Remote Sensing Image Generation From Audio

Zhiyuan Zheng, Jun Chen, *Member, IEEE*, Xiangtao Zheng, *Member, IEEE*, and Xiaoqiang Lu, *Senior Member, IEEE*

***Abstract*— Generating image from other modal data has attracted much attention in cross-modal studies, since the generated image offers intuitive vision information. Unlike the previous works which generate an image from text, a novel task is introduced, generating an image from audio. However, semantic gap intrinsically exists in cross-modal data, which disturbs the generative results. In order to explore the relevance between the audio and image, a novel reranking audio-image translation method is proposed. The proposed method: 1) maps the audio and image into a uniform feature space; 2) designs an audio-audio matching network to match the related audio; and 3) adopts an audio-image matching network for every matched audio to generate a related image, and the most frequent image is voted as the final result. Extensive experiments on two remote sensing cross-modal data sets demonstrate that the proposed method can visualize the content of audio.**

***Index Terms*— Cross-modal, generation, reranking.**

I. Introduction

THIS letter tries to generate remote sensing image from audio. The motivation of this task is inspired by the fact that infants can learn to identify objects through talking without knowing text words. Generating remote sensing image from audio provides a natural way for human to acquire an expected image with talking rather than typing, which enhances the experience of human–computer interaction. The generated remote sensing image concretely reveals the semantic meaning of audios, which can be applicable for auxiliary understanding of geographical information and retrieval systems.

Recently, many methods have been proposed to generate an image from text [1]–[3]. Reed *et al.* [1] utilized generative adversarial network (GAN) [4] to generate an image from text. Zhang *et al.* [2] proposed a coarse-to-fine generative network to get a photo-realistic image from text. Furthermore, an attention mechanism was adopted to improve the effect of fine-grained pixel synthesis based on text [3]. However, speaking is a more natural way than texting for human to express, which is more convenient to control the generation of an image. Thus, an efficient way is to generate an image from audio, which is natural and efficient human–machine interactive.

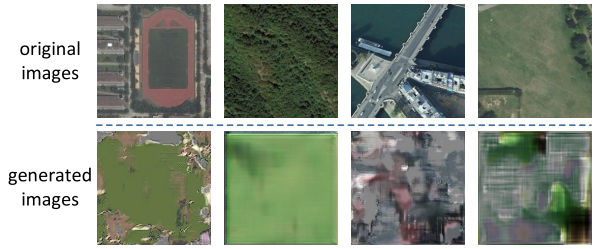

Fig. 1. Generating remote sensing image from audio by the GAN method of [5]. (Top) The original images in data set which show playground, forest, bridge, and meadow. (Bottom) The corresponding generated images.

The methods for generating an image from audio can be roughly divided into: generative adversarial method and feature matching method. The generative adversarial method translates the audio into an image by GAN. For example, Bejiga *et al.* [5] applied GAN to generate retro-remote sensing image from ancient text. Similarly, by replacing the text with audio, a remote sensing image can be generated from an audio. However, the generative adversarial method cannot generate a reasonable image, as shown in Fig. 1. The reasons for the terrible generative result may be listed as follows:

1) The audio is a continuing signal, which cannot be vector-encoded for every word as text does [6]. It is difficult for a GAN to acquire an efficient map between the audio and image.
2) The training of the GAN is unstable and probably leads to collapse [7].

The feature matching method generates an image from audio by measuring the semantic distance between different modal data. For example, Mao *et al.* [8] utilized fused features to measure the semantic correlation between the audio and image. Wang *et al.* [9] proposed collective semantic metric learning framework (CSMLF) to map different modal data into a common feature space. However, it is difficult to find the relevant images due to the semantic gap between different modal data [10].

Compared with generative adversarial method, feature matching method is more stable to learn the semantic correlation between different modal data. The feature matching method only needs to measure the distance of features rather than learn the whole distribution between different

Manuscript received October 11, 2019; revised March 31, 2020; accepted April 21, 2020. Date of publication May 15, 2020; date of current version May 21, 2021. This work was supported in part by the National Natural Science Found for Distinguished Young Scholars under Grant 61925112, in part by the National Key Research and Development Program of China under Grant 2017YFB0502900, in part by the National Natural Science Foundation of China under Grant 61806193, Grant 61772510, and Grant 61876135, in part by the Young Top-Notch Talent Program of Chinese Academy of Sciences under Grant QYZDB-SSW-JSC015, and in part by the CAS "Light of West China" Program under Grant XAB2017B26. *(Corresponding author: Xiangtao Zheng.)*

Zhiyuan Zheng is with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China, and also with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China.

Jun Chen is with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China.

Xiangtao Zheng and Xiaoqiang Lu are with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xiangtaoz@gmail.com).

Color versions of one or more of the figures in this letter are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2020.2992324

modality like the GAN [7]. In addition, feature matching method generates the image from audio based on the nearest feature distance. According to the nearest feature distance, the generated image is retrieved from existing database in which the contained images are photo-realistic. Although the generated image may not match the audio content, it can be ensured that the generated image is photo-realistic.

In order to generate a clear image from audio, a reranking audio-image translation method is proposed. Though many effective reranking methods exist in remote sensing image retrieval, such as those studied by Tang *et al.* [11] who established the reranked list by using editing scheme and optimizing sorted loss, the problem of semantic gap is not concerned. Whereas the reranking mechanism in the proposed method is designed to alleviate the problem of semantic gap between different modal data. The proposed method translates audio to remote sensing image in three stages: feature representation, audio matching, and image generation. First, to reduce the semantic gap between different modal data, both image and audio can be mapped into a uniform feature space. Hash code is exploited for its lower space and computes fast [12], which is more suitable than the original feature for the proposed reranking mechanism with lots of similarity calculation. Then, an audio-audio matching network is trained to match the related audio. Several semantically similar audio are selected by the distance of hash codes. The matched audio shares the same topic with the input audio. Finally, a pretrained audio-image matching network is applied for every matched audio to generate the related image. The final image is generated according to the matched audio and contains the most related vision information with the original input audio. Note that the proposed method generates image directly from audio. Using text as the transition to implement the audio-text-image framework would induce superposed semantic gap and require text annotation.

Overall, the main contributions of our work can be summarized as follows.

1) As far as we know, a novel task which generates remote sensing image from audio is introduced in this letter. It extends the way for humans to interact with the remote sensing image.
2) A novel reranking hash-based audio-image (RHAI) translation method is proposed, which guarantees the clarity of the generated image and outperforms previous related works.
3) A reranking mechanism is embedded in RHAI to narrow the semantic gap between audio and remote sensing image.

We organize the rest of this letter as follows. The proposed method is depicted in Section II. The experimental results are presented in Section III. Section IV summarizes the work and future improvements.

II. Proposed Method

The flowchart of the proposed method is shown in Fig. 2. We elaborate on the proposed method with three steps. First, feature representations of remote sensing image and audio are described. Furthermore, the process of matching semantically similar audio is detailed. Finally, the reranking mechanism is introduced to generate the final image.

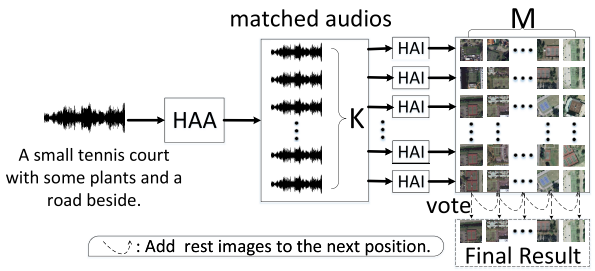

Fig. 2. Flowchart of RHAI which contains three stages. Stage 1: utilizing DNNs to extract hash codes from audio and image. Stage 2: K similar audios are matched to input audio by designed HAA matching network. Stage 3: Through HAI matching network, M images are generated from every matched audio. The most frequent image in each column is voted as the final output after the rest images in previous column are merged.

A. Feature Representation

A deep neural network (DNN) is embedded in the proposed method to get common feature representation from image and audio. Concretely, the pretrained AlexNet is adopted for the image to extract feature, which consists of a down-sample convolutional neural network (CNN) and fully connected (FC) layers. It should be noted that we replace the final FC layer of AlexNet with another which fits the required output length of hash code. Let $f_{\text{image}}(\cdot)$ denote the process of feature extraction of image. Then given a remote sensing image set $X = \{x_i\}_{i=1}^{N_x}$, where N_x denotes the quantity of images, the binary-like feature of single image can be gained through

$$b_i^{(x)} = f_{\text{image}}(x_i) \tag{1}$$

where $b_i^{(x)} \in \mathbb{R}^d$ represents the d-dimensional binary-like feature of the ith image.

Audio is a continuous signal, the words of which cannot be easily distinguished by a computer. To quantify the continuously varying audio signal, Mel-frequency cepstrum coefficients (MFCCs) are adopted as the preprocess of audio data [13]. The delta–deltas feature is extracted from the audio through the method of MFCC, and then flattening the delta–deltas feature to be a vector for the input of FC layers. Given an audio database $Y = \{y_j\}_{j=1}^{N_y}$, where N_y denotes the quantity of audio, the MFCC feature is computed by

$$m_j = \text{flatten}\big(\text{mfcc}\big(y_j\big)\big) \tag{2}$$

where $m_j \in \mathbb{R}^{d_m}$ represents d_m-dimensional delta–deltas feature of jth audio, mfcc denotes the process of MFCC, and flatten denotes the operation of flattening.

After the MFCC feature extraction, the following process of audio is similar with the process of remote sensing image. Let $f_{\text{audio}}(\cdot)$ represent the feature extraction with the neural network, which contains only four FC layers. The d-dimensional binary-like feature $b_j^{(y)} = f_{\text{audio}}(m_j)$.

The ultimate representation of remote sensing image and audio is hash code, which contains only $+1$ and -1. However, nonlinear neural network cannot directly acquire the exact discrete number. With sign function, the binarization is computed

below

$$h_i^{(x)} = \text{sign}\left(b_i^{(x)}\right)$$
$$h_j^{(y)} = \text{sign}\left(b_j^{(y)}\right) \tag{3}$$

where $h_i^{(x)} \in \{-1, +1\}^d$ denotes the hash code of the ith remote sensing image, $h_j^{(y)} \in \{-1, +1\}^d$ denotes the hash code of the jth audio, and d represents the length of hash code.

B. Audio Matching

Given the input audio, a semantically similar audio is selected. As shown in Fig. 3(a), hash-based audio-audio (HAA) is trained to generate appropriate hash codes from audio, the appropriate hash codes should have a small difference in similar audio pair while the difference in dissimilar audio pair should be enlarged. The way to measure "distance" between hash codes of audio is defined as

$$D_h\left(h_{j_1}^{(y)}, h_{j_2}^{(y)}\right) = \frac{1}{2}\left(d - \left\langle h_{j_1}^{(y)}, h_{j_2}^{(y)} \right\rangle\right) \tag{4}$$

where $D_h(\cdot, \cdot)$ represents the Hamming distance between hash codes,$\langle \cdot, \cdot \rangle$ denotes the inner product, j_1 and j_2 are the different number in the range of the quantity of audio data.

For audio-audio matching, A similarity set $S_{\text{AA}} = \{s_{j_1 j_2}\}$ is needed to guide the training process. If the j_1th audio and j_2th audio have the same topic, the similarity label $s_{j_1 j_2} = 1$, otherwise $s_{j_1 j_2} = 0$. The discrete hash code cannot be utilized to update weights of neural network because the sign function is nondifferentiable. However, the binary-like feature generated from neural network can be seen as the approximate representation of discrete hash code. Hence, the loss function is designed based on binary-like feature, and the difference between binary-like features of audio should be reduced if $s_{j_1 j_2} = 1$, or be enlarged if $s_{j_1 j_2} = 0$. We denote the connection between binary-like feature and similarity label by constructing a binomial logistic function

$$P\left(s_{j_1 j_2} \middle| b_{j_1}^{(y)}, b_{j_2}^{(y)}\right) = \begin{cases} \sigma\left(\left\langle b_{j_1}^{(y)}, b_{j_2}^{(y)} \right\rangle\right) & s_{j_1 j_2} = 1 \\ 1 - \sigma\left(\left\langle b_{j_1}^{(y)}, b_{j_2}^{(y)} \right\rangle\right) & s_{j_1 j_2} = 0 \end{cases} \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function $\sigma(y) = (1/(1 + e^{-y}))$, the inner product $\langle b_{j_1}^{(y)}, b_{j_2}^{(y)} \rangle$ reveals the extent of similarity between binary-like features of input audio pair.

With maximum likelihood estimate (MLS), the cross-entropy loss derived from (5) is defined as

$$l_{\text{cross}}^{\text{HAA}} = -\sum_{s_{j_1 j_2} \in S} \left(s_{j_1 j_2}\left\langle b_{j_1}^{(y)}, b_{j_2}^{(y)} \right\rangle - \log\left(1 + e^{\left\langle b_{j_1}^{(y)}, b_{j_2}^{(y)} \right\rangle}\right)\right). \tag{6}$$

Besides that, a quantization loss is adopted to reduce the difference between binary-like feature and binary hash code, euclidean metric is adopted to compute the quantization loss

$$l_q^{\text{HAA}} = \sum_{j_1=1}^{N_{y_1}} \left\| b_{j_1}^{(y)} - h_{j_1}^{(y)} \right\|_2 + \sum_{j_2=1}^{N_{y_2}} \left\| b_{j_2}^{(y)} - h_{j_2}^{(y)} \right\|_2 \tag{7}$$

where the sum of N_{y_1} and N_{y_2} is N_y. In order to get the hash code with maximal information representation, a balance loss is designed as

$$l_b^{\text{HAA}} = \sum_{j_1=1}^{N_{y_1}} \text{avg}\left(b_{j_1}^{(y)}\right) + \sum_{j_2=1}^{N_{y_2}} \text{avg}\left(b_{j_2}^{(y)}\right) \tag{8}$$

where avg$(\cdot)$ calculates the mean value of the elements in vector. Then the integrated loss of HAA is

$$L^{\text{HAA}} = l_{\text{cross}}^{\text{HAA}} + \alpha l_q^{\text{HAA}} + \beta l_b^{\text{HAA}} \tag{9}$$

where α and β denote weighted parameters.

By minimizing the loss, HAA learns the relevance of audio-audio pair while being capable to match related audios to input audio.

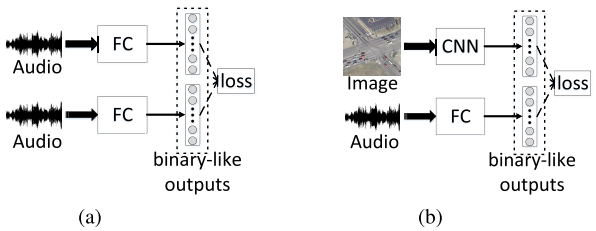

Fig. 3. Two submodules of RHAI, FC denotes FC layers and CNN denotes convolutional neural network such as AlexNet or VGG-net. (a) HAA is to measure the semantic distance between audios. (b) HAI is to measure the semantic distance between audio and image.

C. Remote Sensing Image Generation

The proposed method generate image from every matched audio and select the most frequent image as the final result. Compared with single audio, multiple matched audios can provide more semantic information, which is benefit to generate accurate image. After matching K semantically similar audios by HAA, M related images are generated from every matched audio by hash-based audio-image (HAI). The structure of HAI is shown in Fig. 3(b). HAI is designed to generate appropriate hash codes to measure the semantic distance between audio and image. Similar to HAA, a cross-modal similarity set $S_{\text{AI}} = \{s_{ij}\}$ is constructed in the training of HAI, the loss function of HAI is defined as

$$L^{\text{HAI}} = l_{\text{cross}}^{\text{HAI}} + \alpha l_q^{\text{HAI}} + \beta l_b^{\text{HAI}} \tag{10}$$

where L^{HAI} is calculated by replacing b_{j1}^{y} and $b_{j2}^{(y)}$ in L^{HAA} with $b_i^{(x)}$ and $b_j^{(y)}$.

A reranking mechanism is embedded in the image generation stage. At the first position of M which contains the most related K images to the matched audios, image with the highest number n_1 of occurrences is filled. The rest $K - n_1$ images are combined with the K images at second position of M. Similarly, the most frequent image which appears n_2 times in the $2k - n_1$ images occupies the second position and the rest $2K - n_1 - n_2$ images are shifted to the third position, and so on. Until all M remote sensing images are reranked.

The reason to do so is to reduce the semantic gap between audio and remote sensing image. To be specific, the hard circumstance as shown in Fig. 4 can be ameliorated by our reranking mechanism.

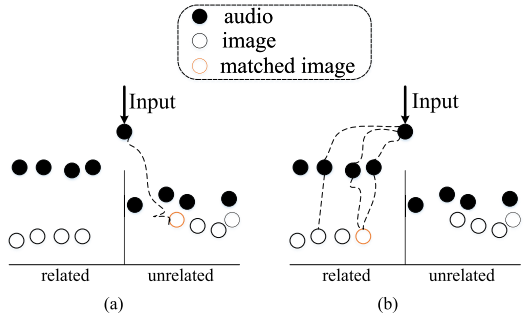

Fig. 4. Generating a remote sensing image from input audio. (a) Unrelated image may have small semantic distance with input audio. (b) With reranking mechanism, the matched audios bridge the gap between input audio and related image. (a) No reranking. (b) With reranking.

D. Computation Complexity

The computational cost of the proposed method includes two parts: 1) finding K related audios to input audio and 2) reranking the related images from the found audios. The first part needs to calculate the hamming distance and sorting, the cost of which is $O(n\log(n))$. In the second part, the cost of retrieving related images is $O(K \cdot m\log(m))$, and counting the most related image requires $O(K)$ computational cost. Then the total computational cost of the proposed method is $O(n\log(n)+(K \cdot m\log(m)+K))$. Here, n denotes the number of audios in retrieval set, m denotes the number of images in retrieval set, and K denotes the number of selective audios in the first stage.

III. Experiments and Analysis

Based on remote sensing image caption data sets [14], audio data are gained by reading the caption. The adopted data sets split into 80% for train and 20% for test. During the procedure of MFCC operation, the number of cepstrum is 12, and the step between successive windows is 0.02 s. We implement all experiments with PyTorch on GTX TITAN X and i7-5930K. Stochastic gradient descent (SGD) is employed to optimize proposed networks. The initial learning rate of HAA is 10^{-2}, which evenly decreases to 10^{-3} in 150 epochs, the weighted parameter α is set to 0.8, and β is set to 1. The initial learning rate of HAI is 10^{-3} and evenly decreases to 10^{-5} in 300 epochs.

A. Data Sets

1) Sydney-Audio: The data set contains seven classes, 613 remote sensing images (RGB format) with related audios which describe the main content of relevant images. The original 500×500 images are resized as 224×224. Through mfcc computing, the audio data are quantified as delta–deltas with 3600 dimensions.

2) UCM-Audio: This data set consists of 2100 remote sensing images which are divided into 21 classes. Every class contains 100 images and every image is expounded by an audio. The sentence in audio describes a complex location relationship of objects in the remote sensing image. The process of the image and audio is the same as above.

B. Evaluation Metrics

The aim of our task is translating audio to semantically similar remote sensing image, which means the generated image should reveal the semantic meaning of audio. We evaluate the semantic correlation between generated image and input audio from two aspects. The first is to evaluate whether the topic of generated image is consistent with that of the input audio. For example, if the input audio describes an airport, the generated image should present the scene about airport, otherwise the result is wrong. Because the generated image is retrieved in image database with label information, the precision and mean average precision (MAP) are employed to evaluate the accuracy of topic matching.

Besides, we use P at 1, P at 5, P at 10 to separately denote the precision of topic matching in the case of generating one, five, and ten remote sensing images.

It is only the basic requirement of our task to ensure the topic consistency achieved between the generated image and input audio. The generated image is also required to highlight detail information (e.g., color, position, and shape) of objects described in audio, which is hard to be evaluated by an objective method. We measure the more complex semantic correlation between the image and audio by human scoring. Concretely, we ask people to classify the generated images from audio as three classes, which are "perfectly matched," "partially matched," and "completely mismatched."

TABLE I
RESULTS OF DIFFERENT METHODS ON SYDNEY-AUDIO

	MAP	P@1	P@5	P@10
PCSMLF [9]	28.43	15.53	15.92	16.41
CMFH [15]	49.46	38.83	38.83	40.58
SePHklr [16]	87.71	88.08	88.15	87.93
DVAN [8]	64.41	68.10	64.14	70.26
TSR [11]	90.28	92.36	92.27	91.49
FAI	94.62	**95.12**	**94.31**	**93.38**
RFAI	**95.33**	93.20	93.58	93.64
HAI	93.21	93.31	92.01	92.01
RHAI	95.31	93.33	93.33	93.32

C. Comparison and Results

Though few studies explore the relationship between audio and remote sensing image, the method of measuring semantic distance between different modality data can be applied to our task. The comparative methods are adopted from cross-modal researches [8], [9], [15], [16] and the reranking research [11]. Except for [8] which used audio data, we modify the input module of the official code of [9], [11], [16], and [15] to fit our data sets. For verifying the improvement effect of RHAI, we evaluate the results of HAI which directly generates the image from audio. We also explore the performance by using original feature in the proposed method, which are denoted by feature-based audio-image (FAI) and reranking FAI (RFAI), they share the same framework with HAI and RHAI.

Tables I and II show the accuracy of topic matching on Sydney-audio and UCM-audio data set. We can see the proposed method gain state-of-the-art result in all metrics. Especially on the UCM-audio data set, the improvement is huge, which indicates the remarkable power of our method for generating topic related image with audio. From Sydney-audio data set to UCM-Audio data set, the results of contrast methods have fallen a lot while the proposed method still acquires high accuracy, which reveals our method is robust to complex data set. The results also reveal that the improvement

TABLE II
RESULTS OF DIFFERENT METHODS ON UCM-AUDIO

	MAP	P@1	P@5	P@10
PCSMLF [9]	7.71	2.86	3.62	4.71
CMFH [15]	17.18	9.52	9.52	9.24
SePHklr [16]	38.63	42.86	39.19	38.24
DVAN [8]	32.38	32.37	33.91	34.34
TSR [11]	86.49	90.03	89.94	89.81
FAI	91.16	94.78	94.39	94.27
RFAI	92.03	94.47	94.16	94.03
HAI	90.72	93.68	93.59	93.48
RHAI	**93.28**	**95.05**	**95.05**	**94.93**

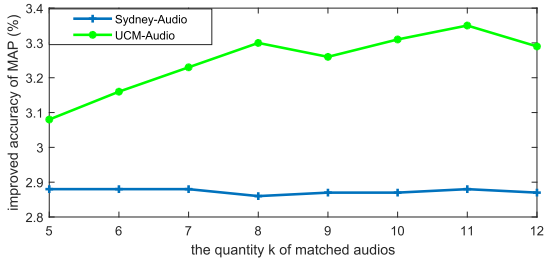

Fig. 5. Variation of improved accuracy of MAP with varying quantity of matched audios. Blue line denotes the variation of improved accuracy on Sydney-audio data set. Green line denotes the variation of improved accuracy on UCM-audio data set.

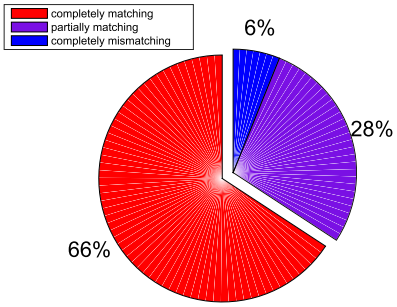

Fig. 6. Results of human evaluation, most of the results are matched perfectly.

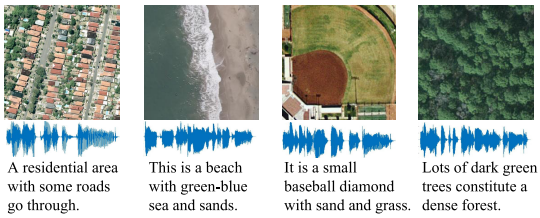

Fig. 7. Some examples of the generated images from input audio.

of using original feature in the proposed reranking mechanism is limited or even useless. And the comparative results of RHAI and two-stage re-ranking (TSR) imply the problem of semantic gap can be well alleviated by the simple reranking mechanism in the proposed method.

It can be seen from the results that RHAI can further improve the accuracy of MAP by 2%–4% on the basis of HAI. The results prove the effectiveness of the proposed reranking method. By adjusting the value of matched audios quantity K, the improved accuracy of MAP varies as shown in Fig. 5. Compared with Sydney-audio data set, UCM-audio data set contains more kinds of data but on which the MAP accuracy can be improved higher, which shows the proposed method has obvious improvement effect on complex data set.

In order to further verify whether the matched image reflects detail information described by audio, we evaluate the test results manually. There are three levels of evaluation criteria: completely matching, partially matching, and completely mismatching. The results of the manual evaluation are shown in Fig. 6. Most of the generated images match the content of related audio. Some of the results are shown in Fig. 7.

IV. CONCLUSION

In this letter, a novel task which generates a remote sensing image from audio is introduced. To complete the task, we propose a method named RHAI to generate a remote sensing image based on the distance between hash codes of image and audio. A novel reranking mechanism is embedded in our method to reduce the semantic gap between image and audio. Compared with other related methods, our proposed method gains the state-of-the-art results on topic matching. And through human evaluation, most of the generated images correspond to the description of sentences in input audio. However, the generated image from the proposed method is retrieved from existing database. In the future, we consider designing a generative network based on the GAN to disentangle the meaning of sentences in audio and to generate a new image.

REFERENCES

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1060–1069.
[2] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.
[3] T. Xu *et al.*, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.
[4] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[5] M. B. Bejiga, F. Melgani, and A. Vascotto, "Retro-remote sensing: Generating images from ancient texts," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 3, pp. 950–960, Mar. 2019.
[6] D. Suris, A. Recasens, D. Bau, D. Harwath, J. Glass, and A. Torralba, "Learning words by drawing images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2029–2038.
[7] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.
[8] G. Mao, Y. Yuan, and L. Xiaoqiang, "Deep cross-modal retrieval for remote sensing image and audio," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens. (PRRS)*, Aug. 2018, pp. 1–7.
[9] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019.
[10] H. Ma, J. Zhu, M. R.-T. Lyu, and I. King, "Bridging the semantic gap between image contents and tags," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 462–473, Aug. 2010.
[11] X. Tang, L. Jiao, W. J. Emery, F. Liu, and D. Zhang, "Two-stage reranking for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5798–5817, Oct. 2017.
[12] Y. Cao, M. Long, B. Liu, and J. Wang, "Deep Cauchy hashing for Hamming space retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1229–1237.
[13] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–55, Jan. 2006.
[14] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.
[15] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5427–5440, Nov. 2016.
[16] Z. Lin, G. Ding, J. Han, and J. Wang, "Cross-view retrieval via probability-based semantics-preserving hashing," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4342–4355, Dec. 2017.