

ASiT: Audio Spectrogram Vision Transformer for General Audio Representation

Sara Atito

Muhammad Awais

Wenwu Wang

Mark D Plumbley

Josef Kittler

CVSSP, University of Surrey, Surrey, UK



Abstract—Vision transformers, which were originally developed for natural language processing, have recently generated significant interest in the computer vision and audio communities due to their flexibility in learning long-range relationships. Constrained by data hungry nature of transformers and limited labelled data most transformer-based models for audio tasks are finetuned from ImageNet pretrained models, despite the huge gap between the natural images domain and audio domain. This has motivated the research in self-supervised pretraining of audio transformers, which reduces the dependency on large amounts of labeled data and focuses on extracting concise representation of the audio spectrograms. In this paper, we propose ASiT, a novel self-supervised transformer for general audio representations that captures local and global contextual information employing group masked model learning and self-distillation. We evaluate our pretrained models on both audio and speech classification tasks including audio event classification, keyword spotting, and speaker identification. We further conduct comprehensive ablation studies, including evaluations of different pretraining strategies. The proposed ASiT framework significantly boosts the performance on all tasks and sets a new state-of-the-art performance on five audio and speech classification tasks, outperforming recent methods, including the approaches that use additional datasets for pretraining. The code and pretrained weights will be made publicly available for the scientific community.

1 INTRODUCTION

Throughout the history of pattern recognition and machine learning, there are numerous examples where the learning methodology developed for one dimensional time varying signals or data have been shown to be equally relevant to two dimensional signal domains and vice versa. The instances include the use of factor analysis in automatic speech recognition, proposed to deal with different speech environments back in 2000 [1], successfully extended to 2D for coping with different poses in face recognition in [2]. More recent is the porting of a new deep neural network (DNN) architecture, the transformer [3], proposed for natural language processing, to the 2D domain of image processing [4]. In this paper, we investigate the reverse task, namely the applicability of techniques developed for 2D image analysis, to a 1D signal domain. In particular, we consider the relevance of transformer based image classification approaches to the problem of audio classification.

Although audio is a 1D signal, working with its spectral representation partly bridges the gap between the 1D and 2D domains. However, the frequency/time nature of a spectrogram differs significantly from a conventional image,

where the statistical properties (pixel relationships) of the signal in any arbitrary direction and its physical meaning are the same. This contrasts with the spectrogram, where the image axes represent different physical phenomena. Moreover, in a spectrogram image, the classes are superimposed, whereas in a scene image containing several objects, the classes are adjacent to each other. These differences are significant enough to raise doubts about the direct applicability of the techniques developed for image analysis to the audio classification problem. These doubts are addressed in the case of convolutional neural networks (CNNs), however, applicability of the attention based model is still an open question with a few works [5, 6, 7, 8] validating the applicability of vision transformers (ViT) for audio.

We shall confine our study to ViT, which are gaining rapid popularity. ViTs are very data hungry [4, 9], and their training is computationally costly. This often necessitates using existing pretrained models for developing new applications, spanning different user domains. The most popular are models pretrained on the ImageNet [10], which are the models of choice, irrespective of their relevance to the application concerned. However, ImageNet pretraining is clearly not ideal for decision making, e.g. in medical domain [11]. This data hungry nature of ViTs has recently been alleviated by group masked model learning (GMML) based self-supervised learning (SSL) methods introduced in seminal work of SiT [9]. At its core, GMML is using the principle of denoising/masked autoencoder [12]. GMML randomly masks groups of connected tokens and leaves groups of connected tokens unmasked. The goal of learning is to recover missing information in the grouped of masked tokens from visible information (context). Contrastive learning [13, 14, 15, 16, 17, 18] is another line of SSL approaches that aim to maximise the similarity between two augmented views of the same image and maximise the disagreement between different images. These approaches use various tricks, e.g. large batch size, memory banks, asymmetric projection heads, different temperatures to boost the performance and more importantly to avoid representation collapse. Due to the recent advance in the SSL approaches, the self-supervised pretraining of DNNs without using labelled data for the first time outperformed supervised pretraining of DNNs for multiple computer vision downstream tasks like classification, segmentation etc. This enabled the use of domain specific pretraining of DNNs

for high performance on several downstream tasks without using labelled data.

The key questions addressed in this paper are: i) can masked image modelling successfully be applied to audio data, given the inherent differences between a 2D image and a spectrogram, and ii) how should the pretraining methodology be adapted to gain from SSL as much as possible. The first question has already been partially answered in earlier studies [5, 6, 7, 8] and we confirm in here that SSL can be applied effectively to the audio classification tasks. However, a straightforward application of SSL is not enough. The use of masked autoencoder does not guarantee that sufficient discriminatory information is retained by the pretrained model. While reconstruction is an essential ingredient, in order to preserve discriminatory information, an SSL method should promote similarity learning. To this end, we propose a novel transformer learning framework, constituted by a teacher-student architecture, which is trained using a distillation approach guided by a contrastive loss, and by reconstruction error produced by an integrated reconstruction head. We argue that the typical global similarity learning using a classification token is insufficient to capture the relevant information, and demonstrate that the proposed local similarity learning mechanism is also crucial to successful pretraining for multilabel audio classification tasks. More specifically, we propose a contrastive loss based learning method by distillation using the alignment of local features derived for the original and corrupted spectrogram as a measure of similarity. The masked spectrogram reconstruction and the combination of global/local similarity learning produces effective pretraining for downstream audio data processing tasks.

We evaluate the proposed SSL pretraining method on several benchmarking audio and speech classification datasets. We demonstrate the ability of the proposed method to learn the underlying properties of audio data using pretraining and finetuning purely on audio datasets, without resorting to auxiliary image data sets. In fact we show that the performance achieved by training exclusively on audio datasets delivers the best results, which exceed the state-of-the-art (SOTA) by a significant margin. In summary, our contributions include:

- An audio spectrogram image transformer (ASiT) for general audio representation.
- A novel self-supervised pretraining method for ASiT, which combines masked spectrogram reconstruction with local-global similarity learning using distillation.
- Extensive evaluation of ASiT on benchmarking datasets, with the results that define a new state of the art performance, demonstrating the merits of the proposed methodology.

2 RELATED WORKS

Unsupervised learning offers advantages in improving the audio representation by leveraging large scale data without labels [19]. SSL aims to learn audio representations with pairs of similar inputs derived from unlabelled data, by exploiting temporal consistency [20] or data augmentation

[21, 22]. After its wide adoption in natural language processing (NLP) and computer vision, SSL is attracting increasing interest from audio community. In early works, SSL has been applied mainly to speech [23, 24, 25, 26]. For example, Speech2Vec [23], which is inspired by Word2Vec [27], learns word embedding from acoustic features of speech, such as mel-frequency cepstral coefficients, using an RNN based encoder-decoder trained with skipgram or continuous bag-of-words approaches.

SSL has been recently used to learn general purpose audio representations. For example, Audio2Vec [28] learns a representation via reconstructing a spectrogram slice in terms of its preceding and following slices, using CNN based backbone and mel-spectrogram as input. In CLAR [29] and COLA [21], which are built on the visual SSL method SimCLR [16], the network is trained to maximize the representations of different views in latent space and reduce the reconstruction loss of the original waveform or spectrogram input, with different augmentations, such as random size cropping, Gaussian noise addition, and mix-back where the incoming patch is mixed with a background patch. These methods often assume that audio segments in temporal neighborhood have more similar representations, while those far away in time have more different representations. Nevertheless, this is not the case for repetitive sounds such as music and car sirens, where distant segments can have similar representations.

Different from the above methods, where training is performed by contrast between positive and negative samples, the distillation based approaches such as BOYL [18], DINO [30] and ATST [31] learn general audio representations, by minimizing the difference (e.g. in mean squared error) between the embeddings of the same input with the contrasts obtained via data augmentation. In this case, no negative samples are required for learning the representation. For example, in BOYL-A [32], BOYL [18] is applied to audio by learning the similarity between the two views with one learned from a randomly cropped single segment and the other from its augmented version obtained with e.g. mixup and random resized crop (RRC). Different from BOYL-A, in ATST [31], the two views were learned from two randomly cropped and then augmented mel-spectrogram segments with a teacher-student transformer architecture. The distillation based methods have been reported to achieve state of the art performance, however, this method can potentially lead to trivial solutions, i.e. collapsed representations, which require careful design of network architectures or training algorithms. For example, in Barlow Twins [33], the outputs of the network twins which take augmented samples as inputs are compared using cross-correlation, which promotes similarity between the learned embeddings from the network twins, while reducing the redundancies of the learned representations.

Inspired from masked language modeling (MLM) [34] and masked image modeling (MIM) [9], several masked acoustic modeling (MAM) approaches such as SSAST [6] and MAE-AST [7] were introduced which learn to reconstruct the masked time-frequency patches of an arbitrary shape from a given spectrogram. This offers potential advantages over the methods such as Audio2Vec [28] which focuses on temporal modelling via reconstructing masked

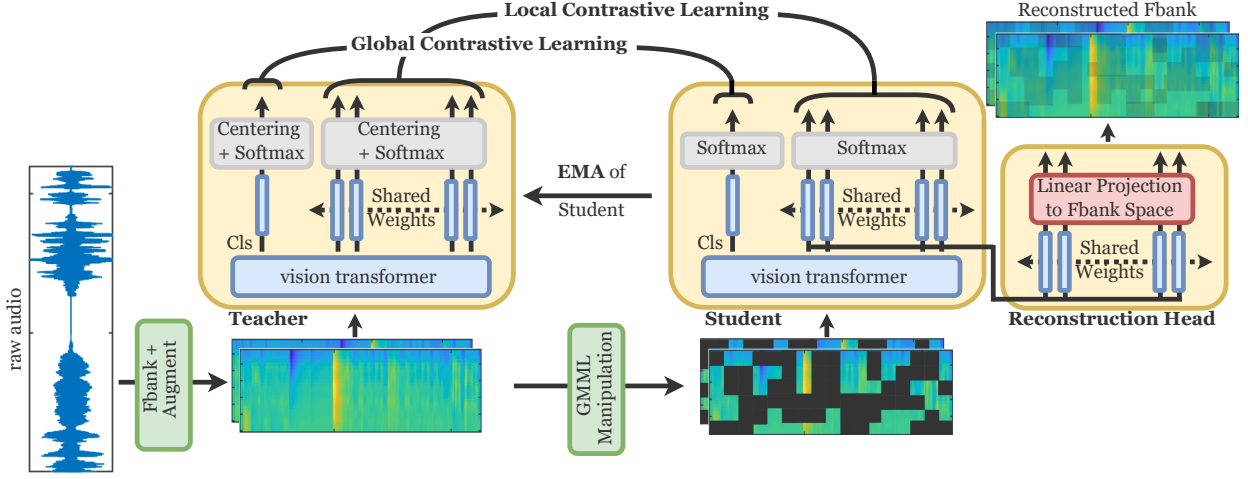


Fig. 1: The proposed self-supervised framework (ASiT). For a given 10-second audio spectrogram, two random augmented views of 6-second each (clean spectrograms) are generated and fed to GMML based manipulation block to obtain the corrupted spectrograms. The clean and corrupted spectrograms are fed to the teacher and student networks, respectively. The recovery of the transformed information from the non-transformed class-token and data-tokens indicates that the network has learnt the semantics of the local as well as the global representation of the given audio and learnt useful inductive bias by learning local statistical correlation in the spectrogram.

temporal spectrogram frames. In SSAST [6], the transformer based AST model [5] is pretrained on unlabelled audio from AudioSet and Librispeech with joint generative and discriminative masked spectrogram patch modeling. SSAST uses a very high masking ratio, hence the vast majority of self-attention is computed on mask tokens. To address this, MAE-AST [7] introduces the encoder-decoder architecture from masked autoencoder into the SSAST, where a large encoder is used to learn on unmasked input while a shallow decoder is used to reconstruct masked input with encoder outputs and masked tokens.

Despite the promising performance of the transformer models in capturing local representations as in SSAST [6], MAE-AST [7], and global representation as in ATST [31], the transformer architecture is limited in capturing optimal local contextual information or global relational information from audio. Modelling both local and global information optimally, nevertheless, can be crucial for detecting transient acoustic events such as gun shots as well as symphony. In this paper, we propose a novel self-supervised pretraining method for local context modelling via token similarity learning, and global representation learning via contrastive instance classification. This allows the transformer model to capture fine-grained region/contextual dependencies as well as global representation learnt from data rather than making any explicit assumption about the inductive bias as is the case in CNNs. This novel framework as a result significantly improves the quality of the learned audio representations.

3 METHODOLOGY

In this section, we introduce ASiT, a self-supervised framework based on vision transformers for general audio representations. Similar to [5], we employed log mel-spectrogram [35] as the input to the ViT instead of using the raw waveform directly. Spectrograms are used extensively in the

fields of audio, speech, and music as they somewhat model human hearing perception and contain abundant low-level acoustic information, which has similar characteristics to images, making them more suitable for ViTs.

Our proposed ASiT framework is based on GMML [36], briefly summarized in Section 3.1, and self-learning of data and class tokens conceptualisation with the incorporation of knowledge distillation [37], explained in Section 3.2.

3.1 Group Masked Model Learning

Masked image modelling, which in principle is similar to the masked language modelling (MLM) used in BERT [34], has been first proposed in SiT [9] and employed in several recent vision [38, 39, 40] and audio [6, 7, 8] works.

The main idea of GMML is to corrupt a group of connected patches representing a “significant” part of a given visual input and recover them by learning a model. The underlying hypothesis is that, by recovering the corrupted parts from the uncorrupted parts based on the context from the whole visual field, the network will implicitly learn the notion of visual integrity. Intuitively the network will only be able to recover the missing information if it learns the characteristic properties of visual stimuli corresponding to specific actions impacting on the visual input.

There are several ways to corrupt the group of the connected patches. For example, replacing the randomly selected patches with zeros, noise, or with patches from other spectrograms. Note that the group of the corrupted patches are manipulated randomly, i.e. the patches with audio or speech and the patches with noise or silence are manipulated with equal probability. Our hypothesis is that each pixel in a given log mel-spectrogram has a semantic meaning associated with it and the network is required to learn to reconstruct the information corresponding to both dominant (speech and audio) as well as non-dominant

(noise and silence) concepts with equal importance, which equip the network in generalizing well for unseen data.

As for the masking strategy, the time and frequency are not treated differently during masking, i.e. we apply the random masking without any prior to allow the model to learn both the temporal and frequency structure of the data.

For spectrogram reconstruction, we propose to use the vision transformer as a group masked autoencoder, i.e. visual transformer autoencoder with group masked model learning. By analogy to autoencoders, our network is trained to reconstruct the input spectrogram through the output tokens of the transformer.

A vision transformer receives as input a sequence of patches obtained by tokenizing the input spectrogram $\mathbf{x} \in \mathbb{R}^{T \times F}$ into n flattened 2D patches of size $p \times p$ pixels, where T and F are the time and frequency resolution of the input spectrogram and n is the total number of patches. The n patches are then flattened and fed to a linear projection, which are then passed to the vision transformer.

To pretrain transformers as group masked autoencoder, the GMML manipulated spectrogram $\hat{\mathbf{x}}$ is first obtained by randomly replacing a significant part of the spectrogram with zeros or with patches from another spectrogram. Note that unlike in natural language processing, partial transformation of a data-token is possible, hence the corrupted connected patches from the spectrogram are not necessarily constrained to be aligned with the tokens, refer to ablation studies in Section 4.4.

The reconstructed spectrogram $\bar{\mathbf{x}}$ is obtained by passing the corrupted spectrogram $\hat{\mathbf{x}}$ to the transformer encoder $E(\cdot)$ (the vision transformer in the student block in Figure 1) and feeding the output to a light decoder $D(\cdot)$ (the reconstruction head block in Figure 1). The decoder consists of 3 fully connected layers; the first two with 2048 neurons and GeLU [41] non-linearity each, and the last bottleneck layer with 256 neurons, followed by a transposed convolution to return back to the spectrogram space. After the pretraining stage, the light decoder $D(\cdot)$ is dropped and only the encoder $E(\cdot)$ is used for the downstream task. The importance of using the representation before the nonlinear projection, i.e. decoder head, is due to loss of information induced by the reconstruction loss. Particularly, the decoder head generally learns task-specific features, i.e. spectrogram reconstruction, which might rescind information that may be useful for the downstream task.

For the spectrogram reconstruction task, we use the ℓ_1 -loss between the original and the reconstructed spectrogram as shown in Equation 1. We compute the loss only on the corrupted pixels, similar to [34, 36].

$$\mathcal{L}(\mathbf{W})_{\text{recons}} = \sum_k^N \left(\sum_i^T \sum_j^F \mathbf{M}_{i,j}^k \times |\mathbf{x}_{i,j}^k - \bar{\mathbf{x}}_{i,j}^k| \right), \quad (1)$$

$$\mathbf{M}_{i,j} = \begin{cases} 1, & \text{if } \mathbf{x}_{i,j} \text{ is manipulated} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where \mathbf{W} denotes the parameters to be learned during training, N is the batch size, and \mathbf{M} is a binary matrix with 1 indicating the manipulated pixels.

3.2 Local and Global Pseudo-Label Learning

Group masked model learning is the key component of ASiT as it implicitly learns the notion of visual integrity. This notion of visual integrity can be further enhanced by using pseudo labels that can be generated automatically, based on some attributes of the data. The idea of assigning a local and global pseudo labels for the spectrogram is inspired from DINO [30] framework. DINO focus on assigning a one global representation for the whole input. Unlike DINO, we propose to use self knowledge distillation to learn a pseudo label for each token in the spectrogram as well as a global pseudo label for the whole spectrogram. Our hypothesis is that modelling only one representation for the whole spectrogram can lead to sub-optimal representation learning. Thus, we investigate the role of learning the appropriate representation for each of the data tokens through knowledge distillation.

In knowledge distillation, a student network $s^\theta(\cdot)$ is trained to match the output of a given teacher network $t^\phi(\cdot)$, where θ and ϕ are the parameters of the student and the teacher networks, respectively. In this work, we employ the same network architecture for both the student and the teacher, i.e. the teacher is a momentum encoder [42] of the student, where the parameters of the teacher network are updated from the past iterations of the student network using exponential moving average of the student weights with the following update rule: $\phi \leftarrow \lambda\phi + (1 - \lambda)\theta$.

In our case, the network architecture of the student comprises three components: A vision transformer backbone $s_b(\cdot)$, followed by two projection heads attached to the output of the transformer. The first projection head $s_{\text{lcl}}(\cdot)$ is for local contrastive learning and the second projection head $s_{\text{gcl}}(\cdot)$ is for global contrastive learning. The local contrastive learning projection head $s_{\text{lcl}}(\cdot)$ consists of 3 fully connected layers; the first two with 2048 neurons and GeLU non-linearity each, and the last bottleneck layer with 256 neurons. The output of the bottleneck layer is ℓ_2 normalised and directly connected to a weight-normalised fully connected classification layer with $K = 1024$ neurons. The global contrastive learning projection head $s_{\text{gcl}}(\cdot)$ has a similar design to $s_{\text{lcl}}(\cdot)$ except that its output layer has $C = 8192$ neurons. The choices of K and C are intuitive in the sense that the local patches have less variability w.r.t each other hence, local similarity can be captured in lower dimensional space ($K = 1024$), while global similarity within the spectrogram is better captured in higher space ($C = 8192$). We expect further improvement by tuning these hyperparameters.

In order to generate a local pseudo label for the input spectrogram \mathbf{x} , the clean spectrogram is fed to the backbone of the teacher network $t_b(\mathbf{x})$ and the output of the data tokens are passed to the local contrastive learning projection head to obtain $\mathbf{z}_t \in \mathbb{R}^{n \times K}$. As for the student network, the GMML-based manipulated spectrogram $\hat{\mathbf{x}}$ is passed to the student network to obtain $\mathbf{z}_s \in \mathbb{R}^{n \times K}$. The task is to match the output of the student network to the output of the teacher network employing the Kullback-Leibler divergence (KL).

Training the student to match the teacher output can easily lead to a trivial constant (i.e. collapsed) embeddings.

To avoid model collapse, we adopted the centring and sharpening of the momentum teacher outputs introduced in [30]. Centring encourages the output to follow a uniform distribution, while the sharpening has the opposite effect. Applying both operations balances their effects, which is sufficient to avoid a collapse in the presence of a momentum teacher. The centre vector is updated using an exponential moving average over the teacher output. The sharpening is obtained by using a low value for the temperature τ_t in the teacher softmax normalisation. The output probability distributions z_t and z_s from the teacher and the student networks are obtained as follows:

$$p_t^{(i,j)} = \frac{\exp(z_t^{(i,j)}/\tau_t)}{\sum_{k=1}^K \exp(z_t^{(i,k)}/\tau_t)} \quad (3)$$

$$p_s^{(i,j)} = \frac{\exp(z_s^{(i,j)}/\tau_s)}{\sum_{k=1}^K \exp(z_s^{(i,k)}/\tau_s)} \quad (4)$$

where z_t and z_s are the class logits for the student and the teacher, $p_t(i, \cdot)$ and $p_s(i, \cdot)$ are the probabilities corresponding to the i -th token output by the teacher and the student, and τ_t and τ_s are the temperature parameters for the teacher and the student, respectively.

Our training objective is to match the output probability of the student p_s with that of teacher p_t . We use the cross entropy measure for this task.

$$\mathcal{L}(\theta)_{\text{lcl}} = \sum_k^N \left(\sum_{i=1}^n \mathbf{T}_i^k \times \sum_{j=1}^K -p_t^k(i, j) \log p_s^k(i, j) \right) \quad (5)$$

$$\mathbf{T}_i = \begin{cases} 1, & \text{if token } i \text{ is manipulated} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where \mathbf{T} is a binary mask with 1 indicating the manipulated/corrupted tokens.

The network is trained to assign a class for each token in the spectrogram and differentiate it from other tokens. In essence, the idea of local contrastive learning is similar to spectrogram reconstruction, but not in the pixel space, but rather in the class space.

As for the global contrastive learning, we follow the same procedure as in the local contrastive learning, but proceeding only on the class token to learn a global representation for the whole spectrogram. For this particular task, two augmented views are required, where the task is to match the output representation of the student from the first augmented view to the output representation of the teacher from the second augmented view and vice versa.

3.3 Putting Together the ASiT Framework

For a given spectrogram, two augmented views are generated and passed to the teacher network. The corrupted version of the two views are obtained and passed to the student network. The output of the data tokens from the backbone of the student network is passed to both the reconstruction and local contrastive learning projection heads. The reconstruction loss $\mathcal{L}(\theta)_{\text{recons}}$ is calculated between the reconstructed spectrograms and the original spectrograms. In addition, the local contrastive learning loss $\mathcal{L}(\theta)_{\text{lcl}}$ is calculated by

matching the output of the data tokens from the student network to the data tokens from the teacher network. Finally, the global contrastive learning loss $\mathcal{L}(\theta)_{\text{gcl}}$ is calculated by matching the cross representation of the two augmented views corresponding to the class token from the student and teacher, respectively. The overall loss is then estimated as follows: $\mathcal{L}(\theta) = \alpha_1 \mathcal{L}(\theta)_{\text{recons}} + \alpha_2 \mathcal{L}(\theta)_{\text{lcl}} + \alpha_3 \mathcal{L}(\theta)_{\text{gcl}}$, where $\alpha_1, \dots, \alpha_3$ are the scaling factors of our multi-task objective function, which are set to 1 for simplicity. We combine the losses of the three pre-text tasks using simple averaging. We believe further improvements can be gained by optimising the weighted sum of the losses or by incorporating the uncertainty weighting approach [43]. Refer to the overall architecture in Figure 1.

4 EXPERIMENTS

We follow the common evaluation protocol to demonstrate the generalisation of the learnt features of the proposed ASiT self-supervised learning approach by pretraining the model in an unsupervised fashion on AudioSet-2M dataset, followed by finetuning on several downstream tasks, including audio event classification, keyword spotting, and speaker identification tasks. We provide the details of the employed datasets and the implementation details of the pretraining and finetuning stages in Section 4.1 and 4.2, respectively. In Section 4.3, we discuss the performance and analysis of the proposed method compared to the state-of-the-art. Further, we conduct several ablation studies to investigate the effect of the individual components of the proposed approach in Section 4.4.

4.1 Datasets

AudioSet (AS-2M, AS-20K) [46] is a collection of YouTube clips for 527 multi-label weakly annotated audio events. The training set has 2 subsets; unbalanced dataset with around 2 million clips (AS-2M) and a class-wise balanced dataset with around 20K clips (AS-20K). The evaluation set has around 20K clips.

Environmental Sound Classification (ESC-50) [47] is a single audio event classification dataset, consists of 2,000 5-second environmental audio recordings organized into 50 classes. We follow the standard 5-fold cross-validation to evaluate our model on this dataset.

Speech Commands (SC-V2, SC-V1) [48] are two datasets for keyword spotting task. SC-V2 consists of 105,829 1-second recordings of 35 common speech commands, split into 84,843, 9,981, and 11,005 samples for the training, validation, and test set, respectively. SC-V1 is similar to SC-V2, but only contains 10 classes of keywords, 1 silence class, and 1 unknown class that includes all the other 20 common speech commands.

VoxCeleb (SID) [49] is a dataset for speaker identification task. It contains around 150K utterances of speech from 1,251 different speakers. The dataset is split into 138,361, 6,904, and 8,251 samples for the training, validation, and test set, respectively.

TABLE 1: Comparison with state-of-the-art works on audio and speech classification tasks. Evaluation metrics are mean average precision (mAP) for AS-2K and accuracy (%) for ESC-5, SC-V1, SC-V2, and SID. \uparrow shows the improvement over best SOTA.

Method	Backbone	Pretraining Data	Transfer Learning				
			AS-20K	ESC-50	SC-V2	SC-V1	SID
<i>Supervised-learning-based methods</i>							
PANNs [44]	CNN	–	27.8	83.3	–	61.8	–
AST [5]	ViT-B	AS-2M	28.6	86.8	96.2	91.6	35.2
<i>Self-supervised-learning-based methods</i>							
COLA [21]	CNN	AS-2M	–	–	98.1	95.5	37.7
SSAST [6]	ViT-B	AS-2M	29.0	84.7	97.8	94.8	57.1
MaskSpec [8]	ViT-B	AS-2M	32.3	89.6	97.7	–	–
ASiT (ours)	ViT-B	AS-2M	35.2 (↑ 2.9)	92.0 (↑ 2.4)	98.8 (↑ 0.7)	98.1 (↑ 2.6)	63.1 (↑ 6.0)
<i>SSL based methods for reference not comparison as they are pretrained on additional speech dataset LS [45]</i>							
SSAST [6]	ViT-B	AS-2M + LS	31.0	88.8	98.0	96.0	64.3
MAE-AST [7]	ViT-B	AS-2M + LS	30.6	90.0	97.9	95.8	63.3

4.2 Implementation Details

The backbone architecture of ASiT employs the base (ViT-B) variant of ViT [4] and pretrained on AudioSet-2M dataset for a fair comparison with the state-of-the-art.

During the pretraining, simple data augmentation techniques are applied as we found that to learn low-level features as well as high-level semantic information, aggressive data augmentation hurts in the pretraining stage. Specifically, the given raw audio is pre-processed as a mono channel under 32,000 Hz sampling rate, then two random chunks of the given audio are cropped and resized to 6 seconds each. The two augmented audio waveforms are then converted into a sequence of 128-dimensional log mel filterbank (fbank) features computed with a 25ms Hanning window that shifts every 10ms. For a 6-second recording, the resulting spectrogram is of 592×128 dimension. The two augmented spectrograms are then normalized to adjust the statistical drifts caused by the aforementioned operations.

For the GMML corruption, the spectrograms are randomly corrupted either by zeros with a total corruption of 70% or by alien concepts (replace with random blocks from another spectrogram) with a total corruption of 30%.

For the optimisation of the self-supervised pretraining, the model is trained using the Adam optimiser [50] with a momentum of 0.9 and batch size of 40 images per GPU, using 4 GeForce RTX 3090 GPUs. The weight decay follows a cosine schedule [51] from 0.04 to 0.4, and the base learning rate is $5e^{-4}$. The model is pretrained for 40 epochs on the AS-2M dataset from scratch. The sharpening parameter of the teacher and the student are set to $\tau_t = 0.07$ and $\tau_s = 0.1$. The teacher is updated using exponential moving average of the student weights with λ following a cosine schedule from 0.996 to 1 during training.

Finally, for the downstream tasks, the projection heads are discarded and finetuning is performed employing the backbone of the pretrained teacher network. For the finetuning, we rely on the AST developer’s default code and hyper-parameters [5] except for the case of SID dataset, we follow the SUPERB evaluation framework [52] with learning rate $1e^{-4}$ and report the accuracy on the test set.

4.3 Results

In Table 1, we compare ASiT to the supervised and self-supervised state-of-the-art approaches in audio event classification. For fair comparison, we pretrained ASiT only

on the AudioSet-2M dataset. As shown in Table 1, pretrained ASiT obtains mAP of 35.2 on AudioSet-20K, which is significantly outperforming supervised learning with an improvement of +6.6 mAP and the state-of-the-art with an improvement of +2.9 mAP.

Further, ASiT achieves the best performance across different audio tasks compared to other approaches. Particularly, we obtain 92.0%, 98.1%, and 98.8% with an improvement of 2.4%, 0.7%, 2.6%, and 6.0% on the ESC-50, speech command v1 and v2, and SID datasets, respectively.

Note that in order to improve the coverage of speech data, SSAST [6] and MAE-AST [7] further used the Librispeech [45] dataset, which has around 1,000 hours of speech. Despite that we only pretrained ASiT on AudioSet-2M dataset, we outperform the aforementioned methods on the SC-V1 and SC-V2 speech tasks with a large margin as shown in Table 1 and obtained on par performance on the SID dataset, showing the generalisability of the proposed self-supervised framework.

4.4 Ablations

In all of the ablation studies, ASiT is pretrained on AS-2M for only 10 epochs employing the small variant of vision transformers, i.e. ViT-S, as the backbone of the student and the teacher of ASiT (unless mentioned otherwise). To assess the quality of the learnt representation, the pretrained models are finetuned on AS-20K and the mean average precision on the validation set is reported.

TABLE 2: Effect of the different components of ASiT for self-supervised pretraining.

Image Corruption	Image Reconstruction	Local Contrastive Learning	Global Contrastive Learning	mAP (AS-20K)
\times	\checkmark	\times	\times	16.6
\times	\times	\times	\checkmark	20.2
\checkmark	\times	\times	\times	23.9
\checkmark	\checkmark	\times	\times	25.5
\checkmark	\times	\checkmark	\checkmark	25.9
\checkmark	\checkmark	\checkmark	\checkmark	27.1
\checkmark	\checkmark	\times	\times	27.6
\checkmark	\checkmark	\checkmark	\checkmark	28.1

Effect of Different Recipes of ASiT. The aim of this ablation study is to investigate the effect of the individual elements of the pretext learning, reported in Table 2. First, we investigate the effectiveness of pretraining transformers

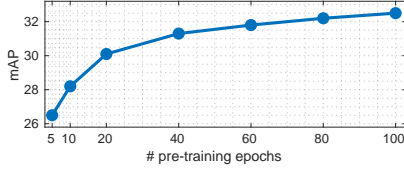


Fig. 2: Ablation studies of the effect of longer pretraining.

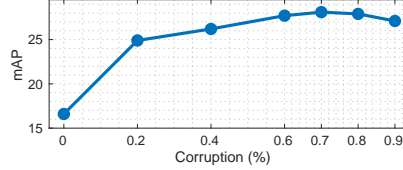


Fig. 3: Effect of percentage of corruption during the pretraining stage.

as an autoencoder to reconstruct the input audio without any sort of corruption, i.e. $D(E(x)) = x$, where x is the input audio, E is the encoder which is ViT-S in our case, and D is a lightweight reconstruction decoder. Expectedly, poor performance is obtained that is slightly better than training the model from scratch. Indeed, this is attributed to the fact that without proper choice of constraints, autoencoders are capable of learning identity mapping, i.e. memorising the input without learning any useful discriminative features.

To regularise the transformer-based autoencoder, we incorporated input corruption along with spectrogram reconstruction, i.e. GMML, where the mAP jumped from 16.6 to 25.5 mAP. We also investigated the effect of individually employing local contrastive learning and global contrastive learning. We found that the best individual task is GMML followed by local contrastive learning and the least effective individual task is global contrastive learning.

Further, we investigated the effect of the different combination of the pre-text tasks. We found that using the spectrogram corruption along with the reconstruction loss on its own as a means of self-supervision provided an effective starting point for efficient downstream task finetuning. Further marginal improvements can be made by extending the range of mechanisms for self-supervised pretraining.

Effect of Longer Self-supervised Pretraining. Figure 2 shows the mAP when the pretrained ASiT is finetuned on AS-20K dataset. We found that longer pretraining leads to systematic performance gain, where the mAP is steadily improving even after 100 epochs of pretraining.

Effect of the Percentage of Audio Corruption. Figure 3 shows the mAP when the models are pretrained with different corruption percentages. We found that the optimal ratio is between 60% to 80%. This behaviour was expected as the masking encourages the network to learn from the uncorrupted patches surrounding the groups of masked tokens in order to recover the missing information.

Effect of Aligning the Corruption with the Patches. We observed that the speed of convergence and generalisation of pretraining improve when the masking of tokens is not aligned to patch boundaries. Particularly, after 10 epochs of pretraining ASiT, the validation mAP on the AS-20K dataset is 26.9 and 28.1 when the corruption is aligned and not aligned with the patch boundaries, respectively. It is worth noting that with longer pretraining, the performance gap reduces between the two strategies.

Effect of the Model Size. In this ablation, we study the impact of model size. Specifically, we employed the small (ViT-S) and base (ViT-B) variants of vision transformer. The small variant has 12 transformer blocks with an embedding dimension of 384 and 3 attention heads. The backbone has

TABLE 3: Mean Average Precision (mAP) on AS-20K pretrained with different weights’ initialization and different backbones.

Pretraining Dataset	Backbone	
	ViT-Small	ViT-Base
Scratch	12.6	14.8
AudioSet-2M (AS-2M)	32.0	35.2
ImageNet-1K (INet)	26.3	30.2
INet \rightarrow AS-2M	31.5	34.5

around 22M parameters in total. As for the base variant of vision transformer, it consists of 12 transformer blocks with an embedding dimension of 768 and 6 attention heads. The backbone has around 85M parameters in total. As shown in Table 3, the performance on the downstream task increases with bigger model, showing that ASiT is model size agnostic, demonstrating that ASiT can unlock the potential of models with higher capacity.

Effect of Different Weights’ Initialisation. Despite the significant discrepancies between image and audio modalities, the common practice for audio classification is to initialize the models from ImageNet pretrained weights. Thus, it is crucial to answer this crucial question in audio domain, is out-of-domain pretraining benefits audio representation learning? For that, in Table 3, we compare the performance with different weights’ initialisation strategies across different models, including (1) from-scratch models, i.e. no pretraining, (2) models pretrained only on AS-2M, (3) models pretrained only on ImageNet-1K, and (4) models pretrained on ImageNet followed by pretraining on AS-2M datasets. As shown in Table 3, the poor performance when training from scratch is expected especially when the training is on small datasets due to the lack of inductive bias in transformers. We observed that ImageNet pretraining alone is not sufficient compared to pretraining on in-domain dataset. Further, we found that ImageNet initialization followed by audio pretraining slightly degrades the performance compared to pretraining only on audio dataset.

5 CONCLUSION

We presented an efficient method of designing an audio signal classifier based on self-supervised pretraining. In our approach audio is represented by a spectrogram and processed as an image by a vision transformer. We demonstrated that the Group Masked Model Learning, that has earlier been shown to be effective for image analysis, provides also a solid basis for self-supervised pretraining for audio classification. However, its core ingredient, namely a masked autoencoder trained using reconstruction loss, has to be bolstered by other information extraction mechanisms.

We proposed a novel transformer learning framework, constituted by a teacher-student architecture trained using a distillation approach. The training minimises a contrastive loss, and reconstruction error. As the typical global similarity learning using a classification token proved insufficient to capture discriminatory information, we proposed local similarity learning using the alignment of local features computed for the original and corrupted spectrogram as a measure of similarity.

The transformer pretraining, afforded by the joint use of masked spectrogram reconstruction and the combination of global/local similarity learning, was instrumental in obtaining promising solutions for downstream audio classification tasks. The results of extensive evaluations of the proposed methodology on benchmarking datasets defined a new state-of-the-art performance, demonstrating the merits of the proposed methodology.

REFERENCES

- [1] Lawrence K. Saul and Mazin G. Rahim. Maximum likelihood and minimum classification error factor analysis for automatic speech recognition. *IEEE Trans. Speech Audio Process.*, 8(2):115–125, 2000. **1**
- [2] Simon J. D. Prince, James H. Elder, Jonathan Warrell, and Fatima M. Felisberti. Tied factor analysis for face recognition across large pose differences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):970–984, 2008. **1**
- [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. **1**
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. **1, 6**
- [5] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. **1, 2, 3, 6**
- [6] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. SSAST: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10699–10709, 2022. **1, 2, 3, 6**
- [7] Alan Baade, Puyuan Peng, and David Harwath. MAE-AST: Masked autoencoding audio spectrogram transformer. *arXiv preprint arXiv:2203.16691*, 2022. **1, 2, 3, 6**
- [8] Dading Chong, Helin Wang, Peilin Zhou, and Qingcheng Zeng. Masked spectrogram prediction for self-supervised audio pre-training. *arXiv preprint arXiv:2204.12768*, 2022. **1, 2, 3, 6**
- [9] Sara Atito, Muhammad Awais, and Josef Kittler. SiT: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021. **1, 2, 3**
- [10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012. **1**
- [11] Sara Atito, Syed Muhammad Anwar, Muhammad Awais, and Josef Kittler. SB-SSL: Slice-based self-supervised transformers for knee abnormality classification from MRI. In *Workshop on Medical Image Learning with Limited and Noisy Data, MICCAI*, pages 86–95. Springer, 2022. **1**
- [12] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103, 2008. **1**
- [13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. **1**
- [14] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. **1**
- [15] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations (ICLR)*, 2019. **1**
- [16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. **1, 2**
- [17] Massimiliano Patacchiola and Amos J Storkey. Self-supervised relational reasoning for representation learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 4003–4014, 2020. **1**
- [18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doherty, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. **1, 2**
- [19] Yong Xu, Qiang Huang, Wenwu Wang, Peter Foster, Siddharth Sigtia, Philip Jackson, and Mark Plumbley. Pre-training audio representations with self-supervision. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 25:1230–1241, 2017. **2**
- [20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. **2**
- [21] Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879. IEEE, 2021. **2, 6**
- [22] Xubo Liu, Qiushi Huang, Xinhao Mei, Tom Ko, H Lillian Tang, Mark Plumbley, and Wenwu Wang. Cl4ac: A contrastive loss for audio captioning. *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2021. **2**
- [23] Yu-An Chung and James Glass. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint arXiv:1803.08976*, 2018. **2**
- [24] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019. **2**
- [25] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020. **2**
- [26] Morgane BRiviere, Armand Joulin, Pierre-Emmanuel Mazare, and Dupoux Emmanuel. Unsupervised pretraining transfers well across languages. *ICASSP*, 2020. **2**
- [27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. **2**
- [28] Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek. Pre-training audio representations with self-supervision. *IEEE Signal Processing Letters*, 27:600–604, 2020. **2**
- [29] Haider Al-Tahan and Yalda Mohsenzadeh. CLAR: Contrastive learning of auditory representations. In *International Conference on Artificial Intelligence and Statistics*, pages 2530–2538. PMLR, 2021. **2**
- [30] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. **2, 4, 5**
- [31] Xian Li and Xiaofei Li. ATST: Audio representation learning with teacher-student transformer. *arXiv preprint arXiv:2204.12076*, 2022. **2, 3**

- [32] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 2
- [33] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. BarlowTwins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2
- [34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3, 4
- [35] Lawrence Rabiner and Ronald Schafer. *Theory and applications of digital speech processing*. Prentice Hall Press, 2010. 3
- [36] Sara Atito, Muhammad Awais, and Josef Kittler. GMM is all you need. *arXiv preprint arXiv:2205.14986*, 2022. 3, 4
- [37] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3
- [38] Hangbo Bao, Li Dong, and Furu Wei. BEIT: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3
- [39] Sara Atito, Muhammad Awais, Ammarah Farooq, Zhenhua Feng, and Josef Kittler. MC-SSL0.0: Towards multi-concept self-supervised learning. *arXiv preprint arXiv:2111.15340*, 2021. 3
- [40] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 3
- [41] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GeLUs). *arXiv preprint arXiv:1606.08415*, 2016. 4
- [42] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 4
- [43] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018. 5
- [44] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 6
- [45] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 6
- [46] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 5
- [47] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1015–1018, 2015. 5
- [48] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018. 5
- [49] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Senior. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020. 5
- [50] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in Adam. *ArXiv*, abs/1711.05101, 2017. 6
- [51] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6
- [52] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhota, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. SU-PERB: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021. 6