Mixup-Based Knowledge Distillation with Causal Intervention for Multi-Task Speech Classification

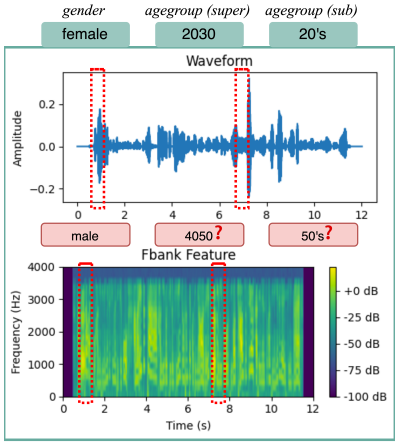
Kwangje Baeg Hyeopwoo Lee Yeomin Yoon Jongmo Kim
Samsung Financial Networks, Seoul, Republic of Korea
{kwangje.baeg, hyeopwoo.lee, yeomin.yoon, jongmo.kim}@samsung.com

Abstract

Speech classification is an essential yet challenging subtask of multitask classification, which determines the gender and age groups of speakers. Existing methods face challenges while extracting the correct features indicative of some age groups that have several ambiguities of age perception in speech. Furthermore, the methods cannot fully understand the causal inferences between speech representation and multilabel spaces. In this study, the causes of ambiguous age group boundaries are attributed to the considerable variability in speech, even within the same age group. Additionally, features that indicate speech from the 20's can be shared by some age groups in their 30's. Therefore, a two-step approach to (1) mixup-based knowledge distillation to remove biased knowledge with causal intervention and (2) hierarchical multi-task learning with causal inference for the age group hierarchy to utilize the shared information of label dependencies is proposed. Empirical experiments on Korean open-set speech corpora demonstrate that the proposed methods yield a significant performance boost in multitask speech classification.

1 Introduction

Human speech contains a wealth of information related to the identity, emotion, gender, height, age, accent, and origin of a speaker from various perspectives, owing to a combination of linguistic and paralinguistic factors[1][2]. Speech classification plays a crucial role in spoken language and audio signal analyses by automatically categorizing or delineating speech into predefined factors. However, general classification maximally discriminates between a number of predefined factors [3], whereas speech classification concerns classifier analysis and design without discrimination based on sensitive features, including age group; thus speech-based classification is challenging [4]. Most of the existing approaches present some difficulties in predicting a speaker's age and classifying the ambiguous boundaries of age groups because age is not a discrete factor and has a subjective nature that poses encapsulation challenges in models [2][5][6][7]. Another reason for such brittleness is label noise, a characteristic of real-world audio data, as depicted in Figure 1. In this study, age-group ambiguity problems are solved by compelling the model to learn the unseen causalities between age groups and mitigate the impact of label noise that otherwise compromises model generalization.



(LTV) D03/J13/S000130/0024.wav

Figure 1: Example of observed label noise in the Korean speech corpus (LTV). The original voice, characterized as female/20-30/20's, is overlaid with an unidentified voice (potentially male/40-50/50's)

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Throughout the study, a multitask learning (MTL) approach is utilized to simultaneously train multiple related tasks and classify target labels, such as gender or superclass and subclass of age groups. Herein, superclass and subclass of age groups are denoted as "agesup" and "agesub," respectively. However, equipping MTL with speech classification is difficult owing to the numerous aspects of information in speech and diverse training techniques used to determine the relationships for each task in the MTL. Therefore, the causalities between agesup (i.e., 20–30, 40–50, and 60) and agesub (i.e., 20's, 30's, 40's, 50's, and 60's) are identified in this study. Furthermore, additional dependency losses are introduced to compel the classification model to learn hierarchically structured relationships, as shown in Figure 2. Additionally, a data-agnostic data augmentation method known as mixup [8] is combined with feature-based knowledge distillation (KD) for improved robustness against noisy label datasets.

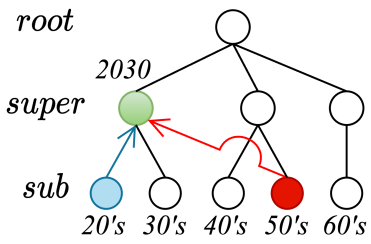


Figure 2: Concept of hierarchical dependency loss. The dependency punishment for forcing the model to learn hierarchical information when conflicting the age-group category from the hierarchy while training independent tasks using the MTL approach is illustrated.

The proposed approach leverages the causal structure among speech features including ambiguous boundaries and improves the model robustness against noise. Causal representation learning (CRL) is an effective approach for extracting invariant and stable causal information. The robustness and generalization performance of machine learning (ML) models are expected to improve by CRL [9]. Under the hypothesis that intrinsic latent factors follow casual models, the performance of speech classification can be improved by learning a causal representation, which is the shared representation used to classify each target task and provide superior performance for independent tasks [10]. The main contributions of this study are summarized as follows:

- The concept of a causal approach to hierarchical MTL of highly variable speech features is implemented while improving the ability of the speech classification model.
- A mixup-based KD method is proposed to acquire a robust representation of the student model trained from a noisy label dataset by transferring knowledge from a pre-trained teacher model, which is trained from a clean dataset.

2 Related work

2.1 Causal representation learning and causal interventions

CRL involves the identification of underlying causal variables and their relationships from high-dimensional observations (such as speech) and investigation of a representation that partially exposes the unknown causal structure [9][11]. Representations that capture the underlying causal factors of data and generalize well to interventions, counterfactual scenarios, and unseen environments are learnt, thereby addressing some of the central challenges faced in ML [9]. Traditionally, representation learning focused on learning mapping from raw data to a lower-dimensional space, ideally preserving the essential characteristics of data while discarding noise. This concept was further developed in CRL with the aim of uncovering the underlying causal structure of data for providing a more robust and transferable representation and describing the relationships between various factors.

Causal interventions are operations in which one or more variables are actively manipulated following a causal mechanism to observe the effects of such manipulations on other variables, while allowing other mechanisms and observations to continue functioning [12][13]. One of the main goals of CRL is to understand the relationship between representations and causal interventions and find interventional data from high-dimensional observations, such as speech, image, and video[14]. Therefore, the integration of CRL with causal interventions offers a deeper understanding of the causal mechanisms underlying the data, facilitates better generalization of ML models, and fosters advancements in numerous domains [9]. Several researchers have proposed methods that utilize causal interventions. Zhang et al.[13] introduced an approach employing causal intervention to eliminate confounding bias in image-level classification, thereby providing enhanced pseudo-masks as a more accurate

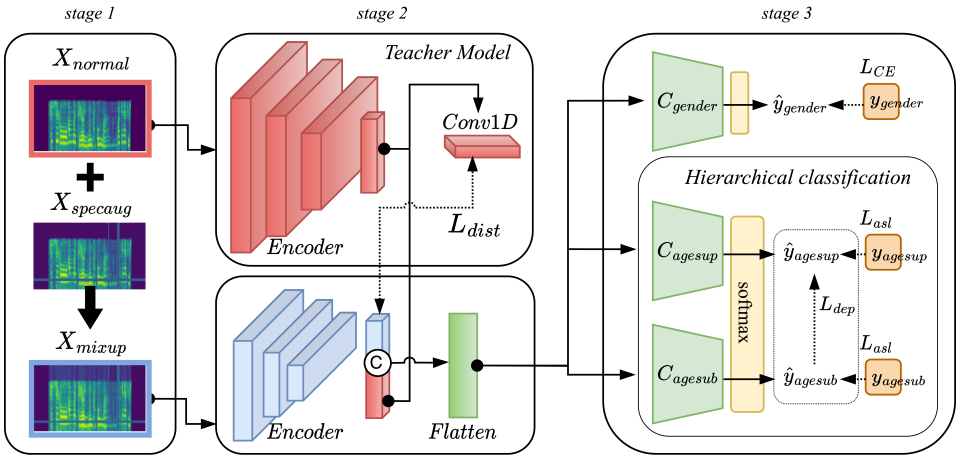


Figure 3: Overview of our proposed Multi-task Speech Classification Model.

ground truth for subsequent semantic segmentation models. Wang et al.[15] presented a video object-grounding model that harnesses causal intervention, aiming to discern object-relevant associations from the vantage of video data generation, and seek genuine causality through backdoor adjustment.

2.2 Knowledge distillation with causal intervention

In ML, KD is the process of training a smaller student model to imitate a larger and more complex teacher model. Primarily, recent distillation techniques have focused on aligning sample representations between teacher and student models but often neglect the adequate transfer of class representations [16]. Fully imitating the representations of a teacher model is not optimal, as the model is typically imperfect and its bias gets transferred to the student model. Therefore, incorporating KD with causal interventions promotes the enhancement of the conventional KD process with causal reasoning derived from interventions and facilitates more robust and interpretable student models. For instance, Deng et al. [17] altered the training process of a student model through interventions that were designed based on the causal understanding derived from a teacher model. Shao et al. [18] developed a multi-teacher causal distillation framework designed to equitably assimilate both classification and localization knowledge throughout the model training process.

3 Methodology

3.1 Architecture

Figure 3 illustrates the overall structure of the proposed model that is capable of being trained in an end-to-end manner. Based on the proposed methods of mixup-based KD and hierarchical MTL, the architecture of the proposed model was built in three stages. In the first stage, two different inputs were generated from the same speech using the temporal frequency mixup operation [19] to imitate the environment of a label noise dataset, while combining temporal information from the same source [20]. For classification problems, the mixup technique demonstrated efficacy in enhancing model robustness through the smoothing of loss landscapes [21]. Nonetheless, its direct application to speech classification encountered challenges because the lengths of audio files differ, making calculations difficult. Hence, unlike traditional mixup, a simple data augmentation method, SpecAugment [22], was directly applied to the feature inputs, such as filter bank coefficients, and the original input was mixed with the augmented feature using a fixed mixup ratio λ, controlling the frequency cut. In this study, λ was set to 0.7. Given an original speech x_{normal} and augmented input $x_{specaug}$, the mixup sample was generated at each time step i and mixup window length T, such that

$$x^i_{mixup} = \lambda x^i_{normal} + (1-\lambda)\frac{1}{T}\sum_{j=i-\frac{T}{2}}^{i+\frac{T}{2}}(x^j_{specaug}) \tag{1}$$

In the second stage, KD was applied during the phase of knowledge transfer from a teacher model. The teacher model was trained using clean label datasets, FCVG, which were recorded in a studio environment to ensure high-quality speech data, setting a benchmark for excellence and reliability in speech data. Experiments were conducted with two basic criteria for the feature-based KD model, the mean squared error (mse) and cosine embedding loss (cos), to measure the discrepancy between the embedding features of the teacher and student models. Utilization of these criteria enabled the student model to inherit the representational power of the teacher model, fostering the learning of rich discriminative features that enhanced its generalization ability for the task at hand. In this study, the features were the results of an attentive statistics pooling operation; therefore, a non-trainable layer (Conv1D) was included to convert the feature map of the teacher model to the shape of the feature map of the student model. Finally, the feature vectors from the teacher and student models were concatenated to train the student classifiers. The objective of KD was formulated as

$$\mathcal{L}_{dist}(S,T) = \frac{1}{N}\sum_{i=1}^{N}(S_i - T_i)^2 \quad \text{or} \quad 1 - \frac{\sum_{i=1}^{N} S_i \cdot T_i}{\|S\| \cdot \|T\|} \tag{2}$$

where T and S denote the feature vectors of the teacher and student models for a given input i, respectively.

In the third stage, an MTL approach that combined hierarchical classification networks was designed. The training objective for MTL was formulated as the sum of the per-task losses as

$$\begin{aligned}\mathcal{L}_{total} = \ \lambda\mathcal{L}_{dist}(S,T) + \ \mathcal{L}_{CE}(H_{gender}, Y_{gender}) + \ \mathcal{L}_{asl}(H_{agesup}, Y_{agesup}) + \\ \mathcal{L}_{asl}(H_{agesub}, y_{agesub}) + \ \mathcal{L}_{dep}(H_{agesup}, y_{agesub})\end{aligned} \tag{3}$$

where λ is set to 0.25. As the proposed approach built different classifiers for each class and sub-hierarchy, the nodes in the proposed model could independently associate with multiple classes. However, this study was focused on CRL, with the aim to find a low-dimensional representation of observations that benefit from predicting multiple tasks. Consequently, the classifiers were forced to learn the shared representation information or the concatenated features from the teacher and student models. The main challenge was to compel each classifier to learn the shared information explicitly, including the hierarchical information of each class. Details on the rationale behind hierarchical multitask classification with CRL and its implementation are discussed in Section 3.2.

3.2 Causal approach to hierarchical multi-task learning

By minimizing the objectives expressed in Equation 3, the network was made to learn classification from the high-quality pre-trained model through KD. From this perspective, the causalities among speech input samples X, prior knowledge of high-quality speech K for training the student model, and target labels Y using a structural causal model (SCM) were formulated. Additionally, the network was developed for hierarchical MTL, and the full causal graph when three different tasks were considered is shown in Figure 4, where Z denotes the shared representation of the teacher and student models extracted from the original observation X and its prior knowledge K, and H denotes task-specific representations based on Z for each target Y. An intermediate process $Y_{agesub} \rightarrow H_{agesub} \rightarrow Y_{agesup}$ conveyed the subclass information to the superclass, whereas the mediation assumption considered the logical dependency from Y_{agesub} to Y_{agesup}. In this study, SCM was applied to empower each classification model to pursue the designed causalities between the three different targets Y.

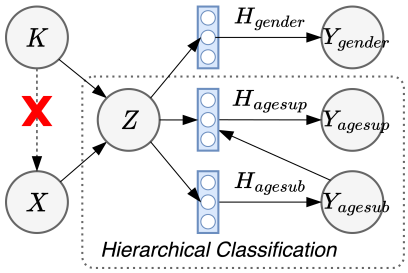


Figure 4: The proposed SCM for hierarchical multi-task classification with causal intervention

$\boldsymbol{X \rightarrow Z \leftarrow K}$**.** The combined representation concatenating X and K for considering the unpredictable label noise was denoted as Z. To pursue this causality, conditional causal intervention $P(Y(X)|do(X))$ was used instead of $P(Y(X)|(X))$ [12]. Notably, this relationship existed due to the independent nature of X and K. Although they originated from the same speech, they were trained separately by different models, teachers, and students.

$\boldsymbol{Z \rightarrow H \rightarrow Y}$**.** This relationship denoted obvious causality; the hidden representation H for each task was predicted based on the shared representation Z. In this study, X could not directly affect the labels Y because the proposed method learned the causal representation Z from observational data, ensuring invariant causal mechanisms between the causal representation and the task labels Y across various tasks. Therefore, Z mediated X and Y via the path $X \rightarrow Z \rightarrow H \rightarrow Y$. The classification was affected by the task-specific representation H through the mediation Z.

$\boldsymbol{Y \rightarrow H \rightarrow Y}$**.** The SCM had direct and indirect causal effects following the path $Y_{agesub} \rightarrow H_{agesup}$ and $Y_{agesub} \rightarrow Y_{agesup}$, respectively. This relationship comprised (1) the weighted summation of the classification loss, L_{CE} & L_{asl}, and (2) hierarchical dependency loss L_{dep}. The cross-entropy loss L_{CE} was used to predict the gender label providing stable performance. However, moving beyond the age group classification for speech, additional constraints were enforced to tackle high negative–positive imbalance for age group label and ground-truth mislabeling issues in classification. Therefore, the single label version of asymmetric loss L_{asl} was used, defined in Equation 4, where p is the output probability of the network [23]. The adjusted probability p_m instituted a hard thresholding mechanism, effectively discarding samples characterized by exceedingly low probabilities. The probability margin $m \geq 0$ was a hyperparameter, and γ denoted the focusing parameter when $\gamma = 0$ yielded binary cross-entropy. The dependency loss, being hierarchy-related, acted as a penalty when predictions were misaligned with a higher hierarchy, specifically agesup. The dependency loss for the super class is given by Equation 5. Here, D and I indicate hierarchy conflicts. Specifically, D_{agesup} is set to 1 when $\hat{y}_{agesup} \neq \hat{y}_{agesub}$, otherwise it is 0. I_{agesup} is 1 if $\hat{y}_{agesup} \neq \hat{y}_{agesup}$, otherwise it is 0. Similarly, I_{agesub} is 1 if $\hat{y}_{agesub} \neq \hat{y}_{agesub}$, otherwise it is 0.

$$\mathcal{L}_{asl} = \left\{ \begin{array}{l} \mathcal{L}_{+} = (1-p)^{\gamma_{+}} \log p \\ \mathcal{L}_{-} = (p_m)^{\gamma_{-}} \log (1-p_m) \quad \text{when} \quad p_m = max(p-m, 0) \end{array} \right\} \tag{4}$$

$$\mathcal{L}_{dep} = -(\mathcal{L}_{agesub})^{D_{agesup} I_{agesub}} \cdot (\mathcal{L}_{agesup})^{D_{agesup} I_{agesup}} \tag{5}$$

4 Experiments and results

Datasets In this study, three AI-Hub open datasets and one in-house call center dataset were used. The wav samples were filtered to retain only those with durations ranging from 1 to 20 s, inclusive. FCVG contained 1,673,214 utterances from 1,958 speakers. The dataset volume was the same in FCVE and FCVG, but only the studio-recorded dataset was used, which contained 121,973 utterances from 112 speakers. In these two datasets, data were originally recorded at a 16 kHz sampling rate and stored in wav format with accompanying age metadata. To ensure consistency with the call-center domain data, a meticulous downsampling process was followed to transform the data to an 8 kHz sampling rate with a 16-bit depth. LTV contained a total of 1,235,302 utterances and speakers were categorized into age groups, such as 10's and 20's, when exact age data were unavailable. The in-house call-center dataset contained 260,344 utterances. All the datasets were executed in an 80-10-10 split. The outcomes presented in Table 1 pertain to the analyses conducted on each evaluation subset.

Experimental settings For data pre-processing, each utterance was converted into 80-dimensional Fbank features. The hidden size of the embeddings from the encoders was 256 after the extraction. For all models, the SpeechBrain framework was used to build the neural networks. The multitask classifiers consisted of a linear dense block with BN, Leaky ReLU, and Dropout. All models were trained with batch sizes of 12 and 10 epochs. The initial learning was 1e-4, and a step decay of rate 0.8 was employed for every 4 epochs from the 2nd to the 10th epoch. The Adam optimizer was employed, and the CE loss was used with a smoothing parameter of 0.1. The evaluation metrics were precision (P), recall (R), and F1 scores (F1).

Baselines For comparison, two different types of models using ECAPA-TDNN[24] and ResNet[25] as encoders were considered to craft teacher–student pairs as: ECAPA-TDNN (residual block size: 8→4) and ResNet (layers: 34→18).

Effects of mixup-based KD (MKD) The baselines trained on FCVG were chosen as the teacher model, and the size of the student model was approximately 50 % of the teacher model. The results

Table 1: Experimental results on FCVG, FCVE, LTV, in-house dataset

Target label			Gender			Agegroup (super class)		
Models	Dataset	Methods	P	R	F1	P	R	F1
ECAPA	FCVG†	base	99.85	99.83	99.84	96.67	96.90	96.78
ResNet	(pre-training)	base	99.86	99.86	99.86	94.13	94.12	94.11
ECAPA	LTV† +FCVE†	base	97.13	97.14	97.13	79.80	68.07	70.01
		+ MKD	97.03	96.76	96.89	81.35	66.14	67.43
		(cos↑ mse↓)	(-0.22)	(-0.21)	(-0.22)	(+0.01)	(-0.05)	(-0.68)
		+ CH	97.35	97.35	97.35	83.45	83.13	81.82
		+ MKD + CH	**98.42**	**98.42**	**98.42**	**86.63**	**86.13**	**85.09**
ResNet	LTV† +FCVE†	base	95.34	95.26	95.26	73.36	73.42	73.00
		+ MKD (cos)	96.11	96.07	96.07	74.78	70.46	73.05
		+ CH	97.55	97.56	97.55	82.38	81.15	77.66
		+ MKD + CH	**98.95**	**98.96**	**98.96**	**88.40**	**87.69**	**87.83**
ECAPA	in-house call center dataset ‡	base (F/T)	96.83	96.82	96.81	62.30	62.88	60.01
		+ CH	98.32	98.29	98.29	75.13	**70.68**	70.15
		+ MKD + CH	**98.59**	**98.59**	**98.59**	**76.04**	70.65	**70.48**

† This paper used datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All data information can be accessed through 'AI-Hub'. (`www.aihub.or.kr`)

– FCVG; Free Conversation Voice (General men and women); 자유대화 음성(일반남여)

– FCVE; Free Conversation Voice (Elderly men and women); 자유대화 음성(노인남여)

– LTV; Low-quality Telephone network Voice recognition data; 저음질 전화망 음성인식 데이터

‡ This is a operational data harvested from the company's in-house call center services.

listed in Table 1 show that MKD retains a substantial part of the performance metrics of the teacher model, although its student model is smaller in size. Thus, MKD successfully encouraged the student model to mimic the teacher's representation, even in the difficulties of the merged LTV and FCVE datasets and the small-scale model. Furthermore, the experiments indicated the potential of MKD in implementing a real-world audio environment with label noise when combined with KD. High-quality speech information was helpful for creating shared information for training three different speech-related tasks.

Effects of Causal approach to Hierarchical multi-task learning (CH) According to Table 1, the proposed method (MKD+CH) significantly improves the performance on all datasets. Particularly, the hierarchical multitask classifier exhibits a larger gain, indicating the strong ability of the method to tackle ambiguous boundaries of age groups and assist classifiers to maximally discriminate the target group in the hierarchical structure. The trained subclass classification model smoothened the superclass classification; thus, dependency punishment forces were used to learn structural information from the hierarchy.

5 Conclusion and limitations

In this study, an end-to-end speech classification model that learns robust representations by enforcing unseen causalities between shared representations and target labels is proposed. Initially, MKD is used with causal intervention, and then a causal approach to hierarchical MTL is proposed to learn hierarchical information without spurious correlations between age groups. The proposed approaches empirically achieve improvements in speech classification, particularly for agesup classification with unseen ambiguous boundaries. Limited to age group information in open-set corpora, the proposed method can only be tested by dividing the age group in increments of 10 years and finding a unidirectional pattern $Y_{agesub} \rightarrow Y_{agesup}$ in the hierarchy. Moreover, the method requires discovery of causalities with other tasks, such as gender, to truly understand speech features and create an invariant representation for speech.

References

[1] Shareef Babu Kalluri, Deepu Vijayasenan, and Sriram Ganapathy. Automatic speaker profiling from short duration speech data. *Speech Communication*, 121:16–28, 2020.

[2] Tessa Bent and Rachael F Holt. Representation of speech variability. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(4):e1434, 2017.

[3] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. Noise-tolerant fair classification. *Advances in neural information processing systems*, 32, 2019.

[4] Ftoon Abu Shaqra, Rehab Duwairi, and Mahmoud Al-Ayyoub. Recognizing emotion from speech based on age and gender using hierarchical models. *Procedia Computer Science*, 151:37–44, 2019.

[5] Zakariya Qawaqneh, Arafat Abu Mallouh, and Buket D Barkana. Age and gender classification from speech and face images by jointly fine-tuned deep neural networks. *Expert Systems with Applications*, 85:76–86, 2017.

[6] Zakariya Qawaqneh, Arafat Abu Mallouh, and Buket D Barkana. Deep neural network framework and transformed mfccs for speaker's age and gender classification. *Knowledge-Based Systems*, 115:5–14, 2017.

[7] Ruben Zazo, Phani Sankar Nidadavolu, Nanxin Chen, Joaquin Gonzalez-Rodriguez, and Najim Dehak. Age estimation in short speech utterances based on lstm recurrent neural networks. *IEEE Access*, 6:22524–22530, 2018.

[8] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[9] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[10] Mengyue Yang, Xinyu Cai, Furui Liu, Weinan Zhang, and Jun Wang. Specify robust causal representation from mixed observations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2978–2987, 2023.

[11] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. icitris: Causal representation learning for instantaneous temporal effects. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

[12] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[13] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020.

[14] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Intervention design for causal representation learning. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

[15] Wei Wang, Junyu Gao, and Changsheng Xu. Weakly-supervised video object grounding via causal intervention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3933–3948, 2022.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[17] Xiang Deng and Zhongfei Zhang. Comprehensive knowledge distillation with causal intervention. *Advances in Neural Information Processing Systems*, 34:22158–22170, 2021.

[18] Feifei Shao, Yawei Luo, Shengjian Wu, Qiyi Li, Fei Gao, Yi Yang, and Jun Xiao. Further improving weakly-supervised object localization via causal knowledge distillation. *arXiv preprint arXiv:2301.01060*, 2023.

[19] Avi Gazneli, Gadi Zimerman, Tal Ridnik, Gilad Sharir, and Asaf Noy. End-to-end audio strikes back: Boosting augmentations towards an efficient audio classification network. *arXiv preprint arXiv:2204.11479*, 2022.

[20] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Contrastive domain adaptation for time-series via temporal mixup. *IEEE Transactions on Artificial Intelligence*, 2023.

[21] Xing Wu, Yifan Jin, Jianjia Wang, Quan Qian, and Yike Guo. Mkd: mixup-based knowledge distillation for mandarin end-to-end speech recognition. *Algorithms*, 15(5):160, 2022.

[22] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[23] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020.

[24] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.