# SSELDNET: A FULLY END-TO-END SAMPLE-LEVEL FRAMEWORK FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

*Daolang Huang*

Aalto University
Department of Computer Science
Espoo, 02150, Finland
daolang.huang@aalto.fi

*Ricardo Falcon Perez*

Aalto University
Department of Signal Processing and Acoustics
Espoo, 02150, Finland
ricardo.falconperez@aalto.fi

## ABSTRACT

Sound event localization and detection (SELD) is a multi-task learning problem that aims to detect different audio events and estimate their corresponding locations. All of the previously proposed SELD systems were based on human-extracted features such as Mel-spectrograms to make the prediction, which required specific prior knowledge in acoustics. In this report, we investigate the possibility to apply representation learning directly to the raw audio and propose an end-to-end sample-level SELD framework. To improve generalization, we applied three data augmentation tricks: sound field rotation, time masking and random audio equalization. The proposed system is evaluated on the TAU-NIGENS Spatial Sound Events 2021 development dataset. The experimental results will be submitted to DCASE 2021 challenge task 3.

***Index Terms*—** Sound event localization and detection, end-to-end, raw audio, time domain, deep learning

## 1. INTRODUCTION

Sound event localization and detection (SELD) is an uprising challenge in the DCASE community, it is a frame-wise based task that aims to simultaneously detect the occurrence of various sound events (SED) and estimate their direction-of-arrival (DOA). It first appeared in DCASE 2019 and was initially treated as separate task learning with individual evaluation metrics. In DCASE 2020 challenge, new evaluation metrics called location-aware detection and class-aware localization were adopted to motivate the development of unified learning systems. After that, some multi-task learning frameworks were proposed to learn joint representations of both SED and DOA. The task of this year stays mostly consistent with former competitions but with non-targeted sound events included in audio samples. These superimposed interference noises make discerning and localizing movable sounds events becomes more challenging.

All the previously proposed systems for SELD used mid-level representations of audio as input formats, such as spectrograms or Mel-frequency cepstral coefficients (MFCCs), which are regarded as a form of prior knowledge constructed by human expertise. However, this feature-based solution requires separate human efforts and is constrained by different parameter settings, e.g. window size, hop size, or filter bank type, which in turn influences the design of model architectures and is considered a sub-optimal solution. Apart from that, audio data is intrinsically disordered among different sound events, thus, it is also a challenge to find the optimal hyper-parameters for hand-engineered features which can adapt to various sound events. To overcome the problem, several papers have investigated the possibilities of directly learning audio representations in the time domain while skipping the construction of acoustic features. This kind of models is often called audio-based end-to-end approaches, which have been explored mainly in speech recognition [1, 2, 3], acoustic scene classification [4, 5] and music auto-tagging [6, 7, 8]. In [9], they tried to build an end-to-end model to do the sound event detection based on the "SampleCNN" architecture, which is widely used in the music auto-tagging domain [10]. However, according to our knowledge, no previous work directly learns the representations based on raw audios in the SELD task.

In this report, we propose an audio-based end-to-end SELD system called **S**ample-level **S**ound **E**vent **L**ocalization and **D**etection network (SSELDnet), which does not depend on the human-extracted features, thus further reducing the human domain knowledge required. We deploy a Sample-level DCNN architecture [7] in the feature embedding part, in which the filter size is several samples long when doing the 1-D convolution in the bottom layers. We also combine the Residual block [11] with the Squeeze-and-Excitation block [12] in the embedding layers, which is proved to be effective for this task. To further improve the results, the original GRU modules in the official baseline system [13] are replaced by "Conformer" blocks [14]. The "Conformer" module is a convolution-augmented transformer architecture, while the transformer [15] can be used to extract long sequence dependencies, and convolution is suitable for refining local features.

The TAU-NIGENS Spatial Sound Events 2021 dataset, used in this competition, only includes 600 60-second audio recordings while containing different interference sound events, making it relatively insufficient for a modern DNN-based system to learn accurate representations. Besides, the input size of raw audio is much larger than human-extracted features, which is more difficult to generalize. To address the issue, applying data augmentation is a necessary step in most of the previously proposed systems [16, 17, 18]. Due to different feature forms, some frequently used data augmentation techniques such as SpecAugment [19] cannot be applied to 1-D raw audios. Besides, some other tricks like mixup [20] and time-stretching may not contribute to the DOA estimation according to our experiments. In fact, there are only limited augmentation methods that can be applied to audio-based system. In our exper-

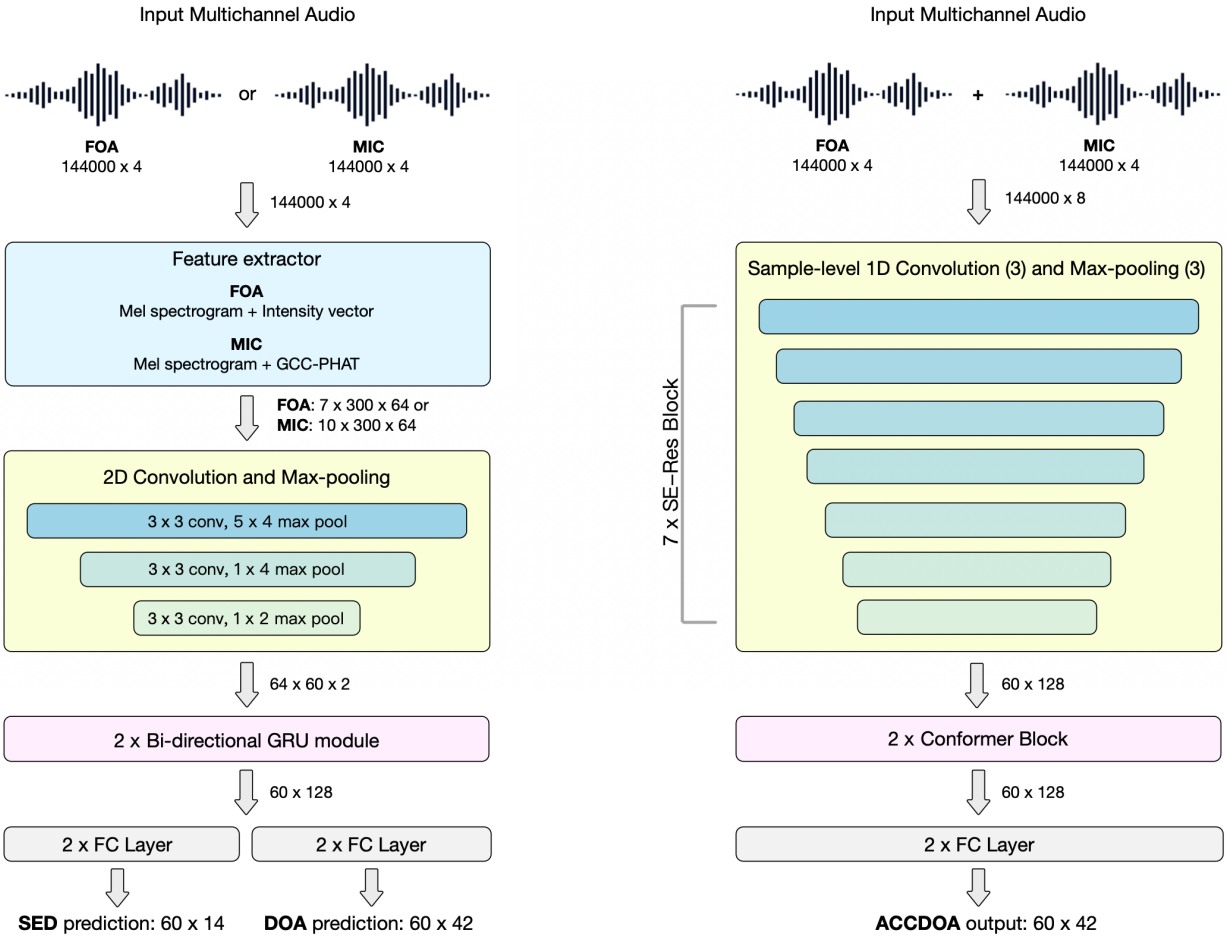


Figure 1: Feature-based CRNN model (left) and proposed audio-based end-to-end SSELDnet (right)

iments, we adopt several approaches to apply data augmentation directly in the time domain. The first method is Sound Field Rotation (SFR) that was first used in [21]. However, they only investigated the feasibility of SFR on the first-order Ambisonics (FOA) dataset. [16] improved their method by adding SFR to the Microphone array (MIC) dataset. Our implementation remains the same with [16]. Another two augmentation tricks that can be applied are Time Masking (TM) and Random Audio Equalization (RAE).

The remainder of the paper is organized as follows. In Section 2, the proposed methods are illustrated in detail, including network architecture and data augmentation. Experimental designs and results are shown in Section 3. Finally, we summarized the paper in Section 4.

## 2. PROPOSED METHOD

### 2.1. Network Architecture

The proposed method is based on a time domain neural network as well as a data augmentation scheme that operates on the raw waveforms. The comparison between our proposed network with the baseline system is shown in Fig. 1. The main difference is that we skip the mel-spectrogram extraction, which consists of a feature engineering part , and instead utilize the raw audio as inputs. Besides, the two GRU modules are replaced by two Conformer blocks. We believe SED and DOA may have some common features that are better preserved in raw audio form. Thus, we follow the design of the baseline system [13] to jointly train the SED and DOA in a shared network and adopt the activity-coupled Cartesian DOA vector (ACCDOA) representation [22] as a single target to predict the SED and DOA simultaneously. Our model only has 2 million parameters, which can be regarded as a light model. The detailed methods will be illustrated below.

- **Sample-level CNN**: SampleCNN was first proposed by [8], which was used for music auto-tagging task. Before that, there were a few raw audio-based solutions, but they all used large size filters, which the system should learn all possible phase variations, resulted in poor performance. In fact, using large-kernel CNN will not improve the generalization ability because it is not efficient at learning acoustic features using raw audio representations. To address this issue, [8] replaced the frame-level kernel with a much smaller sample-level kernel to do the convolution, which resulted in better performance compared with feature-based solution. We adopt their settings by using a

Table 1: Eight transformations of Audio Rotation with Azimuth $\phi$ and Elevation $\theta$, the original channel arrangement is $(C_1, C_2, C_3, C_4)$

| DOA Transformation | MIC | FOA |
|---|---|---|
| $\phi = \phi, \theta = \theta$ | $(C_1, C_2, C_3, C_4)$ | $(C_1, C_2, C_3, C_4)$ |
| $\phi = -\phi - \pi/2, \theta = \theta$ | $(C_4, C_2, C_3, C_1)$ | $(C_1, C_{-4}, C_3, C_{-2})$ |
| $\phi = -\phi + \pi/2, \theta = \theta$ | $(C_1, C_3, C_2, C_4)$ | $(C_1, C_4, C_3, C_2)$ |
| $\phi = \phi + \pi, \theta = \theta$ | $(C_4, C_3, C_2, C_1)$ | $(C_1, C_{-2}, C_3, C_{-4})$ |
| $\phi = \phi - \pi/2, \theta = -\theta$ | $(C_2, C_4, C_1, C_3)$ | $(C_1, C_{-4}, C_{-3}, C_2)$ |
| $\phi = \phi + \pi/2, \theta = -\theta$ | $(C_3, C_1, C_4, C_2)$ | $(C_1, C_4, C_{-3}, C_{-2})$ |
| $\phi = -\phi, \theta = -\theta$ | $(C_2, C_1, C_4, C_3)$ | $(C_1, C_{-2}, C_{-3}, C_4)$ |
| $\phi = -\phi + \pi, \theta = -\theta$ | $(C_3, C_4, C_1, C_2)$ | $(C_1, C_2, C_{-3}, C_{-4})$ |

kernel size of 3 to do the feature embedding.

- **SE-ResNet Module**: Most of the CNN architectures relies on the design of the convolution part. We extend the SampleCNN architecture by bringing in SE-ResNet Module, which is a combination of Squeeze-Excitation block [12] and modified Residual block [11]. The effectiveness of these blocks have been proved in different audio tasks. There was a thorough analysis [7] of using different blocks to do three audio classification tasks, which were music auto-tagging, keyword spotting and acoustic scene tagging. The detailed structure of our SE-ResNet module is shown in Fig. 2 (left).
- **Conformer Block**: The Conformer architecture was proposed in [14], which is a combination of convolution and transformer [15]. It has been first applied to SELD task in [16]. Due to the convolution layers have the property to capture the fine-grained local features while the transformer is capable of learning long-sequence dependencies, Conformer was considered powerful enough to extract both local and global features of audios. In our experiment, we used 2 Conformer blocks after feature embedding. The detailed structure of Conformer is illustrated in Fig. 2 (right).

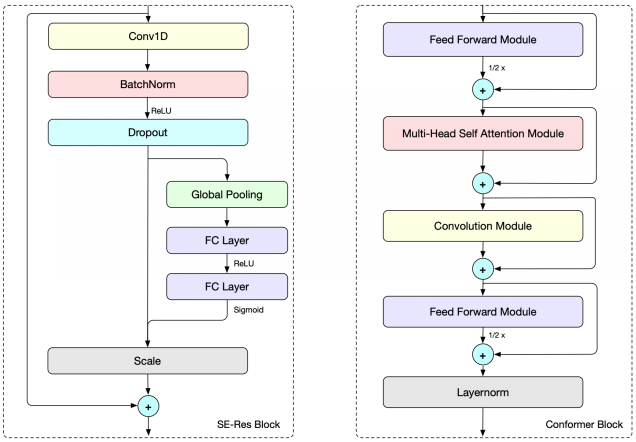


Figure 2: Architecture of SE-Res Block (left) and Conformer Block (right)

### 2.2. Data Augmentation

Data augmentation is an effective way to improve the model generalization and prevent overfitting problem. Furthermore, given the limited data available for the DCASE challenges, data agumentation has had a significant impact in the performance [23, 16, 24, 25]. For the SELD task, some augmentation operations modify the input data as well as the SED or DOA labels. For example, mixup [20] is a common data augmentation strategy, but the DOA labels cannot be mixed up effectively for all cases due to known constraints in the data. More precisely, the number of simultaneous events is limited to two.

In addition, some other audio augmentation tricks like gain shift and polarity inversion could also impact the DOA estimation. The former eliminates information regarding distance to the source, while the latter can affect the phase of the mic signals. Therefore, there is a limited set of augmentation methods that can be reliably applied to our system to ensure the DOA estimation will not be negatively impacted. After a cursory evaluation of multiple techniques, the final system utilises three methods: Sound Field Rotation (SFR), Time Masking (TM), and Random Audio Equalization (RAE).

- **Sound Field Rotation**: SFR is a spatial augmentation method that generates more DOA labels for the same input signals, by doing audio channel swapping and inversion. In [21], they indicated it is an effective way to do the augmentation while not affected DOA information. However, they only did the augmentation on FOA formats, while discarded the information of MIC data. In [16], they analysed the expressions of both two data formats and proposed a complete SFR method that can be applied to both MIC and FOA data. There are only eight valid transformations that can be used to keep the spatial information of MIC data unchanged. The detailed conversion rules are presented in Table 1.
- **Time Masking**: TM is included in SpecAugment [19] as a data augmentation trick for spectrograms. In this place, we directly do the masking in time domain by manipulating audio samples. In our experiments, $t$ consecutive samples $[t_s, t_s + t]$ are masked in a single audio sequence $S$, where $t$ is chosen from a uniform distribution from 0 to the hyper-parameter $T$, and $t_s$ is randomly chosen from $[0, S - t]$.
- **Random Audio Equalization**: For some sound events, the models can rely excessively in a few salient features that do not cover the full audio spectrum. To counter this, we apply RAE to the input signal. In each minibatch, each observation is filtered randomly by either a low pass, high pass, or and octave wide band pass fitlers. In practice, we use biquad IIR filters with order 3, where the frequencies are randomly selected from the set [250, 500, 1000, 2000, 4000] Hz.

Table 2: Evaluation results for the development set using test splits

| Framework | $ER_{20^\circ}$ | $F_{20^\circ}$ | $LE_{CD}$ | $LR_{CD}$ |
|---|---|---|---|---|
| Baseline FOA | 0.69 | 33.9% | 24.1 | 43.9% |
| Baseline MIC | 0.74 | 24.7% | 30.9 | 38.2% |
| **SSELDnet** | 0.74 | 32.6% | 26.4 | 64.2% |
| **SSELDnet + Aug** | 0.71 | 36.8% | 23.3 | 66.8% |

## 3. EXPERIMENTAL EVALUATION

### 3.1. Experimental settings

We use both FOA and MIC data formats in the TAU-NIGENS Spatial Sound Events 2021 dataset, both formats contain 600 60-second, four-channels audio recordings. We adopt ACCDOA representation [22] to jointly learn SED and DOA tasks. The evaluation metrics [26] remain the same with DCASE 2020. For SED task, there are two evaluation metrics, including location-dependent F-score ($F_{\leq T^\circ}$)and Error-Rate ($ER_{\leq T^\circ}$), which only consider predicted events under a certain threshold $T^\circ$, where in this challenge $T = 20$. For DOA task, classification-dependent Localization Error ($LE_{CD}$) and Localization Recall ($LR_{CD}$) are evaluated, where $LE_{CD}$ represents the average angular distance between ground truth and prediction, and $LR_{CD}$ stands for true positive rate of how many locations estimates are detected in a class.

For the hyper-parameter settings, we split one recording into 55 small pieces with window length of 6 seconds and hop length of 1 second to increase dataset size. The sample rate of the signal is set to 24kHz. We use Adam [27] as the optimizer and ReduceLROnPlateau in PyTorch as the learning rate scheduler with the initial learning rate of 0.001. Early stopping is used to prevent overfitting. For the data augmentation, all of the data augmentation tricks are applied online, which means the system randomly chose some hyper-parameters and then apply augmentation before the data is fed into the network. This operation will efficiently reduce the memory usage.

### 3.2. Experimental results

We compare our results with baseline system [13] in test splits. All the results are shown in Table 2.

## 4. CONCLUSION

We propose a fully end-to-end sample-level framework for DCASE 2021 task 3, Sound Event Localization and Detection. We investigate the possibilities that directly take raw audio as input to predict the SELD task. Our system does not require prior knowledge in acoustics to extract hand-engineered features. For the system design, we extend SampleCNN architecture by introducing SE-ResNet block and Conformer block. Further, three data augmentation methods are applied to improve the system generalization. Finally, the experimental results compared with the baseline system on the TAU-NIGENS Spatial Sound Events 2021 dataset are presented.

## 5. REFERENCES


[1] D. Palaz, M. M. Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4295–4299.

[2] D. Palaz, R. Collobert, *et al.*, "Analysis of cnn-based speech recognition system using raw speech as input," Idiap, Tech. Rep., 2015.

[3] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," *arXiv preprint arXiv:1609.03193*, 2016.

[4] T. Purohit and A. Agarwal, "Acoustic scene classification using deep cnn on raw-waveform," *Tech. Rep., DCASE2018 Challenge*, 2018.

[5] D. Salvati, C. Drioli, and G. L. Foresti, "Urban acoustic scene classification using raw waveform convolutional neural networks," DCASE2019 Challenge, Tech. Rep., Jun. 2019.[Online]. Available: http://dcase . . . , Tech. Rep., 2019.

[6] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6964–6968.

[7] T. Kim, J. Lee, and J. Nam, "Comparison and analysis of samplecnn architectures for audio classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 285–297, 2019.

[8] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," *arXiv preprint arXiv:1703.01789*, 2017.

[9] H. Sundar, W. Wang, M. Sun, and C. Wang, "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4642–4646.

[10] T. Kim, J. Lee, and J. Nam, "Sample-level cnn architectures for music auto-tagging using raw waveforms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 366–370.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[13] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[14] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[16] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *arXiv preprint arXiv:2101.02919*, 2021.

[17] K. Shimada, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Sound event localization and detection using activity-coupled cartesian doa vector and rd3net," *arXiv preprint arXiv:2006.12014*, 2020.

[18] T. N. T. Nguyen, N. K. Nguyen, H. Phan, L. Pham, K. Ooi, D. L. Jones, and W.-S. Gan, "A general network architecture for sound event localization and detection using transfer learning and recurrent neural network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 935–939.

[19] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[20] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[21] L. Mazzon, M. Yasuda, Y. Koizumi, and N. Harada, "Sound event localization and detection using foa domain spatial augmentation," in *Proc. of the 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019.

[22] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 915–919.

[23] M. Kim and S. Lee, "A Study on the Data Augment Method considering Room Transfer Functions for Acoustic Scene Classification," p. 6.

[24] H. Wang, Y. Zou, and W. Wang, "Specaugment++: A hidden space data augmentation method for acoustic scene classification," 2021.

[25] T. Inoue, P. Vinayavekhin, S. Wang, D. Wood, A. Munawar, B. Ko, N. Greco, and R. Tachibana, "Shuffling and mixing data augmentation for environmental sound classification," 10 2019.

[26] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv preprint arXiv:2006.01919*, 2020.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.