Cross-Task Transfer for Multimodal Aerial Scene Recognition

Di Hu\textsuperscript{1}, Xuhong Li\textsuperscript{1}, Lichao Mou\textsuperscript{2}, Pu Jin\textsuperscript{3}, Dong Chen\textsuperscript{4}, Liping Jing\textsuperscript{4}, Xiaoxiang Zhu\textsuperscript{2,3}, and Dejing Dou\textsuperscript{1}

\textsuperscript{1} Big Data Laboratory, Baidu Research  \\
\{hudi04,lixuhong,doudejing\}@baidu.com  \\
\textsuperscript{2} German Aerospace Center  \\
\{lichao.mou,xiaoxiang.zhu\}@dlr.de  \\
\textsuperscript{3} Technical University of Munich pu.jin@tum.de  \\
\textsuperscript{4} Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University  \\
\{chendong,lpjing\}@bjtu.edu.cn

Abstract. Aerial scene recognition is a fundamental task in remote sensing and has recently received increased interest. While the visual information from overhead images with powerful models and efficient algorithms yields good performance on scene recognition, additional information is always a bonus, for example, the corresponding audio information. In this paper, for improving the performance on the aerial scene recognition, we explore a novel audiovisual aerial scene recognition task using both images and sounds as input. Based on an observation that some specific sound events are more likely to be heard at a given geographic location, we propose to exploit the knowledge from the sound events to improve the performance on the aerial scene recognition. For this purpose, we have constructed a new dataset named Audio Visual Aerial scene reCognition datasEt (ADVANCE). With the help of this dataset, we evaluate three proposed approaches for transferring the sound event knowledge to the aerial scene recognition task in a multimodal learning framework, and show the benefit of exploiting the audio information for the aerial scene recognition. The source code is publicly available for reproducibility purposes.\textsuperscript{5}

Keywords: Cross-task transfer, aerial scene, multimodal learning

1 Introduction

Scene recognition is a longstanding, hallmark problem in the field of computer vision, and it refers to assigning a scene-level label to an image based on its overall contents. Most scene recognition approaches in the community make use of ground images and have achieved remarkable performance. By contrast, overhead images usually cover larger geographical areas and are capable of offering more comprehensive information from a birds eye view than ground images.

\textsuperscript{5} \url{https://github.com/DTaoo/Multimodal-Aerial-Scene-Recognition}
Fig. 1. Two examples showing aerial scenes soundscapes could be a helpful cue for identifying their scene categories. More details of the audiovisual model please refer to Section 4.4. Here, we use the class activation mapping (CAM) technique to visualize what models are looking.

Hence aerial scene recognition has received increased interest. The success of current state-of-the-art aerial scene understanding models can be attributed to the development of novel convolutional neural networks (CNNs) that aim at learning good visual representations from images.

Albeit successful, these models may not work well in some cases, particularly when they are directly used in worldwide applications, suffering the pervasive factors, such as different remote imaging sensors, lighting conditions, orientations, and seasonal variations. A study in neurobiology reveals that human perception usually benefits from the integration of both visual and auditory knowledge. Inspired by this investigation, we argue that aerial scenes soundscapes are partially free of the aforementioned factors and can be a helpful cue for identifying scene categories (Fig. 1). This is based on an observation that the visual appearance of an aerial scene and its soundscape are closely connected. For instance, sound events like broadcasting, people talking, and perhaps whistling are likely to be heard in all train stations in the world, and cheering and shouting are expected to hear in most sports lands. However, incorporating the sound knowledge into a visual aerial scene recognition model and assessing its contributions to this task still remain underexplored. In addition, it is worth mentioning that with the now widespread availability of smartphones, wearable devices, and audio sharing platforms, geotagged audio data have been easily accessible, which enables us to explore the topic in this paper.

In this work, we are interested in the audiovisual aerial scene recognition task that simultaneously uses both visual and audio messages to identify the scene of a geographical region. To this end, we construct a new dataset, named AuDio Visual Aerial sceNe reCognition datasEt (ADVANCE), providing 5075 paired images and sound clips categorized to 13 scenes, which will be introduced in Section 3, for exploring the aerial scene recognition task. According to our preliminary experiments, simply concatenating representations from the two modalities is not helpful, slightly degrading the recognition performance compared to using a vision-based model. Knowing that sound events are related to scenes, this preliminary result indicates that the model cannot directly learn the underlying relation between the sound events and the scenes. So directly transferring the sound event knowledge to scene recognition may be the key to
making progress. Following this direction, with the multimodal representations, we propose three approaches that can effectively exploit the audio knowledge to solve the aerial scene recognition task, which will be detailed in Section 4. We compare our proposed approaches with baselines in Section 5, showing the benefit of exploiting the sound event knowledge for the aerial scene recognition task.

Thereby, this works contributions are threefold.

- The audiovisual perception of human beings gives us an incentive to investigate a novel audiovisual aerial scene recognition task. We are not aware of any previous work exploring this topic.
- We create an annotated dataset consisting of 5075 geotagged aerial image-sound pairs involving 13 scene classes. This dataset covers a large variety of scenes from across the world.
- We propose three approaches to exploit the audio knowledge, i.e., preserving the capacity of recognizing sound events, constructing a mutual representation in order to learn the underlying relation between sound events and scenes, and directly learning this relation through the posterior probabilities of sound events given a scene. In addition, we validate the effectiveness of these approaches through extensive ablation studies and experiments.

2 Related work

In this section, we briefly review some related works in aerial scene recognition, multimodal learning, and cross-task transfer.

Aerial Scene Recognition. Earlier studies on aerial scene recognition [33,23,24] mainly focused on extracting low-level visual attributes and/or modeling mid-level spatial features [15,17,28]. Recently, deep networks, especially CNNs, have achieved a large development in aerial scene recognition [20,4,6], e.g., Nogueira et al. [20] analysed three possible strategies for fully exploiting the power of the existing CNNs. Meanwhile, there are some other works [16,32] aiming at designing effective network architectures for aerial scene recognition in different cases, e.g., Kyrkou et al. [16] designed a lightweight yet efficient CNN architecture for embedded systems. Moreover, some methods were proposed to solve the problem of the limited collection of aerial images by employing more efficient networks [34,37]. Although these methods have achieved great empirical success, they usually learn scene knowledge from the same modality, i.e., image. Different from previous works, this paper mainly focuses on exploiting multiple modalities (i.e. image and sound) to achieve robust aerial scene recognition performance.

Multimodal Learning. Information in the real world usually comes as different modalities, with each modality being characterized by very distinct statistical properties, e.g., sound and image [3]. An expected way to improve relevant task performance is by integrating the information from different modalities. In past decades, amounts of works have developed promising methods on the related
Fig. 2. The aerial images acquisition and labeling steps.

topics, such as reducing the audio noise by introducing visual lip information for
speech recognition [11,1], improving the performance of facial sentiment recog-
nition by resorting to the voice signal [35,35]. Recently, more attention is paid
to the task of learning to analyze real-world multimodal scenarios [21,12,13] and
events [26,31]. These works have confirmed the advantages of multimodal learn-
ing. In this paper, we proposed to recognize the aerial scene by leveraging the
bridge between scene and sound to help better understand aerial scenes.

Cross-task Transfer. Transferring the learned knowledge from one task to an-
other related task has been approved as an effective way for better data modeling
and messages correlating [7,2,14]. Aytar et al. [2] proposed a teacher-student
framework that transfers the discriminative knowledge of visual recognition to
the representation learning task of sound modality via minimizing the differ-
ences in the distribution of categories. Ehrlich et al. [7] learned a joint feature
representation with the interaction between different tasks (such as PCA and
facial landmark detection) to facilitate facial attributes recognition. Chaplot et
al. [5] utilized a dual-attention unit to align textual and visual representations
with the transferred knowledge of words and objects. Due to the partial cor-
relation of scenes and sound events, Imoto et al. [14] proposed a method for
sound event detection by transferring the knowledge of scenes with soft labels.
Salem et al. [25] proposed to transfer the sound clustering knowledge to the
image recognition task by predicting the distribution of sound clusters from an
overhead image, similarly work can be found in [22]. By contrast, this paper
strives to exploit effective sound event knowledge to facilitate the aerial scene
understanding task.

3 Dataset

To our knowledge, the audiovisual aerial scene recognition task has not been
explored before. Although Salem et al. [25] have ever established a dataset to
explore the correlation between geotagged audio and overhead image, the low-quality images and absent scene labels make it difficult to facilitate the research in this field. To this end, we construct a new dataset, named as ADVANCE\(^6\), which in summary contains 5075 pairs of aerial images and sounds, classified into 13 classes.

The audio data are collected from Freesound\(^7\), where we remove the audio recordings that are shorter than 2 seconds, and extend those that are between 2 and 10 seconds to longer than 10 seconds by replicating the audio content. Each audio recording is attached to the geographic coordinates of the sound being recorded. From the location information, we can download the updated aerial images from Google Earth\(^8\). Then we pair the downloaded aerial image with a randomly extracted 10-second sound clip from the entire audio recording content. Finally, the paired data are labeled according to the annotations from OpenStreetMap\(^9\), also using the attached geographic coordinates from the audio recording. Those annotations have manually been corrected and verified by participants in case that some of them are not up to date. The overview of the establishment is shown in Fig. 2.

Due to the inherent uneven distribution of scene classes, the collected data are strongly unbalanced, which makes difficult the training process. So, two extra steps are designed to alleviate the unbalanced-distribution problem. Firstly we filter out the scenes whose numbers of paired samples are less than 10, such as desert and the site of wind turbines. Then for scenes that have less than 100 samples, we apply a small offset to the original geographic coordinates in four directions. So, correspondingly, four new aerial images are generated from Google Earth and paired with the same audio recording, while for each image, a new 10-second sound clip is randomly extracted from the recording. Fig. 3 reveals the final number of paired samples per class. Moreover, as shown in Fig. 4, the

---

\(^6\) The dataset is now publicly available: [https://zenodo.org/record/3828124](https://zenodo.org/record/3828124)

\(^7\) [https://freesound.org/browse/geotags/](https://freesound.org/browse/geotags/)

\(^8\) [https://earthengine.google.com/](https://earthengine.google.com/)

\(^9\) [https://www.openstreetmap.org/](https://www.openstreetmap.org/)
samples are distributed over the whole world, increasing the diversity of the aerial images and sounds.

4 Methodology

In this paper, we focus on the audiovisual aerial scene recognition task, based on two modalities, i.e., image and audio. We propose to exploit the audio knowledge to better solve the aerial scene recognition task. In this section, we detail our proposed approaches for creating the bridge of knowledge transfer from sound event knowledge to the scene recognition in a multi-modality framework.

Let \((\mathbf{x}, t)\) a paired data sample drawn from the image-audio data distribution \(\mu\), where \(\mathbf{x} = \{\mathbf{v}, \mathbf{a}\}\) contains the image \(\mathbf{v}\) and the paired audio \(\mathbf{a}\), and \(t\) is the scene label. Our built dataset ADVANCE is thus an empirical distribution of the real data distribution \(\mu\). We note \(N_v\) and \(N_a\) as two convolutional networks for extracting representations from images and sound clips respectively, and \(N_v + a\) for the network that concatenates the representations from \(N_v\) and \(N_a\), as shown in Fig. 5.

Then with the extracted representation, a fully-connected layer, whose size depends on the choice of the network \(N_\ast\) and the number of categories, transforms the representation \(N_\ast(\mathbf{x})\) to the label space, and then predicts the probability of each category with an appropriate activation function. We write the output of the entire network before final activation as \(f(\mathbf{x}, N_\ast)\). We use the notation \(f_e(\mathbf{x}, N_\ast)\) for sound event recognition, and \(f_s(\mathbf{x}, N_\ast)\) for aerial scene recognition. Our main task is a supervised learning problem that minimizes the
loss function for aerial scene recognition, defined as

\[ L_s = \mathbb{E}_{(x,t) \sim \mu} \left[ D_{KL}( \mathbf{h}_t \| z(f_s(x, N_s^*)) ) \right], \]

(1)

where \( \mathbf{h}_t \) is the one-hot vector where all elements are null except the \( t \)-th, indicating that the sample \( x \) belongs to the scene \( s_t \), \( z \) is the softmax function and \( D_{KL} \) is the Kullback-Leibler divergence between the one-hot ground truth \( \mathbf{h}_t \) and the predicted probability.

In the rest of this section, we formulate our proposed model architecture for addressing the multimodal scene recognition task, and present our idea of exploiting the audio knowledge following three directions: (1) avoid forgetting the audio knowledge during training by preserving the capacity of recognizing sound events; (2) construct a mutual representation that solves the main task and the sound event recognition task simultaneously, allowing the model to learn the underlying relation between sound events and scenes; (3) directly learn the relation between sound events and scenes. Our total objective function \( L \) is

\[ L = L_s + \alpha L_\Omega, \]

(2)

where \( \alpha \) controls the force of \( L_\Omega \), and \( L_\Omega \) is one of the three approaches that are respectively presented in Section 4.2, 4.3 and 4.4, as illustrated in Fig. 5.

4.1 Overview

For the multimodal learning task with deep networks, we adopt the model architecture that concatenates representations from two deep convolutional networks on images and sound clips. That is, we mainly use \( N_v^a \) for \( N_v \) in Equation (1), and we will focus on this setting to evaluate our proposed approaches of exploiting audio knowledge.

Furthermore, pre-training on related datasets helps accelerate the training process and improving the performance on the new dataset, especially on a relatively small dataset. For our task, the paired data samples are limited, and our preliminary experiments show that the two networks \( N_v \) and \( N_a \) benefit a lot from pre-training on the AID dataset [30] for classifying scenes from aerial images, and AudioSet [9] for recognizing 527 audio events from sound clips [29]. \( N_v^{(0)} \) and \( N_a^{(0)} \) denote the pre-trained networks. We maintain the notations \( N_v \), \( N_a \) and \( N_v^a \) for the networks that are initialized from the pre-trained weights, for the reason that we will mainly base on the networks that are initialized from the pre-trained weights in the following. In the cases where networks are initialized with random weights, we add a dagger notation \( \dagger \), like \( L_s^\dagger \), which differs from \( L_s \) only in the initialization of networks. The performances of randomly initialized networks will also be reported in the experimental section, presenting the contribution from each modality and from the pre-trained model.

4.2 Preservation of Audio Source Knowledge

For our task of aerial scene recognition, audio knowledge is expected to be helpful since the scene information is related to sound events. While initializing the
network by the pre-trained weights is an implicit way of transferring the knowledge to the main task, the audio source knowledge may easily be forgotten during fine-tuning. Without audio source knowledge, the model can hardly recognize the sound events, leading to a random confusion between sound events and scenes.

For preserving the knowledge, we propose to record the soft responses of target samples from the pre-trained model and retain them during fine-tuning. This simple but efficient approach [10] is named as knowledge distillation, for distilling the knowledge from an ensemble of models to a single model, and has also been used in domain adaptation [27] and lifelong learning [18]. All of them encourage to preserve the source knowledge by minimizing the KL divergence between the responses from the pre-trained model and the training model. For avoiding the saturated regions of the softmax, the pre-activations are divided by a large scalar, called temperature [10], to provide smooth responses, with which the knowledge can be easily transferred.

However, for the reason that the audio event task is a multi-label recognition, $f_e(x, N_a)$ is activated by the sigmoid function, instead of softmax, so the knowledge distillation technique is implemented by a sum of KL divergences, instead of a single one:

$$L_{KL|N_a} = \mathbb{E}_{a \sim \mu} \left[ \sum_i \mathcal{C} \left( y_i(a) \left\| \sigma_i(f_e(a, N_a)) \right. \right) \right], \quad (3)$$

![Fig. 5.](image)

Illustration of the main task and three cross-task transfer approaches (best viewed in color). On the left of this figure, our model takes a paired data sample $x$ of an image $v$ and a sound clip $a$ as input, and extracts representations from different combinations of modalities and models (shown in different colors); On the right, the top-left block introduces our main task of aerial scene recognition, and the rest three blocks present the three cross-transfer approaches.
where \( C(p|q) = p \log \left( \frac{p}{q} \right) + (1 - p) \log \left( \frac{1 - p}{1 - q} \right) \) is a KL divergence on binary recognition, \( \sigma \) is the sigmoid function, \( y(a) = \sigma(f_e(a, N_a^{(0)})) \), \( y_i(a) \) is the \( i \)-th element of \( y(a) \), indicating the probability of \( i \)-th sound event happening in sound clip \( a \), predicted by the pre-trained model \( N_a^{(0)} \), while \( \sigma_i(f_e(a, N_a)) \) is predicted by the training model \( N_a \).

From another viewpoint, knowledge distillation with high temperature is equivalent to minimizing the squared Euclidean distance (SQ) between the pre-activated outputs [10]. Thereby, apart from Equation (3), we also propose to preserve the source knowledge by directly comparing the pre-activated outputs:

\[
L_{SQ|N_a} = \mathbb{E}_{a \sim \mu} \left[ \left\| f_e(a, N_a) - f_e(a, N_a^{(0)}) \right\|_2^2 \right].
\] (4)

### 4.3 Audiovisual Representation for Multi-Task

Different from the idea of preserving the knowledge within the audio modality, we encourage our model, along with the visual modality, to learn a mutual representation that recognizes the scenes and the sound events simultaneously. Specifically, we optimize to solve the sound event recognition task using the concatenated representation, with knowledge distillation technique:

\[
L_{KL|N_{v+a}} = \mathbb{E}_{x \sim \mu} \left[ \sum_i C( y_i(a) \| \sigma_i(f_e(x, N_{v+a}))) \right].
\] (5)

Similar to (4), we also propose to minimize the squared Euclidean distance between the pre-activated outputs:

\[
L_{SQ|N_{v+a}} = \mathbb{E}_{x \sim \mu} \left[ \left\| f_e(x, N_{v+a}) - f_e(a, N_a^{(0)}) \right\|_2^2 \right].
\] (6)

This simultaneous multi-task technique is very common within one single modality, such as solving depth estimation, surface normal estimation and semantic segmentation from one single image [8], or recognizing acoustic scenes and sound events from audio [14], hoping that the model can learn the underlying relationships among the tasks. We implement this idea by either Equation (5) or (6), to encourage the model to solve two tasks simultaneously and find the underlying relation between sound events and scenes.

### 4.4 Sound Events in Different Scenes

The two previously proposed approaches are based on the multi-task learning framework, solving two tasks in parallel, either using different or the same representations, in order to preserve the audio source knowledge or learn an underlying relation between aerial scenes and sound events. Here, we propose an explicit way for directly modeling the relation between scenes and sound events, and creating the bridge of transferring the knowledge between two modalities.
We employ the paired image-audio data samples from our built dataset as introduced in Section 3, analyze the happening sound events in each scene, and obtain the posteriors given one scene. Then instead of predicting the probability of sound events by the network, we estimate this probability distribution \( p(e) \) with the help of posteriors \( p(e|s_k) \) and the predicted probability of scenes \( p(s) \):

\[
p(e) = \sum_k p(s_k)p(e|s_k) ,
\]

where \( p(s_k) = z_k(f_s(x, N_{v+a})) \), recalling that \( z \) is the softmax and \( p(s_k) \) is the probability of the \( k \)-th scene, and the posteriors \( p(e|s_k) \) is obtained by averaging \( y(a) = \sigma(f_e(a, N_a^{(0)})) \) over all samples that belong to the scene \( s_k \). This estimation \( p(e) \) is in fact the compound distribution that marginalizes out the probability of scenes, while we search for the optimal scene probability distribution \( p(s) \) (ideally one-hot) through aligning \( p(e) \) with soft responses:

\[
LE_1 = \mathbb{E}_{x \sim \mu} \left[ \sum_i C(y_i(a) \parallel p(e_i)) \right].
\]

Besides estimating the probability of each sound event happening in a specific scene, we also investigate possible concomitant sound events in this scene. Some sound events may largely overlap under a given scene, and this coincidence can be used as a characteristic for recognizing scenes. We propose to extract this characteristic from \( y(a) \) of all samples \( a \) that belong to this specific scene.

We note \( P(e|s_k) \in \mathbb{R}^{n_k \times c} \) as the sound event probabilities of \( n_k \) samples in the scene \( s_k \), where each row of \( P(e|s_k) \) is each sample’s probabilities of sound events in the scene \( s_k \). Then we extract the eigenvector \( d_k \) that corresponds to the largest eigenvalue of the Gram matrix \( P(e|s_k)^T P(e|s_k) \). This eigenvector \( d_k \) indicates the correlated sound events and quantifies their relevance in the scene \( s_k \) by the direction of this vector. We thus propose to align the direction of \( d_t \), the event relevance of the ground truth scene \( s_t \), with the estimated \( p(e) \):

\[
LE_2 = \mathbb{E}_{(x,t) \sim \mu} \left[ \cosine(p(e), d_t) \right] .
\]

Equation (8) and (9) have provided a way of building the connection between scenes and sound events. In the experiments, we usually use them together:

\[
LE = LE_1 + \beta LE_2 ,
\]

where \( \beta \) is a hyper-parameter controlling the importance of \( LE_2 \).

5 Experiments

In this section, we conduct comparison experiments, as well as some ablation studies, to validate the proposed approaches on the ADVANCE dataset.
5.1 Implementation Details

Our built ADVANCE dataset is employed for evaluation, where 70% image-sound pairs are for training, 10% for validation, and 20% for testing. Note that, these three sub-sets do not share audiovisual pairs that are collected from the same coordinate. Before feeding the recognition model, we sub-sample the sound clips at 16 kHz. Then, following [29], the short-term Fourier transform is computed using a window size of 1024 and a hop length of 400. The generated spectrogram is then projected into the log-mel scale to obtain an audio matrix in $\mathbb{R}^{T \times F}$, where the time $T = 400$ and the frequency $F = 64$. Finally, we normalize each feature dimension to have zero mean and unit variance. The image data are all resized into $256 \times 256$, and horizontal flipping, color, and brightness jittering are used as data augmentation means.

In the network setting part, the visual pathway employs the AID pre-trained ResNet-101 for modeling the scene content [30] and the audio pathway adopts the AudioSet pre-trained ResNet-50 for modeling the sound content [29]. The whole network is optimized via an Adam optimizer with a weight decay rate $1 \times 10^{-4}$ and a relatively small learning rate $1 \times 10^{-5}$, as both backbones have been pre-trained from external knowledge. By using grid search strategy, the hyper-parameters of $\alpha$ and $\beta$ are set as 0.1 and 0.001, respectively. We adopt the weighted-averaging precision, recall and F-score metrics for evaluation, which are more convincing when faced with uneven distribution of scene classes.

5.2 Aerial Scene Recognition

Table 1 shows the recognition results of different learning approaches under the unimodal and multimodal scenario, from which we have four points should pay attention to. Firstly, according to the unimodal results, the sound data can provide a certain reference for different scene categories, although it is significantly worse than image-based results. Such phenomenon reminds us that we can take advantage of the audio information to improve recognition results further. Secondly, we recognize that simply using the information from both modalities does not bring benefits but slightly lowers the results (72.85 vs. 72.71 in F-score). This could be because the pre-trained knowledge for audio modality may be forgotten or the audio messages are not fully exploited just with the rough scene labels. Thirdly, when the sound event knowledge is transferred for the scene modeling, we have considerable improvements for all of the proposed approaches. The results of $L_{SQ|Na}$ and $L_{KL|Na}$ show that preserving audio event knowledge is an effective means for better exploiting audio messages for scene recognition, and the performance of $L_{SQ|N_{e+a}}$ and $L_{KL|N_{e+a}}$ demonstrates that transferring the unimodal knowledge of sound events to the multimodal network can help to learn better mutual representation of scene content across modalities. Fourthly, among all the compared approaches, our proposed $L_E$ approach shows the best results, as it better imposes the sound event knowledge by imposing the underlying relation between scenes and sound events.
Table 1. Aerial scene recognition results on the ADVANCE dataset from 5 different runs, where the approaches with the symbol + mean they are weightedly combined with $L_s$. ± means standard deviation.

| Inputs | Sound | Image | Image+Sound |
|--------|-------|-------|-------------|
| Approaches | $L_s$ | $L_s$ | $L_s$ | $L_{SQ|N_a}$ | $L_{KL|N_a}$ | $L_{SQ|N_{a+a}}$ | $L_{KL|N_{a+a}}$ | $L_E$ |
| Precision | 30.46 | 74.05 | 73.93 | 74.41 | 74.89 | 74.96 | 74.98 | 75.25 |
| ±0.66 | ±0.42 | ±0.80 | ±0.62 | ±0.56 | ±0.35 | ±0.33 | ±0.34 |
| Recall | 32.99 | 72.79 | 72.91 | 73.51 | 73.83 | 74.27 | 74.19 | 74.79 |
| ±1.20 | ±0.62 | ±0.74 | ±0.45 | ±0.44 | ±0.37 | ±0.34 | ±0.34 |
| F-score | 28.99 | 72.85 | 72.71 | 73.34 | 73.83 | 74.13 | 74.14 | 74.58 |
| ±0.51 | ±0.57 | ±0.76 | ±0.47 | ±0.47 | ±0.32 | ±0.33 | ±0.25 |

We use the CAM technique [36] to highlight the parts of the input image that make significant contributions to identifying the specific scene category. Fig. 6 shows the comparison of the visualization results and the predicted probabilities of the ground-truth label among different approaches. By resorting to the sound event knowledge, as well as its association with scene information, our proposed model can better localize the salient area of the correct aerial scene and provide a higher predicted probability for the ground-truth category, e.g., the harbour and bridge class.

Apart from the multimodal setting, we have also conducted more experiments under the unimodal settings, shown in Table 2, for presenting the contributions from pre-trained models, and verifying the benefits from the sound event knowledge on the aerial scene recognition. For these unimodal experiments, we keep only one modal data input and set the other to zeros. When only the audio data are considered, the sound event knowledge is transferred within the audio modality and thus $L_{SQ|N_a}$ is equivalent to $L_{SQ|N_{a+a}}$, similarly for the visual modality case. Comparing the results of randomly initializing the weights \(i.e. L_s^i\) and initializing with the pre-trained weights \(i.e. L_s\), we find that initializing the network from the pre-trained model can significantly prompt the performance, which confirms that pre-training from a large-scale dataset benefits the learning task on the small datasets. Another remark from this table is that the results of the three proposed approaches show that both unimodal networks can take advantage of the sound event knowledge to achieve better scene recognition performance. It further validates the generalization of the proposed approaches, either in the multimodal or the unimodal input case. Compared with the multi-task framework of $L_{SQ|N_{a+a}}$ and $L_{KL|N_{a+a}}$, the $L_E$ approach can better utilize the correlation between sound event and scene category via the statistical posteriors.

5.3 Ablation Study

In this subsection, we directly validate the effectiveness of the scene-to-event transfer term $L_{E_1}$ and the event relevance term $L_{E_2}$, without the supervision
from the scene recognition objective of $L_s$. Table 3 shows the comparison results. By resorting to the scene-to-event transfer term, performing sound event recognition can reward the model the ability to distinguish different scenes. When further equipped with the event relevance of the scenes, the model can have higher performance. This demonstrates that cross-task transfer can indeed provide reasonable knowledge if the inherent correlation between these tasks are well exploited and utilized. By contrast, as the multi-task learning approaches do not well take advantage of this knowledge, the scene recognition performance remains at the chance level.

To better illustrate the correlation between aerial scenes and sound events, we further visualize the embedding results. Specifically, we use the well-trained cross-task transfer model to predict the sound event distribution on the testing set. Ideally, the sound event distribution can separate the scenes from each other, since each scene takes a different sound event distribution. Hence, we use t-SNE [19] to visualize the high-dimensional sound event distributions of different scenes. Fig. 7 shows the visualization results, where the points in different color mean different scene categories. As $L_{SQ|Ne}$ is performed within the audio modality, the sound event knowledge cannot well transfer to the entire model,
Table 2. Unimodal aerial scene recognition results on the ADVANCE dataset from 5 different runs, where † means random initialization and the approaches with the symbol + mean they are weightedly combined with $L_s$.

| Modality | Approaches | $L_s$† | $L_s$ | $+L_{SQ|N_{v+a}}$ | $+L_{KL|N_{v+a}}$ | $+L_E$ |
|----------|------------|--------|--------|-------------------|-------------------|--------|
| Sound    | Precision  | 21.15±0.68 | 30.46±0.66 | 31.64±0.65 | 30.00±0.86 | 31.14±0.30 |
|          | Recall     | 24.54±0.67 | 32.99±1.20 | 34.68±0.49 | 34.29±0.35 | 33.80±1.03 |
|          | F-score    | 21.32±0.42 | 28.99±0.51 | 29.31±0.71 | 28.51±0.99 | 29.66±0.13 |
| Image    | Precision  | 64.45±0.97 | 74.05±0.42 | 74.86±0.94 | 74.36±0.85 | 73.97±0.39 |
|          | Recall     | 64.59±1.12 | 72.79±0.62 | 74.11±0.89 | 73.40±0.84 | 73.47±0.52 |
|          | F-score    | 64.04±1.07 | 72.85±0.57 | 73.98±0.92 | 73.52±0.85 | 73.44±0.45 |

Fig. 7. The aerial scene data embeddings indicated by the corresponding sound event distribution, where the points in different color mean in different scene categories.

leading to the mixed scene distribution. By contrast, as $L_{SQ|N_{v+a}}$ transfers the sound event knowledge into the multimodal network, the predicted sound event distribution can separate different scenes to some extent. By introducing the correlation between scenes and events, i.e., $L_E$, different scenes can be further disentangled, which confirms the feasibility and merits of cross task transfer.

6 Conclusions

In this paper, we explore a novel multimodal aerial scene recognition task that considers both visual and audio modality data. We have constructed a dataset, named ADVANCE, with labeled paired audiovisual worldwide samples, for facilitating the research on this topic. Observing that simply combining both modalities cannot easily promote the scene recognition performance, we propose to directly transfer the sound event knowledge to the scene recognition task for the reasons that the sound events are related to the scenes and that this underlying relation is not well exploited. Three kinds of effective cross-task transfer approaches are thus developed to improve the performance on the scene recog-
Table 3. Aerial scene recognition results on the ADVANCE dataset, where only the sound event knowledge is considered in the training stage.

| Approaches  | $L_{E_1}$ | $L_{E_1} + \beta L_{E_2}$ | $L_{KL|N_{v+1}}$ | $L_{SQ|N_{v+1}}$ |
|-------------|----------|---------------------------|-------------------|-------------------|
| Precision   | 43.37±0.59 | 54.23±1.14 | 3.08±0.14 | 2.95±0.07 |
| Recall      | 49.26±0.36 | 52.57±0.72 | 9.69±0.43 | 9.28±0.17 |
| F-score     | 42.50±0.42 | 48.50±0.85 | 4.46±0.20 | 4.24±0.07 |

dition, and amounts of experimental results show their effectiveness, confirming the benefit of exploiting the audio knowledge for the aerial scene recognition.
References

1. Assael, Y.M., Shillingford, B., Whiteson, S., De Freitas, N.: Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599 (2016)
2. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in neural information processing systems. pp. 892–900 (2016)
3. Baltrusaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence 41(2), 423–443 (2018)
4. Castelluccio, M., Poggi, G., Sansone, C., Verdoliva, L.: Land use classification in remote sensing images by convolutional neural networks. arXiv preprint arXiv:1508.00092 (2015)
5. Chaplot, D.S., Lee, L., Salakhutdinov, R., Parikh, D., Batra, D.: Cross-task knowledge transfer for visually-grounded navigation (2018)
6. Cheng, G., Yang, C., Yao, X., Guo, L., Han, J.: When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. IEEE transactions on geoscience and remote sensing 56(5), 2811–2821 (2018)
7. Ehrlich, M., Shields, T.J., Almaev, T., Amer, M.R.: Facial attributes classification using multi-task representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 47–55 (2016)
8. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
9. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Pinkal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017)
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015), http://arxiv.org/abs/1503.02531
11. Hu, D., Li, X., et al.: Temporal multimodal learning in audiovisual speech recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3574–3582 (2016)
12. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9248–9257 (2019)
13. Hu, D., Wang, Z., Xiong, H., Wang, D., Nie, F., Dou, D.: Curriculum audiovisual learning. arXiv preprint arXiv:2001.09414 (2020)
14. Imoto, K., Tonami, N., Koizumi, Y., Yasuda, M., Yamanishi, R., Yamashita, Y.: Sound event detection by multitask learning of sound events and scenes with soft scene labels. arXiv preprint arXiv:2002.05848 (2020)
15. Kato, H., Harada, T.: Image reconstruction from bag-of-visual-words. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 955–962 (2014)
16. Kyrkou, C., Theocharides, T.: Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
17. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06). vol. 2, pp. 2169–2178. IEEE (2006)
18. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 384–393 (2017)
19. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research 9(Nov), 2579–2605 (2008)
20. Nogueira, K., Penatti, O.A., Dos Santos, J.A.: Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recognition 61, 539–556 (2017)
21. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 631–648 (2018)
22. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Learning sight from sound: Ambient sound provides supervision for visual learning. In: International Journal of Computer Vision (IJCV) (2018)
23. Risojević, V., Babić, Z.: Aerial image classification using structural texture similarity. In: 2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). pp. 190–195. IEEE (2011)
24. Risojević, V., Babić, Z.: Orientation difference descriptor for aerial image classification. In: 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP). pp. 150–153. IEEE (2012)
25. Salem, T., Zhai, M., Workman, S., Jacobs, N.: A multimodal approach to mapping soundscapes. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (2018)
26. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 247–263 (2018)
27. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4068–4076 (2015)
28. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3360–3367. IEEE (2010)
29. Wang, Y.: Polyphonic sound event detection with weak labeling. PhD thesis (2018)
30. Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X.: Aid: A benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing 55(7), 3965–3981 (2017)
31. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740 (2020)
32. Yang, K., Liu, Z., Lu, Q., Xia, G.S.: Multi-scale weighted branch network for remote sensing image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–10 (2019)
33. Yang, Y., Newsam, S.: Comparing sift descriptors and gabor texture features for classification of remote sensed imagery. In: 2008 15th IEEE international conference on image processing. pp. 1852–1855. IEEE (2008)
34. Zhang, F., Du, B., Zhang, L.: Scene classification via a gradient boosting random convolutional network framework. IEEE Transactions on Geoscience and Remote Sensing 54(3), 1793–1802 (2015)
35. Zheng, W.L., Liu, W., Lu, Y., Lu, B.L., Cichocki, A.: Emotionmeter: A multi-modal framework for recognizing human emotions. IEEE transactions on cybernetics 49(3), 1110–1122 (2018)
36. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
37. Zou, Q., Ni, L., Zhang, T., Wang, Q.: Deep learning based feature selection for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters 12(11), 2321–2325 (2015)