2023). *Solving audio inverse problems with a diffusion model*.
Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., … Chen, M. (2021). GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International conference on machine learning*.
Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery, 10*.
Oppenheim, A. V., & Schafer, R. W. (2004a). From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine, 21*, 95–106.
Oppenheim, A. V., & Schafer, R. (2004b). From frequency to quefrency: A history of the cepstrum. In *, 21. Signal processing magazine* (pp. 95–106). IEEE.
Oppenlaender, J. (2022). The creativity of text-to-image generation. In *Proceedings of the 25th international academic Mindtrek conference*.
Park, Tae Hong and Lee, Jun Hee and You, Jaeseong and Yoo, Min-Joon and Turner, Johnathan, (2014), Towards soundscape information retrieval (SIR), ICMC.
Phillips, Y., Towsey, M., & Roe, P. (2017). *Visualization of environmental audio using ribbon plots and acoustic state sequences*.
Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., … Pieretti, N. (2011). Soundscape ecology: The science of sound in the landscape. *BioScience, 61*, 203–216.
Pocock, D., Relph, E., & Tuan, Y.-F. (1994). Classics in human geography revisited: Tuan, Y.-F. 1974: Topophilia. Englewood Cliffs, NJ: Prentice-Hall. In *, 18. Progress in human geography - Prog Hum Geogr* (pp. 355–359).
Prestigiacomo, A. (1962). Amplitude contour display of sound spectrograms. *Journal of The Acoustical Society of America, 34*.
Primeau, K. E., & Witt, D. E. (2018). Soundscapes in the past: Investigating sound at the landscape level. *Journal of Archaeological Science: Reports, 19*, 875–885.
Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with CLIP Latents*.
Raymond, C., Kyttä, M., & Stedman, R. (2017). Sense of place, fast and slow: The potential contributions of affordance theory to sense of place. *Frontiers in Psychology, 1674*, 1–14.
Ren, X., Li, Q., Yuan, M., & Shao, S. (2023). How visible street greenery moderates traffic noise to improve acoustic comfort in pedestrian environments. *Landscape and Urban Planning, 238*, 104839.
Reynaud, H., Qiao, M., Dombrowski, M., Day, T., Razavi, R., Gomez, A., … Kainz, B. (2023). *Feature-conditioned cascaded video diffusion models for precise echocardiogram synthesis*.
Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). *High-resolution image synthesis with latent diffusion models*.
Sacchelli, S., & Favaro, M. (2019). A virtual-reality and soundscape-based approach for assessment and management of cultural ecosystem services in urban forest. In *Forests*.
Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., … Norouzi, M. (2022a). *Photorealistic text-to-image diffusion models with deep language understanding*.
Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., … Norouzi, M. (2022b). *Photorealistic text-to-image diffusion models with deep language understanding*.
Salih Can Yurtkulu, Yusuf Hüseyin Şahin, & Gozde Unal. (2019). *Semantic segmentation with extended DeepLabv3 architecture, 2019*. IEEE.
Schafer, R. M. (1977). *The tuning of the world: Toward a theory of soundscape design*.
Schreuder, E., Erp, J., Toet, A., & Kallen, V. (2016). Emotional responses to multisensory environmental stimuli. *SAGE Open, 6*, 1.
Schulte-Fortkamp, B., & Fiebig, A. (2006). Soundscape analysis in a residential area: An evaluation of noise and people's mind. *Acta Acustica united with Acustica, 92*, 875–880.
Shaw, S.-L., & Sui, D. (2021). *Mapping COVID-19 in space and time understanding the spatial and temporal dynamics of a global pandemic: Understanding the spatial and temporal dynamics of a global pandemic*.
Smith, S. (2015). A sense of place: Place, culture and tourism. *Tourism Recreation Research, 40*, 220–233.
Spence, C. (2020). Senses of place: Architectural design for the multisensory mind. *Cognitive Research: Principles and Implications, 5*, 46.
Tan, Johann Kay Ann and Hasegawa, Yoshimi and Lau, Siu-Kit and Tang, Shiu-Keung, (2022). The effects of visual landscape and traffic type on soundscape perception in high-rise residential estates of an urban city. Applied Acoustics, 189, 108580.
Tuan, Y. (1974). *Topophilia: A study of environmental perception, attitudes, and values*.
Tuan, Y.-F. (1975). *Topophilia: A study of environmental perception, attitudes, and values*. SERBIULA (sistema Librum 2.0), 65.
Watts, G., Khan, A., & Pheasant, R. (2016). Influence of soundscape and interior design on anxiety and perceived tranquillity of patients in a healthcare setting. *Applied Acoustics, 104*, 135–141.
Williams, A., Lattner, S., & Barthet, M. (2023). *Sound-and-image-informed music artwork generation using text-to-image models*.
Wilson, C. J., & Soranzo, A. (2015). The use of virtual reality in psychology: A case study in visual perception. *Computational and Mathematical Methods in Medicine, 2015*, 151702.
Wróżyński, R., Sojka, M., & Pyszny, K. (2016). The application of GIS and 3D graphic software to visual impact assessment of wind turbines. *Renewable Energy, 96*, 625–635.
Wu, H.-H., Seetharaman, P., Kumar, K., & Bello, J. (2021). *Wav2CLIP: Learning robust audio representations from CLIP*.
Wu, Zherong, Ma, Peifeng, Zheng, Yi, Gu, … Lin. (2023). Automatic detection and classification of land subsidence in deltaic metropolitan areas using distributed scatterer InSAR and Oriented R-CNN}. *Remote Sensing of Environment, 290*, 113545.
Xu, Xiaowei, Chen, Yinrong, Zhang, Chen, … Manickam. (2021). A novel approach for scene classification from remote sensing images using deep learning methods. *European Journal of Remote Sensing, 54*(Sup 2), 383–395.
Yang, L., Zhang, Z., Yang, S., Hong, S., Xu, R., Zhao, Y., … Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications. In *ACM computing surveys*.
Yildirim, Y., & Arefi, M. (2022). Sense of place and sound: Revisiting from multidisciplinary outlook. *Sustainability, 14*, 11508.
Zhao, T., Liang, X., Wei, T., Huang, Z., & Biljecki, F. (2023). Sensing urban soundscapes from street view imagery. *Computers, Environment and Urban Systems, 99*, 101915.
Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision, 127*.